1 Probabilistic preliminaries

These notes are written with my fellow analysis graduate students (as opposed to probability students) in mind. Acquaintance with probability theory is helpful but hopefully not strictly assumed. Folland’s book contains a nice “dictionary” between analysis and probability.

A random variable $X$ is a measurable function from some unseen probability space $(\Omega, \mathcal{P})$ to the real or complex numbers. $\int_{\Omega} X \, d\mathcal{P}$ is written $E_X$. We will talk about three modes of convergence of random variables:

1. $X_n \to X$ almost surely (i.e. almost everywhere),
2. $X_n \to X$ in probability (i.e. in measure),
3. $X_n \to X$ in mean (or “convergence of the mean”). This is not $L^1(\Omega, \mathcal{P})$ convergence! Rather, by this we mean convergence of the integrals $EX_n \to EX$.

While the last may not be familiar from analysis, it is will be very convenient for us to investigate this rather weak mode of convergence. We will see how this can often be upgraded to one of the other two with the tool of concentration of measure.
Finally, given a sequence of “events” (measurable subsets of $\Omega$) $E_n$, we say that $E_n$ holds asymptotically almost surely if they exhaust the measure of $\Omega$. There is actually a hierarchy of how asymptotically almost sure about something one can be! We say that an event $E = E_n$ depending on $n$

1. holds asymptotically almost surely if $P(E) = 1 - o(1)$,
2. holds with high probability if there exist $C, A > 0$ such that $P(E) \geq 1 - Cn^{-A}$, and
3. holds with overwhelming probability if for any $A > 0$ there exists $C_A > 0$ such that $P(E) \geq 1 - C_An^{-A}$.

That is, the probability of the complement of an event holding with high probability has polynomial decay, while complements of events holding with overwhelming probability decay faster than any polynomial. An example of the latter is a sequence of events whose complements decay in probability sub-exponentially.

2 Description of the result

We define a sequence of random Hermitian matrices with size going to infinity as follows:

**Definition** (Wigner ensemble). Let $M(i,j)$, $1 \leq i \leq j$ be a family of independent complex valued mean zero random variables bounded in modulus by $T$, with $M(i,i)$ real for all $i \geq 1$ and $\text{Var}(M(i,j)) = E|M(i,j)|^2 = 1$ for all $i < j$. For each $j < i$, let $M(i,j) = M(j,i)$. For each $n \geq 1$, let $M_n$ be the random matrix with entries $M(i,j)$, $1 \leq i, j \leq n$.

(The uniform bound $T$ on the entries will be convenient but is not necessary. I may include a later section on reducing this hypothesis to a finite moment bound on the entries via a truncation argument.)

**Remark** There are other natural ways to get a random Hermitian matrix: for instance, by taking a rectangular matrix $M$ with all entries independent, and forming the Hermitian matrix $M^*M$ (cf. Wishart ensembles and the Marcenko-Pastur law for their limiting spectral measure). We could also draw a random diagonal matrix $D$ and conjugate it by a random unitary matrix $U$ drawn from Haar measure, or even let one of $D, U$ be deterministic. This example runs backwards from the usual problems of random matrix theory, which are to deduce distributional information on the spectrum from the joint distribution of the entries.

This random matrix has $n$ random eigenvalues on the real line (these random variables will be **dependent**). Can we make some limiting statement about these as $n$ goes to infinity?

What kind of statement could this be? It is not a single random variable we are trying to track, but an increasing number of them. A reasonable task would be to determine the limiting behavior of the spectral edge, that is, the largest eigenvalue $\lambda_1(M_n)$ (this is an ordinary sequence of scalar valued random variables). It turns out that $\lambda_1(M_n)$ grows like $\sqrt{n}$; more precisely, if we set all of the entries above the diagonal to have variance 1, then $\lambda_1(M_n)/\sqrt{n}$ converges to 2 almost surely.
Similarly, $\lambda_n(M_n)/\sqrt{n}$ converges to $-2$ almost surely. Indeed, these results are actually obtained by a refinement of the the “moment method” proof of the semicircular law that we will give presently; see section 4.5

This suggests we track the eigenvalues of the normalized matrix $\frac{1}{\sqrt{n}}M_n$. As $n$ goes to infinity, they will all clutter into the interval $[-2, 2]$. What is their local density or spacing? Getting the joint distribution for the eigenvalues turns out the be hard. Instead we will settle on tracking the “bulk” limiting behavior. We phrase this with duality: if we select a bounded continuous test function $f$, can we identify a limit for the scalar random variable $\frac{1}{n} \sum_{i=1}^{n} f(\lambda_i(\frac{1}{\sqrt{n}}M_n))$? (One then obtains the asymptotic portion of eigenvalues in an interval $I \subset [-2, 2]$ by setting $f$ to be (a regularization of) the indicator function $1_I$.) Note how this weak formulation has succeeded in giving us a sequence of scalar valued random variables, which is more tractable to study than a sequence of vectors in spaces of increasing dimension.

Indeed, we will show that for such $f$,

$$\frac{1}{n} \sum_{i=1}^{n} f(\lambda_i(\frac{1}{\sqrt{n}}M_n)) \longrightarrow \int_{\mathbb{R}} f(x) d\mu_{sc}(x) \quad (1)$$

where $d\mu_{sc}(x) = \frac{2}{\pi} \sqrt{4 - x^2} dx$ is the semicircular measure. The density for this measure shows us the “bulk” limiting distribution of eigenvalues in $[-2, 2]$.

**Remark** We used knowledge of the asymptotics for the spectral edge to select the “right” scaling of $1/\sqrt{n}$ for the eigenvalues to converge into a compact set. It is not necessary to have knowledge of the spectral edge; we’ll see that this scaling emerges naturally in the proof of the semicircular law.

### 3 Statement of result

We can re-express the left hand side of (1) in terms of a (random) probability measure:

**Definition** (Empirical spectral measure (ESM)). For a Hermitian matrix $n \times n$ matrix $M_n$, let $\lambda_1(\frac{1}{\sqrt{n}}M_n) \geq \cdots \geq \lambda_n(\frac{1}{\sqrt{n}}M_n)$ be the ordered (random) eigenvalues of $\frac{1}{\sqrt{n}}M_n$. Define the empirical spectral measure (ESM) to be the random probability measure

$$\mu_n := \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i(\frac{1}{\sqrt{n}}M_n)}$$

**Theorem 3.1** The sequence of (random) empirical spectral measures of a sequence of Wigner matrices $M_n$ converge weakly almost surely to the (deterministic) semicircular distribution supported on $[-2, 2]$. That is, for any bounded continuous test function $f$,

$$\int_{\mathbb{R}} f d\mu_n \longrightarrow \int_{\mathbb{R}} f d\mu_{sc} = \int_{-2}^{2} f(x) \frac{1}{2\pi} \sqrt{4 - x^2} dx$$

almost surely.
How does one demonstrate weak or vague convergence of a sequence of probability measures \( \mu_n \)? A general pattern is to first restrict attention to certain dense classes of test functions, and to upgrade to general test functions via a continuity theorem. For instance:

1. With the family of test functions \( \{ x \mapsto e^{ixt} : t \in \mathbb{R} \} \) we are considering the Fourier transforms (or characteristic functions) of the measures \( \mu_n \). For this case we have the Levy continuity theorem, one direction of which states that to prove weak convergence of the measures \( \mu_n \), it is enough to verify pointwise convergence of the characteristic functions \( \phi_\mu(t) := \int_\mathbb{R} e^{itx} d\mu(x) \).

2. With the family \( \{ x \mapsto x^k : 0 \leq k \in \mathbb{Z} \} \) is to consider the moments of the measures \( \mu_n \). For this case we have the Carleson continuity theorem, which states that a sequence of sub-Gaussian measures converge weakly if and only if all of their moments converge.

3. With the family \( \{ x \mapsto \frac{1}{x-z} : z \in \mathbb{C} \setminus \mathbb{R} \} \), we are considering the Stieltjes or Cauchy transforms of the measures, for which we have a Stieltjes continuity theorem similar to the Levy continuity theorem.

The Fourier analytic method is the “textbook” method for proving the central limit theorem: that for a measure \( \mu \) with first and second moment conditions \( \int_\mathbb{R} x d\mu(x) = 0, \int_\mathbb{R} x^2 d\mu(x) = 1 \), the measures \( \mu_n := \frac{1}{\sqrt{n}} \mu^{*n} \) converge weakly to the standard Gaussian measure. This is ultimately because the Fourier transform interacts nicely with convolution; stated in terms of random variables, characteristic functions have a nice homomorphism property for sums of independent random variables. These advantages are not present when studying complex exponential moments of spectra of matrices, however, essentially because for matrices it is generally not true that \( e^{A+B} = e^A e^B \).

However, the second and third classes of test functions can both be applied to give two different proofs of the semicircular law.

### 3.1 Reversal of quantifiers

John Garnett asked if we could reverse the order of quantifiers in the theorem; that is, do we have almost surely that \( \mu_n \to \mu_{sc} \) weakly? We sketched an argument for the affirmative that I flesh out here. It will be convenient to phrase this equivalently as vague convergence (the two modes are equivalent since the limit measure is a probability measure):

**Theorem 3.2** Almost surely, for any \( f \in C_c(\mathbb{R}) \) we have that

\[
\int_\mathbb{R} f(x) d\mu_n(x) \to \int_\mathbb{R} f(x) d\mu_{sc}(x).
\]

**Proof** For each \( m \), let \( \{ \phi_k^m \}_{k=1}^\infty \) be a countable dense set in the space of continuous functions supported on \([-m,m]\). For each \( m,k \), let \( E_{m,k} \) be the event in \( \Omega \) that \( (\phi_k^m, \mu_n) \to (\phi_k^m, \mu_{sc}) \). By Theorem 3.1 we have \( P(E_{m,k}) = 1 \) for all \( m,k \). Forming the master event

\[
E = \bigcap_{m \in \mathbb{N}} \bigcap_{k \in \mathbb{N}} E_{m,k}
\]
we have $P(E) = 1$.

Now let $f \in C_c(\mathbb{R})$ and let $\epsilon > 0$. Then there is an $m$ sufficiently large such that $\text{supp}(f) \subset [-m, m]$. We can approximate $f$ to within $\epsilon$ in supremum norm by some $\phi^m_k$. Then on $E$ we have

$$\limsup |(f, \mu_n - \mu_{sc})| \leq \limsup |(\phi^m_k, \mu_n - \mu_{sc})| + 2\epsilon = 2\epsilon.$$  

As $\epsilon > 0$ was arbitrary we obtain the claim. \qed

4 Moment method

4.1 Reduction to consideration of moments

It will be enough to show $\int x^k d\mu_n \longrightarrow \int x^k d\mu_{sc}$

almost surely for all $k \geq 0$. Indeed, let $f$ be a bounded continuous test function on $\mathbb{R}$, let $[-R, R]$ be an interval containing $[-2, 2]$, and let $\delta > 0$. Then by the Weierstrass approximation theorem there exists a polynomial $P_\delta$ such that $\sup_{[-R, R]} |f(x) - P_\delta(x)| < \delta$. Now we have

$$|\int f d\mu_n - \int f d\mu_{sc}| \leq |\int P_\delta(x) d(\mu_n - \mu_{sc})| + \delta + |\int_{[-R, R]} f(x) - P_\delta(x) d\mu_n|.$$  

The first term converges almost surely to zero by assumption. The last term can be bounded

$$|\int_{[-R, R]} f(x) - P_\delta(x) d\mu_n| \leq C \int_{[-R, R]} |x|^p d\mu_n$$  

$$\leq CR^{-p-2q} \int x^{2(p+q)} d\mu_n$$  

for some constant $C$, where $p$ is the degree of $P_\delta$ and $q > 0$ is arbitrary. Again by assumption we have

$$\limsup_{n \to \infty} |\int_{[-R, R]} f(x) - P_\delta(x) d\mu_n| \leq CR^{-p-2q} \int x^{2(p+q)} d\mu_{sc}$$  

$$\leq CR^{-p-2q} 2^{2(p+q)} = CR^{-p-2q(2+q)}(2/R)^{2q}$$  

almost surely. Letting $q$ go to infinity and then letting $\delta$ go to zero we have the result.

Remark We have essentially proved (one direction of) the Carleman continuity theorem: a sequence of sub-Gaussian measures $\mu_n$ on $\mathbb{R}$ converges weakly to another measure $\mu$ if and only if all moments converge. (This is analogous to the well-known Levy continuity theorem that a sequence of measures converge weakly if and only if their Fourier transforms converge pointwise.) The almost sure sub-Gaussian hypothesis can be established for the empirical spectral measures with results on
the almost sure convergence of the spectral edge and concentration of measure. The above proof is a bit more specific to our case, using only that all moments are finite (the sub-Gaussian hypothesis amounts to assuming a certain growth-condition on the moments) and crucially using that the limit measure is supported on a compact set.

4.2 Convergence of the mean

The core of the proof (or the fun part at least) is to show that the expected moments of the ESMs converge to the moments of the semicircular measure. After this is accomplished, some kind of concentration of measure is needed to upgrade this to convergence in probability or almost sure convergence. We have the trace formula

\[
\mathbb{E} \int x^k d\mu_n = \frac{1}{n} \sum_{i=1}^{n} \lambda_i^k \\
= \frac{1}{n} \text{tr} \left( \frac{1}{\sqrt{n}} M_n^k \right) \\
= \frac{1}{n^{k/2+1}} \text{Etr} (M_n^k) \\
= \frac{1}{n^{k/2+1}} \sum_{1 \leq i_1, \ldots, i_k \leq n} \mathbb{E}M(i_1, i_2)M(i_2, i_3) \cdots M(i_k, i_1).
\]

This expression for the trace of a power of a matrix is straightforward to prove by induction. We have hence reexpressed the moments in terms of cycles of length \( k \) on vertices labeled 1 to \( n \). Each step has a weight given by the corresponding matrix entry.

[Draw pictures of walks corresponding to terms in e.g. \( \mathbb{E} \int x^5 d\mu_n \).]

We will proceed by restructuring this sum through a classification of cycles of length \( k \) on vertices labeled 1, \ldots, \( n \). A cycle is specified by a \( k \)-tuple of vertex labels \( i = (i_1, \ldots, i_k) \in \{1, \ldots, n\}^n \). We denote its contribution to the expected moment by

\[
C(i) = \mathbb{E}M(i_1, i_2) \cdots M(i_k, i_1).
\]

We let \( V(i) = \{i_1, \ldots, i_k\} \) be the set of vertices visited by the cycle, without multiplicities. Let \( E(i) = \{(i_1, i_2), \ldots, (i_k, i_1)\} \) be the directed edges / steps with multiplicities, and \( \tilde{E}(i) \) be the list of edges without multiplicity or orientation. We call the undirected graph \( G(i) = (V(i), \tilde{E}(i)) \) the skeleton of the cycle \( i \).

[Again, draw out an example of a cycle, its vertex and edge sets, and the skeleton graph.]

Recall that for independent random variables \( X, Y \) we have \( \mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y) \) (whenever these expressions are defined). By the mean zero hypothesis on the entries \( M(i, j) \) we have that \( C(i) = 0 \) unless for all \( \tilde{e} \in E(i) \), either \( \tilde{e} \) or its reverse appear at least twice in the cycle. Hence we have that \( |\tilde{E}(i)| \leq \lfloor k/2 \rfloor \). It then follows that \( |V(i)| \leq \lfloor k/2 \rfloor + 1 \), where we use the following
**FACT:** Let $G = (V, E)$ be an undirected connected graph. Then $|V| \leq |E| + 1$, with equality if and only if $G$ is a tree.

Finally, we say that two cycles are equivalent if there is a permutation of the underlying vertices that takes one cycle to the other. That is, $i \sim j$ if $\exists$ a bijection $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ such that $(j_1, \ldots, j_k) = (\pi(i_1), \ldots, \pi(i_k))$. Note that two cycles can only be equivalent if they take the same number of steps $k$, and if they visit the same number of distinct vertices $t$.

Let $W_{k,t}$ be a set of equivalence class representatives of cycles of $k$ steps on $t$ distinct vertices. A natural choice for these representatives is provided by cycles on the vertices 1 to $t$ that start at 1, and such that each time the cycle visits a vertex it has not visited before, the label on that vertex is one larger than the largest label of the vertices it has seen already.

We rewrite the sum

$$E \int x^k d\mu_n = \frac{1}{n^{k/2+1}} \sum_{i \in \{1, \ldots, n\}^k} EM(i_1, i_2) \cdots M(i_k, i_1)$$

$$= \frac{1}{n^{k/2+1}} \sum_{t=1}^{|k/2|+1} \sum_{i \in W_{k,t}} \sum_{j \sim i} E \prod_{\vec{e} \in E(j)} M(\vec{e}).$$

This structuring of the sum reflects that a cycle of length $t$ is uniquely determined by selecting $t$ ordered vertices and an equivalence class representative from $W_{k,t}$.

It turns out that most of the terms on the right hand side are $o_{k,n \to \infty}(1)$. Indeed, from the uniform bound $T$ on the entries we have

$$|E \prod_{\vec{e} \in E(j)} M(\vec{e})| \leq T^k.$$ 

We also have that for any cycle $i$ on $t$ vertices, the number of equivalent cycles is the number of ways to choose $t$ vertices from $n$ vertices. Hence

$$|\{j : j \sim i\}| = \frac{n!}{(n-t)!} \sim n^t.$$ 

Putting this together, we can bound the $t$th summand

$$\left| \frac{1}{n^{k/2+1}} \sum_{i \in W_{k,t}} \sum_{j \sim i} E \prod_{\vec{e} \in E(j)} M(\vec{e}) \right| \leq |W_{k,t}| T^k \frac{n^t}{n^{k/2+1}}$$

which is $o_k(1)$ for all $t \leq \lfloor k/2 \rfloor$ when $k$ is even, and all $t \leq \lfloor k/2 \rfloor + 1$ when $k$ is odd. Hence, for $k$ odd we conclude
\[ \mathbb{E} \int x^k d\mu_n \rightarrow 0 \]
as \( n \rightarrow \infty \). Hence we can already see the symmetry of the limit measure about the \( y \)-axis. For \( k \) even, we are left with
\[
\mathbb{E} \int x^k d\mu_n = \frac{1}{n^{k/2+1}} \sum_{i \in W_{k/2+1}} \sum_{j \sim i} \mathbb{E} \prod_{\vec{e} \in E(j)} M(\vec{e}) + o_k(1).
\]
The cycles in this sum are of \( k \) steps, visiting \( k/2 + 1 \) vertices, with each edge in the skeleton graph traversed at least twice (in some direction); again by the FACT the skeleton graphs are trees on \( k/2 + 1 \) vertices with \( k/2 \) edges that are traversed exactly twice each, once in each direction (for if an edge were traversed twice in the same direction, there would be an undirected cycle in the skeleton). For the expectation of the product of matrix entries we then have
\[
\mathbb{E} \prod_{\vec{e} \in E(j)} M(\vec{e}) = \mathbb{E} \prod_{e \in E(j)} M(e)M(e) = \prod_{e \in E(j)} \mathbb{E}|M(e)|^2 = 1
\]
where we have used the hypotheses that the matrix is Hermitian, and the above-diagonal entries are independent with variance 1. Hence we have
\[
\mathbb{E} \int x^k d\mu_n = \frac{1}{n^{k/2+1}} |W_{k/2+1}| \frac{n!}{(n - (k/2 + 1))!} + o_k(1) = |W_{k/2+1}| + o_k(1).
\]
It follows that, asymptotically, the even moments are the ESM are given by the number of equivalence classes of cycles of length \( k \) on \( k/2 + 1 \) vertices. Our next task is to count these.

Remark While the problem of counting cycles of length \( k \) on \( n \) vertices may seem complicated at first, note that much of the complexity is only \( k \)-dependent. We are only interested in the asymptotics for \( n \) large, so crude bounds on the \( k \)-complexity were sufficient.

4.3 Catalan numbers
In this section we will show that for each \( k \geq 1 \),
\[ |W_{2k,k+1}| = \int x^{2k} d\mu_{sc}. \]

Lemma 4.1 The equivalence class representatives for walks of length \( 2k \) on \( k + 1 \) vertices are in bijection with Dyck paths, which are walks on the nonnegative integers starting and ending at 0, and which move one unit up or down at each step.
Proof Indeed, at each step of one of our equivalence class representatives, the options are to (1) go to the next unexplored vertex (which has been determined already as we are counting equivalence classes of cycles, not cycles), or (2) return to a vertex already visited along an edge that has been used once, and there can only be one such edge. We encode these options as $+1$ and $-1$ in a walk on the integers starting at 0. Since the cycle ends at the starting vertex, the walk on the integers ends at 0. A step to $-1$ from 0 would mean there is an edge leading away from 0 that has been traversed once before. Since the walk started at 0, this would imply there is a cycle in the skeleton graph, which is a contradiction. We have hence described how to draw a Dyck path from one of our equivalence class representatives. Conversely, given a Dyck path, we start a cycle from the vertex 1 and move to the next unvisited vertex each time the Dyck path moves up 1, and move back along the available edge each time the Dyck path moves down 1. We end at the first vertex. □

Lemma 4.2 Let $C_k$ denote the number of Dyck paths of length $2k$. We have that

$$C_k = \frac{1}{k+1} \binom{2k}{k}.$$  

(The $C_k$ are called Catalan numbers.)

Proof (Picture-aided.) First note that the number of walks of length $2k$ on the integers starting and ending at 0 is $\binom{2k}{k}$. For each walk that is not a Dyck path, let $m$ be the first step the walk is at $-1$ (where at the 0th step the walk is at 0). If we denote by $w(i)$ the location of the walk at the $i$th step, we form a new walk $\tilde{w}$ by setting $\tilde{w}(i) = w(i)$ for each $i > m$, and $\tilde{w}(i) = -1 - w(i)$ for each $0 \leq i \leq m$. Then $\tilde{w}(2k) = -1$. Hence, the walks that are not Dyck paths are in bijection with walks that start at 0 and end at $-1$, of which there are $\binom{2k}{k+1}$. It follows that the number of Dyck paths is $\binom{2k}{k} - \binom{2k}{k+1} = C_k$. □

Lemma 4.3 For each $k$ even,

$$\int x^k d\mu_{sc} = C_{k/2}$$

and for each $k$ odd, the integral is 0.

Proof Trig substitution. □

4.4 Almost sure convergence of moments to Catalan numbers

Let $m_k$ denote the moments of the semicircular distribution. Our goal here is to show that

$$|\int x^k d\mu_n - m_k| \rightarrow 0$$

almost surely. By the Borel-Cantelli lemma it will be sufficient to show that the probabilities of the bad events are summable:

$$\sum_{n=1}^{\infty} P(|\int x^k d\mu_n - m_k| > \epsilon) < \infty.$$
We can use the triangle inequality and our result on convergence of the moments in mean to bound
the summands on the left hand side by
\[ P(|\int x^k d\mu_n - E \int x^k d\mu_n| + |E \int x^k d\mu_n - m_k| > \epsilon) \leq P(|\int x^k d\mu_n - E \int x^k d\mu_n| > 2\epsilon) \]
for all \( n \) sufficiently large. By Chebyshev’s inequality, it will now suffice to get a bound on the
variance of the random variable \( \int x^k d\mu_n \) that is summable in \( n \).

Expanding out the variance with the trace formula, we have
\[
\text{Var} \int x^k d\mu_n = E(\int x^k d\mu_n)^2 - (E \int x^k d\mu_n)^2
\]
\[
= \frac{1}{n^{k+2}} \sum_{i,j \in \{1,\ldots,n\}^k} C(i,j) - C(i)C(j)
\]
where as before
\[
C(i) = EM(i_1, i_2) \cdots M(i_k, i_1)
\]
and
\[
C(i, j) = EM(i_1, i_2) \cdots M(i_k, i_1)M(j_1, j_2) \cdots M(j_k, j_1)
\]
is the contribution from the two cycles of length \( k \) executed one after the other.

Now if the cycles \( i \) and \( j \) do not share any edges, by independence we have \( C(i, j) = C(i)C(j) \), and
the summands vanish. Hence, the only pairs of cycles that contribute to the sum share at least
one edge. It follows that the number of vertices the pair of cycles can visit is bounded by two less
than the sum of the largest number of vertices they can individually visit without the expressions
vanishing (which we previously found to be \( k/2+1 \)). This gives a bound of \( n^k \) for the combinatorial
weight of pairs of cycles that contribute a nonzero expectation to the sum. Hence,
\[
\text{Var} \int x^k d\mu_n = O(n^{-2})
\]
and by Chebyshev’s inequality
\[
P(|\int x^k d\mu_n - E \int x^k d\mu_n| > \epsilon) = O_\epsilon(n^{-2})
\]
which is summable for each \( \epsilon > 0 \). The moment method proof of Theorem 3.1 is complete.

4.5 Moment method bounds for convergence of the operator norm

Understanding the asymptotics of the moments of the ESM is also a first step towards understanding
the asymptotics of the spectral edge \( ||M_n||_{op} \). Indeed, while the starting point for the moment
method was the identity
the starting point for studying the spectral edge is the family of inequalities
\[ \text{tr}(M_n^k) \leq ||M_n||_{op}^k \leq n\text{tr}(M_n^k) \]
for \( k \) ranging over positive even integers. Rearranging we have
\[ n^{-1/k}\text{tr}(M_n^{1/k}) \leq ||M_n||_{op} \leq \text{tr}(M_n^{1/k}). \]
We would like to optimize this family of inequalities by taking \( k \) slowly increasing with \( n \). Inserting the asymptotic value \( n^{k/2+1}C_{k/2} \) for \( \text{tr}(M_n^k) \) and using Stirling’s formula on \( C_{k/2} \) suggests that we can show
\[ \frac{1}{\sqrt{n}}||M_n||_{op} \to 2 \]
in some sense. However, in taking \( k \) increasing, we need to mind the \( k \)-dependent error term in our asymptotic that
\[ \text{tr}(M_n^k) = n^{k/2+1}(C_{k/2} + o_{k,n\to\infty}(1)) \]
with high probability (one can show this actually holds with overwhelming probability using Talagrand’s inequality on the Schatten \( k \)-norms of \( M_n \)). Hence, to proceed, we need to establish reasonable \( k \)-dependence for the error term in the asymptotic for moments through a more careful investigation of cycles that don’t walk around fat trees. See for instance [Ta2012].

5 Stieltjes transform method

Much of our treatment of this method closely follows that of [Ta2012].

For the Stieltjes transforms of the ESMs we have
\[ s_n(z) := \int_{\mathbb{R}} \frac{1}{x-z}d\mu_n(z) = \frac{1}{n}\text{tr}(\frac{1}{\sqrt{n}}M_n - zI_n)^{-1} \]
for any \( z \in \mathbb{C} \setminus \text{supp}(\mu_n) \). We first review some general properties of Stieltjes transforms of (random and deterministic) probability measures on \( \mathbb{R} \), and prove the Stieltjes continuity theorem, which will allow us to deduce Theorem 3.1 from almost sure convergence of the above random variable to the Stieltjes transform of the semicircular measure.

5.1 Properties of the Stieltjes transform

Let \( \mu \) be a probability measure on \( \mathbb{R} \). In this section we denote the Stieltjes transform
\[ s_\mu(z) = \int_{\mathbb{R}} \frac{1}{x-z}d\mu(x) \]
which is defined for any \( z \in \mathbb{C} \setminus \text{supp}(\mu) \), and in particular for any \( z \) in the upper or lower half plane. We list some basic properties:
Proposition 5.1 Let $\mu$ a probability measure on $\mathbb{R}$ and let $s_\mu(z)$ denote its Stieltjes transform. Then we have

1. $s_\mu(z) = s_\mu(\bar{z})$
2. $|s_\mu(z)| \leq \frac{1}{|\text{Im}z|}$ pointwise
3. $s_\mu(z) = \frac{-1+o_\mu(1)}{z}$, where the asymptotic $o_\mu(1)$ is for $z$ approaching infinity non-tangentially in the upper (or lower) half plane (viewed as a wedge of the sphere); i.e., $z$ approaches infinity with $|\text{Re}z|/|\text{Im}z|$ bounded.
4. $s_\mu(z)$ is complex analytic on the upper and lower half planes. (Proof: Morera’s theorem, using DCT and Fubini.)
5. (regularity estimates) $|\frac{d}{dz}s_\mu(z)| = O_{j}(\frac{1}{|\text{Im}z| j+1})$. (Proof: DCT.)
6. $\text{Im} s_\mu(\lambda + i\epsilon) = \pi \mu \ast P_\epsilon(\lambda)$, where $P_\epsilon(x) = \frac{1}{\pi \sqrt{x^2 + \epsilon^2}}$ is the Poisson kernel, or Cauchy distribution, with parameter $\epsilon$. In particular we see that $s_\mu$ maps the upper half plane to the upper half plane (which with analyticity means that $s_\mu$ is a Herglotz function).

Proof 1. and 2. are immediate from examination of the integrand.

From 6., a standard approximate identity argument gives us a way to recover the measure $\mu$ from the function $s_\mu$:

Proposition 5.2 (Inversion formula)

$\text{Im} s_\mu(\lambda + i\epsilon)d\lambda \rightarrow \pi d\mu(\lambda)$

as $\epsilon \rightarrow 0^+$ in the vague topology.

We give a short probabilistic argument for the interested analyst:

Proof Let $X$ be a random variable with distribution $\mu$ and $\{Y_\epsilon : \epsilon > 0\}$ a family of random variables independent of $X$ with Cauchy($\epsilon$) distribution. Then $X + Y_\epsilon$ converges in probability to $X$ as $\epsilon \rightarrow 0^+$, and hence converges in distribution, which the same as vague convergence of the distribution $\mu \ast P_\epsilon(\lambda)$ to $\mu$. \qed

Some of the above properties help us establish

Theorem 5.3 (Stieltjes continuity theorem) Let $\mu_n$ be a sequence of probability measures on $\mathbb{R}$.

1. If $\mu_n$ converges weakly to another probability measure $\mu$ on $\mathbb{R}$, then $s_{\mu_n}(z) \rightarrow s_\mu(z)$ for all $z \in \mathbb{C} \setminus \mathbb{R}$.
2. If $s_{\mu_n}(z) \rightarrow s(z)$ for all $z \in \mathbb{C} \setminus \mathbb{R}$, then $s$ is the Stieltjes transform of a sub-probability measure $\mu$, and $\mu_n \rightarrow \mu$.
3. If for a sequence of random probability measures $\mu_n$ on $\mathbb{R}$ we have $s_{\mu_n}(z) \rightarrow s_\mu(z)$ almost surely for all $z \in \mathbb{C} \setminus \mathbb{R}$ for some deterministic probability measure $\mu$, then $\mu_n \rightarrow \mu$ weakly almost surely.
Proof 1. is immediate from $x \mapsto 1/(x-z)$ bounded and continuous for each $z$ off the real line.

For 2., let $\mu_{n_k}$ be a subsequence. Then we have a subsubsequence $\mu_{n_{k_j}} \rightarrow \mu$ for some sub-probability measure $\mu$ by the variously named Banach/Alaoglu/Helly Selection theorem. It follows from $\text{Re}1/(x-z), \text{Im}1/(x-z) \in C_0(\mathbb{R})$ for all $z$ off $\mathbb{R}$ that $s_{\mu_{n_{k_j}}}(z) \rightarrow s_\mu(z)$ for all $z \in \mathbb{C}\setminus\mathbb{R}$. Hence we have $s(z) = s_\mu(z)$ for any $\mu$ that is a subsubsequential limit of the sequence $\mu_n$. It follows from the Inversion formula (5.2) that all subsubsequential limits are the same sub-probability measure (call it $\mu$), and hence that $\mu_n \rightarrow \mu$.

For 3., we first define a metric on probability measures. Fix a sequence $z_m \rightarrow z_0$ in $\mathbb{C}\setminus\mathbb{R}$ with $z_m \neq z_0$ for all $i$. For probability measures $\nu_1, \nu_2$ on $\mathbb{R}$, let

$$\rho(\nu_1, \nu_2) = \sum_m \frac{1}{2^m} |s_{\nu_1}(z_m) - s_{\nu_2}(z_m)|.$$

Then for a sequence of probability measures $\nu_n$ and $\nu$, $\rho(\nu_n, \nu) \rightarrow 0$ implies that $\nu_n \rightarrow \nu$ weakly. Indeed, $\rho(\nu_n, \nu) \rightarrow 0$ implies that $s_{\nu_n}(z_m) \rightarrow s_\nu(z_m)$ for all $m$. Now for a subsequence $\nu_{n_k}$ there is a subsubsequence $\nu_{n_{k_j}} \rightarrow \theta$ for some sub-probability measure $\theta$. Vague convergence implies $s_{\nu_{n_{k_j}}}(z) \rightarrow s_\theta(z)$ for all $z \in \mathbb{C}\setminus\mathbb{R}$. Hence for any $\theta$ thus obtained, $s_\theta = s_\nu$ on a set with an accumulation point, and hence $s_\theta(z) = s_\nu(z)$ for all $z \in \mathbb{C}\setminus\mathbb{R}$. By the inversion formula we have that any such $\theta$ is equal to $\nu$, and hence that for the entire sequence we have $\nu_n \rightarrow \nu$. Since $\nu$ is a probability measure we have $\nu_n \rightarrow \nu$ weakly.

Now we turn to our sequence $\mu_n$ of random probability measures. From the assumption $s_{\mu_n}(z) \rightarrow s_\mu(z)$ almost surely, we have in particular that $s_{\mu_n}(z_m) \rightarrow s_\mu(z_m)$ for all $m$. For each $m$, let $E_m$ be the event $\{s_{\mu_n}(z_m) \rightarrow s_\mu(z_m)\} \subset \Omega$, and let $E = \bigcap_m E_m$. Then $P(E) = 1$, and on $E$ we have $\rho(\mu_n, \mu) \rightarrow 0$ by the uniform boundedness of $s_{\mu_n}(z_m)$ by $1/\delta_0$ where $\delta_0 = \inf_m |\text{Im}z_m|$. So on $E$, $\mu_n \rightarrow \mu$ weakly, and we have the desired result. \qed

Remark A quantitative Stieltjes continuity theorem is given in appendix C of [TaVu2011].

5.2 Concentration of measure tools

We use this section to state without proof three concentration of measure inequalities that will be useful in the next section.

Theorem 5.4 (Hoeffding’s inequality)
Let $Y_1, \ldots, Y_n$ be independent random variables, with each $Y_i$ bounded in magnitude by some $b_i < \infty$. Let $S_n = \sum_{i=1}^n Y_i$. Then

$$P(|S_n - \mathbb{E}S_n| \geq \epsilon) \leq Ce^{-c\epsilon^2/|\bar{b}|^2}$$

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

Nicholas Cook Two proofs of Wigner’s semicircular law June 5, 2012

13
**Theorem 5.5 (McDiarmid’s inequality)**

Let $F : S_1 \times \cdots \times S_n \to \mathbb{C}$ be a function on a product space which is Lipschitz in the following sense: for each $1 \leq i \leq n$ there exists $b_i < \infty$ such that

$$|F(x_1, \ldots, x_i, \ldots, x_n) - F(x_1, \ldots, x'_i, \ldots, x_n)| \leq b_i$$

for any $(x_1, \ldots, x_n) \in S_1 \times S_n$ and any other $x'_i \in S_i$. That is, varying only the $i$th coordinate can only change $F$ by $b_i$.

Let $Y_1, \ldots, Y_n$ be independent random variables taking values in $S_1, \ldots, S_n$. Then for any $\epsilon > 0$ we have

$$\mathbb{P}(|F(Y) - \mathbb{E}F(Y)| \geq \epsilon) \leq C e^{-\alpha^2/|\vec{b}|^2}$$

for some absolute constants $C, \alpha > 0$.

(One may notice the above implies Hoeffding’s inequality; indeed, the former can be viewed as a tensorization of the latter.)

**Theorem 5.6 (Talagrand’s inequality)**

Let $F : \mathbb{C}^n \to \mathbb{C}$ be a convex 1-Lipschitz function. Let $X = (X_1, \ldots, X_n)$ be a random vector in $\mathbb{C}^n$ with independent components of magnitude $O(1)$. Let $\text{MF}(X)$ be a median for the random variable $F(X)$. Then for any $\epsilon > 0$ we have

$$\mathbb{P}(|F(X) - \text{MF}(X)| \geq \epsilon) \leq C e^{-\alpha^2}$$

for some absolute constance $C, \alpha > 0$.

**5.3 A recurrence relation**

Fix $z = a + ib$ in the upper half plane. All subsequent asymptotics will depend on $z$, and this dependence will be suppressed (the dependence on $z$ becomes important in further extensions of the theory, such as the local semicircular law, where $z$ is allowed to be of size $n^{-1} \log^O(1)/n$).

In this section we will derive a recurrence relation for

$$\mathbb{E}s_n(z) = \mathbb{E}\frac{1}{n} \text{tr}(\frac{1}{\sqrt{n}}M_n - zI_n)^{-1}.$$

Along the way we’ll show (more than sufficient) concentration of measure for $s_n(z)$ to upgrade convergence of the mean to almost sure convergence. By our Stieltjes continuity theorem this will complete the proof.

For sake of simplicity of the algebra we take the entries to be identically distributed. With this assumption we can permute rows and columns under the expectation to see that

$$\mathbb{E}[(\frac{1}{\sqrt{n}}M_n - zI_n)^{-1}]_{ii} = \mathbb{E}[(\frac{1}{\sqrt{n}}M_n - zI_n)^{-1}]_{nn}$$
for each $1 \leq i \leq n$. Hence,

$$\mathbb{E}s_n(z) = \mathbb{E}[(\frac{1}{\sqrt{n}} M_n - zI_n)^{-1}]_{nn}.$$ 

Our next step is to use a linear algebraic identity:

**Proposition 5.7** Let $A_n$ be an $n \times n$ invertible matrix, such that the upper-left $n - 1 \times n - 1$ minor $A_{n-1}$ is also invertible. The we have the identity

$$A_n^{-1} = \frac{1}{a_{nn} - Y_n^* A_{n-1}^{-1} X_n}$$

where $a_{nn}$ is the bottom-left entry of $A$, $Y_n \in \mathbb{C}^{n-1}$ is the (transpose of the) last row of $A_n$ without the final entry $a_{nn}$, and $X_n \in \mathbb{C}^{n-1}$ is the last column of $A_n$ without the final entry $a_{nn}$.

**Proof** Solve $A_n v = e_n$, where $e_n$ is the nth standard basis vector, for the last entry of $v$, after writing $v = (v', v_n) \in \mathbb{C}^{n-1} \times \mathbb{C}$ and $A_n$ in blocks of $A_{n-1}, X_n, Y_n, a_{nn}$.

Applying the above proposition with $A_n = \frac{1}{\sqrt{n}} M_n - zI_n$, we have

$$\mathbb{E}s_n(z) = \mathbb{E}\left[\frac{-1}{z + \frac{1}{n} X_n^* R_{n-1} X_n - \frac{1}{\sqrt{n}} \xi_{nn}}\right]$$

where $X_n \in \mathbb{C}^{n-1}$ is the last column of $M_n$ with last entry removed, and

$$R_{n-1} := (\frac{1}{\sqrt{n}} M_{n-1} - zI_{n-1})^{-1}.$$ 

(Note that $\frac{1}{\sqrt{n}} M_{n-1} - zI_{n-1}$ is invertible by $\text{Im} z > 0$.)

We claim that

$$\frac{1}{n} X_n^* R_{n-1} X_n = \mathbb{E}s_n(z) + o(1)$$

with overwhelming probability. (Recall the hierarchy of how asymptotically almost sure we can be about an event depending on $n$.)

We will actually argue each step in the following chain of approximations:

$$\frac{1}{n} X_n^* R_{n-1} X_n \approx \mathbb{E}(\frac{1}{n} X_n^* R_{n-1} X_n | M_{n-1})$$  \hspace{1cm} (3)

$$= \frac{1}{n} \text{tr} R_{n-1}$$  \hspace{1cm} (4)

$$\approx s_{n-1}(z)$$  \hspace{1cm} (5)

$$\approx s_n(z)$$  \hspace{1cm} (6)

$$\approx \mathbb{E}s_n(z)$$  \hspace{1cm} (7)
where we understand \( LHS \approx RHS \) to mean \( LHS = RHS + o(1) \) with overwhelming probability ((5) and (6) actually hold asymptotically deterministically).

We prove these in the order (4) (5) (6) (7) (3), saving the concentration of measure statements (7) and (3) for last.

**Proof of (4)**

\[
\mathbf{E}(\frac{1}{n}X_n^*R_{n-1}X_n|M_{n-1}) = \frac{1}{n} \sum_{i,j=1}^{n-1} \mathbf{E}(\overline{\xi_{in}(R_{n-1})_{ij}\xi_{jn}}|M_{n-1}) 
\]

\[
= \frac{1}{n} \sum_{i,j=1}^{n-1} (R_{n-1})_{ij} \mathbf{E}(\overline{\xi_{in}\xi_{jn}}|M_{n-1}) 
\]

\[
= \frac{1}{n} \sum_{i,j=1}^{n-1} (R_{n-1})_{ij} \mathbf{E}(\overline{\xi_{in}\xi_{jn}}) 
\]

\[
= \frac{1}{n} \sum_{i,j=1}^{n-1} (R_{n-1})_{ij} \delta_{ij} 
\]

\[
= \frac{1}{n} \text{tr} R_{n-1} 
\]

□

**Proof of (5)**

First note

\[
\frac{1}{n} \text{tr} R_{n-1} = \frac{1}{n} \left( \frac{1}{\sqrt{n}}M_{n-1} - zI_{n-1} \right)^{-1} 
\]

\[
= \sqrt{\frac{n-1}{n}} s_{n-1}(\sqrt{\frac{n}{n-1}z}). 
\]

Now by the regularity estimates for the Stieltjes transform in Proposition (5.1),

\[
\sqrt{\frac{n-1}{n}} s_{n-1}(\sqrt{\frac{n}{n-1}z}) = (1 + O(1/n))s_{n-1}(z)(1 + O(1/n)) 
\]

\[
= s_{n-1}(z) + O(1/n). 
\]

So we have \( \frac{1}{n} \text{tr} R_{n-1} = s_{n-1}(z) + O(1/n) \) almost surely. □

We state (6) as a lemma:

**Lemma 5.8 (Predecessor comparison)**

\[
s_n(z) = s_{n-1}(z) + O(1/n). 
\]

(deterministically).
Proof We consider the expression
\[
\text{tr} \left( \frac{1}{\sqrt{n}} M_n - z I_n \right)^{-1} - \text{tr} \left( \frac{1}{\sqrt{n}} M_{n-1} - z I_{n-1} \right)^{-1}.
\] (8)

On the one hand we can rewrite this using the spectral theorem as
\[
\sum_{j=1}^{n} \frac{1}{\sqrt{n}} \lambda_j(M_n) - z - \sum_{j=1}^{n-1} \frac{1}{\sqrt{n}} \lambda_j(M_{n-1}) - z.
\] (9)

An important quick corollary of the Courant-Fischer minimax formula for the eigenvalues of a Hermitian matrix \(M_n\) is the Cauchy interlacing law, that the eigenvalues \(\lambda_1(M_{n-1}) \geq \cdots \geq \lambda_{n-1}(M_{n-1})\) of \(M_{n-1}\) intersperse the eigenvalues \(\lambda_1(M_n) \geq \cdots \geq \lambda_n(M_n)\) of \(M_n\). That is, for each \(1 \leq j \leq n-1\),
\[
\lambda_j(M_n) \geq \lambda_j(M_{n-1}) \geq \lambda_{j+1}(M_n).
\]
(See section 1.3 of [Ta2012].) Hence, (9) is an alternating sum of evaluations of the function
\[
f_z(x) = \frac{1}{x - z}.
\]

For \(z = a + ib\) fixed with \(b > 0\), this function has total variation \(O(1)\) (including the \(z\) dependence it has total variation \(O(\frac{|a|+|b|}{b^2 n})\)). So (8) is \(O(1)\).

On the other hand we can rewrite (8) as
\[
ns_n(z) - \text{tr}R_{n-1}.
\]
The second term was shown in the proof of (5) to be \(s_{n-1}(z) + O(1/n)\). Putting all of this together we have proved the lemma.

The lemma gives a stability for the Stieltjes transform of a matrix that one may find surprising: the last row and column of \(M_n\) can only influence the value of \(s_n(z)\) by \(O(1/n)\), even if they contain unbounded entries. Note also that this holds deterministically. This same proof shows that the above holds for the Stieltjes transform of any \(n - 1 \times n - 1\) minor of \(M_n\), so that any single row or column of \(M_n\) can only influence the value of \(s_n(z)\) by \(O(1/n)\).

It is at this point that one should suspect the statistic \(s_n(z)\) of the matrix \(M_n\) to exhibit very strong measure concentration. Indeed, a general principle is that concentration of measure holds for functions that depend in a sufficiently regular way on lots of independent random variables (this principle was explicitly stated in [Ta1995]).

Reaching for our concentration of measure toolbox, we find that McDiarmid’s inequality (5.5), at least, fits our needs.

Proof of (7)
For each \(1 \leq i \leq n\), let \(Y_i \in \mathbf{C}^i\) be the first \(i\) entries of the \(i\)th column of \(M_n\). Take \(s_n(z)\), viewed
as a function of the columns of the upper triangle of $M_n$, for $F$ in the statement of McDiarmid’s inequality. Lemma 5.8 (and the subsequent discussion) give us bounds $b_i = O(1/n)$ for the variation of $s_n(z)$ along each coordinate. It follows from McDiarmid’s inequality that

$$P(|s_n(z) - Es_n(z)| \geq \epsilon) \leq Ce^{-cn\epsilon^2}$$

(10)

for possibly different values of $C, c > 0$. Taking $\epsilon$ to be a sufficiently slowly decaying function of $n$ (such as $n^{-0.01}$) gives the claim. □

The above also provides sufficient concentration for $s_n(z)$ to upgrade convergence in mean to almost sure convergence. Indeed, the right hand side of (10) is summable, so from the Borel-Cantelli lemma we have that for any $\epsilon > 0$,

$$\mathbb{P}(|s_n(z) - Es_n(z)| \geq \epsilon \text{ infinitely often}) = 0$$

It follows that $s_n(z)$ and its mean are converging together almost surely.

It remains to establish (3), ie that

$$\frac{1}{n} X_n^* R_{n-1} X_n = \mathbb{E}(\frac{1}{n} X_n^* R_{n-1} X_n | M_{n-1}) + o(1)$$

(11)

with overwhelming probability.

Recall that $X_n$ is a random vector in $\mathbb{C}^{n-1}$ with bounded iid components, and $R_{n-1}$ has operator norm $O(1)$ by the spectral theorem and $b = \text{Im} z > 0$. We also make the crucial observation that $X_n$ is independent of $R_{n-1}$ (this observation and the predecessor comparison lemma (5.8) are the most important ingredients of the Stieltjes transform approach).

It hence suffices to prove concentration for a general random variable of the form $X^*RX$, with $R$ a deterministic $n \times n$ matrix with operator norm $O(1)$ and $X$ a random vector with bounded iid components.

$$||X||^2 = \sum_{i=1}^{n} |\xi_i|^2$$

is a sum of independent bounded random variables with mean 1. From Hoeffding’s inequality (5.4), we have that $X$ has magnitude $O(\sqrt{n})$ with overwhelming probability. It follows that $X^*RX$ has magnitude $O(n)$ with overwhelming probability.

Observe that for the case of $R$ positive semi-definite with operator norm $O(1)$, $(X^*RX)^{1/2} = ||R^{1/2}X||$ is a convex $O(1)$-Lipschitz function of $X$. Applying Talagrand’s inequality (5.6) we have

$$\mathbb{P}(|(X^*RX)^{1/2} - \mathbb{M}(X^*RX)^{1/2}| \geq \epsilon) \leq Ce^{-c\epsilon^2}$$

for any $\epsilon > 0$ and possibly modified constants $C, c > 0$. Moreover, since $(X^*RX)^{1/2} = O(\sqrt{n})$ with overwhelming probability, we must have $\mathbb{M}(X^*RX)^{1/2} = O(\sqrt{n})$. We can hence convert the above tail estimate to a tail estimate for $X^*RX$:
\[ Ce^{-cn^2} \geq P\left( \left| (X^*RX)^{1/2} - M(X^*RX)^{1/2} \right| \geq \epsilon \right) \]
\[ = P\left( \left| X^*RX - MX^*RX \right| \geq \epsilon ((X^*RX)^{1/2} + M(X^*RX)^{1/2})) \right) \]
\[ \geq \frac{1}{2} P\left( \left| X^*RX - MX^*RX \right| \geq \epsilon C' \sqrt{n} \right) \]

for some \( C' > 0 \) sufficiently large (any \( C' \geq 2M(X^*RX)^{1/2} \) will probably). This gives the desired tail bound for \( X^*RX \) around the median, for some modified constants, which in turn gives the bound

\[ P\left( \left| \frac{1}{n} X^*RX - \mathbb{E} \frac{1}{n} X^*RX \right| \geq \epsilon \right) \leq Ce^{-cn^2} \]

again for modified constants \( C, c > 0 \).

The above bound extends to Hermitian \( R \) with bounded operator norm by the triangle inequality, and to general \( R \) with bounded operator norm by a density argument. Taking \( \epsilon \) to be a slowly decaying function of \( n \) (as in the proof of (7)), we obtain

\[ \frac{1}{n} X^*RX = \mathbb{E} X^*RX + o(1) \]

with overwhelming probability.

**Remark** While Talagrand’s inequality is a powerful general purpose concentration of measure tool, there are also inequalities for the specific setting of quadratic forms due to Hanson and Wright [HaWr1971], which were used in [ErScYa2010] to establish concentration for the Stieltjes transform for \( z \) small depending on \( n \) (a regime for which the above argument fails).

Applying this with \( X_n \) and \( R_{n-1} \) and undoing the conditioning on \( R_{n-1} \) we obtain (11), and hence

\[ \frac{1}{n} X_n^*R_{n-1}X_n = \mathbb{E} s_n(z) + o(1) \]

(12)

with overwhelming probability. Substituting this into the recurrence relation (2), and noting that the term \( \xi_{nn}/\sqrt{n} = o(1) \) almost surely, we obtain an approximate recurrence relation:

\[ \mathbb{E} s_n(z) = \frac{-1}{z + \mathbb{E} s_n(z) + o(1)} + o(1) \]

where we have used that the denominator of (2) is bounded away from zero to throw the contribution of the bad event on which (12) does not hold into the additive error \( o(1) \). The boundedness of the denominator away from zero also allows us to use Taylor expansion to bring the error in the denominator up to combine with the other error term:

\[ \mathbb{E} s_n(z) = \frac{-1}{z + \mathbb{E} s_n(z)} + o(1). \]
5.4 Qualitative endgame

In the last section we did not spend much effort keeping track of the precise decay of the error terms, but it is possible to do so. From here, we could also conclude a quantitative statement on the convergence of $s_n(z)$ to the Stieltjes transform of the semicircular law, which when combined with a quantitative version of the Stieltjes continuity theorem can be used to show a polynomial rate of convergence of the empirical spectral measures (in the Lévy metric, say). We point the interested reader to [Ba1993]. In this section we settle on simple qualitative methods (the Arzelá-Ascoli theorem) to show convergence to a fixed point of the recurrence relation.

Indeed, the functions $E_n s_n(z)$ are locally uniformly bounded in the upper half plane (by the reciprocal of the distance to the real line of the compact set in question). By Fubini’s theorem and the regularity of the Stieltjes transform we see also that they are locally uniformly continuous. By the Arzelá-Ascoli theorem there is a subsequence converging locally uniformly to a limit $s$. This will be a Herglotz function. Taking limits in (13) (observing the denominator is bounded away from zero independent of $n$ for all $n$ sufficiently large) we obtain an exact equation

$$s(z) = \frac{-1}{z + s(z)}$$

which we solve to obtain

$$s(z) = \frac{-z \pm \sqrt{z^2 - 4}}{2}.$$

To figure out which branch to take, we first note that from part 3 of Proposition 5.1 we have that for each $n$

$$s_n(z) = \frac{-1 + o_n(1)}{z}$$

where the asymptotic is for $z$ approaching infinite non-tangentially in the upper half plane. We need the asymptotic to hold uniformly over $n$, and this can be established from almost sure uniform tightness of the measures $\mu_n$. This in turn can be established using the almost sure convergence of the second moments, say, of the empirical spectral measures, which is a very easy case of the moment method (cycles of length two, and pairs thereof (for a variance bound), are easy to count!). It is perhaps unfortunate that we need some element of the moment method ideas in the Stieltjes transform proof; I am currently unaware of alternative route to getting the above asymptotic. (The treatment in [Ta2012] has the advantage of having proved the almost sure convergence of the spectral edge — from which the desired asymptotic easily follows — before attacking the semicircular law.)

We hence obtain a non-tangential asymptotic for $s$:

$$s(z) = \frac{-1 + o(1)}{z}.$$
we see that the appropriate branch matching the asymptotic is

\[ s(z) = -
\frac{z + \sqrt{z^2 - 4}}{2} \]  

(14)

where \( \sqrt{z^2 - 4} \) is the branch of the square root with branch cut at \([-2, 2]\) and which is asymptotic to \( z \) at infinity.

As we have identified the only possible subsequential limit of the \( E_{s_n} \), we conclude that the entire sequence converges locally uniformly to (14). From the concentration of measure established in (10) we have that \( s_n(z) \) converges almost surely to \( s(z) \).

Finally, one can see that

\[ \frac{s(a + ib) - s(a - ib)}{2\pi i} \]

converges to the density of the semicircular law

\[ \frac{1}{2\pi} \sqrt{4 - a^2}^+ \]

as \( b \to 0^+ \), which by the inversion formula of Proposition 5.2 concludes the proof of the semicircular law.

5.5 Notes

We took \( z \) to be fixed of size 1. Stronger, quantitative local semicircular laws address the asymptotic density of eigenvalues for intervals as short as \( \log^{O(1)} n/n \) in length. It turns out that to do this one must have control of the Stieltjes transforms in the region \( z \gg \log^{O(1)} n/n \). For this one needs a suitable quantitative Stieltjes continuity theorem for this range of \( z \) – see [ErScYa2009] and [TaVu2011].

References


