5. Continuity of eigenvalues

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

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

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

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

(a) Rank inequality: *Suppose $rank(A-B) = 1$. Then $\sup_{x \in \mathbb{R}} |F_A(x) - F_B(x)| \leq \frac{1}{n}$.*
(b) Hoffman-Wielandt inequality: *Let $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_n$ and $\mu_1 \geq \mu_2 \geq \ldots \geq \mu_n$ be the eigenvalues of A and B respectively. Then, $\sum_{k=1}^n |\lambda_k - \mu_k|^2 \leq tr(A-B)^2$.*
(c) Bound on Lévy distance: *$\mathcal{D}(L_A, L_B) \leq \sqrt[3]{\frac{1}{n} tr(A-B)^2}$.*

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

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

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

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

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

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

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

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

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

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

These inequalities will be used many times later.

6. WSL for Wigner matrices by method of moments

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

(1) To study the expected ESD, we shall have to look at (5). However, if $X_{i,j}$ do not have higher moments, exprssions such as $\mathbf{E}[X_{i_1,i_2} \ldots X_{i_{2q},i_1}]$ may not exist. For instance, if i_r are all equal, this becomes $\mathbf{E}[X_{1,1}^{2q}]$.
(2) Secondly, the evaluation of $\mathbf{E}[X_{i_1,i_2} \ldots X_{i_{2q},i_1}]$ used Wick formulas, which apply to joint Gaussians only.
(3) Lastly, we must prove the result for the ESD itself, not just the expected ESD.

We now indicate how these problems are to be addressed.

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

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

Define,

$Y_{i,j} = X_{i,j}\mathbf{1}_{|X_{i,j}|\leq A}$. $\qquad |Y_{i,j}| \leq A$. Perhaps not centered or scaled right.

$Z_{i,j} = Y_{i,j} - \alpha_A$. $\qquad |Z_{i,j}| \leq 2A$. Centered, but entries have variance β_A, not 1.

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

(A) By part (c) of Lemma 28

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

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

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

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

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

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

Combining (A)-(D), we get

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

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

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

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

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

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

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

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

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

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

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

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

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

There are atmost $\lfloor p/2 \rfloor$ distinct (counting without multiplicities) edges in G since each must be repeated at least twice. Hence, the number of vertices in G is atmost $\lfloor p/2 \rfloor + 1$. If p is even this is attained if and only if G is a tree and $\mathbf{i}$ is a depth-first search of G (hence each edge is traversed twice, once in each direction). If p is odd and there are $\lfloor p/2 \rfloor + 1$ vertices, then some edge will have to be traversed three times exactly, and that is not possible if G is a tree (since $\mathbf{i}$ starts and ends at the same vertex). Note that because of the way our isomorphism works, isomorphism of trees is really isomorphism of plane trees[5].

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

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

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

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

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

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

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