Random Matrices:
Wigner and Marchenko-Pastur Theorems

Steven P. Lalley

May 22, 2019

1 Wigner's Theorem

The mathematical study of large random matrices began with a famous paper of Eugene Wigner published in 1955. Wigner had earlier suggested, on partly heuristic grounds, that the distribution of eigenvalues of a large matrix with random entries could perhaps shed light on the problem of the distribution of energy levels in large, complex quantum mechanical systems. In his 1955 paper, he showed that in fact, if the above-diagonal entries entries of a large, symmetric square matrix are independent and identically distributed with mean 0 and variance 1 then with high probability the distribution of eigenvalues will closely follow a certain probability distribution known as the *semi-circle law*.

Definition 1.1. An $N \times N$ *Wigner matrix* X is a real symmetric matrix whose above-diagonal entries $X(i,j)$, where $1 \leq i \leq j \leq N$, are independent real random variables such that

(a) the diagonal entries $X(i,i)$ are i.i.d. with mean 0, and
(b) the off-diagonal entries $X(i,j)$ are i.i.d. with mean 0 and variance 1.

The Spectral Theorem for real symmetric matrices states that for any such $N \times N$ matrix there is a complete set $\lambda_1, \lambda_2, \cdots, \lambda_N$ of real eigenvalues, with corresponding real unit eigenvectors $u_1, u_2, \cdots, u_N$ forming a complete orthonormal basis of $\mathbb{R}^N$.

Definition 1.2. The *empirical spectral distribution* F^M of a diagonalizable $N \times N$ matrix M with eigenvalues $\lambda_1, \lambda_2, \cdots, \lambda_N$ is the uniform distribution on the set of eigenvalues, that is,

$$F^M = N^{-1} \sum_{i=1}^{N} \delta_{\lambda_i}.$$

Theorem 1.3. *(Wigner) Let $X = X^{(N)}$ be a sequence of $N \times N$ Wigner matrices, and for each N define*

$$M = M^{(N)} = X/\sqrt{N}.$$

Then as $N \to \infty$ the empirical spectral distribution F^M of M converges weakly to the semicircle law *with density*

$$p(t) = \frac{\sqrt{4-t^2}}{2\pi} \mathbf{1}_{[-2,2]}(t). \tag{1.1}$$

More precisely, if $\{\lambda_i\}_{1 \le i \le N}$ are the eigenvalues of M, then for any bounded, continuous function $f : \mathbb{R} \to \mathbb{R}$ and any $\varepsilon > 0$,

$$\lim_{N\to\infty} P\{|N^{-1} \sum_{i=1}^{N} f(\lambda_i) - \int f(t) p(t)\, d\tau| \ge \varepsilon\} = 0 \tag{1.2}$$

The proof will be broken into two distinct parts: first, in sections 4–5 we will show that the theorem is true under the additional hypotheses that the random variables $X_{i,i}$ and $X_{i,j}$ have expectation zero and finite moments of all orders; second, in section 7 we will show that these hypotheses are extraneous.

2 Sample Covariance Matrices

The *sample covariance matrix* of a sample $Y_1, Y_2, \ldots, Y_N$ of independent, identically distributed random (column) vectors is usually defined by

$$S = (N-1)^{-1} \sum_{j=1}^{N} (Y_j - \bar{Y})(Y_j - \bar{Y})^T. \tag{2.1}$$

Here $\bar{Y}$ denotes the sample mean $\bar{Y} = N^{-1} \sum_j Y_j$, and the superscript T denotes matrix transpose. In the particular case where the random vectors Y_i are i.i.d. with the $p-$dimensional multivariate normal distribution $N(0, \Sigma)$ with positive-definite population covariance matrix Σ, the distribution of the random matrix S is known as the *Wishart distribution* $W(\Sigma, N-1)$. It can be shown (exercise!) that in this case the sample covariance matrix has the same distribution as

$$S = (N-1)^{-1} \sum_{j=1}^{N-1} Y_j Y_j^T \tag{2.2}$$

Theorem 2.1. *(Marchenko & Pastur) Let G be a probability distribution on $\mathbb{R}$ with mean 0 and variance* 1*. Assume that Y is a $p \times n$ random matrix whose entries are independent, identically distributed with common distribution G, and define*

$$S = n^{-1} Y Y^T. \tag{2.3}$$

Iff $p, n \to \infty$ in such a way that $p/n \to y \in (0,1)$ then the empirical spectral distribution F^S converges weakly to the Marchenko-Pastur distribution *with density*

$$f_y(x) = \frac{\sqrt{(b-x)(x-a)}}{2\pi x y} \mathbf{1}_{[a,b]}(x) \tag{2.4}$$

where

$$a = a_y = (1-\sqrt{y})^2 \quad \textit{and}$$
$$b = b_y = (1+\sqrt{y})^2.$$

The proof, in section 6 below, will follow the same strategy as that of Wigner's theorem.

3 Method of Moments; Szego's Theorem

3.1 Method of Moments

The proofs of Theorems 1.3 and 2.1 will be based on the *method of moments* (see Appendix A). The strategy will be to show that for each integer $k = 1, 2, \dots$, the kth moment of the empirical spectral distribution converges (in probability) to the kth moment of the corresponding limit law (the semi-circle law in Theorem 1.3, the Marchenko-Pastur law in Theorem 2.1). The key to this strategy is that moments of the empirical spectral distribution can be represented as matrix traces: in particular, if M is an $N \times N$ diagonizable matrix with empirical spectral distribution F^M, then

$$\boxed{\int \lambda^k F^M(d\lambda) = N^{-1}\mathrm{Tr}\, M^k}. \tag{3.1}$$

The trace $\mathrm{Tr}\, M^k$ has, in turn, a combinatorial interpretation as a sum over *closed paths* of length k in the *augmented* complete graph K_N^+. (A *closed path* is a path in a graph that begins and ends at the same vertex. The *complete graph* K_N on N vertices is the graph with vertex set $[N] = \{1, 2, 3, \dots, N\}$ and an edge connecting every pair of distinct vertices. The *augmented complete graph* K_N^+ has the same vertices and edges as the complete graph K_N, and in addition a loop connecting each vertex to itself. Thus, a path in K_N^+ is an arbitrary sequence of vertices $i_1 i_2 \cdots i_m$, whereas a path in K_N is a sequence of vertices in which no vertex appears twice consecutively.) Denote by $\mathscr{P}_i^k$ the set of all length$-k$ closed paths

$$\gamma = i = i_0, i_1, i_2, \dots, i_{k-1}, i_k = i$$

in K_N^+ that begin and end at vertex i, and for each such path γ define the *weight*

$$w(\gamma) = w_M(\gamma) = \prod_{j=0}^{k-1} M_{i_j, i_{j+1}}. \tag{3.2}$$

Then

$$\boxed{\mathrm{Tr}\, M^k = \sum_{i=1}^{N} \sum_{\gamma \in \mathscr{P}_i^k} w(\gamma)}. \tag{3.3}$$

3.2 Szego's Theorem

To illustrate how the method of moments works in a simple case, we prove a classical theorem of Szego regarding the empirical spectral distribution of a *Toeplitz matrix.* A Toeplitz matrix is a square matrix whose entries are constant down diagonals:

$$T = T_N = \begin{pmatrix} a_0 & a_1 & a_2 & \cdots & a_{N-1} & a_N \\ a_{-1} & a_0 & a_1 & \cdots & a_{N-2} & a_{N-1} \\ a_{-2} & a_{-1} & a_0 & \cdots & a_{N-3} & a_{N-2} \\ & & & \cdots & & \\ a_{-N} & a_{-N+1} & a_{-N+2} & \cdots & a_{-1} & a_0 \end{pmatrix}. \tag{3.4}$$

The Fourier series $A(\theta) := \sum_{n=-\infty}^{\infty} a_n e^{in\theta}$ whose coefficients are the entries of T is sometimes called the *symbol* of the infinite Toeplitz matrix T_∞. The *spectral distribution* is the distribution of $A(\Theta)$, where Θ is a random variable uniformly distributed on the interval $[-\pi, \pi]$.

Theorem 3.1. *(Szego) Assume that only finitely many of the coefficients a_n are nonzero, that is, the symbol $A(\theta)$ is a trig polynomial. Then the empirical spectral distribution F_N of the Toeplitz matrix T_N converges as $N \to \infty$ to the spectral distribution, that is, for every bounded continuous function $f : \mathbb{C} \to \mathbb{R}$,*

$$\lim_{N\to\infty} \int f(t)\, F_N(dt) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(A(\theta))\, d\theta. \tag{3.5}$$

Proof. The kth moment of the empirical spectral distribution F_N is $N^{-1}\times$the trace of T_N^k, by (3.1), and by (3.3), the trace $\text{Tr } T_N^k$ is the sum of the $T-$weights of all length-k closed paths. By hypothesis, there exists $M < \infty$ such that $a_n = 0$ if $|n| > M$. Thus, only closed paths in which the steps are of size $\leq M$ in magnitude will have nonzero weights. Moreover, for any indices i, j between $kM + 1$ and $N - kM - 1$ there is a one-to-one correspondence beween length$-k$ paths beginning and ending at i and length$-k$ paths beginning and ending at j, and because T_N is Toeplitz, this correspondence preserves path-weight, that is, if $\gamma \sim \gamma'$ then $w(\gamma) = w(\gamma')$. Consequently,

$$\sum_{\gamma \in \mathscr{P}_i^k} w(\gamma) = \sum_{\gamma' \in \mathscr{P}_j^k} w(\gamma') = (a * a * \cdots * a)_0 = (2\pi)^{-1} \int_{-\pi}^{\pi} A(\theta)^k\, d\theta$$

where $*$ denotes sequence convolution. Thus, as $N \to \infty$,

$$N^{-1}\text{Tr } T_N^k \longrightarrow (2\pi)^{-1} \int_{-\pi}^{\pi} A(\theta)^k\, d\theta.$$

This proves that the moments of the empirical spectral distribution converge to the corresponding moments of the spectral distribution. Since the spectral distribution is uniquely determined by its moments (this follows because it has bounded support), the result follows. □

Exercise 3.2. Show that Szego's Theorem is true under the weaker hypothesis that $\sum_{-\infty}^{\infty} |a_n| < \infty$.

Exercise 3.3. Let M_N be the random symmetric tridiagonal matrix

$$M_N = \begin{pmatrix} X_{1,1} & X_{1,2} & 0 & 0 & \cdots & 0 & 0 \\ X_{1,2} & X_{2,2} & X_{2,3} & 0 & \cdots & 0 & 0 \\ 0 & X_{2,3} & X_{3,3} & X_{3,4} & \cdots & 0 & 0 \\ & & & & \cdots & & \\ 0 & 0 & 0 & 0 & \cdots & X_{N-1,N-1} & X_{N-1,N} \\ 0 & 0 & 0 & 0 & \cdots & X_{N-1,N} & X_{N,N} \end{pmatrix}$$

whose entries $X_{i,j}$, where $j = i$ or $j = i+1$, are independent, identically distributed $N(0,1)$. (Note: Since M_N is symmetric, the eigenvalues are real.)

(A) * Prove that as $N \to \infty$ the empirical spectral distribution of M_N converges to a nontrivial, nonrandom limit. HINT: WLLN for $m-$dependent random variables.

(B) ** Identify the limit distribution.

4 Semi-Circle Law and the Catalan Numbers

Definition 4.1. The nth *Catalan number* is defined by

$$\kappa_n = \binom{2n}{n} \Big/ (n+1). \tag{4.1}$$

Exercise 4.2. (A) Show that κ_n is the number of *Dyck words* of length $2n$. A Dyck word is a sequence $x_1 x_2 \cdots x_{2n}$ consisting of n +1's and n −1's such that no initial segment of the string has more −1's than +1's, that is, such that for every $1 \leq m \leq 2n$,

$$\sum_{j=1}^{m} x_j \geq 0$$

HINT: Reflection Principle. (B) Show that κ_n is the number of expressions containing n pairs of parentheses which are correctly matched: For instance, if $n = 3$,

((()))   ()(())   ()()()   (())()   (()())

(C) Show that κ_n is the number of rooted ordered binary trees with $n+1$ leaves. (See your local combinatoricist for the definition.) HINT: Show that there is a bijection with the set of Dyck words, and use the result of part (A). (D) Do problem 6.19 in R. Stanley, *Enumerative Combinatorics*, vol. 2.

Proposition 4.3. *Let $p(t)$ be the semi-circle density* (1.1)*. Then all odd moments of p are zero, and the even moments are the Catalan numbers:*

$$\kappa_n = \int_{-2}^{2} t^{2n} p(t)\, dt \quad \forall\ n = 0, 1, 2, \dots \tag{4.2}$$

Proof. Calculus exercise. Hint: Use the change of variable $t = 2\cos\theta$ to rewrite the integral as an integral involving powers of $\sin\theta$ and $\cos\theta$. Then try either integration by parts, or rewrite the sines and cosines in terms of $e^{\pm i\theta}$ and use the orthogonality of the complex exponentials. □

Remark 4.4. If the semi-circle density (1.1) seems vaguely familiar, it is because it is closely related to the probability generating function of the first return time to the origin by a simple random walk on the integers. Let S_n be a simple random walk started at 0; thus, the increments $X_n = S_n - S_{n-1}$ are independent, identically distributed Rademacher-$\frac{1}{2}$ (that is, ±1 with probabilities $\frac{1}{2}, \frac{1}{2}$). Define

$$T = \min\{n \geq 1 : S_n = 0\} \tag{4.3}$$

Then the probability generating function of T is (see 312 notes)

$$Ez^T = 2 - \sqrt{4 - z^2} = \sum_{n=1}^{\infty} z^{2n} \binom{2n}{n} \Big/ \{(n+1)2^{2n}\}. \tag{4.4}$$

The last power series expansion follows either from Newton's binomial formula or a direct combinatorial argument (exercise). Note the appearance of the Catalan numbers.

5 Proof of Wigner's Theorem

In this section we will prove Theorem 1.3 under the following additional assumption on the distributions of the entries $X_{i,j}$. Later we will show how truncation methods can be used to remove this assumption.

Assumption 5.1. *The random variables $(X_{i,j})_{j \geq i}$ are independent. The random variables $(X_{i,j})_{j>i}$ have distribution F with mean zero, variance* 1*, and all moments finite, and the (real) random variables $X_{i,i}$ have distribution G with mean zero and all moments finite.*

5.1 Combinatorial Preliminaries

The proof will be based on the method of moments. We will show that for each $k = 1, 2, \dots,$ the random variable $N^{-1}\mathrm{Tr}\, M^k$ converges in probability to the kth moment of the semi-circle law. To accomplish this, it suffices, by Chebyshev's inequality, to show that the mean $N^{-1}E\mathrm{Tr}\,M^k$ converges to the kth moment of the semi-circle law, and that the variance $N^{-2}\mathrm{Var}(\mathrm{Tr}\,M^k)$ converges to 0. By equation (3.3), the expected trace is a sum over closed paths of length k in the augmented complete graph K_N^+. We will begin by showing that the contribution to the expectation of an closed path γ depends only on its *type*.

Definition 5.2. The *type* of a path $\gamma = i_0 i_1 i_2 \cdots i_k$ of length k is the sequence $\tau(\gamma) = m_0 m_1 m_2 \cdots m_k$, where for each j the vertex i_j is the m_jth *distinct* vertex encountered in the sequence $i_0, i_1, \dots, i_k$. (Observe that for a *closed* path γ, that is, a path with the same initial and final vertices, that last entry m_k of $\tau(\gamma)$ is 1.)

Example 5.3. The type of the path $uvquuquvu$ is $1,2,3,1,1,3,1,2,1$.

Lemma 5.4. *If γ and γ' are two closed paths in the augmented complete graph K_N^+ with the same type, then there is a permutation σ of the the vertex set $[N]$ that maps γ to γ'.*

Proof. Easy exercise. □

Corollary 5.5. *If γ and γ' are two closed paths in K_N^+ with the same type, then*

$$Ew_M(\gamma) = Ew_M(\gamma') \tag{5.1}$$

Proof. This follows directly from the preceding lemma, because for any permutation σ of the index set $[N]$, the distribution of the random matrix $(M_{\sigma(i),\sigma(j)})$ is the same as that of $(M_{i,j})$. This is obvious, because the off-diagonal entries of M are i.i.d., as are the diagonal entries. Note: In fact, this holds more generally for any random matrix ensemble that is invariant under either unitary or orthogonal transformations, because a permutation matrix is both orthogonal and unitary. □

Corollary 5.6. *Let $\mathscr{T}_k$ be the set of all path types of length$-k$ closed paths. For each type τ, let $H(\tau,N)$ be the number of closed paths in K_N^+ with type τ, and let $Ew_M(\tau)$ be the common value of the expectations* (5.1). *Then*

$$E\mathrm{Tr}\,M^k = \sum_{\tau\in\mathscr{T}_k} H(\tau,N)Ew_M(\tau). \tag{5.2}$$

Moreover, if $G(\tau)$ is the number of distinct integers in the sequence τ (that is, the number of distinct vertices in a representative path), then

$$H(\tau,N) = N(N-1)(N-2)\cdots(N-G(\tau)+1) := (N)_{G(\tau)}. \tag{5.3}$$

Proof. Another easy exercise. □

Lemma 5.7. *Let γ be a closed path in K_N^+. If $Ew_M(\gamma)\neq 0$ then every edge or loop e crossed by γ at least once must be crossed at least twice.*

Proof. If the edge ij is crossed by γ just once (say in the forward direction), then the factor $X_{i,j}$ occurs to the first power in the weight $w_M(\gamma)$ (see (3.2)), and the factor $X_{j,i} = X^*_{i,j}$ does not ocur at all. Since the random variables $X_{i,j}$ above or on the main diagonal are independent, all with mean zero, it follows that the expectation $Ew_M(\gamma) = 0$. □

For each closed path γ in the *augmented* complete graph K^+_N, define $\Gamma(\gamma)$ to be the subgraph of the complete graph K_N consisting of the vertices visited and edges crossed (in either direction) by γ. (Note: Loops are not included.) Clearly $\Gamma(\gamma)$ is itself a connected graph, because the path γ visits every vertex and edge. By Lemma 5.4, if two closed paths γ, γ' are of the same type, then $\Gamma(\gamma)$ and $\Gamma(\gamma')$ are isomorphic; moreover, if γ crosses every edge in $\Gamma(\gamma)$ at least twice, then γ' crosses every edge in $\Gamma(\gamma')$ at least twice.

Lemma 5.8. *If Γ is a connected graph , then*

$$\#\textit{(vertices)} \le \#\textit{(edges)} + 1, \tag{5.4}$$

and equality holds if and only if Γ is a tree *(that is, a graph with no nontrivial cycles).*

Remark 5.9. Together with the equation (5.3), this will be used to show that the dominant contribution to the sum (5.2) comes from path-types for which representative paths γ are such that $\Gamma(\gamma)$ is a tree and γ crosses every edge of $\Gamma(\gamma)$ exactly twice. The key observation is that if γ is a closed path of length k (that is, it makes k steps) that crosses every edge in $\Gamma(\gamma)$ at least twice, then $\Gamma(\gamma)$ has no more than $[k/2]$ edges; moreover, if k is even then the graph $\Gamma(\gamma)$ will have *fewer* than $k/2$ distinct edges unless the path γ croses every edge *exactly* twice.

Proof of Lemma 5.8. Consider first the case where Γ is a tree. I will show, by induction on the number of vertices in the tree Γ, that

$$\#(\text{vertices}) = \#(\text{edges}) + 1.$$

First note that, since Γ has no nontrivial cycles, the removal of any edge must disconnect Γ (why?). Also, since Γ contains only finitely many edges, it must have an *end*, that is, a vertex v that is incident to only one edge e (why?) Now remove both v and e from the graph; the resulting graph Γ' is still a tree (why?) but has one fewer vertex than Γ. Hence, the induction hypothesis implies that the desired formula holds for Γ', and it therefore follows that it must also hold for Γ.

Now consider the case where the graph Γ has at least one cycle. Then removal of one edge on this cycle will leave the graph *connected*, with the same number of vertices but one fewer edge. Continue to remove edges from cycles, one at a time, until there are no more cycles. Then the resulting graph will be a tree, with the same vertex set as the original graph Γ but with strictly fewer edges. □

Remark 5.10. Here is another argument for inequality (5.4). This argument has the virtue that it extends easily to *pairs* of paths γ, γ' – see the proof of Proposition 5.13 below. Let $\gamma = i_0 i_1 \ldots i_{2k}$ be a path of length $2k$. Define a corresponding 0-1 sequence $\nu(\gamma) = n_0 n_1 \cdots n_{2k}$ as follows: If γ visits vertex i_j for the first time at step j, set $n_j = 1$, otherwise set $n_j = 0$. Now consider the jth step: If it is across an edge that was previously crossed (in either direction), then vertex i_j must have been visited earlier, and so $n_j = 0$. Hence, if the path γ crosses every edge at least twice, then

$$\sum_{j=1}^{2k} n_j \le k,$$

because there are $2k$ edge crossings. Thus, the total number $\sum_{j=0}^{2k} n_j$ of vertices visited by γ cannot be larger than $k+1$.

Lemma 5.11. *Let $\mathscr{G}_k$ be the set of closed path types $\tau \in \mathscr{T}_{2k}$ with the following properties:*

(i) For some (and hence every) closed path γ of type $\tau(\gamma) = \tau$, the graph $\Gamma(\gamma)$ is a tree.
(ii) The path γ contains no loops, and crosses every edge in $\Gamma(\gamma)$ exactly *twice, once in each direction.*

Then $\mathscr{G}_k$ has cardinality

$$\#\mathscr{G}_k = \kappa_k, \tag{5.5}$$

where κ_k is the kth Catalan number.

Note: If $\Gamma(\gamma)$ is a tree and if γ crosses every edge in $\Gamma(\gamma)$ exactly twice, then it must necessarily cross every edge once in each direction, because otherwise the graph $\Gamma(\gamma)$ would contain a nontrivial cycle.

Proof. Let $\gamma = i_0 i_1 \cdots i_{2k}$ be an element of $\mathscr{G}_k$. By Lemma 5.8, the path γ visits $k+1$ distinct vertices, including the initial vertex i_0. Define a ± 1 sequence $s(\gamma) = s_1 s_2 \cdots s_{2k}$ as follows:

(+) $s_j = +1$ if γ visits i_{j-1} for the first time on the $(j-1)$th step;
(-) $s_j = -1$ otherwise.

This is a Dyck word, because for every $m \le 2k$ the sum $\sum_{j=1}^{m} s_j$ counts the number of vertices visited exactly once by γ in the first m steps (why?). Conversely, for every Dyck word s of length $2k$, there is a closed path γ in the graph K_N such that $s(\gamma) = s$ *provided* $N \ge k+1$ (why?). Hence, the result follows from Exercise 4.2. □

5.2 Convergence of Means

Proposition 5.12. *If Assumption 5.1 holds, then for every integer $k \ge 0$,*

$$\lim_{N\to\infty} N^{-1} E\mathrm{Tr}\, M^{2k} = \kappa_k \quad \textit{and} \tag{5.6}$$

$$\lim_{N\to\infty} N^{-1} E\mathrm{Tr}\, M^{2k+1} = 0 \tag{5.7}$$

where κ_k is the kth Catalan number.

Proof. The starting point is formula (5.2). According to this formula, the expectation is a sum over closed path types of the appropriate length. By Lemma 5.7, the only path types τ that contribute to this sum are those for which representative closed paths γ cross every edge twice. I will use Lemma 5.8 to determine the types that make the *dominant* contribution to the sum (5.2).

Case 1: Odd Moments. Since only paths γ that make $2k+1$ steps and cross every edge at least twice contribute to the expectation $E\mathrm{Tr}\, M^{2k+1}$, such paths have graphs $\Gamma(\gamma)$ with no more than k edges. Consequently, by Lemma 5.8, $\Gamma(\gamma)$ has no more than $k+1$ vertices, and so equation (5.3) implies that

$$H(\tau(\gamma), N) \le N^{k+1}.$$

Now consider the expectation $Ew_M(\gamma)$. The factors in the product

$$w_M(\gamma) = \prod_{j=0}^{2k} M_{i_j,i_{j+1}} = N^{-(2k+1)/2} \prod_{j=0}^{2k} X_{i_j,i_{j+1}}$$

can be grouped by edge (or loop, if there are factors $X_{i,i}$ from the diagonal). By Assumption 5.1, the distributions of the diagonal and off-diagonal entries have finite moments of all orders, so Hölder's inequality implies that

$$|Ew_M(\tau(\gamma))| = |Ew_M(\gamma)| \le N^{-(2k+1)/2} \max(E|X_{1,2}|^{2k+1}, E|X_{1,1}|^{2k+1}).$$

Therefore,

$$\begin{aligned} N^{-1}|E\mathrm{Tr}\, M^{2k+1}| &\le N^{-1} \sum_{\tau} H(\tau, N)|Ew_M(\tau)| \\ &\le N^{-1} N^{k+1} N^{-(2k+1)/2} \max(E|X_{1,2}|^{2k}, E|X_{1,1}|^{2k}) \\ &= O(N^{-1/2}) \end{aligned}$$

Note: The upper bound can be sharpened by a factor of N^{-1}, because with a bit more work (Exercise!) it can be shown that if γ is a closed path with an odd number $2k+1$ of steps that crosses every edge at least twice, then $\Gamma(\gamma)$ must contain a cycle, and so cannot have more than k vertices.

Case 2: Even Moments. In this case, I will use use Lemma 5.8 to show that for large N, the *dominant* contribution comes from types $\tau \in \mathcal{G}_k$ First, observe that if $\tau \in \mathcal{G}_k$, then for any closed path γ of type $\tau(\gamma) = \tau$ the product contains k *distinct* factors $|X_{l,m}|^2/N$, all with $m > l$. This is because the path γ crosses every edge in $\Gamma(\gamma)$ exactly twice, once in each direction, and contains no loops. (Recall that the random matrix X is Hermitian, so

$X_{l,m} = X^*_{m,l}$.) Since the k factors are distinct, they are independent, and since they are also identically distributed with variance 1 it follows that

$$Ew_M(\tau) = N^{-k}(E|X_{1,2}|^2)^k = N^{-k}.$$

Now consider an arbitrary path type $\tau \in \mathscr{T}_{2k}$, and let γ be a closed path of type τ. As in Case 1, Assumption 5.1 and Hölder's inequality imply that

$$|Ew_M(\gamma)| \leq N^{-k} \max(E|X_{1,2}|^{2k}, E|X_{1,1}|^{2k}).$$

Observe that this is of the same order of magnitude as the expectation N^{-1} for types $\tau \in \mathscr{G}_k$.

Finally, consider the factors $H(\tau, N) = (N)_{G(\tau)}$ in the formula (5.2). By Lemma 5.8, $G(\tau) \leq k+1$ for all types, and the inequality is strict *except* for types $\tau \in \mathscr{G}_k$. Hence, the dominant contribution to the sum (5.2) comes from types $\tau \in \mathscr{G}_k$. Therefore, by Lemma 5.11,

$$E\text{Tr}\, M^{2k} = \sum_{\tau \in \mathscr{T}_{2k}} H(\tau, N) Ew_M(\tau) \sim \sum_{\tau \in \mathscr{G}_k} H(\tau, N) N^{-k} \sim N\kappa_k.$$

□

5.3 Bound on the Variance

By Lemma 5.7 and Proposition 5.12, the *expected* moments of the empirical spectral distribution F^M converge to the corresponding moments of the semi-circle law. Thus, to complete the proof of Wigner's theorem, it suffices to show that the *variances* of these moments converge to zero as $N \to \infty$.

Proposition 5.13. *If Assumption 5.1 holds, then there exist constants $C_k < \infty$ such that*

$$\text{Var}(N^{-1}\text{Tr}\, M^k) \leq C_k/N. \tag{5.8}$$

Proof. By the elementary formula $\text{Var}(Y) = E(Y^2) - (EY)^2$, bounding the variance can be accomplished by comparing the first and second moments of Tr M^k. By equation (3.1),

$$E(\text{Tr}\, M^k)^2 = \sum_\gamma \sum_{\gamma'} Ew_M(\gamma) w_M(\gamma') \quad \text{and}$$
$$(E\text{Tr}\, M^k)^2 = \sum_\gamma \sum_{\gamma'} Ew_M(\gamma) Ew_M(\gamma'),$$

where the sums are over the set of all pairs γ, γ' of closed paths of length k in the augmented complete graph K_N^+. Now for any pair γ, γ' with no vertices (and hence no edges) in common, the products $w_M(\gamma)$ and $w_M(\gamma')$ have no factors in common, and consequently

$$Ew_M(\gamma)w_M(\gamma') = Ew_M(\gamma)Ew_M(\gamma').$$

Consequently, the difference $E(\text{Tr } M^k)^2 - (E\text{Tr } M^k)^2$ is entirely due to pairs γ, γ' that have at least one vertex in common. Furthermore, for the same reason as in Lemma 5.7, $Ew_M(\gamma)w_M(\gamma') = 0$ unless every edge in $\Gamma(\gamma) \cup \Gamma(\gamma')$ is crossed at least twice (either twice by γ, twice by γ, or once by each).

Consider a pair of closed paths γ, γ' such that γ and γ' share at least one vertex, and such that every edge crossed at least once by either γ or γ' is crossed at least twice. Since γ and γ' share at least one vertex, the graph[1] $\Gamma(\gamma) \cup \Gamma(\gamma')$ is *connected*, and since each edge in this graph is crossed at least twice, the total number of edges in $\Gamma(\gamma) \cup \Gamma(\gamma')$ is no larger than k. Consequently, by Lemma 5.8, the number of vertices in $\Gamma(\gamma) \cup \Gamma(\gamma')$ cannot exceed $k+1$. Therefore, the total number of such pairs γ, γ' is no larger than N^{k+1}, and so their aggregate contribution to the expectation $E(\text{Tr } M^k)^2$ is bounded in magnitude by

$$N^{k+1} \max_{\gamma,\gamma'} |Ew_M(\gamma)w_M(\gamma')| \leq C_k N$$

where $C_k = \max_{j \leq k} \max(E|X_{11}|^j, E|X_{1,2}|^j)$. □

6 Proof of the Marchenko-Pastur Theorem

6.1 Trace Formula for Covariance Matrices

The basic strategy will be the same as for Wigner's theorem, but the combinatorics is somewhat different. The starting point is once again a combinatorial formula for Tr S^k, where S is defined by equation (2.3):

$$S = n^{-1} Y Y^T. \tag{6.1}$$

Recall that Y is a $p \times n$ random matrix whose entries are i.i.d. with mean 0 and variance 1. Thus, the entries $Y_{i,j}$ of Y have row indices $i \in [p]$ and column indices$j \in [n]$. Define $B = B(p,n)$ to be the *complete bipartite graph* with partitioned vertex set $[p] \cup [n]$: that is, the graph whose edges join arbitrary pairs $i \in [p]$ and $j \in [n]$. It is helpful to think of the vertices as being arranged on two levels, an upper level $[p]$ and a lower level $[n]$; with this convention, paths in the graph make alternating up and down steps. Denote by $\mathscr{Q}$ the set of all closed paths

$$\gamma = i_0, j_1, i_1, j_2, \cdots i_{k-1}, j_k, i_k = i_0 \tag{6.2}$$

[1] Recall that $\Gamma(\gamma)$ is the subgraph of the complete graph K_N consisting of all vertices visited and all edges crossed (in either direction) by γ.

in $B(p,n)$ of length $2k+1$ beginning and ending at a vertex $i \in [p]$, and for each such path define the *weight* (or Y*−weight*) of γ by

$$w(\gamma) = w_Y(\gamma) = \prod_{m=1}^{k} Y_{i_{m-1},j_m} Y_{j_m,i_m}. \tag{6.3}$$

Then

$$\boxed{\text{Tr } S^k = \sum_{\gamma \in \mathscr{Q}} w_Y(\gamma)}. \tag{6.4}$$

6.2 Enumeration of Paths

The proof of the Marchenko-Pastur Theorem will be accomplished by showing that for each $k \geq 0$ (a) the expectation ETr S^k converges to the kth moment of the Marchenko-Pastur distribution (2.4), and (b) the variance VarTr S^k converges to 0. Given (a), the proof of (b) is nearly the same as in the proof of Wigner's theorem, so I will omit it. The main item, then, is the analysis of the expectation ETr S^k as $n, p \to \infty$ with $p/n \to y$.

Lemma 6.1. *Let $\gamma \in \mathscr{Q}$ be a closed path in $B(p,n)$ that begins and ends at a vertex $i \in [p]$. If $Ew_Y(\gamma) \neq 0$ then every edge crossed by γ must be crossed at least twice.*

Proof. See Lemma 5.7. □

For any path γ in $B(n,p)$ with initial vertex in $[p]$, the type $\tau(\gamma)$ and the associated subgraph $\Gamma(\gamma)$ of $B(n,p)$ are defined as earlier. Since paths γ jump back and forth between the disjoint vertex sets $[p]$ and $[n]$, even entries of $\tau(\gamma)$ count vertices in $[p]$ and odd entries count vertices in $[n]$.(Note: Indices start at 0.) Thus, it will be convenient to extract the even- and odd- entry subsequences τ_E and τ_O of τ, and to write $\tau = (\tau_E, \tau_O)$ (this is an abuse of notation – actually τ is the sequence obtained by *shuffling* τ_E and τ_O). Lemma 5.4 translates to the bipartite setting as follows:

Lemma 6.2. *If γ and γ' are two closed paths in the complete bipartite graph $B(n,p)$ with the same type $\tau(\gamma) = \tau(\gamma')$, then there are permutations σ, σ' of the vertex sets $[n]$ and $[p]$ that jointly map γ to γ'. Consequently, if γ, γ' have the same type then*

$$Ew_M(\gamma) = Ew_M(\gamma'). \tag{6.5}$$

Corollary 6.3. *Let $\mathscr{T}_k^B$ be the set of all closed path types of length $2k+1$. For each $\tau \in \mathscr{T}_K^B$ let $H^B(\tau, n, p)$ be the number of closed paths in $B(n,p)$ of type τ, and let $Ew_S(\tau)$ be the common value of the expectations* (6.5)*. Then*

$$E\text{Tr } S^k = \sum_{\tau \in \mathscr{T}_k^B} H^B(\tau, n, p) Ew_S(\tau). \tag{6.6}$$

Moreover, if $G_O(\tau)$ and $G_E(\tau)$ are the number of distinct entries in the even/odd sequences τ_O and τ_E, respectively, then

$$H(\tau, n, p) = (p)_{G_E(\tau)}(n)_{G_O(\tau)}. \tag{6.7}$$

Recall that $(n)_m = n(n-1)\cdots(n-m+1)$. The proof of Corollary 6.3 is trivial, but the difference in formulas (5.3) and (6.7) is what ultimately accounts for the difference between the limit distributions in the Wigner and Marchenko-Pastur theorems. Note that $G_O(\tau)$ and $G_E(\tau)$ are the numbers of distinct vertices in the lower and upper levels $[n]$ and $[p]$, respectively, visited by a path γ of type τ.

Lemma 6.4. *Let γ be a closed path in the bipartite graph $B(n,p)$ of length $2k+1$ (thus, γ makes $2k$ steps, each across an edge). If γ crosses each edge of $\Gamma(\gamma)$ at least twice, then the number of distinct vertices in $\Gamma(\gamma)$ is no larger than $k+1$:*

$$G_O(\tau(\gamma)) + G_E(\tau(\gamma)) \leq k+1, \tag{6.8}$$

and equality holds if and only if $G(\gamma)$ is a tree and γ crosses every edge of $\Gamma(\gamma)$ exactly twice, once in each direction.

This can be proved by the same argument as in Lemma 5.8, and the result serves the same purpose in the proof of the theorem – it implies that the main contribution to the sum (6.6) comes from path types τ for which the corresponding graph Γ is a tree. To see this, keep in mind that p and n are of the same order of magnitude, so (6.7) will be of maximum order of magnitude when $G_O(\tau(\gamma)) + G_E(\tau(\gamma))$ is maximal. Enumeration of these types is the main combinatorial task:

Lemma 6.5. *Let $\mathscr{G}_k^B(r)$ be the set of closed path types $\tau \in \mathscr{T}_{2k}^B$ with the following properties:*

(i) For some (and hence every) closed path γ of type $\tau(\gamma) = \tau$, the graph $\Gamma(\gamma)$ is a tree.
(ii) The path γ crosses every edge in $\Gamma(\gamma)$ exactly *twice, once in each direction.*
(iii) $G_E(\tau) = r+1$ and $G_O(\tau) = k-r$.

(Note that $G_E(\tau) \geq 1$ for any allowable type, because all paths start and end on the upper level $[p]$.) Then $\mathscr{G}_k^B(r)$ has cardinality

$$\#\mathscr{G}_k^B(r) = \binom{k-1}{r}\binom{k}{r}\Big/(r+1). \tag{6.9}$$

Proof. First, I will show that $\mathscr{G}_k^B(r)$ can be put into one-to-one correspondence with the set $\mathscr{A}_k^+(r,r)$ of sequences

$$s = d_1u_1d_2u_2\cdots d_ku_k$$

satisfying the following properties:

(A) Each $d_i \in \{0,-1\}$, and each $u_i \in \{0,1\}$.

(r,r) $\sum_{i=1}^{k} u_1 = r$ and $\sum_{i=1}^{k} d_i = -r$.
(+) $\sum_{i=1}^{m}(u_i + d_i) + d_{m+1} \geq 0$ for all $m < k$.

Then I will show that the set $\mathscr{A}_k^+(r,r)$ has cardinality (6.9).

Step 1: Bijection $\mathscr{G}_k^B(r) \leftrightarrow \mathscr{A}_k^+(r,r)$. Let γ be a path satisfying hypotheses (i)–(iii) of the lemma. This path crossses k edges, once in each direction, and alternates between the vertex sets $[p]$ and $[n]$. After an even number $2i$ of edge crossings, γ is at a vertex in $[p]$; if this vertex is new – that is, if it is being visited for the first time – then set $u_i = +1$, and otherwise set $u_i = 0$. After an odd number $2i-1$ of edge crossings, γ is at a vertex in $[n]$, having just exited a vertex $v \in [p]$. If this edge crossing is the *last* exit from vertex v, then set $d_i = -1$; otherwise, set $d_i = 0$. It is clear that this sequence s has properties (1), (2), (3) above, and it is also clear that if γ and γ' have the same type, then the corresponding sequences s and s' are the same. Thus, we have constructed a mapping $\varphi : \mathscr{G}_k^B(r) \to \mathscr{A}_k^+(r,r)$.

It remains to prove that φ is a bijection. This we will accomplish by exhibiting an inverse mapping $\psi : \mathscr{A}_k^+(r,r) \to \mathscr{G}_k^B(r)$. Several facts about the algorithm used to produce φ are relevant: If γ is a path through vertices $i_0, j_1, i_1, j_2, \dots$ as in (6.2), with $i_m \in [p]$ amd $j_m \in [n]$, then

(a) Vertex j_m is new *unless* $d_m = -1$.
(b) Vertex j_m is exited for the last time if and only if $u_m = 0$.
(c) If vertex i_m is *not* new, that is, if $u_m = 0$, then i_m is the *last* vertex in $[p]$ visited by γ that has not been exited for the last time, that is, by an edge marked $d = -1$.
(d) If vertex j_m is *not* new, then j_m is the *last* vertex in $[n]$ visited by γ that has not been exited for the last time.

These all follow from the hypotheses that $\Gamma(\gamma)$ is a tree, and that γ crosses every edge of $\Gamma(\gamma)$ exactly twice, once in each direction. Consider, for instance, assertion (a): If $d_m = -1$, then vertex i_{m-1} is being exited for the last time, across the edge $i_{m-1}j_m$; this edge must have been crossed once earlier, so j_m is not new. Similarly, if $d_m = 0$, then vertex i_{m-1} will be revisited at some later time; if j_m were *not* new, then either γ would have a nontrivial cycle or edge $i_{m-1}j_m$ would be crossed ≥ 3 times.

Exercise 6.6. Give detailed proofs of (a)–(d).

Assertions (a)–(d) provide an algorithm for constructing the type sequences τ_E, τ_O from the sequence s. By (a) and the definition of the sequence s, the markers $u_m = 1$ and $d_l = 0$ indicate which entries of τ_E and τ_O correspond to new vertices; each such entry must be the max+1 of the previous entries. By (b) and the definition of s, the markers $u_m = 0$ and $d_l = -1$ indicate which entries of τ_E and τ_O correspond to vertices being exited for the last time. Consequently, (c) and (d) allow one to deduce which previously visited vertex is being visited when the current vertex is not new.

Step 2: Enumeration of $\mathscr{A}_k^+(r,r)$. This will make use of the *Reflection Principle*, in much the same way as in the classical ballot problem. Note first that sequences $s \in \mathscr{A}_k^+(r,r)$ must have $d_1 = 0$ and $u_k = 0$, because of the constraints (+) and (r,r), so we are really enumerating the set of sequences

$$s = 0(u_1 d_2)(u_2 d_3) \cdots (u_{k-1} d_k)0 \tag{6.10}$$

that satisfy the properties (A), (r,r), and (+).

The first step is to enumerate the sequences of the form (6.10) that satisfy property (A) and count constraints (r,s) such as the constraint (r,r) above. Thus, for each integer $k \geq 1$ and integers $r, s \geq 0$, let $\mathscr{A}_{k-1}(r,s)$ be the set of sequences s of the form (6.10) such that:

(A) Each $d_i \in \{0, -1\}$, and each $u_i \in \{0, 1\}$.

(r,s) $\sum_{i=1}^k u_1 = r$ and $\sum_{i=1}^k d_i = -s$.

This set is easily enumerated: There are $k-1$ "down" slots, which must be filled with $k-s-1$ 0's and s (-1)'s, and there are $k-1$ "up" slots, which must be filled with $k-r-1$ 0's and r (+1)'s, so

$$\#(\mathscr{A}_{k-1}(r,s)) = \binom{k-1}{r}\binom{k-1}{s}. \tag{6.11}$$

The proof of (6.9) will be completed by showing that

$$\#\mathscr{A}_k^+(r,r) = \#\mathscr{A}_{k-1}(r,r) - \#\mathscr{A}_{k-1}(r-1,r+1). \tag{6.12}$$

To see this, observe that every sequence in $\mathscr{A}_k^+(r,r)$ is an element of $\mathscr{A}_{k-1}(r,r)$, by (6.10). Hence, it is enough to show that the set of sequences in $\mathscr{A}_{k-1}(r,r)$ that do *not* satisfy the nonnegativity constraint (+) is in one-to-one correspondence with $\mathscr{A}_{k-1}(r-1,r+1)$. If a sequence $s \in \mathscr{A}_k(r,r)$ does not satisfy (+), then there must be a *first* time $\tau < k$ such that

$$\sum_{i=1}^{\tau-1} (u_i + d_{i+1}) = -1; \tag{6.13}$$

moreover, the remaining terms must sum to $+1$, because the overall sum is 0:

$$\sum_{i=\tau}^{k-1} (u_i + d_{i+1}) = +1. \tag{6.14}$$

Reflect this part of the path by reversing all pairs $u_j d_{j+1}$ and then negating: thus, for each $u_j d_{j+1}$ with $j \geq \tau$,

$$\begin{gathered}(+,0) \mapsto (0,-)\\ (+,-) \mapsto (+,-)\\ (0,-) \mapsto (+,0)\\ (0,0) \mapsto (0,0).\end{gathered}$$

This mapping sends sequences s in $\mathscr{A}_{k-1}(r,r)$ that do not satisfy (+) to sequences s' in $\mathscr{A}_{k-1}(r-1),r+1)$; it is invertible because for any sequence $s' \in \mathscr{A}_{k-1}(r-1,r+1)$, the reflection rule can be applied in reverse – after the first time τ such that the sum (6.13) reaches the level -1, flip the remaining pairs according to the same rules as above to recover a sequence $s \in \mathscr{A}_{k-1}(r-1,r-1)$. □

6.3 Moments of the Marchenko-Pastur Law

Proposition 6.7. *The kth moment of the Marchenko-Pastur density $f_y(x)$ defined by* (2.4) *is*

$$\int x^k f_y(x)\,dx = \sum_{r=0}^{k} y^r \binom{k-1}{r}\binom{k}{r}\Big/(r+1). \tag{6.15}$$

Proof. Another calculus exercise, this one somewhat harder than the last. □

6.4 Convergence of Means

Proposition 6.8. *Let X be a random $p \times n$ matrix whose entries $X_{i,j}$ are independent, identically distributed random variables with distribution G with mean zero, variance* 1*, and all moments finite. Define $S = XX^T/n$. As $n, p \to \infty$ with $p/n \to y \in (0,\infty)$,*

$$Ep^{-1}\mathrm{Tr}\, S^k \longrightarrow \sum_{r=0}^{k} y^r \binom{k-1}{r}\binom{k}{r}\Big/(r+1).$$

Proof. This is very similar to the proof of the corresponding result in the Wigner matrix setting. Lemma 6.4, together with a simple argument based on the Hölder inequality, implies that the dominant contribution to the sum (6.6) comes from types τ such that paths γ of type $\tau(\gamma) = \tau$ have graphs $\Gamma(\gamma)$ that are trees, and cross every edge exactly twice, once in each direction. For these types τ, the products $w_S(\tau)$ contain k factors $|X_{i,j}|^2/n^k$; these are independent, with mean 1, so the expectation $Ew_S(\gamma) = 1/n^k$ for each type $\tau \in \mathscr{G}_k^B(r)$. Consequently,

$$E\mathrm{Tr}\, S^k \sim \sum_{r=0}^{k-1} \sum_{\tau \in \mathscr{G}_k^B(r)} H^B(\tau,n,p)/n^k = \sum_{r=0}^{k-1} \sum_{\tau \in \mathscr{G}_k^B(r)} (p)_{r+1}(n)_{k-r}/n^k$$

The result now follows directly from Lemma 6.5. □

7 Truncation Techniques

7.1 Perturbation Inequalities

In section 5 we proved Wigner's theorem under the assumption that the entries $X_{i,j}$ have expectation zero and all moments finite. In this section, we show how these hypotheses can be removed. The main tools are the following inequalities on the difference between the empirical spectral distributions of two Hermitian matrices A, B. In the first, $L(F, G)$ denotes the Lévy distance between the probability distributions F, G (see Background notes).

Proposition 7.1. *(Perturbation Inequality) Let A and B be Hermitian operators on $V = \mathbb{C}^n$ relative to the standard inner product, and let F^A and F^B be their empirical spectral distributions. Then*

$$L(F^A, F^B) \leq n^{-1}\text{rank}(A - B). \tag{7.1}$$

Proposition 7.2. *(Hoffman-Wielandt Inequality) Let A and B be $n \times n$ Hermitian matrices with eigenvalues λ_i and μ_i, respectively, listed in decreasing order. Then*

$$\sum_{i=1}^{n} (\lambda_i - \mu_i)^2 \leq \text{Tr } (A - B)^2. \tag{7.2}$$

See Appendix C below for proofs.

7.2 Wigner's Theorem for Real Symmetric Matrices

Assume now that the entries $X_{i,j}$ are *real*, so that $X_{i,j} = X_{j,i}$,and that the hypotheses of Theorem 1.3 are satisfied. Thus, the common distribution G of the off-diagonal entries $X_{i,j}$ has finite second moment, but the distribution H of the diagonal entries is arbitrary (it need not even have a finite first moment). The first step will be to show that *truncating* either the diagonal or off-diagonal entries has only a small effect on the empirical spectral distribution .

Lemma 7.3. *For a fixed constant $0 < C < \infty$, set*

$$\tilde{X}_{i,i} = X_{i,i} \mathbf{1}_{[-C,C]}(X_{i,i})$$

and define $\tilde{X}$ and $\tilde{M} = \tilde{X}/\sqrt{N}$ to be the matrices obtained from X by changing the diagonal entries $X_{i,i}$ of X to $\tilde{X}_{i,i}$. Then for any $\varepsilon > 0$ there exists $C = C_{\varepsilon,F} < \infty$ such that with probability approaching 1 as $N \to \infty$,

$$\|F^M - F^{\tilde{M}}\|_\infty < \varepsilon. \tag{7.3}$$

Proof. This follows from the perturbation inequality (C.12). For any $\varepsilon > 0$, if C is sufficiently large then with probability approaching one as $N \to \infty$, fewer than εN of the diagonal entries will be changed. Consequently, the difference $M - \tilde{M}$ will have fewer than εN nonzero entries, and hence will be of rank $< \varepsilon N$. Therefore, the perturbation inequality (C.12) implies (7.3). □

Lemma 7.4. *Assume that the distribution G of the off-diagonal entries $X_{i,j}$ has finite second moment. For any constant $0 < C < \infty$, let*

$$\tilde{X}_{i,j} = X_{i,j}\mathbf{1}_{[-C,C]}(X_{i,j}) \quad \textit{for } j \neq i, \tag{7.4}$$

and define $\tilde{X}$ and $\tilde{M} = \tilde{X}/\sqrt{N}$ to be the matrix obtained from X by changing the off-diagonal entries $X_{i,j}$ of X to $\tilde{X}_{i,j}$. Then for any $\varepsilon > 0$ there exists $C = C_{\varepsilon,F} < \infty$ such that with probability approaching 1 as $N \to \infty$,

$$\sup_{x\in\mathbb{R}} |F^M(x+\varepsilon) - F^{\tilde{M}}(x)| < \varepsilon \quad \textit{and} \tag{7.5}$$

$$\sup_{x\in\mathbb{R}} |F^M(x-\varepsilon) - F^{\tilde{M}}(x)| < \varepsilon. \tag{7.6}$$

Remark 7.5. The slightly odd-looking conclusion is a way of asserting that the empirical spectral distributions of M and $\tilde{M}$ are close in the *weak topology* (the topology of convergence in distribution).

Proof. This follows from the trace inequality (C.13). Denote by λ_i and μ_i the eigenvalues of M and $\tilde{M}$, respectively (in decreasing order). The difference $M - \tilde{M}$ has nonzero entries only in those i, j for which $i \neq j$ and $|X_{i,j}| > C$, and so

$$\mathrm{Tr}\,(M-\tilde{M})^2 = N^{-1}\sum_{i\neq j} X_{i,j}^2 \mathbf{1}\{|X_{i,j}| > C\}.$$

Taking expectation gives

$$\begin{aligned} E\mathrm{Tr}\,(M-\tilde{M})^2 &= N^{-1}\sum_{i\neq j} EX_{i,j}^2 \mathbf{1}\{|X_{i,j}| > C\} \\ &\leq N^{-1}N(N-1)EX_{1,2}^2\mathbf{1}\{|X_{1,2}| > C\} \\ &\leq \delta^2 N \end{aligned}$$

provided C is sufficiently large. (This follows because we have assumed that the distribution of the random variables $X_{i,j}$ has finite second moment). The Markov inequality now implies that

$$P\{\mathrm{Tr}\,(M-\tilde{M})^2 \geq \delta N\} \leq \delta.$$

Hence, by the trace inequality (C.13),

$$P\{\sum_{i=1}^{N} (\lambda_i - \mu_i)^2 \geq \delta N\} \leq \delta.$$

This (with $\delta = \varepsilon^2$) implies the inequalities (7.5)-(**??**). (Exercise: Explain why.) □

For any constant $C < \infty$, the truncated random variables $\tilde{X}_{i,i}$ and $\tilde{X}_{i,j}$ defined in Lemmas 7.3 and 7.4 have finite moments of all order. However, they need not have expectation zero, so the version of Wigner's theorem proved in section 5 doesn't apply directly.

Lemma 7.6. *Suppose that Wigner's theorem holds for Wigner matrices $X_{i,i} \sim G$ and $X_{i,j} \sim H$. Then for any constants $a, b \in \mathbb{R}$, the theorem also holds for Wigner matrices with diagonal entries $X_{i,i} + a$ and off-diagonal entries $X_{i,j} + b$.*

Proof. First, consider the effect of adding a constant b to *all* of the entries of X: this changes the matrix to $\tilde{X} = X + b\mathbf{1}$, where $\mathbf{1}$ is the matrix with all entries 1. The matrix $\mathbf{1}$ has rank 1, so the perturbation inequality (C.12) implies that the Lévy distance between the empirical spectral distributions $F^M, F^{\tilde{M}}$ of $M = X/\sqrt{N}$ and $\tilde{M} = \tilde{X}/\sqrt{N}$ is bounded by $1/N$. Hence, if F^M is close to the semi-circle law in Lévy distance, then so is $F^{\tilde{M}}$.

Now consider the effect of adding a constant a to every diagonal entry. This changes X to $\tilde{X} = X + aI$, and thus M to $\tilde{M} = M + aI/N$. Clearly,

$$\operatorname{Tr}\,(M - \tilde{M})^2 = a^2/N;$$

consequently, by Proposition C.13, the eigenvalues λ_i and μ_i (listed in order) satisfy

$$\sum_i |\lambda_i - \mu_i|^2 \le a^2/N.$$

This implies that the Lévy distance between the empirical spectral distributions of M and $\tilde{M}$ converges to 0 as $N \to \infty$. □

Proof of Wigner's Theorem. This is a direct consequence of Lemmas 7.3, 7.4, and 7.6. □

A Weak Convergence

Definition A.1. Let $(\mathscr{X}, d)$ be a complete, separable metric space (also known as a *Polish space*). The *Borel σ−algebra* on $\mathscr{X}$ is the minimal σ−algebra containing the open (and hence also closed) subsets of $\mathscr{X}$. If μ_n and μ are finite Borel measures on $\mathscr{X}$, then μ_n converges *weakly* (or *in distribution*), written

$$\mu_n \Longrightarrow \mu,$$

if for every bounded, continuous function $f : \mathscr{X} \to \mathbb{R}$,

$$\lim_{n\to\infty} \int f\, d\mu_n = \int f\, d\mu. \tag{A.1}$$

Proposition A.2. *(Weierstrass) Let $\mathscr{X}$ be a compact subset of $\mathbb{R}^d$. Then the polynomial functions in d variables on $\mathscr{X}$ are uniformly dense in $C(\mathscr{X})$, that is, for every function $f \in C(\mathscr{X})$ and every $\varepsilon > 0$ there is a polynomial $p(x) = p(x_1, x_2, \dots, x_d)$ such that*

$$\|f - p\|_\infty < \varepsilon.$$

Proposition A.3. *Let μ, ν be finite Borel measures both with compact support $K \subset \mathbb{R}^d$. If μ and ν have the same moments, that is, if for every monomial $x_1^{n_1} x_2^{n_2} \dots x_d^{n_d}$,*

$$\int x_1^{n_1} x_2^{n_2} \dots x_d^{n_d} \, d\mu = \int x_1^{n_1} x_2^{n_2} \dots x_d^{n_d} \, d\nu, \tag{A.2}$$

then they are equal as measures. More generally, if μ and ν are finite positive Borel measures on $\mathbb{R}^d$ with finite moment generating functions in a neighborhood of the origin, then equality of moments (A.2) *implies that $\mu = \nu$.*

Remark A.4. In general, a probability measure is *not* uniquely determined by its moments, even when they are all finite. For further information on this subject, see, e.g., Rudin, *Real and Complex Analysis*, chapter on the Denjoy-Carleman theorem.

Proof. Consider first the case where both μ and ν are supported by a compact set K. Equality of moments (A.2) implies that $\int f \, d\mu = \int f \, d\nu$ for every polynomial f, and therefore, by the Weierstrass theorem, for all continuous functions f on K. This in turn implies that $\mu(F) = \nu(F)$ for every rectangle F, by an easy approximation argument, and therefore for every Borel set F.

Now consider the case where both measures have finite moment generating functions in a neighborhood of 0, that is,

$$\int e^{\theta^T x} \, d\mu(x) + \int e^{\theta^T x} \, d\nu(x) < \infty$$

for all θ in a neighborhood of the origin in $\mathbb{R}^d$. In this case the moment generating functions extend to complex arguments $z = (z_1, z_2, \dots, z_d)$, and define holomorphic (see Remark A.5 below) functions of d variables:

$$\varphi_\mu(z) := \int e^{z^T x} \, d\mu(x) \quad \text{and} \quad \varphi_\nu(z) := \int e^{z^T x} \, d\nu(x). \tag{A.3}$$

Equality of moments implies that these two holomorphic functions have the same power series coefficients, and therefore are equal in a neighborhood of zero. It now follows that the two measures are the same (e.g., by the Fourier inversion theorem). □

Remark A.5. That the functions $\varphi_\mu(z)$ and φ_ν defined by (A.3) are holomorphic in their arguments $z_1, z_2, \dots, z_d$ is a consequence of the *Cauchy, Morrera,* and *Fubini* theorems, by a standard argument, using the fact that $e^{z^T x}$ is holomorphic in z for each x. The Morrera theorem states that a function $f(\zeta)$ is holomorphic in a domain $\Omega \subset \mathbb{C}$ if it integrates to 0 on every closed curve γ that is contractible in Ω. The Cauchy integral theorem asserts that if $f(\zeta)$ is holomorphic in a domain Ω then it integrates to zero on every closed curve γ that is contractible in Ω. Thus, if $f(\zeta, x)$ is holomorphic in ζ for each $x \in \mathscr{X}$, and μ is a Borel measure on $\mathscr{X}$, then $\int_{\mathscr{X}} f(\zeta, x)\, d\mu(x)$ is holomorphic in ζ provided the conditions of the Fubini theorem are satisfied.

Proposition A.6. *(Method of Moments) Let μ be a probability measure on $\mathbb{R}^d$ with compact support K, and let μ_n be a sequence of probability measures on $\mathbb{R}^d$ whose moments converge as $n \to \infty$ to the moments of μ. Then*

$$\mu_n \Longrightarrow \mu.$$

Proof. Convergence of moments implies that for every *polynomial* $p(\mathbf{x}) = p(x_1, x_2, \cdots, x_d)$,

$$\lim_{n\to\infty} \int p(\mathbf{x})\, d\mu_n(\mathbf{x}) = \int p(\mathbf{x})\, d\mu(\mathbf{x})$$

The density of polynomials in the space $C(K)$ of continuous functions on K (Weierstrass' theorem) and an elementary argument now show that for every continuous function $f : K \to \mathbb{R}$,

$$\lim_{n\to\infty} \int f(\mathbf{x})\, d\mu_n(\mathbf{x}) = \int f(\mathbf{x})\, d\mu(\mathbf{x}).$$

□

The method of moments is the most useful tool for proving convergence of empirical spectral distributions in the theory of random matrices. There is one other useful tool, the *Stieltjes transform*:

Definition A.7. Let μ be a finite, positive Borel measure on $\mathbb{R}$. The Stieltjes transform $F_\mu(z)$ is the holomorphic (see Remark A.5) function of $z \in \mathbb{C} \setminus \mathbb{R}$ defined by

$$F_\mu(z) = \int (x-z)^{-1}\, d\mu(x).$$

Remark A.8. Since the probability measure μ is supported by $\mathbb{R}$, the Stieltjes transform satisfies

$$F_\mu(\bar{z}) = \overline{F_\mu(z)}.$$

Proposition A.9. *Let μ_n be a tight sequence of probability measures on $\mathbb{R}$. If the Stieltjes transforms $F_n(z)$ of the measures μ_n converge for all z in a set $A \subset \mathbb{C} \setminus \mathbb{R}$ with an accumulation point in $\mathbb{C} \setminus \mathbb{R}$ to a limit function $F(z)$ defined on A, then the sequence μ_n converges weakly to a probability measure whose Stieltjes transform agrees with $F(z)$ on A.*

Proof. Since the sequence μ_n is tight, every subsequence has a weakly convergent subsequence. For every convergent subsequence μ_k, the Stieltjes transforms $F_k(z)$ converge to the Stieltjes transform of the limit measure for all $z \in \mathbb{C} \setminus \mathbb{R}$, because $x \mapsto (x-z)^{-1}$ is a bounded, continuous function of $x \in \mathbb{R}$. (Note: This function is complex-valued, but its real and imaginary parts will be bounded, continuous, real-valued functions.) Thus, the limit measure must have a Stieltjes transform that agrees with $F(z)$ on the set A. Since the set A has an accumulation point, analyticity guarantees that the Stieltjes transform of the limit measure is uniquely determined by its values $F(z)$ on A. □

B Matrix Theory: Trace of a Square Matrix

The *trace* of a square matrix $M = (m_{i,j})$, denoted by tr(M), is the sum $\sum_i m_{i,i}$ of its diagonal entries. The following is elementary but important:

Proposition B.1. *Let $A = (a_{i,j})_{m\times n}$ and $B = (b_{i,j})_{n\times m}$ be real or complex square matrices. Then*

$$tr(AB) = tr(BA) = \sum_{i=1}^{m}\sum_{j=1}^{n} a_{i,j} b_{j,i}. \tag{B.1}$$

Consequently, if A and B are similar square matrices, that is, if there is an invertible matrix U such that $B = U^{-1}AU$, then

$$tr(A) = tr(B). \tag{B.2}$$

A square matrix A is *diagonalizable* if there exist an invertible matrix U amd a diagonal matrix D such that $A = U^{-1}DU$. You should recall that A is diagonalizable if and only if the underlying vector space has a basis consisting of eigenvectors of A; in this case, the diagonalization $A = U^{-1}DU$ is obtained by letting U be the matrix whose columns are the eigenvectors, and D the diagonal matrix whose diagonal entries are the corresponding eigenvalues. Thus, for diagonalizable A, the trace tr(A) is just the sum of the eigenvalues. More generally:

Corollary B.2. *Assume that A is diagonalizable. Then for any integer $k \geq 0$,*

$$tr(A^k) = \sum_i \lambda_i^k. \tag{B.3}$$

Furthermore, for any analytic function $f(z)$ defined by a power series $f(z) = \sum_{k=0}^{\infty} a_k z^k$ with radious of convergence $R > 0$, if A has spectral radius $< R$ (that is, if all eigenvalues of A have absolute value $< R$) then

$$tr(f(A)) = \sum_i f(\lambda_i). \tag{B.4}$$

Both assertions are easy consequences of Proposition B.1. Relation (B.3) is especially useful in conjunction with the method of moments, as it gives an effective way of computing the moments of the *empirical spectral distribution* (defined below). Similarly, relation (B.4) gives a handle on various transforms of the empirical spectral distribution, in particular, the Stieltjes transform.

Definition B.3. If A is a diagonalizable $n \times n$ matrix with eigenvalues λ_i, the *empirical spectral distribution* F^A of A is defined to be the uniform distribution on the eigenvalues (counted according to multiplicity), that is

$$F^A := n^{-1} \sum_{i=1}^{n} \delta_{\lambda_i}. \tag{B.5}$$

Relation (B.3) implies

$$\boxed{\int \lambda^k F^A(d\lambda) = n^{-k} \mathrm{tr}(A^k).} \tag{B.6}$$

C Hermitian, Unitary, and Orthogonal Matrices

C.1 Spectral Theorems

A *Hermitian matrix* is a square complex matrix that equals its conjugate transpose. A matrix with *real* entries is therefore Hermitian if and only if it is *symmetric.* A *unitary matrix* is a square complex matrix whose conjugate transpose is its inverse. In spectral problems, it is often advantageous to take a coordinate-free approach, using an *inner product* to define Hermitian and unitary operators. Thus, let V be a complex vector space, and let $\langle \cdot, \cdot \rangle$ be a complex inner product on V. Recall the definition (see Rudin, *Real and Complex Analysis,* ch. 4):

Definition C.1. An inner product on a complex vector space V is a mapping $V \times V \to \mathbb{C}$ that satisfies

(a) $\langle u, v \rangle = \overline{\langle v, u \rangle}$.
(b) $\langle \alpha u + \alpha' u', v \rangle = \alpha \langle u, v \rangle + \alpha' \langle u', v \rangle$.
(c) $\langle u, u \rangle > 0$ for all $u \neq 0$.
(d) $\langle 0, 0 \rangle = 0$.

The difference between the real and complex cases is rule (a); this together with (b) implies that $\langle u, \alpha v \rangle = \bar{\alpha} \langle u, v \rangle$. The natural inner product on $\mathbb{C}^n$ is $\langle u, v \rangle = \sum_i u_i \bar{v}_i$. More generally, the natural inner product on the space $L^2(\mu)$ of square-integrable complex-valued functions on a measure space $(\Omega, \mathscr{F}, \mu)$ is

$$\langle f, g \rangle = \int f \bar{g} \, d\mu;$$

when μ is a probability measure this is a complex analogue of the covariance. An *inner product space* is a vector space equipped with an inner product. Two vectors u, v in an inner product space are said to be *orthogonal* if $\langle u, v\rangle = 0$. An *orthonormal set* is a set of unit vectors such that any two are orthogonal. You should recall that the *Gram-Schmidt algorithm* produces orthonormal bases.

Definition C.2. Let V be a finite-dimensional inner product space. For any operator (i.e., linear transformation) $T : V \to V$ the *adjoint* T^* is the unique linear transformation such that

$$\langle Tu, v\rangle = \langle u, Tv\rangle \quad \forall\, u, v \in V. \tag{C.1}$$

The linear transformation T is called *Hermitian* if $T = T^*$, and *unitary* if $T^{-1} = T^*$, equivalently,

$$\langle Tu, v\rangle = \langle u, Tv\rangle \quad \text{(Hermitian)} \tag{C.2}$$

$$\langle Tu, Tv\rangle = \langle u, v\rangle \quad \text{(Unitary)} \tag{C.3}$$

Theorem C.3. *(Spectral Theorem for Hermitian Operators) Let T be a Hermitian operator on a finite-dimensional inner product space V. Then all eigenvalues λ_i of T are real, and there is an orthonormal basis $\{u_i\}$ consisting of eigenvectors of T. Thus,*

$$Tv = \sum_i \lambda_i \langle v, u_i\rangle u_i \quad \forall\, v \in V. \tag{C.4}$$

Theorem C.4. *(Spectral Theorem for Unitary Operators) Let T be a unitary operator on a finite-dimensional inner product space V. Then all eigenvalues λ_i of T have absolute value* 1*, and there is an orthonormal basis $\{u_i\}$ consisting of eigenvectors of T. Thus,*

$$Tv = \sum_i \lambda_i \langle v, u_i\rangle u_i \quad \forall\, v \in V. \tag{C.5}$$

Some of the important elements of the proofs are laid out below. Consider first the case where T is Hermitian. Suppose that $Tv = \lambda v$; then

$$\begin{aligned}
\lambda\langle v, v\rangle &= \langle \lambda v, v\rangle \\
&= \langle Tv, v\rangle \\
&= \langle v, Tv\rangle \\
&= \langle v, \lambda v\rangle \\
&= \bar{\lambda}\langle v, v\rangle.
\end{aligned}$$

Thus, $\lambda = \bar{\lambda}$, so all eigenvalues of T are real. A similar argument shows that eigenvalues of unitary operators must be complex numbers of absolute value 1.

Next, there is the notion of an *invariant subspace*: a linear subspace W of V is invariant for T if $TW \subset W$. If T is Hermitian (respectively, unitary) and W is an invariant subspace, then the restriction $T|W$ of T to W is also Hermitian (respectively, unitary). Also, if T is invertible, as is the case if T is unitary, then W is an invariant subspace if and only if $TW = W$. The following is an easy exercise:

Proposition C.5. *Let T be either Hermitian or unitary. If T has an invariant subspace W, then the orthogonal complement*[2] *$W^{\perp}$ of W is also an invariant subspace for T.*

Proof of Theorems C.3–C.4. The proof is by induction on the dimension of V. Dimension 1 is trivial. Now every linear operator T on a complex vector space V has at least one eigenvector v. (Proof: The characteristic polynomial $\det(\lambda I - T)$ has a zero, since $\mathbb{C}$ is algebraically complete. For any such root λ the linear transformation $\lambda I - T$ must be singular, by Proposition **??**, and so the equation $(\lambda I - T)v = 0$ has a nonzero solution v.) If v is an eigenvector of T, then the one-dimensional subspace of V spanned by v is invariant, and so its orthogonal complement W is also invariant. But dimension(W) is less than dimension(V), so the induction hypothesis applies to $T|W$: in particular, W has an orthonormal basis consisting of eigenvectors of T. When augmented by the vector $v/\sqrt{\langle v, v\rangle}$, this gives an orthonormal basis of V made up entirely of eigenvectors of T. □

C.2 Orthogonal Matrices: Spectral Theory

An *orthogonal matrix* is a unitary matrix whose entries are real, equivalently, a real matrix whose transpose is its inverse. Because an orthogonal matrix is unitary, the Spectral Theorem for unitary operators implies that its eigenvalues are complex numbers of modulus 1, and that there is an orthonormal basis of eigenvectors. This doesn't tell the whole story, however, because in many circumstances one is interested in the action of an orthonormal matrix on a *real* vector space. An orthogonal linear transformation of a real inner product space need not have real eigenvectors: for instance, the matrix

$$R_\theta := \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

acting on $\mathbb{R}^2$ has no nonzero eigenvectors unless θ is an integer multiple of 2π, because R_θ rotates every nonzero vector through an angle of θ.

Lemma C.6. *Let T be an orthogonal $n \times n$ matrix, and let v be a (possibly complex) eigenvector with eigenvalue λ. Then the complex conjugate $\bar{v}$ is also an eigenvector, with eigenvalue $\bar{\lambda}$.*

The proof is trivial, but the result is important because it implies the following structure theorem for orthogonal linear transformations of real inner product spaces.

Corollary C.7. *Let T be an orthogonal $n \times n$ matrix acting on the real inner product space $\mathbb{R}^n$. Then $\mathbb{R}^n$ decomposes as an orthogonal direct sum of one- or two-dimensional*

[2] The orthogonal complement $W^{\perp}$ is defined to be the set of all vectors u such that u is orthogonal to every $w \in W$.

invariant subspaces for T, on each of which T acts as a rotation matrix R_θ. In other words, in a suitable orthonormal basis T is represented by a matrix in block form (where all but the last two blocks are of size 2×2*)*

$$\begin{pmatrix} R_{\theta_1} & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & R_{\theta_2} & 0 & \cdots & 0 & 0 & 0 \\ & \cdots & & & & & \\ 0 & 0 & 0 & \cdots & R_{\theta_k} & 0 & 0 \\ 0 & 0 & 0 & \cdots & 0 & -I & 0 \\ 0 & 0 & 0 & \cdots & 0 & 0 & I \end{pmatrix}$$

Proof. The only possible real eigenvalues are ±1. On the space of eigenvectors with eigenvector $+1$ the matrix T acts as the identity, and on the space of eigenvectors with eigenvector $+1$ the matrix T acts as $(-1)\times$ the identity. Consequently, each of these subspaces splits as a direct sum of one-dimensional subspaces.

Let $v = u + iw$ be a complex eigenvector with real and imaginary parts u, w and eigenvalue $\lambda = e^{i\theta}$. By Lemma C.6, $\bar{v} = u - iw$ is an eigenvector with eigenvalue $e^{-i\theta}$. Adding and subtracting the eigenvector equations for these two eigenvectors shows that the two-dimensional real subspace of $\mathbb{R}^n$ spanned by u, w is invariant for T, and that the restriction of T to this subspace is just the rotation by θ. It is routine to check that the two-dimensional invariant subspaces obtained in this manner are mutually orthogonal, and that each is orthogonal to the one-dimensional invariant subspaces corresponding to eigenvalues ±1. □

C.3 Minimax Characterization of Eigenvalues

Let $T : V \to V$ be a Hermitian operator on a finite-dimensional vector space V of dimension n. According to the Spectral Theorem C.3, the eigenvalues λ_i of T are real, and there is an orthonormal basis consisting of eigenvectors u_i. Because the eigenvectors are real, they are linearly ordered:

$$\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n. \tag{C.6}$$

Proposition C.8.

$$\lambda_n = \max_{u:|u|=1} \langle Tu, u\rangle \quad \textit{and} \tag{C.7}$$

$$\lambda_1 = \min_{u:|u|=1} \langle Tu, u\rangle \tag{C.8}$$

Proof. Any unit vector u has an expansion $u = \sum_i \alpha_1 u_i$ in the eigenvectors of T, where the (complex) scalars α_i satisfy $\sum_i |\alpha_i|^2 = 1$. It follows that

$$Tu = \sum_i \alpha_i \lambda_i u_i \quad \Longrightarrow \quad \langle Tu, u\rangle = \sum_i \lambda_i |\alpha_i|^2$$

Since u is a unit vector, the assignment $i \mapsto |\alpha_i|^2$ is a probability distribution on the index set $[n]$. Clearly, the probability distribution that maximizes the expectation $\sum_i \lambda_i |\alpha_i|^2$ puts all of its mass on the indices i for which λ_i is maximal. Thus, the maximal expectation is λ_n. Similarly, the minimum expectation is λ_1. □

The minimax characterization is a generalization of Proposition C.8 to the entire spectrum. This characterization is best described using the terminology of game theory. The game is as follows: First, I pick a linear subspace $W \subset V$ of dimension k; then you pick a unit vector u in W. I pay you $\langle Tu, u\rangle$. If we both behave rationally (not always a sure thing on my end, but for the sake of argument let's assume that in this case I do) then I should choose the subspace spanned by the eigenvectors $u_1, u_2, \dots, u_k$, and then you should choose $u = u_k$, so that the payoff is λ_k. That this is in fact the optimal strategy is the content of the minimax theorem:

Theorem C.9. *(Minimax Characterization of Eigenvalues)*

$$\begin{aligned} \lambda_k &= \min_{W:\dim(W)=k} \; \max_{u\in W:|u|=1} \langle Tu, u\rangle \qquad \text{(C.9)} \\ &= \max_{W:\dim(W)=n-k} \; \min_{u\in W:|u|=1} \langle Tu, u\rangle \end{aligned}$$

Proof. The second equality is obtained from the first by applying the result to the Hermitian operator $-T$, so only the first equality need be proved. It is clear that the right side of (C.9) is no larger than λ_k (see the comments preceding the statement of the theorem), so what must be proved is that for any subspace W of dimension k,

$$\max_{u\in W:|u|=1} \langle Tu, u\rangle \ge \lambda_k.$$

Let u_i be an orthonormal basis of V such that $Tu_i = \lambda_i u_i$, where the eigenvalues λ_i are arranged in increasing order as in (C.6). Let U be the linear subspace of V spanned by the vectors $u_k, u_{k+1}, \cdots, u_n$; since the vectors u_i are linearly independent, the subspace U has dimension $n-k+1$. Hence,

$$\dim(W) + \dim(U) = k + (n-k+1) = n+1 > \dim(V),$$

and so the subspaces U and W must have a nonzero vector in common, and consequently a *unit* vector u in common. Since $u \in U$, it is a linear combination of $u_k, u_{k+1}, \cdots, u_n$:

$$u = \sum_{i=k}^{n} a_i u_i \quad \text{where} \quad \sum_{i=k}^{n} a_i^2 = 1.$$

Consequently,

$$\langle Tu, u\rangle = \sum_{i=k}^{n} \lambda_i a_i^2;$$

since the eigenvalues λ_i are listed in increasing order, and since the coefficients a_i^2 constitute a probability distribution on the indices $i = k, k+1, \cdots, n$, it follows that

$$\langle Tu, u\rangle \geq \lambda_k.$$

□

Corollary C.10. *(Eigenvalue interlacing) Let $T : V \to V$ be Hermitian, and let $W \subset V$ be a linear subspace of dimension $n-1$, where $n = \dim(V)$. Then the eigenvalues of the restriction $T|W$ are interlaced with those of T on V: that is, if the eigenvalues of T are $\lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n$ and the eigenvalues of $T|W$ are $\mu_1 \leq \mu_2 \leq \cdots \leq \mu_{n-1}$, then*

$$\lambda_1 \leq \mu_1 \leq \lambda_2 \leq \mu_2 \leq \cdots \leq \lambda_{n-1} \leq \mu_{n-1} \leq \lambda_n. \tag{C.10}$$

Proof. It suffices to prove that $\mu_k \geq \lambda_k$, because the reverse inequalities then follow by considering $-T$. By Theorem C.9,

$$\begin{aligned} \lambda_k &= \min_{S\subset V:\dim(S)=k} \max_{u\in S:|u|=1} \langle Tu, u\rangle \quad \text{and} \\ \mu_k &= \min_{S\subset W:\dim(S)=k} \max_{u\in S:|u|=1} \langle Tu, u\rangle, \end{aligned}$$

where the minima are taken over linear subspaces S of V and W, respectively. Since $W \subset V$, every linear subspace of W is a linear subspace of V, and so the first min is taken over a larger collection than the second. Thus, $\mu_k \geq \lambda_k$. □

C.4 Empirical spectral distributions

Recall that the *empirical spectral distribution* of a diagonalizable matrix is the uniform distribution on eigenvalues (counted according to multiplicity). The empirical spectral distribution is the object of primary interest in the study of random matrices. Thus, it is useful to know how changes in the entries of a matrix affect the empirical spectral distribution. In this section we give two useful bounds on the magnitude of the change in the empirical spectral distribution under certain types of matrix perturbations.

Definition C.11. The *Lévy distance* between two probability distributions μ, ν on $\mathbb{R}$ with cumulative distribution functions F_μ and F_ν is defined to be

$$L(\mu, \nu) := \inf\{\varepsilon > 0 : F(x-\varepsilon) - \varepsilon \leq G(x) \leq F(x+\varepsilon) + \varepsilon\}. \tag{C.11}$$

Observe that if $\|F - G\|_\infty < \varepsilon$ then $L(F, G) < \varepsilon$; the converse, however, need not hold. Moreover, the Lévy distance characterizes convergence in distribution, that is, $\lim_{n\to\infty} L(F_n, F) = 0$ if and only if $F_n \Longrightarrow F$.

Corollary C.12. *(Perturbation Inequality) Let A and B be Hermitian operators on $V = \mathbb{C}^n$ relative to the standard inner product, and let F^A and F^B be their empirical spectral distributions. Then*

$$L(F^A, F^B) \leq n^{-1}\mathrm{rank}(A-B). \tag{C.12}$$

Proof. It suffices to prove this for Hermitian operators A, B that differ by a rank-1 operator $\Delta = A - B$, because operators that differ by a rank-k operator can be connected by a chain of $k+1$ Hermitian operators whose successive differences are all rank-1. The operator Δ is Hermitian, so if it is rank-1 then it has a single nonzero eigenvalue δ, with corresponding eigenvector w. Let W be the $(n-1)-$dimensional subspace orthogonal to w; then $\Delta|W = 0$ and so the restrictions $A|W$ and $B|W$ are identical. Let $\mu_1 \leq \cdots \leq \mu_{n-1}$ be the eigenvalues of $A|W = B|W$. By the Eigenvalue Interlacing Theorem (Corollary C.10), the (ordered) eigenvalues λ_i^B of B are interlaced with the sequence μ_i, and so are the eigenvalues λ_i^A of A. Consequently, it is impossible for either

$$\lambda_{k+2}^A < \lambda_k^B \quad \text{or} \quad \lambda_{k+2}^B < \lambda_k^A$$

to occur for any k. □

Proposition C.13. *(Hoffman-Wielandt Inequality) Let A and B be $n \times n$ Hermitian matrices with eigenvalues λ_i and μ_i, respectively, listed in decreasing order. Then*

$$\sum_{i=1}^{n} (\lambda_i - \mu_i)^2 \leq \mathrm{Tr}\,(A-B)^2. \tag{C.13}$$

Proof. Expand the squares on both sides of the inequality to obtain the equivalent statement

$$\sum_i (\lambda_i^2 - 2\lambda_i\mu_i + \mu_i^2) \leq \mathrm{Tr}\,A^2 + \mathrm{Tr}\,B^2 - 2\mathrm{Tr}\,AB \tag{C.14}$$

(this also uses the fact that Tr AB = Tr BA). SInce Tr $A^2 = \sum \lambda_i^2$ and Tr $B = \sum \mu_i^2$, proving inequality (C.14) is tantamount to proving

$$\mathrm{Tr}\,AB \leq \sum \lambda_i\mu_i. \tag{C.15}$$

Neither the trace nor the spectrum of a diagonizable matrix depends on which basis for the vector space is used. Consequently, we can work in the orthonormal basis of eigenvectors of A, that is to say, we may assume that A is diagonal. Since B is also Hermitian, there is a unitary matrix U that diagonizes B. Thus,

$$A = \mathrm{diag}(\lambda_1, \lambda_2, \ldots, \lambda_n) \quad \text{and}$$
$$B = U\mathrm{diag}(\mu_1, \mu_2, \ldots, \mu_n)U^*,$$

and so (C.15) is equivalent to

$$\sum_i \sum_j \lambda_i\mu_j |U_{i,j}|^2 \leq \sum \lambda_i\mu_i \tag{C.16}$$

Now U is unitary, so the matrix $(|U_{i,j}|^2)_{i,j}$ is doubly stochastic (U unitary means that the rows and columns are orthonormal). Hence, to prove inequality (C.16) it will suffice to prove that

$$\sum_i \sum_j \lambda_i \mu_j p_{i,j} \leq \sum_i \lambda_i \mu_i \tag{C.17}$$

where $p_{i,j}$ is any doubly stochastic matrix.

There are various ways to prove (C.17). Following is a short and painless proof that uses the *Birkhoff-von Neumann theorem* on doubly stochastic matrices. This theorem states that *every doubly stochastic matrix is a convex combination of permutation matrices.* To see how the Birkhoff-von Neumann theorem applies to (C.17), consider the problem of maximizing the left side over all doubly stochastic matrices $p_{i,j}$. Since the left side is linear in the variables $p_{i,j}$, the Birkhoff-von Neumann theorem implies that to prove (C.17) it suffices to check that

$$\max_\sigma \sum_i \lambda_i \mu_{\sigma(i)} = \sum_i \lambda_i \mu_i,$$

where the max is over the set of all permutations σ. For this, just check that if σ is not the identity permutation, then the left side can be increased (or at least not decreased) by switching two indices i, i' where $\lambda_i, \lambda_{i'}$ and $\mu_i, \mu_{i'}$ are in opposite relative order. □

Exercise C.14. (Challenging) Prove Birkhoff theorem, that is, show that every doubly stochastic matrix can be written as a convex combination of permutation matrices.

Example:

$$\begin{pmatrix} .3 & .7 \\ .7 & .3 \end{pmatrix} = .3 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix} + .7 \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$