Random matrix theory - an introduction

May 22, 2024

1 Matrices and their spectra

1.1 Linear operators and their matrices

Let V be a finite-dimensional linear space over $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$. If $\{e(1), \dots, e(n)\}$ is a basis for V and $\phi \in V$, then there exist unique $\phi_1, \dots, \phi_n \in \mathbb{K}$ such that

$$\phi = \sum_{i=1}^{n} \phi_i e(i).$$

Let V, W be finite-dimensional linear spaces over $\mathbb{K}$ equipped with bases $\{e(1), \dots, e(n)\}$ and $\{f(1), \dots, f(m)\}$. Let $\mathcal{L}(V, W)$ be the space of linear operators $A : V \to W$. Then for each $A \in \mathcal{L}(V, W)$ there exist unique $A_{ij} \in \mathbb{K}$ with $1 \le i \le m$ and $1 \le j \le n$ such that

$$(A\phi)_i = \sum_{j=1}^{n} A_{ij}\phi_j \qquad (1 \le i \le m).$$

We call $(A_{ij})_{1\le i\le m,\ 1\le j\le n}$ the *matrix* of A and we call A_{ij} the (i, j)-th *entry* of the matrix. One has

$$(AB)_{ik} = \sum_{j} A_{ij}B_{jk}.$$

The *trace* of an operator $A \in \mathcal{L}(V) := \mathcal{L}(V, V)$

$$\mathrm{tr}(A) := \sum_{i} A_{ii}$$

does not depend on the choice of the basis and satisfies $\mathrm{tr}(AB) = \mathrm{tr}(BA)$. Let $A \in \mathcal{L}(V)$. By definition, $0 \neq \phi \in V$ is an *eigenvector* with *eigenvalue* $\lambda \in \mathbb{K}$ if $A\phi = \lambda\phi$. We call

$$\sigma(A) := \big\{\lambda : \lambda \text{ is an eigenvalue of } A\big\}$$

the *spectrum* of A. One has

$$\sigma(A) := \big\{\lambda : (\lambda - A) \text{ is not invertible}\big\}.$$

Lemma 1 (Nonempty spectrum) *Assume that $A \in \mathcal{L}(V)$ and $\mathbb{K} = \mathbb{C}$. Then $\sigma(A) \neq \emptyset$.*

Proof (sketch) One has $\sigma(A) := \big\{\lambda : \det(\lambda - A) = 0\big\}$. The equation $\det(\lambda - A) = 0$ is a polynomial in λ of degree $\dim(V)$ which is guaranteed to have at least one complex root. ∎

1.2 Inner product spaces

Let V be a finite-dimensional linear space over $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$. An *inner product* on V is a map that assigns to two vectors $\phi, \psi \in V$ a number $\langle \phi, \psi \rangle \in \mathbb{K}$ such that

(i) $\psi \mapsto \langle \phi, \psi \rangle$ is linear,

(ii) $\langle \phi, \psi \rangle = \overline{\langle \psi, \phi \rangle}$,

(iii) $\langle \phi, \phi \rangle \geq 0$,

(iv) $\langle \phi, \phi \rangle = 0 \quad \Rightarrow \quad \phi = 0$.

Here $\bar{c}$ denotes the complex conjugate of a number $c \in \mathbb{K} = \mathbb{R}$ or $\mathbb{C}$. In the complex case, $\phi \mapsto \langle \phi, \psi \rangle$ is not linear but *colinear* in the sense that

$$\langle c_1 \phi(1) + c_2 \phi(2), \psi \rangle = \bar{c}_1 \langle \phi(1), \psi \rangle + \bar{c}_2 \langle \phi(2), \psi \rangle.$$

The norm associated with the inner product is $|\phi| := \sqrt{\langle \phi, \phi \rangle}$. A basis $\{e(1), \ldots, e(n)\}$ is *orthogonal* if $\langle e(i), e(j) \rangle = 0$ for $i \neq j$ and *orthonormal* if in addition $\langle e(i), e(i) \rangle = 1$ for each i. For each $\phi \in V$ we define $\langle \phi | \in \mathcal{L}(V, \mathbb{K})$ and $|\phi\rangle \in \mathcal{L}(\mathbb{K}, V)$ by

$$\langle \phi | \psi := \langle \phi, \psi \rangle \quad \text{and} \quad |\phi\rangle c := c\phi.$$

The space $V' := \mathcal{L}(V, \mathbb{K})$ is called the *dual linear space* of V. On the other hand, $\mathcal{L}(\mathbb{K}, V)$ can naturally be identified with V itself. Also, $\langle \phi | \, | \psi \rangle \in \mathcal{L}(\mathbb{K}, \mathbb{K}) \cong \mathbb{K}$ can be identified with the number $\langle \phi, \psi \rangle$. For $A \in \mathcal{L}(V)$ one has

$$A|\psi\rangle = |A\psi\rangle.$$

The coordinates of a vector and operator with respect to an *orthonormal* basis are given by

$$\phi_i = \langle e(i), \phi \rangle \quad \text{and} \quad A_{ij} = \langle e(i), Ae(j) \rangle = \langle e(i) | A | e(j) \rangle.$$

Note that $|\phi\rangle\langle\psi| \in \mathcal{L}(V, V) = \mathcal{L}(V)$. One has

$$A = \sum_{i,j} A_{ij} |e(i)\rangle\langle e(j)|,$$

and the operators $|e(i)\rangle\langle e(j)|$ form a basis for $\mathcal{L}(V)$. In particular

$$1 = \sum_i |e(i)\rangle\langle e(i)| \tag{1}$$

is the identity operator.

Each $A \in \mathcal{L}(V)$ has a unique *adjoint* $A^* \in \mathcal{L}(V)$ such that

$$\langle \phi, A\psi \rangle = \langle A^* \phi, \psi \rangle \qquad (\phi, \psi \in V).$$

the map $A \mapsto A^*$ is colinear with

$$(A^*)^* = A \quad \text{and} \quad (AB)^* = B^* A^*.$$

In coordinates

$$A^*_{ij} = \overline{A_{ji}}.$$

An operator A is *normal* if it commutes with its adjoint:

$$AA^* = A^* A.$$

We say that A is *hermitian* if $A^* = A$.

Theorem 2 (Diagonalisation of normal operators) *Assume that* $\mathbb{K} = \mathbb{C}$. *Then* $A \in \mathcal{L}(V)$ *is normal if and only if there exists an orthonormal basis* $\{e(1), \ldots, e(n)\}$ *and complex numbers* $\lambda_1, \ldots, \lambda_n$ *such that*

$$A = \sum_i \lambda_i |e(i)\rangle\langle e(i)|. \tag{2}$$

Moreover, A *is hermitian if and only if* $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$.

Proof It is straightforward to check an operator of the form (2) is normal, and hermitian if and only if $\lambda_1, \ldots, \lambda_n \in \mathbb{R}$. It remains to show each normal operator can be written in the form (2). Assume A is normal. Then for each $\phi \in V$,

$$\langle A^*\phi, A^*\phi\rangle = \langle \phi, AA^*\phi\rangle = \langle \phi, A^*A\phi\rangle = \langle A\phi, A\phi\rangle,$$

which shows that

$$|A^*\phi| = |A\phi|.$$

Since $\mathbb{K} = \mathbb{C}$, the operator A has at least one eigenvector ϕ with some eigenvalue λ. Then

$$A\phi = \lambda\phi \quad \Rightarrow \quad |(A-\lambda)\phi| = 0 \quad \Rightarrow \quad |(A-\lambda)^*\phi| = 0 \quad \Rightarrow \quad A^*\phi = \lambda^*\phi.$$

Let $\{\phi\}^\perp := \{\psi \in V : \langle \phi, \psi\rangle = 0\}$. Then

$$\psi \in \{\phi\}^\perp \quad \Rightarrow \quad \langle \phi, A\psi\rangle = \langle A^*\phi, \psi\rangle = \lambda\langle \phi, \psi\rangle = 0 \quad \Rightarrow \quad A\psi \in \{\phi\}^\perp.$$

Now A restricted to $\{\phi\}^\perp$, is again a normal operator so repeating the argument we can find an orthonormal basis of eigenvectors. ∎

Although our proof of Theorem 2 used the complex numbers in an essential way, for *symmetric* real matrices one can prove something similar.

Theorem 3 (Diagonalisation of symmetric matrices) *Let* V *be a real vector space and let* $A \in \mathcal{L}(V)$ *satisfy* $A^* = A$. *Then there exists an orthonormal basis* $\{e(1), \ldots, e(n)\}$ *consisting of eigenvectors of* A.

Proof This will follow from the same arguments as in the proof of Theorem 2 provided we show that each symmetric real matrix has at least one eigenvector. By compactness, the function $\phi \mapsto \langle \phi, A\phi\rangle$ assumes its maximum over the ball surface $\{\phi \in V : |\phi| = 1\}$ in some point ψ. This means that in the point ψ the derivatives of the function $\phi \mapsto \langle \phi, A\phi\rangle$ in directions tangential to the ball are all zero. By the method of Lagrange multipliers, there exists a $\lambda \in \mathbb{R}$ such that derivatives of the function $\phi \mapsto \langle \phi, A\phi\rangle - \lambda\langle \phi, \phi\rangle$ in the point ψ are zero in *all* directions. Thus, for our ψ, we can find $\lambda \in \mathbb{R}$ so that

$$\tfrac{\partial}{\partial\varepsilon}\big[\langle \psi + \varepsilon\phi, A(\psi + \varepsilon\phi)\rangle - \lambda\langle \psi + \varepsilon\phi, \psi + \varepsilon\phi\rangle\big]\Big|_{\varepsilon=0} = 0 \qquad (\phi \in V).$$

This gives

$$\langle \psi, A\phi\rangle + \langle \phi, A\psi\rangle - \lambda\langle \psi, \phi\rangle - \lambda\langle \phi, \psi\rangle = 0 \qquad (\phi \in V).$$

Using the fact that $A^* = A$ and dividing out a factor 2, we get

$$\langle \phi, A\psi\rangle - \lambda\langle \phi, \psi\rangle = 0 \qquad (\phi \in V).$$

But this says that $\langle \phi, A\psi - \lambda\psi\rangle = 0$ for all $\phi \in V$, which means that $A\psi = \lambda\psi$, i.e., we have found an eigenvector. ∎

1.3 The operator norm

Let V be a finite-dimensional linear space over $\mathbb{K} = \mathbb{R}$ or $\mathbb{C}$. The *operator norm* of an operator $A \in \mathcal{L}(V)$ is defined as

$$\|A\|_{\mathrm{op}} := \sup_{|\phi| \leq 1} |A\phi|.$$

By linearity, $\|A\|_{\mathrm{op}}$ is the least constant for which that the inequality

$$|A\phi| \leq \|A\|_{\mathrm{op}} \cdot |\phi|$$

holds for all $\phi \in V$. From this, it is easy to see that

$$\|AB\|_{\mathrm{op}} \leq \|A\|_{\mathrm{op}} \cdot \|B\|_{\mathrm{op}}.$$

Lemma 4 (Operator norm and spectrum) *If A is normal and $\mathbb{K} = \mathbb{C}$, then*

$$\|A\|_{\mathrm{op}} = \sup\big\{|\lambda| : \lambda \in \sigma(A)\big\}. \tag{3}$$

Proof Let $C := \sup\big\{|\lambda| : \lambda \in \sigma(A)\big\}$. If ϕ is an eigenvector with eigenvalue λ, then

$$|A\phi| = |\lambda| \cdot |\phi|,$$

so $|\lambda| \leq \|A\|_{\mathrm{op}}$ for each $\lambda \in \sigma(A)$, which implies $\|A\|_{\mathrm{op}} \geq C$. To prove the other inequality, we use that by Theorem 2 there exists an orthonormal basis $\{e(1), \ldots, e(n)\}$ of eigenvectors. Denoting the corresponding eigenvalues by $\lambda_1, \ldots, \lambda_n$, we can write for arbitrary $\phi \in V$,

$$|A\phi|^2 = \Big|\sum_{i=1}^{n} \lambda_i \phi_i e(i)\Big|^2 = \sum_{ij} \overline{\lambda}_i \overline{\phi}_i \lambda_j \phi_j \langle e(i), e(j)\rangle = \sum_i |\lambda_i|^2 \cdot |\phi_i|^2 \leq C^2 \sum_i |\phi_i|^2 = C^2 |\phi|^2.$$

It follows that $|A\phi| \leq C|\phi|$ for all ϕ and hence $\|A\|_{\mathrm{op}} \leq C$. ∎

Remark The assumption in Lemma 4 that A is normal cannot be dropped. Equip $\mathbb{R}^2$ or $\mathbb{C}^2$ with the standard basis, inner product, and associated euclidean norm, and consider the linear operator A given by the matrix

$$A = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}.$$

Then $\|A\|_{\mathrm{op}} = 1$ while $\sigma(A) = \{0\}$.

2 Random matrices

2.1 Matrix ensembles

The central question we will be interested in is the following:

> **Question** Let M be a random matrix of size $n \times n$. What can we say about its spectrum as $n \to \infty$?

To make this question more precise, we must say what we mean with a "random matrix", i.e., we must describe its law, and we must also be more specific about what we want to know about its spectrum. In random matrix theory, a sequence of probability laws on the spaces of $n \times n$ real or complex matrices is called a *matrix ensemble*. There are several natural choices. In what follows, for each integer $n \geq 1$, we consider a random matrix $M = (\xi_{ij})_{1 \leq i,j \leq n}$.

- **I.i.d. matrix ensembles** These are ensembles where the entries are i.i.d. according to some common law. Examples are:

  - *Bernoulli ensemble* ξ_{ij} uniformly distributed on $\{-1, 1\}$,
  - *Real Gaussian ensemble* ξ_{ij} standard normally distributed on $\mathbb{R}$,
  - *Complex Gaussian ensemble* ξ_{ij} standard normally distributed on $\mathbb{C} \cong \mathbb{R}^2$.

- **Symmetric Wigner matrix ensembles** These are ensembles where $(\xi_{ij})_{i \geq j}$ are independent real random variables and $\xi_{ij} := \xi_{ji}$ for $i > j$. Examples are:
  - *Symmetric Bernoulli ensemble* $(\xi_{ij})_{i \leq j}$ uniformly distributed on $\{-1, 1\}$,
  - *Gaussian orthogonal ensemble (GOE)* $(\xi_{ij})_{i \leq j}$ standard normally distributed on $\mathbb{R}$, and $(\xi_{ii})_{1 \leq i \leq n}$ normally distributed with mean zero and variance 2.

- **Hermitian Wigner matrix ensembles** These are ensembles where $(\xi_{ij})_{i<j}$ are independent complex random variables, $(\xi_{ii})_{1 \leq i \leq n}$ are independent real random variables, and $\xi_{ij} := \overline{\xi_{ji}}$ for $i > j$. The symmetric Wigner matrix ensembles are special cases of this. An important additional example is:
  - *Gaussian unitary ensemble (GUE)* $(\xi_{ij})_{i \leq j}$ standard normally distributed on $\mathbb{C} \cong \mathbb{R}^2$, and $(\xi_{ii})_{1 \leq i \leq n}$ standard normally distributed on $\mathbb{R}$.

Note that in all these examples, we can start with an infinite matrix

$$M = (\xi_{ij})_{i,j \in \mathbb{N}_+},$$

and then define $M_n := (\xi_{ij})_{1 \leq i,j \leq n}$. This provides a natural coupling between matrixes of different size.

Question If V is an inner product space and $\{e(1), \ldots, e(n)\}$ is an orthonormal basis, then each of the "ensembles" above naturally defines a probability law on $\mathcal{L}(V)$, when we identify a linear operator with its matrix. For which of the probability distributions above (if any) is this law independent of the choice of the orthonormal basis? I suspect this may be true for GOE and GUE (and is probably known). Indeed, for GUE this seems to follow from the first displayed formula in Section 3.3 of Terence tao's book. For GOE there may be a similar formula. Note that in general (for example for the Bernoulli ensembles), if we change the basis, then for the new matrix it will not even be true that the (upper diagional) matrix entries are independent.

Question Does there exist a random matrix ensemble where $(\xi_{ij})_{1 \leq i,j \leq n}$ are normal, but not self-adjoint? The space of normal linear operators is not nice in the sense that the sum of two normal operators is in general not normal.

2.2 Limit behaviour of the spectrum

One of the highlights of the course is the proof of Theorem 2.4.2 in the book, which reads:

> **Wigner semicircle law** Let $M = (\xi_{ij})_{i,j \in \mathbb{N}_+}$ be an infinite hermitian Wigner matrix. For each $n \geq 1$, let $M_n := (\xi_{ij})_{1 \leq i,j \leq n}$ and let $\lambda_1(M_n) \leq \cdots \leq \lambda_n(M_n)$ denote its eigenvalues. Let
>
> $$\mu_n := \frac{1}{n} \sum_{j=1}^{n} \delta_{\lambda_j(M_n)/\sqrt{n}} \qquad (n \geq 1).$$
>
> Then almost surely $\mu_n \Rightarrow \mu$, where $\Rightarrow$ denotes weak convergence of probability measures on $\mathbb{R}$ and μ is the *semicircle law*
>
> $$\mu(\mathrm{d}x) := \frac{1}{2\pi} \sqrt{0 \vee (4 - |x|^2)}\, \mathrm{d}x.$$

Note that we are able to formulate this as almost sure convergence only due to the coupling described above. Of course, almost sure convergence to a deterministic limit also implies convergence in probability.

The theorem holds for the symmetric Bernoulli ensemble and the GOE and GUE ensembles, as well as many other hermitian Wigner matrix ensembles. Thus, similar to the central limit theorem, the Wigner semicircle law is highly *universal*. For i.i.d. matrix ensembles the limit law is different, however: in this case the limit law is the uniform distribution on the ball $\{z \in \mathbb{C} : |z| \leq 2\}$. This limit law is technically more difficult to prove, so we will not see its proof in this course.

The Wigner semicircle law implies that for large n, with high probability, most of the eigenvalues of M_n lie between $-2\sqrt{n}$ and $2\sqrt{n}$. Since there are n eigenvalues, their average spacing is $4/\sqrt{n}$. There are other questions about the spectrum one could be interested in, such as:

- Let $x_n \in \mathbb{R}$ satisfy $x_n/\sqrt{n} \to x \in (-2, 2)$. Consider the point process

$$\big\{\lambda \in \mathbb{R} : x_n + \lambda/\sqrt{n} \in \sigma(M_n)\big\}.$$

  Does this point process have a limit (in law)?

- Same question as above, but for a suitable sequence x_n such that $x_n/\sqrt{n} \to 2$, perhaps with a different scaling of space to compensate for the lower density of points near the end of the spectrum.

- What can we say about $\|M_n\|_{\mathrm{op}}$, i.e., the largest eigenvalue in absolute value?. Is it true that $\|M_n\|_{\mathrm{op}}/\sqrt{n} \to 2$?

These questions have been studied in detail and a lot is known. We will only look at the last question. We will prove that indeed $\|M_n\|_{\mathrm{op}}$ is with high probability close to $2\sqrt{n}$. The methods used to prove this turn out to be very useful also later when we prove the Wigner semicircle law.

2.3 Why would we care?

The book we use is written by Terence Tao, who is a pure mathematician who has worked in a variety of fields such as number theory. For him, there is one clear reason for studying random matrices:

> It is a rich field with beautiful mathematics and many interesting problems.

There are also more practical reasons why one could be interested in random matrices. Some of the first people to study them were in fact physicists who were interested in the absorption spectrum of large atoms. One of the first great achievements of quantum mechanics was the description of the spectrum of the hydrogen atom, which has just one electron. The spectrum of helium, which has two electrons, is already much more complicated. As one moves up in the periodic system of elements things quickly get really messy. Mathematically, the possible energy values of an electron orbiting an atom are given by the spectrum of a hermitian matrix. This matrix becomes so complicated that at some point a physicist wondered what would happen if one assumes it is completely random -and actually got reasonable results.

Apart from the reasons mentioned above, one may also ask if within mathematics, random matrix theory is more or less isolated, or on the other hand connected to lots of other problems (for which one may then again have practical reasons to study them). The answer seems to lie a bit in the middle. On the one hand, interesting connections have been found to other subjects such as the totally asymetric exclusion process (TASEP) or free probability. On the

other hand, it is fair to say that within probability theory, the theory of random matrices is a bit of an outlier. On conferences, one sometimes meets the true specialists in random matrix theory, that seem to live a bit in a world of their own, with limited interaction with the rest of probability. This is mostly concerned with very subtle questions of the field, however, which require difficult and long proofs using specialised methods. As for the basic topics covered by Tao's book, it is probably a good thing for every probabilist to know a bit about them.

2.4 How do we start

We will focus on hermitian Wigner ensembles. Our first aim will be to prove that $\|M_n\|_{\mathrm{op}}$ is with high probability close to $2\sqrt{n}$. As mentioned before, the techniques we use for that will also be useful when proving the Wigner semicircle law. Thus, perhaps surprisingly, the strategy is to first prove some sort of (weak) law of large numbers, as a first step towards something that is a bit like the CLT. Actually, there is much more similarity than might be expected at first sight between the classical limit theorems of probability theory (LLN, CLT) and their random matrix "counterparts" (LLN for $\|M_n\|_{\mathrm{op}}$, Wigner semicircle law).

Let $X_1, \dots, X_n$ be i.i.d. real random variables with finite mean μ. Then the weak law of large numbers says that $\frac{1}{n}\sum_{i=1}^n X_i$ is with high probability close to μ. One can generalise the problem and ask what we can say about $F_n(X_1, \dots, X_n)$, where $F_n : \mathbb{R}^n \to \mathbb{R}$ is a sequence of "nice" functions. For such functions, one often observes that for large n, the distribution of $F_n(X_1, \dots, X_n)$ is closely concentrated around one deterministic value. If the F_n are not linear, then it may be difficult to say what that value is precisely. But it turns out that there are nice methods that give sufficient conditions for the law of $F_n(X_1, \dots, X_n)$ to be closely concentrated around one deterministic value, *even if we can's say precisely what value that is.* This is the problem of *concentration of measure*. There exists a large literature on this and it obviously has many applications in probability theory outside of random matrix theory. Our first aim will be *Talagrand's concentration inequality* (Theorem 2.1.13 in the book).

3 The semicircle law

3.1 Weak convergence

A *compactification* of a topological space E is a compact topological space $\overline{E}$ such that E is a dense subset of $\overline{E}$ and the topology on E is the induced topology from its embedding in $\overline{E}$. A topological space is *metrisable* if there exists a metric that generates the topology. A *Polish space* is a separable topological space E with the property that there exists a complete metric d_{c} that generates the topology. Warning: on a Polish space, there may also exists metrics d that are not complete but that nevertheless generate the topology on E. We say that E is *locally compact* if for every $x \in E$, there exist an open set O and compact set C such that $x \in O \subset C$. A G_δ-subset is a set that can be written as a countable intersection of open sets.

Theorem 5 (Compactifications of Polish spaces) *Let E be a metrisable topological space. Then the following statements are equivalent.*

(i) *E is Polish.*

(ii) *There exists a metrisable compactification $\overline{E}$ of E such that E is a G_δ-subset of $\overline{E}$.*

(iii) *For each metrisable compactification $\overline{E}$ of E, it holds that E is a G_δ-subset of $\overline{E}$.*

A Polish space is locally compact if and only if the following two equivalent statements hold.

(i) *There exists a metrisable compactification $\overline{E}$ of E such that E is an open of $\overline{E}$.*

(ii) *For each metrisable compactification $\overline{E}$ of E, it holds that E is an open of $\overline{E}$.*

We always equip a Polish space with the Borel σ-field $\mathcal{B}(E)$, which is generated by the open sets. For any Polish space E, we let $\mathcal{C}_{\mathrm{b}}(E)$ and $B_{\mathrm{b}}(E)$ denote the spaces of real bounded functions on E that are continuous and measurable, respectively. If E is compact, then $\mathcal{C}_{\mathrm{b}}(E) = \mathcal{C}(E)$, the space of all continuous real functions on E. If E is compact, then the space $\mathcal{C}(E)$, equipped with the supremumnorm $\|f\|_\infty := \sup_{x\in E} |f(x)|$ is separable and complete.

Proposition 6 (Borel σ-field) *Assume that E is a Polish space. Assume that $f_i \in B_{\mathrm{b}}(E)$ for all $i \in \mathbb{N}$ and that $(f_i)_{i\in\mathbb{N}}$ separate points. Then $(f_i)_{i\in\mathbb{N}}$ generate the Borel σ-field.*

We let $\mathcal{P}(E)$ denote the space of probability measures on E, equipped with the topology of weak convergence. For any $A \subset E$ and $\varepsilon > 0$, we set

$$A^\varepsilon := \big\{x \in E : d(x, A) < \varepsilon\big\} \quad \text{with} \quad d(x, A) := \inf_{y\in A} d(x, y). \tag{4}$$

The *Prohorov metric* d_{P} on $\mathcal{P}(E)$ is defined as

$$d_{\mathrm{P}}(\mu, \nu) := \inf\big\{\varepsilon > 0 : \mu(A) \le \nu(A^\varepsilon) + \varepsilon \ \forall A \in \mathcal{B}(E)\big\}. \tag{5}$$

The Prohorov metric has an alternative definition in terms of *coupling*. Indeed,

$$d_{\mathrm{P}}(\mu, \nu) = \inf\big\{\varepsilon > 0 : \exists \text{ r.v.'s } X, Y \text{ with laws } \mu, \nu \text{ s.t. } \mathbb{P}[d(X, Y) \ge \varepsilon] \le \varepsilon\big\}. \tag{6}$$

One can prove that

$$d_{\mathrm{P}}(\mu_n, \mu) \underset{n\to\infty}{\longrightarrow} 0 \quad \Leftrightarrow \quad \int \mu_n(\mathrm{d}x) f(x) \underset{n\to\infty}{\longrightarrow} \int \mu(\mathrm{d}x) f(x) \quad \forall f \in \mathcal{C}_{\mathrm{b}}(E).$$

The topology generated by $d_{\rm P}$ is called the topology of *weak convergence*. Convergence in this topology is denoted as $\mu_n \Rightarrow \mu$. The space $\mathcal{P}(E)$ is separable and complete in $d_{\rm P}$. So $\mathcal{P}(E)$ is Polish. One can check that

$$\begin{array}{rcl} \mu_n \underset{n\to\infty}{\Longrightarrow} \mu & \Leftrightarrow & \liminf\limits_{n\to\infty} \mu_n(O) \geq \mu(O) \quad \forall \text{open } O \subset E \\ & \Leftrightarrow & \limsup\limits_{n\to\infty} \mu_n(C) \leq \mu(C) \quad \forall \text{closed } C \subset E. \end{array} \tag{7}$$

For any measurable $A \subset E$, we let $\mathrm{int}(A)$ and $\overline{A}$ denote its interior and closure and we call $\partial A := \overline{A} \backslash \mathrm{int}(A)$ its boundary. It follows from (7) that

$$\mu_n \underset{n\to\infty}{\Longrightarrow} \mu, \ A \in \mathcal{B}(E), \text{ and } \mu(\partial A) = 0 \quad \text{imply} \quad \mu_n(A) \underset{n\to\infty}{\longrightarrow} \mu(A). \tag{8}$$

Theorem 7 (Skorohod representation theorem) *Let E be a Polish space and let $\mu_n, \mu \in \mathcal{P}(E)$. Then one has $\mu_n \Rightarrow \mu$ if and only if there exist, on some probability space, random variables X_n, X with laws μ_n, μ, such that $X_n \to X$ a.s.*

Recall that a subset of a topological space is *precompact* if its closure is compact.

Theorem 8 (Prohorov) *A subset $\mathcal{A} \subset \mathcal{P}(E)$ is precompact if and only if it is tight, i.e., for every $\varepsilon > 0$, there exists a compact set $C \subset E$ such that $\mu(C) \geq 1 - \varepsilon$ for all $\mu \in \mathcal{A}$.*

If $\overline{E}$ is a compactification of E, then we have the natural identification

$$\mathcal{P}(E) \cong \{\mu \in \mathcal{P}(\overline{E}) : \mu(E) = 1\}. \tag{9}$$

Using (7), one can check that the topology of weak convergence on $\mathcal{P}(E)$ coincides with the induced topology from its embedding in $\mathcal{P}(\overline{E})$. By Theorem 5, it follows that $\mathcal{P}(E)$ is a G_δ-subset of $\mathcal{P}(\overline{E})$.

3.2 Random measures

Let E be a Polish space. We equip $\mathcal{P}(E)$, of course, with the Borel σ-algebra. The next lemma says that this coincides with many other natural choices for the σ-algebra on $\mathcal{P}(E)$.

Lemma 9 (Measurable sets of measures) *Let $L_f : \mathcal{P}(E) \to \mathbb{R}$ be defined by $L_f(\mu) := \int \mu(\mathrm{d}x) f(x)$. Then the Borel σ-field on $\mathcal{P}(E)$ is generated by*

$$1.\ \big\{L_f : f \in \mathcal{C}_{\rm b}(E)\big\}, \quad 2.\ \big\{L_{1_O} : O \subset E \text{ open}\big\}, \quad 3.\ \big\{L_{1_A} : A \subset E \text{ measurable}\big\}.$$

Proof If $f \in \mathcal{C}_{\rm b}(E)$, then $L_f \in \mathcal{C}_{\rm b}(\mathcal{P}(E))$, so L_f is in particular measurable. If $O \subset E$ is open, then there exist continuous $f_n : E \to [0,1]$ that increase to 1_O, which implies that L_{1_O} is measurable. Now $\{A \in \mathcal{B}(E) : L_{1_A} \text{ is measurable}\}$ is a λ-system that contains the π-system consisting of all open subsets of E, so by the π/λ-theorem this set equals $\mathcal{B}(E)$. In particular, L_{1_A} is measurable for all $A \in \mathcal{B}(E)$.

Let $\overline{E}$ be a metrisable compactification. Since $\mathcal{C}(\overline{E})$ is separable, we can choose $f_i \in \mathcal{C}(\overline{E})$ $(i \in \mathbb{N})$ such that $\{f_i : i \in \mathbb{N}\}$ is dense in $\mathcal{C}(\overline{E})$. The functions L_{f_i} are continuous (and hence in particular measurable) and separate points, so by Proposition 6 they generate the Borel σ-field on $\mathcal{P}(E)$. Similarly, since E is separable, there exists a countable base $\{O_i : i \in \mathbb{N}\}$ of the topology. Since $\{L_{1_{O_i}} : i \in \mathbb{N}\}$ separate points, by Proposition 6 they generate the Borel σ-field on $\mathcal{P}(E)$. ∎

If E is a Polish space, then the same is true for $\mathcal{P}(E)$ and hence also for $\mathcal{P}(\mathcal{P}(E))$. A random variable with values in $\mathcal{P}(E)$ is a random probability measure; its law is then an

element of $\mathcal{P}(\mathcal{P}(E))$. If μ is a random variable with values in $\mathcal{P}(E)$, then its *mean* is the probability measure $E[\mu]$ on E defined as

$$\int_E \mathbb{E}[\mu](\mathrm{d}x) f(x) := \mathbb{E}\big[\int_E \mu(\mathrm{d}x) f(x)\big] \qquad \forall f \in B_{\mathrm{b}}(E). \tag{10}$$

By Lemma 9, the right-hand side of this equation is well-defined. Once can check that it is linear in f and continuous with respect to increasing sequences. It follows that

$$\mathcal{B}(E) \ni A \mapsto \mathbb{E}\big[\int_E \mu(\mathrm{d}x) 1_A(x)\big]$$

is a measure. Recalling the definition of the integral, we see that the integral of a bounded measurable function with respect to this measure is given by the right-hand side of (10).

3.3 The semicircle law

Let $(\xi_{ij})_{i,j\in\mathbb{N}_+}$ be real random variables such that

(i) $(\xi_{ij})_{i<j}$ are i.i.d. with mean zero and variance one,

(ii) $(\xi_{ij})_{i=j}$ are i.i.d. with mean zero and finite variance, and independent of $(\xi_{ij})_{i<j}$,

(iii) $\xi_{ji} = \xi_{ij}$ $(i < j)$.

Let $M_n := (\xi_{ij})_{1\leq i,j\leq n}$, let $\lambda_1(n) \geq \cdots \geq \lambda_n(n)$ be its eigenvalues, let μ_n be the random variables with values in $\mathcal{P}(\mathbb{R})$ defined by

$$\mu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i(n)/\sqrt{n}}, \tag{11}$$

and let μ_{sc} denote the semicircle law

$$\mu_{\mathrm{sc}}(\mathrm{d}x) := \frac{1}{2\pi}\big(4 - |x|^2\big)_+^{1/2}\mathrm{d}x. \tag{12}$$

Then Theorem 2.4.2 from the book says that

$$\mu_n \underset{n\to\infty}{\Longrightarrow} \mu_{\mathrm{sc}} \quad \text{a.s.} \tag{13}$$

The book explains how to define a.s. convergence for random variables taking values in a σ-compact metrisable space, i.e., a metrisable space that is the union of countably many compact sets. The book does not say what topology we should choose on $\mathcal{P}(\mathbb{R})$. This is very sloppy indeed: one has to specify the topology, otherwise the convergence statement has no meaning. From Exercise 2.4.1, we understand that what is meant is, of course, that we equip $\mathcal{P}(\mathbb{R})$ with the topology of weak convergence. Alas:

Lemma 10 (Lack of σ-compactness) *The space $\mathcal{P}(\mathbb{R})$ is not σ-compact.*

Proof According to the Baire category theorem, each complete metrisable space has the Baire property, which says that if $(C_i)_{i\in\mathbb{N}}$ are closed sets with empty interior, then also $\bigcup_{i\in\mathbb{N}} C_i$ has empty interior. We will show that in $\mathcal{P}(\mathbb{R})$, each compact set has empty interior. It follows that $\mathcal{P}(\mathbb{R})$ cannot be the union of countable many compact sets, since by the Baire category theorem this would imply that $\mathcal{P}(\mathbb{R})$ has empty interior.

It remains to show that each compact subset $\mathcal{A}$ of $\mathcal{P}(\mathbb{R})$ has empty interior. Let ε_n be positive constants, tending to zero. By Prohorov's theorem (Theorem 8), for each n we can find a compact set $C_n \subset \mathbb{R}$ such that $\mu(C_n) \geq 1 - \varepsilon_n/2$ for all $\mu \in \mathcal{A}$. Since $\mathbb{R}$ is not compact,

for each n we can choose $x_n \in \mathbb{R}\backslash C_n$. Now fix $\mu \in \mathcal{A}$ and define $\mu_n := (1-\varepsilon_n)\mu + \varepsilon_n \delta_{x_n}$. Then $\mu_n \notin \mathcal{A}$ while $\mu_n \Rightarrow \mu$, which proves that $\mathcal{A}$ has empty interior. ∎

The argument above shows quite generally that if E is a Polish space that is not compact, then $\mathcal{P}(E)$ is not σ-compact and not locally compact either. This seems unpleasant, but $\mathcal{P}(E)$ is still a Polish space which is for most probabilistic purposes sufficient. Also, in view of (9) we can equivalently view the random measures in (11) as random variables with values in the space $\mathcal{P}(\overline{\mathbb{R}})$ of probability measures on the extended real line $\overline{\mathbb{R}} := [-\infty, \infty]$. In view of Prohorov's theorem (Theorem 8), the space $\mathcal{P}(\overline{\mathbb{R}})$ is compact. In fact, $\mathcal{P}(\overline{\mathbb{R}})$ is a compactification of $\mathcal{P}(\mathbb{R})$ and the latter is a G_δ-subset, but not an open subset of the former.

3.4 Some simplifications

Let us say that a sequence $(n_j)_{j\in\mathbb{N}}$ is *lacunary* if there exists a $c > 1$ such that $n_{j+1}/n_j \geq c$ for all $j \in \mathbb{N}$.

Proposition 11 (Lacunary subsequence) *In order to prove (13), it suffices to show that*

$$\mu_{n_j} \underset{j\to\infty}{\Longrightarrow} \mu_{\rm sc} \quad \text{a.s.} \tag{14}$$

for any lacunary sequence $(n_j)_{j\in\mathbb{N}}$.

To prove this, we will use the following lemma.

Lemma 12 (Convergence of distribution functions) *Let μ_n, μ be probability measures on $\mathbb{R}$. Let $F_n(r) := \mu_n\big((r,\infty)\big)$ and $F(r) := \mu_n\big((r,\infty)\big)$ $(r \in \mathbb{R})$ and assume that F is continuous. Then $\mu_n \Rightarrow \mu$ if and only if*

$$F_n(r_n) \to F(r) \qquad \forall r_n, r \in \mathbb{R} \text{ s.t. } r_n \to r. \tag{15}$$

It suffices to prove this for a countable collection of real constants r_n^i, r^i $(i, n \in \mathbb{N})$ such that $r_n^i \to r^i$ for all i and $\{r^i : i \in \mathbb{N}\}$ is dense in $\mathbb{R}$.

Proof (sketch) Since by assumption F is continuous, it follows from (7) and (8) that $\mu_n \Rightarrow \mu$ if and only if

$$F_n(r) \to F(r) \qquad \forall r \in \mathbb{R}. \tag{16}$$

Using the fact that the functions F_n, F are nondecreasing, it is not hard to show that (16) implies the seemingly stronger (15). To prove the final statement, assume that (15) holds for some choice of r_n^i, r^i $(i, n \in \mathbb{N})$. By Prohorov, $\{\mu_n : n \in \mathbb{N}\}$ is a precompact subset of $\mathcal{P}(\overline{\mathbb{R}})$, so to prove that $\mu_n \Rightarrow \mu$ it suffices to show that μ is the only cluster point in $\mathcal{P}(\overline{\mathbb{R}})$ of the sequence μ_n. By our assumptions and what we have already proved, each cluster point μ_* satisfies $\mu_*\big((r^i,\infty)\big) = \mu\big((r^i,\infty)\big)$ for all $i \in \mathbb{N}$, which implies $\mu_* = \mu$ since $\{r^i : i \in \mathbb{N}\}$ is dense. ∎

Proof or Proposition 11 The proof of Proposition 11 makes use of Exercise 1.3.14, which asks to prove *Cauchy's interlacing law*

$$\lambda_{i+1}(n) \leq \lambda_i(n-1) \leq \lambda_i(n). \tag{17}$$

Let us set

$$F(r) := \mu_{\rm sc}\big((r,\infty)\big) \quad \text{and} \quad N_n(\lambda) := \sup\{i : \lambda_i(n) > \lambda\}, \tag{18}$$

i.e., $N_n(\lambda)$ is the number of eigenvalues of M_n that are greater than λ. By Lemma 12, to prove the proposition, it suffices show that almost surely

$$\frac{1}{n} N_n(r\sqrt{n}) \underset{n\to\infty}{\longrightarrow} F(r) \qquad \forall r \in \mathbb{Q}. \tag{19}$$

It follows from (17) that

$$N_n(\lambda) \le N_{n+1}(\lambda) \le N_n(\lambda) + 1. \tag{20}$$

In particular, this implies that

$$N_n(\lambda) \le N_m(\lambda) \qquad (n \le m). \tag{21}$$

Fix $c > 1$ and let $(n_j)_{j\in\mathbb{N}}$ be the lacunary sequence defined as $n_0 := 1$ and $n_j := \inf\{n > n_{j-1} : n/n_{j-1} \ge c\}$. Then for each $n \in \mathbb{N}$, there is a $j \in \mathbb{N}$ such that

$$n_j \le n < n_{j+1}. \tag{22}$$

By (21), this implies

$$N_{n_j}(\lambda) \le N_n(\lambda) \le N_{n_{j+1}}(\lambda) \qquad (\lambda \in \mathbb{R}), \tag{23}$$

which allows us to estimate, for any $r \le 0$,

$$\frac{1}{n_{j+1}} N_{n_j}(r\sqrt{n_j}) \le \frac{1}{n} N_{n_j}(r\sqrt{n}) \le \frac{1}{n} N_n(r\sqrt{n}) \le \frac{1}{n} N_{n_{j+1}}(r\sqrt{n}) \le \frac{1}{n_j} N_{n_{j+1}}(r\sqrt{n_{j+1}}). \tag{24}$$

Using our assumption (14), Lemma 12, and the fact that $n_{j+1}/n_j \to c$ as $j \to \infty$, it follows that

$$c^{-1}F(r) \le \liminf_{n\to\infty} \frac{1}{n} N_n(r\sqrt{n}) \le \limsup_{n\to\infty} \frac{1}{n} N_n(r\sqrt{n}) \le cF(r) \quad \text{a.s.} \qquad (r \le 0). \tag{25}$$

In a similar way, for $r \ge 0$, we can estimate

$$\frac{1}{n_{j+1}} N_{n_j}(r\sqrt{n_{j+1}}) \le \frac{1}{n} N_n(r\sqrt{n}) \le \frac{1}{n_j} N_{n_{j+1}}(r\sqrt{n_j}), \tag{26}$$

which yields

$$c^{-1}F(c^{1/2}r) \le \liminf_{n\to\infty} \frac{1}{n} N_n(r\sqrt{n}) \le \limsup_{n\to\infty} \frac{1}{n} N_n(r\sqrt{n}) \le cF(c^{-1/2}r) \quad \text{a.s.} \qquad (r \ge 0). \tag{27}$$

Since (25) and (27) hold for all $c, r \in \mathbb{Q}$ with $c > 1$ and $r \le 0$ or $r \ge 0$, respectively, we conclude that a.s.

$$\frac{1}{n} N_n(r\sqrt{n}) \underset{n\to\infty}{\longrightarrow} F(r) \qquad (r \in \mathbb{Q}), \tag{28}$$

which by Lemma 12 implies that $\mu_n \Rightarrow \mu$ a.s. ∎

Proposition 13 (Cut-off argument) *In order to prove (13), we can without loss of generality assume that there exists a $K < \infty$ such that $|\xi_{ij}| \le K$ a.s. for all $i, j \in \mathbb{N}_+$.*

The proof of Proposition 13 needs some preparations. For any symmetric $n \times n$ matrix A, we let $\lambda_1(A) \ge \cdots \ge \lambda_n(A)$ denote its eigenvalues, and in line with (11), we set

$$\mu(A) := \frac{1}{n} \sum_{i=1}^{n} \delta_{\lambda_i(A)/\sqrt{n}}. \tag{29}$$

We make use of the *Weilandt-Hoffmann inequality* which says that

$$\sum_{i=1}^{n} \big|\lambda_i(A) - \lambda_i(B)\big|^2 \le \|A - B\|_{\mathrm{F}}^2, \tag{30}$$

where

$$\|A\|_{\mathrm{F}}^2 := \sum_{i=1}^{n} \sum_{j=1}^{n} |A_{ij}|^2 = \mathrm{tr}(A^*A) \tag{31}$$

is the Frobenius norm of A.

Lemma 14 (Distance between eigenvalue distributions) *Let A, B be symmetric $n \times n$ matrices. Then the Prohorov distance between $\mu(A)$ and $\mu(B)$ satisfies*

$$d_{\mathrm{P}}\big(\mu(A), \mu(B)\big) \leq \Big(\frac{1}{n^2}\|A - B\|_{\mathrm{F}}^2\Big)^{1/3}. \tag{32}$$

Proof Let I be uninformly distributed on $\{1, \dots, n\}$ and let $X := \lambda_I(A)/\sqrt{n}$ and $Y := \lambda_I(B)/\sqrt{n}$. Then X and Y are distributed according to the probability laws $\mu(A)$ and $\mu(B)$, respectively, and the Weilandt-Hoffmann inequality (30) allows us to estimate

$$\mathbb{E}\big[|X - Y|^2\big] = \frac{1}{n}\sum_{i=1}^{n} \big|\lambda_i(A)/\sqrt{n} - \lambda_i(B)/\sqrt{n}\big|^2 \leq \frac{1}{n^2}\|A - B\|_{\mathrm{F}}^2. \tag{33}$$

Since

$$\mathbb{E}\big[|X - Y|^2\big] \geq \varepsilon^2 \mathbb{P}\big[|X - Y| \geq \varepsilon\big], \tag{34}$$

this implies that

$$\mathbb{P}\big[|X - Y| \geq \varepsilon\big] \leq \frac{1}{\varepsilon^2 n^2}\|A - B\|_{\mathrm{F}}^2. \tag{35}$$

The right-hand side of this equation equals ε if

$$\varepsilon = \Big(\frac{1}{n^2}\|A - B\|_{\mathrm{F}}^2\Big)^{1/3}. \tag{36}$$

In view of (6), this implies (32). ∎

Proof of Proposition 13 We first prove the statement under the additional assumption that ξ_{ij} is equally distributed with $-\xi_{ij}$ for all i, j. For each $K > 0$, we define

$$\xi_{ij}^{\leq K} := 1_{\{|\xi_{ij}| \leq K\}}\xi_{ij} \quad \text{and} \quad \xi_{ij}^{> K} := 1_{\{|\xi_{ij}| > K\}}\xi_{ij}. \tag{37}$$

We let $M_n^{\leq K}$ and $M_n^{> K}$ denote the corresponding $n \times n$ matrices. Since ξ_{ij} is equally distributed with $-\xi_{ij}$ the same is true for $\xi_{ij}^{\leq K}$ which implies that these random variables have mean zero. We denote their variance by

$$\sigma_K^2 := \mathbb{E}\big[|\xi_{ij}^{\leq K}|^2\big] \qquad (i \neq j), \tag{38}$$

which increases to one as $K \to \infty$. Let $\mu_{\mathrm{sc}}^{\sigma}$ denote the image of the semicircle law μ_{sc} under the map $\lambda \mapsto \sigma\lambda$. By the triangle inequality for the Prohorov metric

$$d_{\mathrm{P}}\big(\mu(M_n), \mu_{\mathrm{sc}}\big) \leq d_{\mathrm{P}}\big(\mu(M_n), \mu(M_n^{\leq K})\big) + d_{\mathrm{P}}\big(\mu(M_n^{\leq K}), \mu_{\mathrm{sc}}^{\sigma_K}\big) + d_{\mathrm{P}}\big(\mu_{\mathrm{sc}}^{\sigma_K}, \mu_{\mathrm{sc}}\big). \tag{39}$$

Since μ_{sc} is concentrated on $[-2, 2]$, it is easy to see that $d_{\mathrm{P}}\big(\mu_{\mathrm{sc}}^{\sigma_K}, \mu_{\mathrm{sc}}\big) \leq 2(1 - \sigma_K)$, which tends to zero as $K \to \infty$. Our assumptions imply that the second term tends to zero for each $K < \infty$. By Lemma 14, we can estimate the first term as

$$d_{\mathrm{P}}\big(\mu(M_n), \mu(M_n^{\leq K})\big) \leq \Big(\frac{1}{n^2}\|M_n^{> K}\|_{\mathrm{F}}^2\Big)^{1/3}. \tag{40}$$

Here

$$\|M_n^{> K}\|_{\mathrm{F}}^2 = \sum_{i=1}^{n}\sum_{j=1}^{n} 1_{\{|\xi_{ij}| > K\}}|\xi_{ij}|^2. \tag{41}$$

Let

$$\varepsilon_K := \mathbb{E}\big[|\xi_{ij}|^2 1_{\{|\xi_{ij}| > K\}}\big]. \tag{42}$$

By assumption, ξ_{ij} has mean zero and variance one, so its second moment is finite and hence $\lim_{K\to\infty} \varepsilon_K = 0$ by dominated convergence. The strong law of large numbers tells us that

$$\lim_{n\to\infty} \frac{1}{n^2} \|M_n^K\|_{\rm F}^2 = \varepsilon_K \quad \text{a.s.} \tag{43}$$

Inserting our estimates in (39), we see that

$$\limsup_{n\to\infty} d_{\rm P}\big(\mu(M_n), \mu_{\rm sc}\big) \leq \varepsilon_K^{1/3} + 2(1-\sigma_K) \quad \text{a.s.} \quad \forall K < \infty, \tag{44}$$

which implies that $\mu(M_n) \Rightarrow \mu_{\rm sc}$ a.s.

If we drop the assumption that ξ_{ij} is equally distributed with $-\xi_{ij}$ for all i, j, then the proof becomes a bit more complicated. Let I_n denote the identity matrix and let U_n denote the matrix whose entries are all equal to one. Let

$$\delta_K := \mathbb{E}[\xi_{ij}^{>K}] \qquad (i \neq j) \quad \text{and} \quad \delta'_K := \mathbb{E}[\xi_{ij}^{>K}] \qquad (i = j). \tag{45}$$

Then the entries of

$$\tilde{M}_n^{\leq K} := M_n^{\leq K} + \delta_K U_n + (\delta'_K - \delta_K) I_n \tag{46}$$

have mean zero, and the off-diagonal entries have some variance $\tilde{\sigma}_K^2$ that converges to one as $K \to \infty$. By Lemma 14,

$$d_{\rm P}\big(\mu(M_n^{\leq K}), \mu(\tilde{M}_n^{\leq K})\big) \leq \Big(\frac{1}{n^2}\|\delta_K U_n + (\delta'_K - \delta_K) I_n\|_{\rm F}^2\Big)^{1/3}, \tag{47}$$

where

$$\frac{1}{n^2}\|\delta_K U_n + (\delta'_K - \delta_K) I_n\|_{\rm F}^2 \underset{n\to\infty}{\longrightarrow} \delta_K^2 \underset{K\to\infty}{\longrightarrow} 0. \tag{48}$$

Using this, the proof above can easily be generalised to the case that the law of the ξ_{ij}'s is not symmetric. ∎

3.5 Eigenvalue inequalities

We still need to provide the proof of Cauchy's interlacing law (17) and the Weilandt-Hoffmann inequality (30).

Theorem 15 (Minmax formula) *Let V be a complex inner product space of dimension $\dim(V) = n$, let $A \in \mathcal{L}(V)$ satisfy $A^* = A$, and let $\lambda_1 \geq \cdots \geq \lambda_n$ be its eigenvalues. Then*

$$\lambda_i := \sup_{W:\,\dim(W)=i} \inf_{\substack{\phi\in W\\ |\phi|=1}} \langle\phi|A\phi\rangle. \tag{49}$$

Proof Let $\big\{e(1), \ldots, e(n)\big\}$ be an orthonormal basis made up of eigenvectors corresponding to the eigenvalues $\lambda_1, \ldots, \lambda_n$ and let ϕ_i denote the i-th coordinate of a vector ϕ with respect to this basis. Then

$$\langle\phi|A\phi\rangle = \sum_{i=1}^n |\phi_i|^2 \lambda_i. \tag{50}$$

If $|\phi| = 1$, then $\big(|\phi_1|^2, \ldots, |\phi_n|^2\big)$ is a probability distribution on $\{1, \ldots, n\}$, so $\langle\phi|A\phi\rangle$ is the mean of the function $j \mapsto \lambda_j$ with respect to this probability distribution.

If we choose for W the linear space spanned by $e(1), \ldots, e(i)$, then $\{\phi \in W : |\phi| = 1\}$ is the set of vectors for which this probability distribution is concentrated on $\{1, \ldots, i\}$. It follows that

$$\langle\phi|A\phi\rangle \geq \lambda_i \quad \forall \phi \in W \text{ s.t. } |\phi| = 1, \tag{51}$$

with equality if and only if ϕ is a multiple of $e(i)$. We have thus found a linear subspace W of dimension i such that

$$\lambda_i = \inf_{\substack{\phi\in W\\ |\phi|=1}} \langle\phi|A\phi\rangle. \tag{52}$$

To complete the proof, we need to show that if W is an arbitrary linear subspace of dimension i, then

$$\lambda_i \geq \inf_{\substack{\phi\in W\\ |\phi|=1}} \langle\phi|A\phi\rangle. \tag{53}$$

Let

$$W' := \big\{\phi \in W : \phi_j = 0 \ \forall j < i\big\}. \tag{54}$$

Since W has dimension i, the space W' has dimension at least one. If $\phi \in W'$ satisfies $|\phi| = 1$, then the probability distribution $\big(|\phi_1|^2, \ldots, |\phi_n|^2\big)$ is concentrated on $\{i, \ldots, n\}$ and hence $\langle\phi, |\phi\rangle \leq \lambda_i$, proving (53). ∎

We can now also prove Cauchy's interlacing law (17).

Proof of Cauchy's interlacing law Let $\{e(1), \ldots, e(n)\}$ be the standard basis in $\mathbb{R}^n$. We associate M_n with the linear operator

$$(M_n\phi)_i := \sum_{j=1}^{n} \xi_{ij}\phi_j \qquad (\phi \in \mathbb{R}^n). \tag{55}$$

Let V' be the linear subspace spanned by $e(1), \ldots, e(n-1)$. Then Theorem 15 tells us that

$$\lambda_i(n-1) = \sup_{\substack{W:\,\dim(W)=i\\ W\subset V'}} \inf_{\substack{\phi\in W\\ |\phi|=1}} \langle\phi|M_n\phi\rangle \leq \sup_{W:\,\dim(W)=i} \inf_{\substack{\phi\in W\\ |\phi|=1}} \langle\phi|M_n\phi\rangle = \lambda_i(n). \tag{56}$$

The other inequality in (17) follows by applying Theorem 15 to $-M_n$. ∎

The Weilandt-Hoffmann inequality can also be derived from Theorem 15, but this is a lot of work, see Exercises 1.3.5 and 1.3.6 in the book. We will give another derivation based on Exercise 1.3.11 from the book. Our proof will not be completely rigorous. In Section 1.3.4 of the book some hints are given on how to make this rigorous but it is too much work to fill in all the details. The main idea of the proof is quite elegant, however. Assume that A^t is a hermitian matrix that depends smoothly on a parameter $t \in [0, 1]$, and that for each $t \in [0, 1]$, all eigenvalues of A^t are simple, so that we can order them $\lambda^t(1) > \cdots > \lambda^t(n)$. For each t, we can choose an orthonormal basis $\phi^t(1), \ldots, \phi^t(n)\}$ of eigenvectors. Since the eigenvalues are distinct, the normalised eigenvectors are unique up to a *phase factor*, i.e., a multiplicative constant of absolute value one. The projection operators $|\phi^t(i)\rangle\langle\phi^t(i)|$ on the corresponding eigenspaces do not have this ambiguity. Let us write

$$\dot{A}^t := \tfrac{\partial}{\partial t}A^t \quad \text{and} \quad \dot{\lambda}^t(i) := \tfrac{\partial}{\partial t}\lambda^t(i). \tag{57}$$

Then *Hadamard's first variational formula* says that

$$\dot{\lambda}^t(i) = \langle\phi^t(i)|\dot{A}^t|\phi^t(i)\rangle. \tag{58}$$

Note that the right-hand side of this equation is the trace of the product of $\dot{A}^t$ and $|\phi^t(i)\rangle\langle\phi^t(i)|$, which is well-defined (without ambiguity due to a phase factor). The derivation of (58) in the book is not correct, but let us believe it for the moment and see how it can be used to derive the Weilandt-Hoffmann inequality (30).

Proof of the Weilandt-Hoffmann inequality We will show that

$$\sqrt{\sum_{i=1}^{n} \big|\lambda_i(A) - \lambda_i(B)\big|^2} \leq \|A - B\|_{\mathrm{F}} \tag{59}$$

The idea is to write

$$A^t := A + tC \quad \big(t \in [0,1]\big) \quad \text{with} \quad C := B - A, \tag{60}$$

to write $\lambda(i) := \lambda_i(A)$ and $\lambda^t(i) := \lambda_i(A^t)$, and to show that

$$\tfrac{\partial}{\partial t} \sqrt{\sum_{i=1}^{n} \big|\lambda^t(i) - \lambda(i)\big|^2} \leq \|C\|_{\mathrm{F}} \qquad \big(t \in [0,1]\big). \tag{61}$$

The technical difficulty is that we need to show that without loss of generality we can assume that the eigenvalues are simple for all $t \in [0,1]$ and that this implies that all quantities depend smoothly on t and Hadamard's first variational formula is valid. In Section 1.3.4 of the book, some simple dimension counting argument is given that makes this plausible. For the technical details, one has to look elsewhere in the literature.

Assuming everything is all right, we calculate the left-hand side of (61) using Hadamard's first variational formula (58) with $\dot{A}^t = C$, which gives

$$\frac{\sum_{j=1}^{n} \big(\lambda^t(j) - \lambda(j)\big) \langle \phi^t(j) | C \phi^t(j) \rangle}{\sqrt{\sum_{i=1}^{n} \big|\lambda^t(i) - \lambda(i)\big|^2}}. \tag{62}$$

We recognize that this is the inner product in $\mathbb{R}^n$ of the function $j \mapsto \langle \phi^t(j) | C \phi^t(j) \rangle$ with a vector of unit length, which by Cauchy-Schwarz can be estimated from above by

$$\sqrt{\sum_{j=1}^{n} \big|\langle \phi^t(j) | C \phi^t(j) \rangle\big|^2}. \tag{63}$$

Here, using (1),

$$\begin{aligned} \sum_{j=1}^{n} \big|\langle \phi^t(j) | C \phi^t(j) \rangle\big|^2 &\leq \sum_{i=1}^{n} \sum_{j=1}^{n} \big|\langle \phi^t(i) | C | \phi^t(j) \rangle\big|^2 \\ &= \sum_{i=1}^{n} \sum_{j=1}^{n} \langle \phi^t(j) | C^* | \phi^t(i) \rangle \langle \phi^t(i) | C | \phi^t(j) \rangle = \sum_{j=1}^{n} \langle \phi^t(j) | C^* C | \phi^t(j) \rangle = \|C\|_{\mathrm{F}}^2. \end{aligned} \tag{64}$$

■

3.6 The moment method

Since the trace is independent of the choice of the basis, and M_n can be diagonalised with the eigenvalues on its diagonal, we observe that

$$\mathrm{tr}(M_n^k) = \sum_{i=1}^{k} \lambda_i^k. \tag{65}$$

It follows that

$$\int \mu_n(\mathrm{d}\lambda) \lambda^k = \frac{1}{n} \sum_{i=1}^{k} \big(\frac{\lambda_i}{\sqrt{n}}\big)^k = \frac{1}{n} \mathrm{tr}\big((M_n/\sqrt{n})^k\big). \tag{66}$$

Taking expectations, this implies that

$$\mathbb{E}\big[\int \mu_n(\mathrm{d}\lambda)\lambda^k\big] = \frac{1}{n}\sum_{i=1}^{k}\big(\frac{\lambda_i}{\sqrt{n}}\big)^k = \frac{1}{n}\mathbb{E}\big[\mathrm{tr}\big((M_n/\sqrt{n})^k\big)\big]. \tag{67}$$

Our aim is to prove that

$$\mu_n \underset{n\to\infty}{\Longrightarrow} \mu_{\mathrm{sc}} \quad \text{a.s.} \tag{68}$$

By Proposition 13, we can (and will) without loss of generality assume that $|\xi_{ij}| \leq K$ a.s. for some $K < \infty$.

Proposition 16 (Convergence of the mean and concentration) *In order to prove (68), it suffices to prove that*

$$\frac{1}{n}\mathbb{E}\big[\mathrm{tr}\big((M_n/\sqrt{n})^k\big)\big] \underset{n\to\infty}{\longrightarrow} \int \mu_{\mathrm{sc}}(\mathrm{d}\lambda)\lambda^k \qquad (k \geq 1) \tag{69}$$

and that for each $k \geq 1$, there exists constants $c_k > 0$ and $C_k < \infty$ such that

$$\frac{1}{n^2}\mathrm{Var}\big(\mathrm{tr}\big((M_n/\sqrt{n})^k\big)\big) \leq C_k n^{-c_k} \qquad (n \geq 1). \tag{70}$$

Proof By Proposition 11, in order to prove (68), it suffices to prove that

$$\mu_{n_j} \underset{j\to\infty}{\Longrightarrow} \mu_{\mathrm{sc}} \quad \text{a.s.} \tag{71}$$

for an arbitrary lacunary sequence $(n_j)_{j\in\mathbb{N}}$. We claim that it suffices to prove that

$$\int \mu_{n_j}(\mathrm{d}\lambda)\lambda^k \underset{j\to\infty}{\Longrightarrow} \int \mu_{\mathrm{sc}}(\mathrm{d}\lambda)\lambda^k \quad \text{a.s.} \quad (k \geq 1) \tag{72}$$

for an arbitrary lacunary sequence $(n_j)_{j\in\mathbb{N}}$. To see this, we use the strong Bai-Yin theorem (Theorem 2.3.24 in the book), which says that

$$\limsup_{n\to\infty} \sup_{1\leq i\leq n} |\lambda_i|/\sqrt{n} \leq 2. \tag{73}$$

It follows that almost surely, the sequence of measures $\{\mu_{n_j} : j \geq 1\}$ is tight, and each weak cluster point μ_* is concentrated on $[-2, 2]$. If we combine this with (72), then we obtain that almost surely, each weak cluster point μ_* is concentrated on $[-2, 2]$ and satisfies

$$\int \mu_*(\mathrm{d}\lambda)\lambda^k = \int \mu_{\mathrm{sc}}(\mathrm{d}\lambda)\lambda^k \quad (k \geq 1). \tag{74}$$

Since a probability measure on $[-2, 2]$ is uniquely determined by its moments, it follows that almost surely, the sequence of measures $\{\mu_{n_j} : j \geq 1\}$ is tight and μ_{sc} is its only cluster point. This implies (71).

It remains to show that (69) and (70) imply (72). From now on, we fix $k \geq 1$ and a lacunary sequence $(n_j)_{j\in\mathbb{N}}$ with $n_{j+1}/n_j \geq c$ $(j \in \mathbb{N})$ for some $c > 1$. Note that this implies that

$$n_j \geq c^j \qquad (j \in \mathbb{N}). \tag{75}$$

To ease notation, let us write

$$X_j := \int \mu_{n_j}(\mathrm{d}\lambda)\lambda^k = \frac{1}{n_j}\mathrm{tr}\big((M_{n_j}/\sqrt{n_k})^k\big) \quad \text{and} \quad X := \int \mu_{\mathrm{sc}}(\mathrm{d}\lambda)\lambda^k. \tag{76}$$

Then formulas (69) and (70) say that

$$\mathbb{E}[X_j] \underset{j\to\infty}{\longrightarrow} X \quad \text{and} \quad \mathrm{Var}(X_j) \le C_k n_j^{-c_k} \le C_k c^{-c_k j}. \tag{77}$$

By Chebyshev's inequality,

$$\mathbb{P}\big[\big|X_j - \mathbb{E}[X_j]\big| \ge \varepsilon\big] \le \mathrm{Var}(X_j)/\varepsilon^2 \le \varepsilon^{-2} C_k c^{-c_k j}. \tag{78}$$

Since $\sum_j \varepsilon^{-2} C_k c^{-c_k j} < \infty$, by Borel-Cantelli, this implies that almost surely, for each $\varepsilon > 0$, the set

$$\big\{j : \big|X_j - \mathbb{E}[X_j]\big| \ge \varepsilon\big\} \tag{79}$$

is finite. It follows that

$$X - \varepsilon \le \liminf_{j\to\infty} X_j \le \limsup_{j\to\infty} X_j \le X + \varepsilon \quad \text{a.s.} \tag{80}$$

for each $\varepsilon > 0$, which implies (72). ∎

Our first aim is now to check the condition (70). We need one preparatory lemma.

Proposition 17 (Concentration of measure) *For each $k \ge 1$, there exists constants $c_k > 0$ and $C_k < \infty$ such that (70) holds.*

Proof attempt (unsuccesful) We try to prove the statement only for even k following Exercise 2.4.7 (i) in the book. For all even $k \ge 2$,

$$\|A\|_{\mathrm{S}_k} := \mathrm{tr}(A^k)^{1/k} \tag{81}$$

defines a norm on the space of real symmetric $n \times n$ matrices, called the k-Schatten norm. In particular,

$$\|A\|_{\mathrm{S}_2} = \|A\|_{\mathrm{F}}^2 := \sum_{i=1}^n \sum_{j=1}^n |A_{ij}|^2 \tag{82}$$

is the Frobenius norm defined in (31). The hint of Exercise 2.4.7 (i) is to use that the k-Schatten norm, as any norm, is a convex function, and to apply Talagrand's inequality (Theorem 2.1.13 in the book). Talagrand's inequality applies to convex functions that are defined on $\mathbb{R}^m$ and are 1-Lipschitz with respect to the euclidean norm. We can naturally identify the space of all real symmetric $n \times n$ matrices with the space $\mathbb{R}^{n(n+1)/2}$. In view of (82), the euclidean norm then corresponds to the Frobenius norm which is the 2-Schatten norm. If a symmetric matrix A has eigenvalues $\vec{\lambda} = (\lambda_1, \dots, \lambda_n)$, then $\mathrm{tr}(A^k)^{1/k} = \|\vec{\lambda}\|_k$ is just the ℓ_k-norm of the vector $\vec{\lambda}$. Since on ℓ^p spaces, we have the inequality $\|\cdot\|_p \le \|\cdot\|_q$ for $1 \le q \le p \le \infty$, we have

$$\|A\|_{\mathrm{S}_k} \le \|A\|_{\mathrm{S}_2}$$

for all even $k \ge 2$, which shows that $\|\cdot\|_{\mathrm{S}_k}$ is 1-Lipschitz with respect to the euclidean norm on $\mathbb{R}^{n(n+1)/2}$. Thus Talagrand's inequality is applicable.

Talagrand's inequality now tells us that there exist constants $c > 0$ and $C < \infty$ such that for any random symmetric matrix A whose entries satisfy $|A_{ij}| \le a$ a.s.,

$$\mathbb{P}\big[\big|\mathrm{tr}(A^k)^{1/k} - \mathbb{M}[\mathrm{tr}(A^k)^{1/k}]\big| \ge \varepsilon a\big] \le C e^{-c\varepsilon^2} \qquad (\varepsilon > 0), \tag{83}$$

where $\mathbb{M}[X]$ denotes the median[1] We want to apply this to $A = M_n/\sqrt{n}$. Recall that the entries of M_n satisfy $|\xi_{ij}| \le K$ a.s., so that the entries of $M_n/\sqrt{n}$ satisfy $|\xi_{ij}|/\sqrt{n} \le K/\sqrt{n}$ a.s. Setting $a = K/\sqrt{n}$ gives

$$\mathbb{P}\big[\big|\mathrm{tr}((M_n/\sqrt{n})^k)^{1/k} - \mathbb{M}[\mathrm{tr}((M_n/\sqrt{n})^k)^{1/k}]\big| \ge \varepsilon K/\sqrt{n}\big] \le C e^{-c\varepsilon^2} \qquad (\varepsilon > 0). \tag{84}$$

[1] Defined as $\sup\{m : \mathbb{P}[X < m] < \frac{1}{2}\}$ or $\inf\{m : \mathbb{P}[X < m] > \frac{1}{2}\}$, it doesn't matter which definition we use.

This is close in spirit to (70), but really deriving (70) from (84) is not so easy and in fact cannot be done without some additional estimate on the size of $\mathbb{M}[\mathrm{tr}((M_n/\sqrt{n})^k)^{1/k}]$. To see this, let us set

$$T := \mathbb{M}[\mathrm{tr}((M_n/\sqrt{n})^k)] \quad \text{so that} \quad \mathbb{M}[\mathrm{tr}((M_n/\sqrt{n})^k)^{1/k}] = \mathbb{M}[\mathrm{tr}((M_n/\sqrt{n})^k)]^{1/k} = T^{1/k}. \tag{85}$$

In order to prove (70), we need that the standard deviation of $\mathrm{tr}((M_n/\sqrt{n})^k)$ is of order $\sqrt{n^2 n^{-c_k}} = n^{1-c_k/2}$ for some $c_k > 0$. If we want to derive this from (84), then (at the very least) we should be capable of showing that the probability that

$$\mathrm{tr}((M_n/\sqrt{n})^k) - T \geq L n^{1-c_k} \tag{86}$$

is small when L is large. Now let us see what kind of upper estimates on $\mathrm{tr}((M_n/\sqrt{n})^k) - T$ can be derived from (84). Formula (84) tells us that

$$\mathbb{P}\big[\mathrm{tr}((M_n/\sqrt{n})^k)^{1/k} \geq T^{1/k} + \varepsilon K/\sqrt{n}\big] \leq C e^{-c\varepsilon^2}, \tag{87}$$

or equivalently,

$$\mathbb{P}\big[\mathrm{tr}((M_n/\sqrt{n})^k) - T \geq (T^{1/k} + \varepsilon K/\sqrt{n})^k - T\big] \leq C e^{-c\varepsilon^2}. \tag{88}$$

Here

$$(T^{1/k} + \varepsilon K/\sqrt{n})^k - T = \int_0^{\varepsilon K/\sqrt{n}} \mathrm{d}z \, \tfrac{\partial}{\partial z}(T^{1/k} + z)^k = \int_0^{\varepsilon K/\sqrt{n}} \mathrm{d}z \, k(T^{1/k} + z)^{k-1}. \tag{89}$$

In view of (69), we expect T to be of order n as $n \to \infty$. If that is true, then the expression in (89) is of order $n^{-1/2}(n^{1/k})^{k-1} = n^{1-1/2-1/k}$. Thus, if we use that T is of order n, then formula (84) allows us to show that the probability in (86) is small when L is large. With some extra work, it seems that now one should be able to derive (70). However, if we have no additional information about the size of T, then it seems such a conclusion cannot be drawn. Indeed, if T would be of order n^2, then the expression in (89) would be of order $n^{-1/2}(n^{2/k})^{k-1} = n^{2-1/2-2/k}$, which for $k \geq 4$ is larger n^{1-c_k}. ∎

About Exercise 2.4.8 in the book Here one has to show that as $n \to \infty$,

$$\mathbb{E}\big[\big(\mathrm{tr}(M_n^k)\big)^2\big] = n^{k+2} + o(n^{k+2}). \tag{90}$$

Similar to formula (2.70) in the book, we can rewrite the left-hand side as

$$\sum_{1 \leq i_1, \ldots, i_n \leq n} \sum_{1 \leq j_1, \ldots, j_n \leq n} \mathbb{E}\big[\xi_{i_1 i_2} \cdots \xi_{i_k i_1} \xi_{j_1 j_2} \cdots \xi_{j_k j_1}\big], \tag{91}$$

where now we sum over all possible combinations to choose *two* cycles $i_1, \ldots, i_k$ and $j_1, \ldots, j_k$. These cycles form a graph where each edge of the graph is used at least once by a step of one of the cycles. As in Subsection 2.3.4 in the book, no edge of the graph can be used precisely once (by both cycles together), since different edges are independent and the ξ_{ij}'s have mean zero, so such terms all give zero. Now you can group the terms you get according to how many distinct vertices your graoh has. You then get a factor $n(n-1)\cdots$ for each vertex that you can choose. From this, you see that the leading order terms come from graphs that have the maximal number of distinct vertices while satsfying the constraint that each edge used at least twice. Now one can see that these are precisely the graphs that consist of two disjoint components, each of which has the tree structure of Subsection 2.3.4. These have $2(k/2+1) = k+2$ vertices so the leading order term is of order n^{k+2}. The prefactor is precisely $C_{k/2}^2$, because you choose two cycles as before.

4 The Stieltjes transform

4.1 Set-up

We work in the set-up of Subsection 3.3. We recall that $(\xi_{ij})_{i,j\in\mathbb{N}_+}$ are real random variables such that

(i) $(\xi_{ij})_{i<j}$ are i.i.d. with mean zero and variance one,

(ii) $(\xi_{ij})_{i=j}$ are i.i.d. with mean zero and finite variance, and independent of $(\xi_{ij})_{i<j}$,

(iii) $\xi_{ji} = \xi_{ij}$ $(i < j)$.

We set $M_n := (\xi_{ij})_{1\leq i,j\leq n}$, we let $\lambda_1(n) \geq \cdots \geq \lambda_n(n)$ denote its eigenvalues, and let μ_n be the random probability measures on $\mathbb{R}$ defined by

$$\mu_n := \frac{1}{n}\sum_{i=1}^{n} \delta_{\lambda_i(n)/\sqrt{n}}, \tag{92}$$

Then Theorem 2.4.2 from the book says that

$$\mu_n \underset{n\to\infty}{\Longrightarrow} \mu_{\mathrm{sc}} \quad \text{a.s.}, \tag{93}$$

where μ_{sc} denotes the semicircle law

$$\mu_{\mathrm{sc}}(\mathrm{d}x) := \frac{1}{2\pi}\big(4 - |x|^2\big)_+^{1/2}\mathrm{d}x. \tag{94}$$

Formula (93) implies that

$$\mathbb{E}\big[\mu_n\big] \underset{n\to\infty}{\Longrightarrow} \mu_{\mathrm{sc}} \quad \text{a.s.}. \tag{95}$$

Conversely, it is possible to derive (93) from (95) if one can show that the random measures μ_n are with high probability close to their mean $\mathbb{E}[\mu_n]$. In the previous section, we have seen how (95) can be proved by showing that the moments of $\mathbb{E}[\mu_n]$ converge to those of μ_{sc}. In the present section, we will take another approach and show that the Stieltjes transform of $\mathbb{E}[\mu_n]$ converge to that of μ_{sc}.

4.2 The Stieltjes transform

If μ is a probability measure on $\mathbb{R}$, then its *Stieltjes transform* is the function $s_\mu : \mathbb{C}\backslash\mathbb{R} \to \mathbb{C}$ defined as

$$s_\mu(z) := \int_{\mathbb{R}} \mu(\mathrm{d}x)\, \frac{1}{x - z} \qquad (z \in \mathbb{C}\backslash\mathbb{R}). \tag{96}$$

We observe that

$$\Big|\frac{1}{x-z}\Big| = \frac{1}{\sqrt{\big(x - \Re(z)\big)^2 + \Im(z)^2}} \leq \frac{1}{|\Im(z)|} \qquad (x \in \mathbb{R},\ z \in \mathbb{C}), \tag{97}$$

so the integrand in (96) is absolutely integrable for each $z \notin \mathbb{R}$, and

$$\big|s_\mu(z)\big| \leq \frac{1}{|\Im(z)|} \qquad (x \in \mathbb{R},\ z \in \mathbb{C}). \tag{98}$$

Also,

$$s_\mu(\overline{z}) = \overline{s_\mu(z)} \qquad (z \in \mathbb{C}\backslash\mathbb{R}), \tag{99}$$

so it suffices to know s_μ on the upper half-plane

$$C^+ := \big\{z \in \mathbb{C} : \Im(z) > 0\big\}. \tag{100}$$

One can check that $z \mapsto s_\mu(z)$ is complex differentiable on $\mathbb{C}\backslash\mathbb{R}$, i.e., the limit

$$\tfrac{\partial}{\partial z} s_\mu(z) := \lim_{w\to 0} w^{-1}\big(s_\mu(z+w) - s_\mu(w)\big) \tag{101}$$

(with w running through $\mathbb{C}$) exists for all $z \in \mathbb{C}\backslash\mathbb{R}$. We recall from complex number theory that this implies that s_μ is complex analytic on $\mathbb{C}\backslash\mathbb{R}$. In particular, it is infinitely differentiable. Since

$$\Im\Big(\frac{1}{x-z}\Big) = \Im\Big(\frac{\overline{x-z}}{|x-z|^2}\Big) = \frac{\Im(z)}{|x-z|^2}, \tag{102}$$

we observe that s_μ maps $\mathcal{C}^+$ into itself. The following lemma shows that the measure μ is uniquely determined by s_μ.

Lemma 18 (Stieltjes transform near the real axis) *For each real $\varepsilon > 0$, let $h_\varepsilon : \mathbb{R} \to \mathbb{R}$ be defined by*

$$h_\varepsilon(x) := \frac{1}{\pi}\Im\big(s_\mu(x+i\varepsilon)\big) \qquad (x \in \mathbb{R}). \tag{103}$$

Then h_ε takes values in $[0,\infty)$ and setting $\mu_\varepsilon(\mathrm{d}x) := h_\varepsilon(x)\mathrm{d}x$ defines a probability measure on $\mathbb{R}$. These probability measures satisfy

$$\mu_\varepsilon \underset{\varepsilon\to 0}{\Longrightarrow} \mu \tag{104}$$

where $\Rightarrow$ denotes weak convergence of probability laws on $\mathbb{R}$.

Proof (sketch) We observe that

$$h_\varepsilon(y) = \int_{\mathbb{R}} \mu(\mathrm{d}x) P_\varepsilon(y-x) \quad \text{with} \quad P_\varepsilon(x) := \frac{1}{\pi}\Im\Big(\frac{1}{x-i\varepsilon}\Big) = \frac{1}{\pi}\frac{\varepsilon}{x^2+\varepsilon^2}. \tag{105}$$

It is well-known that the measures $P_\varepsilon(x)\mathrm{d}x$ are probability measures that approximate the delta-measure as $\varepsilon \to 0$, so the convolutions $P_\varepsilon * \mu$ approximate μ. ∎

Similarly to Lemma 18, one can show that a sequence of probability measures μ_n converges weakly to a limit μ if and only if s_{μ_n} converges pointwise to s_μ, i.e.,

$$\mu_n \underset{n\to\infty}{\Longrightarrow} \mu \quad \Leftrightarrow \quad s_{\mu_n}(z) \underset{n\to\infty}{\longrightarrow} s_\mu(z) \quad \forall z \in \mathbb{C}^+. \tag{106}$$

See Exercise 2.4.10 in the book.

4.3 The Stieltjes transform method

We want to prove (95) using (106). Recall from Theorem 2 that a linear operator A on a complex finite dimensional inner product space is normal (i.e., commutes with its adjoint) if and only it is diagonal with respect to some orthonormal basis $\{e(1),\dots,e(n)\}$. In this case

$$A = \sum_{i=1}^n \lambda_i |e(i)\rangle\langle e(i)|, \tag{107}$$

where $\lambda_1,\dots,\lambda_n$ are the eigenvalues of A. For the moment method (see Subsection 3.6), we made use of the fact that

$$\mathrm{tr}(A^k) = \sum_{i=1}^n \lambda_i^k \tag{108}$$

for all integers $k \geq 1$. Formula (108) follows from the fact that for any linear operator B one has $\mathrm{tr}(A^k) = \sum_{i=1}^n (A^k)_{ii}$ holds with respect to any basis, so in particular also with respect to the basis that makes A diagonal. This argument is not restricted to positive integer powers only. It clearly also holds for $k = 0$ and if $\lambda_i \neq 0$ for all $1 \leq i \leq n$, which is equivalent to A being invertible, then it also holds for negative integer powers. We observe that for each $z \in \mathbb{C}$, the linear operator

$$\frac{1}{\sqrt{n}} M_n - zI \tag{109}$$

is normal. Since the eigenvalues of M_n are real, the operator in (109) is invertible if $z \notin \mathbb{R}$, so (108) tells us that

$$\mathrm{tr}\Big(\big(\frac{1}{\sqrt{n}} M_n - zI \big)^{-1} \Big) = \sum_{i=1}^n (\lambda_i(n)/\sqrt{n} - z)^{-1}, \tag{110}$$

and hence the Stieltjes transform of the probability measure $\mathbb{E}[\mu_n]$, with μ_n defined in (92), is given by

$$\begin{aligned} s_n(z) &:= s_{\mathbb{E}[\mu_n]}(z) = \mathbb{E}\Big[\int_R \mu_n(\mathrm{d}x) \, \frac{1}{x - z} \Big] \\ &= \mathbb{E}\Big[\frac{1}{n} \sum_{i=1}^n (\lambda_i(n)/\sqrt{n} - z)^{-1} \Big] = \frac{1}{n} \mathbb{E}\Big[\mathrm{tr}\Big(\big(\frac{1}{\sqrt{n}} M_n - zI \big)^{-1} \Big) \Big]. \end{aligned} \tag{111}$$