Lectures 5–8: Wick's Theorem and the Semicircle Law

Pair partitions, counting paths, and the Stieltjes transform

OUTLINE L5 Wick's theorem $\longrightarrow$ the Gaussian Wigner matrix $\longrightarrow$ $\|Mu\|^2$ for a fixed direction

L6 The semicircle law $\longrightarrow$ the strategy $\longrightarrow$ $\mathbb{E}\,\mathrm{tr}\,M^2$, $\mathbb{E}\,\mathrm{tr}\,M^4$ by hand $\longrightarrow$ traces are sums over paths $\longrightarrow$ the same two moments by paths $\longrightarrow$ every moment

L7 The Stieltjes transform $\longrightarrow$ counting eigenvalues by a contour integral $\longrightarrow$ what to compute $\longrightarrow$ a generating function at $z = \infty$

L8 Schur complement $\longrightarrow$ the self-consistent equation $\longrightarrow$ the trace concentrates $\longrightarrow$ solving it: the semicircle law

HOW TO READ Sections marked $*$ are not lectured. They are complete, they are examinable, and they carry most of the exercises. Exercises are typed A and B; there are three per lecture.

1 Wick's theorem, and Gaussian random matrices

Q. What can be computed about a random matrix without touching its eigenvalues?

Every moment of a Gaussian vector is a counting problem, and the entries of a symmetric Gaussian matrix are a Gaussian vector. So anything polynomial in the entries — the length of Mu for a fixed vector u, for instance — can be computed exactly.

1.1 Pair partitions

Definition 1.1 (Pair partition)**.** A *pair partition* of $\{1, \dots, 2q\}$ is a partition into q blocks of size two.

Theorem 1.2 (Wick)**.** *Let $X \sim \mathcal{N}(0, \Sigma)$ on $\mathbb{R}^d$ and $a_1, \dots, a_m \in \{1, \dots, d\}$. If m is odd then $\mathbb{E}[X_{a_1} \cdots X_{a_m}] = 0$. If $m = 2q$ then*

$$\boxed{\mathbb{E}\big[X_{a_1} X_{a_2} \cdots X_{a_{2q}}\big] \;=\; \sum_{\textit{pair partitions } P \textit{ of } \{1,\dots,2q\}} \; \prod_{\{r,s\} \in P} \Sigma_{a_r a_s}}$$

Proof. By Corollary 1.7 of Lectures 1–2 every linear functional of X is a one-dimensional Gaussian, so X has moments of every order, φ_X is infinitely differentiable, and its derivatives may be taken inside the expectation. Since $\partial_{\xi_j} e^{-i\langle x, \xi \rangle} = -i x_j e^{-i\langle x, \xi\rangle}$, multiplication by X_j acts on the transform as $i\,\partial_{\xi_j}$:

$$\mathbb{E}\big[X_{a_1} \cdots X_{a_m}\big] \;=\; i^m\, \partial_{\xi_{a_1}} \cdots \partial_{\xi_{a_m}} \varphi_X(\xi)\big|_{\xi = 0}. \tag{1}$$

By Definition 1.3 of Lectures 1–2, $\varphi_X(\xi) = \exp\big(-\tfrac{1}{2}\xi^\top \Sigma \xi\big)$ — for every $\Sigma \succeq 0$, degenerate cases included — and differentiating it once,

$$\partial_{\xi_i} \varphi_X(\xi) \;=\; -\sum_j \Sigma_{ij}\, \xi_j\, \varphi_X(\xi). \tag{2}$$

Apply $\partial_{\xi_{a_2}} \cdots \partial_{\xi_{a_m}}$ to (2) taken at $i = a_1$. The factor ξ_j is linear, so Leibniz leaves terms of two kinds: those in which every derivative falls on φ_X, which still carry a factor ξ_j; and, for each

$s \in \{2, \dots, m\}$, the one in which $\partial_{\xi_{a_s}}$ falls on ξ_j and produces δ_{ja_s}. Setting $\xi = 0$ kills the first kind, leaving

$$\partial_{\xi_{a_1}} \cdots \partial_{\xi_{a_m}} \varphi_X\big|_0 \;=\; -\sum_{s=2}^{m} \Sigma_{a_1 a_s}\; \partial_{\{\xi_{a_r}:\, r\neq 1,s\}} \varphi_X\big|_0.$$

Rewrite both sides with (1) and multiply through by i^m; since $i^2 = -1$ the two minus signs cancel and

$$\mathbb{E}\big[X_{a_1} \cdots X_{a_m}\big] \;=\; \sum_{s=2}^{m} \Sigma_{a_1 a_s}\; \mathbb{E}\Big[\prod_{r\neq 1,s} X_{a_r}\Big] : \tag{3}$$

pair the first factor with each of the others in turn, and recurse. Induct on m. Both sides vanish for $m = 1$, and for $m = 2$ the claim is the definition of Σ. For $m > 2$, (3) drops the number of factors by two, so parity is preserved: an odd product reduces to $\mathbb{E}[X_a] = 0$, and an even one to a sum over pair partitions of the remaining $2q - 2$ indices, each prefixed by $\Sigma_{a_1 a_s}$ recording the block containing 1. Summing over s enumerates every pair partition of $\{1, \dots, 2q\}$ once. ∎

Remark 1.3. Reading (3) with $F(X) = X_{a_2} \cdots X_{a_m}$ and noting that $\partial_j F$ deletes the factor X_j gives

$$\mathbb{E}\big[X_i\, F(X)\big] \;=\; \sum_j \Sigma_{ij}\, \mathbb{E}\big[\partial_j F(X)\big],$$

Gaussian integration by parts, which for $d = 1$ and $\Sigma = 1$ is $\mathbb{E}[gF(g)] = \mathbb{E}[F'(g)]$ — Lemma 1.10 of Lectures 1–2, and Stein's lemma in Lecture 4. It is Exercise 2.3 of Lectures 3–4; here it falls out of Wick's recursion for free.

1.2 Three examples

Corollary 1.4. *There are* $(2q-1)!! = (2q-1)(2q-3)\cdots 3 \cdot 1$ *pair partitions of* $2q$ *points, and* $\mathbb{E}[g^{2q}] = (2q-1)!!$ *for* $g \sim \mathcal{N}(0,1)$.

Proof. Take $d = 1$ and $\Sigma = 1$ in Theorem 1.2: every product of covariances equals 1, so the moment is the number of pair partitions. That number is $(2q-1)!!$ because the point 1 has $2q-1$ choices of partner and deleting that block leaves $2q-2$ points. ∎

Example 1.5 (Four indices)**.** For $q = 2$ the three pair partitions of $\{1,2,3,4\}$ are $\{12\}\{34\}$, $\{13\}\{24\}$, $\{14\}\{23\}$, so

$$\mathbb{E}\big[X_1X_2X_3X_4\big] = \Sigma_{12}\Sigma_{34} + \Sigma_{13}\Sigma_{24} + \Sigma_{14}\Sigma_{23}.$$

Setting all four indices equal recovers $\mathbb{E}[X_1^4] = 3\Sigma_{11}^2$.

Example 1.6 (A quadratic form)**.** Let $X \sim \mathcal{N}(0,\Sigma)$ on $\mathbb{R}^d$ and let A be symmetric. Then $\mathbb{E}[X^\top AX] = \sum_{ij} A_{ij}\mathbb{E}[X_iX_j] = \operatorname{tr}(A\Sigma)$, no Wick needed. For the second moment,

$$\mathbb{E}\big[(X^\top AX)^2\big] = \sum_{ijkl} A_{ij}A_{kl}\,\mathbb{E}\big[X_iX_jX_kX_l\big] = \sum_{ijkl} A_{ij}A_{kl}\big(\Sigma_{ij}\Sigma_{kl} + \Sigma_{ik}\Sigma_{jl} + \Sigma_{il}\Sigma_{jk}\big)$$

by Example 1.5. The first term is $(\operatorname{tr} A\Sigma)^2$; the other two are each $\operatorname{tr}(A\Sigma A\Sigma)$, by symmetry of A and Σ. Hence

$$\operatorname{Var}\big(X^\top AX\big) = 2\operatorname{tr}\big((A\Sigma)^2\big), \qquad \text{and in particular} \qquad \operatorname{Var}\big(\|X\|^2\big) = 2\operatorname{tr}\Sigma^2.$$

1.3 The Gaussian Wigner matrix

Definition 1.7 (Gaussian Wigner matrix)**.** Let A be $N \times N$ with i.i.d. $\mathcal{N}(0,1)$ entries and put

$$M = \frac{1}{\sqrt{2N}}\big(A + A^\top\big).$$

Then M is symmetric, its entries $\{M_{ij}\}_{i\le j}$ are independent centered Gaussians with $\mathbb{E}M_{ij}^2 = \frac{1}{N}$ for $i<j$ and $\mathbb{E}M_{ii}^2 = \frac{2}{N}$, and for *all* four indices

$$\mathbb{E}\big[M_{ij}M_{kl}\big] = \tfrac{1}{N}\big(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\big). \tag{4}$$

This ensemble is the *Gaussian orthogonal ensemble*, or GOE, for the reason in Remark 1.8. Everything below is (4) and Theorem 1.2; the construction itself never appears again. The two delta terms are the two ways of matching $\{i,j\}$ with $\{k,l\}$, and they are there because $M_{ij} = M_{ji}$: an entry may be met twice under either name. The diagonal variance $\frac{2}{N}$ comes out of the symmetrization (Exercise 1.9), and it is exactly what makes (4) hold at $i=j=k=l$ without a special case. The N diagonal entries are a vanishing fraction of the N^2.

Remark 1.8 (Why "orthogonal")*.* For any fixed orthogonal O, $OAO^\top$ again has i.i.d. $\mathcal{N}(0,1)$ entries — each entry is a unit linear functional of A, and Theorem 1.5 of Lectures 1–2 makes the collection jointly Gaussian with the same covariance — so $OMO^\top \overset{d}{=} M$. The law of the spectrum therefore cannot depend on the basis. This is the matrix form of the rotational invariance of Corollary 1.6 of Lectures 1–2.

Exercise 1.9 (TYPE A)**.** Verify Definition 1.7 from the construction: with A having i.i.d. $\mathcal{N}(0,1)$ entries and $M = \frac{1}{\sqrt{2N}}(A + A^\top)$, compute $\mathbb{E}[M_{ij}M_{kl}]$ directly and obtain (4). Check the case $i=j=k=l$ separately and confirm that the diagonal convention $\mathbb{E}M_{ii}^2 = \frac{2}{N}$ is forced, not chosen.

1.4 One direction at a time

Before the eigenvalues, a question with the same shape but no linear algebra in it: fix a unit vector u and ask how long Mu is.

Proposition 1.10. *Let M be as in Definition 1.7 and let $u \in \mathbb{R}^N$ with $\|u\| = 1$. Then Mu is a centered Gaussian vector with covariance*

$$C \;=\; \mathbb{E}\big[(Mu)(Mu)^\top\big] \;=\; \frac{1}{N}\big(I + uu^\top\big),$$

and consequently

$$\mathbb{E}\big\|Mu\big\|^2 = 1 + \frac{1}{N}, \qquad \operatorname{Var}\big\|Mu\big\|^2 = \frac{2}{N} + \frac{6}{N^2}.$$

Proof. Each coordinate $(Mu)_i = \sum_j M_{ij}u_j$ is a linear functional of the entries of M, so by Theorem 1.5 of Lectures 1–2 the vector Mu is centered Gaussian. Its covariance is (4) contracted with u twice:

$$\mathbb{E}\big[(Mu)_i(Mu)_k\big] = \sum_{j,l} u_j u_l \cdot \tfrac{1}{N}\big(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\big) = \tfrac{1}{N}\big(\delta_{ik}\|u\|^2 + u_i u_k\big),$$

which is C. Then $\mathbb{E}\|Mu\|^2 = \operatorname{tr} C = \frac{1}{N}(N+1)$. For the variance, Example 1.6 with $A = I$ and $\Sigma = C$ gives $\operatorname{Var}\|Mu\|^2 = 2\operatorname{tr} C^2$, and since $u^\top u = 1$,

$$\big(I + uu^\top\big)^2 = I + 2uu^\top + u(u^\top u)u^\top = I + 3uu^\top, \qquad \operatorname{tr} C^2 = \tfrac{1}{N^2}\big(N+3\big). \qquad \blacksquare$$

Remark 1.11. Two things are worth reading off. First, the answer does not depend on u — it cannot, because $OMO^\top \overset{d}{=} M$ by Remark 1.8, so the law of $\|Mu\|^2$ is the same in every direction. Second, the standard deviation is $\sqrt{2/N}$ against a mean of 1: for large N a Gaussian Wigner matrix maps *every* unit vector to a vector of length very close to 1. That is already a statement about eigenvalues, since $\|Mu\|^2 = u^\top M^2 u$ is a weighted average of the λ_i^2 with weights $\langle u, \phi_i\rangle^2$ summing to 1, where ϕ_i are the eigenvectors. Averaging over directions instead of fixing one leads to $\frac{1}{N}\operatorname{tr} M^2$, which is taken up in §2.3.

1.5 $*$ Cumulants

Not lectured. Theorem 1.2 is the Gaussian instance of a formula that holds for any random vector with moments. This section states that formula; the exercises develop the properties that make it useful.

Definition 1.12 (Joint cumulants)**.** For $Y_1, \dots, Y_m$ with all moments finite,

$$\kappa\big(Y_1, \dots, Y_m\big) \;=\; \frac{\partial^m}{\partial t_1 \cdots \partial t_m} \log \mathbb{E}\big[e^{t_1 Y_1 + \cdots + t_m Y_m}\big]\bigg|_{t=0}.$$

Moments are derivatives of the transform; cumulants are derivatives of its logarithm. The mth cumulant of a single variable is $\kappa(Y, \dots, Y)$ with m arguments; when the argument is itself a sum, keeping the m slots visible is what makes multilinearity usable.

Stated without proof: the moment–cumulant formula. For any $Y_1, \dots, Y_m$ with moments,

$$\mathbb{E}\big[Y_1 Y_2 \cdots Y_m\big] \;=\; \sum_{\pi \in \mathcal{P}(m)} \prod_{B \in \pi} \kappa\big(Y_i : i \in B\big),$$

where $\mathcal{P}(m)$ is the set of *all* partitions of $\{1, \dots, m\}$ into non-empty blocks, of which there are $1, 2, 5, 15, 52, \dots$

For $m = 2$ this reads $\mathbb{E}[Y_1 Y_2] = \mathbb{E}Y_1 \mathbb{E}Y_2 + \operatorname{Cov}(Y_1, Y_2)$, the two partitions of $\{1, 2\}$. If the Y_i are coordinates of a centered Gaussian vector, every joint cumulant of order other than two vanishes (Exercise 1.14), so a partition contributes nothing unless all of its blocks have size two. Those are the pair partitions, and the moment–cumulant formula becomes Theorem 1.2. Of the fifteen partitions of a four-element set, three survive, which is Example 1.5.

Remark 1.13. For a non-Gaussian vector the larger blocks do not vanish, $\log \mathbb{E}[e^{\langle t, X\rangle}]$ is no longer a polynomial, and the recursion (3) acquires a tail that never terminates. Cumulants then measure the distance from Gaussian, and because mixed cumulants of independent variables vanish, they are the natural bookkeeping for sums of independent variables.

Exercise 1.14 (TYPE B)**.** Prove the following directly from Definition 1.12.

(a) **Multilinearity.** κ is symmetric in its arguments and linear in each, and for $m \geq 2$ nothing changes when a constant is added to any argument. In particular

$$\kappa(cY, \dots, cY) = c^m \, \kappa(Y, \dots, Y).$$

(b) **Independence.** If the arguments split into two non-empty groups that are independent of one another, the cumulant vanishes. Deduce that for independent Y, Z,

$$\kappa(Y + Z, \dots, Y + Z) = \kappa(Y, \dots, Y) + \kappa(Z, \dots, Z).$$

(c) **The Gaussian is where they stop.** For $X \sim \mathcal{N}(0, \Sigma)$, $\kappa(X_a, X_b) = \Sigma_{ab}$ and every joint cumulant of order $\neq 2$ vanishes.

For (a), give the two variables separate arguments s, s' in the log-transform and apply the chain rule once. For (c), use Definition 1.3 of Lectures 1–2 with $\xi = it$.

Exercise 1.15 (TYPE B)**.** Let $X_1, X_2, \dots$ be i.i.d. with mean μ and variance σ^2, and put

$$S_n = \frac{1}{\sqrt{n}} \sum_{i=1}^{n} (X_i - \mu).$$

Using only parts (a) and (b) of Exercise 1.14, show that for $m \geq 2$, with m arguments in each cumulant,

$$\kappa(S_n, \dots, S_n) = n^{1-m/2}\, \kappa(X_1, \dots, X_1).$$

Deduce that $\kappa(S_n, S_n) = \sigma^2$ for every n, while every cumulant of order ≥ 3 tends to 0.

This is the central limit theorem seen through cumulants: the $\sqrt{n}$ normalization is exactly the rescaling that freezes the second cumulant and kills everything above it.

ASSIGNED **Lecture 5**, due at the start of Lecture 6: Exercises 1.9, 1.14, 1.15. *Three problems. The first checks the definition; the other two are the cumulant calculation we did not have time for in class.*

2 The semicircle law

Q. Where are the eigenvalues of a large random symmetric matrix?

Their moments are normalized traces. A trace is a sum over closed paths, and at leading order only the paths that run along a tree contribute. Pairing the two steps on each edge of such a path gives a non-crossing pair partition, and there are Catalan-many of those.

2.1 The theorem

Recall Definition 1.7: $M = \frac{1}{\sqrt{2N}}(A + A^\top)$ with A having i.i.d. $\mathcal{N}(0,1)$ entries, so that M is symmetric, its entries $\{M_{ij}\}_{i \leq j}$ are independent centered Gaussians, and

$$\mathbb{E}\big[M_{ij} M_{kl}\big] = \tfrac{1}{N}\big(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\big), \qquad \text{in particular} \qquad \mathbb{E}M_{ij}^2 = \tfrac{1}{N}\ (i \neq j), \quad \mathbb{E}M_{ii}^2 = \tfrac{2}{N}.$$

M has N real eigenvalues $\lambda_1, \dots, \lambda_N$. Figure 1 shows their histogram.

Theorem 2.1 (Wigner)**.** *Let $\rho_{sc}(x) = \frac{1}{2\pi}\sqrt{4 - x^2}$ on $[-2, 2]$. For every $k \geq 0$,*

$$\frac{1}{N} \sum_{i=1}^{N} \lambda_i^k \longrightarrow \int_{-2}^{2} x^k \rho_{sc}(x)\, dx \qquad \textit{in probability.}$$

Stated without proof: the method of moments. A probability density with bounded support is determined by its moments. Consequently Theorem 2.1 implies $\frac{1}{N}\sum_i f(\lambda_i) \to \int f \rho_{sc}$ in probability for every bounded continuous f: the histogram converges to ρ_{sc}.

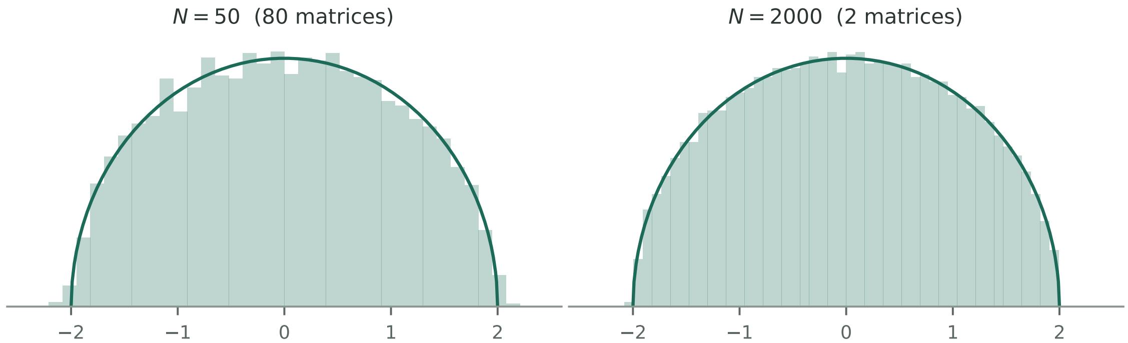

Figure 1: Histograms of the eigenvalues of Gaussian Wigner matrices, normalized to have area one, with the curve $\frac{1}{2\pi}\sqrt{4-x^2}$ drawn over them.

2.2 The strategy

The proof rests on four ideas.

1. **Moments converge to moments.** If the eigenvalues have a limiting distribution with density ρ, in the sense that $\frac{1}{N}\sum_i f(\lambda_i) \to \int f\rho$, then taking $f(x) = x^k$ the eigenvalue moments $\frac{1}{N}\sum_i \lambda_i^k$ should converge to the moments of ρ. And the eigenvalue moments are traces (Lemma 2.2).
2. **The average is a Gaussian computation plus combinatorics.** $\operatorname{tr} M^k$ is a polynomial in the entries of M, so Theorem 1.2 computes $\mathbb{E}\operatorname{tr} M^k$. What is left is to organize the sum, which is a counting problem (§2.3–§2.6).
3. **The fluctuations are small.** If $\operatorname{Var}\left(\frac{1}{N}\operatorname{tr} M^k\right)/\left(\frac{1}{N}\mathbb{E}\operatorname{tr} M^k\right)^2 \to 0$, Chebyshev's inequality turns convergence of the mean into convergence in probability (§2.7).
4. **Identify the limit.** The limits of the means turn out to be the Catalan numbers, and these are the moments of ρ_{sc} (Proposition 2.17, Exercise 2.21).

Lemma 2.2. *For every $k \geq 1$,*

$$\frac{1}{N}\sum_{i=1}^{N} \lambda_i^k \;=\; \frac{1}{N}\operatorname{tr} M^k.$$

Proof. Diagonalize: $M = O\Lambda O^\top$ with O orthogonal and Λ diagonal, so $M^k = O\Lambda^k O^\top$ and $\operatorname{tr} M^k = \operatorname{tr}\Lambda^k = \sum_i \lambda_i^k$ by cyclicity of the trace. ∎

Remark 2.3 (The empirical spectral measure). The "limiting distribution" is the limit of the probability measure $\hat\rho_M = \frac{1}{N}\sum_i \delta_{\lambda_i}$ that puts mass $\frac{1}{N}$ at each eigenvalue. The argument below never needs it, and works with the numbers $\frac{1}{N}\operatorname{tr} M^k$ directly. Neither does Lecture 7, where the object is the function m_N and a density falls out of a contour integral.

2.3 Two moments by hand

Averaging $\|Mu\|^2 = u^\top M^2 u$ of Proposition 1.10 over the vectors of an orthonormal basis gives $\frac{1}{N}\operatorname{tr} M^2$, the first moment to compute.

Proposition 2.4. $\mathbb{E}\operatorname{tr} M^2 = N+1$.

Proof. $\operatorname{tr} M^2 = \sum_{i,j} M_{ij}M_{ji}$, and (4) gives $\mathbb{E}[M_{ij}M_{ji}] = \frac{1}{N}(\delta_{ij}\delta_{ji} + \delta_{ii}\delta_{jj}) = \frac{1}{N}(1+\delta_{ij})$, so

$$\mathbb{E}\operatorname{tr} M^2 = \frac{1}{N}\sum_{i,j}\big(1+\delta_{ij}\big) = \frac{1}{N}\big(N^2+N\big) = N+1. \qquad \blacksquare$$

For the fourth moment Theorem 1.2 offers three pair partitions of the four factors.

Proposition 2.5. $\mathbb{E}\operatorname{tr} M^4 = 2N + 5 + \dfrac{5}{N}$.

Proof. Expanding, $\operatorname{tr} M^4 = \sum_{i,j,k,l} M_{ij}M_{jk}M_{kl}M_{li}$, and Theorem 1.2 splits each term into the three pair partitions of the four factors. Take them in turn.

The pairing $\{12\}\{34\}$. By (4), $\mathbb{E}[M_{ij}M_{jk}] = \frac{1}{N}(\delta_{ij}\delta_{jk} + \delta_{ik})$ and $\mathbb{E}[M_{kl}M_{li}] = \frac{1}{N}(\delta_{kl}\delta_{li} + \delta_{ki})$. Multiplying out gives four terms; summing each over i,j,k,l counts the index tuples it allows:

$$\underbrace{\delta_{ij}\delta_{jk}\delta_{kl}\delta_{li}}_{\text{all equal: } N} + \underbrace{\delta_{ij}\delta_{jk}\delta_{ki}}_{i=j=k:\ N^2} + \underbrace{\delta_{ik}\delta_{kl}\delta_{li}}_{i=k=l:\ N^2} + \underbrace{\delta_{ik}}_{i=k:\ N^3},$$

so the contribution is $\frac{1}{N^2}(N^3 + 2N^2 + N) = N + 2 + \frac{1}{N}$.

The pairing $\{14\}\{23\}$. The same computation with $\mathbb{E}[M_{ij}M_{li}] = \frac{1}{N}(\delta_{il}\delta_{ij} + \delta_{jl})$ and $\mathbb{E}[M_{jk}M_{kl}] = \frac{1}{N}(\delta_{jk}\delta_{kl} + \delta_{jl})$ gives $N + 2 + \frac{1}{N}$ again.

The pairing $\{13\}\{24\}$. Here

$$\mathbb{E}[M_{ij}M_{kl}]\,\mathbb{E}[M_{jk}M_{li}] = \tfrac{1}{N^2}\big(\delta_{ik}\delta_{jl} + \delta_{il}\delta_{jk}\big)\big(\delta_{jl}\delta_{ki} + \delta_{ji}\delta_{kl}\big),$$

whose four terms allow N^2, N, N and N tuples respectively — the first is $\delta_{ik}\delta_{jl}$, two constraints, and each of the other three forces all four indices equal. The contribution is $\frac{1}{N^2}(N^2+3N) = 1+\frac{3}{N}$.

Adding, $\mathbb{E}\operatorname{tr} M^4 = 2(N+2+\frac{1}{N}) + 1 + \frac{3}{N} = 2N + 5 + \frac{5}{N}$. $\blacksquare$

Dividing by N, $\frac{1}{N}\mathbb{E}\operatorname{tr} M^2 \to 1$ and $\frac{1}{N}\mathbb{E}\operatorname{tr} M^4 \to 2$. Two of the three pairings contributed at order N and the third only at order 1. The rest of this section explains which pairings win, and why.

Exercise 2.6 (TYPE A)**.** Show that $\mathbb{E}\operatorname{tr} M^{2k+1} = 0$ for every $k \geq 0$, with no computation — say why. Then compute $\mathbb{E}[(\operatorname{tr} M)^2]$.

2.4 Matrix products are sums over paths

For matrices A and B, $(AB)_{ij} = \sum_l A_{il}B_{lj}$: a sum over all ways to go from i to j in two steps, via some l, weighted by the product of the entries used. Iterating gives the same description of any power.

Definition 2.7 (Paths)**.** A *path of length m from i to j* is a sequence of vertices in $\{1,\dots,N\}$,

$$\pi = \big(\pi(0), \pi(1), \dots, \pi(m)\big), \qquad \pi(0) = i, \quad \pi(m) = j,$$

also written $\pi(0) \to \pi(1) \to \cdots \to \pi(m)$. It is *closed* if $\pi(m) = \pi(0)$. Its ath step goes from $\pi(a-1)$ to $\pi(a)$ and *traverses* the edge $\{\pi(a-1), \pi(a)\}$, which is a loop if $\pi(a-1) = \pi(a)$. Its *weight* is

$$W(\pi) = \prod_{a=1}^{m} M_{\pi(a-1)\pi(a)}.$$

The entry on an edge does not depend on the direction of traversal, because $M_{ab} = M_{ba}$. Expanding the products,

$$(M^m)_{ij} = \sum_{\pi:\, i \to j} W(\pi), \qquad \text{so} \qquad \mathbb{E}\,\mathrm{tr}\, M^m = \sum_{i=1}^{N} \sum_{\pi:\, i \to i} \mathbb{E}\, W(\pi),$$

the inner sums running over paths of length m from i to j, respectively from i back to i. We record this as

$$\mathbb{E}\,\mathrm{tr}\, M^m = \sum_{\pi \text{ closed, length } m} \mathbb{E}\, W(\pi). \tag{5}$$

Lemma 2.8. *Let π be a closed path of length m. Then $\mathbb{E}W(\pi) = 0$ unless π traverses every edge, loops included, an even number of times. In every case $|\mathbb{E}W(\pi)| \le c_m N^{-m/2}$, with c_m depending only on m.*

Proof. Group the factors of $W(\pi)$ by edge. Entries on distinct edges are independent, so $\mathbb{E}W(\pi) = \prod_e \mathbb{E}[M_e^{n_e}]$, where n_e is the number of traversals of e and $\sum_e n_e = m$. Odd moments of a centered Gaussian vanish, and $|\mathbb{E}M_e^n| \le (2/N)^{n/2}\mathbb{E}|g|^n$. ■

In particular only closed paths that traverse every edge at least twice contribute to (5).

2.5 The same two moments, by paths

$m = 2$. A closed path of length 2 is $i \to j \to i$: it traverses the edge $\{i, j\}$ twice, out and back. There are $N(N-1)$ such paths with $i \ne j$, each with $\mathbb{E}M_{ij}^2 = \frac{1}{N}$, and N loops, each with $\mathbb{E}M_{ii}^2 = \frac{2}{N}$: again $\mathbb{E}\,\mathrm{tr}\, M^2 = N - 1 + 2 = N + 1$.

$m = 4$. Sort the closed paths of length 4 that traverse every edge at least twice by their shape:

shape	example	vertices	contribution
two edges from one vertex	$i \to j \to i \to l \to i$	3	$N(N-1)(N-2) \cdot N^{-2} = N + O(1)$
a path out and back	$i \to j \to k \to j \to i$	3	$N(N-1)(N-2) \cdot N^{-2} = N + O(1)$
one edge, four times	$i \to j \to i \to j \to i$	2	$N(N-1) \cdot 3N^{-2} = O(1)$
loops, other coincidences		≤ 2	$O(1)$

The first two shapes use three distinct vertices and two distinct edges, each traversed exactly twice, once out and once back; they are the ones of order N. The graph they trace out is a *tree*. The third uses one edge four times and has only two vertices.

From a path to a pair partition. In a path that traverses each edge exactly twice, pair the two steps that traverse the same edge. For $i \to j \to i \to l \to i$ this pairs steps $1, 2$ and steps $3, 4$: the partition $\{12\}\{34\}$. For $i \to j \to k \to j \to i$ it pairs steps $2, 3$ and steps $1, 4$: the partition $\{14\}\{23\}$. These are exactly the two pairings that contributed at order N in Proposition 2.5. The third pairing, $\{13\}\{24\}$, would need step 1 and step 3 on one edge and steps $2, 4$ on another; in a closed path $i \to j \to k \to l \to i$ that forces $i = k$ and $j = l$, the single-edge path, with only two vertices. What distinguishes it is the following.

Definition 2.9 (Crossing)**.** Draw the $2q$ points of a pair partition on a line and join each block by an arc above it. Two blocks $\{a, b\}$ and $\{c, d\}$ *cross* if $a < c < b < d$. A pair partition with no crossing blocks is *non-crossing*; there are $1, 2, 5, 14, \ldots$ of them on $2, 4, 6, 8$ points, against $1, 3, 15, 105$ pair partitions in total.

Remark 2.10 (What a crossing costs). A path of order N had three vertices, a path compatible with the crossing pairing only two. Each vertex is a free index and is worth a factor of N. The crossing pairing loses one vertex, and with it one power of N.

Two warnings about the correspondence. First, a path determines its pairing only at leading order: the single-edge path $i \to j \to i \to j \to i$ traverses its edge four times, so Theorem 1.2 pairs its steps in all three ways, $\mathbb{E}M_{ij}^4 = 3(\mathbb{E}M_{ij}^2)^2$, and no single pairing belongs to it. Second, "crossing" refers only to the arcs of a pair partition; the steps of a path *traverse* edges.

Remark 2.11 (Real versus complex). The two delta terms in (4) come from $M_{ij} = M_{ji}$. A complex Hermitian matrix has $\mathbb{E}[H_{ij}H_{kl}] = \frac{1}{N}\delta_{il}\delta_{jk}$, one term instead of two, and then the crossing partition leaves three constraints rather than two and costs N^{-2}. One extra way of gluing two entries together is the whole difference between the two ensembles.

2.6 Every moment

The odd moments vanish: $\mathbb{E}\,\mathrm{tr}\, M^{2k+1} = 0$, because the weight of a path of odd length is a product of an odd number of centered jointly Gaussian variables. From now on the power is even, $2k$.

Fix a starting vertex. Every closed path starts somewhere, and the starting point can be fixed at no cost.

Lemma 2.12. *For every vertex i,*

$$\frac{1}{N}\mathbb{E}\,\mathrm{tr}\, M^{2k} = \mathbb{E}\big(M^{2k}\big)_{ii} = \sum_{\pi:\, i \to i} \mathbb{E}\, W(\pi).$$

Proof. For a permutation matrix Q, the matrix $QMQ^\top$ has the same law as M (Remark 1.8), and $(QM^{2k}Q^\top)_{ii}$ is another diagonal entry of M^{2k}. So all N diagonal entries of M^{2k} have the same law, and their average $\frac{1}{N}\,\mathrm{tr}\, M^{2k}$ has the same mean as any one of them. ∎

Fixing the starting vertex removes the factor $\frac{1}{N}$, and the whole argument is now a count of vertices. By Lemma 2.8 with $m = 2k$, $\mathbb{E}W(\pi) = 0$ unless π traverses every edge an even number of times, and $|\mathbb{E}W(\pi)| \le c_{2k}N^{-k}$ always.

Lemma 2.13 (A tree is the sparsest connected graph). *Let π be a path of length $2k$ from i to i that traverses every edge at least twice, and let v be its number of distinct vertices, i included. Then $v \le k+1$, with equality if and only if the edges it traverses form a tree with k edges, each traversed exactly twice.*

Proof. Each edge takes at least two of the $2k$ steps, so there are at most k distinct edges. The graph they form is connected, because a single path runs through all of it, and a connected graph on v vertices has at least $v-1$ edges that are not loops, with equality exactly for a tree. So $v \le k+1$, and equality forces k distinct edges, none a loop, forming a tree, each traversed exactly twice. ∎

Call a path as in the equality case a *tree path.* Deleting an edge disconnects a tree, so a path in a tree that returns to its start traverses each edge alternately in each direction: a tree path traverses each of its k edges once out and once back.

Proposition 2.14 (Only tree paths survive)**.** *Let $T_{k,N}$ be the number of tree paths of length $2k$ from i. Then*

$$\frac{1}{N}\mathbb{E}\,\mathrm{tr}\,M^{2k} = N^{-k}\,T_{k,N} + O\Big(\frac{1}{N}\Big).$$

Proof. Sort the paths $\pi : i \to i$ by their number v of distinct vertices. By Lemma 2.8, only paths traversing every edge at least twice contribute, and by Lemma 2.13 these have $v \le k+1$.

Paths with v vertices are at most $c_{k,v}N^{v-1}$ in number. Record which of the times $0, 1, \dots, 2k$ visit the same vertex: there are at most $c_{k,v}$ such patterns, depending only on k and v. Given the pattern, the $v-1$ vertices other than i are chosen from $N-1$.

So the paths with $v \le k$ contribute at most $c_{k,v}N^{v-1} \cdot c_{2k}N^{-k} = O(1/N)$ in total, by Lemma 2.8. The paths with $v = k+1$ are the tree paths. A tree has no loops, so each of the k edges of a tree path joins two distinct vertices, carries an entry of variance $\frac{1}{N}$, and is traversed exactly twice: $\mathbb{E}W(\pi) = N^{-k}$ exactly. ■

It remains to count tree paths, and this is where the non-crossing pair partitions enter. The *pairing* of a tree path pairs the two steps that traverse each edge.

Lemma 2.15. *The pairing of a tree path is non-crossing. Conversely, each non-crossing pair partition of the $2k$ steps is the pairing of exactly $(N-1)(N-2)\cdots(N-k)$ tree paths from i.*

Proof. Suppose the pairing of a tree path had crossing blocks $\{a, b\}$ and $\{c, d\}$, $a < c < b < d$, and let e be the edge traversed at steps a and b. Between steps a and b the path is on the far side of e, and after step b it is on the near side for good, since e is traversed only twice. The edge traversed at step c therefore lies on the far side, yet it is traversed again at step d, from the near side. That is impossible.

For the converse, induct on k. For $k = 1$ the tree paths are $i \to j \to i$ with $j \ne i$, and there are $N-1$ of them. A non-crossing pair partition P has a block of adjacent steps $\{r, r+1\}$: take a block $\{a, b\}$ with $b - a$ smallest, and note that a point strictly inside it would be paired inside it. In a tree path with pairing P, steps r and $r+1$ traverse the same edge out and back, so $\pi(r+1) = \pi(r-1)$, and $\pi(r)$ is visited at no other time: it is a leaf, and it is not i. Deleting those two steps leaves a tree path of length $2k-2$ from i whose pairing is P with that block removed, still non-crossing. Conversely, any such shorter tree path extends by inserting an excursion to a new leaf after its $(r-1)$st step, and the leaf can be any of the $N-k$ vertices it has not yet used. So the count is $(N-1)\cdots(N-k+1)\cdot(N-k)$. ■

Proposition 2.16. *Let C_k be the number of non-crossing pair partitions of $2k$ points. Then*

$$\frac{1}{N}\mathbb{E}\,\mathrm{tr}\,M^{2k} = C_k\Big(1-\frac{1}{N}\Big)\Big(1-\frac{2}{N}\Big)\cdots\Big(1-\frac{k}{N}\Big) + O\Big(\frac{1}{N}\Big) \;\longrightarrow\; C_k.$$

Proof. By Lemma 2.15, $T_{k,N} = C_k(N-1)\cdots(N-k)$. Multiply by N^{-k} and apply Proposition 2.14. ■

Proposition 2.17 (Catalan)**.** *Let C_k be the number of non-crossing pair partitions of $2k$ points, with $C_0 = 1$. Then $C_k = \sum_{j=0}^{k-1} C_jC_{k-1-j}$, so $C_1, \dots, C_5 = 1, 2, 5, 14, 42$. The generating function $C(w) = \sum_k C_kw^k$ satisfies*

$$C(w) = 1 + w\,C(w)^2, \tag{6}$$

so that $C(w) = \frac{1-\sqrt{1-4w}}{2w}$ and $C_k = \frac{1}{k+1}\binom{2k}{k}$.

Proof. Let $\{1, b\}$ be the block containing the point 1. A block joining a point strictly between 1 and b to a point after b would cross $\{1, b\}$, so the points strictly inside pair among themselves and those after b pair among themselves. Hence $b = 2j + 2$ for some j, the inside carries any of C_j partitions and the outside any of C_{k-1-j}, and summing over j gives the recursion. Multiplying by w^k and summing over $k \geq 1$ turns the convolution into a product, $C - 1 = wC^2$. Of the two roots, only $\frac{1-\sqrt{1-4w}}{2w}$ is finite at $w = 0$, and the binomial series for $\sqrt{1-4w}$ gives the closed form. ■

2.7 The fluctuations are small

Proposition 2.18. *For every m,* $\operatorname{Var}\left(\frac{1}{N} \operatorname{tr} M^m\right) = O(1/N)$.

Proof. By (5), $\operatorname{Var}(\operatorname{tr} M^m) = \sum_{\pi,\pi'} \operatorname{Cov}\big(W(\pi), W(\pi')\big)$ over pairs of closed paths of length m. If π and π' traverse no common edge, $W(\pi)$ and $W(\pi')$ are functions of disjoint sets of entries, hence independent, and the covariance is 0. If some edge is traversed an odd number of times by the two paths together, then $\mathbb{E}[W(\pi)W(\pi')] = 0$, and one of $\mathbb{E}W(\pi)$, $\mathbb{E}W(\pi')$ is 0, by the argument of Lemma 2.8. So the surviving pairs share an edge, which makes the union of the two paths connected, and traverse every edge at least twice in total, which gives the union at most m edges. By the argument of Lemma 2.13 it then has at most $m + 1$ vertices. There are at most cN^{m+1} such pairs, each covariance is at most $c'N^{-m}$ by Lemma 2.8, and so $\operatorname{Var}(\operatorname{tr} M^m) = O(N)$. ■

Since $\frac{1}{N}\mathbb{E} \operatorname{tr} M^{2k} \to C_k > 0$, this gives $\operatorname{Var}/(\mathbb{E})^2 \to 0$, which is idea 3. The bound is not sharp — in fact $\operatorname{Var}(\operatorname{tr} M^m) = O(1)$ — but it is all idea 3 needs.

Proof of Theorem 2.1. By Lemma 2.2 and Propositions 2.16 and 2.17, the mean of $\frac{1}{N}\sum_i \lambda_i^m$ tends to C_k if $m = 2k$ and is 0 if m is odd. By Exercise 2.21 these are the moments of ρ_{sc}. By Proposition 2.18 and Chebyshev's inequality, $\frac{1}{N}\sum_i \lambda_i^m$ is within $o(1)$ of its mean with probability tending to one. ■

Remark 2.19 (What was used). Gaussianity entered only through Lemma 2.8. The proof of Proposition 2.14 used only that an edge traversed once gives 0 and that moments are bounded, and a tree path sees nothing about an entry but its variance. The same proof works for any symmetric matrix with independent centered entries, variance $\frac{1}{N}$ off the diagonal, and $\mathbb{E}|M_{ij}|^p \leq c_p N^{-p/2}$. Only the second moment of an entry survives the leading order, which is why the semicircle is universal.

Exercise 2.20 (TYPE A)**.** Draw all 15 pair partitions of 6 points, mark the crossings, and check that exactly 5 are non-crossing. For each of those five, draw a tree path with that pairing and check that it has 4 vertices.

Exercise 2.21 (TYPE B)**.** Show that ρ_{sc} is a probability density and that its odd moments vanish. Then check by direct integration that

$$\int_{-2}^{2} x^2 \,\rho_{sc}(x)\,dx = 1, \qquad \int_{-2}^{2} x^4 \,\rho_{sc}(x)\,dx = 2,$$

which are C_1 and C_2. *Substitute $x = 2\sin\theta$. The general statement $\int x^{2k}\rho_{sc} = C_k$ is true and is what Theorem 2.1 needs; you are checking the first two cases.*

ASSIGNED **Lecture 6**, due at the start of Lecture 7: Exercises 2.6, 2.20, 2.21. *Three problems. Do 2.20 with a pen and paper, not a computer — the picture is the point.*

3 The Stieltjes transform

Q. The counting proof produced a list of numbers, and then recognized them. Where does the *density* $\frac{1}{2\pi}\sqrt{4-x^2}$ come from?

From a function of one complex variable. The Stieltjes transform of M encodes the whole spectrum, a contour integral reads eigenvalue counts back off it, and the density is the imaginary part of its limit on the real axis. This lecture sets that up; the next one finds the limit.

The counting proof left four things undone.

(i) It never produced a **density**. It produced a sequence of numbers that we recognized; perturb the model and there is nothing to recognize.

(ii) It requires **all moments finite**.

(iii) It gives **no local information** — nothing about one eigenvalue, a gap, or the edge.

(iv) It does not survive **deformation**: unequal variances, or a rank-one spike, and the combinatorics does not adapt.

All four are repaired by replacing the polynomials x^k with a different family of test functions.

3.1 The Stieltjes transform of a matrix

Definition 3.1 (Resolvent and Stieltjes transform)**.** Let M be a symmetric $N \times N$ matrix with eigenvalues $\lambda_1, \dots, \lambda_N$. For $z \in \mathbb{C}$ not an eigenvalue, the *resolvent* is the matrix $(M - zI)^{-1}$, and the *Stieltjes transform* of M is

$$m_N(z) \;=\; \frac{1}{N}\operatorname{tr}\big(M - zI\big)^{-1} \;=\; \frac{1}{N}\sum_{i=1}^{N}\frac{1}{\lambda_i - z}.$$

The second equality is the trace lemma again: diagonalizing $M = O\Lambda O^\top$ gives $(M - zI)^{-1} = O(\Lambda - zI)^{-1}O^\top$, whose eigenvalues are the numbers $\frac{1}{\lambda_i - z}$.

Two facts will be used constantly. First, for z off the real axis the resolvent exists, because M is symmetric and so has real eigenvalues, and

$$\big\|(M - zI)^{-1}\big\| = \max_i \frac{1}{|\lambda_i - z|} \le \frac{1}{|\Im z|}, \qquad \text{hence} \qquad \big|m_N(z)\big| \le \frac{1}{|\Im z|}. \tag{7}$$

Second, m_N is analytic on $\mathbb{C}$ minus the spectrum, being a rational function of z. Third, for $\Im z > 0$ every term $\Im\frac{1}{\lambda_i - z} = \frac{\Im z}{|\lambda_i - z|^2}$ is positive, so $\Im m_N(z) > 0$.

When M is random, so is m_N: it is the transform of the matrix in hand. Its mean $\mathbb{E}\, m_N(z)$ is a deterministic function of z, also with positive imaginary part and modulus at most $1/|\Im z|$, and the expectation will always be written out, so that m_N never means anything but the random quantity. Two things will be shown about it. At each fixed z off the real axis, $m_N(z)$ is close to $\mathbb{E}\, m_N(z)$ (§4.3), with an error that blows up as $\Im z \to 0$ — at fixed N the random transform becomes a comb of spikes as z approaches the axis while its mean stays smooth, so the closeness is genuinely a fixed-$\Im z$ statement. And $\mathbb{E}\, m_N(z)$ converges as $N \to \infty$ (§4.4).

Remark 3.2. Ridge regression in §2.5 of Lectures 1–2 is $(H^\top H + \lambda I)^{-1} H^\top y$, and the matrix inverted there is a resolvent at $z = -\lambda$.

3.2 Counting eigenvalues with the Cauchy integral formula

Nothing is lost in passing from M to m_N: the eigenvalues can be read back off. The question we actually want to answer is a counting question — *how many eigenvalues lie in* (a, b)*?* — and the Cauchy integral formula answers it directly.

Recall. If f is analytic inside and on a positively oriented simple closed contour Γ except for finitely many poles, none on Γ, then $\frac{1}{2\pi i}\oint_\Gamma f(z)\,dz$ is the sum of the residues at the poles inside. The Cauchy integral formula is the case $f(z) = \frac{g(z)}{z-w}$ with g analytic, whose residue at w is $g(w)$.

Proposition 3.3 (The eigenvalues are the poles)**.** *m_N has a simple pole at each distinct eigenvalue λ of M, with residue $-k/N$, where k is the multiplicity of λ; it is analytic elsewhere. Consequently, for every positively oriented simple closed contour Γ with no eigenvalue on it,*

$$-\frac{N}{2\pi i}\oint_\Gamma m_N(z)\,dz \;=\; \#\{i : \lambda_i \text{ inside } \Gamma\},$$

counted with multiplicity.

Proof. Group equal eigenvalues: $m_N(z) = \frac{1}{N}\sum_i \frac{1}{\lambda_i - z} = \frac{1}{N}\sum_\lambda \frac{-k_\lambda}{z-\lambda}$, a finite sum of simple poles with residues $-k_\lambda/N$. Apply the residue theorem and multiply by N. ■

So the spectrum and the transform carry the same information: the eigenvalues are the poles, the multiplicities are the residues, and integrating m_N around a region *counts* eigenvalues in it.

Proposition 3.3 allows any contour. Take the one adapted to the question and let it collapse onto the real axis. What comes out is the inversion formula for the Stieltjes transform, in the form we want: an eigenvalue count.

Theorem 3.4 (Inversion, as a count)**.** *Let M be symmetric with eigenvalues $\lambda_1, \dots, \lambda_N$, let $a < b$ be real numbers that are not eigenvalues of M, and let $\varepsilon > 0$. Then*

$$\#\{i : \lambda_i \in (a,b)\} = \frac{N}{\pi}\int_a^b \Im\, m_N(x + i\varepsilon)\,dx + \frac{1}{\pi}\sum_{i=1}^N \Big(\arctan\frac{\varepsilon}{\lambda_i - a} - \arctan\frac{\varepsilon}{\lambda_i - b}\Big).$$

Consequently, writing $d_i = \min(|\lambda_i - a|, |\lambda_i - b|)$ for the distance from λ_i to the endpoints,

$$\left|\; \#\{i : \lambda_i \in (a,b)\} - \frac{N}{\pi}\int_a^b \Im\, m_N(x+i\varepsilon)\,dx \;\right| \;\le\; \sum_{i=1}^N \min\Big(1,\; \frac{2\varepsilon}{\pi d_i}\Big).$$

Proof. Let Γ_ε be the boundary of the rectangle $[a,b]\times[-\varepsilon,\varepsilon]$, positively oriented. No eigenvalue lies on it: the eigenvalues are real, and the only real points of Γ_ε are a and b. The eigenvalues inside are exactly those in (a,b), so Proposition 3.3 gives

$$\#\{i : \lambda_i \in (a,b)\} = -\frac{N}{2\pi i}\oint_{\Gamma_\varepsilon} m_N(z)\,dz.$$

We compute all four edges exactly. Because the λ_i are real,

$$\overline{m_N(z)} = \frac{1}{N}\sum_i \overline{\Big(\frac{1}{\lambda_i - z}\Big)} = \frac{1}{N}\sum_i \frac{1}{\lambda_i - \bar z} = m_N(\bar z),$$

so the bottom edge, traversed left to right at height $-\varepsilon$, contributes $\int_a^b \overline{m_N(x+i\varepsilon)}\,dx$, while the top edge, traversed right to left at height $+\varepsilon$, contributes $-\int_a^b m_N(x+i\varepsilon)\,dx$. Their sum is

$$\int_a^b \Big(\overline{m_N(x+i\varepsilon)} - m_N(x+i\varepsilon)\Big)dx = -2i\int_a^b \Im\, m_N(x+i\varepsilon)\,dx,$$

and multiplying by $-\frac{N}{2\pi i}$ turns this into $\frac{N}{\pi}\int_a^b \Im\, m_N(x+i\varepsilon)\,dx$. *The horizontal edges of the contour are the inversion formula.*

On the left edge $z = a + iy$ with y running from ε down to $-\varepsilon$ and $dz = i\,dy$, so it contributes $-i\int_{-\varepsilon}^{\varepsilon} m_N(a+iy)\,dy$. For a single term of m_N put $u = \lambda_i - a \neq 0$; then $\frac{1}{u-iy} = \frac{u+iy}{u^2+y^2}$, the iy part is odd in y and integrates to zero, and

$$\int_{-\varepsilon}^{\varepsilon} \frac{u\,dy}{u^2+y^2} = 2\arctan\frac{\varepsilon}{u}.$$

So the left edge contributes $-\frac{2i}{N}\sum_i \arctan\frac{\varepsilon}{\lambda_i - a}$, and the right edge, traversed upward, contributes $+\frac{2i}{N}\sum_i \arctan\frac{\varepsilon}{\lambda_i - b}$ by the same computation. Multiplying by $-\frac{N}{2\pi i}$ gives the two arctangent sums in the statement.

For the bound, $|\arctan t| \le \min(\frac{\pi}{2}, |t|)$, so the ith term of the correction is at most

$$\frac{1}{\pi}\Big[\min\Big(\frac{\pi}{2}, \frac{\varepsilon}{|\lambda_i - a|}\Big) + \min\Big(\frac{\pi}{2}, \frac{\varepsilon}{|\lambda_i - b|}\Big)\Big] \le \min\Big(1, \frac{2\varepsilon}{\pi d_i}\Big).$$

■

The correction is a sum over eigenvalues in which those far from the endpoints contribute $O(\varepsilon/d_i)$ and those within ε of an endpoint contribute $O(1)$. Divide by N and split the sum at a threshold η:

Corollary 3.5 (The proportion)**.** *For every $\varepsilon > 0$ and $\eta > 0$,*

$$\left|\ \frac{1}{N}\#\{i : \lambda_i \in (a,b)\} - \frac{1}{\pi}\int_a^b \Im\, m_N(x+i\varepsilon)\,dx\ \right| \ \le\ \frac{2\varepsilon}{\pi\eta} + \frac{\#\{i : d_i \le \eta\}}{N}.$$

Proof. In the sum of Theorem 3.4, the terms with $d_i > \eta$ are each at most $\frac{2\varepsilon}{\pi\eta}$ and there are at most N of them; the terms with $d_i \le \eta$ are each at most 1. ■

The second term is the one to worry about: it is a statement about where the eigenvalues are, which is what we are trying to find out. The way around this is that the transform controls its own local counts.

Lemma 3.6 (The transform bounds local counts)**.** *For every real a and every $\eta > 0$,*

$$\frac{1}{N}\#\{i : |\lambda_i - a| \le \eta\}\ \le\ 2\eta\,\Im\, m_N(a+i\eta).$$

Proof. $\Im m_N(a+i\eta) = \frac{1}{N}\sum_i \frac{\eta}{(\lambda_i - a)^2 + \eta^2}$, every term is positive, and each term with $|\lambda_i - a| \le \eta$ is at least $\frac{1}{N}\cdot\frac{\eta}{2\eta^2} = \frac{1}{2\eta N}$. ■

So the proportion of eigenvalues within η of a is at most $2\eta\,\Im m_N(a+i\eta)$, and m_N is the very quantity the rest of the lecture computes. Nothing about a density needs to be assumed; it will be read off the limit.

Proposition 3.7 (From the transform to the proportion)**.** *Suppose there is a function m on the upper half plane with $\Im m \le K$ such that, for every $\varepsilon > 0$ and every $L < \infty$,*

$$\sup_{|x|\le L} \big|\mathbb{E}\, m_N(x+i\varepsilon) - m(x+i\varepsilon)\big| \;\longrightarrow\; 0 \qquad \text{and} \qquad \sup_{|x|\le L} \mathbb{E}\big|m_N(x+i\varepsilon) - \mathbb{E}\, m_N(x+i\varepsilon)\big| \;\longrightarrow\; 0$$

as $N \to \infty$. Then for all real $a < b$ the limit

$$\ell(a,b) = \lim_{\varepsilon\downarrow 0} \frac{1}{\pi}\int_a^b \Im\, m(x+i\varepsilon)\,dx$$

exists, and

$$\mathbb{E}\left|\;\frac{1}{N}\#\{i : \lambda_i \in (a,b)\} - \ell(a,b)\;\right| \;\longrightarrow\; 0.$$

Proof. Fix $\varepsilon, \eta > 0$ and write $p_N = \frac{1}{N}\#\{i : \lambda_i \in (a,b)\}$. Corollary 3.5 together with Lemma 3.6 at a and at b gives, for every realization of M,

$$\left|\,p_N - \frac{1}{\pi}\int_a^b \Im\, m_N(x+i\varepsilon)\,dx\,\right| \;\le\; \frac{2\varepsilon}{\pi\eta} + 2\eta\Big(\Im\, m_N(a+i\eta) + \Im\, m_N(b+i\eta)\Big).$$

Inside the integral replace m_N by m. Since $|\Im m_N - \Im m| \le |m_N - \mathbb{E} m_N| + |\mathbb{E} m_N - m|$, writing

$$r_N(\varepsilon) = \sup_{x\in[a,b]} \mathbb{E}\big|m_N(x+i\varepsilon) - \mathbb{E}\, m_N(x+i\varepsilon)\big| + \sup_{x\in[a,b]} \big|\mathbb{E}\, m_N(x+i\varepsilon) - m(x+i\varepsilon)\big|$$

we have $\mathbb{E}\big|\frac{1}{\pi}\int_a^b (\Im m_N - \Im m)(x+i\varepsilon)\,dx\big| \le \frac{b-a}{\pi}\, r_N(\varepsilon)$. Take expectations in the first display:

$$\mathbb{E}\left|\,p_N - \frac{1}{\pi}\int_a^b \Im\, m(x+i\varepsilon)\,dx\,\right| \le \frac{2\varepsilon}{\pi\eta} + 2\eta\Big(\Im\,\mathbb{E} m_N(a+i\eta) + \Im\,\mathbb{E} m_N(b+i\eta)\Big) + \frac{b-a}{\pi}\, r_N(\varepsilon).$$

Let $N \to \infty$. Then $r_N(\varepsilon) \to 0$ by hypothesis, and $\Im\,\mathbb{E} m_N(a+i\eta) \to \Im m(a+i\eta) \le K$, likewise at b. Hence

$$\limsup_{N\to\infty}\; \mathbb{E}\left|\,p_N - \frac{1}{\pi}\int_a^b \Im\, m(x+i\varepsilon)\,dx\,\right| \;\le\; \frac{2\varepsilon}{\pi\eta} + 4K\eta \;=\; \Big(\frac{2}{\pi} + 4K\Big)\sqrt{\varepsilon}$$

on choosing $\eta = \sqrt{\varepsilon}$. Applying this with two values $\varepsilon, \varepsilon'$ and the triangle inequality through the same p_N shows that $\frac{1}{\pi}\int_a^b \Im m(x+i\varepsilon)\,dx$ and the same at ε' differ by at most $(\frac{2}{\pi} + 4K)(\sqrt{\varepsilon} + \sqrt{\varepsilon'})$; so the limit $\ell(a,b)$ exists. Then $\limsup_N \mathbb{E}|p_N - \ell| \le (\frac{2}{\pi} + 4K)\sqrt{\varepsilon} + \big|\frac{1}{\pi}\int_a^b \Im m(x+i\varepsilon)\,dx - \ell\big|$ for every $\varepsilon > 0$, and the right side tends to 0 as $\varepsilon \downarrow 0$. ■

Remark 3.8 (The order of the limits). In the proof, $N \to \infty$ is taken at fixed ε, and only afterwards does $\varepsilon \downarrow 0$. That order is forced: the error in Corollary 3.5 is small uniformly in N, but only once $\varepsilon \ll \eta$ and η is small, and neither has anything to do with N. The reverse order asks a different question. At fixed N, Theorem 3.4 says the error in the *count* is $\sum_i \min(1, \frac{2\varepsilon}{\pi d_i})$, which with N eigenvalues at spacing of order $\frac{1}{N}$ is of order $\varepsilon N \log N$; so counting exactly needs ε below the spacing. That is the local regime, and Exercise 3.12 shows what $\Im m_N$ looks like there: a comb of spikes, one per eigenvalue. Exercise 3.11 is a picture of both regimes, with the crossover at $\varepsilon \sim 1/N$.

Theorem 3.4 is the only place a density will come from, and it is now proved rather than quoted. It also tells us exactly what to compute. Lecture 8 computes

$$\boxed{\rho(x) \;=\; \lim_{\varepsilon\downarrow 0}\ \lim_{N\to\infty}\ \frac{1}{\pi}\,\Im\,\mathbb{E}\,m_N(x+i\varepsilon).} \tag{8}$$

Proposition 3.7 is what entitles the answer to be called a density: once $\mathbb{E}m_N \to m$ is known, the proportion of eigenvalues in any interval converges to the integral of the right-hand side of (8) over that interval, with no further assumption. Two comments on (8). The expectation is harmless: at fixed ε, m_N is within $O(1/N)$ of its mean in L^1, which is §4.3, so the box could equally be written with m_N and the inner limit taken in L^1. And the order of the limits is not a technicality, it is the content: taking $\varepsilon \downarrow 0$ first, at fixed N, returns the N individual eigenvalues and no density at all.

3.3 The transform at infinity: a generating function for traces

The other end of the picture is $z \to \infty$. For $|z| > \max_i |\lambda_i|$ the Neumann series $(M - zI)^{-1} = -\sum_{k\ge 0} M^k/z^{k+1}$ converges, and taking traces gives

$$m_N(z) = -\sum_{k\ge 0} \frac{1}{z^{k+1}}\cdot\frac{1}{N}\operatorname{tr} M^k, \qquad \text{that is} \qquad -\,z\,m_N(z) = \sum_{k\ge 0}\Big(\frac{1}{N}\operatorname{tr} M^k\Big)\frac{1}{z^k}. \tag{9}$$

So near infinity the normalized transform is the generating function of the normalized trace moments in the variable $1/z$: it carries exactly the information §2 computed, one coefficient at a time. The $k = 0$ term gives

$$m_N(z) = -\frac{1}{z} + O\big(|z|^{-2}\big),$$

which matches the behavior of the limit found in §4.4, Lemma 4.9(b), and is where Remark 4.16 starts.

By §2.4, $\operatorname{tr} M^k$ is a sum over closed paths of length k, so:

Remark 3.9 (What the resolvent is). $-z\,m_N(z)$ is the generating function of closed paths in M, counted by length and divided by N, with $1/z$ marking a step. The two proofs of Theorem 2.1 are two ways of reading one function: §2 computes its Taylor coefficients at $z = \infty$ one at a time, by counting paths; §4.2 finds an equation for the whole function without computing any coefficient. For the semicircle the coefficients are the Catalan numbers, $m(z) = -\sum_{k\ge 0} C_k z^{-(2k+1)}$, the odd moments vanishing by symmetry.

Exercise 3.10 (TYPE A)**.** Compute $m_N(z) = \frac{1}{N}\sum_{i=1}^N \frac{1}{\lambda_i - z}$ in closed form in two cases: (a) $M = 0$; (b) M with $N/2$ eigenvalues equal to -1 and $N/2$ equal to $+1$. In each case verify the bound (7), and check that $m_N(z) = -\frac{1}{z} + O(|z|^{-2})$ as $z \to \infty$.

Exercise 3.11 (TYPE A)**.** Numerically, for a GOE matrix at $N = 2000$, plot $\frac{1}{\pi}\Im\,m_N(\lambda+i\varepsilon)$ against $\lambda \in [-3, 3]$ for $\varepsilon = 1$, 0.1 and 0.001, with ρ_{sc} overlaid. *At* $\varepsilon = 1$ *the curve is a smoothed semicircle; at* $\varepsilon = 0.001$ *it is a forest of spikes at the individual eigenvalues. Somewhere in between is the global law, and the crossover is at* $\varepsilon \sim 1/N$*.*

Exercise 3.12 (TYPE B)**.** Each eigenvalue is visible in the transform, and so is its multiplicity. Let M be symmetric with eigenvalues $\lambda_1, \dots, \lambda_N$. Show that for $\lambda \in \mathbb{R}$ and $\varepsilon > 0$,

$$\Im\,m_N(\lambda + i\varepsilon) = \frac{1}{N}\sum_{i=1}^N \frac{\varepsilon}{(\lambda_i - \lambda)^2 + \varepsilon^2} \;>\; 0,$$

so $\Im m_N > 0$ on the upper half plane, and deduce that for every $\lambda \in \mathbb{R}$

$$\lim_{\varepsilon \downarrow 0} \varepsilon \Im m_N(\lambda + i\varepsilon) = \frac{1}{N} \#\{i : \lambda_i = \lambda\}.$$

Each eigenvalue contributes a bump of height $\frac{1}{N\varepsilon}$ and width ε to $\Im m_N$. Use this to explain the picture in Exercise 3.11: why the bumps must overlap before $\frac{1}{\pi}\Im m_N$ can look like a curve, and why that forces $\varepsilon \gg 1/N$.

ASSIGNED **Lecture 7**, due at the start of Lecture 8: Exercises 3.10, 3.11, 3.12. *Three problems, one of them numerical. Nothing here needs Lecture 8.*

4 The semicircle law by the Stieltjes transform

Q. How do we find the limit of $\mathbb{E}m_N(z)$ without computing a single moment?

From a fixed-point equation. Deleting one row and column of M expresses one diagonal entry of the resolvent through the resolvent of the smaller matrix; averaging turns that into an equation the transform must satisfy, up to an error that concentration kills. The equation is stable, so $\mathbb{E}m_N$ converges to its solution, and Remark 4.11 reads the semicircle off the solution.

Throughout, M is a Gaussian Wigner matrix, z is fixed with $\Im z = \varepsilon > 0$, and m_N, $\mathbb{E}m_N$ are as in Definition 3.1. The lecture has four parts: a piece of linear algebra (§4.1), the derivation of the equation with explicit error bounds (§4.2), the one probabilistic estimate that derivation needs (§4.3), and the solution of the equation together with the theorem it proves (§4.4). By the end, Proposition 3.7 has everything it asks for.

4.1 The Schur complement

The mechanism of the whole lecture is that one diagonal entry of a resolvent can be written in terms of the resolvent of a smaller matrix. That is linear algebra, and it holds for any block matrix.

Lemma 4.1 (Schur complement)**.** *Let*

$$A = \begin{pmatrix} P & Q \\ R & S \end{pmatrix}$$

be a square matrix in block form with S square and invertible, and put $T = P - QS^{-1}R$, the Schur complement *of S in A. Then A is invertible if and only if T is, and in that case*

$$A^{-1} = \begin{pmatrix} T^{-1} & -T^{-1}QS^{-1} \\ -S^{-1}RT^{-1} & S^{-1} + S^{-1}RT^{-1}QS^{-1} \end{pmatrix}.$$

In particular the upper-left block of A^{-1} is T^{-1}.

Proof. Eliminate Q by a block row operation:

$$\begin{pmatrix} I & -QS^{-1} \\ 0 & I \end{pmatrix} \begin{pmatrix} P & Q \\ R & S \end{pmatrix} = \begin{pmatrix} T & 0 \\ R & S \end{pmatrix} =: L.$$

The left factor U is invertible, with $U^{-1} = \begin{pmatrix} I & QS^{-1} \\ 0 & I \end{pmatrix}$, so A is invertible exactly when L is, and L is block lower triangular with diagonal blocks T and S, hence invertible exactly when T is. Then $A^{-1} = L^{-1}U$, and

$$L^{-1} = \begin{pmatrix} T^{-1} & 0 \\ -S^{-1}RT^{-1} & S^{-1} \end{pmatrix}, \qquad L^{-1}U = \begin{pmatrix} T^{-1} & -T^{-1}QS^{-1} \\ -S^{-1}RT^{-1} & S^{-1} + S^{-1}RT^{-1}QS^{-1} \end{pmatrix},$$

which is the stated inverse. ∎

Taking P to be the single entry in position $(1,1)$ makes T a scalar, and the lemma reads as follows.

Corollary 4.2. *Let $A = \begin{pmatrix} \alpha & b^\top \\ b & D \end{pmatrix}$ with α scalar, b a column vector and D invertible. If $\alpha - b^\top D^{-1} b \neq 0$ then A is invertible,*

$$g := \left(A^{-1}\right)_{11} = \frac{1}{\alpha - b^\top D^{-1} b}, \qquad \textit{and} \qquad \operatorname{tr} A^{-1} = \operatorname{tr} D^{-1} + g\left(1 + b^\top D^{-2} b\right).$$

Proof. The first formula is the upper-left block. The lower-right block is $D^{-1} + g\, D^{-1}bb^\top D^{-1}$, whose trace is $\operatorname{tr} D^{-1} + g\, b^\top D^{-2} b$; add the corner entry g. ∎

Apply this to $A = M - zI$. Write $M^{(1)}$ for M with its first row and column deleted, and $a = (M_{21}, \dots, M_{N1})^\top \in \mathbb{R}^{N-1}$ for the rest of the first column. Then $\alpha = M_{11} - z$, $b = a$, $D = M^{(1)} - zI$, and

$$\left(\left(M - zI\right)^{-1}\right)_{11} = \frac{1}{M_{11} - z - a^\top \left(M^{(1)} - zI\right)^{-1} a}. \tag{10}$$

Remark 4.3. This is the same identity as §2.4 of Lectures 1–2, where the Schur complement gave the conditional variance of a Gaussian vector. There the small matrix was a covariance; here it is a resolvent.

4.2 The self-consistent equation

Fix z with $\Im z > 0$ and write $R = (M^{(1)} - zI)^{-1}$. Write $\varepsilon = \Im z$. The argument has four steps, of which only the second is probabilistic, and it ends with a quantitative statement.

Step 0: one entry is enough. For a permutation matrix Q, $QMQ^\top$ has the same law as M (Remark 1.8) and $(Q(M - zI)^{-1}Q^\top)_{11}$ is another diagonal entry of the resolvent, so all N diagonal entries have the same law and

$$\mathbb{E}\, m_N(z) = \mathbb{E}\left[\left((M - zI)^{-1}\right)_{11}\right]. \tag{11}$$

Step 1: Schur. By (10), that entry is $\left(M_{11} - z - a^\top Ra\right)^{-1}$. Note first that the denominator can never be small: $a^\top Ra = \sum_k \frac{\langle a, u_k\rangle^2}{\mu_k - z}$, where μ_k and u_k are the eigenvalues and eigenvectors of $M^{(1)}$, so $\Im(a^\top Ra) > 0$, and therefore

$$\left|M_{11} - z - a^\top Ra\right| \;\geq\; \Im z. \tag{12}$$

Everything below is a perturbation of that denominator, and (12) converts it into a perturbation of the entry: if $|w|, |w'| \geq \Im z$ then $|w^{-1} - w'^{-1}| \leq |w - w'|/(\Im z)^2$.

Step 2: the quadratic form concentrates. The vector a consists of the entries $M_{21}, \dots, M_{N1}$, and $M^{(1)}$ consists of entries M_{ij} with $i, j \geq 2$. These are disjoint sets of entries of M, so a is independent of $M^{(1)}$, and hence of R. Also $a \sim \mathcal{N}(0, \frac{1}{N} I_{N-1})$.

Lemma 4.4. *Let R be a (possibly complex) symmetric $(N-1)\times(N-1)$ matrix and $a \sim \mathcal{N}(0, \frac{1}{N} I_{N-1})$. Then*

$$\mathbb{E}\big[a^\top R a\big] = \frac{1}{N} \operatorname{tr} R, \qquad \mathbb{E}\Big|a^\top R a - \frac{1}{N}\operatorname{tr} R\Big|^2 = \frac{2}{N^2}\sum_{r,s} |R_{rs}|^2.$$

In particular, if $R = (M^{(1)} - zI)^{-1}$ then the right-hand side is at most $\dfrac{2}{N(\Im z)^2}$.

Proof. $\mathbb{E}[a_r a_s] = \frac{\delta_{rs}}{N}$ gives the mean. For the second moment, $a^\top R a - \frac{1}{N}\operatorname{tr} R = \sum_{r,s} R_{rs}\big(a_r a_s - \frac{\delta_{rs}}{N}\big)$, so

$$\mathbb{E}\Big|a^\top R a - \tfrac{1}{N}\operatorname{tr} R\Big|^2 = \sum_{r,s,t,u} R_{rs}\overline{R_{tu}}\,\Big(\mathbb{E}[a_r a_s a_t a_u] - \frac{\delta_{rs}\delta_{tu}}{N^2}\Big).$$

By Theorem 1.2, $\mathbb{E}[a_r a_s a_t a_u] = \frac{1}{N^2}(\delta_{rs}\delta_{tu} + \delta_{rt}\delta_{su} + \delta_{ru}\delta_{st})$, so the bracket is $\frac{1}{N^2}(\delta_{rt}\delta_{su} + \delta_{ru}\delta_{st})$ and the sum is

$$\frac{1}{N^2}\Big(\sum_{r,s} R_{rs}\overline{R_{rs}} + \sum_{r,s} R_{rs}\overline{R_{sr}}\Big) = \frac{2}{N^2}\sum_{r,s}|R_{rs}|^2,$$

using $R_{sr} = R_{rs}$. Finally $\sum_{r,s}|R_{rs}|^2 = \operatorname{tr}(R\bar{R}^\top)$ is the sum of the squared singular values of R, each at most $\|R\| \leq 1/\Im z$ by (7), and there are $N-1$ of them. ■

So $a^\top R a = \frac{1}{N}\operatorname{tr} R + \varepsilon_N$ with $\mathbb{E}|\varepsilon_N|^2 \leq \frac{2}{N(\Im z)^2}$: the quadratic form is its own average, to $O(N^{-1/2})$ in L^2. *This is the only place probability enters.*

Step 3: the smaller trace is the same trace. What appears in Step 1 is $\frac{1}{N}\operatorname{tr} R$, a resolvent trace for the matrix one dimension down; what we want is $m_N = \frac{1}{N}\operatorname{tr}(M - zI)^{-1}$. They differ by little, and the Schur complement says exactly how. By Corollary 4.2 with $g = ((M - zI)^{-1})_{11}$,

$$\operatorname{tr}\big(M - zI\big)^{-1} - \operatorname{tr} R = g\,\big(1 + a^\top R^2 a\big).$$

Two facts bound the right side. First, the imaginary part of the denominator in Step 1 is $-\varepsilon - \Im(a^\top R a)$, so $|g| \leq \big(\varepsilon + \Im(a^\top R a)\big)^{-1}$. Second, $R - \bar{R} = R\big((M^{(1)} - \bar{z}I) - (M^{(1)} - zI)\big)\bar{R} = 2i\varepsilon\, R\bar{R}$, so $\Im(a^\top R a) = \varepsilon\, a^\top R\bar{R} a = \varepsilon\|Ra\|^2$, because R is symmetric and a is real; while $a^\top R^2 a = (Ra)^\top(Ra)$ has modulus at most $\|Ra\|^2$. Writing $t = \Im(a^\top R a) \geq 0$,

$$|g|\,\big|1 + a^\top R^2 a\big| \;\leq\; \frac{1 + t/\varepsilon}{\varepsilon + t} \;=\; \frac{1}{\varepsilon},$$

that is,

$$\Big|\operatorname{tr}\big(M - zI\big)^{-1} - \operatorname{tr}\big(M^{(1)} - zI\big)^{-1}\Big| \;\leq\; \frac{1}{\varepsilon}. \tag{13}$$

Deleting a row and column moves the trace of the resolvent by at most one term's worth. Dividing by N, $|m_N - \frac{1}{N}\operatorname{tr} R| \leq \frac{1}{N\varepsilon}$.

Step 4: assemble. Put everything in the Step 1 denominator that is not $-z - m_N$ into one remainder:

$$M_{11} - z - a^\top R a = -z - m_N + \Delta, \qquad \Delta = M_{11} - \Big(a^\top R a - \frac{1}{N}\operatorname{tr} R\Big) + \Big(m_N - \frac{1}{N}\operatorname{tr} R\Big).$$

By $\mathbb{E}M_{11}^2 = \frac{2}{N}$, Lemma 4.4 and (13), using $\mathbb{E}|X| \le (\mathbb{E}X^2)^{1/2}$ on the first two terms,

$$\mathbb{E}|\Delta| \;\le\; \sqrt{\frac{2}{N}} + \frac{\sqrt{2}}{\varepsilon\sqrt{N}} + \frac{1}{N\varepsilon}.$$

Both $-z - m_N + \Delta$ (by (12)) and $-z - m_N$ (because $\Im m_N > 0$) have modulus at least ε, so by the remark after (12)

$$\Big|\big((M - zI)^{-1}\big)_{11} - \frac{1}{-z - m_N}\Big| \;\le\; \frac{|\Delta|}{\varepsilon^2}.$$

The same remark replaces m_N by its mean, which also has $\Im \mathbb{E}m_N > 0$: $\big|\frac{1}{-z-m_N} - \frac{1}{-z-\mathbb{E}m_N}\big| \le \frac{|m_N - \mathbb{E}m_N|}{\varepsilon^2}$. Now take expectations, using (11) on the left:

$$\Big|\,\mathbb{E}m_N(z) - \frac{1}{-z - \mathbb{E}m_N(z)}\,\Big| \;\le\; \frac{\mathbb{E}|\Delta| + \mathbb{E}|m_N - \mathbb{E}m_N|}{\varepsilon^2} \;=:\; \delta_N(\varepsilon). \tag{14}$$

One term is not yet controlled, $\mathbb{E}|m_N - \mathbb{E}m_N|$: the random transform must be close to its own mean. That is §4.3, which gives $\mathbb{E}|m_N - \mathbb{E}m_N| \le \frac{3}{N\varepsilon^2}$. Altogether, for $0 < \varepsilon \le 1$ and every z with $\Im z = \varepsilon$,

$$\delta_N(\varepsilon) \;=\; \frac{1}{\varepsilon^2}\Big[\sqrt{\frac{2}{N}} + \frac{\sqrt{2}}{\varepsilon\sqrt{N}} + \frac{1}{N\varepsilon} + \frac{3}{N\varepsilon^2}\Big] \;\le\; \frac{7}{\varepsilon^4\sqrt{N}}, \tag{15}$$

which depends on z only through ε. So $\mathbb{E}m_N$ satisfies

$$\boxed{m(z) = \frac{1}{-z - m(z)}, \qquad \text{that is} \qquad m^2 + zm + 1 = 0,} \tag{16}$$

up to an error of order $N^{-1/2}$. Section 4.4 shows the equation is stable, so that $\mathbb{E}m_N$ converges to its solution.

Equation (16) is a fixed-point equation for the answer, derived without knowing the answer. The moment method had to be told what to recognize.

4.3 The trace concentrates

The estimate Step 4 still owes us is that m_N is close to its mean $\mathbb{E}m_N$. It is the last missing piece, and it is worth doing properly, because the size of the answer is surprising.

Here is the shape of the problem. $Nm_N(z) = \sum_i \frac{1}{\lambda_i - z}$ is a sum of N terms, each of modulus at most $1/\Im z$. If the λ_i were independent, such a sum would have variance of order N, and m_N would fluctuate at the scale $N^{-1/2}$. We will find that the variance of the whole sum is $O(1)$: the eigenvalues are rigid enough that N of them together fluctuate no more than one of them does.

Lemma 4.5 (Gaussian Poincaré inequality)**.** *Let $X_1, \dots, X_n$ be independent, $X_i \sim \mathcal{N}(0, \sigma_i^2)$, and let $f : \mathbb{R}^n \to \mathbb{R}$ be continuously differentiable with bounded gradient. Then*

$$\operatorname{Var} f(X) \;\le\; \frac{\pi^2}{8} \sum_{i=1}^{n} \sigma_i^2 \, \mathbb{E}\Big[\big(\partial_i f(X)\big)^2\Big].$$

Proof. Replacing $f(x)$ by $f(\sigma_1 x_1, \dots, \sigma_n x_n)$, we may take every $\sigma_i = 1$. Let Y be an independent copy of X and interpolate along a rotation: for $\theta \in [0, \frac{\pi}{2}]$ put

$$X_\theta = X \sin\theta + Y \cos\theta, \qquad X'_\theta = X\cos\theta - Y\sin\theta.$$

For fixed θ the pair (X_θ, X'_θ) is a linear image of the standard Gaussian vector (X, Y), hence jointly Gaussian, with $\mathbb{E}[X_\theta X_\theta^\top] = \mathbb{E}[X'_\theta (X'_\theta)^\top] = I$ and $\mathbb{E}[X_\theta (X'_\theta)^\top] = (\sin\theta\cos\theta - \cos\theta\sin\theta)I = 0$: so X_θ and X'_θ are independent standard Gaussian vectors. Since $\frac{d}{d\theta}X_\theta = X'_\theta$, with $X_{\pi/2} = X$ and $X_0 = Y$,

$$f(X) - f(Y) = \int_0^{\pi/2} \langle \nabla f(X_\theta), X'_\theta \rangle \, d\theta,$$

and Cauchy–Schwarz on an interval of length $\frac{\pi}{2}$ gives $(f(X) - f(Y))^2 \le \frac{\pi}{2}\int_0^{\pi/2} \langle \nabla f(X_\theta), X'_\theta\rangle^2 d\theta$. Take expectations; the integrand is bounded, so $\mathbb{E}$ and $\int d\theta$ commute. For fixed θ, condition on X_θ: by the independence just established, $\mathbb{E}\langle \nabla f(X_\theta), X'_\theta\rangle^2 = \mathbb{E}\|\nabla f(X_\theta)\|^2 = \mathbb{E}\|\nabla f(X)\|^2$. Hence $\mathbb{E}[(f(X) - f(Y))^2] \le \frac{\pi^2}{4}\mathbb{E}\|\nabla f\|^2$, and since X, Y are independent copies, $\operatorname{Var} f(X) = \frac{1}{2}\mathbb{E}[(f(X) - f(Y))^2]$. ∎

The constant $\frac{\pi^2}{8}$ can be improved to 1, which is sharp, but nothing below depends on it.

Note what the right-hand side is and is not. It is not n times a typical squared derivative — it is the *sum*. A function of n independent Gaussians whose n partial derivatives are each of size $n^{-1/2}$ has variance $O(1)$, however large n is. Our f will be exactly of that type.

Proposition 4.6. *Let M be a Gaussian Wigner matrix and let $\Im z = \varepsilon > 0$. Then*

$$\operatorname{Var}\left(\Re \operatorname{tr}(M - zI)^{-1}\right) \;\le\; \frac{\pi^2}{4\,\varepsilon^4},$$

and the same bound holds for the imaginary part. Consequently

$$\mathbb{E}\big|m_N(z) - \mathbb{E}\, m_N(z)\big|^2 \;\le\; \frac{\pi^2}{2N^2\varepsilon^4}, \qquad \mathbb{E}\big|m_N(z) - \mathbb{E}\, m_N(z)\big| \;\le\; \frac{\pi}{\sqrt{2}\,N\varepsilon^2} \;<\; \frac{3}{N\varepsilon^2}.$$

Proof. The independent coordinates are the entries M_{ij} with $i \le j$, with $\operatorname{Var} M_{ij} = \frac{1}{N}$ off the diagonal and $\frac{2}{N}$ on it. Write $R = (M - zI)^{-1}$ and $f(M) = \operatorname{tr} R$. Differentiating $\operatorname{tr}(M + tH - zI)^{-1}$ at $t = 0$ with the resolvent identity gives

$$\frac{d}{dt}\Big|_{t=0} \operatorname{tr}\big(M + tH - zI\big)^{-1} = -\operatorname{tr}\big(RHR\big) = -\operatorname{tr}\big(HR^2\big).$$

Taking $H = E_{ij} + E_{ji}$ for $i < j$ and $H = E_{ii}$ for the diagonal,

$$\frac{\partial f}{\partial M_{ij}} = -2\big(R^2\big)_{ij} \quad (i < j), \qquad \frac{\partial f}{\partial M_{ii}} = -\big(R^2\big)_{ii}.$$

So the Poincaré sum is

$$\sum_{i<j} \frac{1}{N}\, 4\big|(R^2)_{ij}\big|^2 + \sum_i \frac{2}{N}\big|(R^2)_{ii}\big|^2 = \frac{2}{N}\sum_{i,j}\big|(R^2)_{ij}\big|^2.$$

Now diagonalize: $M = O\Lambda O^\top$ makes $R^2 = O(\Lambda - zI)^{-2}O^\top$, whose eigenvalues are $(\lambda_i - z)^{-2}$, and the quantity $\sum_{i,j}|(R^2)_{ij}|^2$ is unchanged by conjugating with O. Hence

$$\sum_{i,j}\big|(R^2)_{ij}\big|^2 = \sum_i \frac{1}{|\lambda_i - z|^4} \;\le\; \frac{N}{(\Im z)^4},$$

and the Poincaré sum is at most $\frac{2}{\varepsilon^4}$, with no N left; Lemma 4.5 multiplies it by $\frac{\pi^2}{8}$. Since $|\partial_{ij}\Re f| \le |\partial_{ij} f| \le 2/\varepsilon^2$, the lemma applies to $\Re f$ and to $\Im f$ separately, and adding the two gives the second display after dividing by N^2; the last inequality is Cauchy–Schwarz. ∎

The N's cancelled twice, in opposite directions: each derivative is of size $N^{-1/2}$ because the entries have variance $\frac{1}{N}$, and there are N^2 of them, but their squares sum to tr of a resolvent power, which is N times a bounded number rather than N^2 times one. The content is the identity $\sum_{i,j} |(R^2)_{ij}|^2 = \sum_i |\lambda_i - z|^{-4}$: the gradient of a trace is not a generic vector, it is another spectral sum.

Remark 4.7 (This is where rigidity shows up)*.* $\operatorname{Var}\operatorname{tr}(M - zI)^{-1} = O(1)$ says that linear statistics $\sum_i f(\lambda_i)$ of a random matrix satisfy a central limit theorem *with no normalization at all* — there is no $\frac{1}{\sqrt{N}}$, because the sum of N terms is already of constant order. For independent points that would be false. It is the first quantitative sign that the eigenvalues repel each other, which Lecture 8 makes exact.

Remark 4.8 (Gaussianity is a convenience, not the reason)*.* A weaker bound holds for any Wigner matrix with independent entries and no Gaussian assumption. Let $\mathcal{F}_k$ be generated by the first k rows and columns of M, and split $\operatorname{tr} R - \mathbb{E}\operatorname{tr} R = \sum_{k=1}^{N} D_k$ into martingale differences $D_k = \mathbb{E}[\operatorname{tr} R \mid \mathcal{F}_k] - \mathbb{E}[\operatorname{tr} R \mid \mathcal{F}_{k-1}]$. The matrix $M^{(k)}$ with the kth row and column removed does not involve the kth row, so $\operatorname{tr}(M^{(k)} - zI)^{-1}$ has the same conditional expectation given $\mathcal{F}_k$ and given $\mathcal{F}_{k-1}$, and may be subtracted inside both terms; (13) then bounds $|D_k| \le \frac{2}{\Im z}$. Martingale differences are orthogonal, so $\operatorname{Var}\operatorname{tr} R = O(N)$ and $\operatorname{Var}\left(\frac{1}{N}\operatorname{tr} R\right) = O(1/N)$. That is already enough for Step 4. The Gaussian argument is what improves $O(1/N)$ to the true $O(1/N^2)$, and it is needed for the rate in (15).

4.4 Solving the equation, and the theorem

The two roots. For $\Im z > 0$ the equation $w^2 + zw + 1 = 0$ has two roots, with product 1 and sum $-z$. Neither lies on the unit circle: if $|w| = 1$ then the other root is $1/w = \bar{w}$, and $z = -(w + \bar{w}) = -2\Re w$ is real. So exactly one root lies inside the unit disk. Call it $m(z)$, and the other one $m'(z) = 1/m(z)$. This is the definition of m. The familiar formula with a square root comes out below, together with which square root it is.

Lemma 4.9 (The good root)**.** *Let $\Im z = \varepsilon > 0$.*

(a) $|m(z)| < 1$, and m is continuous (indeed analytic) on the upper half plane.

(b) For $|z| > 1$, $|m(z)| \le \dfrac{1}{|z| - 1}$ and $\left|m(z) + \dfrac{1}{z}\right| \le \dfrac{1}{|z|\,(|z| - 1)^2}$; in particular $m(z) = -\dfrac{1}{z} + O(|z|^{-3})$.

(c) $\Im m(z) > 0$ and $\Im m'(z) < -\varepsilon$.

(d) $m(z) = \dfrac{-z + \sqrt{z^2 - 4}}{2}$, where $\sqrt{z^2 - 4}$ is the square root with positive imaginary part. (A number $w \notin [0, \infty)$ has two square roots, negatives of each other, so exactly one has positive imaginary part; and $z^2 - 4 \notin [0, \infty)$ whenever $\Im z > 0$, so this names one number for every z in the upper half plane, continuously. It is not the principal branch: that one has its cut where $z^2 - 4 < 0$, which includes the whole imaginary axis, and $\frac{-z + \sqrt{z^2 - 4}}{2}$ with the principal root switches from the small root to the large one as z crosses it.) Moreover m extends continuously to the closed upper half plane, with $\Im m(\lambda + i0) = \frac{1}{2}\sqrt{4 - \lambda^2}$ for $|\lambda| \le 2$ and $\Im m(\lambda + i0) = 0$ for $|\lambda| \ge 2$.

Proof. (a) $|m| < 1$ is the definition. The discriminant $z^2 - 4$ does not vanish for $\Im z > 0$, so near any point the two roots are given by two analytic functions of z; neither ever meets the unit circle, so the one inside the disk stays inside, and it is m.

(b) From $m' = -z - m$ and $|m| < 1$, $|m'| \geq |z| - 1$, hence $|m| = 1/|m'| \leq \frac{1}{|z|-1}$: the small root is small because the big one is $-z$ up to a bounded error and their product is 1. For the sharper statement write the equation as $m(-z-m) = 1$, i.e. $m = \frac{1}{-z-m}$. Then

$$m + \frac{1}{z} = \frac{1}{-z-m} + \frac{1}{z} = \frac{m}{z\,(z+m)}, \qquad \text{so} \qquad \left|m + \frac{1}{z}\right| \leq \frac{|m|}{|z|\,(|z|-1)} \leq \frac{1}{|z|\,(|z|-1)^2}.$$

(c) From $m + m' = -z$, $\Im m + \Im m' = -\varepsilon$; from $m' = 1/m$, $\Im m' = -\Im m/|m|^2$. Together, $\Im m\,(1 - |m|^{-2}) = -\varepsilon$, and since $|m| < 1$ the bracket is negative, so $\Im m > 0$. Then $\Im m' = -\varepsilon - \Im m < -\varepsilon$.

(d) Put $s = m - m'$. Then $s^2 = (m+m')^2 - 4mm' = z^2 - 4$, so s is one of the two square roots of $z^2 - 4$, and $\Im s = \Im m - \Im m' > 0$ by (c). Hence $m = \frac{(m+m')+(m-m')}{2} = \frac{-z+s}{2}$, which is the formula, with the branch now forced rather than chosen. For the boundary values let $z = \lambda + i\varepsilon \to \lambda$. If $|\lambda| < 2$ then $z^2 - 4 \to \lambda^2 - 4 < 0$, and the square root with positive imaginary part tends to $i\sqrt{4-\lambda^2}$. If $|\lambda| > 2$ then $z^2 - 4 \to \lambda^2 - 4 > 0$ with $\Im(z^2-4) = 2\lambda\varepsilon$, so from above when $\lambda > 2$ and from below when $\lambda < -2$; the square root tends to $+\sqrt{\lambda^2-4}$ and to $-\sqrt{\lambda^2-4}$ respectively, and $m(\lambda + i0)$ is real. At $\lambda = \pm 2$ the square root tends to 0. In each case the limit is attained continuously, and the formulas for $\Im m(\lambda + i0)$ follow. ∎

Remark 4.10 (Why $-1/z$). Part (b) is worth seeing without the square root. The formula in (d) has $\sqrt{z^2-4} \approx +z$ for large z in the upper half plane (the sign is fixed by the imaginary part), so the two terms in $-z + \sqrt{z^2-4}$ cancel at leading order and $-1/z$ is what survives at the next one: $\sqrt{z^2-4} = z\sqrt{1-4/z^2} = z - \frac{2}{z} + O(|z|^{-3})$. That is correct but easy to get wrong. The proof of (b) says the same thing with no branch to track: the roots multiply to 1, one of them is $-z$ up to a bounded error, so the other is $-1/z$ up to a relative error of order $|z|^{-2}$.

Remark 4.11 (Reading the density off the equation). The boundary values in (d) can be read directly from $w^2 + zw + 1 = 0$, with no square root. At a real point $z = x$ the coefficients are real, so the two roots are either both real or a conjugate pair. If $\Im m(x + i0) \neq 0$ they are conjugate, so $m' = \bar{m}$ and $|m|^2 = mm' = 1$: *on the support, m lies on the unit circle.* Write $m = e^{i\theta}$ and divide the equation by $e^{i\theta}$:

$$e^{i\theta} + x + e^{-i\theta} = 0 \qquad \Longrightarrow \qquad x = -2\cos\theta, \qquad \Im\, m(x+i0) = \sin\theta = \tfrac{1}{2}\sqrt{4-x^2}.$$

If instead both roots are real, $\Im m(x+i0) = 0$. So the limiting density is $\frac{1}{\pi}\sin\theta$ at $x = -2\cos\theta$ and zero elsewhere. This is where the name comes from: the graph of ρ_{sc} is an ellipse, but $m(x+i0)$ traces the upper unit semicircle as x runs from -2 to 2, and the density is its height divided by π.

Stability. An approximate solution of (16) in the upper half plane is close to m, and the error only gets divided by ε.

Lemma 4.12 (Stability of the fixed point)**.** *Let $\Im z = \varepsilon > 0$, and let $w \in \mathbb{C}$ satisfy $\Im w \geq 0$ and $\left|w - \frac{1}{-z-w}\right| \leq \delta$. Then*

$$|w - m(z)| \;\leq\; \frac{\delta}{\varepsilon}\,\big(|z| + |w|\big).$$

Proof. Multiplying the hypothesis by $|z+w|$ gives $|w^2 + zw + 1| \leq \delta\,|z+w|$, and $w^2 + zw + 1 = (w-m)(w-m')$. By Lemma 4.9(c), $|w - m'| \geq \Im w - \Im m' > \varepsilon$. Divide. ∎

Theorem 4.13 (Convergence of the transform)**.** *For $\Im z = \varepsilon \in (0,1]$,*

$$\big|\mathbb{E}\, m_N(z) - m(z)\big| \;\le\; \frac{\delta_N(\varepsilon)}{\varepsilon}\Big(|z| + \frac{1}{\varepsilon}\Big) \;\le\; \frac{7\,\big(|z| + \varepsilon^{-1}\big)}{\varepsilon^5\sqrt{N}}.$$

In particular $\mathbb{E}\, m_N(x + i\varepsilon) \to m(x + i\varepsilon)$ uniformly for x in bounded sets.

Proof. Apply Lemma 4.12 to $w = \mathbb{E}m_N(z)$, which has $\Im\mathbb{E}m_N > 0$ and $|\mathbb{E}m_N| \le \frac{1}{\varepsilon}$ by (7), with $\delta = \delta_N(\varepsilon)$ from (14) and (15). ■

The theorem. Everything Proposition 3.7 asks for is now in hand.

Theorem 4.14 (The semicircle law, by the Stieltjes transform)**.** *Let M be a Gaussian Wigner matrix with eigenvalues $\lambda_1, \dots, \lambda_N$. For every $a < b$,*

$$\mathbb{E}\left|\,\frac{1}{N}\#\{i : \lambda_i \in (a,b)\} - \int_a^b \rho_{sc}(x)\,dx\,\right| \;\longrightarrow\; 0, \qquad \rho_{sc}(x) = \frac{1}{2\pi}\sqrt{(4 - x^2)_+}\,.$$

Proof. Check the hypotheses of Proposition 3.7. The first, $\mathbb{E}m_N(x + i\varepsilon) \to m(x + i\varepsilon)$ uniformly on bounded sets of x, is Theorem 4.13. The second, $\sup_x \mathbb{E}|m_N(x + i\varepsilon) - \mathbb{E}m_N(x + i\varepsilon)| \to 0$, is Proposition 4.6, whose bound $\frac{3}{N\varepsilon^2}$ does not involve x. The third, $\Im m \le K$, holds with $K = 1$ by Lemma 4.9(a). So the proposition gives $\mathbb{E}|\frac{1}{N}\#\{\lambda_i \in (a,b)\} - \ell(a,b)| \to 0$ with $\ell(a,b) = \lim_{\varepsilon\downarrow 0} \frac{1}{\pi}\int_a^b \Im m(x + i\varepsilon)\,dx$. It remains to identify ℓ. By Lemma 4.9(d), m is continuous on the compact set $[a,b]\times[0,1]$, hence uniformly continuous there, so $\Im m(x+i\varepsilon) \to \Im m(x+i0)$ uniformly on $[a,b]$ and

$$\ell(a,b) = \frac{1}{\pi}\int_a^b \Im\, m(x + i0)\,dx = \frac{1}{\pi}\int_a^b \tfrac{1}{2}\sqrt{(4 - x^2)_+}\,dx = \int_a^b \rho_{sc}(x)\,dx. \qquad ■$$

That is Theorem 2.1 again, in a sharper form — the proportion of eigenvalues in every interval, in L^1 — and the density arrived by contour integration rather than by recognition.

Remark 4.15 (What was used). Wick's theorem (Lecture 5), in Lemma 4.4; the residue theorem, in Proposition 3.3; and two facts proved here, the Schur complement and the Poincaré inequality. Nothing else is quoted. The two places where an expectation was moved past an integral — $\int_a^b dx$ in Proposition 3.7 and $\int_0^{\pi/2} d\theta$ in Lemma 4.5 — involve integrands bounded by a constant depending only on ε, and in the first case $x \mapsto m_N(x + i\varepsilon)$ is Lipschitz with the deterministic constant ε^{-2} by the resolvent identity, so the integral is a uniform limit of Riemann sums. The rate $N^{-1/2}$ in (15) is not sharp: the truth is N^{-1}, and the loss is in bounding $\mathbb{E}|\Delta|$ by first moments rather than expanding to second order.

Remark 4.16 (The two proofs are the same equation). Expanding m for large $|z|$ as in (9) gives $m(z) = -\sum_{k\ge 0} C_k z^{-(2k+1)}$, so with $w = z^{-2}$ we have $-zm(z) = C(w)$. Substituting into the Catalan relation (6) returns $-zm = 1 + m^2$, which is (16). *The recursion that counts non-crossing pair partitions and the self-consistent equation for the resolvent are one equation written twice.*

Return to the four complaints. The density came out directly. Only the second moment of the entries was used, so nothing needed all moments. Unequal variances $\mathbb{E}M_{ij}^2 = S_{ij}/N$ turn (16) into a system of N coupled equations rather than one, which is a change of size and not of method. A rank-one spike is a rank-one perturbation of the resolvent. And shrinking $\Im z$ toward the real axis, rather than holding it fixed, is precisely what converts a global law into a local one.

Exercise 4.17 (TYPE A)**.** Take $N = 2$ and $A = \begin{pmatrix} \alpha & b \\ b & d \end{pmatrix}$ with $d \neq 0$ and $\alpha d - b^2 \neq 0$. Invert A directly and check both statements of Corollary 4.2: $(A^{-1})_{11} = \frac{1}{\alpha - b^2/d}$, and $\operatorname{tr} A^{-1} = \frac{1}{d} + (A^{-1})_{11}\big(1 + \frac{b^2}{d^2}\big)$. Then take $A = M - zI$ for a symmetric 2×2 matrix M and verify (13) by hand: $\big|\operatorname{tr}(M - zI)^{-1} - \frac{1}{M_{22} - z}\big| \leq \frac{1}{\Im z}$.

Exercise 4.18 (TYPE A)**.** Let M be a Gaussian Wigner matrix and $\sigma > 0$. The entries of σM have variance σ^2/N off the diagonal. Show, by rescaling z rather than redoing any estimate, that the transform of σM converges to the root of

$$\sigma^2 m^2 + zm + 1 = 0$$

in the disk $|m| < 1/\sigma$, and use the argument of Remark 4.11 to read off the limiting density $\frac{1}{2\pi\sigma^2}\sqrt{(4\sigma^2 - x^2)_+}$: a semicircle of radius 2σ. *The variance of an entry enters the equation in exactly one place.*

Exercise 4.19 (TYPE B)**.** Apply Lemma 4.5 exactly as in the proof of Proposition 4.6, but to $f(M) = \operatorname{tr} M^2 = \sum_{i,j} M_{ij}^2$, and show

$$\operatorname{Var}\big(\operatorname{tr} M^2\big) \;\leq\; \frac{\pi^2}{8}\Big(8 + \frac{8}{N}\Big).$$

Compute $\partial f/\partial M_{ij}$ for $i < j$ and for $i = j$, remembering that $M_{ij} = M_{ji}$ is one variable. The exact value is $4 + \frac{4}{N}$, by Wick. The point is that Poincaré sees the $O(1)$ with no combinatorics at all — the rigidity of Remark 4.7, for the simplest linear statistic there is.

ASSIGNED **Lecture 8**, due at the start of Lecture 9: Exercises 4.17, 4.18, 4.19. *Three problems. The second is the whole lecture in miniature.*

Lecture 9 derives the exact joint law of the eigenvalues and reads it as a gas of repelling charges; the semicircle appears a third time, as its equilibrium measure.