MATH 247A: INTRODUCTION TO RANDOM MATRIX THEORY

TODD KEMP

CONTENTS

Date: December 5, 2016.

For the next ten weeks, we will be studying the eigenvalues of random matrices. A *random matrix* is simply a matrix (for now square) all of whose entries are random variables. That is: X is an $n \times n$ matrix with $\{X_{ij}\}_{1\le i,j\le n}$ random variables on some probability space $(\Omega, \mathscr{F}, \mathbb{P})$. Alternatively, one can think of X as a random *vector* in $\mathbb{R}^{n^2}$ (or $\mathbb{C}^{n^2}$), although it is better to think of it as taking values in $M_n(\mathbb{R})$ or $M_n(\mathbb{C})$ (to keep the matrix structure in mind). For any fixed instance $\omega \in \Omega$, then, $X(\omega)$ is an $n \times n$ matrix and (maybe) has eigenvalues $\lambda_1(\omega), \ldots, \lambda_n(\omega)$. So the eigenvalues are also random variables. We seek to understand aspects of the distributions of the λ_i from knowledge of the distribution of the X_{ij}. To begin, in order to guarantee that the matrix actually has eigenvalues (and they accurately capture the behavior of the linear transformation X), we will make the assumption that X is a symmetric / Hermitian matrix.

1. Wigner Matrices

We begin by fixing an infinite family of real-valued random variables $\{Y_{ij}\}_{j\ge i\ge 1}$. Then we can define a sequence of symmetric random matrices $\mathbf{Y}_n$ by

$$[\mathbf{Y}_n]_{ij} = \begin{cases} Y_{ij}, & i \le j \\ Y_{ji}, & i > j \end{cases}.$$

The matrix $\mathbf{Y}_n$ is symmetric and so has n real eigenvalues, which we write in increasing order $\lambda_1(\mathbf{Y}_n) \le \cdots \le \lambda_n(\mathbf{Y}_n)$. In this very general setup, little can be said about these eigenvalues. The class of matrices we are going to begin studying, *Wigner matrices*, are given by the following conditions.

- We assume that the $\{Y_{ij}\}_{1\le i\le j}$ are independent.
- We assume that the diagonal entries $\{Y_{ii}\}_{i\ge 1}$ are identically-distributed, and the off-diagonal entries $\{Y_{ij}\}_{1\le i<j}$ are identically-distributed.
- We assume that $\mathbb{E}(Y_{ij}^2) < \infty$ for all i, j. (I.e. $r_2 = \max\{\mathbb{E}(Y_{11}^2), \mathbb{E}(Y_{12}^2)\} < \infty$.)

It is not just for convenience that we separate out the diagonal terms; as we will see, they really do not contribute to the eigenvalues in the limit as $n \to \infty$. It will also be convenient, at least at the start, to strengthen the final assumption to moments of *all* orders: for $k \ge 1$, set

$$r_k = \max\{\mathbb{E}(Y_{11}^k), \mathbb{E}(Y_{12}^k)\}.$$

To begin, we will assume that $r_k < \infty$ for each k; we will weaken this assumption later. Note: in the presence of higher moments, much of the following will not actually require identically-distributed entries. Rather, uniform bounds on all moments suffice. Herein, we will satisfy ourselves with the i.i.d. case.

One variation we will allow from i.i.d. is a uniform scaling of the matrices as n increases. That is: let α_n be a sequence of positive scalars, and set

$$\mathbf{X}_n = \alpha_n \mathbf{Y}_n.$$

In fact, there is a natural choice for the scaling behavior, in order for there to be limiting behavior of eigenvalues. That is to say: we would like to arrange (if possible) that both $\lambda_1(\mathbf{X}_n)$ and $\lambda_n(\mathbf{X}_n)$ converge to finite numbers (different from each other). An easy way to test this is suggested by the following simple calculation:

$$\lambda_1^2 + \cdots + \lambda_n^2 \le n \cdot \max_{j=1}^{n}\{\lambda_j^2\} = n \cdot \max\{\lambda_1^2, \lambda_n^2\}.$$

Hence, in order for λ_1 and λ_n to both converge to distinct constants, it is necessary for the sequence $\frac{1}{n}(\lambda_1^2 + \cdots + \lambda_n^2)$ to be bounded. Fortunately, this sequence can be calculated without explicit reference to eigenvalues: it is the (normalized) *Hilbert-Schmidt* norm (or *Fröbenius* norm) of a symmetric matrix $\mathbf{X}$:

$$\|\mathbf{X}\|_2^2 = \frac{1}{n}\sum_{i=1}^{n} \lambda_i(\mathbf{X}_n)^2 = \frac{1}{n}\sum_{1\le i,j\le n} X_{ij}^2.$$

Exercise 1.0.1. *Verify the second equality above, by showing (using the spectral theorem) that both expressions are equal to the quantity* $\frac{1}{n}\operatorname{Tr}(\mathbf{X}^2)$.

For our *random* matrix $\mathbf{X}_n$ above, then, we can calculate the expected value of this norm:

$$\begin{aligned}\mathbb{E}\|\mathbf{X}_n\|_2^2 = \frac{1}{n}\alpha_n^2 \sum_{1\le i,j\le n}\mathbb{E}(Y_{ij}^2) &= \frac{\alpha_n^2}{n}\sum_{i=1}^{n}\mathbb{E}(Y_{ii})^2 + \frac{2\alpha_n^2}{n}\sum_{1\le i<j\le n}\mathbb{E}(Y_{ij}^2)\\ &= \alpha_n^2 \cdot \mathbb{E}(Y_{11}^2) + (n-1)\alpha_n^2 \cdot \mathbb{E}(Y_{12}^2).\end{aligned}$$

We now have two cases. If $\mathbb{E}(Y_{12}^2) = 0$ (meaning the off-diagonal terms are all 0 $a.s.$) then we see the "correct" scaling for α_n is $\alpha_n \sim 1$. This is a boring case: the matrices $\mathbf{X}_n$ are diagonal, with all diagonal entries identically distributed. Thus, these entries are also the eigenvalues, and so the distribution of eigenvalues is given by the common distribution of the diagonal entries. We ignore this case, and therefore assume that $\mathbb{E}(Y_{12}^2) > 0$. Hence, in order for $\mathbb{E}\|\mathbf{X}_n\|_2^2$ to be a bounded sequence (that does not converge to 0), we must have $\alpha_n \sim n^{-1/2}$.

Definition 1.1. *Let* $\{Y_{ij}\}_{1\le i\le j}$ *and* $\mathbf{Y}_n$ *be as above, with* $r_2 < \infty$ *and* $\mathbb{E}(Y_{12}^2) > 0$. *Then the matrices* $\mathbf{X}_n = n^{-1/2}\mathbf{Y}_n$ *are* **Wigner matrices***.*

It is standard to abuse notation and refer to the sequence $\mathbf{X}_n$ as a *a Wigner matrix*. The preceding calculation shows that, if $\mathbf{X}_n$ is a Wigner matrix, then the expected Hilbert-Schmidt norm $\mathbb{E}\|\mathbf{X}_n\|_2^2$ converges (as $n \to \infty$) to the second moment of the (off-diagonal) entries. As explained above, this is prerequisite to the bulk convergence of the eigenvalues. As we will shortly see, it is also sufficient.

Consider the following example. Take the entires Y_{ij} to be $N(0,1)$ normal random variables. These are easy to simulate with MATLAB. Figure 1 shows the histogram of all $n = 4000$ eigenvalues of one instance of the corresponding Gaussian Wigner matrix $\mathbf{X}_{4000}$. The plot suggests that

$\lambda_1(\mathbf{X}_n) \to -2$ while $\lambda_n(\mathbf{X}_n) \to 2$ in this case. Moreover, although random fluctuations remain, it appears that the histogram of eigenvalues (sometimes called the *density* of eigenvalues) converges to a deterministic shape. In fact, this is *universally* true. There is a universal probability distribution σ_t such that the density of eigenvalues of *any* Wigner matrix (with second moment t) converges to σ_t. The limiting distribution is known as **Wigner's semicircle law**:

$$\sigma_t(dx) = \frac{1}{2\pi t}\sqrt{(4t - x^2)_+}\, dx.$$

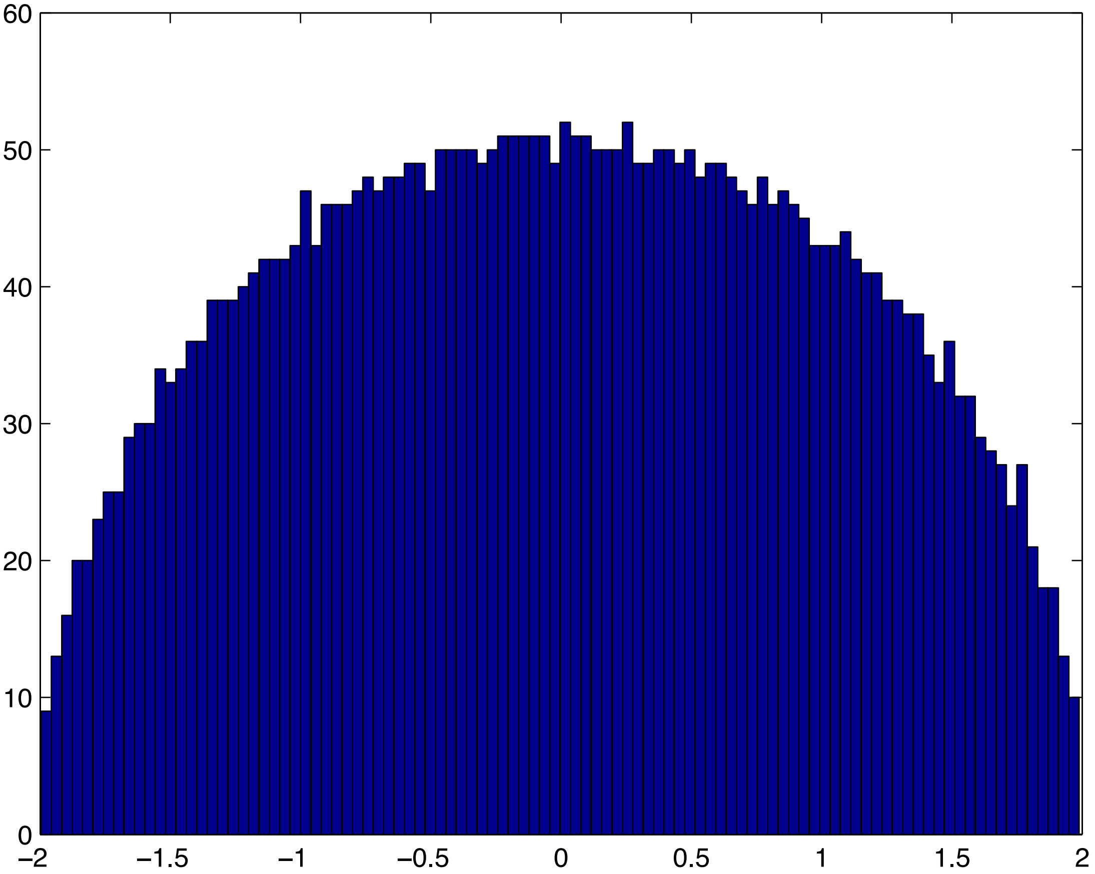


FIGURE 1. The density of eigenvalues of an instance of $\mathbf{X}_{4000}$, a Gaussian Wigner matrix.

2. Wigner's Semicircle Law

Theorem 2.1 (Wigner's Semicircle Law). *Let $\mathbf{X}_n = n^{-1/2}\mathbf{Y}_n$ be a sequence of Wigner matrices, with entries satisfying $\mathbb{E}(Y_{ij}) = 0$ for all i, j and $\mathbb{E}(Y_{12}^2) = t$. Let $I \subset \mathbb{R}$ be an interval. Define the random variables*

$$E_n(I) = \frac{\#\left(\{\lambda_1(\mathbf{X}_n), \ldots, \lambda_n(\mathbf{X}_n)\} \cap I\right)}{n}$$

Then $E_n(I) \to \sigma_t(I)$ in probability as $n \to \infty$.

The first part of this course is devoted to proving Wigner's Semicircle Law. The key observation (that Wigner made) is that one can study the behavior of the random variables $E_n(I)$ without computing the eigenvalues directly. This is accomplished by reinterpreting the theorem in terms of a *random measure*, the **empirical law of eigenvalues**.

Definition 2.2. *Let $\mathbf{X}_n$ be a Wigner matrix. Its* **empirical law of eigenvalues** *$\mu_{\mathbf{X}_n}$ is the random discrete probability measure*

$$\mu_{\mathbf{X}_n} = \frac{1}{n}\sum_{j=1}^{n} \delta_{\lambda_j(\mathbf{X}_n)}.$$

That is: $\mu_{\mathbf{X}_n}$ is defined as the (random) probability measure such that, for any continuous function $f \in C(\mathbb{R})$, the integral $\int f\, d\mu_{\mathbf{X}_n}$ is the random variable

$$\int f\, d\mu_{\mathbf{X}_n} = \frac{1}{n}\sum_{j=1}^{n} f(\lambda_j(\mathbf{X}_n)).$$

Note that the random variables $E_n(I)$ in Theorem 2.1 are given by $E_n(I) = \int \mathbb{1}_I\, d\mu_{\mathbf{X}_n}$. Although $\mathbb{1}_I$ is not a continuous function, a simple approximation argument shows that the following theorem (which we also call Wigner's semicircle law) is a stronger version of Theorem 2.1.

Theorem 2.3 (Wigner's Semicircle Law). *Let $\mathbf{X}_n = n^{-1/2}\mathbf{Y}_n$ be a sequence of Wigner matrices, with entries satisfying $\mathbb{E}(Y_{ij}) = 0$ for all i, j and $\mathbb{E}(Y_{12}^2) = t$. Then the empirical law of eigenvalues $\mu_{\mathbf{X}_n}$ converges in probability to σ_t as $n \to \infty$. Precisely: for any $f \in C_b(\mathbb{R})$ (continuous bounded functions) and each $\epsilon > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(\left|\int f\, d\mu_{\mathbf{X}_n} - \int f\, d\sigma_t\right| > \epsilon\right) = 0.$$

In this formulation, we can use the spectral theorem to eliminate the explicit appearance of eigenvalues in the law $\mu_{\mathbf{X}_n}$. Diagonalize $\mathbf{X}_n = U_n^\top \mathbf{\Lambda}_n U_n$. Then (by definition)

$$\begin{aligned}\int f\, d\mu_{\mathbf{X}_n} = \frac{1}{n}\sum_{j=1}^{n} f(\lambda_j(\mathbf{X}_n)) &= \frac{1}{n}\sum_{j=1}^{n} f([\mathbf{\Lambda}_n]_{jj}) \\ &= \frac{1}{n}\operatorname{Tr} f(\mathbf{\Lambda_n}) = \frac{1}{n}\operatorname{Tr}\left(U_n^\top f(\mathbf{\Lambda_n}) U_n\right) = \frac{1}{n}\operatorname{Tr} f(\mathbf{X}_n).\end{aligned}$$

The last equality is the statement of the spectral theorem. Usually we would use it in reverse to *define* $f(\mathbf{X}_n)$ for measurable f. However, in this case, we will take f to be a polynomial. (Since both $\mu_{\mathbf{X}_n}$ and σ_t are compactly-supported, any polynomial is equal to a $C_b(\mathbb{R})$ function on their supports, so this is consistent with Theorem 2.3.) This leads us to the third, a priori weaker form of Wigner's law.

Theorem 2.4 (Wigner's law for matrix moments)**.** *Let $\mathbf{X}_n = n^{-1/2}\mathbf{Y}_n$ be a sequence of Wigner matrices, with entries satisfying $\mathbb{E}(Y_{ij}) = 0$ for all i,j and $\mathbb{E}(Y_{12}^2) = t$. Let f be a polynomial. Then the random variables $\int f\,d\mu_{\mathbf{X}_n}$ converge to $\int f\,d\sigma_t$ in probability as $n\to\infty$. Equivalently: for fixed $k\in\mathbb{N}$ and $\epsilon>0$,*

$$\lim_{n\to\infty}\mathbb{P}\left(\left|\frac{1}{n}\operatorname{Tr}(\mathbf{X}_n^k) - \int x^k\,\sigma_t(dx)\right| > \epsilon\right) = 0.$$

Going back from Theorem 2.4 to Theorem 2.3 is, in principal, just a matter of approximating any C_b function by polynomials on the supports of $\mu_{\mathbf{X}_n}$ and σ_t. However, even though $\mu_{\mathbf{X}_n}$ has finite support, there is no obvious bound on $\bigcup_n \operatorname{supp}\mu_{\mathbf{X}_n}$, and this makes such an approximation scheme tricky. In fact, when we eventually fully prove Theorem 2.3 (and hence Theorem 2.1 – approximating $\mathbb{1}_I$ by C_b functions is routine in this case), we will take a different approach entirely. For now, however, it will be useful to have some intuition for the theorem, so we begin by presenting a scheme for proving Theorem 2.4. The place to begin is calculating the moments $\int x^k\,\sigma_t(dx)$.

Exercise 2.4.1. *Let σ_t be the semicircle law of variance t defined above.*

(a) *With a suitable change of variables, show that $\int x^k\sigma_t(dx) = t^{k/2}\int x^k\sigma_1(dx)$.*

(b) *Let $m_k = \int x^k\,\sigma_1(dx)$. By symmetry, $m_{2k+1} = 0$ for all k. Use a trigonometric substitution to show that*

$$m_0 = 1, \quad m_{2k} = \frac{2(2k-1)}{k+1}m_{2(k-1)}.$$

This recursion completely determines the even moments; show that, in fact,

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

The numbers C_k in Exercise 2.4.1 are the *Catalan numbers*. No one would disagree that they form the most important integer sequence in combinatorics. In fact, we will use some of their combinatorial interpretations to begin our study of Theorem 2.4.

3. Markov's Inequality and Convergence of Expectation

To prove Theorem 2.4, we will begin by showing convergence of *expectations*. Once this is done, the result can be proved by showing the variance tends to 0, because of:

Lemma 3.1 (Markov's Inequality)**.** *Let $Y \geq 0$ be a random variable with $\mathbb{E}(Y) < \infty$. Then for any $\epsilon > 0$,*

$$\mathbb{P}(Y > \epsilon) \leq \frac{\mathbb{E}(Y)}{\epsilon}.$$

Proof. We simply note that $\mathbb{P}(Y > \epsilon) = \mathbb{E}(\mathbb{1}_{\{Y>\epsilon\}})$. Now, on the event $\{Y > \epsilon\}$, $\frac{1}{\epsilon}Y > 1$; on the complement of this event, $\mathbb{1}_{\{Y>\epsilon\}} = 0$ while $\frac{1}{\epsilon}Y \geq 0$. Altogether, this means that we have the pointwise bound

$$\frac{1}{\epsilon}Y \geq \mathbb{1}_{\{Y>\epsilon\}}.$$

Taking expectations of both sides yields the result. □

Corollary 3.2. *Suppose X_n is a sequence of random variables with $\mathbb{E}(X_n^2) < \infty$ for each n. Suppose that $\lim_{n\to\infty} \mathbb{E}(X_n) = m$, and that $\lim_{n\to\infty} \mathrm{Var}(X_n) = 0$. Then $X_n \to m$ in probability: for each $\epsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|X_n - m| > \epsilon) = 0$.*

Proof. Applying Markov's inequality to $Y = |X_n - m|$ (with ϵ^2 in place of ϵ),

$$\mathbb{P}(|X_n - m| > \epsilon) = \mathbb{P}((X_n - m)^2 > \epsilon^2) \leq \frac{\mathbb{E}((X_n - m)^2)}{\epsilon^2}.$$

Now, using the triangle inequality for the $L^2(\mathbb{P})$-norm, we have

$$\begin{aligned}\mathbb{E}((X_n - m))^2 = \|X_n - m\|_{L^2(\mathbb{P})}^2 &= \|X_n - \mathbb{E}(X_n) + \mathbb{E}(X_n) - m\|_{L^2(\mathbb{P})}^2 \\ &\leq \left(\|X_n - \mathbb{E}(X_n))\|_{L^2(\mathbb{P})} + \|\mathbb{E}(X_n) - m\|_{L^2(\mathbb{P})}\right)^2.\end{aligned}$$

Both $\mathbb{E}(X_n)$ and m are constants, so the second $L^2(\mathbb{P})$ norm is simply $|\mathbb{E}(X_n) - m|$; the first term, on the other hand, is

$$\|X_n - \mathbb{E}(X_n)\|_{L^2(\mathbb{P})} = \sqrt{\mathbb{E}[(X_n - \mathbb{E}(X_n))^2]} = \sqrt{\mathrm{Var}(X_n)}.$$

Hence, by Markov's inequality, we have

$$\mathbb{P}(|X_n - m| > \epsilon) \leq \frac{(\sqrt{\mathrm{Var}(X_n)} + |\mathbb{E}(X_n) - m|)^2}{\epsilon^2}.$$

The right-hand-side tends to 0 by assumption; this proves the result. □

Remark 3.3. Using the Borel-Cantelli lemma, one can strengthen the result of Corollary 3.2: indeed, $X_n \to m$ almost surely, and so almost sure convergence also applies to Wigner's theorem. (This requires some additional estimates showing that the *rate* of convergence in probability is summably fast.) It is customary to state Wigner's theorem in terms of convergence in probability, though, because our setup (choosing all entries of all the matrices from one common probability space) is a little artificial. If the entries of Y_n and Y_{n+1} need not come from the same probability space, then almost sure convergence has no meaning; but convergence in probability still makes perfect sense.

Applying Corollary 3.2 to the terms in Theorem 2.4, this gives a two-step outline for a method of proof.

Step 1. For each odd k, show that $\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,(\mathbf{X}_n^k) \to 0$ as $n \to \infty$; for each even k, show that $\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,(\mathbf{X}_n^k) \to C_{k/2}$ as $n \to \infty$.

Step 2. For each k, show that $\mathrm{Var}\left(\frac{1}{n}\,\mathrm{Tr}\,(\mathbf{X}_n^k)\right) \to 0$ as $n \to \infty$.

We will presently follow through with those two steps in the next section. The proof is quite combinatorial. We will also present a completely different, more analytical, proof of Theorem 2.3 in later lectures.

Remark 3.4. Note that the statement of Theorem 2.4 does not require higher moments of $\mathrm{Tr}\,(\mathbf{X}_n^k)$ to exist – the result is convergence in probability (and this is one benefit of stating the convergence this way). However: Step 1 above involves expectations, and so we will be taking expectations of terms that involve kth powers of the variables Y_{ij}. Thus, in order to follow through with this program, we will need to make the additional assumption that $r_k < \infty$ for all k; the entries of $\mathbf{X}_n$ possess moments of *all* orders. This is a technical assumption that can be removed after the fact with appropriate cut-offs; when we delve into the analytical proof later, we will explain this.

4. First Proof of Wigner's Semicircle Law

In this section, we give a complete proof of Theorem 2.4, under the assumption that all moments are finite.

4.1. **Convergence of matrix moments in Expectation.** We begin with Step 1: convergence in expectation.

Proposition 4.1. *Let $\{Y_{ij}\}_{1\le i\le j}$ be independent random variables, with $\{Y_{ii}\}_{i\ge 1}$ identically distributed and $\{Y_{ij}\}_{1\le i<j}$ identically distributed. Suppose that $r_k = \max\{\mathbb{E}(|Y_{11}|^k), \mathbb{E}(|Y_{12}|^k)\} < \infty$ for each $k\in\mathbb{N}$. Suppose further than $\mathbb{E}(Y_{ij}) = 0$ for all i,j and set $t = \mathbb{E}(Y_{12}^2)$. If $i > j$, define $Y_{ij} \equiv Y_{ji}$, and let $\mathbf{Y}_n$ be the $n\times n$ matrix with $[\mathbf{Y}_n]_{ij} = Y_{ij}$ for $1 \le i, j \le n$. Let $\mathbf{X}_n = n^{-1/2}\mathbf{Y}_n$ be the corresponding Wigner matrix. Then*

$$\lim_{n\to\infty} \frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,(\mathbf{X}_n^k) = \begin{cases} t^{k/2}C_{k/2}, & k \text{ even} \\ 0, & k \text{ odd} \end{cases}.$$

Proof. To begin the proof, we expand the expected trace terms in terms of the entries. First, we have

$$\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,(\mathbf{X}_n^k) = \frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,[(n^{-1/2}\mathbf{Y}_n)^k] = n^{-k/2-1}\mathbb{E}\,\mathrm{Tr}\,(\mathbf{Y}_n^k). \tag{4.1}$$

Now, from the (repeated) definition of matrix multiplication, we have for any $1\le i,j\le n$

$$[\mathbf{Y}_n^k]_{ij} = \sum_{1\le i_2,\ldots,i_k\le n} Y_{ii_2}Y_{i_2i_3}\cdots Y_{i_{k-1}i_k}Y_{i_kj}.$$

Summing over the diagonal entries and taking the expectation, we get the expected trace:

$$\mathbb{E}\,\mathrm{Tr}\,(\mathbf{Y}_n^k) = \sum_{i_1=1}^n \mathbb{E}([\mathbf{Y}_n^k]_{i_1i_1}) = \sum_{1\le i_1,i_2,\ldots,i_k\le n} \mathbb{E}(Y_{i_1i_2}Y_{i_2i_3}\cdots Y_{i_ki_1}) \equiv \sum_{\mathbf{i}\in[n]^k}\mathbb{E}(Y_{\mathbf{i}}), \tag{4.2}$$

where $[n] = \{1,\ldots,n\}$, and for $\mathbf{i} = (i_1,\ldots,i_k)$ we define $Y_{\mathbf{i}} = Y_{i_1i_2}\cdots Y_{i_ki_1}$. The term $\mathbb{E}(Y_{\mathbf{i}})$ is determined by the k-index $\mathbf{i}$, but only very weakly. The sum is better indexed by a collection of *walks on graphs* arising from the indices.

Definition 4.2. *Let $\mathbf{i}\in[n]^k$ be a k-index, $\mathbf{i} = (i_1,i_2,\ldots,i_k)$. Define a graph $G_{\mathbf{i}}$ as follows: the vertices $V_{\mathbf{i}}$ are the* distinct *elements of $\{i_1,i_2,\ldots,i_k\}$, and the edges $E_{\mathbf{i}}$ are the* distinct *pairs among $\{i_1,i_2\},\{i_2,i_3\},\ldots,\{i_{k-1},i_k\},\{i_k,i_1\}$. The path $w_{\mathbf{i}}$ is the sequence*

$$w_{\mathbf{i}} = (\{i_1,i_2\},\{i_2,i_3\},\ldots,\{i_{k-1},i_k\},\{i_k,i_1\})$$

of edges.

For example, take $\mathbf{i} = (1,2,2,3,5,2,4,1,4,2)$; then

$$V_{\mathbf{i}} = \{1,2,3,4,5\}, \qquad E_{\mathbf{i}} = \{\{1,2\},\{1,4\},\{2,2\},\{2,3\},\{2,4\},\{2,5\},\{3,5\}\}.$$

The index $\mathbf{i}$ which defines the graph now also defines a *closed walk* $w_{\mathbf{i}}$ on this graph. For this example, we have $Y_{\mathbf{i}} = Y_{12}Y_{22}Y_{23}Y_{35}Y_{52}Y_{24}Y_{41}Y_{14}Y_{42}Y_{21}$, which we can interpret as the walk $w_{\mathbf{i}}$ pictured below, in Figure 2. By definition, the walk $w_{\mathbf{i}}$ visits edge of $G_{\mathbf{i}}$ (including any self-edges present), beginning and ending at the beginning vertex. In particular, this means that the graph $G_{\mathbf{i}}$ is connected. The walk $w_{\mathbf{i}}$ encodes a labeling of the edges: the number of times each edge is traversed. We will denote this statistic also as $w_{\mathbf{i}}(e)$ for each edge e. In the above example, $w_{\mathbf{i}}(e)$ is equal to 2 on the edges $e\in\{\{1,2\},\{1,4\},\{2,4\}\}$, and is equal to 1 for all the other edges

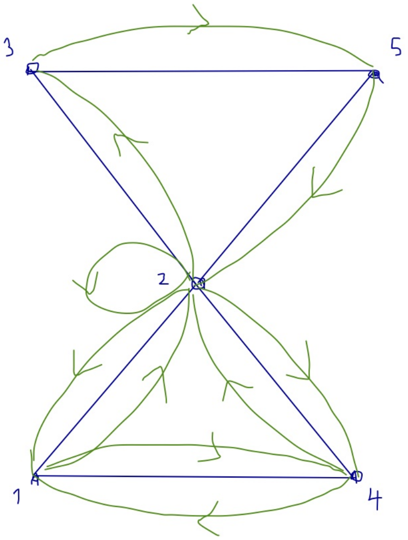


FIGURE 2. The graph and walk corresponding to the index $\mathbf{i} = (1,2,2,3,5,2,4,1,4,2)$.

(including the self-edge $\{2,2\}$). These numbers are actually evident in the expansion $Y_\mathbf{i}$: using the fact that $Y_{ij} = Y_{ji}$, we have

$$Y_\mathbf{i} = Y_{12}^2 Y_{14}^2 Y_{24}^2 Y_{22} Y_{23} Y_{25} Y_{35},$$

and in each case the exponent of Y_{ij} is equal to $w_\mathbf{i}(\{i,j\})$. This is true in general (essentially by definition): if we define $w_\mathbf{i}(\{i,j\}) = 0$ if the pair $\{i,j\} \notin E_\mathbf{i}$, then

$$Y_\mathbf{i} = \prod_{1\le i\le j\le n} Y_{ij}^{w_\mathbf{i}(\{i,j\})}. \tag{4.3}$$

Now, all the variables Y_{ij} are independent. Since we allow the diagonal vs. off-diagonal terms to have different distributions, we should make a distinction between the *self*-edges $E_\mathbf{i}^s = \{\{i,i\} \in E_\mathbf{i}\}$ and *connecting*-edges $E_\mathbf{i}^c = \{\{i,j\} \in E_\mathbf{i} \colon i \ne j\}$. Then we have

$$\mathbb{E}(Y_\mathbf{i}) = \prod_{1\le i\le j\le n} \mathbb{E}(Y_{ij}^{w_\mathbf{i}(\{i,j\})}) = \prod_{e_s\in E_\mathbf{i}^s} \mathbb{E}(Y_{11}^{w_\mathbf{i}(e_s)}) \cdot \prod_{e_c\in E_\mathbf{i}^c} \mathbb{E}(Y_{12}^{w_\mathbf{i}(e_c)}). \tag{4.4}$$

That is: the value of $\mathbb{E}(Y_\mathbf{i})$ is determined by the pair $(G_\mathbf{i}, w_\mathbf{i})$ (or better yet $(E_\mathbf{i}, w_\mathbf{i})$). Let us denote the common value in Equation 4.4 as $\Pi(G_\mathbf{i}, w_\mathbf{i})$.

For any k-index $\mathbf{i}$, the connected oriented graph $G_\mathbf{i}$ has at most k vertices. Since $w_\mathbf{i}$ records all the non-zero exponents in $Y_\mathbf{i}$ (that sum to the number of terms k), we have $|w_\mathbf{i}| \equiv \sum_{e\in E_\mathbf{i}} w_\mathbf{i}(e) = k$. Motivated by these conditions, let $\mathcal{G}_k$ denote the set of all pairs (G,w) where $G = (V,E)$ is a connected graph with at most k vertices, and w is a closed walk covering G satisfying $|w| = k$. Using Equation 4.4, we can reindex the sum of Equation 4.2 as

$$\mathbb{E}\,\mathrm{Tr}\,(\mathbf{Y}_n^k) = \sum_{(G,w)\in\mathcal{G}_k} \sum_{\substack{\mathbf{i}\in[n]^k \\ (G_\mathbf{i},w_\mathbf{i})=(G,w)}} \mathbb{E}(Y_\mathbf{i}) = \sum_{(G,w)\in\mathcal{G}_k} \Pi(G,w)\cdot \#\{\mathbf{i}\in[n]^k \colon (G_\mathbf{i},w_\mathbf{i}) = (G,w)\}. \tag{4.5}$$

Combining with the renormalization of Equation 4.1, we have

$$\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,(\mathbf{X}_n^k) = \sum_{(G,w)\in\mathcal{G}_k} \Pi(G,w)\cdot \frac{\#\{\mathbf{i}\in[n]^k \colon (G_\mathbf{i},w_\mathbf{i}) = (G,w)\}}{n^{k/2+1}}. \tag{4.6}$$

It is, in fact, quite easy to count the sets of k-indices in Equation 4.6. For any $(G, w) \in \mathcal{G}_k$, an index with that corresponding graph G and walk w is completely determined by assigning which distinct values from $[n]$ appear at the vertices of G. For example, consider the pair (G, w) in Figure 2. If $\mathbf{i} = (i_1, i_2, \ldots, i_k)$ is a k-index with this graph and walk, then reading along the walk we must have

$$\mathbf{i} = (i_1, i_2, i_2, i_3, i_5, i_2, i_4, i_1, i_4, i_2, i_1)$$

and so the set of all such $\mathbf{i}$ is determined by assigning 5 distinct values from $[n]$ to the indices i_1, i_2, i_3, i_4, i_5. There are $n(n-1)(n-2)(n-3)(n-4)$ ways of doing this. Following this example, in general we have:

Lemma 4.3. *Given $(G, w) \in \mathcal{G}_k$, denote by $|G|$ the number of vertices in G. Then*

$$\#\{\mathbf{i} \in [n]^k \colon (G_\mathbf{i}, w_\mathbf{i}) = (G, w)\} = n(n-1)\cdots(n-|G|+1).$$

Hence, Equation 4.6 can be written simply as

$$\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,(\mathbf{X}_n^k) = \sum_{(G,w)\in\mathcal{G}_k} \Pi(G, w) \cdot \frac{n(n-1)\cdots(n-|G|+1)}{n^{k/2}+1}. \tag{4.7}$$

The summation is finite (since k is fixed as $n \to \infty$), and so we are left to determined the values $\Pi(G, w)$. We begin with a simple observation: let $(G, w) \in \mathcal{G}_k$ and suppose there exists and edge $e = \{i, j\}$ with $w(e) = 1$. This means that, in the expression 4.4 for $\Pi(G, w)$, a singleton term $\mathbb{E}(Y_{ij}^{w(e)}) = \mathbb{E}(Y_{ij})$ appears. By assumption, the variables Y_{ij} are all centered, and so this term is 0; hence, the product $\Pi(G, w) = 0$ for any such pair (G, w). This reduces the sum in Equation 4.6 considerably, since we need only consider those w that cross each edge at least twice. We record this condition as $w \geq 2$, so

$$\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,(\mathbf{X}_n^k) = \sum_{\substack{(G,w)\in\mathcal{G}_k \\ w\geq 2}} \Pi(G, w) \cdot \frac{n(n-1)\cdots(n-|G|+1)}{n^{k/2+1}}. \tag{4.8}$$

The condition $w \geq 2$ restricts those graphs that can appear. Since $|w_\mathbf{i}| = k$, if each edge in $G_\mathbf{i}$ is traversed at least twice, this means that the number of edges is $\leq k/2$.

Exercise 4.3.1. *Let $G = (V, E)$ be a connected finite graph. Show that $|G| = \#V \leq \#E + 1$, and that $|G| = \#V = \#E + 1$ if and only if G is a plane tree.*

Hint: the claim is obvious for $\#V = 1$. The inequality can be proved fairly easily by induction. The equality case also follows by studying this induction (if G has a loop, removing the neighboring edges of a single vertex reduces the total number of edges by at least three).

Using Exercise 4.3.1, we see that for any graph $G = (V, E)$ appearing in the sum in Equation 4.8, $|G| \leq k/2 + 1$. The product $n(n-1)\cdots(n-|G|+1)$ is asymptotically equal to $n^{|G|}$, and so we see that the sequence $n \mapsto \frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,(\mathbf{X}_n^k)$ is bounded. What's more, suppose that k is odd. Since $|G| \leq \#E + 1 \leq k/2 + 1$ and $|G|$ is an integer, it follows that $|G| \leq (k-1)/2 + 1 = k/2 + 1/2$. Hence, in this case, all the terms in the (finite n-independent) sum in Equation 4.8 are $O(n^{-1/2})$. Thus, we have already proved that

$$\lim_{n\to\infty} \mathbb{E}\,\mathrm{Tr}\,(\mathbf{X}_n^k) = 0, \qquad k \text{ odd.}$$

Henceforth, we assume that k is even. In this case, it is still true that most of the terms in the sum in Equation 4.8 are 0. The following proposition testifies to this fact.

Proposition 4.4. *Let $(G,w) \in \mathcal{G}_k$ with $w \ge 2$.*

(a) *If there exists a self-edge $e \in E_s$ in G, then $|G| \le k/2$.*
(b) *If there exists an edge e in G with $w(e) \ge 3$, then $|G| \le k/2$.*

Proof. (a) Since the graph $G = (V,E)$ contains a loop, it is not a tree; it follows from Exercise 4.3.1 that $\#V < \#E + 1$. But $w \ge 2$ implies that $\#E \le k/2$, and so $\#V < k/2 + 1$, and so $|G| = \#V \le k/2$.

(b) The sum of w over all edges E in G is k. Hence, the sum of w over $E \setminus \{e\}$ is $\le k - 3$. Since $w \ge 2$, this means that the number of edges excepting e is $\le (k-3)/2$; hence, $\#E \le (k-3)/2 + 1 = (k-1)/2$. By the result of Exercise 4.3.1, this means that $\#V \le (k-1)/2 + 1 = (k+1)/2$. Since k is even, it follows that $|G| = \#V \le k/2$.

□

Combining proposition 4.4 with Equation 4.8 suggest we drastically refine the set $\mathcal{G}_k$. Set $\mathcal{G}_k^{k/2+1}$ to be the set of pairs $(G,w) \in \mathcal{G}_k$ where G has $k/2+1$ vertices, contains no self-edges, and the walk w crosses every edge exactly 2 times. Then

$$\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,(\mathbf{X}_n^k) = \sum_{(G,w)\in\mathcal{G}_k^{k/2+1}} \Pi(G,w)\cdot\frac{n(n-1)\cdots(n-|G|+1)}{n^{k/2+1}} + O_k(n^{-1}), \tag{4.9}$$

where $O_k(n^{-1})$ means that the absolute difference of the terms is $\le B_k/n$ for some n-independent constant B_k. Since $|G| = k/2+1$ and $n(n-1)\cdots(n-k/2+1) \sim n^{k/2+1}$, it follows therefore that

$$\lim_{n\to\infty}\mathbb{E}\,\mathrm{Tr}\,(\mathbf{X}_n^k) = \sum_{(G,w)\in\mathcal{G}_k^{k/2+1}} \Pi(G,w).$$

Let $(G,w) \in \mathcal{G}_k^{k/2+1}$. Since w traverses each edge exactly twice, the number of edges in G is $k/2$. Since the number of vertices is $k/2+1$, Exercise 4.3.1 shows that G is a tree. In particular there are no self-edges (as we saw already in Proposition 4.4) and so the value of $\Pi(G,w)$ in Equation 4.4 is

$$\Pi(G,w) = \prod_{e_c\in E^c}\mathbb{E}(Y_{12}^{w(e_c)}) = \prod_{e_c\in E^c}\mathbb{E}(Y_{12}^2) = t^{\#E} = t^{k/2}. \tag{4.10}$$

Hence, we finally have

$$\lim_{n\to\infty}\mathbb{E}\,\mathrm{Tr}\,(\mathbf{X}_n^k) = t^{k/2}\cdot\#\mathcal{G}_k^{k/2+1}. \tag{4.11}$$

Finally, we must enumerate the set $\mathcal{G}_k^{k/2+1}$. To do so, we introduce another combinatorial structure: *Dyck paths*. Given a pair $(G,w) \in \mathcal{G}_k^{k/2+1}$, define a sequence $\mathbf{d} = \mathbf{d}(G,w) \in \{+1,-1\}^k$ recursively as follows. Let $d_1 = +1$. For $1 < j \le k$, if $w_j \notin \{w_1,\ldots,w_{j-1}\}$, set $d_j = +1$; otherwise, set $d_j = -1$; then $\mathbf{d}(G,w) = (d_1,\ldots,d_k)$. For example, with $k = 10$ suppose that G is the graph in Figure 3, with walk

$$w = (\{1,2\},\{2,3\},\{3,2\},\{2,4\},\{4,5\},\{5,4\},\{4,6\},\{6,4\},\{4,2\},\{2,1\}).$$

Then $\mathbf{d}(G,w) = (+1,+1,-1,+1,+1,-1,+1,-1,-1,-1)$. One can interpret $\mathbf{d}(G,w)$ as a lattice path by looking at the successive sums: set $P_0 = (0,0)$ and $P_j = (j, d_1 + \cdots + d_j)$ for $1 \le j \le k$; then the piecewise linear path connecting $P_0, P_1, \ldots, P_k$ is a lattice path. Since $(G,w) \in \mathcal{G}_k^2$, each edge appears exactly two times in w, meaning that the ± 1s come in pairs in $\mathbf{d}(G,w)$. Hence $d_1 + \cdots + d_k = 0$. What's more, for any edge e, the -1 assigned to its second

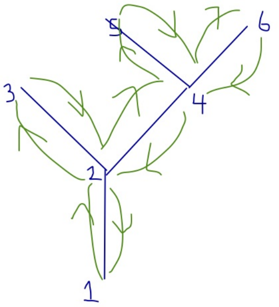


FIGURE 3. A graph with associated walk in $\mathcal{G}_{10}^2$.

appearance in w comes *after* the $+1$ corresponding to its first appearance; this means that the partial sums $d_1 + \cdots + d_j$ are all ≥ 0. That is: $\mathbf{d}(G,w)$ is a **Dyck path**.

Exercise 4.4.1. *Let k be even and let $\mathcal{D}_k$ denote the set of Dyck paths of length k*

$$\mathcal{D}_k = \{(d_1,\ldots,d_k) \in \{\pm 1\} \colon \sum_{i=1}^{k} d_i \geq 0 \, \text{for } 1 \leq j \leq j, \text{ and} \sum_{i=1}^{k} d_i = 0\}.$$

For $(G,w) \in \mathcal{G}_k^{k/2+1}$, the above discussion shows that $\mathbf{d}(G,w) \in \mathcal{D}_k$. Show that $(G,w) \mapsto \mathbf{d}(G,w)$ is a bijection $\mathcal{G}_k^{k/2+1} \to \mathcal{D}_k$ by finding its explicit inverse.

It is well-known that $\#D_k = C_{k/2}$ are enumerated by Catalan numbers. (For proof, see Stanley's Enumerative Combinatorics Vol. 2; or Lecture 7 from Math 247A Spring 2011, available at `www.math.ucsd.edu/~tkemp/247ASp11`.) Combining this with Equation 4.10 concludes the proof. □

4.2. **Convergence of Matrix Moments in Probability.** Now, to complete the proof of Theorem 2.4, we must now proceed with Step 2: we must show that $\frac{1}{n}\,\mathrm{Tr}\,(\mathbf{X}_n^k)$ has variance tending to 0 as $n \to \infty$. Using the ideas of the above proof, this is quite straightforward. We will actually show a little more: we will find the optimal rate that it tends to 0.

Proposition 4.5. *Let $\{Y_{ij}\}_{1\leq i\leq j}$ be independent random variables, with $\{Y_{ii}\}_{i\geq 1}$ identically distributed and $\{Y_{ij}\}_{1\leq i<j}$ identically distributed. Suppose that $r_k = \max\{\mathbb{E}(|Y_{11}|^k), \mathbb{E}(|Y_{12}|^k)\} < \infty$ for each $k \in \mathbb{N}$. Suppose further than $\mathbb{E}(Y_{ij}) = 0$ for all i,j. If $i > j$, define $Y_{ij} \equiv Y_{ji}$, and let $\mathbf{Y}_n$ be the $n \times n$ matrix with $[\mathbf{Y}_n]_{ij} = Y_{ij}$ for $1 \leq i,j \leq n$. Let $\mathbf{X}_n = n^{-1/2}\mathbf{Y}_n$ be the corresponding Wigner matrix. Then*

$$\mathrm{Var}\left(\frac{1}{n}\,\mathrm{Tr}\,(\mathbf{X}_n^k)\right) = O_k\left(\frac{1}{n^2}\right).$$

Again, to be precise, saying $F(k,n) = O_k(1/n^2)$ means that, for each k, there is a constant $B_k < \infty$ so that $F(k,n) \leq B_k/n^2$.

Proof. We proceed as above, expanding the variance in terms of the entries of the matrix.

$$\begin{aligned}\mathrm{Var}\left(\frac{1}{n}\,\mathrm{Tr}\,(\mathbf{X}_n^k)\right) &= \mathbb{E}\left[\frac{1}{n}\cdot\frac{1}{n^{k/2}}\,\mathrm{Tr}\,(\mathbf{Y}_n^k)\right]^2 - \left(\mathbb{E}\left[\frac{1}{n}\cdot\frac{1}{n^{k/2}}\,\mathrm{Tr}\,(\mathbf{Y}_n^k)\right]\right)^2 \\ &= \frac{1}{n^{k+2}}\left\{\mathbb{E}\left[\,\mathrm{Tr}\,(\mathbf{Y}_n^k)\right]^2 - \left(\mathbb{E}\,\mathrm{Tr}\,(\mathbf{Y}_n^k)\right)^2\right\}.\end{aligned}$$

Expanding the trace exactly as above, and squaring, gives

$$\mathbb{E}\left[\operatorname{Tr}(\mathbf{Y}_n^k)\right]^2 = \sum_{\mathbf{i},\mathbf{j}\in[n]^k} \mathbb{E}(Y_\mathbf{i} Y_\mathbf{j}), \quad \left(\mathbb{E}\operatorname{Tr}(\mathbf{Y}_n^k)\right)^2 = \sum_{\mathbf{i},\mathbf{j}\in[n]^k} \mathbb{E}(Y_\mathbf{i})\mathbb{E}(Y_\mathbf{j}).$$

Thus the variance in question is

$$\operatorname{Var}\left(\frac{1}{n}\operatorname{Tr}(\mathbf{X}_n^k)\right) = \frac{1}{n^{k+2}} \sum_{\mathbf{i},\mathbf{j}\in[n]^k} \left[\mathbb{E}(Y_\mathbf{i} Y_\mathbf{j}) - \mathbb{E}(Y_\mathbf{i})\mathbb{E}(Y_\mathbf{j})\right].$$

Now, as above, the values $\mathbb{E}(Y_\mathbf{i} Y_\mathbf{j})$ and $\mathbb{E}(Y_\mathbf{i})\mathbb{E}(Y_\mathbf{j})$ only depend on $\mathbf{i}$ and $\mathbf{j}$ through a certain graph structure underlying the $2k$-tuple $\mathbf{i},\mathbf{j}$. Indeed, let $\mathbf{G}_{\mathbf{i}\#\mathbf{j}} \equiv \mathbf{G}_\mathbf{i} \cup \mathbf{G}_\mathbf{j}$, where the union of two graphs (whose vertices are chosen from the same underlying set, in this case $[n]$) is the graph whose vertex set is the union of the vertex sets, and whose edge set is the union of the edge sets. For example, if $\mathbf{i} = (1,2,1,3,2,3)$ and $\mathbf{j} = (2,4,5,1,4,3)$, then

$$Y_\mathbf{i} Y_\mathbf{j} = Y_{12}Y_{21}Y_{13}Y_{32}Y_{23}Y_{31} \cdot Y_{24}Y_{45}Y_{51}Y_{14}Y_{43}Y_{32}$$

and so

$$G_{\mathbf{i}\#\mathbf{j}} = \left(\{1,2,3,4,5\}, \{\{1,2\},\{1,3\},\{2,3\},\{2,4\},\{4,5\},\{1,5\},\{1,4\},\{3,4\}\}\right).$$

The word $Y_\mathbf{i} Y_\mathbf{j}$ also gives rise to the two graph walks $w_\mathbf{i}$ and $w_\mathbf{j}$; as the above example shows, they do not constitute a single walk of length $2k$, since there is no a priori reason for the endpoint of the first walk to coincide with the starting point of the second walk.

Since we can recover $(G_\mathbf{i}, w_\mathbf{i})$ and $(G_\mathbf{j}, w_\mathbf{j})$ from $(G_{\mathbf{i}\#\mathbf{j}}, w_\mathbf{i}, w_\mathbf{j})$, the same reasoning as in the previous proposition shows that the quantity $\mathbb{E}(Y_\mathbf{i} Y_\mathbf{j}) - \mathbb{E}(Y_\mathbf{i})\mathbb{E}(Y_\mathbf{j})$ is determined by the data $(G_{\mathbf{i}\#\mathbf{j}}, w_\mathbf{i}, w_\mathbf{j})$. Denote this common value as

$$\mathbb{E}(Y_\mathbf{i} Y_\mathbf{j}) - \mathbb{E}(Y_\mathbf{i})\mathbb{E}(Y_\mathbf{j}) = \pi(G_{\mathbf{i}\#\mathbf{j}}, w_\mathbf{i}, w_\mathbf{j}).$$

Let $\mathcal{G}_{k,k}$ be the set of connected graphs G with $\le 2k$ vertices, together with two paths each of length k whose union covers G. This means we can expand (as above)

$$\operatorname{Var}\left(\frac{1}{n}\operatorname{Tr}(\mathbf{X}_n^k)\right) = \frac{1}{n^{k+2}} \sum_{(G,w,w')\in\mathcal{G}_{k,k}} \pi(G,w,w')\cdot\#\left\{(\mathbf{i},\mathbf{j})\in[n]^{2k} \colon (G_{\mathbf{i}\#\mathbf{j}}, w_\mathbf{i}, w_\mathbf{j}) = (G,w,w')\right\}.$$

As before, let $E^s_{\mathbf{i}\#\mathbf{j}}$ denote the self edges and $E^c_{\mathbf{i}\#\mathbf{j}}$ the connecting edges in $G_{\mathbf{i}\#\mathbf{j}}$. For any edge e in $G_{\mathbf{i}\#\mathbf{j}}$, let $w_{\mathbf{i}\#\mathbf{j}}(e)$ denote the number of times the edge e is traversed by *either* of the two paths $w_\mathbf{i}$ and $w_\mathbf{j}$. Then, as in (4.4),

$$\begin{aligned}\mathbb{E}(Y_\mathbf{i} Y_\mathbf{j}) - \mathbb{E}(Y_\mathbf{i})\mathbb{E}(Y_\mathbf{j}) &= \prod_{e_s\in E^s_{\mathbf{i}\#\mathbf{j}}} \mathbb{E}(Y_{11}^{w_{\mathbf{i}\#\mathbf{j}}(e_s)}) \cdot \prod_{e_c\in E^c_{\mathbf{i}\#\mathbf{j}}} \mathbb{E}(Y_{12}^{w_{\mathbf{i}\#\mathbf{j}}(e_c)}) \\ &\quad - \prod_{e_s\in E^s_\mathbf{i}} \mathbb{E}(Y_{11}^{w_\mathbf{i}(e_s)}) \cdot \prod_{e_c\in E^c_\mathbf{i}} \mathbb{E}(Y_{12}^{w_\mathbf{i}(e_c)}) \cdot \prod_{e_s\in E^s_\mathbf{j}} \mathbb{E}(Y_{11}^{w_\mathbf{j}(e_s)}) \cdot \prod_{e_c\in E^c_\mathbf{j}} \mathbb{E}(Y_{12}^{w_\mathbf{j}(e_c)}).\end{aligned}$$

The key thing is to note that

$$\sum_{e\in G_{\mathbf{i}\#\mathbf{j}}} w_{\mathbf{i}\#\mathbf{j}}(e) = 2k = \sum_{e\in G_\mathbf{i}} w_\mathbf{i}(e) + \sum_{e\in G_\mathbf{j}} w_\mathbf{j}(e);$$

that is, the sum of all exponents in either term $\mathbb{E}(Y_\mathbf{i} Y_\mathbf{j})$ or $\mathbb{E}(Y_\mathbf{i})\mathbb{E}(Y_\mathbf{j})$ is the total length of the words, which is $2k$. Thus, each of these terms is bounded (in modulus) above by something of the form $r_{m_1}\cdots r_{m_\ell}$ for some positive integers $m_1,\ldots m_\ell$ for which $m_1+\cdots+m_\ell = 2k$. (Recall that

$r_m = \max\{\mathbb{E}(|Y_{11}|^m), \mathbb{E}(|Y_{12}|^m)\}$.) There are only finitely many such integer partitions of $2k$, and so the maximum M_{2k} over all of them is finite. We therefore have the (blunt) upper bound

$$|\pi(G_{\mathbf{i}\#\mathbf{j}}, w_{\mathbf{i}}, w_{\mathbf{j}})| = |\mathbb{E}(Y_{\mathbf{i}}Y_{\mathbf{j}}) - \mathbb{E}(Y_{\mathbf{i}})\mathbb{E}(Y_{\mathbf{j}})| \le 2M_{2k}, \quad \forall\, \mathbf{i}, \mathbf{j} \in [n]^k.$$

That being said, we can (as in the previous proposition) show that many of these terms are identically 0. Indeed, following the reasoning from above: by construction, every edge in the joined graph $G_{\mathbf{i}\#\mathbf{j}}$ is traversed at least once by the union of the two paths $w_{\mathbf{i}}$ and $w_{\mathbf{j}}$. Suppose that e is an edge that is traversed only *once*. This means that $w_{\mathbf{i}\#\mathbf{j}}(e) = 1$, and so it follows that the two values $w_{\mathbf{i}}(e), w_{\mathbf{j}}(e)$ are $\{0, 1\}$. Hence, the above expansion and the fact that $\mathbb{E}(Y_{11}) = \mathbb{E}(Y_{12}) = 0$ show that $\pi(G_{\mathbf{i}\#\mathbf{j}}, w_{\mathbf{i}}, w_{\mathbf{j}}) = 0$ in this case. If we think if a path as a labeling of edges (counting the number of times an edge is traversed), then this means the variance sum reduces to

$$\mathrm{Var}\left(\frac{1}{n}\,\mathrm{Tr}\,(\mathbf{X}_n^k)\right) = \sum_{\substack{(G,w,w')\in\mathcal{G}_{k,k}\\ w+w'\ge 2}} \pi(G, w, w') \cdot \frac{\#\left\{(\mathbf{i}, \mathbf{j}) \in [n]^{2k} \colon (G_{\mathbf{i}\#\mathbf{j}}, w_{\mathbf{i}}, w_{\mathbf{j}}) = (G, w, w')\right\}}{n^{k+2}}.$$

The enumeration of the number of $2k$-tuples yielding a certain graph with two walks is the same as in the previous proposition: the structure (G, w, w') specifies the $2k$-tuple precisely once we select the $|G|$ distinct indices for the vertices. So, as before, this ratio becomes

$$\frac{n(n-1)\cdots(n-|G|+1)}{n^{k+2}}.$$

Now, we have the condition $w + w' \ge 2$, meaning every edge is traversed at least twice. Since there are k steps in each of the two paths, this means there are at most k edges. Appealing again to Exercise 4.3.1, it follows that $|G| \le k + 1$. Hence, $n(n-1)\cdots(n-|G|+1) \le n^{|G|} \le n^{k+1}$, and so we have already proved that

$$\mathrm{Var}\left(\frac{1}{n}\,\mathrm{Tr}\,(\mathbf{X}_n^k)\right) \le \sum_{\substack{(G,w,w')\in\mathcal{G}_{k,k}\\ w+w'\ge 2}} \pi(G, w, w') \cdot \frac{n^{k+1}}{n^{k+2}} \le \frac{1}{n} \cdot 2M_{2k} \cdot \#\mathcal{G}_{k,k}.$$

The (potentially enormous) number $B_k = 2M_{2k} \cdot \#\mathcal{G}_{k,k}$ is independent of n, and so we have already proved that the variance is $= O_k(1/n)$, which is enough to prove Theorem 2.4 (convergence in probability).

We will go one step further and show that there are also no terms with $|G| = k + 1$. From Exercise 4.3.1 again, since there are at most k edges, it follows that there are exactly k edges and so G is a tree; this means that $w + w' = 2$, every edge is traversed exactly two times. It then follows that $G_{\mathbf{i}}$ and $G_{\mathbf{j}}$ share no edges in common. Indeed, suppose e is a shared edge. Each walk $w_{\mathbf{i}}, w_{\mathbf{j}}$ traverses all edges of its graph $G_{\mathbf{i}}, G_{\mathbf{j}}$. Since $w_{\mathbf{i}}(e) + w_{\mathbf{j}}(e) = 2$, the values $(w_{\mathbf{i}}(e), w_{\mathbf{j}}(e))$ are either $(2, 0)$, $(1, 1)$, or $(0, 2)$. The first and last cases are impossible: in the first case, this would mean that $w_{\mathbf{j}}$ does not traverse e, which is an edge in $G_{\mathbf{j}}$, contradicting the definition of $w_{\mathbf{j}}$. But $(1, 1)$ is also impossible: since the union graph $G_{\mathbf{i}\#\mathbf{j}}$ is a tree, each subgraph $G_{\mathbf{i}}$ and $G_{\mathbf{j}}$ is a tree, and since the walks $w_{\mathbf{i}}$ and $w_{\mathbf{j}}$ cover each edge and return to their starting points, each edge must be traversed an even number of times (as there are no loops).

Hence, we see that the only graph walks $(G, w, w') \in \mathcal{G}_{k,k}$ with $|G| = k + 1$ must have the edge sets covered by w and w' distinct. In other words, if $(G_{\mathbf{i}\#\mathbf{j}}, w_{\mathbf{i}}, w_{\mathbf{j}}) = (G, w, w')$, then the edge sets do not intersect: $\{\{i_1, i_2\}, \ldots, \{i_k, i_1\}\} \cap \{\{j_1, j_2\}, \ldots, \{j_k, j_1\}\} = \varnothing$. But that means that the products $Y_{\mathbf{i}}$ and $Y_{\mathbf{j}}$ are independent, and so $\pi(G_{\mathbf{i}\#\mathbf{j}}, w_{\mathbf{i}}, w_{\mathbf{j}}) = \mathbb{E}(Y_{\mathbf{i}}Y_{\mathbf{j}}) - \mathbb{E}(Y_{\mathbf{i}})\mathbb{E}(Y_{\mathbf{j}}) = 0$. Thus, we have shown that, for any $(G,w,w')\in\mathcal{G}_{k,k}$ with $k+1$ vertices, $\pi(G,w,w')=0$. So, we actually have

$$\mathrm{Var}\left(\frac{1}{n}\,\mathrm{Tr}\,(\mathbf{X}_n^k)\right)=\sum_{\substack{(G,w,w')\in\mathcal{G}_{k,k}\\ w+w'\geq 2,|G|\leq k}}\pi(G,w,w')\cdot\frac{\#\left\{(\mathbf{i},\mathbf{j})\in[n]^{2k}\colon (G_{\mathbf{i}\#\mathbf{j}},w_{\mathbf{i}},w_{\mathbf{j}})=(G,w,w')\right\}}{n^{k+2}},$$

and so the same counting argument now bounded the variance by

$$\mathrm{Var}\left(\frac{1}{n}\,\mathrm{Tr}\,(\mathbf{X}_n^k)\right)\leq\sum_{\substack{(G,w,w')\in\mathcal{G}_{k,k}\\ w+w'\geq 2,|G|\leq k}}\pi(G,w,w')\cdot\frac{n^k}{n^{k+2}}\leq\frac{1}{n^2}\cdot 2M_{2k}\cdot\#\mathcal{G}_{k,k}.$$

This proves the proposition. □

Remark 4.6. We showed not only that $\mathrm{Var}[\frac{1}{n}\,\mathrm{Tr}\,(\mathbf{X}_n^k)]\to 0$ as $n\to\infty$, giving convergence in probability by Chebyshev's inequality; we actually proved that the variance is $O(1/n^2)$. This is relevant because $\sum_n\frac{1}{n^2}<\infty$. It therefore follows from the Borel-Cantelli lemma that

$$\frac{1}{n}\,\mathrm{Tr}\,(\mathbf{X}_n^k)\to\int x^k\,\sigma_t(dx)\ a.s.$$

as $n\to\infty$, with the caveat that this almost sure convergence only makes sense if we artificially sample all entries of all matrices from the same probability space.

4.3. **Weak Convergence.**

We have now proved Theorem 2.4, which was a weaker form of Theorem 2.3. In fact, we can now fairly easily prove this stronger theorem. Let us first remark that we may easily fix the variance of the entries t to be $t=1$: an elementary scaling argument then extends Theorem 2.3 to the general case. With this convention in hand, we begin with the following useful cutoff lemma.

Lemma 4.7. *Let $k\in\mathbb{N}$ and $\epsilon>0$. Then for any $b>4$,*

$$\limsup_{n\to\infty}\mathbb{P}\left(\int_{|x|>b}|x|^k\,\mu_{\mathbf{X}_n}(dx)>\epsilon\right)=0.$$

Proof. First, by Markov's inequality, we have

$$\mathbb{P}\left(\int_{|x|>b}|x|^k\,\mu_{\mathbf{X}_n}(dx)>\epsilon\right)\leq\frac{1}{\epsilon}\mathbb{E}\left(\int_{|x|>b}|x|^k\,\mu_{\mathbf{X}_n}(dx)\right).$$

Now, let ν be the random measure $\nu(dx)=|x|^k\,\mu_{\mathbf{X}_n}(dx)$. Markov's inequality (which applies to all positive measures, not just probability measures) now shows that

$$\int_{|x|>b}|x|^k\,\mu_{\mathbf{X}_n}(dx)=\nu\{x\colon|x|>b\}=\nu\{x\colon|x|^k>b^k\}\leq\frac{1}{b^k}\int|x|^k\,\nu(dx)=\frac{1}{b^k}\int x^{2k}\,\mu_{\mathbf{X}_n}(dx).$$

So, taking expectations, we have

$$\mathbb{P}\left(\int_{|x|>b}|x|^k\,\mu_{\mathbf{X}_n}(dx)>\epsilon\right)\leq\frac{1}{\epsilon b^k}\mathbb{E}\left(\int x^{2k}\,\mu_{\mathbf{X}_n}(dx)\right)=\frac{1}{\epsilon b^k}\cdot\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,(\mathbf{X}_n^k).$$

Now, by Proposition 4.1, the right hand side converges to $C_k/\epsilon b^k$ where C_k is the Catalan number, which is bounded by 4^k. Hence, it follows that

$$\limsup_{n\to\infty}\mathbb{P}\left(\int_{|x|>b}|x|^k\,\mu_{\mathbf{X}_n}(dx)>\epsilon\right)\leq\frac{1}{\epsilon}\left(\frac{4}{b}\right)^k.$$

On the other hand, when $|x| > b > 4 > 1$, the function $k \mapsto |x|^k$ is strictly increasing, which means that the sequence of $\limsup$s on the left-hand-side is increasing. But this sequence decays exponentially since $4/b < 1$. The only way this is possible is if the sequence of $\limsup$s is constantly 0, as claimed. □

Proof of Theorem 2.3. Fix a bounded continuous function $f \in C_b(\mathbb{R})$, fix $\epsilon > 0$, and fix $b > 4$. By the Weierstrass approximation theorem, there is a polynomial P_ϵ such that

$$\sup_{|x|\le b} |f(x) - P_\epsilon(x)| < \frac{\epsilon}{6}.$$

Now, we have the triangle inequality estimates

$$\left|\int f\, d\mu_{\mathbf{X}_n} - \int f\, d\sigma_1\right| \le \left|\int f\, d\mu_{\mathbf{X}_n} - \int P_\epsilon\, d\mu_{\mathbf{X}_n}\right| + \left|\int P_\epsilon\, d\mu_{\mathbf{X}_n} - \int P_\epsilon\, d\sigma_1\right| \\ + \left|\int P_\epsilon\, d\sigma_1 - \int f\, d\sigma_1\right|.$$

Hence, the event $\{\left|\int f\, d\mu_{\mathbf{X}_n} - \int f\, d\sigma_1\right| > \epsilon\}$ is contained in the union of the three events that each of the four above terms is $> \epsilon/3$. But this means that

$$\mathbb{P}\left(\left|\int f\, d\mu_{\mathbf{X}_n} - \int f\, d\sigma_1\right| > \epsilon\right) \le \mathbb{P}\left(\left|\int f\, d\mu_{\mathbf{X}_n} - \int P_\epsilon\, d\mu_{\mathbf{X}_n}\right| > \epsilon/3\right) \\ + \mathbb{P}\left(\left|\int P_\epsilon\, d\mu_{\mathbf{X}_n} - \int P_\epsilon\, d\sigma_1\right| > \epsilon/3\right) \\ + \mathbb{P}\left(\left|\int P_\epsilon\, d\sigma_1 - \int f\, d\sigma_1\right| > \epsilon/3\right).$$

By construction, $|P_\epsilon - f| < \epsilon/6$ on $[-b, b]$, which includes the support $[-2, 2]$ of σ_1; thus, the last term is identically 0. For the first term, we break up the integral over $[-b, b]$ and its complement:

$$\left|\int (f - P_\epsilon)\, d\mu_{\mathbf{X}_n}\right| \le \int_{|x|\le b} |f(x) - P_\epsilon(x)|\, \mu_{\mathbf{X}_n}(dx) + \int_{|x|>b} |f(x) - P_\epsilon(x)|\, \mu_{\mathbf{X}_n}(dx).$$

By the same reasoning as above, we can estimate

$$\mathbb{P}\left(\left|\int f\, d\mu_{\mathbf{X}_n} - \int P_\epsilon\, d\mu_{\mathbf{X}_n}\right| > \epsilon/3\right) \le \mathbb{P}\left(\int |f - P_\epsilon| \mathbb{1}_{|x|\le b}\, d\mu_{\mathbf{X}_n} > \epsilon/6\right) \\ + \mathbb{P}\left(\int |f - P_\epsilon| \mathbb{1}_{|x|>b}\, d\mu_{\mathbf{X}_n} > \epsilon/6\right).$$

Again, by construction, $|f - P_\epsilon| < \epsilon/6$ on $[-b, b]$, and so since $\mu_{\mathbf{X}_n}$ is a probability measure, the first term is identically 0. We therefore have the estimate

$$\mathbb{P}\left(\left|\int f\, d\mu_{\mathbf{X}_n} - \int f\, d\sigma_1\right| > \epsilon\right) \le \mathbb{P}\left(\int |f - P_\epsilon| \mathbb{1}_{|x|>b}\, d\mu_{\mathbf{X}_n} > \epsilon/6\right) \tag{4.12}$$

$$+ \mathbb{P}\left(\left|\int P_\epsilon\, d\mu_{\mathbf{X}_n} - \int P_\epsilon\, d\sigma_1\right| > \epsilon/3\right). \tag{4.13}$$

That the terms in (4.13) tend to 0 as $n \to \infty$ follows immediately from Theorem 2.4 (since convergence in probability respects addition and scalar multiplication). So we are left only to estimate (4.12). We do this as follows. Let $k = \deg P_\epsilon$. Since f is bounded, we have $|f(x) -$

$P_\epsilon(x)| \le \|f\|_\infty + |P_\epsilon(x)|$, and on the set $|x| > b$, this is $\le c|x|^k$ for some constant c (since $b > 0$). This means that

$$\mathbb{P}\left(\int |f - P_\epsilon|\, 1_{|x|\ge b}\, d\mu_{\mathbf{X}_n} > \epsilon/6\right) \le \mathbb{P}\left(\int c|x|^k \mathbb{1}_{|x|\ge b}\, \mu_{\mathbf{X}_n}(dx) > \epsilon/6\right),$$

and therefore, by Lemma 4.7, this sequence has $\limsup_{n\to\infty} = 0$. This concludes the proof. □

Remark 4.8. From (4.13) and our proof of Theorem 2.4, we actually have decay rate $O_k(1/n^2)$ for (4.13). A much more careful analysis of the proof of Lemma 4.7 shows that (4.12) also decays summably fast, and so we can actually conclude almost sure weak convergence in general. We will reprove this result in a different way later in these notes.

Finally, this brings us to the proof of Theorem 2.1.

Proof of Theorem 2.1. The random variable $E_n(I)$ in the statement of the theorem can be written as

$$E_n(I) = \frac{1}{n} \cdot \#\{j \in [n] \colon \lambda_j(\mathbf{X}_n) \in I\} = \mu_{\mathbf{X}_n}(I).$$

That is to say, the desired conclusion is that $\mu_{\mathbf{X}_n}(I) \to \sigma_t(I)$ in probability for all intervals I. By an easy scaling argument, we can take $t = 1$, in which case we have proved above that $\mu_{\mathbf{X}_n} \to \sigma_1$ weakly in probability. By a standard convergence theorem in probability theory (c.f. Theorem 25.8 in Billingsley's "Probability and Measure" (Third Edition)), this implies that $\mu_{\mathbf{X}_n}(A) \to \sigma_1(A)$ in probability for every σ_1-continuous measurable seat A. But since σ_1 has a continuous density, *every* measurable set is a σ_1-continuous set. This proves the theorem. □

5. Removing Moment Assumptions

In the proof we presented of Wigner's theorem, we were forced to assume that all the moments of the i.i.d. random variables Y_{11} and Y_{12} are finite. In this section, we concern ourselves with removing these unnecessary assumptions; all that is truly required is $r_2 = \max\{\mathbb{E}(Y_{11}^2), \mathbb{E}(Y_{12}^2)\} < \infty$. To that end, the idea is as follows: begin with a Wigner matrix $\mathbf{X}_n = n^{-1/2}\mathbf{Y}_n$. We will find an approximating Wigner matrix $\hat{\mathbf{X}}_n = n^{-1/2}\hat{\mathbf{Y}}_n$ with entries having all moments finite, such that the empirical laws $\mu_{\mathbf{X}_n}$ and $\mu_{\hat{\mathbf{X}}_n}$ are "close". To be a little more precise: for simplicity, let us standardize so that we may assume that the off-diagonal entries of $\mathbf{Y}_n$ have unit variance. Our goal is to show that $\int f\, d\mu_{\mathbf{X}_n} \to \int f\, d\sigma_1$ in probability, for each $f \in C_b(\mathbb{R})$. To that end, fix $\epsilon > 0$ from the outset. For any approximating Wigner matrix $\hat{\mathbf{X}}_n$, if we may arrange that $\left|\int f\, d\mu_{\mathbf{X}_n} - \int f\, d\mu_{\hat{\mathbf{X}}_n}\right| \le \epsilon/2$ and $\left|\int f\, d\mu_{\hat{\mathbf{X}}_n} - \int f\, d\sigma_1\right| \le \epsilon/2$, then by the triangle inequality $\left|\int f\, d\mu_{\mathbf{X}_n} - \int f\, d\sigma_1\right| \le \epsilon$. The contrapositive of this implication says that

$$\begin{aligned}&\left\{\left|\int f\, d\mu_{\mathbf{X}_n} - \int f\, d\sigma_1\right| < \epsilon\right\}\\ \subseteq &\left\{\left|\int f\, d\mu_{\mathbf{X}_n} - \int f\, d\mu_{\hat{\mathbf{X}}_n}\right| > \epsilon/2\right\} \cup \left\{\left|\int f\, d\mu_{\hat{\mathbf{X}}_n} - \int f\, d\sigma_1\right| > \epsilon/2\right\}.\end{aligned}$$

Hence

$$\begin{aligned}&\mathbb{P}\left(\left|\int f\, d\mu_{\mathbf{X}_n} - \int f\, d\sigma_1\right| < \epsilon\right)\\ \le &\mathbb{P}\left(\left|\int f\, d\mu_{\mathbf{X}_n} - \int f\, d\mu_{\hat{\mathbf{X}}_n}\right| > \epsilon/2\right) + \mathbb{P}\left(\left|\int f\, d\mu_{\hat{\mathbf{X}}_n} - \int f\, d\sigma_1\right| > \epsilon/2\right).\end{aligned} \tag{5.1}$$

Since $\mathbf{X}_n$ is a Wigner matrix possessing all finite moments, we know that the second term above tends to 0. Hence, in order to prove that $\mu_{\mathbf{X}_n} \to \sigma_1$ weakly in probability, it suffices to show that we can find an approximating Wigner matrix $\hat{\mathbf{X}}_n$ with all finite moments such that $\int f\, d\mu_{\mathbf{X}_n} - \int f\, d\mu_{\hat{\mathbf{X}}_n}$ is small (uniformly in n) for given $f \in C_b(\mathbb{R})$; this is the sense of "close" we need. Actually, it will turn out that we can find such an approximating Wigner matrix provided we narrow the scope of the test functions f a little bit.

Definition 5.1. *A function $f \in C_b(\mathbb{R})$ is said to be* **Lipschitz** *if*

$$\|f\|_{\mathrm{Lip}} \equiv \sup_{x\neq y} \frac{|f(x)-f(y)|}{|x-y|} + \sup_x |f(x)| < \infty.$$

The set of Lipschitz functions is denoted $\mathrm{Lip}(\mathbb{R})$. *The quantity* $\|\cdot\|_{\mathrm{Lip}}$ *is a norm on* $\mathrm{Lip}(\mathbb{R})$, *and makes it a Banach space.*

Lipschitz functions are far more regular than generic continuous functions. (For example, they are differentiable almost everywhere.) Schemes like the Weierstraß approximation theorem can be used to approximate generic continuous functions by Lipschitz functions; as such, restricting test functions to be Lipschitz still metrizes weak convergence. We state this standard theorem here without proof.

Proposition 5.2. *Suppose μ_n, ν_n are sequences of measures on $\mathbb{R}$. Suppose that $\int f\, d\mu_n - \int f\, d\nu_n \to 0$ for each $f \in \mathrm{Lip}(\mathbb{R})$. Then $\int f\, d\mu_n - \int f\, d\nu_n \to 0$ for each $f \in C_b(\mathbb{R})$.*

Thus, we will freely assume the test functions are Lipschitz from now on. This is convenient, due to the following.

Lemma 5.3. *Let A, B be symmetric $n \times n$ matrices, with eigenvalues $\lambda_1^A \le \cdots \le \lambda_n^A$ and $\lambda_1^B \le \cdots \le \lambda_n^B$. Denote by μ_A and μ_B the empirical laws of these eigenbalues. Let $f \in \mathrm{Lip}(\mathbb{R})$. Then*

$$\left|\int f\,d\mu_A - \int f\,d\mu_B\right| \le \|f\|_{\mathrm{Lip}} \left(\frac{1}{n}\sum_{i=1}^n (\lambda_i^A - \lambda_i^B)^2\right)^{1/2}.$$

Proof. By definition $\int f\,d\mu_A - \int f\,d\mu_B = \frac{1}{n}\sum_{i=1}^n [f(\lambda_i^A) - f(\lambda_i^B)]$. Hence we have the straightforward estimate

$$\left|\int f\,d\mu_A - \int f\,d\mu_B\right| \le \frac{1}{n}\sum_{i=1}^n |f(\lambda_i^A) - f(\lambda_i^B)| = \frac{1}{n}\sum_{i=1}^n \|f\|_{\mathrm{Lip}}|\lambda_i^A - \lambda_i^B|,$$

where we have used the fact that $|f(x) - f(y)| \le \|f\|_{\mathrm{Lip}}|x-y|$. (This inequality does not require the term $\sup_x |f(x)|$ in the definition of $\|f\|_{\mathrm{Lip}}$; this term is included to make $\|\cdot\|_{\mathrm{Lip}(\mathbb{R})}$ into a norm, for without it all constant functions would have "norm" 0.) The proof is completed by noting the equivalence of the ℓ^1 and ℓ^2 norms on $\mathbb{R}^n$: for any vector $\mathbf{v} = [v_1, \ldots, v_n] \in \mathbb{R}^n$,

$$\|\mathbf{v}\|_1 = \sum_{i=1}^n |v_i| = [1, 1, \ldots, 1] \cdot [|v_1|, |v_2|, \ldots, |v_n|] \le \|[1, \ldots, 1]\|_2 \|\mathbf{v}\|_2 = \sqrt{n}\|\mathbf{v}\|_2.$$

Hence

$$\frac{1}{n}\sum_{i=1}^n |\lambda_i^A - \lambda_i^B| \le \frac{1}{\sqrt{n}}\left(\sum_{i=1}^n |\lambda_i^A - \lambda_i^B|^2\right)^{1/2}.$$

□

The quantity on the right-hand-side of the estimate in Lemma 5.3 can be further estimated by a simple expression in terms of the matrices A and B.

Lemma 5.4 (Hoffman–Wielandt)**.** *Let A, B be $n \times n$ symmetric matrices, with eigenvalues $\lambda_1^A \le \cdots \le \lambda_n^A$ and $\lambda_1^B \le \cdots \le \lambda_n^B$. Then*

$$\sum_{i=1}^n (\lambda_i^A - \lambda_i^B)^2 \le \mathrm{Tr}\,[(A-B)^2].$$

Proof. Begin by diagonalizing $A = U^\top \Lambda^A U$ and $B = V^\top \Lambda^B V$ where U, V are orthogonal matrices (with columns that are normalized eigenvectors of A and B) and Λ^A, Λ^B diagonal with $[\Lambda^A]_{jj} = \lambda_j^A$ and $[\Lambda^B]_{jj} = \lambda_j^B$. Thus

$$\mathrm{Tr}\,(AB) = \mathrm{Tr}\,(U^\top \Lambda^A U V^\top \Lambda^B V) = \mathrm{Tr}\,[(VU^\top)\Lambda^A (UV^\top)\Lambda^B].$$

Set $W = UV^\top$, which is also an orthogonal matrix. Then

$$\begin{aligned}\mathrm{Tr}\,(AB) = \mathrm{Tr}\,(W^\top \Lambda^A W \Lambda^B) &= \sum_{1\le i,j,k,\ell \le n} [W^\top]_{ij}[\Lambda^A]_{jk}[W]_{k\ell}[\Lambda^B]_{\ell i} \\ &= \sum_{1\le i,j,k,\ell \le n} [W]_{ji}\lambda_j^A \delta_{jk}[W]_{k\ell}\lambda_i^B \delta_{\ell i} \\ &= \sum_{1 \le i,j \le n} \lambda_j^A \lambda_i^B [W]_{ji}^2.\end{aligned}$$

Now, W is an orthogonal matrix, which means that $\sum_i [W]_{ji}^2 = 1$ for each j and $\sum_j [W]_{ji}^2 = 1$ for each i. Set $\upsilon_{ji} = [W]_{ji}^2$, so that the matrix $[\upsilon_{ji}]_{j,i}$ is doubly-stochastic. Let $\mathcal{D}_n$ denote the set of doubly-stochastic $n \times n$ matrices; thus, we have

$$\mathrm{Tr}\,(AB) = \sum_{1\le i,j\le n} \lambda_j^A \lambda_i^B \upsilon_{ij} \le \sup_{[\nu_{ij}]\in\mathcal{D}_n} \sum_{1\le i,j\le n} \lambda_i^A \lambda_j^B \nu_{ij}.$$

The set $\mathcal{D}_n$ is convex (it is easily checked that any convex combination of doubly-stochastic matrices is doubly-stochastic). The function $[\nu_{ij}] \mapsto \sum_{i,j} \lambda_i^A \lambda_j^B \nu_{ij}$ is a linear function, and hence its supremum on the convex set $\mathcal{D}_n$ is achieved at an extreme point of $\mathcal{D}_n$.

Claim. The extreme points of $\mathcal{D}_n$ are the permutation matrices.

This is the statement of the Birkhoff-von Neumann theorem. For a simple proof, see `http://mingus.la.asu.edu/~hurlbert/papers/SPBVNT.pdf`. The idea is: if any row contains two non-zero entries, one can increase one and decrease the other preserving the row sum; then one can appropriately increase/decrease non-zero elements of those columns to preserve the column sums; then modify elements of the adjusted rows; continuing this way must result in a closed path through the entries of the matrix (since there are only finitely-many). In the end, one can then perturb the matrix to produce a small line segment staying in $\mathcal{D}_n$. The same observation works for columns; thus, extreme points must have exactly one non-zero entry in each row and column: hence, the extreme points are permutation matrices. Thus, we have

$$\mathrm{Tr}\,(AB) \le \max_{\sigma\in S_n} \sum_{i=1}^n \lambda_i^A \lambda_{\sigma(i)}^B.$$

Because the sequences λ_i^A and λ_i^B are non-decreasing, this maximum is achieved when σ is the identity permutation. To see why, first consider the case $n = 2$. Given $x_1 \le x_2$ and $y_1 \le y_2$, note that

$$(x_1y_1 + x_2y_2) - (x_1y_2 + x_2y_1) = x_1(y_1 - y_2) + x_2(y_2 - y_1) = (x_2 - x_1)(y_2 - y_1) \ge 0.$$

That is, $x_1y_2 + x_2y_1 \le x_1y_1 + x_2y_2$. Now, let σ be any permutation not equal to the identity. Then there is some pair $i < j$ with $\sigma(i) > \sigma(j)$. Let σ' be the new permutation with $\sigma'(i) = \sigma(j)$ and $\sigma'(j) = \sigma(i)$, and $\sigma = \sigma'$ on $[n] \setminus \{i, j\}$. Then the preceding argument shows that $\sum_{i=1}^n \lambda_i^A \lambda_{\sigma(i)}^B \le \sum_{i=1}^n \lambda_i^A \lambda_{\sigma'(i)}^B$. The permutation σ' has one fewer order-reversal than σ; iterating this process shows that $\sigma = \mathrm{id}$ maximizes the sum. In particular, taking negatives shows that

$$-\sum_{i=1}^n \lambda_i^A \lambda_i^B \le -\,\mathrm{Tr}\,(AB).$$

Finally, since $\mathrm{Tr}\,[A^2] = \sum_i (\lambda_i^A)^2$ and $\mathrm{Tr}\,[B^2] = \sum_i (\lambda_i^B)^2$, we have

$$\begin{aligned}\sum_{i=1}^n (\lambda_i^A - \lambda_i^B)^2 &= \sum_{i=1}^n (\lambda_i^A)^2 + \sum_{i=1}^n (\lambda_i^B)^2 - 2\sum_{i=1}^n \lambda_i^A \lambda_i^B \\ &= \mathrm{Tr}\,(A^2) + \mathrm{Tr}\,(B^2) - 2\sum_{i=1}^n \lambda_i^A \lambda_i^B \\ &\le \mathrm{Tr}\,(A^2) + \mathrm{Tr}\,(B^2) - 2\,\mathrm{Tr}\,(AB) = \mathrm{Tr}\,[(A-B)^2].\end{aligned}$$

□

Now, let $A = \mathbf{X}_n$ and let $B = \hat{\mathbf{X}}_n$ be an approximating Wigner matrix. Combining Lemmas 5.3 and 5.4, we have for any Lipschitz function f,

$$\left|\int f\,d\mu_{\mathbf{X}_n} - \int f\,d\mu_{\hat{\mathbf{X}}_n}\right| \le \|f\|_{\mathrm{Lip}} \left(\frac{1}{n}\mathrm{Tr}\,[(\mathbf{X}_n - \hat{\mathbf{X}}_n)^2]\right)^{1/2}. \tag{5.2}$$

Now fix $\delta > 0$. Then using Markov's inequality,

$$\mathbb{P}\left(\frac{1}{n}\mathrm{Tr}\,[(\mathbf{X}_n - \hat{\mathbf{X}}_n)^2] > \delta\right) \le \frac{1}{\delta}\cdot\mathbb{E}\left(\frac{1}{n}\mathrm{Tr}\,[(\mathbf{X}_n - \hat{\mathbf{X}}_n)^2]\right).$$

Letting $\hat{\mathbf{X}}_n = n^{-1/2}\hat{\mathbf{Y}}_n$ where $\hat{\mathbf{Y}}_n$ has entries $\hat{Y}_{ij}$, we can expand this expected trace as

$$\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,[(\mathbf{X}_n - \hat{\mathbf{X}}_n)^2] = \frac{1}{n^2}\mathbb{E}\,\mathrm{Tr}\,[(\mathbf{Y}_n - \hat{\mathbf{Y}}_n)^2] = \frac{1}{n^2}\sum_{1\le i,j\le n}\mathbb{E}[(Y_{ij} - \hat{Y}_{ij})^2].$$

Breaking the sum up in terms of diagonal and off-diagonal terms, and using identical distributions, we get

$$\begin{aligned} &\frac{1}{n^2}\sum_{i=1}^n \mathbb{E}[(Y_{ii} - \hat{Y}_{ii})^2] + \frac{2}{n^2}\sum_{1\le i<j\le n}\mathbb{E}[(Y_{ij} - \hat{Y}_{ij})^2] \\ =\; &\frac{1}{n}\mathbb{E}[(Y_{11} - \hat{Y}_{11})^2] + \left(1 - \frac{1}{n}\right)\mathbb{E}[(Y_{12} - \hat{Y}_{12})^2] \\ \le\; &\mathbb{E}[(Y_{11} - \hat{Y}_{11})^2] + \mathbb{E}[(Y_{12} - \hat{Y}_{12})^2]. \end{aligned}$$

So, using Equation 5.2 setting $\delta = (\epsilon/2\|f\|_{\mathrm{Lip}})^2$ we have

$$\begin{aligned} \mathbb{P}\left(\left|\int f\,d\mu_{\mathbf{X}_n} - \int f\,d\mu_{\hat{\mathbf{X}}_n}\right| > \epsilon/2\right) &\le \mathbb{P}\left(\|f\|_{\mathrm{Lip}}\left(\frac{1}{n}\mathrm{Tr}\,[(\mathbf{X}_n - \hat{\mathbf{X}}_n)^2]\right)^{1/2} > \epsilon/2\right) \\ &\le \frac{4\|f\|^2_{\mathrm{Lip}}}{\epsilon^2}\left(\mathbb{E}[(Y_{11} - \hat{Y}_{11})^2] + \mathbb{E}[(Y_{12} - \hat{Y}_{12})^2]\right). \end{aligned} \tag{5.3}$$

This latter estimate is uniform in n. We see, therefore, that it suffices to construct the approximating $\hat{\mathbf{Y}}_n$ in such a way that the entries are close in the L^2 sense: i.e. we must be able to choose Y_{ij} with $\mathbb{E}[(Y_{ij} - \hat{Y}_{ij})^2] < \epsilon^3/16\|f\|^2_{\mathrm{Lip}}$.

We are now in a position to define appropriate approximating Wigner matrices $\hat{\mathbf{X}}_n$. Fix a cut-off constant $C > 0$. Morally, we want to simply define the entries of the approximating matrix to be $Y_{ij}\mathbb{1}_{|Y_{ij}|\le C}$, which are bounded and hence have moments of all orders. This cut-off does not preserve mean or variance, though, so we must standardize. For $1 \le i,j \le n$, define

$$\hat{Y}_{ij} = \frac{1}{\sigma_{ij}(C)}\left(Y_{ij}\mathbb{1}_{|Y_{ij}|\le C} - \mathbb{E}(Y_{ij}\mathbb{1}_{|Y_{ij}|\le C})\right), \tag{5.4}$$

where $\sigma_{ij}(C)^2 = \mathrm{Var}\left(Y_{ij}\mathbb{1}_{|Y_{ij}|\le C}\right)$ when $i \ne j$ and $\sigma_{ii}(C) = 1$. Note: it is possible that, for small C and $i \ne j$, $\sigma_{ij}(C) = \sigma_{12}(C) = 0$, but it is > 0 for all sufficiently large C (and we assume C is large), so $\hat{Y}_{ij}$ is meaningful. Let $\hat{\mathbf{Y}}_n$ have entries $\hat{Y}_{ij}$. Of course, the $\hat{Y}_{ii}$ are all $i.i.d.$ as are the $\hat{Y}_{ij}$ with $i < j$. These entries are centered and the off-diagonal ones have unit variance; they are all bounded random variables, hence all moments are finite. Thus, setting $\hat{\mathbf{X}}_n = n^{-1/2}\hat{\mathbf{Y}}_n$, $\hat{\mathbf{X}}_n$ is a Wigner matrix with all moments finite, and so Theorem 2.3 (Wigner's Semicircle Law), even with moment assumptions, holds for $\hat{\mathbf{X}}_n$: for any $f \in C_b(\mathbb{R})$, $\int f\, d\mu_{\hat{\mathbf{X}}_n} \to \int f\, d\sigma_1$ in probability as $n \to \infty$.

Thus, combining Equation 5.1 with Inequality 5.3, we merely need to show that, for sufficiently large $C > 0$, the random variables $Y_{1j} - \hat{Y}_{1j}$ (for $j = 1, 2$) are small in L^2-norm. Well,

$$Y_{ij}\mathbb{1}_{|Y_{ij}|\le C} - \mathbb{E}(Y_{ij}\mathbb{1}_{|Y_{ij}|\le C}) = Y_{ij} - (Y_{ij}\mathbb{1}_{|Y_{ij}|>C} - \mathbb{E}(Y_{ij}\mathbb{1}_{|Y_{ij}|>C}))$$

where we have used the fact that $\mathbb{E}(Y_{ij}) = 0$ in the last equality. Hence

$$Y_{ij} - \hat{Y}_{ij} = \left(1 - \frac{1}{\sigma_{ij}(C)}\right) Y_{ij} + \frac{1}{\sigma_{ij}(C)}\left(Y_{ij}\mathbb{1}_{|Y_{ij}|>C} - \mathbb{E}(Y_{ij}\mathbb{1}_{|Y_{ij}|>C})\right). \tag{5.5}$$

The key point is that the random variable $Y_{ij}\mathbb{1}_{|Y_{ij}|>C}$ converges to 0 in L^2 as $C \to \infty$. This is because

$$\mathbb{E}\left[(Y_{ij}\mathbb{1}_{|Y_{ij}|>C})^2\right] = \int_{|Y_{ij}|>C} Y_{ij}^2\, d\mathbb{P},$$

and since $\mathbb{E}(Y_{ij}^2) < \infty$ (i.e. $Y_{ij}^2 \in L^1(\mathbb{P})$), it suffices to show that $\mathbb{P}(|Y_{ij}| > C) \to 0$ as $C \to \infty$. But this follows by Markov's inequality (which we apply to Y_{ij}^2 here just for æsthetic reasons):

$$\mathbb{P}(|Y_{ij}| > C) = \mathbb{P}(Y_{ij}^2 > C^2) \le \frac{\mathbb{E}(Y_{ij}^2)}{C^2}.$$

Thus, $Y_{ij}\mathbb{1}_{|Y_{ij}|>C} \to 0$ in L^2 as $n \to \infty$, and so $Y_{ij}\mathbb{1}_{|Y_{ij}|\le C} \to Y_{ij}$ in L^2 as $C \to \infty$. In particular, it then follows that $\sigma_{ij}(C) \to 1$ as $C \to \infty$. Convergence in $L^2(\mathbb{P})$ implies convergence in $L^1(\mathbb{P})$, and so $\mathbb{E}(Y_{ij}\mathbb{1}_{|Y_{ij}|>C}) \to 0$. Altogether, this shows that $Y_{ij} - \hat{Y}_{ij} \to 0$ in L^2, which completes the proof.

□

6. Largest Eigenvalue

In the previous section, we saw that we could remove the technical assumption that all moments are finite – Wigner's semicircle law holds true for Wigner matrices that have only moments of order 2 and lower finite. This indicates that Wigner's theorem is very stable: the statistic we are measuring (the density of eigenvalues *in the bulk*) is very universal, and not sensitive to small fluctuations that result from heavier tailed entries. The fluctuations and large deviations of $\mu_{\mathbf{X}_n}$ around the semicircular distribution as $n \to \infty$, however, are heavily depending on the distribution of the entries.

Let us normalize the variance of the (off-diagonal) entries, so that the resulting empirical law of eigenvalues is σ_1. The support of this measure is $[-2, 2]$. This suggests that the largest eigenvalue should converge to 2. Indeed, *half* of this statement follows from Wigner's law in general.

Lemma 6.1. *Let $\mathbf{X}_n$ be a Wigner matrix (with normalized variance of off-diagonal entries). Let $\lambda_n(\mathbf{X}_n)$ be the largest eigenvalue of $\mathbf{X}_n$. Then for any $\delta > 0$,*

$$\lim_{n\to\infty} \mathbb{P}(\lambda_n(\mathbf{X}_n) < 2 - \delta) = 0.$$

Proof. Fix $\delta > 0$, and let f be a continuous (or even C_c^∞, ergo Lipschitz) function supported in $[2-\delta, 2]$, satisfying $\int f\, d\sigma_1 = 1$. If $\lambda_n(\mathbf{X}_n) < 2 - \delta$, then $\mu_{\mathbf{X}_n}$ is supported in $(-\infty, 2-\delta)$, and so $\int f\, d\mu_{\mathbf{X}_n} = 0$. On the other hand, since $\int f\, d\sigma_1 = 1 > \frac{1}{2}$, we have

$$\mathbb{P}(\lambda_n(\mathbf{X}_n) < 2 - \delta) \le \mathbb{P}\left(\int f\, d\mu_{\mathbf{X}_n} = 0\right) \le \mathbb{P}\left(\left|\int f\, d\mu_{\mathbf{X}_n} - \int f\, d\sigma_1\right| > \frac{1}{2}\right),$$

and the last quantity tends to 0 as $n \to \infty$ by Wigner's law. □

To prove that $\lambda_n(\mathbf{X}_n) \to 2$ in probability, it therefore suffices to prove a complementary estimate for the probability that $\lambda_n(\mathbf{X}_n) > 2 + \delta$. As it happens, *this can fail to be true*. If the entries of $\mathbf{X}_n$ have heavy tails, the largest eigenvalue may fail to converge at all. This is a case of *local statistics* having large fluctuations. However, if we assume that all moments of $\mathbf{X}_n$ are finite (and sufficiently bounded), then the largest eigenvalue converges to 2.

Theorem 6.2. *Let $\mathbf{X}_n$ be a Wigner matrix with normalized variance of off-diagonal entries and* bounded *entries. Then $\lambda_n(\mathbf{X}_n) \to 2$ in probability. In fact, for any $\epsilon, \delta > 0$,*

$$\lim_{n\to\infty} \mathbb{P}\left(n^{1/6-\epsilon}(\lambda_n(\mathbf{X}_n) - 2) > \delta\right) = 0.$$

Remark 6.3.
(1) As with Wigner's theorem, the strong moment assumptions are not actually necessary for the statement to be true; rather, they make the proof convenient, and then can be removed afterward with a careful cutoff argument. However, in this case, it is known that second moments alone are not enough: in general, $\lambda_n(\mathbf{X}_n) \to 2$ if and only if the *fourth* moments of the entries are finite. (This was proved by Bai and Yin in the Annals of Probability in 1988). If there are no fourth moments, then it is known that the largest eigenvalue does not converge to 2; indeed, it has a Poissonian distribution (proved by Soshnikov in Electronic Communications in Probability, 2004).
(2) The precise statement in Theorem 6.2 shows not only that the largest eigenvalue converges to 2, but that the maximum displacement of this eigenvalue above 2 is about $O(n^{-1/6})$. In fact, this is not tight. It was shown by Vu (Combinatorica, 2007) that the displacement is $O(n^{-1/4})$. Under the additional assumption that the entries have *symmetric* distributions, it is known (Sinai-Soshnikov, 1998) that the maximum displacement is $O(n^{-2/3})$, and that

this is tight: at this scale, the largest eigenvalue has a non-trivial distribution. (Removing the symmetry condition is one of the front-line research problems in random matrix theory; for the most recent progress, see Péché-Soshnikov, 2008.) We will discuss this later in the course, in the special case of Gaussian entries.

Proof of Theorem 6.2. The idea is to use moment estimates, as follows. We wish to estimate $\mathbb{P}(\lambda_n > 2 + \delta n^{-1/6+\epsilon})$ (for fixed $\epsilon, \delta > 0$). Well, for any $k \in \mathbb{N}$,

$$\lambda_n > 2 + \delta n^{-1/6+\epsilon} \implies \lambda_n^{2k} > (2 + \delta n^{-1/6+\epsilon})^{2k} \implies \lambda_1^{2k} + \cdots + \lambda_n^{2k} > (2 + \delta n^{-1/6+\epsilon})^{2k}.$$

But $\lambda_1^{2k} + \cdots + \lambda_n^{2k} = \operatorname{Tr}(\mathbf{X}_n^{2k})$. Hence, using Markov's inequality, we have

$$\mathbb{P}(\lambda_n(\mathbf{X}_n) > 2 + \delta n^{-1/6+\epsilon}) \leq \mathbb{P}\left(\operatorname{Tr}(\mathbf{X}_n^{2k}) > (2 + \delta n^{-1/6+\epsilon})^{2k} \right) \leq \frac{\mathbb{E}\operatorname{Tr}(\mathbf{X}_n^{2k})}{(2 + \delta n^{-1/6+\epsilon})^{2k}}. \tag{6.1}$$

Equation 6.1 holds for any k; we have freedom to choose k as $n \to \infty$. The idea of this proof (which is due to Füredi and Komlós) is to let $k = k(n)$ grow with n in a precisely controlled fashion.

To begin, we revisit the proof of Wigner's theorem to estimate $\mathbb{E}\operatorname{Tr}(\mathbf{X}_n^k)$. Recall Equation 4.8, which shows (substituting $2k$ for k)

$$\frac{1}{n}\mathbb{E}\operatorname{Tr}(\mathbf{X}_n^{2k}) = \sum_{\substack{(G,w)\in\mathcal{G}_{2k}\\ w\geq 2}} \Pi(G,w) \cdot \frac{n(n-1)\cdots(n-|G|+1)}{n^{k+1}}.$$

Here $\mathcal{G}_{2k}$ denotes the set of pairs (G,w) where G is a connected graph with $\leq 2k$ vertices, w is a closed walk on G of length $2k$; the condition $w \geq 2$ means that w crosses each edge in G at least 2 times. The coefficients $\Pi(G,w)$ are given in Equation 4.4:

$$\Pi(G,w) = \prod_{e_s\in E^s(G)} \mathbb{E}(Y_{11}^{w(e_s)}) \cdot \prod_{e_c\in E^c(G)} \mathbb{E}(Y_{12}^{w(e_c)}),$$

where $E^s(G)$ is the set of self-edges $\{i,i\}$ in G, and $E^c(G)$ is the set of connecting-edge $\{i,j\}$ with $i \neq j$ in G; here $w(e)$ denotes the number of times the word w traverses the edge e. It is useful here to further decompose this sum according to the number of vertices in G. Let $\mathcal{G}_{2k}^t$ denote the subset of $\mathcal{G}_{2k}$ consisting of those pairs (G,w) where the number of vertices $|G|$ in G is t, and $w(e) \geq 2$ on each edge e of G. Since the length of w is $2k$, the number of edges in each G is $\leq k$. According to Exercise 4.3.1, it follows that $\mathcal{G}_{2k}^t$ (with condition $w \geq 2$) is empty if $t > k+1$. So we have the following refinement of Equation 4.8

$$\frac{1}{n}\mathbb{E}\operatorname{Tr}(\mathbf{X}_n^{2k}) = \sum_{t=1}^{k+1} \frac{n(n-1)\cdots(n-t+1)}{n^{k+1}} \sum_{(G,w)\in\mathcal{G}_{2k}^t} \Pi(G,w). \tag{6.2}$$

Then we can estimate

$$\frac{1}{n}\mathbb{E}\operatorname{Tr}(\mathbf{X}_n^{2k}) = \left|\frac{1}{n}\mathbb{E}\operatorname{Tr}(\mathbf{X}_n^{2k})\right| \leq \sum_{t=1}^{k+1} n^{t-(k+1)} \cdot \#\mathcal{G}_{2k}^t \cdot \sup_{(G,w)\in\mathcal{G}_{2k}^t} |\Pi(G,w)|. \tag{6.3}$$

First let us upper-bound the value of $|\Pi(G,w)|$. Let $M = \max\{\|Y_{11}\|_\infty, \|Y_{12}\|_\infty\}$. For fixed (G,w), break up the set of edges E of G into $E = E_0 \sqcup E_1 \sqcup E_2$, where E_0 are the self-edges and $E_2 = \{e \in E^c \colon w(e) = 2\}$; denote $\ell = |E_2|$. Since $\sum_{e\in E} w(e) = 2k$, we then have

$\sum_{e\in E_0\sqcup E_1} w(e) = 2k-2\ell$. Now, for any edge e, $|\mathbb{E}(Y_{ij}^{w(e)})| \le M^{w(e)}$ since $|Y_{ij}|\le M$ *a.s.* Thus, we have

$$|\Pi(G,w)| \le \prod_{e_0\in E_0} M^{w(e_0)} \prod_{e_1\in E_1} M^{w(e_1)} \prod_{e_2\in E_2} \mathbb{E}(Y_{12}^2) = M^{2k-2\ell} \tag{6.4}$$

because of the normalization $\mathbb{E}(Y_{12}^2)=1$. To get a handle on the statistic ℓ, first consider the new graph G' whose vertices are the same as those in G, but whose edges are the connecting edges $E_1\sqcup E_2$. Since $(G,w)\in\mathcal{G}_{2k}^t$, $t=|G|=|G'|$, and (cf. Exercise 4.3.1) the number of edges in G' is $\ge t-1$; that is $|E_1|+\ell\ge t-1$. Hence, we have

$$2k = \sum_{e\in E} w(e) \ge \sum_{e\in E_1} w(e) + \sum_{e\in E_2} w(e) \ge 3|E_1| + 2\ell \ge 3(t-\ell-1)+2\ell,$$

where the second inequality follows from the fact that every edge in E_1 is traversed ≥ 3 times by w. Simplifying yields $2k\ge 3t-3-\ell$, and so $3k-3t+3\ge k-\ell$. Hence $2(k-\ell)\le 6(k-t+1)$, and (assuming without loss of generality that $M\ge 1$) Inequality 6.5 yields

$$|\Pi(G,w)| \le M^{6(k-t+1)}. \tag{6.5}$$

This is true for all $(G,w)\in\mathcal{G}_{2k}^t$; so combining with Inequality 6.2 we find that

$$\left|\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,(\mathbf{X}_n^{2k})\right| \le \sum_{t=1}^{k+1} n^{t-(k+1)} M^{6(k-t+1)}\cdot \#\mathcal{G}_{2k}^t = \sum_{t=1}^{k+1}\left(\frac{M^6}{n}\right)^{k-t+1}\cdot\#\mathcal{G}_{2k}^t. \tag{6.6}$$

We are left to estimate the size of $\mathcal{G}_{2k}^t$.

Proposition 6.4. *For $t\le k+1$,*

$$\#\mathcal{G}_{2k}^t \le \binom{2k}{2t-2} C_{t-1} t^{4(k-t+1)}$$

where $C_{t-1}=\frac{1}{t}\binom{2t-2}{t-1}$ is the Catalan number.

The proof of Proposition 6.4 is fairly involved; we reserve it until after have used the result to complete the proof of Theorem 6.2. Together, Proposition 6.4 and Equation 6.6 yield

$$\left|\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,(\mathbf{X}_n^{2k})\right| \le \sum_{t=1}^{k+1}\left(\frac{M^6}{n}\right)^{k-t+1}\binom{2k}{2t-2}C_{t-1}t^{4(k-t+1)}.$$

Recombining and reindexing using $r=k-t+1$, noting that $\binom{2k}{2(k-r+1)-2}=\binom{2k}{2k-2k}=\binom{2k}{2r}$, this becomes

$$\left|\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,(\mathbf{X}_n^{2k})\right| \le \sum_{r=0}^{k}\left(\frac{M^6(k-r+1)^4}{n}\right)^r\binom{2k}{2r}C_{k-r}.$$

To simplify matters a little, we replace $k-r+1\le k$ (valid for $r\ge 1$, but when $r=0$ the exponent r makes the overestimate $k+1>k$ irrelevant). So, define

$$S(n,k,r) = \left(\frac{M^6k^4}{n}\right)^r\binom{2k}{2r}C_{k-r}.$$

Then we wish to bound $\sum_{r=0}^k S(n,k,r)$. We will estimate this by a geometric series, by estimating the successive ratios. For $1\le r\le k$,

$$\frac{S(n,k,r)}{S(n,k,r-1)} = \frac{M^6k^4}{n}\cdot\frac{\binom{2k}{2r}}{\binom{2k}{2(r-1)}}\frac{C_{k-r}}{C_{k-(r-1)}}. \tag{6.7}$$

Expanding out the radio of binomial coefficients gives

$$\frac{\frac{(2k)(2k-1)\cdots(2k-2r+1)}{(2r)!}}{\frac{(2k)(2k-1)\cdots(2k-2r+3)}{(2r-2)!}} = \frac{(2k-2r+2)(2k-2r+1)}{(2r)(2r-1)} \le 2k^2.$$

Similarly, for any $j \ge 1$,

$$\frac{C_{j-1}}{C_j} = \frac{\frac{1}{j}\binom{2j-2}{j-1}}{\frac{1}{j+1}\binom{2j}{j}} = \frac{j+1}{j}\frac{\frac{(2j-2)(2j-3)\cdots(j)}{(j-1)!}}{\frac{(2j)(2j-1)\cdots(j+1)}{j!}} = \frac{j+1}{2(2j-1)} \le 1. \tag{6.8}$$

(The ratio is close to $\frac{1}{4}$ when j is large.) Using this with $j-1 = k-r$, Equation 6.7 yields

$$\frac{S(n,k,r)}{S(n,k,r-1)} \le \frac{M^6k^4}{n}\cdot 2k^2 = \frac{2(Mk)^6}{n}.$$

Hence, for all r, we have

$$S(n,k,r) \le \left(\frac{2(Mk)^6}{n}\right)^r S(n,k,0) = \left(\frac{2(Mk)^6}{n}\right)^r C_k. \tag{6.9}$$

Now, let $k = k(n)$ be a function of n which satisfies

$$cn^{1/6} \le k(n) \le \frac{1}{M}\left(\frac{n}{4}\right)^{1/6} \tag{6.10}$$

for some $c > 0$. The upper bound on $k(n)$ is designed so the $\frac{2(Mk(n))^6}{n} \le \frac{1}{2}$; this, together with Inequality 6.9, gives

$$\left|\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,(\mathbf{X}_n^{2k(n)})\right| \le \sum_{r=0}^{k(n)} S(n,k(n),r) \le \sum_{r=0}^{k(n)}\left(\frac{1}{2}\right)^r C_{k(n)} \le \sum_{r=0}^{\infty}\left(\frac{1}{2}\right)^r C_{k(n)} = 2C_{k(n)}.$$

From Equation 6.8, we have $C_j/C_{j-1} = 2(2j-1)/(j+1) \le 4$, so $C_j \le 4^j$. With $j = k(n)$, plugging this into Inequality 6.1 (noting the missing $1/n$ to be accounted for) gives

$$\mathbb{P}(\lambda_n(\mathbf{X}_n) > 2 + \delta n^{-1/6} + \epsilon) \le n\cdot\frac{2\cdot 4^{k(n)}}{(2+\delta n^{-1/6+\epsilon})^{2k(n)}} = 2n\left(1+\frac{\delta}{2}n^{-1/6+\epsilon}\right)^{-2k(n)}.$$

Since $1+\frac{\delta}{2}n^{-1/6+\epsilon} > 1$, this quantity *increases* if we *decrease* the exponent. Using the lower bound in Inequality 6.10, this shows that

$$\mathbb{P}(\lambda_n(\mathbf{X}_n) > 2 + \delta n^{-1/6+\epsilon}) \le 2n\left(1+\frac{\delta}{2}n^{-1/6+\epsilon}\right)^{-2cn^{1/6}}.$$

It is now a simple matter of calculus to check that the right-hand-side converges to 0, proving the theorem. □

Exercise 6.4.1. *Refine the statement of Theorem 6.2 to show that, for any $\delta > 0$,*

$$\lim_{n\to\infty}\mathbb{P}(n^{1/6}\rho(n)(\lambda_n(\mathbf{X}_n)-2) > \delta) = 0$$

for any function $\rho > 0$ which satisfies

$$\lim_{n\to\infty}(\log n)\rho(n) = 0, \qquad \liminf_{n\to\infty} n^{1/12}\rho(n) > 0.$$

For example, one can take $\rho(n) = 1/\log(n)^{1+\epsilon}$. [Hint: use the Taylor approximation $\log(1+x) \ge x - \frac{1}{2}x^2$*.]*

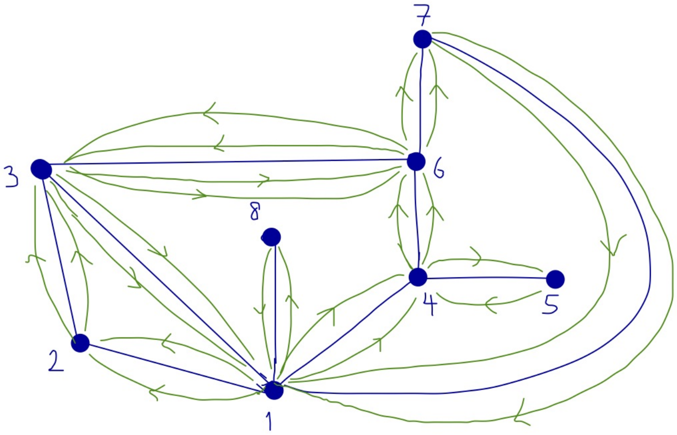
Proof of Proposition 6.4. We will count $\mathcal{G}_{2k}^t$ by producing a mapping into a set that is easier to enumerate (and estimate). The idea, originally due to Füredi and Komlós (1981), is to assign *codewords* to elements $(G, w) \in \mathcal{G}_{2k}^t$. To illustrate, we will consider a (fairly complicated) example throughout this discussion. Consider the pair $(G, w) \in \mathcal{G}_{22}^8$ with walk

$$w = 1231454636718146367123$$

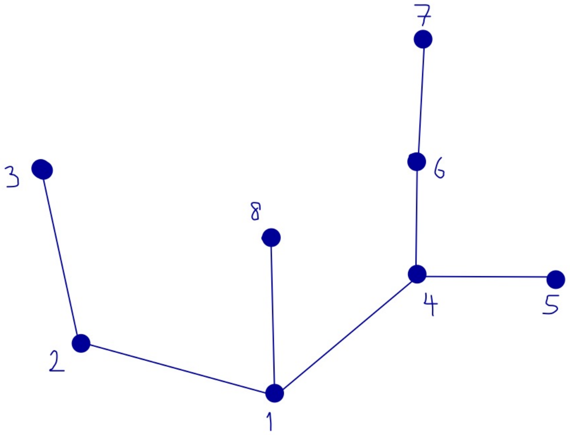
where, recall, the walk automatically returns to 1 at the end. It might be more convenient to list the consecutive edges in the walk:

$$w \sim 12, 23, 31, 14, 45, 54, 46, 63, 36, 67, 71, 18, 81, 14, 46, 63, 36, 67, 71, 12, 23, 31.$$

Figure 6 gives a diagram of the pair (G, w).



To begin our coding procedure, we produce a spanning tree for G: this is provided by w. Let $T(G, w)$ be the tree whose vertices are the vertices of G; the edges of $T(G, w)$ are only those edges $w_i w_{i+1}$ in $w = w_1 \cdots w_{2k}$ such that $w_{i+1} \notin \{w_1, \ldots, w_i\}$. (I.e. we include an edge from G in $T(G, w)$ only if it is the first edge traversed by w to reach its tip.) The spanning tree for the preceding example is displayed in Figure 6.



Now, here is a first attempt at a coding algorithm for the pair (G, w). We produce a *preliminary codeword* $c(G, w) \in \{+, -, 1, \ldots, t\}^{2k}$ as follows. Thinking of w as a sequence of $2k$ edges, we assign to each edge a symbol as follows: to the first appearance of each $T(G, w)$-edge in w, assign a $+$; to the second appearance of each $T(G, w)$-edge in w, assign a $-$. Refer to the other edges in the sequence w as *neutral* edges; these are the edges in G that are not in $T(G, w)$. If uv is a neutral edge, assign it the symbol v. In our example above, this procedure produces the preliminary codeword

$$c(G,w) = + + 1 + + - +36 + 1 + - - -36 - 1 - -1. \tag{6.11}$$

Let us pause here to make some easy observations: because $T(G,w)$ is a tree, the number of edges is $t-1$. Each edge appears a first time exactly once, and a second time exactly once, so the number of $+$s and the number of $-$s in $c(G,w)$ are both equal to $t-1$. This leaves $2k-2(t-1) = 2(k-t+1)$ neutral edges in w. We record this for future use:

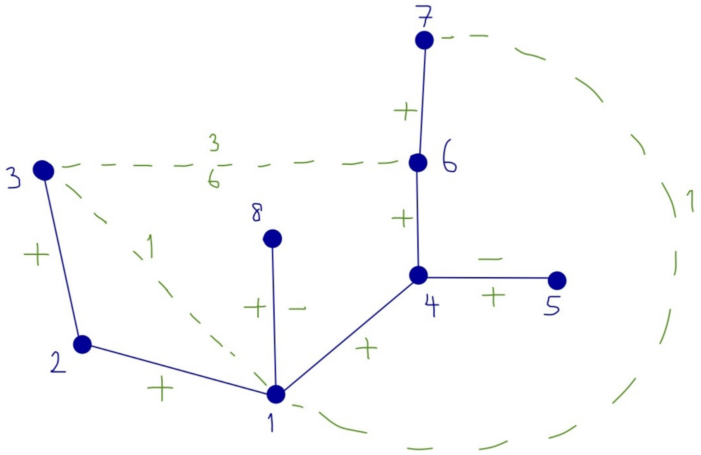
$$(G,w) \in \mathcal{G}_{2k}^t \implies \begin{cases} \#\{+ \in c(G,w)\} = \#\{- \in c(G,w)\} = t-1, \\ \#\{\text{netural edges in } w\} = 2(k-t+1). \end{cases} \tag{6.12}$$

Now, if $c\colon \mathcal{G}_{2k}^t \to \{+,-,1,\ldots,t\}^{2k}$ were injective, we could attempt to estimate the size of its range to get an upper-bound on $\#\mathcal{G}_{2k}^t$. Unfortunately, c is not injective, as our example above demonstrates. Let us attempt to "decode" the preliminary codeword in Equation 6.11, repeated here for convenience

$$+ + 1 + + - +36 + 1 + - - -36 - 1 - -1$$

- Each $+$ indicates an edge to a new vertex; so the initial $++$ indicates the edges $12, 23$.
- Whenever a symbol from $\{1,\ldots,t\}$ appears, the next edge connects to that vertex, so we follow with 31.
- Another $++$ indicates two new vertices, yielding $14, 45$.
- The $-$ means the next edge in w must be its second appearance; so it must be an edge in the tree *already generated* that has a $+$. There is only one such edge: 45. Hence, the next edge in w is 54.
- The following $+36+1+-$, by similar reasoning to the above, must now give $46, 63, 36, 67, 71, 18, 81$.

We have decoded the initial segment $+ + 1 + + - +36 + 1 + -$ to build up the initial segment $12, 23, 31, 14, 45, 54, 46, 63, 36, 67, 71, 18, 81$ of w. It is instructive to draw the (partial) spanning tree with (multiple) labels obtained from this procedure. This is demonstrated in Figure 6.



But now, we cannot continue the decoding procedure. The next symbol in $c(G,w)$ is a $-$, and we are sitting at vertex 1, which is adjacent to *two* edges that, thus far, have $+$s. Based on the data we have, we cannot decide whether the next edge is 12 or 24. Indeed, one can quickly check that there are two possible endings for the walk w,

$$14, 46, 63, 36, 67, 71, 12, 23, 31 \quad \textit{or} \quad 12, 23, 33, 36, 67, 71, 14, 46, 61.$$

Thus, the function c is not injective. Following this procedure, it is easy to see that each $+$ yields a unique next vertex, as does each appearance of a symbol in $\{1,\ldots,t\}$; the only trouble may arise with some $-$ labels.

Definition 6.5. *Let $(G,w)\in\mathcal{G}^t_{2k}$. A vertex u in G is a* **critical vertex** *if there is an edge uv in w such that, in the preliminary codeword $c(G,w)$, the label of uv is $-$, while there are at least two edges adjacent to u whose labels in $c(G,w)$* before *uv are both $+$.*

The set of critical vertices is well defined by (G,w). One way to "fix" the problem would be as follows. Given (G,w), produce $c(G,w)$. Produce a new codeword $c'(G,w)$ as follows: for each critical vertex u, and any edge uv in w that is labeled $-$ by $c(G,w)$, replace this label by $-_v$ in $c'(G,w)$. One can easily check that this clears up the ambiguity, and the map $(G,w)\mapsto c'(G,w)$ is injective. This fix, however, is too wasteful: easy estimates on the range of c' are not useful for the estimates needed in Equation 6.6. We need to be a little more clever.

To find a more subtle fix, first we make a basic observation. Any subword of $c(G,w)$ of the form $++\cdots+--\cdots-$ (with equal numbers of $+$s and $-$s) can always be decoded, regardless of its position relative to other symbols. Such subwords yield isolated chains within $T(G,w)$. In our running example, the edges 45 and 18 are of this form; each corresponds to a subword $+-$ in $c(G,w)$. We may therefore ignore such isolated segments, for the purposes of decoding. Hence, we produce a *reduced* codeword $c_r(G,w)$ as follows: from $c(G,w)$, successively remove all subwords of the form $++\cdots+--\cdots-$. There is some ambiguity about the order to do this: for example, $+++--+--\,uv$ can be reduced as

$$+(++--)+--\,uv \quad\mapsto\quad ++--\,uv=(++--)uv \quad\mapsto\quad uv$$

or

$$+++--(+-)-uv \quad\mapsto\quad +++---\,uv=(+++---)uv \quad\mapsto\quad uv$$

(or others). One can check by a simple induction argument that such successive reductions always result in the same reduced codeword $c_r(G,w)$.

Note, in our running example, the critical vertex 1 is followed in the word w by the isolated chain $18,81$, yielding a $+-$ in $c(G,w)$ just before the undecodable next $-$. There was no problem decoding this $+-$ (there is *never* any problem decoding a $+-$).

Definition 6.6. *An edge in w is called* **important** *if, in the reduced codeword $c_r(G,w)$, it is the final edge labeled $-$ in a string of $-$s following a critical vertex.*

For example, suppose u is a critical vertex, and v is another vertex. If the preliminary codeword continues after reaching u with

- $-+v---$ then the first edge is important (the codeword is reduced).
- $-++--v$ then the first edge is important (the reduced codeword is $u-v$).
- $-+---v$ then the fifth edge is important (the reduced codeword is $u---v$).
- $---+v-$ then the third edge is important (the codeword is reduced).

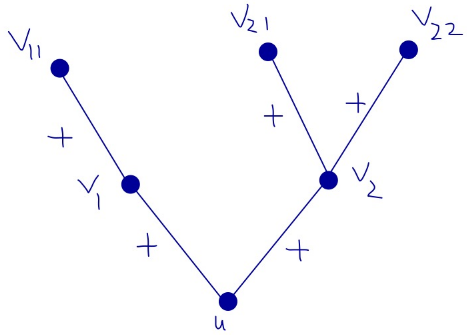
In our running example, the only critical vertix is 1; the final segment of $c(G,w)$ is $1+---36-1--1$, and this reduces in $c_r(G,w)$ to the reduced segment $1--36-1--1$. The important edge is the one labeled by the *second* $-$ sign after the 1; in the word w, this is the 46 (in the 15th position).

Definition 6.7. *Given a $(G,w)\in\mathcal{G}^t_{2k}$, define the* Füredi-Komlós codeword *$c_{FK}(G,w)$ as follows. In $c(G,w)$, replace the label $-$ of each important edge uv by the label $-_v$.*

So, in our running example, we have

$$c_{FK}(G,w) = ++1++-+36+1+---_{6}36-1--1.$$

The codeword $c_{FK}(G,w)$ usually requires many fewer symbols than the codeword $c'(G,w)$. For example, suppose that u is a critical vertex, and we have reached u with the adjacent segment of the spanning tree shown in Figure 6.



Suppose the walk w continues uv_2v_{21}, and so the preliminary codeword $c(G,w)$ continues $--$. Hence, since both u and v_2 are critical vertices, $c'(G,w)$ is modified to $-_{v_2}-_{v_{21}}$. On the other hand, $c_{FK}(G,w)$ only changes the label of the final $-$, yielding $--_{v_{21}}$. Of course, since $T(G,w)$ is a tree, knowing the final vertex in the path of $-$s determines the entire path.

Proposition 6.8. *The map $c_{FK}\colon \mathcal{G}^t_{2k} \to \{+,-,1,2,\ldots,t,-_1,-_2,\ldots,-_t\}^{2k}$ is injective.*

Proof of Proposition 6.8. We need to show that we can decode (G,w) from $c_{FK}(G,w)$. In fact, it suffices to show that we can recover $c'(G,w)$ from $c_{FK}(G,w)$, since (as described above) decoding $c'(G,w)$ is easy. The only points at which $c'(G,w)$ and $c_{FK}(G,w)$ may differ are strings of the form $\mathbf{d}--\cdots-$ in $c(G,w)$ following a critical vertex, where $\mathbf{d}$ is a sequence of $+$s and $-$s that reduces to $\varnothing$ in $c_r(G,w)$. (One can easily check that $\mathbf{d}$ is a Dyck path.) The initial redundant sequence $\mathbf{d}$ can be labeled independently of its position following a critical vertex, and so these labels remain the same in both $c'(G,w)$ and $c_{FK}(G,w)$. In $c_{FK}(G,w)$, only the final $-$ is changed to a $-_v$ for some vertex v in G. With this label in place, the information we have about w is: w begins at u, follows the subtree decoded from $\mathbf{d}$ back again to u, and then follows a chain of vertices *in* $T(G,w)$ (since they are labeled by $-$s) to v. Since $T(G,w)$ is a tree, there is only one path joining u and v, and hence all of the vertices in this path are determined by the label $-_v$ at the end. In particular, all label changes required in $c'(G,w)$ are dictated, and so we recover $c'(G,w)$ as required. □

Thus, $c_{FK}\colon \mathcal{G}^t_{2k} \to \{+,-,1,\ldots,t,-_1,\ldots,-_t\}^{2k}$ is injective. It now behooves us to bound the number of possible Füredi-Komlós codewords c_{FK}. Let $\mathcal{G}^{t,i}_{2k}$ denote the subset of $\mathcal{G}^t_{2k}$ consisting of those walks with precisely i important edges. Let I denote the maximum number of important edges possible in a walk of length $2k$ (so $I \le 2k$; we will momentarily give a sharp bound on I); so $\#\mathcal{G}^t_{2k} = \sum_{i=0}^{I} \#\mathcal{G}^{t,i}_{2k}$. So we can proceed to bound the image $c_{FK}(\mathcal{G}^{t,i}_{2k})$ – Füredi-Komlós codewords with precisely i important edges. We do this in two steps.

- Given any *preliminary codeword* c, we can determine the *locations* of the important edges without knowledge of their FK-labelings. Indeed, proceed to attempt to decode c: if there is an important edge, then at some point previous to it there is a critical vertex with an edge labeled $-$ following. We will not be able to decode the word at this point; so we simply look at the final $-$ in the reduced codeword following this point, and that is the location of the first important edge. Now, we proceed to "decode" from this point (actually decoding requires knowledge of where to go after the critical vertex; but we can still easily check if the next segment of c would allow decoding). Repeating this process, we can find all the locations of the i important edges. Hence, knowing c, there are at most t^i possible FK-codewords corresponding to it, since each important edge is labeled with a i_v for some $v \in \{1, \ldots, t\}$. That is:

$$\#\mathcal{G}_{2k}^{t,i} \le \#c(\mathcal{G}_{2k}^{t,i}) \cdot t^i. \tag{6.13}$$

- So we now have to count the number of preliminary codewords. This is easy to estimate. First, note that there are $t-1$ +s and $t-1$ −s (cf. Equation 6.12); we have to choose the $2(t-1)$ positions for them among the $2k$ total; there are $\binom{2k}{2t-2}$ such choices. Once the positions are chosen, we note that the sequence of +s and −s forms a Dyck path: each closing $-$ follows its opening $+$; hence, the number of ways to insert them is the Catalan number C_{t-1}. Finally, for each neutral vertex, we must choose a symbol from $\{1, \ldots, t\}$; there are complicated constraints on which ones may appear and in what order, but a simple upper bound for the number of ways to do this is $t^{2(k-t+1)}$ since (again by Equation 6.12) there are $2(k-t+1)$ neutral edges. Altogether, then, we have

$$\#c(\mathcal{G}_{2k}^{t,i}) \le \binom{2k}{2t-2} C_{t-1} t^{2(k-t+1)}. \tag{6.14}$$

Combining Equations 6.13 and 6.14, we have the upper-bound

$$\#\mathcal{G}_{2k}^{t} \le \binom{2k}{2t-2} C_{t-1} t^{2(k-t+1)} \sum_{i=0}^{I} t^i. \tag{6.15}$$

All we have left to do is bound the number of important edges. A first observation is as follows: suppose there are no neutral edges (which, as is easy to see, implies that $t = k+1$ since the graph must be a tree). Then all the symbols in the preliminary codeword are ±, and the reduced codeword is empty (since the Dyck path they form can be successively reduced to $\varnothing$). Then there are no important edges. So, in this case, $I = 0$, and our estimate gives

$$\#\mathcal{G}_{2k}^{t} \le \binom{2k}{2t-2} C_{t-1} t^{2(k-(t-1))} = \binom{2k}{2k} C_k t^{2(k-k)} = C_k.$$

This is the exact count we did as part of the proof of Wigner's theorem, so we have a sharp bound in this case. This case highlights that, in general, the maximal number of important edges I is controlled by the number of neutral edges.

Claim 6.9. *In any codeword, there is a neutral edge preceding the first important edge; there is a neutral edge between any two consecutive important edges; there is a neutral edge following the final important edge.*

Given Claim 6.9, it follows that in a codeword with i important edges, the number of neutral edges is at least $i+1$; this means that, in general, the maximal number of important edges is $I \le \#\{\text{neutral edges}\} - 1 = 2(k-t+1) - 1$ by Equation 6.12. Then we have

$$\sum_{i=0}^{I} t^i \le \sum_{i=0}^{2(k-t+1)-1} t^i \le t^{2(k-t+1)},$$

and plugging this into Equation 6.15 completes the proof of Proposition 6.4. So we are left to verify the claim.

- If an initial segment of the codeword consists of only $\pm$s, then this segment is eliminated in the reduced codewords; hence none of the $-$s can label important edges. Thus, there must be a neutral edge preceding the first important edge.
- Suppose e_1 and e_2 are two consecutive important edges, and suppose there is no neutral edge between them. Hence each $+$ following e_1 has its $-$ before e_2; so in the reduced codewords between e_1 and e_2, there are only $-$s. But e_1 is important, which means by definition that it is the *final* $-$ following a critical vertex in the reduced codeword, producing a contradiction.
- Let e be the final important edge, and let v be the critical vertex that defines it. Hence, at the point the walk reaches v before traversing e, there are *two* $+$ edges in the reduced codeword emanating from v. If there are no neutral edges after e, then remainder of the walk is along edges into the spanning tree, which means the walk cannot reach both $+$ edges (since doing so would require a loop). Since both $+$ edges must be crossed again by the walk, this is a contradiction.

□

7. The Stiletjes Transform

We have seen that some interesting combinatorial objects crop up in the study of the behaviour of random eigenvalues. We are now going to proceed in a different direciton, and introduce some *analytical* tools to study said eigenvalues. The first such tool is a transform that is, in many ways, analogous to the Fourier transform (i.e. characteristic function) of a probability measure.

Definition 7.1. *Let μ be a positive finite measure on $\mathbb{R}$. The* Stieltjes transform *of μ is the function*

$$S_\mu(z) = \int_\mathbb{R} \frac{\mu(dt)}{t-z}, \quad z \in \mathbb{C}\setminus\mathbb{R}.$$

Note: for fixed $z = x+iy$ with $y \neq 0$, we have

$$t \mapsto \frac{1}{t-z} = \frac{1}{t-x-iy} = \frac{t-x+iy}{(t-x)^2+y^2}$$

is a continuous function, and is bounded (with real and imaginary parts bounded by $\frac{1}{2|y|}$ and $\frac{1}{|y|}$ respectively). Hence, since μ is a finite measure, the integral $S_\mu(z)$ exists for any $z \notin \mathbb{R}$, and we have $|S_\mu(z)| \le \mu(\mathbb{R})/|\Im z|$. By similar reasoning, with careful use of the dominated convergence theorem, one can see that S_μ is complex analytic on $\mathbb{C}\setminus\mathbb{R}$. What's more, another application of the dominated convergenc theorem yields

$$\lim_{|z|\to\infty} zS_\mu(z) = \int_\mathbb{R} \lim_{|z|\to\infty} \frac{z}{t-z}\mu(dt) = -\int_\mathbb{R} \mu(dt) = -\mu(\mathbb{R}). \tag{7.1}$$

In fact, morally speaking, S_μ is (a change of variables from) the *moment-generating function* of the measure μ. Suppose that μ is actually compactly-supported, and let $m_n(\mu) = \int_\mathbb{R} t^n\,\mu(dt)$. Then if $\operatorname{supp}\mu \subseteq [-R,R]$, we have $|m_n(\mu)| \le R^n$, and so the generating function $u \mapsto \sum_n m_n u^n$ has a positive radius of convergence ($\ge 1/R$). Well, in this case, using the geometric series expansion

$$\frac{1}{t-z} = -\sum_{n=0}^\infty \frac{t^n}{z^{n+1}},$$

we have

$$S_\mu(z) = \int_{-R}^R \frac{\mu(dt)}{t-z} = -\int_{-R}^R \sum_{n=0}^\infty \frac{t^n}{z^{n+1}}\,\mu(dt).$$

So long as $|z| > R$, the sum is unformly convergent, and so we may interchange the integral and the sum

$$S_\mu(z) = -\sum_{n=0}^\infty \frac{1}{z^{n+1}} \int_{-R}^R t^n\,\mu(dt) = -\sum_{n=0}^\infty \frac{m_n(\mu)}{z^{n+1}}.$$

Thus, in a neighborhood of ∞, S_μ is a power series in $1/z$ whose coefficients (up to a shift and a $-$) are the moments of μ. This can be useful in calculating Stieltjes transforms.

Example. Consider the semicircle law $\sigma_1(dt) = \frac{1}{2\pi}\sqrt{(4-t^2)_+}\,dt$. From Exercise 2.4.1, we know that $m_{2n-1}(\sigma_1) = 0$ for $n \ge 1$ while $m_{2n}(\sigma_1) = C_n$ are the Catalan numbers. Since $C_n \le 4^n$, this shows that $m_n(\sigma_1) \le 2^n$, and so for $|z| > 2$, we have

$$S_{\sigma_1}(z) = -\sum_{n=0}^\infty \frac{C_n}{z^{2n+1}}. \tag{7.2}$$

This sum can actually be evaluated in closed-form, by utilizing the *Catalan recurrence relation*:

$$C_n = \sum_{j=1}^{n} C_{n-j}C_{j-1}, \quad n \geq 1$$

which is easily proved by induction. (This is kind of backwards: usually one proves that the number of Dyck paths of length $2n$ satisfies the Catalan recurrence relation, which therefore implies they are counted by Catalan number!) Hence we have

$$S_{\sigma_1}(z) = -\frac{C_0}{z} - \sum_{n=1}^{\infty} \frac{C_n}{z^{2n+1}} = -\frac{1}{z} - \sum_{n=1}^{\infty} \frac{1}{z^{2n+1}} \sum_{j=1}^{n} C_{n-j}C_{j-1}.$$

For the remaining double sum, it is convenient to make the change of variables $m = n-1$; so the sum becomes

$$S_{\sigma_1}(z) = -\frac{1}{z} - \sum_{m=0}^{\infty} \frac{1}{z^{2m+3}} \sum_{j=1}^{m+1} C_{m+1-j}C_{j-1}.$$

Now we reindex the internal sum by $i = j-1$, yielding

$$S_{\sigma_1}(z) = -\frac{1}{z} - \frac{1}{z}\sum_{m=0}^{\infty} \frac{1}{z^{2m+2}} \sum_{i=0}^{m} C_{m-i}C_{i}.$$

For each i in the internal sum, we distribute

$$\frac{1}{z^{2m+2}} = \frac{1}{z^{2(m-i)+1}}\frac{1}{z^{2i+1}}$$

so that

$$S_{\sigma_1}(z) = -\frac{1}{z} - \frac{1}{z}\sum_{m=0}^{\infty}\sum_{i=0}^{m} \left(\frac{C_{m-i}}{z^{2(m-i)+1}}\right)\left(\frac{C_i}{z^{2i+1}}\right).$$

The double sum is a series of the form $\sum_{m=0}^{\infty}\sum_{i=0}^{m} a_{m-i}a_i$ (where $a_i = C_i/z^{2i+1}$), which is a reindexing of the sum $\sum_{m=0}^{\infty}\sum_{k=0}^{\infty} a_m a_k = \left(\sum_{m=0}^{\infty} a_m\right)^2$. Thus, we have

$$S_{\sigma_1}(z) = -\frac{1}{z} - \frac{1}{z}\left(\sum_{m=0}^{\infty}\frac{C_m}{z^{2m+1}}\right)^2 = -\frac{1}{z} - \frac{1}{z}S_{\sigma_1}(z)^2$$

where the last equality follows from Equation 7.2. In other words, for each $z \in \mathbb{C} \setminus \mathbb{R}$, $S_{\sigma_1}(z)$ satisfies the quadratic equation

$$S_{\sigma_1}(z)^2 + zS_{\sigma_1} + 1 = 0.$$

The solutions are

$$S_{\sigma_1}(z) = \frac{-z \pm \sqrt{z^2-4}}{2}.$$

Noting, from Equation 7.1, that $S_{\sigma_1}(z) \to 0$ as $|z| \to \infty$, we see the correct sign is $+$; and one can easily check that

$$zS_{\sigma_1}(z) = \frac{z(-z+\sqrt{z^2-4})}{2} = \frac{z((z^2-4)-z^2)}{2(\sqrt{z^2-4}+z)} = \frac{-2z}{\sqrt{z^2-4}+z} \to -1 \text{ as } |z| \to \infty.$$

In the above example, we verified that, for $|z| > 2$, $S_{\sigma_1}(z) = \frac{1}{2}(-z+\sqrt{z^2-4})$. The function $z \mapsto \frac{1}{2}(-z+\sqrt{z^2-4})$ is analytic everywhere on $\mathbb{C} \setminus \mathbb{R}$, however (in fact everywhere except at $[-2,2]$). Since S_{σ_1} is also analytic on $\mathbb{C} \setminus \mathbb{R}$, it follows that the two agree everywhere. This is one of the nice features of the Stieltjes transform: one only need calculate it on a neighborhood of ∞ (or, indeed, on any set that contains an accumulation point) to determine its value everywhere.

Another important feature of the Stieltjes transform is that, as with the characteristic function, the measure μ is determined by S_μ in a simple (and analytically robust) way.

Theorem 7.2 (Stieltjes inversion formula)**.** *Let μ be a positive finite measure on $\mathbb{R}$. For $\epsilon > 0$, define a measure μ_ϵ by*

$$\mu_\epsilon(dx) = \frac{1}{\pi}\Im S_\mu(x+i\epsilon)\,dx.$$

Then μ_ϵ is a positive measure with $\mu_\epsilon(\mathbb{R}) = \mu(\mathbb{R})$, and $\mu_\epsilon \to \mu$ weakly as $\epsilon \downarrow 0$. In particular, if I is an open interval in $\mathbb{R}$ and μ does not have an atom in ∂I, then

$$\mu(I) = \lim_{\epsilon\downarrow 0}\frac{1}{\pi}\int_I \Im S_\mu(x+i\epsilon)\,dx.$$

Proof. First note that the map $\mu \mapsto S_\mu$ is a linear map from the cone of positive finite measures into the space of analytic functions on $\mathbb{C}\setminus\mathbb{R}$. Hence, to verify that it is one-to-one, we need only check that its kernel is 0. Suppose, then, that $S_\mu(z) = 0$ for all z. In particular, this means $S_\mu(i) = 0$. Well

$$\Im S_\mu(i) = \Im\int_{\mathbb{R}}\frac{\mu(dt)}{t-i} = \int_{\mathbb{R}}\frac{\mu(dt)}{t^2+1}$$

and so if this is 0, then the positive measure μ is 0. So, we can in principle recover μ from S_μ. Now, assume S_μ is not identically 0. Then $\mu(\mathbb{R}) > 0$, and $\frac{1}{\mu(\mathbb{R})}S_\mu = S_{\mu/\mu(\mathbb{R})}$; thus, without loss of generality, we assume that μ is a probability measure. We can now calculate in general

$$\Im S_\mu(x+i\epsilon) = \Im\int_{\mathbb{R}}\frac{\mu(dx)}{x+i\epsilon-t} = \int_{\mathbb{R}}\frac{\epsilon}{(x-t)^2+\epsilon^2}\,\mu(dx).$$

Thus $\mu_\epsilon = \chi_\epsilon * \mu$ where $\chi_\epsilon(dx) = \frac{\epsilon\,dx}{\pi(x^2+\epsilon^2)}$ is the Cauchy distribution. A convolution of two probability measures is a probability measure, verifying that $\mu_\epsilon(\mathbb{R}) = 1$. Moreover, it is well-known that the density of χ_ϵ forms an apporximate identity sequence as $\epsilon \downarrow 0$; hence, $\mu_\epsilon = \chi_\epsilon * \mu$ converges weakly to μ as $\epsilon \downarrow 0$. The second statement (in terms of measures of intervals I) follows by a standard approximation of $\mathbb{1}_I$ by C_c^∞ functions. □

Remark 7.3. The approximating measures μ_ϵ in Theorem 7.2 all have smooth densities $\rho_\epsilon(x) = \frac{1}{\pi}\Im S_\mu(x+i\epsilon)$. If the measure μ also has a density $\mu(dx) = \rho(x)\,dx$, then in particular μ has no atoms; the statement of the theorem thus implies that, in this case, $\rho_\epsilon \to \rho$ pointwise as $\epsilon \downarrow 0$.

Example. As calculated above, the Stieltjes transform of the semicircle law σ_1 is $S_{\sigma_1}(z) = \frac{1}{2}(-z+\sqrt{z^2-4})$. Hence, the approximating measures from the Stieltjes inversion formula have densities

$$\begin{aligned}\rho_\epsilon(x) = \frac{1}{\pi}\Im S_\mu(x+i\epsilon) &= \frac{1}{2\pi}\Im\left(-(x+i\epsilon)+\sqrt{(x+i\epsilon)^2-4}\right)\\ &= -\frac{\epsilon}{2\pi}+\frac{1}{2\pi}\Im\sqrt{(x+i\epsilon)^2-4}.\end{aligned}$$

Taking the limit inside the square root, this yields

$$\lim_{\epsilon\downarrow 0}\rho_\epsilon(x) = \frac{1}{2\pi}\Im\sqrt{x^2-4} = \frac{1}{2\pi}\sqrt{(4-x^2)_+}$$

yielding the density of σ_1, as expected.

8. The Stieltjes Transform and Convergence of Measures

In order to fully characterize the robust features of the Stieltjes transform with respect to convergent sequence of measures, it is convenient to introduce a less-well-known form of convergence.

8.1. Vague Convergence.

Definition 8.1. *Say that a sequence μ_n of measures on $\mathbb{R}$ converge* **vaguely** *to a measure μ on $\mathbb{R}$ if, for each $f \in C_c(\mathbb{R})$, $\int_{\mathbb{R}} f\, d\mu_n \to \int_{\mathbb{R}} f\, d\mu$.*

Vague convergence is a slight weakening of weak convergence. Since constants are no longer allowed test functions, vague convergence is not required to preserve total mass: it may allow some mass to escape at $\pm\infty$, while weak convergence does not. For example, the sequence of point-masses δ_n converges vaguely to 0, but does not converge weakly. This is the only difference between the two: if the measures μ_n and μ are all probability measures, then vague convergence implies weak convergence.

Lemma 8.2. *Let μ_n be a sequence of probability measures, and suppose $\mu_n \to \mu$ vaguely where μ is a probability measure. Then $\mu_n \to \mu$ weakly.*

Proof. Fix $\epsilon > 0$. Because $\mu(\mathbb{R}) = 1$ and the nested sets $[-R, R]$ increase to $\mathbb{R}$ as $R \to \infty$, there is an R_ϵ such that $\mu([-R_\epsilon, R_\epsilon]) > 1 - \epsilon/2$. Let $R'_\epsilon > R_\epsilon$, and fix a positive C_c test function ψ_ϵ that equals 1 on $[-R_\epsilon, R_\epsilon]$, equals 0 outside of $[-R'_\epsilon, R'_\epsilon]$, and is always ≤ 1. Now, by assumption $\mu_n \to \mu$ vaguely. Hence

$$\begin{aligned}\mu_n([-R'_\epsilon, R'_\epsilon]) = \int \mathbb{1}_{[-R'_\epsilon, R'_\epsilon]}\, d\mu_n \geq \int \psi_\epsilon\, d\mu_n \to \int \psi_\epsilon\, d\mu \geq \int \mathbb{1}_{[-R_\epsilon, R_\epsilon]} \\ = \mu([-R_\epsilon, R_\epsilon]) > 1 - \epsilon/2.\end{aligned}$$

Therefore, there is $N_\epsilon \in \mathbb{N}$ so that for all $n > N_\epsilon$, $\mu_n([-R'_\epsilon, R'_\epsilon]) > 1 - \epsilon$. Denote the interval $[-R'_\epsilon, R'_\epsilon]$ as I_ϵ; so for large enough n, μ_n satisfy $\mu_n(I_\epsilon) > 1 - \epsilon$, and also $\mu(I_\epsilon) > 1 - \epsilon$.

Now, let $g \in C_b(\mathbb{R})$. Fix a positive test function $\varphi_\epsilon \in C_c(\mathbb{R})$ that is equal to 1 on I_ϵ, and $\varphi \leq 1$. Then $g\varphi_\epsilon \in C_c(\mathbb{R})$, and so by the assumption of vague convergence, there is an $N'_\epsilon \in \mathbb{N}$ such that, for $n > N'_\epsilon$,

$$\left| \int g\varphi_\epsilon\, d\mu_n - \int g\varphi_\epsilon\, d\mu \right| < \|g\|_\infty \epsilon. \tag{8.1}$$

Now we compute

$$\begin{aligned}\int g\, d\mu_n &= \int g \cdot (\varphi_\epsilon + (1 - \varphi_\epsilon))\, d\mu_n = \int g\varphi_\epsilon\, d\mu_n + \int g \cdot (1 - \varphi_\epsilon)\, d\mu_n \\ \int g\, d\mu &= \int g \cdot (\varphi_\epsilon + (1 - \varphi_\epsilon))\, d\mu = \int g\varphi_\epsilon\, d\mu + \int g \cdot (1 - \varphi_\epsilon)\, d\mu.\end{aligned}$$

Subtracting, we get

$$\left| \int g\, d\mu_n - \int g\, d\mu \right| \leq \left| \int g\varphi_\epsilon\, d\mu_n - \int g\varphi_\epsilon\, d\mu \right| + \int |g| \cdot (1 - \varphi_\epsilon)\, d\mu_n + \int |g| \cdot (1 - \varphi_\epsilon)\, d\mu.$$

The first term is $< \|g\|_\infty \epsilon$ for $n > N'_\epsilon$ by Equation 8.1. The function $1 - \varphi_\epsilon$ is 0 on I_ϵ, and since $\varphi_\epsilon \leq 1$, we have $1 - \varphi_\epsilon \leq \mathbb{1}_{I_\epsilon^c}$. By construction, $\mu_{I_\epsilon^c} < \epsilon$, and so

$$\int |g| \cdot (1 - \varphi_\epsilon)\, d\mu_n \leq \|g\|_\infty \int (1 - \varphi_\epsilon)\, d\mu_n \leq \|g\|_\infty \int 1_{I_\epsilon^c}\, d\mu_n = \|g\|_\infty \mu_n(I_\epsilon) < \|g\|_\infty \epsilon,$$

the final inequality holding true for $n > N_\epsilon$. Similarly $\int |g| \cdot (1-\varphi_\epsilon)\,d\mu < \|g\|_\infty \epsilon$ for $n > N_\epsilon$. Altogether, we have

$$\left| \int g\,d\mu_n - \int g\,d\mu \right| < 3\|g\|_\infty \epsilon, \quad \text{for } n > \max\{N_\epsilon, N'_\epsilon\}.$$

It follows that $\int g\,d\mu_n \to \int g\,d\mu$. We have therefore shown that $\mu_n \to \mu$ weakly. □

Remark 8.3. The statement (and proof) of Lemma 8.2 remain valid for *finite* measures μ_n, μ, *provided there is a uniform upper bound on the total mass of* μ_n.

Remark 8.4. Since $C_c \subset C_b$, we can define other convergence notions relative to other classes of test functions in between these two; while they may give different results for general measures, Lemma 8.2 shows that they will all be equivalent to weak convergence when the measures involved are all probability measures. In particular, many authors define vague convergen in terms of $C_0(\mathbb{R})$, the space of continuous functions that tend to 0 at $\pm\infty$, rather than the smaller subspace $C_c(\mathbb{R})$. This is a useful convention for us on two counts: the "test function" $\frac{1}{z-t}$ in the definition of the Stieltjes transform is in C_0 but not in C_c, and also the space C_0 is a nicer space than C_c from a functional analytic point of view. We expound on this last point below.

From a functional analytic point of view, vague convergence is more natural than weak convergence. Let X be a Banach space. Its dual space X^* is the space of bounded linear functionals $X \to \mathbb{C}$. The dual space has a norm, given by

$$\|x^*\|_{X^*} = \sup_{x\in X,\ \|x\|_X=1} |x^*(x)|.$$

This norm makes X^* into a Banach space (even if X itself is not complete but merely a normed vector space). There are, however, other topologies on X^* that are important. The **weak* topology** on X^*, denoted (X^*, wk^*), is determined by the following convergence rule: a sequence x_n^* in X^* converges to $x \in X^*$ if $x_n^*(x) \to x^*(x)$ for each $x \in X$. This is strictly weaker than the weak topology, where convergence means $\|x_n^* - x\|_{X^*} \to 0$; i.e. $\sup_{x\in X,\ \|x\|_X=1} |x_n^*(x) - x^*(x)| \to 0$. The terms for each fixed x can converge even if the supremum doesn't; weak*-convergence is sometimes called simply pointwise convergence.

The connection between these abstract notions and our present setting is as follows. The space $X = C_0(\mathbb{R})$ is a Banach space (with the uniform norm $\|f\|_\infty = \sup_t |f(t)|$). According to the Riesz Representation theorem, its dual space X^* can be identified with the space $M(\mathbb{R})$ of complex Radon measures on $\mathbb{R}$. (Note: any finite positive measure on $\mathbb{R}$ is automatically Radon.) The identification is the usual one: a measure $\mu \in M(\mathbb{R})$ is identified with the linear functional $f \mapsto \int_\mathbb{R} f\,d\mu$ on $C_0(\mathbb{R})$. As such, in $(C_0(\mathbb{R})^*, wk^*)$, the convergence $\mu_n \to \mu$ is given by

$$\int_\mathbb{R} f\,d\mu_n \to \int_\mathbb{R} f\,d\mu, \quad \forall\, f \in C_0(\mathbb{R}).$$

That is, **weak* convergence in $M(\mathbb{R}) \cong C_0(\mathbb{R})^*$ is equal to vague convergence**. This is the reason vague convergence is natural: it has a nice functional analytic description. The larger space $C_b(\mathbb{R})$ is also a Banach space in the uniform norm; however, the dual space $C_b(\mathbb{R})^*$ cannot be naturally identified as a space of measures. (For example: $C_0(\mathbb{R})$ is a closed subspace of $C_b(\mathbb{R})$; by the Hahn-Banach theorem, there exists a bounded linear functional $\Lambda \in C_b(\mathbb{R})^*$ that is identically 0 on $C_0(\mathbb{R})$, and $\Lambda(1) = 1$. It is easy to check that Λ cannot be given by integration against a measure.)

The main reason the weak* topology is useful is the following important theorem.

Theorem 8.5 (Banach-Alaoglu theorem). *Let X be a normed space, and let $B(X^*)$ denote the closed unit ball in X^*: $B(X^*) = \{x^* \in X^* \colon \|x^*\|_{X^*} \le 1\}$. Then $B(X^*)$ is* **compact** *in (X^*, wk^*).*

Compact sets are hard to come by in infinite-dimensional spaces. In the X^*-topology, $B(X^*)$ is definitely not compact (if X is infinite-dimensional). Thinking in terms of sequential compactness, the Banach-Alaoglu theorem therefore has the following important consequence for vague convergence.

Lemma 8.6. *Let μ_n be a collection of finite positive measures on $\mathbb{R}$ with $\mu_n(\mathbb{R}) \le 1$. Then there exists a subsequence μ_{n_k} that converges vaguely to some finite positive measure μ.*

Proof. Let $M_{\le 1}(\mathbb{R})$ denote the set of positive measures μ on $\mathbb{R}$ with $\mu(\mathbb{R}) \le 1$. As noted above, $M_{\le 1}(\mathbb{R}) \subset M(\mathbb{R}) = C_0(\mathbb{R})^*$. Now, for any $f \in C_0(\mathbb{R})$ with $\|f\|_\infty = 1$, and any $\mu \in M_{\le 1}(\mathbb{R})$, we have

$$\left|\int_{\mathbb{R}} f\,d\mu\right| \le \int_{\mathbb{R}} \|f\|_\infty\,d\mu = \|f\|_\infty \mu(\mathbb{R}) \le 1.$$

Taking the supremum over all such f shows that $\mu \in B(C_0(\mathbb{R})^*)$; thus $M_{\le 1}(\mathbb{R})$ is contained in the closed unit ball of $C_0(\mathbb{R})^*$.

Now, suppose that μ_n is a sequence in $M_{\le 1}$ that converges vaguely to $\mu \in M(\mathbb{R})$. First note that μ must be a positive measure: given any $f \in C_c(\mathbb{R}) \subset C_0(\mathbb{R})$, we have (by the assumption of vague convergence) that $\int f\,d\mu = \lim_{n\to\infty} f\,d\mu_n \ge 0$, and so μ is positive on all compact subsets of $\mathbb{R}$, hence is a positive measure. Now, fix a sequence $f_k \in C_c(\mathbb{R})$ that increases to 1 pointwise. Then

$$\mu(\mathbb{R}) = \int_{\mathbb{R}} d\mu = \int_{\mathbb{R}} \lim_{k\to\infty} f_k\,d\mu \le \limsup_{k\to\infty} \int f_k\,d\mu = \limsup_{k\to\infty} \lim_{n\to\infty} \int f_k\,d\mu_n,$$

where the inequality follows from Fatou's lemma. Because $f_k \le 1$, and $\mu_n \in M_{\le 1}(\mathbb{R})$, all the terms in this double sequence are ≤ 1, and so the $\limsup$ is ≤ 1, as desired. This shows that $\mu \in M_{\le 1}(\mathbb{R})$.

Thus, we have shown that $M_{\le 1}$ is a subset of $B(C_0(\mathbb{R})^*)$ that is closed under vague convergence: i.e. it is closed in $(C_0(\mathbb{R})^*, wk^*)$. A closed subset of a compact set is compact, and therefore by the Banach-Alaoglu theorem, $M_{\le 1}$ is compact in the weak* topology. This is precisely the statement of the lemma (in sequential compactness form). □

Remark 8.7. Compactness and sequential compactness are not generally equivalent, but they are equivalent in metric spaces. In fact, for any *separable* normed space X (such as $C_0(\mathbb{R})$), the unit ball $(B(X^*), wk^*)$ is metrizable: fix a countable dense subset $\{x_n\}_{n=1}^\infty$ of X. Then

$$d(x^*, y^*) = \sum_{n=1}^\infty 2^{-n} \frac{|x^*(x_n) - x^*(x_n)|}{1 + |x^*(x_n) - y^*(x_n)|}$$

is a metric on $B(X^*)$ which can easily be seen to yield the weak* topology. Indeed, this is probably the best approach to prove the Banach-Alaoglu theorem: one can use a diagonalization argument mimicking the standard proof of the Arzelà-Ascoli theorem.

Lemma 8.6 and Remark 8.7 show that the set $M_{\le 1}(\mathbb{R})$ of positive measures with mass ≤ 1 is a compact metric space in the vague topology. It is therefore worth noting the following simple fact about compact metric spaces.

Lemma 8.8. *Let (M, d) be any compact metric space. If μ_n is a sequence in M and $\mu \in M$, if μ_n does not converge to μ (i.e. $d(\mu_n, \mu)$ does not converge to 0), then there exists a subsequence $\{\mu_{n_k}\}$ that converges to some $\mu' \in M$ with $\mu' \neq \mu$.*

Proof. Since μ_n does not converge to μ, there is some $\epsilon > 0$ such that $d(\mu_n, \mu) \geq \epsilon$ for infinitely many n. So there is a subsequence contained in the closed set $\{\nu \in M : d(\nu, \mu) \geq \epsilon\}$. Since M is compact, this closed set is compact, and hence there is a further subsequence which converges in this set – ergo to a limit μ' with $d(\mu, \mu') \geq \epsilon$. □

8.2. **Robustness of the Stieltjes transform.** We are now ready to prove the main results on convergence of Stieltjes transforms of sequences of measures.

Proposition 8.9. *Let μ_n be a sequence of probability measures on $\mathbb{R}$.*

(a) *If μ_n converges vaguely to the probability measure μ, then $S_{\mu_n}(z) \to S_\mu(z)$ for each $z \in \mathbb{C} \setminus \mathbb{R}$.*
(b) *Conversely, suppose that $\lim_{n\to\infty} S_{\mu_n}(z) = S(z)$ exists for each $z \in \mathbb{C} \setminus \mathbb{R}$. Then there exists a finite positive measure μ with $\mu(\mathbb{R}) \leq 1$ such that $S = S_\mu$, and $\mu_n \to \mu$ vaguely.*

Proof. We noted following the definition of the Stieltjes transform that the function $t \mapsto \frac{1}{t-z}$ is continuous and bounded; it is also clear that it is in $C_0(\mathbb{R})$, and hence, part (a) follows by the definition of vague convergence. To prove part (b), begin by selecting from μ_n a subsequence μ_{n_k} that converges vaguely to some sub-probability measure μ, as per Lemma 8.6. Then, as in part (a), it follows that $S_{\mu_{n_k}}(z) \to S_\mu(z)$ for each $z \in \mathbb{C} \setminus \mathbb{R}$. Thus $S(z) = S_\mu(z)$ for each $z \in \mathbb{C} \setminus \mathbb{R}$. By the Stieltjes inversion formula, if $S(z) = S_\nu(z)$ for some finite positive measure ν, then $\nu = \mu$. So we have shown that every vaguely-convergent subsequence of μ_n converges to the same limit μ. It follows from Lemma 8.8 that $\mu_n \to \mu$ vaguely. □

Remark 8.10. It would be more desirable to conclude in part (b) that S is the Stieltjes transform of a probability measure. But this cannot be true in general: for example, with $\mu_n = \delta_n$, the sequence $S_{\delta_n}(z) = \frac{1}{n-z}$ converges pointwise to 0, the Stieltjes transform of the 0 measure. The problem is that the set $M_1(\mathbb{R})$ of probability measures is not compact in the vague topology; indeed, it is not closed, since mass can escape at $\pm\infty$. Unfortunately $M_1(\mathbb{R})$ is also not compact in the weak topology, for the same reason: by letting mass escape to ∞, it is easy to find sequences in $M_1(\mathbb{R})$ that have no weakly convergent subsequences (for example δ_n).

Our main application of the Stieltjes transform will be to the empirical eigenvalue measure of a Wigner matrix. As such, we must deal with S_μ as a *random variable*, since μ will generally be a random measure. The following result should be called the **Stieltjes continuity theorem** for random probability measures.

Theorem 8.11. *Let μ_n and μ be random probability measures on $\mathbb{R}$. Then μ_n converges weakly almost surely to μ if and only if $S_{\mu_n}(z) \to S_\mu(z)$ almost surely for each $z \in \mathbb{C} \setminus \mathbb{R}$.*

Remark 8.12. We could state this theorem just as well for convergence in probability in each case; almost sure convergence is stronger, and generally easier to work with.

Proof. For the “only if” direction, we simply apply Proposition 8.9(a) pointwise. That is to say: by assumption $\mu_n \to \mu$ weakly a.s., meaning there is an event Ω with probability $\mathbb{P}(\Omega) = 1$ such that, for all $\omega \in \Omega$, $\mu_n(\omega) \to \mu(\omega)$ weakly. Now, fix $z \in \mathbb{C} \setminus \mathbb{R}$; then the function $f_z(t) = \frac{1}{t-z}$ is in $C_b(\mathbb{R})$. Thus, by definition of weak convergence, we have

$$S_{\mu_n(\omega)}(z) = \int f_z \, d\mu_n(\omega) \to \int f_z \, d\mu(\omega) = S_{\mu(\omega)}(z),$$

establishing the desired statement.

The converse "if" direction requires slightly more work; this proof was constructed by Yinchu Zhu in Fall 2013. By assumption, $S_{\mu_n}(z) \to S_\mu(z)$ a.s. for each $z \in \mathbb{C} \setminus \mathbb{R}$. To be precise, this means that, for each $z \in \mathbb{C} \setminus \mathbb{R}$, there is an event Ω_z of full probability $\mathbb{P}(\Omega_z) = 1$ such that $S_{\mu_n(\omega)}(z) \to S_{\mu(\omega)}(z)$ for all $\omega \in \Omega_z$. We would like to find a single z-independent full probability event Ω such that $S_{\mu_n(\omega)}(z) \to S_{\mu(\omega)}(z)$ for all $z \in \mathbb{C} \setminus \mathbb{R}$ and $\omega \in \Omega$. Näively, we should take $\Omega = \cap_{z \in \mathbb{C}\setminus\mathbb{R}} \Omega_z$; but $\mathbb{C} \setminus \mathbb{R}$ is uncountable, so this intersection need not be full probability (indeed, it need not be measurable, or could well be empty). Instead, we take a countable dense subset. Let $\mathbb{Q}(i) = \{x' + iy' \colon x', y' \in \mathbb{Q}\}$. Then $\mathbb{Q}(i) \setminus \mathbb{Q}$ is dense in $\mathbb{C} \setminus \mathbb{R}$. Define

$$\Omega \equiv \bigcap_{z' \in \mathbb{Q}(i) \setminus \mathbb{Q}} \Omega_{z'}.$$

Then $\mathbb{P}(\Omega) = 1$, and essentially by definition we have $S_{\mu_n(\omega)}(z') \to S_{\mu(\omega)}(z')$ for all $\omega \in \Omega$ and $z' \in \mathbb{Q}(i) \setminus \mathbb{Q}$.

Claim 8.13. *For any $\omega \in \Omega$ and $z \in \mathbb{C} \setminus \mathbb{R}$, $S_{\mu_n(\omega)}(z) \to S_{\mu(\omega)}(z)$ as $n \to \infty$.*

To prove the claim, we use the triangle inequality in the usual way: for any $z' \in \mathbb{Q}(i) \setminus \mathbb{Q}$,

$$|S_{\mu_n(\omega)}(z) - S_{\mu(\omega)}(z)| \le |S_{\mu_n(\omega)}(z) - S_{\mu_n(\omega)}(z')| + |S_{\mu_n(\omega)}(z') - S_{\mu(\omega)}(z')| + |S_{\mu(\omega)}(z') - S_{\mu(\omega)}(z)|.$$

By assumption, the middle term tends to 0 as $n \to \infty$. For the first and last terms, simply note the following: for any probability measure ν on $\mathbb{R}$,

$$|S_\nu(z) - S_\nu(z')| = \left| \int \left(\frac{1}{t-z} - \frac{1}{t-z'} \right) \nu(dt) \right| = \left| \int \frac{z - z'}{(t-z)(t-z')}\, \nu(dt) \right| \le \int \frac{|z-z'|}{|t-z||t-z'|}\, \nu(dt).$$

Since t is real, $|t - z| \ge |\Im z|$, and so we have the estimate

$$|S_\nu(z) - S_\nu(z')| \le \int \frac{|z-z'|}{|\Im z||\Im z'|}\, \nu(dt) = \frac{|z-z'|}{|\Im z||\Im z'|}.$$

Applying this to the first and third terms above (with $\nu = \mu_n(\omega)$ or $\nu = \mu(\omega)$), we therefore have the estimate

$$|S_{\mu_n(\omega)}(z) - S_{\mu(\omega)}(z)| \le |S_{\mu_n(\omega)}(z') - S_{\mu(\omega)}(z')| + \frac{2|z-z'|}{|\Im z||\Im z'|} \tag{8.2}$$

which holds for any z'.

Now, fix $\epsilon > 0$. To simplify, let us assume $\Im z > 0$ (a nearly identical argument works in the case $\Im z < 0$). Let ¿0 be such that $\frac{1}{\delta} > \Im z + \frac{4}{\epsilon}$. By the density of $\mathbb{Q}(i) \setminus \mathbb{Q}$ in $\mathbb{C} \setminus \mathbb{R}$, there is a rational point z' with $|z - z'| < (\Im z)^2 \delta$. Then we also have $|\Im z - \Im z'| = |\Im(z - z')| \le |z - z'| < (\Im z)^2 \delta$, and so, in particular, $\Im z' > \Im z - (\Im z)^2 \delta = \Im z(1 - \Im z \delta)$. By the assumption on δ, $1 - \Im z \delta > 4/\epsilon > 0$. Thus

$$\frac{2|z-z'|}{|\Im z||\Im z'|} = \frac{2|z-z'|}{\Im z \cdot \Im z'} < \frac{2(\Im z)^2 \delta}{\Im z \cdot \Im z (1 - \Im z \delta)} = \frac{2\delta}{1 - \Im z \delta} < \frac{2\delta}{4/\epsilon} = \frac{\epsilon}{2}.$$

Now, since $S_{\mu_n(\omega)}(z') \to S_{\mu(\omega)}(z')$, there is an $N(\omega) \in \mathbb{N}$ such that, for all $n \ge N(\omega)$, $|S_{\mu_n(\omega)}(z') - S_{\mu(\omega)}(z')| < \frac{\epsilon}{2}$. From (8.2), it therefore follows that, for $n \ge N(\omega)$, $|S_{\mu_n(\omega)}(z) - S_{\mu(\omega)}(z)| < \epsilon$. This establishes the claim.

Now with Claim 8.13 in hand: from Proposition 8.9(b), we may conclude that, for each $\omega \in \Omega$, $\mu_n(\omega) \to \mu(\omega)$ vaguely. Since both $\mu_n(\omega)$ and $\mu(\omega)$ are known a priori to be probability measures, it therefore follows from Lemma 8.2 that $\mu_n(\omega) \to \mu(\omega)$ weakly for all $\omega \in \Omega$. Since $\mathbb{P}(\Omega) = 1$, this proves the theorem. □

8.3. **The Stieltjes Transform of an Empirical Eigenvalue Measure.** Let $\mathbf{X}_n$ be a Wigner matrix, with (random) empirical law of eigenvalues $\mu_{\mathbf{X}_n}$. Let $\bar{\mu}_{\mathbf{X}_n}$ denote the *averaged* empirical eigenvalue law: that is, $\bar{\mu}_{\mathbf{X}_n}$ is determined by

$$\int_{\mathbb{R}} f\, d\bar{\mu}_{\mathbf{X}_n} = \mathbb{E}\left(\int f\, d\mu_{\mathbf{X}_n}\right) \quad \forall\, f \in C_b(\mathbb{R}). \tag{8.3}$$

The Stieltjes transforms of these measures are easily expressible in terms of the matrix $\mathbf{X}_n$, without explicit reference to the eigenvalues $\lambda_i(\mathbf{X}_n)$. Indeed, for $z \in \mathbb{C} \setminus \mathbb{R}$,

$$S_{\mu_{\mathbf{X}_n}}(z) = \int_{\mathbb{R}} \frac{1}{t-z}\, \mu_{\mathbf{X}_n}(dt) = \frac{1}{n}\sum_{i=1}^n \frac{1}{\lambda_i(\mathbf{X}_n) - z}.$$

Now, for a real symmetrix matrix $\mathbf{X}$, the eigenvalues $\lambda_i(\mathbf{X})$ are real, and so the matrix $\mathbf{X} - z\mathbf{I}$ (where $\mathbf{I}$ is the $n \times n$ identity matrix) is invertible for each $z \in \mathbb{C} \setminus \mathbb{R}$. Accordingly, define

$$S_{\mathbf{X}}(z) = (\mathbf{X} - z\mathbf{I})^{-1}.$$

(This function is often called the *resolvent* of $\mathbf{X}$.) By the spectral theorem, the eigenvalues of $S_{\mathbf{X}}(z)$ are $(\lambda_i(\mathbf{X}) - z)^{-1}$. Thus, we have

$$S_{\mu_{\mathbf{X}_n}}(z) = \frac{1}{n}\sum_{i=1}^n \lambda_i(S_{\mathbf{X}_n}(z)) = \frac{1}{n}\operatorname{Tr} S_{\mathbf{X}_n}(z). \tag{8.4}$$

Now, since the function $f_z(t) = \frac{1}{t-z}$ is in $C_b(\mathbb{R})$, Equation 8.3 yields

$$\mathbb{E}S_{\mu_{\mathbf{X}_n}} = \mathbb{E}\left(\int f_z\, d\mu_{\mathbf{X}_n}\right) = \int f_z\, d\bar{\mu}_{\mathbf{X}_n} = S_{\bar{\mu}_{\mathbf{X}_n}}(z),$$

and so we have the complementary formula

$$S_{\bar{\mu}_{\mathbf{X}_n}}(z) = \frac{1}{n}\mathbb{E}\operatorname{Tr} S_{\mathbf{X}_n}(z). \tag{8.5}$$

We will now proceed to use (matrix-valued) calculus on the functions $\mathbf{X} \mapsto S_{\mathbf{X}}(z)$ (for fixed z) to develop an entirely different approach to Wigner's Semicircle Law.

9. The Stieltjes Transform and the Averaged Empirical Law of Eigenvalues

9.1. **Gaussian Random Matrices.** To get a sense for how the Stieltjes transform may be used to prove Wigner's theorem, we begin by considering a very special matrix model. We take a Wigner matrix $\mathbf{X}_n = n^{-1/2}\mathbf{Y}_n$ with the following characteristics:

$$Y_{ii} = 0, \qquad Y_{ij} = Y_{ji} \sim N(0,1) \text{ for } i < j.$$

That is: we assume that the diagonal entries are 0 (after all, we know they do not contribute to the limit), and the off-diagonal entries are standard normal random variables, independent above the main diagonal. Now, we may identify the space of symmetric, 0-diagonal $n \times n$ matrices with $\mathbb{R}^{n(n-1)/2}$ in the obvious manner. Under this identification, $\mathbf{Y}_n$ becomes a random vector whose joint law is the standard $n(n-1)/2$-dimensional Gaussian.

With this in mind, let us attempt to calculuate the Stieltjes transform of $\bar{\mu}_{\mathbf{X}_n}$, a la Equation 8.5. First, note the following identity for the resolvent $S_{\mathbf{X}}$ of any matrix $\mathbf{X}$. Since $S_{\mathbf{X}}(z) = (\mathbf{X} - z\mathbf{I})^{-1}$, we have $(\mathbf{X} - z\mathbf{I})S_{\mathbf{X}}(z) = \mathbf{I}$. In other words, $\mathbf{X}S_{\mathbf{X}}(z) - zS_{\mathbf{X}}(z) = \mathbf{I}$, and so

$$S_{\mathbf{X}}(z) = \frac{1}{z}(\mathbf{X}S_{\mathbf{X}}(z) - \mathbf{I}). \tag{9.1}$$

Taking the normalized trace, we then have

$$\frac{1}{n}\operatorname{Tr} S_{\mathbf{X}}(z) = \frac{1}{nz}\operatorname{Tr} \mathbf{X}S_{\mathbf{X}}(z) - \frac{1}{z}. \tag{9.2}$$

Now, as with the Gaussian matrix above, assume that $\mathbf{X}$ has 0 diagonal. In accordance with thinking of $\mathbf{X}$ as a vector in $\mathbb{R}^{n(n-1)/2}$, we introduce some temporary new notation: for $z \in \mathbb{C} \setminus \mathbb{R}$, let $\mathbf{F}^z(\mathbf{X}) = S_{\mathbf{X}}(z)$; that is, we think of $\mathbf{X} \mapsto S_{\mathbf{X}}(z)$ as a ($\mathbb{C} \setminus \mathbb{R}$-parametrized) vector-field $\mathbf{F}^z$. The inverse of a (complex) symmetric matrix is (complex) symmetric, and so $\mathbf{F}^z(\mathbf{X})$ is symmetric. It generally has non-zero diagonal, however. Nevertheless, let us write out the trace term:

$$\operatorname{Tr} \mathbf{X}S_{\mathbf{X}}(z) = \operatorname{Tr} \mathbf{X}\mathbf{F}^z(\mathbf{X}) = \sum_{i=1}^n [\mathbf{X}\mathbf{F}^z(\mathbf{X})]_{ii} = \sum_{i,j} [\mathbf{X}]_{ij}[\mathbf{F}^z(\mathbf{X})]_{ji}.$$

Since $[\mathbf{X}]_{ii} = 0$, and since both $\mathbf{X}$ and $\mathbf{F}^z(\mathbf{X})$ are symmetric, we can rewrite this as

$$\sum_{i \neq j} [\mathbf{X}]_{ij}[\mathbf{F}^z(\mathbf{X})]_{ij} = 2\sum_{i<j} [\mathbf{X}]_{ij}[\mathbf{F}^z(\mathbf{X})]_{ij}.$$

So, although $\mathbf{F}^z$ takes values in a space larger than $\mathbb{R}^{n(n-1)/2}$, the calculation only requires its values on $\mathbb{R}^{n(n-1)/2}$, and so we restrict our attention to *those* components of $\mathbf{F}^z$. Let us now use notation more becoming of a Euclidean space: the product above is nothing else but the dot product of $\mathbf{X}$ with the vector field $\mathbf{F}^z(\mathbf{X})$.

In the case $\mathbf{X} = \mathbf{X}_n$, the Gaussian Wigner matrix above, let γ denote the joint law of the random vector $\mathbf{X} \in \mathbb{R}^{n(n-1)/2}$. Then, taking expectation, we have (from Equation 9.2)

$$\begin{aligned}\frac{1}{n}\mathbb{E}\operatorname{Tr} S_{\mathbf{X}_n}(z) = \frac{1}{nz}\mathbb{E}\operatorname{Tr} \mathbf{X}_n S_{\mathbf{X}_n} - \frac{1}{z} &= \frac{2}{nz}\mathbb{E}\left(\mathbf{X}_n \cdot \mathbf{F}^z(\mathbf{X}_n)\right) - \frac{1}{z} \\ &= -\frac{1}{z} + \frac{2}{nz}\int_{\mathbb{R}^{n(n-1)/2}} \mathbf{x} \cdot \mathbf{F}^z(\mathbf{x})\,\gamma(d\mathbf{x}).\end{aligned}$$

Evaluating this integral can be accomplished through a standard integration by parts formula for Gaussians.

Lemma 9.1 (Gaussian integration by parts)**.** *Let γ_t denote the standard Gaussian law on $\mathbb{R}^m$, with variance $t>0$:*

$$\gamma_t(d\mathbf{x}) = (2\pi t)^{-m/2} e^{-|\mathbf{x}|^2/2t}\, d\mathbf{x}.$$

Let $\mathbf{F}\colon \mathbb{R}^m \to \mathbb{R}^m$ be C^1 with F_j and $\partial_j F_j$ of polynomial growth for all j. Then

$$\int \mathbf{x}\cdot\mathbf{F}(\mathbf{x})\,\gamma_t(d\mathbf{x}) = t\int \nabla\cdot\mathbf{F}(\mathbf{x})\,\gamma_t(d\mathbf{x})$$

Proof. We will prove the Lemma term by term in the dot-product sum; that is, for each j,

$$\int x_j F_j(\mathbf{x})\,\gamma_t(d\mathbf{x}) = t\int \partial_j F_j(\mathbf{x})\,\gamma_t(d\mathbf{x}).$$

As such, we set $F=F_j$ and only deal with a single function. We simply integrate by parts on the right-hand-side:

$$\begin{aligned}\int \partial_j F\,d\gamma_t &= (2\pi t)^{-m/2}\int (\partial_j F)(\mathbf{x})e^{-|\mathbf{x}|^2/2t}\,d\mathbf{x}\\ &= -(2\pi t)^{-m/2}\int F(\mathbf{x})\partial_j(e^{-|\mathbf{x}|^2/2t})\,d\mathbf{x}\\ &= -(2\pi t)^{-m/2}\int F(\mathbf{x})(-x_j/t)e^{-|\mathbf{x}|^2/2t}\,d\mathbf{x}\\ &= \frac{1}{t}\int x_j F(\mathbf{x})\,\gamma_t(d\mathbf{x}),\end{aligned}$$

where the integration by parts is justified since $\partial_j F \in L^1(\gamma_t)$ and $F\in L^1(x_j\gamma_t)$ due to the polynomial-growth assumption. □

The Gaussian Wigner matrix $\mathbf{X}_n = n^{-1/2}\mathbf{Y}_n$ is built out of standard normal entries Y_{ij}, with variance 1. Thus the variance of the entries in $\mathbf{X}_n$ is $\frac{1}{n}$. The vector field $\mathbf{F}^z$ has the form $\mathbf{F}^z(\mathbf{X}) = S_{\mathbf{X}}(z) = (\mathbf{X}-z\mathbf{I})^{-1}$ which is a bounded function of $\mathbf{X}$. The calculations below show that its partial derivatives are also bounded, so we may use Lemma 9.1.

$$\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,S_{\mathbf{X}_n}(z) = -\frac{1}{z} + \frac{2}{n^2 z}\int_{\mathbb{R}^{n(n-1)/2}} \nabla\cdot\mathbf{F}^z(\mathbf{x})\,\gamma_{1/n}(d\mathbf{x}) = -\frac{1}{z}+\frac{2}{n^2 z}\mathbb{E}\left(\nabla\cdot\mathbf{F}^z(\mathbf{X}_n)\right).$$

We must now calculate the divergence $\nabla\cdot\mathbf{F}^z$, where $[\mathbf{F}^z(\mathbf{X})]_{ij} = [S_{\mathbf{X}}(z)]_{ij}$ for $i<j$. To begin, we calculate the partial derivatives of the (vector-valued) function $\mathbf{F}^z$. By definition

$$\partial_{ij}\mathbf{F}^z(\mathbf{X}) = \lim_{h\to 0}\frac{1}{h}\left[\mathbf{F}^z(\mathbf{X}+h\mathbf{E}_{ij}) - \mathbf{F}^z(\mathbf{X})\right],$$

where $\mathbf{E}_{ij}$ is the standard basis vector in the ij-direction. Under the identification of $\mathbb{R}^{n(n-1)/2}$ with the space of symmetric 0-diagonal matrices, this means that $\mathbf{E}_{ij}$ is the matrix with 1s in the ij and ji slots, and 0s elsewhere. The difference then becomes

$$S_{\mathbf{X}+h\mathbf{E}_{ij}}(z) - S_{\mathbf{X}}(z) = (\mathbf{X}+h\mathbf{E}_{ij}-z\mathbf{I})^{-1} - (\mathbf{X}-z\mathbf{I})^{-1}.$$

We now use the fact that $A^{-1}-B^{-1} = A^{-1}(B-A)B^{-1}$ for two invertible matrices A,B to conclude that

$$S_{\mathbf{X}+h\mathbf{E}_{ij}}(z) - S_{\mathbf{X}}(z) = (\mathbf{X}+h\mathbf{E}_{ij}-z\mathbf{I})^{-1}\left(-h\mathbf{E}_{ij}\right)(\mathbf{X}-z\mathbf{I})^{-1}.$$

Dividing by h and taking the limit yields

$$\partial_{ij}\mathbf{F}^z(\mathbf{X}) = -S_{\mathbf{X}}(z)\mathbf{E}_{ij}S_{\mathbf{X}}(z).$$

Now, the divergence in question is

$$\nabla \cdot \mathbf{F}^z(\mathbf{X}) = \sum_{i<j} \partial_{ij} \mathbf{F}^z_{ij}(\mathbf{X}) = -\sum_{i<j} [S_{\mathbf{X}}(z)\mathbf{E}_{ij}S_{\mathbf{X}}(z)]_{ij}.$$

Well, for any matrix A,

$$[A\mathbf{E}_{ij}A]_{ij} = \sum_{k,\ell} A_{ik}[E_{ij}]_{k\ell}A_{\ell j} = \sum_{k,\ell} A_{ik}(\delta_{ik}\delta_{j\ell} + \delta_{i\ell}\delta_{jk})A_{\ell j} = A_{ii}A_{jj} + A_{ij}A_{ji}.$$

Thus, we have the divergence of $\mathbf{F}^z$:

$$\nabla \cdot \mathbf{F}^z(\mathbf{X}) = -\sum_{i<j} \left([S_{\mathbf{X}}(z)]_{ii}[S_{\mathbf{X}}(z)]_{jj} + [S_{\mathbf{X}}(z)]_{ij}^2\right). \tag{9.3}$$

Combining with the preceding calculations yields

$$\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\, S_{\mathbf{X}_n}(z) = -\frac{1}{z} - \frac{2}{n^2 z}\mathbb{E}\left(\sum_{1\le i<j\le n} ([S_{\mathbf{X}_n}(z)]_{ii}[S_{\mathbf{X}_n}(z)]_{jj}) + \mathbb{E}\left([S_{\mathbf{X}_n}(z)]_{ij}^2\right)\right). \tag{9.4}$$

Let $s_{ij} = [S_{\mathbf{X}_n}(z)]_{ij}$; then $s_{ij} = s_{ji}$. So, the first term in the above sum is

$$2\sum_{i<j} s_{ii}s_{jj} = \sum_{i\neq j} s_{ii}s_{jj} = \sum_{i,j} s_{ii}s_{jj} - \sum_i s_{ii}^2 = \left(\sum_i s_{ii}\right)^2 - \sum_i s_{ii}^2.$$

The second term is

$$2\sum_{i<j} s_{ij}^2 = \sum_{i\neq j} s_{ij}^2 = \sum_{i,j} s_{ij}^2 - \sum_i s_{ii}^2.$$

Two of these terms can be expressed in terms of traces of powers of $S_{\mathbf{X}}(z)$:

$$\sum_i s_{ii} = \mathrm{Tr}\, S_{\mathbf{X}}(z), \qquad \sum_{i,j} s_{ij}^2 = \mathrm{Tr}\,(S_{\mathbf{X}}(z)^2).$$

Plugging these into Equation 9.4 yields

$$\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\, S_{\mathbf{X}_n}(z) = -\frac{1}{z} - \frac{1}{n^2 z}\mathbb{E}\left((\mathrm{Tr}\, S_{\mathbf{X}_n}(z))^2 + \mathrm{Tr}\,(S_{\mathbf{X}_n}(z)^2) - 2\sum_{i=1}^n [S_{\mathbf{X}_n}(z)]_{ii}^2\right). \tag{9.5}$$

It is convenient to rewrite this as

$$1 + z\cdot\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\, S_{\mathbf{X}_n}(z) + \mathbb{E}\left(\left(\frac{1}{n}\mathrm{Tr}\, S_{\mathbf{X}_n}(z)\right)^2\right) = -\frac{1}{n^2}\mathbb{E}\left(\mathrm{Tr}\,(S_{\mathbf{X}_n}(z)^2) - 2\sum_{i=1}^n [S_{\mathbf{X}_n}(z)]_{ii}^2\right).$$

Employing Equation 8.4, the left-hand-side can be rewritten as

$$1 + z\mathbb{E}S_{\mu_{\mathbf{X}_n}}(z) + \mathbb{E}(S_{\mu_{\mathbf{X}_n}}(z)^2).$$

Let us now estimate the right-hand-side. For the second term, we can make a blunt estimate:

$$\left|\sum_{i=1}^n [S_{\mathbf{X}_n}(z)]_{ii}^2\right| \le \sum_{1\le i,j\le n} |[S_{\mathbf{X}_n}(z)]_{ij}|^2 = \mathrm{Tr}\,(S_{\mathbf{X}_n}(z)^* S_{\mathbf{X}_n}(z)).$$

Similarly, the trace term is bounded in absolute value by the same quantity. Hence, the right-hand-side is bounded in absolute value as follows:

$$\left|1+z\cdot\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,S_{\mathbf{X}_n}(z)+\mathbb{E}\left(\left(\frac{1}{n}\,\mathrm{Tr}\,S_{\mathbf{X}_n}(z)\right)^2\right)\right|\le\frac{3}{n^2}\mathbb{E}\,\mathrm{Tr}\,(S_{\mathbf{X}_n}(z)^*S_{\mathbf{X}_n}(z)). \tag{9.6}$$

But $S_{\mathbf{X}_n}(z)^*=S_{\mathbf{X}_n}(\bar{z})$ since $\mathbf{X}_n$ is a real matrix. So we have

$$\begin{aligned}\frac{1}{n^2}\mathbb{E}\,\mathrm{Tr}\,(S_{\mathbf{X}_n}(z)S_{\mathbf{X}_n}(\bar{z}))=\frac{1}{n^2}\mathbb{E}\,\mathrm{Tr}\,|\mathbf{X}_n-z\mathbf{I}|^{-2}&=\frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^n|\lambda_i(\mathbf{X}_n)-z|^{-2}\right)\\&=\frac{1}{n}\int\frac{1}{|t-z|^2}\,\bar{\mu}_{\mathbf{X}_n}(dt).\end{aligned}$$

Note that, for fixed $z\in\mathbb{C}\setminus\mathbb{R}$, we have $|t-z|^{-2}\le|\Im z|^{-2}$ for all $t\in\mathbb{R}$; thus, since $\bar{\mu}_{\mathbf{X}_n}$ is a probability measure, it follows that the last quantity is $\le\frac{1}{|\Im z|^2 n}$. Equation 9.6 therefore shows that

$$1+z\mathbb{E}S_{\mu_{\mathbf{X}_n}}(z)+\mathbb{E}(S_{\mu_{\mathbf{X}_n}}(z)^2)\to 0\quad\text{as}\quad n\to\infty. \tag{9.7}$$

Now, at this point, we need to make a leap of faith (which we will, in the next lectures, proceed to justify). Suppose that we can bring the expectation inside the square in this term; that is, suppose that it holds true that

$$\mathbb{E}(S_{\mu_{\mathbf{X}_n}}(z)^2)-\left(\mathbb{E}S_{\mu_{\mathbf{X}_n}}(z)\right)^2\to 0\quad\text{as}\quad n\to\infty. \tag{9.8}$$

Combining Equations 9.7 and 9.8, and utilizing the fact that $\mathbb{E}S_{\mu_{\mathbf{X}_n}}(z)=S_{\bar{\mu}_{\mathbf{X}_n}}(z)$ (cf. Equation 8.5), we would then have, for each $z\in\mathbb{C}\setminus\mathbb{R}$

$$1+zS_{\bar{\mu}_{\mathbf{X}_n}}(z)+S_{\bar{\mu}_{\mathbf{X}_n}}(z)^2\to 0\quad\text{as}\quad n\to\infty.$$

For fixed z, the sequence $\bar{S}_n(z)=S_{\bar{\mu}_{\mathbf{X}_n}}(z)$ is contained in the closed ball of radius $\frac{1}{|\Im z|^2}$ in $\mathbb{C}$. This ball is compact, hence there is a subsequence $\bar{S}_{n_k}(z)$ that converges to a limit $\bar{S}(z)$. This limit then satisfies the quadratic equation

$$1+z\bar{S}(z)+\bar{S}(z)^2=0$$

whose solutions are

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

Now, $\bar{S}(z)$ is the limit of a sequence of Stieltjes transforms of probability measures; hence, by Proposition 8.9(b), there is a positive finite measure μ such that $\bar{S}(z)=S_\mu(z)$. It follows, then, since $\bar{S}(z)$ is a Stieltjes transform, that for $\Im z>0$,

$$\Im\bar{S}(z)=\Im S_\mu(z)=\Im\int_{\mathbb{R}}\frac{\mu(dt)}{t-z}=\int_{\mathbb{R}}\frac{y}{(t-x)^2+y^2}\,\mu(dt)>0.$$

Hence, this shows that the correct sign choice is $+$, uniformly: we have

$$\bar{S}(z)=\frac{-z+\sqrt{z^2-4}}{2}=S_{\sigma_1}(z),$$

the Stieltjes transform of the semicircle law. Thus, we have shown that every convergent subsequence of $S_{\bar{\mu}_{\mathbf{X}_n}}(z)$ converges to $S_{\sigma_1}(z)$; it follows that the limit exists, and

$$\lim_{n\to\infty}S_{\bar{\mu}_{\mathbf{X}_n}}(z)=S_{\sigma_1}(z),\quad z\in\mathbb{C}\setminus\mathbb{R}.$$

Finally, by Theorem 8.11, it follows that $\bar{\mu}_{\mathbf{X}_n} \to \sigma_1$ weakly, which verifies Wigner's semicircle law (on the level of expectations) for the Gaussian Wigner matrix $\mathbf{X}_n$.

Now, the key to this argument was the interchange of Equation 9.8. But looking once more at this equation, we see what it says is that $\mathrm{Var} S_{\mu_{\mathbf{X}_n}}(z) \to 0$. In fact, if we know this, then the theorem just proved – that the *averaged* empirical eigenvalue distribution $\bar{\mu}_{\mathbf{X}_n}$ converges weakly to σ_1 – yields stronger convergence. Indeed, we will see that this kind of *concentration of measure* assumption implies that the random measure $\mu_{\mathbf{X}_n}$ converges weakly almost surely to σ_1, giving us the full power of Wigner's semicircle law.

Before we can get there, we need to explore some so-called *coercive functional inequalities* for providing concentration of measure. First, let us consider how to work this same kind of argument for a more general Wigner matrix.

9.2. **More General Wigner Matrices.** In the vector calculus approach taken in the previous section, Gaussian integration by parts (Lemma 9.1) played an important role. In fact, this integration by parts formula characterizes Gaussians (this is Stein's lemma), and so for more general Wigner matrices, we need a somewhat different approach.

The first step is a general linear algebra lemma, regarding the resolvent function $S_{\mathbf{X}}(z)$ of a generic symmetric matrix $\mathbf{X}$.

Lemma 9.2. *Let $\mathbf{X} \in \mathscr{X}_n$ be a symmetric $n \times n$ matrix, with columns $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n]$. For $1 \le j \le n$, denote by $\tilde{\mathbf{x}}_j \in \mathbb{R}^{n-1}$ the vector achieved by deleting the jth component of $\mathbf{x}_j$, and let $\mathbf{X}^{(j)} \in \mathscr{X}_n$ denote the symmetric matrix obtained by deleting the jth column and row from $\mathbf{X}$. Then, for $z \in \mathbb{C} \setminus \mathbb{R}$,*

$$[S_{\mathbf{X}}(z)]_{jj} = \frac{1}{\mathbf{X}_{jj} - z - \mathbf{x}_j^\top S_{\mathbf{X}^{(j)}}(z)\mathbf{x}_j}.$$

Proof. To be clear: the statement is that

$$[(\mathbf{X} - zI_n)^{-1}]_{jj} = \frac{1}{\mathbf{X}_{jj} - z - \mathbf{x}_j^\top (\mathbf{X}^{(j)} - zI_{n-1})^{-1}\mathbf{x}_j}.$$

To prove this, we use Cramer's rule, which tells us directly that

$$[(\mathbf{X} - zI_n)^{-1}]_{jj} = \frac{\det(\mathbf{X}^{(j)} - zI_{n-1})}{\det(\mathbf{X} - zI_n)}.$$

Now, write out $\mathbf{X} - zI_n$ in block form:

$$\mathbf{X} - zI_n = \begin{bmatrix} \mathbf{X}^{(n)} - zI_{n-1} & \tilde{\mathbf{x}}_n \\ \tilde{\mathbf{x}}_j^\top & \mathbf{X}_{nn} - z \end{bmatrix}.$$

Now we use the following block-diagonal matrix identity: if A, B, C, D are matrices of appropriate dimension and A is invertible, then

$$\det \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \det(A) \cdot \det(D - CA^{-1}B).$$

Applying this with $A = \mathbf{X}^{(n)} - zI_{n-1}$, $B = C^\top = \tilde{\mathbf{x}}_n$, and $D = \mathbf{X}_{nn} - z$ proves the result in the case $j = n$; other j follow from the same argument by first conjugating $\mathbf{X} - zI$ by the permutation matrix that interchanges columns j and n. □

Now, let $\mathbf{X}_n = \frac{1}{\sqrt{n}}\mathbf{Y}_n$ be a Wigner matrix. As above, we may assume that the diagonal is identically $0 = [\mathbf{X}]_{jj}$. We continue to use the notations $\tilde{\mathbf{x}}_j$ and $\mathbf{X}^{(j)}$ from Lemma 9.2, suppressing the explicit n-dependence unless necessary for clarity. The lemma therefore implies that

$$\frac{1}{n}\operatorname{Tr}(S_{\mathbf{X}_n}) = \frac{1}{n}\sum_{j=1}^{n}\frac{1}{-z-\tilde{\mathbf{x}}_j^\top(\mathbf{X}^{(j)}-zI)^{-1}\tilde{\mathbf{x}}_j}.$$

Now, multiply through as follows:

$$\begin{aligned}\left(z+\frac{1}{n}\operatorname{Tr}(S_{\mathbf{X}_n}(z))\right)\cdot\frac{1}{n}\operatorname{Tr}(S_{\mathbf{X}_n}) &= \left(z+\frac{1}{n}\operatorname{Tr}(S_{\mathbf{X}_n}(z))\right)\cdot\frac{1}{n}\sum_{j=1}^{n}\frac{1}{-z-\tilde{\mathbf{x}}_j^\top(\mathbf{X}^{(j)}-zI)^{-1}\tilde{\mathbf{x}}_j}\\ &= -1+\delta_n(z),\end{aligned}$$

where

$$\delta_n(z) = \frac{1}{n}\sum_{j=1}^{n}\frac{\epsilon_{j,n}(z)}{-z-\frac{1}{n}\operatorname{Tr}(S_{\mathbf{X}_n}(z))+\epsilon_{j,n}(z)},$$

and

$$\epsilon_{j,n}(z) = \frac{1}{n}\operatorname{Tr}(S_{\mathbf{X}_n}(z)) - \tilde{\mathbf{x}}_j^\top(\mathbf{X}^{(j)}-zI)^{-1}\tilde{\mathbf{x}}_j.$$

Now, as in the previous section, we have $\frac{1}{n}\operatorname{Tr}(S_{\mathbf{X}_n}(z)) = S_{\mu_{\mathbf{X}_n}}(z)$, and so we have

$$S_{\mu_{\mathbf{X}_n}}(z)^2 + zS_{\mu_{\mathbf{X}_n}}(z) + 1 = \delta_n(z).$$

Taking expectations, we see that the argument (that $S_{\overline{\mu}_{\mathbf{X}_n}}(z) \to S_{\sigma_1}(z)$ for all $z \in \mathbb{C}\setminus\mathbb{R}$) proceeds exactly as it did following (9.7), provided we can show that $\mathbb{E}(\delta_n(z)) \to 0$. In fact, we will show that $\delta_n(z) \to 0$ in probability for each $z \in \mathbb{C}\setminus\mathbb{R}$. To do so, it suffices to prove that $\sup_{j\le n}|\epsilon_{j,n}(z)| \to 0$ in probability as $n\to\infty$. This is somewhat tricky, but not very deep to prove: see page 49 of “An Introduction to Random Matrices” by Anderson, Guionnet, and Zeitouni.

10. Logarithmic Sobolev Inequalities

We consider here some ideas that fundamentally come from information theory. Let μ, ν be probability measures on $\mathbb{R}^m$. The **entropy of** ν relative to μ is

$$\mathrm{Ent}_\mu(\nu) = \int_{\mathbb{R}^m} \frac{d\nu}{d\mu} \log \frac{d\nu}{d\mu}\, d\mu$$

if ν is absolutely continuous with respect to μ, and $\mathrm{Ent}_\mu(\nu) = +\infty$ if not. We will utilize entropy, thinking of the input not as a measure ν, but as its density $f = d\nu/d\mu$. In fact, we can be even more general than this. Let $f\colon \mathbb{R}^m \to \mathbb{R}_+$ be a non-negative function in $L^1(\mu)$. Then $\hat{f} = f/\|f\|_1$ (where $\|f\|_1 = \|f\|_{L^1(\mu)}$) is a probability density with respect to μ. So we have

$$\begin{aligned}\mathrm{Ent}_\mu(\hat{f}\, d\mu) = \int_{\mathbb{R}^m} \hat{f} \log \hat{f}\, d\mu &= \int_{\mathbb{R}^m} \frac{f}{\|f\|_1} \log \frac{f}{\|f\|_1}\, d\mu \\ &= \frac{1}{\|f\|_1}\left(\int_{\mathbb{R}^m} f \log f\, d\mu - \log \|f\|_1 \int_{\mathbb{R}^m} f\, d\mu\right).\end{aligned}$$

This allows us to define the entropy of a non-negative function, regardless of whether it is a probability density. Abusing notation slightly (and leaving out the global factor of $1/\|f\|_1$ as is customary), we define

$$\mathrm{Ent}_\mu(f) = \int_{\mathbb{R}^m} f \log f\, d\mu - \int_{\mathbb{R}^m} f\, d\mu \cdot \log \int_{\mathbb{R}^m} f\, d\mu.$$

This quantity is not necessarily finite, even for $f \in L^1(\mu)$. It is, however, finite, provided $\int f \log(1+f)\, d\mu < \infty$. This condition defines an *Orlicz space*. In general, for $0 < p < \infty$, set

$$L^p \log L(\mu) = \{f\colon \int |f|^p \log(1+|f|)\, d\mu < \infty\}.$$

For any p and any $\epsilon > 0$, $L^p(\mu) \supset L^p \log L(\mu) \supset L^{p+\epsilon}(\mu)$; $L^p \log L$ is *infinitesimally smaller* than L^p. In particular, if $f \in L^1 \log L$ then $f \in L^1$, and so $f \in L^1 \log L$ implies that $\mathrm{Ent}_\mu(f) < \infty$.

The function $[0,\infty) \ni x \mapsto \varphi(x) = x \log x$ is convex. Note that

$$\mathrm{Ent}_\mu(f) = \int_{\mathbb{R}^m} \varphi(f)\, d\mu - \varphi\left(\int_{\mathbb{R}^m} f\, d\mu\right)$$

and since μ is a probability measure, by Jensen's inequality we have $\mathrm{Ent}_\mu(f) \geq 0$ for all f. It is also useful to note that, for scalars $\alpha > 0$,

$$\mathrm{Ent}_\mu(\alpha f) = \int_{\mathbb{R}^m} (\alpha f) \log \frac{\alpha f}{\int_{\mathbb{R}^m} \alpha f\, d\mu}\, d\mu = \alpha \mathrm{Ent}_\mu(f).$$

That is, Ent_μ is a positive functional, homogeneous of degree 1.

Definition 10.1. *The probability measure μ on $\mathbb{R}^m$ is said to satisfy the* **logarithmic Sobolev inequality** *with constant $c > 0$ if, for all sufficiently smooth $f \in L^2 \log L$,*

$$\mathrm{Ent}_\mu(f^2) \leq 2c \int_{\mathbb{R}^m} |\nabla f|^2\, d\mu. \tag{LSI}$$

On the right-hand-side, ∇f is the usual gradient of f, and $|\cdot|$ denotes the Euclidean length. The integral of $|\nabla f|^2$ is a measure of the *energy* in f (relative to the state μ). It might be better to refer to Inequality LSI as an **energy-entropy** inequality. The inequality was discovered by L. Gross, in a context where it made sense to compare it to Sobolev inequalities (which are used to prove smoothness properties of solutions of differential equations).

Remark 10.2. Since $f \mapsto \int |\nabla f|^2\, d\mu$ is homogeneous of order 2, we must take $\mathrm{Ent}_\mu(f^2)$ in LSI; if the two sides scale differently, no inequality could possibly hold in general. An alternative formulation, however, is to state Inequality LSI in terms of the function $g = f^2$. Using the chain rule, we have

$$\nabla f = \nabla(\sqrt{g}) = \frac{1}{2\sqrt{g}}\nabla g$$

and so we may restate the log-Sobolev inequality as

$$\mathrm{Ent}_\mu(g) \le \frac{c}{2}\int_{\mathbb{R}^m} \frac{|\nabla g|}{g}\, d\mu \tag{LSI'}$$

which should hold for all $g = f^2$ where $f \in L^2 \log L$ is sufficiently smooth; this is equivalent to $g \in L^1 \log L$ being sufficiently smooth. The quantities in Inequality LSI' are even more natural from an information theory perspective: the left-hand-side is Entropy, while the right-hand-side is known as *Fisher information*.

The first main theorem (and the one most relevant to our cause) about log-Sobolev inequalities is that Gaussian measures satisfy LSIs.

Theorem 10.3. *For $t > 0$, the Gaussian measure $\gamma_t(d\mathbf{x}) = (2\pi t)^{-m/2}e^{-|\mathbf{x}|^2/2t}\, d\mathbf{x}$ on $\mathbb{R}^m$ satisfies Inequality LSI with constant $c = t$.*

Remark 10.4. Probably the most important feature of this theorem (and the main reason LSIs are useful) is that the constant c is *independent of dimension*.

Remark 10.5. In fact, it is sufficient to prove the Gaussian measure γ_1 satisfies the LSI with constant $c = 1$. Assuming this is the case, the general statement of Theorem 10.3 follows by a change of variables. Indeed, for any measurable function $\varphi\colon \mathbb{R} \to \mathbb{R}$, we have

$$\int_{\mathbb{R}^m} (\varphi \circ f)\, d\gamma_t = (2\pi t)^{-m/2} \int_{\mathbb{R}^m} \varphi(f(\mathbf{x}))\, e^{-|\mathbf{x}|^2/2t}\, d\mathbf{x} = (2\pi t)^{-m/2}\int_{\mathbb{R}^m} \varphi(f(\sqrt{t}\mathbf{y}))e^{-|\mathbf{y}|^2/2}t^{m/2}d\mathbf{y}$$

where we have substituted $\mathbf{y} = \mathbf{x}/\sqrt{t}$. Setting $f_t(\mathbf{x}) = f(\sqrt{t}\mathbf{x})$, this shows that

$$\int_{\mathbb{R}^m} (\varphi \circ f)\, d\gamma_t = \int_{\mathbb{R}^m} (\varphi \circ f_t)\, d\gamma_1. \tag{10.1}$$

Applying this with $\varphi(x) = x\log x$ in the first term and $\varphi(x) = x$ in each half of the product in the second term of Ent_{γ_t}, this shows $\mathrm{Ent}_{\gamma_t}(f) = \mathrm{Ent}_{\gamma_1}(f_t)$. Scaling the argument preserves the spaces $L^p \log L$ (easy to check), and also preserves all smoothness classes, so by the assumption of LSI for γ_1, we have

$$\mathrm{Ent}_{\gamma_1}(f_t) \le 2\int_{\mathbb{R}^m} |\nabla f_t|^2\, d\gamma_1.$$

Now, note that

$$\nabla f_t(\mathbf{x}) = \nabla(f(\sqrt{t}\mathbf{x})) = \sqrt{t}(\nabla f)(\sqrt{t}\mathbf{x}) = \sqrt{t}(\nabla f)_t(\mathbf{x}).$$

Hence

$$\mathrm{Ent}_{\gamma_1}(f_t) \le 2\int_{\mathbb{R}^m} |\sqrt{t}(\nabla f)_t|^2\, d\gamma_1 = 2t\int_{\mathbb{R}^m} |\nabla f|_t^2\, d\gamma_1.$$

Now, applying Equation 10.1 with $\varphi(x) = x^2$ to the function $|\nabla f|$, the last term becomes $\int_{\mathbb{R}^m} |\nabla f|^2\, d\gamma_t$, proving the desired inequality.

The proof of Theorem 10.3 requires a diversion into *heat kernel analysis*. Let us take a look at the right-hand-side of Inequality LSI. Using Gaussian integration by parts,

$$\int_{\mathbb{R}^m} |\nabla f|^2\,d\gamma_1 = (2\pi)^{-m/2}\sum_{j=1}^m \int_{\mathbb{R}^m} (\partial_j f(\mathbf{x}))^2 e^{-|\mathbf{x}|^2/2}\,d\mathbf{x} = -(2\pi)^{-m/2}\sum_{j=1}^m f(\mathbf{x})\partial_j[\partial_j f(\mathbf{x})e^{-|\mathbf{x}|^2/2}]\,d\mathbf{x}.$$

From the product rule we have

$$\partial_j[\partial_j f(\mathbf{x})e^{-|\mathbf{x}|^2/2}] = \partial_j^2 f(\mathbf{x})e^{-|\mathbf{x}|^2/2} - x_j\partial_j f(\mathbf{x})e^{-|\mathbf{x}|^2/2}.$$

The operator $\Delta = \sum_{j=1}^m \partial_j^2$ is the *Laplacian* on $\mathbb{R}^m$. The above equations say that

$$\int_{\mathbb{R}^m} |\nabla f|^2\,d\gamma_1 = -\int_{\mathbb{R}^m} f(\mathbf{x})(\Delta - \mathbf{x}\cdot\nabla)f(\mathbf{x})\,\gamma_1(d\mathbf{x}). \tag{10.2}$$

The operator $L = \Delta - \mathbf{x}\cdot\nabla$ is called the **Ornstein-Uhlenbeck operator**. It is a first-order perturbation of the Laplacian. One might hope that it possesses similar properties to the Laplacian; indeed, in many ways, it is better. To understand this statement, we consider the *heat equation for* L.

10.1. **Heat kernels.** The classical **heat equation** on $\mathbb{R}^m$ is the following PDE:

$$\partial_t u - \Delta u = 0, \quad t > 0.$$

A solution is a function $u\colon \mathbb{R}_+ \times \mathbb{R}^m \to \mathbb{R}$ which satisfies the equation (perhaps in a weak sense, a priori. Typically we are interested in an *initial value problem*: we want a solution $u(t,\mathbf{x}) = u_t(\mathbf{x})$ satisfying the heat equation, and with a given initial condition $\lim_{t\downarrow 0} u_t = f$ in an appropriate sense. On all of $\mathbb{R}^m$, no specialized tools are necessary to solve the heat equation: it is given by the **heat kernel**. That is, for any $L^p(\mathbb{R}^m)$ function f, define $H_t f = \gamma_t * f$. It is an easy exercise that $u(t,\mathbf{x}) = H_t f(\mathbf{x})$ satisfies the heat equation, and the initial condition $\lim_{t\downarrow 0} H_t f = f$ in $L^p(\mathbb{R}^m)$-sense (and even stronger senses if f is nice enough).

In fact, we can do much the same with the Ornstein-Uhlenbeck operator L. It is, in fact, more adapted to the Gaussian measure γ_1 than is the Laplacian.

Definition 10.6. *Let $f \in L^2(\mathbb{R}^m,\gamma_1)$. Say that a function $u\colon \mathbb{R}_+ \times \mathbb{R}^m \to \mathbb{R}$ is a solution to the* Ornstein-Uhlenbeck heat equation *with initial condition f if $u_t(\mathbf{x}) = u(t,\mathbf{x})$ is in $L^2(\mathbb{R}^m,\gamma_1)$ for each $t > 0$ and*

$$\begin{aligned} \partial_t u - Lu = 0, \quad t > 0 \\ \lim_{t\downarrow 0} \|u_t - f\|_{L^2(\gamma_1)} = 0. \end{aligned}$$

It is a remarkable theorem (essentially discovered by Mehler in the 19th century) that the OU-heat equation is also solved by a heat kernel. For $f \in L^2(\mathbb{R}^m,\gamma_1)$ (note that this allows f to grow relatively fast at ∞), define

$$P_t f(\mathbf{x}) = \int_{\mathbb{R}^m} f(e^{-t}\mathbf{x} + \sqrt{1-e^{-2t}}\mathbf{y})\,\gamma_1(d\mathbf{y}). \tag{10.3}$$

This *Mehler formula* has a beautiful geometric interpretation. For fixed t, choose $\theta \in [0,\pi/2]$ with $\cos\theta = e^{-t}$. Then $f(e^{-t}\mathbf{x} + \sqrt{1-e^{-2t}\mathbf{y}}) = f(\cos\theta\mathbf{x} + \sin\theta\mathbf{y})$. This is a function of two $\mathbb{R}^m$-variables, while f is a function of just one. So we introduce two operators between the one-variable and two-variable L^2-spaces:

$$\Phi\colon L^2(\gamma_1) \to L^2(\gamma_1 \times \gamma_1)\colon \quad (\Phi f)(\mathbf{x},\mathbf{y}) = f(\mathbf{x})$$

$$\Phi^*\colon L^2(\gamma_1 \times \gamma_1) \to L^2(\gamma_1)\colon \qquad (\Phi F)(\mathbf{x}) = \int F(\mathbf{x},\mathbf{y})\,\gamma_1(d\mathbf{y}).$$

Also define the block rotation $R_\theta\colon \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}^m \times \mathbb{R}^m$ by

$$R_\theta(\mathbf{x},\mathbf{y}) = (\cos\theta\mathbf{x} - \sin\theta\mathbf{y}, \sin\theta\mathbf{x} + \cos\theta\mathbf{y}).$$

Then R_θ also acts dually on function $F\colon \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}$ via

$$(R_\theta^* F)(\mathbf{x},\mathbf{y}) = F(R_\theta^{-1}(\mathbf{x},\mathbf{y})).$$

Putting these pieces together, we see that

$$P_t = \Phi^* R_\theta^* \,\Phi. \tag{10.4}$$

So, up to the maps Φ and Φ^*, P_t is just a rotation of coordinates by the angle $\theta = \cos^{-1}(e^{-t})$. This is useful because all three composite maps in P_t are very well-behaved on L^2.

Lemma 10.7. *Let $\theta \in [0,\pi/2]$, and define Φ, Φ^*, and R_θ^* as above.*

(a) *The map Φ is an L^2-isometry $\|\Phi f\|_{L^2(\gamma_1\times\gamma_1)} = \|f\|_{L^2(\gamma_1)}$.*

(b) *The map Φ^* is an L^2-contraction $\|\Phi^* F\|_{L^2(\gamma_1)} \le \|F\|_{L^2(\gamma_1\times\gamma_1)}$.*

(c) *Each of the maps R_θ^* is an isometry on $L^2(\gamma_1 \times \gamma_1)$. Moreover, the map*

$$[0,\pi/2] \to L^2(\gamma_1 \times \gamma_1)\colon \theta \mapsto R_\theta^*$$

is uniformly continuous.

Proof. (a) Just calculate

$$\|\Phi f\|^2_{L^2(\gamma_1\times\gamma_1)} = \int_{\mathbb{R}^{2m}} |(\Phi f)(\mathbf{x},\mathbf{y})|^2\,(\gamma_1\times\gamma_1)(d\mathbf{x}d\mathbf{y}) = \int_{\mathbb{R}^{2m}} |f(\mathbf{x})|^2\,\gamma_1(d\mathbf{x})\,\gamma_1(d\mathbf{y})$$

where Fubini's theorem has been used since the integrand is positive. The integrand does not depend on $\mathbf{y}$, and γ_1 is a probability measure, so integrating out the $\mathbf{y}$ yields $\|f\|^2_{L^2(\gamma_1)}$.

(b) Again we calculate

$$\|\Phi F\|^2_{L^2(\gamma_1)} = \int_{\mathbb{R}^m} |(\Phi F)(\mathbf{x})|^2\,\gamma_1(d\mathbf{x}) = \int_{\mathbb{R}^m} \left|\int_{\mathbb{R}^m} F(\mathbf{x},\mathbf{y})\,\gamma_1(d\mathbf{y})\right|^2 \gamma_1(d\mathbf{x}).$$

The function $x \mapsto |x|^2$ is convex, and γ_1 is a probability measure, so the inside integral satisfies, for (almost) every $\mathbf{x}$,

$$\left|\int_{\mathbb{R}^m} F(\mathbf{x},\mathbf{y})\,\gamma_1(d\mathbf{y})\right|^2 \le \int_{\mathbb{R}^m} |F(\mathbf{x},\mathbf{y})|^2\,\gamma_1(d\mathbf{y}).$$

Combining, and using Fubini's theorem (again justified by the positive integrand) yields the result.

(c) We have

$$\|R_\theta^* F\|^2_{L^2(\gamma_1\times\gamma_1)} = \int_{\mathbb{R}^{2m}} |F(R_\theta^{-1}(\mathbf{x},\mathbf{y}))|^2\,(\gamma_1\times\gamma_1)(d\mathbf{x}d\mathbf{y}).$$

Now we make the change of variables $(\mathbf{u},\mathbf{v}) = R_\theta^{-1}(\mathbf{x},\mathbf{y})$. This is a rotation, so has determinant 1; thus $d\mathbf{u}d\mathbf{v} = d\mathbf{x}d\mathbf{y}$. The density of the Gaussian measure transforms as

$$(\gamma_1\times\gamma_1)(d\mathbf{x}d\mathbf{y}) = (2\pi t)^{-m} e^{-(|\mathbf{x}|^2+|\mathbf{y}|^2)/2t}\,d\mathbf{x}d\mathbf{y}.$$

But the Euclidean length is invariant under rotations, so $|\mathbf{x}|^2 + |\mathbf{y}|^2 = |\mathbf{u}|^2 + |\mathbf{v}|^2$, and so we see that $\gamma_1 \times \gamma_1$ is invariant under the coordinate transformation. (Gaussian measures are invariant under rotations.) Thus

$$\int_{\mathbb{R}^{2m}} |F(R_\theta^{-1}(\mathbf{x},\mathbf{y}))|^2\,(\gamma_1\times\gamma_1)(d\mathbf{x}d\mathbf{y}) = \int_{\mathbb{R}^{2m}} |F(\mathbf{u},\mathbf{v})|^2\,(\gamma_1\times\gamma_1)(d\mathbf{u}d\mathbf{v}) = \|F\|^2_{L^2(\gamma_1\times\gamma_1)}.$$

proving that R_θ^* is an isometry.

Now, let $\theta,\phi\in[0,\pi/2]$. Fix $\epsilon>0$, and choose a continuous approximating function $G\in C_b(\mathbb{R})$ with $\|F-G\|_{L^2(\gamma_1\times\gamma_1)}<\frac{\epsilon}{3}$. The maps R_θ^* and R_ϕ^* are linear, and so we have

$$\begin{aligned}\|R_\theta^*F - R_\theta^*G\|_2 &= \|R_\theta^*(F-G)\|_2 = \|F-G\|_2 < \frac{\epsilon}{3}\\ \|R_\phi^*F - R_\phi^*G\|_2 &= \|R_\phi^*(F-G)\|_2 = \|F-G\|_2 < \frac{\epsilon}{3}.\end{aligned}$$

Thus

$$\begin{aligned}\|R_\theta^*F - R_\phi^*F\|_2 &\le \|R_\theta^*F - R_\theta^*G\|_2 + \|R_\theta^*G - R_\phi^*G\|_2 + \|R_\phi^*G - R_\phi^*F\|_2\\ &\le \frac{2}{3}\epsilon + \|R_\theta^*G - R_\phi^*G\|_2.\end{aligned}$$

Now,

$$\|R_\theta^*G - R_\phi^*G\|_2^2 = \int_{\mathbb{R}^{2m}} |G(R_\theta^{-1}(\mathbf{x},\mathbf{y})) - G(R_\phi^{-1}(\mathbf{x},\mathbf{y}))|^2\,(\gamma_1\times\gamma_1)(d\mathbf{x}d\mathbf{y}).$$

Since G is continuous, the integrand converges uniformly to 0 as $\phi\to\theta$, and is bounded by $2\|G\|_\infty$; so by the dominated convergence theorem, for $|\phi-\theta|$ sufficiently small we can make $\|R_\theta^*G - R_\phi^*G\|_2 < \frac{\epsilon}{3}$. This shows that $\theta\mapsto R_\theta^*$ is uniformly continuous. □

Corollary 10.8. *The operator P_t of Equations 10.3 and 10.4 is a contraction on $L^2(\gamma_1)$ for each $t\ge 0$. Moreover, for any $f\in L^2(\gamma_1)$,*

(1) *$P_tf\to f$ in $L^2(\gamma_1)$ as $t\downarrow 0$, and $P_tf\to\int f\,d\gamma_1$ in $L^2(\gamma_1)$ as $t\uparrow\infty$,*

(2) *If $f\ge 0$, then $\int P_tf\,d\gamma_1 = \int f\,d\gamma_1$ for all $t\ge 0$.*

Proof. From Equation 10.4, $P_t = \Phi^* R_\theta^*\Phi$. Lemma 10.7 shows that this is a composition of isometries and a contraction, hence is a contraction. By part (c) of that lemma, for any $f\in L^2(\gamma_1)$, $R_\phi^*\Phi f\to R_\theta^*\Phi f$ in $L^2(\gamma_1\times\gamma_1)$ as $\phi\to\theta$. Since Φ^* is a contraction, we also have $P_tf = \Phi^*R_\phi^*\Phi f\to\Phi^*R_\theta^*\Phi f$ in $L^2(\gamma_1)$ as $\phi\to\theta$, where $\theta=\cos^{-1}(e^{-t})$ is a continuous function of t. The two limits in (1) are just the special cases $\theta=0$ (corresponding to $t\downarrow 0$) and $\theta=\pi/2$ (corresponding to $t\uparrow\infty$. In the former case, since $R_0 = Id$, we have $P_tf\to\Phi^*\Phi f = f$ as the reader can quickly verify. In the latter case, $R_{\pi/2}^*F(\mathbf{x},\mathbf{y}) = F(\mathbf{y},-\mathbf{x})$, and so

$$(\Phi^*R_{\pi/2}\,\Phi f)(\mathbf{x}) = \int f(\mathbf{y})\,\gamma_1(d\mathbf{y})$$

is a constant function, the γ_1-expectation of f.

For part (2), we simply integrate and use Fubini's theorem (on the positive integrant $P_t f$). With $\cos\theta = e^{-t}$ as usual,

$$\begin{aligned}\int_{\mathbb{R}^m} P_t f \, d\gamma_1 &= \int_{\mathbb{R}^m} \left(\int_{\mathbb{R}^m} f(e^{-t}\mathbf{x} + \sqrt{1-e^{-2t}}\mathbf{y}) \, \gamma_1(d\mathbf{y}) \right) \gamma_1(d\mathbf{x}) \\ &= \int_{\mathbb{R}^{2m}} f(\cos\theta \mathbf{x} + \sin\theta \mathbf{y}) \, (\gamma_1 \times \gamma_1)(d\mathbf{x} d\mathbf{y}).\end{aligned}$$

Changing variables and using the rotational invariance of the Gaussian measure $\gamma_1 \times \gamma_1$ as in the proof of Lemma 10.7(c), this is the same as $\int_{\mathbb{R}^{2m}} f \, d(\gamma_1 \times \gamma_1)$, which (by integrating out the second $\mathbb{R}^m$-variable with respect to which f is constant) is equal to $\int_{\mathbb{R}^m} f \, d\gamma_1$ as claimed. □

This finally brings us to the purpose of P_t: it is the OU-heat kernel.

Proposition 10.9. *Let $f \in L^2(\mathbb{R}^m, \gamma_1)$, and for $t > 0$ set $u_t(\mathbf{x}) = u(t, \mathbf{x}) = P_t f(\mathbf{x})$. Then u solves the Ornstein-Uhlenbeck heat equation with initial value f.*

Proof. By Corollary 10.8, $\|u_t\| = \|P_t f\|_2 \le \|f\|_2$ so u_t is in L^2; also $u_t = P_t f$ converges to f in L^2 as $t \downarrow 0$. Hence, we are left to verify that u satisfies the OU-heat equation, $\partial_t u = Lu$. For the duration of the proof we will assume that $f \in C_c^2(\mathbb{R}^m)$; this assumption can be removed by a straightforward approximation afterward. Thus, all the derivatives of f are bounded. We leave it to the reader to provide the (easy dominated convergence theorem) justification that we can differentiate $P_t f(\mathbf{x})$ under the integral in both t and $\mathbf{x}$. As such, we compute

$$\partial_t f(e^{-t}\mathbf{x} + \sqrt{1-e^{-2t}}\mathbf{y}) = \left(-e^{-t}\mathbf{x} + e^{-2t}(1-e^{-2t})^{-1/2}\mathbf{y}\right) \cdot \nabla f(e^{-t}\mathbf{x} + \sqrt{1-e^{-2t}}\mathbf{y}).$$

Thus,

$$\begin{aligned}\partial_t u(t,\mathbf{x}) = -e^{-t} &\int_{\mathbb{R}^m} \mathbf{x} \cdot \nabla f(e^{-t}\mathbf{x} + \sqrt{1-e^{-2t}}\mathbf{y})\, \gamma_1(d\mathbf{y}) \\ &+ e^{-2t}(1-e^{-2t})^{-1/2} \int_{\mathbb{R}^m} \mathbf{y} \cdot \nabla f(e^{-t}\mathbf{x} + \sqrt{1-e^{-2t}}\mathbf{y})\, \gamma_1(d\mathbf{y}).\end{aligned}$$

For the space derivatives, we have

$$\frac{\partial}{\partial x_j} f(e^{-t}\mathbf{x} + \sqrt{1-e^{-2t}}\mathbf{y}) = e^{-t}(\partial_j f)(e^{-t}\mathbf{x} + \sqrt{1-e^{-2t}}\mathbf{y}).$$

Mutiplying by x_j, adding up over $1 \le j \le m$, and integrating with respect to γ_1 shows that the first term in $\partial_t u(t,\mathbf{x})$ is $-\mathbf{x} \cdot \nabla u(t,\mathbf{x})$. That is:

$$\partial_t u(t,\mathbf{x}) = -\mathbf{x} \cdot \nabla u(t,\mathbf{x}) + e^{-2t}(1-e^{-2t})^{-1/2} \int_{\mathbb{R}^m} \mathbf{y} \cdot \nabla f(e^{-t}\mathbf{x} + \sqrt{1-e^{-2t}}\mathbf{y})\, \gamma_1(d\mathbf{y}). \quad (10.5)$$

For the second term, we now do Gaussian integration by parts. For fixed $\mathbf{x}$, let $\mathbf{F}(\mathbf{y}) = \nabla f(e^{-t}\mathbf{x} + \sqrt{1-e^{-2t}}\mathbf{y})$. Then the above integral is equal to

$$\int_{\mathbb{R}^m} \mathbf{y} \cdot \mathbf{F}(\mathbf{y})\, \gamma_1(d\mathbf{y}) = \int_{\mathbb{R}^m} \nabla \cdot \mathbf{F}\, d\gamma_1$$

where the equality follows from Lemma 9.1. Well,

$$\nabla \cdot \mathbf{F}(\mathbf{y}) = \sum_{j=1}^m \frac{\partial}{\partial y_j}(\partial_j f)(e^{-t}\mathbf{x} + \sqrt{1-e^{-2t}}\mathbf{y}) = \sqrt{1-e^{-2t}} \sum_{j=1}^n (\partial_j^2 f)(e^{-t}\mathbf{x} + \sqrt{1-e^{-2t}}\mathbf{y}).$$

Combining with Equation 10.5 gives

$$\partial_t u(t,\mathbf{x}) = -\mathbf{x} \cdot \nabla u(t,\mathbf{x}) + e^{-2t} \int_{\mathbb{R}^m} \Delta f(e^{-t}\mathbf{x} + \sqrt{1-e^{-2t}}\mathbf{y})\, \gamma_1(d\mathbf{y}).$$

Finally, note that

$$\Delta u(t,\mathbf{x}) = \sum_{j=1}^n \frac{\partial^2}{\partial x_j^2} \int_{\mathbb{R}^m} f(e^{-t}\mathbf{x} + \sqrt{1-e^{-2t}}\mathbf{y})\, \gamma_1(d\mathbf{y}) = e^{-2t} \int_{\mathbb{R}^m} \Delta f(e^{-t}\mathbf{x} + \sqrt{1-e^{-2t}}\mathbf{y})\, \gamma_1(d\mathbf{y}),$$

showing that $\partial_t u(t,\mathbf{x}) = -\mathbf{x} \cdot \nabla u(t,\mathbf{x}) + \Delta u(t,\mathbf{x}) = Lu(t,\mathbf{x})$. Thus, $u = P_t f$ satisfies the OU-heat equation, completing the proof. □

Remark 10.10. An alternate approach, avoiding approximation of f by smoother functions, is to rewrite the operator P_tf by doing a change of variables. One can rewrite it in terms of a kernel agains *Lebesgue* measure: the reader should verify that

$$P_tf(\mathbf{x}) = \int_{\mathbb{R}^m} M_t(\mathbf{x},\mathbf{y})f(\mathbf{y})\,d\mathbf{y} \tag{10.6}$$

where

$$M_t(\mathbf{x},\mathbf{y}) = \pi^{-m/2}(1-e^{-2t})^{-m/2}\exp\left\{-\frac{|\mathbf{y}-e^{-t}\mathbf{x}|^2}{1-e^{-2t}}\right\}. \tag{10.7}$$

In this form, one simply has to verify that, for fixed $\mathbf{y}$, the function $\mathbf{x}\to M_t(\mathbf{x},\mathbf{y})$ satisfies the OU-heat equation. The functions $\partial_t M_t(\mathbf{x},\mathbf{y})$ and $\Delta_{\mathbf{x}}M_t(\mathbf{x},\mathbf{y})$ still have Gaussian tail decay, and a straightforward dominated converge argument shows one can differentiate under the integral in Equation 10.6, proving the result for generic $f\in L^2(\gamma_1)$.

10.2. **Proof of the Gaussian log-Sobolev inequality, Theorem 10.3.** Here we will show that any sufficiently smooth non-negative function $f\in L^2(\gamma_1)$ satisfies the log-Sobolev inequality in the L^1 form of Equation LSI':

$$\mathrm{Ent}_{\gamma_1}(g) \le \frac{1}{2}\int \frac{|\nabla g|}{g}\,d\gamma_1. \tag{10.8}$$

To begin, we make a cutoff approximation. Fix $\epsilon>0$, and assume $\epsilon\le f\le\frac{1}{\epsilon}$. For example, let h_ϵ be a smooth, positive function that is equal to 1 on the set where $2\epsilon\le g\le\frac{1}{2\epsilon}$, equal to 0 where $g<\epsilon$ or $g>\frac{1}{\epsilon}$, and has bounded partial derivatives; then set $f=h_\epsilon g+\epsilon$. If we can prove Inequality 10.8 with f in place of g, then we can use standard limit theorems to send $\epsilon\downarrow 0$ on both sides of the inequality to conclude it holds in general.

Let $\varphi(x)=x\log x$. By definition

$$\begin{aligned}\mathrm{Ent}_{\gamma_1}(f) = \int f\log f\,d\gamma_1 - \int f\,d\gamma_1\cdot\log\int f\,d\gamma_1 &= \int\varphi(f)\,d\gamma_1 - \varphi\left(\int f\,d\gamma_1\right)\\ &= \int\left[\varphi(f)-\varphi\left(\int f\,d\gamma_1\right)\right]\,d\gamma_1.\end{aligned}$$

Now, from Corollary 10.8, we have $\lim_{t\downarrow 0}P_tf=f$ while $\lim_{t\to\infty}P_tf=\int f\,d\gamma_1$; hence, using the dominated convergence theorem (justified by the assumption $\epsilon\le f\le\frac{1}{\epsilon}$, which implies that $\epsilon\le P_tf\le\frac{1}{\epsilon}$ from Equation 10.3, and so $\log P_tf$ is bounded)

$$\mathrm{Ent}_{\gamma_1}(f) = \lim_{t\downarrow 0}\int\varphi(P_tf)\,d\gamma_1 - \lim_{t\to\infty}\int\varphi(P_tf)\,d\gamma_1.$$

The clever trick here is to use the Fundamental Theorem of Calculus (which applies since the function $t\mapsto\int\varphi(P_tf)\,d\gamma_1$ is continuous and bounded) to write this as

$$\mathrm{Ent}_{\gamma_1}(f) = -\int_0^\infty\frac{d}{dt}\left(\int\varphi(P_tf)\,d\gamma_1\right)\,dt = -\int_0^\infty\frac{d}{dt}\left(\int P_tf\log P_tf\,d\gamma_1\right)\,dt. \tag{10.9}$$

Now, for fixed $t>0$, we differentiate under the inside integral (again justified by dominated convergence, using the boundedness assumption on f).

$$\frac{d}{dt}\left(\int P_tf\log P_tf\,d\gamma_1\right) = \int\partial_t\left(P_tf\log P_tf\right)\,d\gamma_1.$$

Let $u_t = P_t f$. By Proposition 10.9, u_t solves the OU-heat equation. So, we have

$$\partial_t(P_t f \log P_t f) = \partial_t(u_t \log u_t) = (\partial_t u_t)\log u_t + \partial_t u_t = (Lu_t)\log u_t + Lu_t.$$

Integrating,

$$\int \partial_t(u_t \log u_t)\, d\gamma_1 = \int (Lu_t)\log u_t\, d\gamma_1 + \int Lu_t\, d\gamma_1. \tag{10.10}$$

Now, remember where L came from in the first place: Gaussian integration by parts. Polarizing Equation 10.2, we have for any g, h sufficiently smooth

$$\int \nabla g \cdot \nabla h\, d\gamma_1 = -\int g\, Lh\, d\gamma_1. \tag{10.11}$$

In particular, this means

$$\int Lu_t\, d\gamma_1 = \int 1\, Lu_t\, d\gamma_1 = -\int \nabla 1 \cdot \nabla h\, d\gamma_1 = 0,$$

while

$$\int (Lu_t)\log u_t\, d\gamma_1 = -\int \nabla u_t \cdot \nabla(\log u_t)\, d\gamma_1 = -\int \frac{|\nabla u_t|^2}{u_t}\, d\gamma_1.$$

Combining with Equation 10.10 gives

$$\frac{d}{dt}\left(\int P_t f \log P_t f\, d\gamma_1\right) = -\int \frac{|\nabla P_t f|^2}{P_t f}\, d\gamma_1. \tag{10.12}$$

We now utilize the Mehler formula for the action of P_t to calculate $\nabla P_t f$:

$$\begin{aligned}\partial_j P_t f(\mathbf{x}) = \frac{\partial}{\partial x_j}\int f(e^{-t}\mathbf{x} + \sqrt{1-e^{-2t}}\mathbf{y})\,\gamma_1(d\mathbf{y}) &= e^{-t}\int \partial_j f(e^{-t}\mathbf{x} + \sqrt{1-e^{-2t}}\mathbf{y})\,\gamma_1(d\mathbf{y}) \\ &= e^{-t}P_t(\partial_j f)(\mathbf{x}).\end{aligned}$$

In other words, $\nabla P_t f = e^{-t}P_t(\nabla f)$ where P_t is defined on vector-valued functions componentwise $P_t(f_1, \dots, f_m) = (P_t f_1, \dots, P_t f_m)$. In particular, if $\mathbf{v}$ is any vector in $\mathbb{R}^m$, we have by the linearity of P_t

$$\nabla P_t f \cdot \mathbf{v} = e^{-t}P_t(\nabla f \cdot \mathbf{v}). \tag{10.13}$$

Now, for any functions $h \le \tilde{h}$, it is easy to verify that $P_t h \le P_t \tilde{h}$. It follows also that $|P_t h| \le P_t|h|$. Using these facts and the Cauchy-Schwarz inequality in Equation 10.13 yields

$$|\nabla P_t f \cdot \mathbf{v}| \le e^{-t}P_t(|\nabla f \cdot \mathbf{v}|) \le e^{-t}P_t(|\nabla f|\,|\mathbf{v}|) = |\mathbf{v}|e^{-t}P_t(|\nabla f|).$$

Taking $\mathbf{v} = \nabla P_t f$ shows that

$$|\nabla P_t f|^2 \le |\nabla P_t f|\, e^{-t}P_t(|\nabla f|).$$

If $|\nabla P_t f| = 0$ then it is clearly $\le e^{-t}P_t(|\nabla f|)$. Otherwise, we can divide through by it to find that

$$|\nabla P_t f| \le e^{-t}P_t(|\nabla f|). \tag{10.14}$$

We further estimate this pointwise bound as follows. For fixed t and $\mathbf{x}$, set $h(\mathbf{y}) = |\nabla f(e^{-t}\mathbf{x} + \sqrt{1-e^{-2t}}\mathbf{y})|$. Similarly, set $r(\mathbf{y})^2 = f(e^{-t}\mathbf{x} + \sqrt{1-e^{-2t}}\mathbf{y})$. Then we have

$$\begin{aligned}P_t(|\nabla f|)^2 = \left(\int h\, d\gamma_1\right)^2 = \left(\int \frac{h}{r}\cdot r\, d\gamma_1\right)^2 &\le \int \left(\frac{h}{r}\right)^2 d\gamma_1 \cdot \int r^2\, d\gamma_1 \\ &= P_t\left(\frac{|\nabla f|^2}{f}\right) P_t f \end{aligned} \tag{10.15}$$

where the inequality is the Cauchy-Schwarz inequality. Combining Equations 10.14 and 10.15 yields

$$|\nabla P_t f|^2 \le e^{-2t} P_t\left(\frac{|\nabla f|^2}{f}\right) P_t f$$

and combining this with Equation 10.12 gives

$$-\frac{d}{dt}\left(\int P_t f \log P_t f \, d\gamma_1\right) = \int \frac{|\nabla P_t f|^2}{P_t f}\, d\gamma_1 \le e^{-2t} \int P_t\left(\frac{|\nabla f|^2}{f}\right) d\gamma_1.$$

The function $|\nabla f|^2/f$ is ≥ 0, and so by Corollary 10.8(2), it follows that

$$\int P_t\left(\frac{|\nabla f|^2}{f}\right) d\gamma_1 = \int \frac{|\nabla f|^2}{f}\, d\gamma_1.$$

Integrating over $t \in (0,\infty)$, Equation 10.9 implies that

$$\mathrm{Ent}_{\gamma_1}(f) = -\frac{d}{dt}\left(\int P_t f \log P_t f \, d\gamma_1\right) \le \int_0^\infty e^{-2t}\left(\int \frac{|\nabla f|^2}{f}\, d\gamma_1\right) dt. \tag{10.16}$$

Since $\int_0^\infty e^{-2t}\, dt = \frac{1}{2}$, this proves the result.

Remark 10.11. This argument extends beyond Gaussian measures, to some degree. Suppose that μ has a density of the form $Z^{-1}e^{-U}$ where $U > 0$ is a smooth "potential" and $Z = \int e^{-U(\mathbf{x})}\, d\mathbf{x} < \infty$. One may integrate by parts to find the associated heat operator $L_U = \Delta - \nabla U \cdot \nabla$. It is not always possible to write down an explicit heat kernel P_t^U for the L_U-heat equation, but most of the preceding argument follows through. Without an explicit formula, one cannot prove an equality like Equation 10.13, but it is still possible to prove the subsequent inequality in a form $|\nabla P_t^U f| \le e^{-ct} P_t^U(|\nabla f|)$ with an appropriate constant – *provided that* $\mathbf{x} \mapsto U(\mathbf{x}) - |\mathbf{x}|^2/2c$ *is convex*. Following the rest of the proof, one sees that such *log-concave* measures do satisfy the log-Sobolev inequality with constant c. (This argument is due to Ledoux, with generalizations more appropriate for manifolds other than Euclidean space due to Bakry and Emery.) This is basically the largest class of measures known to satisfy LSIs, with many caveats: for example, an important theorem of Stroock and Holley is that if μ satisfies a log-Sobolev inequality and h is a μ-probability density that is bounded above and below, then $h\,\mu$ also satisfies a log-Sobolev inequality (with a constant determined by $\sup h - \inf h$).

11. THE HERBST CONCENTRATION INEQUALITY

The main point of introducing the log-Sobolev inequality is the following concentration of measure which it entails.

Theorem 11.1 (Herbst). *Let μ be a probability measure on $\mathbb{R}^m$ satisfying the logarithmic Sobolev inequality LSI with constant c. Let $F\colon \mathbb{R}^m \to \mathbb{R}$ be a Lipschitz function. Then for all $\lambda \in \mathbb{R}$,*

$$\int e^{\lambda(F-\mathbb{E}_\mu(F))}\,d\mu \le e^{c\lambda^2\|F\|^2_{\mathrm{Lip}}/2}. \tag{11.1}$$

It follows that, for all ¿0,

$$\mu\{|F-\mathbb{E}_\mu(F)| \ge \delta\} \le 2e^{-\delta^2/2c\|F\|^2_{\mathrm{Lip}}}. \tag{11.2}$$

Remark 11.2. The key feature of the concentration of measure for the random variable F above is that the behavior is *independent of dimension* m. This is the benefit of the log-Sobolev inequality: it is dimension-independent.

Proof. The implication from Equation 11.1 to 11.2 is a very standard exponential moment bound. In general, by Markov's inequality, for any random variable F possessing finite exponential moments,

$$\mathbb{P}(|F-\mathbb{E}(F)|>\delta) = \mathbb{P}\left(e^{\lambda|F-\mathbb{E}(F)|} > e^{\lambda\delta}\right) \le \frac{1}{e^{\lambda\delta}}\mathbb{E}\left(e^{\lambda|F-\mathbb{E}(F)|}\right).$$

We can estimate this expectation by

$$\begin{aligned}\mathbb{E}(e^{\lambda|F-\mathbb{E}(F)|}) &= \mathbb{E}\left(e^{\lambda(F-\mathbb{E}(F))}\mathbb{1}_{\{F\ge\mathbb{E}(F)\}}\right) + \mathbb{E}\left(e^{-\lambda(F-\mathbb{E}(F))}\mathbb{1}_{\{F\le\mathbb{E}(F)\}}\right)\\ &\le \mathbb{E}\left(e^{\lambda(F-\mathbb{E}(F))}\right) + \mathbb{E}\left(e^{-\lambda(F-\mathbb{E}(F))}\right).\end{aligned}$$

In the case $\mathbb{P} = \mu$ at hand, where Inequality 11.1 holds for Lipschitz F (for all $\lambda \in \mathbb{R}$), each of these terms is bounded above by $e^{c\lambda^2\|F\|^2_{\mathrm{Lip}}/2}$, and so we have

$$\mu\{|F-\mathbb{E}_\mu(F)| \ge \delta\} \le 2e^{-\lambda\delta}e^{c\lambda^2\|F\|^2_{\mathrm{Lip}}/2}$$

for any $\lambda \in \mathbb{R}$. By elementary calculus, the function $\lambda \mapsto c\lambda^2\|F\|^2_{\mathrm{Lip}}/2 - \lambda\delta$ achieves its minimal value at $\lambda =$ ¡$c\|F\|^2_{\mathrm{Lip}}$, and this value is $\delta^2/2c\|F\|^2_{\mathrm{Lip}}$, thus proving that Inequality 11.2 follows from Inequality 11.1.

To prove Inequality 11.1, we will first approximate F by a C^∞ function G whose partial derivatives are bounded. (We will show how to do this at the end of the proof.) We will, in fact, prove the inequality

$$\int e^{\lambda(G-\mathbb{E}_\mu(G))}\,d\mu \le e^{c\lambda^2\||\nabla G|^2\|_\infty/2}, \qquad \lambda \in \mathbb{R} \tag{11.3}$$

for all such G. First note that, by taking $\tilde{G} = G - \mathbb{E}_\mu(G)$, $\nabla\tilde{G} = \nabla G$; so it suffices to prove Inequality 11.3 under the assumption that $\mathbb{E}_\mu(G) = 0$. Now, for fixed $\lambda \in \mathbb{R}$, set

$$f_\lambda = e^{\lambda G/2 - c\lambda^2\||\nabla G|^2\|_\infty/4} \quad \text{and so} \quad \nabla f_\lambda = f_\lambda \cdot \frac{\lambda}{2}\nabla G.$$

Then the μ-energy of f_λ is

$$\int |\nabla f_\lambda|^2\,d\mu = \frac{\lambda^2}{4}\int |\nabla G|^2 f_\lambda^2\,d\mu \le \frac{\lambda^2}{4}\||\nabla G|^2\|_\infty \int f_\lambda^2\,d\mu. \tag{11.4}$$

Now, set $\Lambda(\lambda) = \int f_\lambda^2\, d\mu$. Note that

$$\frac{\partial}{\partial\lambda} f_\lambda^2 = f_\lambda^2 \cdot (G - c\lambda \| |\nabla G|^2 \|_\infty),$$

which is bounded uniformly for λ in any compact set (since G is bounded); hence, by the dominated convergence theorem, Λ is differentiable and

$$\Lambda'(\lambda) = \int \frac{\partial}{\partial\lambda} f_\lambda^2\, d\mu = \int f_\lambda^2 \cdot (G - c\lambda \| |\nabla G|^2 \|_\infty)\, d\mu. \tag{11.5}$$

Thus, we have

$$\begin{aligned} \lambda\Lambda'(\lambda) &= \int f_\lambda^2 \cdot (\lambda G - c\lambda^2 \| |\nabla G|^2 \|_\infty)\, d\mu \\ &= \int f_\lambda^2 \cdot (\lambda G - \frac{c}{2}\lambda^2 \| |\nabla G|^2 \|_\infty)\, d\mu - \frac{c}{2}\lambda^2 \| |\nabla G|^2 \|_\infty \int f_\lambda^2\, d\mu \\ &= \int f_\lambda^2 \log f_\lambda^2\, d\mu - \frac{c}{2}\lambda^2 \| |\nabla G|^2 \|_\infty \Lambda(\lambda). \end{aligned} \tag{11.6}$$

Note also that

$$\mathrm{Ent}_\mu(f_\lambda^2) = \int f_\lambda^2 \log f_\lambda^2\, d\mu - \int f_\lambda^2\, d\mu \cdot \log \int f_\lambda^2\, d\mu = \int f_\lambda^2 \log f_\lambda^2\, d\mu - \Lambda(\lambda)\log\Lambda(\lambda). \tag{11.7}$$

Combining Equations 11.6 and 11.7 gives

$$\mathrm{Ent}_\mu(f_\lambda^2) = \lambda\Lambda'(\lambda) - \Lambda(\lambda)\log\Lambda(\lambda) + \frac{c}{2}\lambda^2 \| |\nabla G|^2 \|_\infty \Lambda(\lambda). \tag{11.8}$$

Now using the log-Sobolev inequality LSI applied to f_λ, Equations 11.4 and 11.8 combine to give

$$\lambda\Lambda'(\lambda) - \Lambda(\lambda)\log\Lambda(\lambda) + \frac{c}{2}\lambda^2 \| |\nabla G|^2 \|_\infty \Lambda(\lambda) = \mathrm{Ent}_\mu(f_\lambda^2) \le 2c \int |\nabla f_\lambda|^2\, d\mu \le \frac{c}{2}\lambda^2 \| |\nabla G|^2 \|_\infty \Lambda(\lambda).$$

That is, miraculously, we simply have

$$\lambda\Lambda'(\lambda) - \Lambda(\lambda)\log\Lambda(\lambda) \le 0, \quad \lambda \in \mathbb{R}. \tag{11.9}$$

Notice that $f_\lambda^2 > 0$ and so $\Lambda(\lambda) > 0$ for all $\lambda \in \mathbb{R}$. So we divide through by $\Lambda(\lambda)$, giving

$$\lambda \frac{\Lambda'(\lambda)}{\Lambda(\lambda)} \le \log\Lambda(\lambda). \tag{11.10}$$

Now, define

$$H(\lambda) = \begin{cases} \frac{1}{\lambda}\log\Lambda(\lambda), & \lambda \ne 0 \\ 0, & \lambda = 0 \end{cases}.$$

Since Λ is strictly positive and differentiable, H is differentiable at all points except possibly $\lambda = 0$. At this point, we at least have

$$\lim_{\lambda\to 0} H(\lambda) = \lim_{\lambda\to 0} \frac{\log\Lambda(\lambda)}{\lambda} = \frac{d}{d\lambda}\log\Lambda(\lambda)|_{\lambda=0} = \frac{\Lambda'(0)}{\Lambda(0)}$$

where the second equality follows from the fact that $f_0 \equiv 1$ so $\Lambda(0) = \int f_0^2\, d\mu = 1$, and so $\log\Lambda(0) = 0$. Similarly, from Equation 11.5, we have $\Lambda'(0) = \int f_0^2 \cdot G\, d\mu = \int G\, d\mu = \mathbb{E}_\mu(G) = 0$ by assumption. Thus, H is continuous at 0. For $\lambda \ne 0$, we have

$$H'(\lambda) = \frac{d}{d\lambda}\frac{\log\Lambda(\lambda)}{\lambda} = -\frac{1}{\lambda^2}\log\Lambda(\lambda) + \frac{1}{\lambda}\frac{\Lambda'(\lambda)}{\Lambda(\lambda)} = \frac{1}{\lambda^2}\left[\lambda\frac{\Lambda'(\lambda)}{\Lambda(\lambda)} - \log\Lambda(\lambda)\right] \le 0$$

where the inequality follows from 11.10. So, H is continuous, differentiable except (possibly) at 0, and has non-positive derivatve at all other points; it follows that H is non-increasing on $\mathbb{R}$. That is, for $\lambda > 0$, $0 = H(0) \geq H(\lambda) = \log\Lambda(\lambda)/\lambda$ which implies that $0 \geq \log\Lambda(\lambda)$, so $\Lambda(\lambda) \leq 1$. Similarly, for $\lambda < 0$, $0 = H(0) \leq H(\lambda) = \log\Lambda(\lambda)/\lambda$ which implies that $\log\Lambda(\lambda) \leq 0$ and so $\Lambda(\lambda) \leq 1$ once more. Ergo, for all $\lambda \in \mathbb{R}$, we have

$$1 \geq \Lambda(\lambda) = \int f_\lambda^2\, d\mu = \int e^{\lambda G - c\lambda^2 \||\nabla G|^2\|_\infty/2}\, d\mu = e^{-c\lambda^2\||\nabla G|^2\|_\infty/2}\int e^{\lambda G}\, d\mu$$

and this proves Inequality 11.3 (in the case $\mathbb{E}_\mu(G) = 0$), as desired.

To complete the proof, we must show that Inequality 11.3 implies Inequality 11.1. Let F be Lipschitz. Let $\psi \in C_c^\infty(\mathbb{R}^m)$ be a bump function: $\psi \geq 0$ with $\operatorname{supp}\psi \subseteq B_1$ (the unit ball), and $\int \psi(\mathbf{x})\, d\mathbf{x} = 1$. For each $\epsilon > 0$, set $\psi_\epsilon(\mathbf{x}) = \epsilon^{-m}\psi(\mathbf{x}/\epsilon)$, so that $\operatorname{supp}\psi_\epsilon \subseteq B_\epsilon$ but we still have $\int \psi_\epsilon(\mathbf{x})\, d\mathbf{x} = 1$. Now, define $F_\epsilon = F * \psi_\epsilon$. Then F_ϵ is C^∞. For any $\mathbf{x}$

$$F_\epsilon(\mathbf{x}) - F(\mathbf{x}) = \int F(\mathbf{x}-\mathbf{y})\,\psi_\epsilon(\mathbf{y})\, d\mathbf{y} - F(\mathbf{x}) = \int [F(\mathbf{x}-\mathbf{y}) - F(\mathbf{x})]\,\psi_\epsilon(\mathbf{y})\, d\mathbf{y}$$

where the second equality uses the fact that ψ_ϵ is a probability density. Now, since F is Lipschitz, we have $|F(\mathbf{x}-\mathbf{y}) - F(\mathbf{x})| \leq \|F\|_{\mathrm{Lip}}|\mathbf{y}|$ uniformly in $\mathbf{x}$. Therefore

$$|F_\epsilon(\mathbf{x}) - F(\mathbf{x})| \leq \int |F(\mathbf{x}-\mathbf{y}) - F(\mathbf{x})|\,\psi_\epsilon(\mathbf{y})\, d\mathbf{y} \leq \|F\|_{\mathrm{Lip}} \int |\mathbf{y}|\,\psi_\epsilon(\mathbf{y})\, d\mathbf{y}.$$

Since $\operatorname{supp}\psi_\epsilon \subseteq B_\epsilon$, this last integral is bounded above by

$$\int_{B_\epsilon} |\mathbf{y}|\psi_\epsilon(\mathbf{y})\, d\mathbf{y} \leq \int_{B_\epsilon} \epsilon\psi_\epsilon(\mathbf{y})\, d\mathbf{y} = \epsilon.$$

This shows that $F_\epsilon \to F$ *uniformly*. Moreover, note that for any $\mathbf{x}, \mathbf{x}' \in \mathbb{R}^m$,

$$\begin{aligned}|F_\epsilon(\mathbf{x}) - F_\epsilon(\mathbf{x}')| &\leq \int |F(\mathbf{x}-\mathbf{y}) - F(\mathbf{x}'-\mathbf{y})|\,\psi_\epsilon(\mathbf{y})\, d\mathbf{y} \\ &\leq \int \|F\|_{\mathrm{Lip}}|(\mathbf{x}-\mathbf{y}) - (\mathbf{x}'-\mathbf{y})|\,\psi_\epsilon(\mathbf{y})\, d\mathbf{y} \\ &= \|F\|_{\mathrm{Lip}}|\mathbf{x}-\mathbf{x}'| \qquad (11.11)\end{aligned}$$

which shows that $\|F_\epsilon\|_{\mathrm{Lip}} \leq \|F\|_{\mathrm{Lip}}$.

Now, by a simple application of the mean value theorem, $\|F_\epsilon\|_{\mathrm{Lip}} \leq \||\nabla F_\epsilon|\|_\infty$, but this inequality runs the wrong direction. Instead, we use the following linearization trick: for any pair of vectors $\mathbf{u}, \mathbf{v}$, since $0 \leq |\mathbf{v}-\mathbf{u}|^2 = |\mathbf{v}|^2 - 2\mathbf{v}\cdot\mathbf{u} + |\mathbf{u}|^2$ with equality if and only if $\mathbf{u} = \mathbf{v}$, the quantity $|\mathbf{v}|^2$ can be defined variationally as

$$|\mathbf{v}|^2 = \sup_{\mathbf{u}\in\mathbb{R}^m} \left[2\mathbf{v}\cdot\mathbf{u} - |\mathbf{u}|^2\right].$$

Taking $\mathbf{v} = \nabla F_\epsilon(\mathbf{x})$ for a fixed $\mathbf{x}$, this means that

$$|\nabla F_\epsilon(\mathbf{x})|^2 = \sup_{\mathbf{u}\in\mathbb{R}^m} \left[2\nabla F_\epsilon(\mathbf{x})\cdot\mathbf{u} - |\mathbf{u}|^2\right].$$

Now, the dot-product here is the directional derivative (since F_ϵ is C^∞, ergo differentiable):

$$\nabla F_\epsilon(\mathbf{x})\cdot\mathbf{u} = D_{\mathbf{u}}F_\epsilon(\mathbf{x}) = \lim_{t\to 0}\frac{F_\epsilon(\mathbf{x}+t\mathbf{u}) - F_\epsilon(\mathbf{x})}{t} \leq \sup_{t>0}\left|\frac{F_\epsilon(\mathbf{x}+t\mathbf{u}) - F_\epsilon(\mathbf{x})}{t}\right| \leq \|F_\epsilon\|_{\mathrm{Lip}}|\mathbf{u}|.$$

Hence, since $\|F_\epsilon\|_{\mathrm{Lip}} \le \|F\|_{\mathrm{Lip}}$ by Equation 11.11,

$$|\nabla F_\epsilon(\mathbf{x})|^2 \le \sup_{\mathbf{u}\in\mathbb{R}^m} \left[2\|F\|_{\mathrm{Lip}}|\mathbf{u}| - |\mathbf{u}|^2\right] = \|F\|_{\mathrm{Lip}}^2$$

holds for all $\mathbf{x}$. This shows that $\||\nabla F_\epsilon|^2\|_\infty \le \|F\|_{\mathrm{Lip}}^2$. In particular, this shows that F_ϵ has bounded partial derivatives.

Hence, applying Inequality 11.3 to $G = F_\epsilon$, we have

$$\int e^{\lambda(F_\epsilon - \mathbb{E}_\mu(F_\epsilon))}\, d\mu \le e^{c\lambda^2\||\nabla F_\epsilon|^2\|_\infty/2} \le e^{c\lambda^2\|F\|_{\mathrm{Lip}}^2/2}. \tag{11.12}$$

Since $F_\epsilon \to F$ uniformly, and F is Lipschitz (ergo bounded), for all sufficiently small $\epsilon > 0$ we have $\|F_\epsilon\|_\infty \le \|F\|_\infty + 1$. Hence, by the dominated convergence theorem, $\mathbb{E}_\mu(F_\epsilon) \to \mathbb{E}_\mu(F)$ as $\epsilon \downarrow 0$. It then follows that $F_\epsilon - \mathbb{E}_\mu(F_\epsilon) \to F - \mathbb{E}_\mu(F)$ uniformly, and so for any $\lambda \in \mathbb{R}$,

$$e^{\lambda(F_\epsilon - \mathbb{E}_\mu(F_\epsilon))} \to e^{\lambda(F - \mathbb{E}_\mu(F))} \quad \text{uniformly as } \epsilon \downarrow 0.$$

The boundedness the right-hand-side therefore yields, by the dominated convergence theorem, that

$$\lim_{\epsilon\downarrow 0} \int e^{\lambda(F_\epsilon - \mathbb{E}_\mu(F_\epsilon))}\, d\mu = \int e^{\lambda(F - \mathbb{E}_\mu(F))}\, d\mu.$$

Combining this with Inequality 11.12 yields Inequality 11.1 for F, as desired. □

Remark 11.3. Theorem 11.1 is stated and proved for *globally* Lipschitz functions F; in particular, this means F is bounded, and we used this heavily in the proof. This will suffice for the applications we need it for, but it is worth mentioning that the theorem holds true for *locally* Lipschitz functions: i.e. F for which $\sup_{\mathbf{x}\ne\mathbf{y}} \frac{|F(\mathbf{x})-F(\mathbf{y})|}{|\mathbf{x}-\mathbf{y}|} < \infty$, but F is not necessarily bounded (it can grow sub-linearly at ∞). The proof need only be modified in the last part, approximating F by a smooth G with bounded partial derivatives. One must first cut off F so $|F|$ takes no values greater than $1/\epsilon$, and then mollify. (F_ϵ cannot converge uniformly to F any longer, so it is also more convenient to use a Gaussian mollifier for the sake of explicit computations.) The remainder of the proof follows much like above, until the last part, showing that the exponential moments converge appropriately as $\epsilon \downarrow 0$; this becomes quite tricky, as the dominated convergence theorem is no longer easily applicable. Instead, one uses the concentration of measure (Inequality 11.2) known to hold for the smoothed F_ϵ to prove the family $\{F_\epsilon\}$ is *uniformly integrable* for all small $\epsilon > 0$. This, in conjunction with Fatou's lemma, allows the completion of the limiting argument.

12. Concentration for Random Matrices

We wish to apply Herbst's concentration inequality to the Stieltjes transform of a random Gaussian matrix. To do so, we need to relate Lipschitz functions of the *entries* of the matrix to functions of the *eigenvalues*.

12.1. **Continuity of Eigenvalues.** Let X_n denote the linear space of $n\times n$ symmetric real matrices. As a vector space, $\mathscr{X}_n$ can be identified with $\mathbb{R}^{n(n+1)/2}$. However, the more natural inner product for $\mathscr{X}_n$ is

$$\langle \mathbf{X},\mathbf{Y}\rangle = \mathrm{Tr}\,[\mathbf{X}\mathbf{Y}] = \sum_{1\le i,j\le n}\mathbf{X}_{ij}\mathbf{Y}_{ij} = \sum_{i=1}^n \mathbf{X}_{ii}\mathbf{Y}_{ii} + 2\sum_{1\le i<j\le n}\mathbf{X}_{ij}\mathbf{Y}_{ij}.$$

We use this inner-product to define the norm on symmetrix matrices:

$$\|\mathbf{X}\|_2 = \langle \mathbf{X},\mathbf{X}\rangle^{1/2} = \left(\mathrm{Tr}\,[\mathbf{X}^2]\right)^{1/2}.$$

If we want this to match up with the standard Euclidean norm, the correct identification $\mathscr{X}_N \to \mathbb{R}^{N(N+1)/2}$ is

$$\mathbf{X}\mapsto (\mathbf{X}_{11},\ldots,\mathbf{X}_{nn},\tfrac{1}{\sqrt{2}}\mathbf{X}_{12},\tfrac{1}{\sqrt{2}}\mathbf{X}_{13},\ldots,\tfrac{1}{\sqrt{2}}\mathbf{X}_{n-1,n}).$$

That is: $\|\mathbf{X}\|_2 \le \sqrt{2}\|\mathbf{X}\|$ (where $\|\mathbf{X}\|^2 = \sum_{1\le i\le j\le n}\mathbf{X}_{ij}^2$ is the usual Euclidean norm). It is this norm that comes into play in the Lipschitz norm we used in the discussion of log Sobolev inequalities and the Herbst inequality, so it is the norm we should use presently.

Lemma 12.1. *Let $g\colon \mathbb{R}^n\to\mathbb{R}$ be a Lipschitz function. Extend g to a function on $\mathscr{X}_n$ (the space of symmetric $n\times n$ matrices) as follows: letting $\lambda_1(\mathbf{X}),\ldots,\lambda_n(\mathbf{X})$ denote the eigenvalues as usual, set $\tilde{g}\colon \mathscr{X}_n\to\mathbb{R}$ to be*

$$\tilde{g}(\mathbf{X}) = g(\lambda_1(\mathbf{X}),\ldots,\lambda_n(\mathbf{X})).$$

Then $\tilde{g}$ is Lipschitz, with $\|\tilde{g}\|_{\mathrm{Lip}}\le\sqrt{2}\|g\|_{\mathrm{Lip}}$.

Proof. Recall the Hoffman–Wielandt lemma, Lemma 5.4, which states that for any pair $\mathbf{X},\mathbf{Y}\in\mathscr{X}_n$,

$$\sum_{i=1}^n |\lambda_i(\mathbf{X})-\lambda_i(\mathbf{Y})|^2 \le \mathrm{Tr}\,[(\mathbf{X}-\mathbf{Y})^2].$$

Thus, we have simply

$$\begin{aligned}|\tilde{g}(\mathbf{X})-\tilde{g}(\mathbf{Y})| &= |g(\lambda_1(\mathbf{X}),\ldots,\lambda_n(\mathbf{X})) - g(\lambda_1(\mathbf{Y}),\ldots,\lambda_n(\mathbf{Y}))|\\ &\le \|g\|_{\mathrm{Lip}}\|(\lambda_1(\mathbf{X})-\lambda_1(\mathbf{Y}),\ldots,\lambda_n(\mathbf{X})-\lambda_n(\mathbf{Y}))\|_{\mathbb{R}^n}\\ &= \|g\|_{\mathrm{Lip}}\left(\sum_{i=1}^n|\lambda_i(\mathbf{X})-\lambda_i(\mathbf{Y})|^2\right)^{1/2}\\ &\le \|g\|_{\mathrm{Lip}}\left(\mathrm{Tr}\,[(\mathbf{X}-\mathbf{Y})^2]\right)^{1/2} = \|g\|_{\mathrm{Lip}}\|\mathbf{X}-\mathbf{Y}\|_2.\end{aligned}$$

The result follows from the inequality $\|\mathbf{X}-\mathbf{Y}\|_2\le\sqrt{2}\|\mathbf{X}-\mathbf{Y}\|$ discussed above. □

Corollary 12.2. *Let $f\in\mathrm{Lip}(\mathbb{R})$. Extend f to a map $f_{\mathrm{Tr}}\colon\mathscr{X}_n\to\mathbb{R}$ via*

$$f_{\mathrm{Tr}}(\mathbf{X}) = \mathrm{Tr}\,[f(\mathbf{X})],$$

where $f(\mathbf{X})$ is defined in the usual way using the Spectral Theorem. Then f_{Tr} is Lipschitz, with $\|f_{\mathrm{Tr}}\|_{\mathrm{Lip}}\le\sqrt{2n}\|f\|_{\mathrm{Lip}}$.

Proof. note that

$$f_{\mathrm{Tr}}(\mathbf{X}) = \sum_{i=1}^{n} f(\lambda_i(\mathbf{X})) = \tilde{g}(\mathbf{X})$$

where $g(\lambda_1, \dots, \lambda_n) = f(\lambda_1) + \cdots + f(\lambda_n)$. Well,

$$\begin{aligned} |g(\lambda_1, \dots, \lambda_n) - g(\mu_1, \dots, \mu_n)| &\leq |f(\lambda_1) - f(\mu_1)| + \cdots + |f(\lambda_n) - f(\mu_n)| \\ &\leq \|f\|_{\mathrm{Lip}}|\lambda_1 - \mu_1| + \cdots + \|f\|_{\mathrm{Lip}}|\lambda_n - \mu_n| \\ &\leq \|f\|_{\mathrm{Lip}} \cdot \sqrt{n}\|(\lambda_1 - \mu_1, \dots, \lambda_n - \mu_n)\|. \end{aligned}$$

This shows that $\|g\|_{\mathrm{Lip}} \leq \sqrt{n}\|f\|_{\mathrm{Lip}}$. The result follows from Lemma 12.1. □

12.2. **Concentration of the Empirical Eigenvalue Distribution.** Let $\mathbf{X}_n$ be a Wigner matrix. Denote by $\nu_{\mathbf{X}_n}$ the *joint-law of entries* of $\mathbf{X}_n$; this is a probability measure on $\mathbb{R}^{n^2}$ (or $\mathbb{R}^{n(n+1)/2}$ if we like). We will suppose that $\nu_{\mathbf{X}_n}$ satisfies a logarithmic Sobolev inequality LSI with constant c. Our primary example is the Gaussian Wigner matrix from Section 9.1. Recall this was defined by $\mathbf{X}_n = n^{-1/2}\mathbf{Y}_n$ where $\mathbf{Y}_{ii} = 0$ and, for $i < j$, $\mathbf{X}_n \sim N(0,1)$. This means that $\nu_{\mathbf{X}_n}$ should properly be thought of as a probability measure on $\mathbb{R}^{n(n-1)/2}$, and as such it is a standard Gaussian measure with variance $\frac{1}{n}$: i.e. $\nu_{\mathbf{X}_n} = \gamma_{\frac{1}{n}}$ on $\mathbb{R}^{n(n-1)/2}$. By Theorem 10.3, this law satisfies the LSI with constant $c = \frac{1}{n}$. Hence, by the Herbst inequality (Theorem 11.1), if F is any Lipschitz function on $\mathbb{R}^{n(n-1)/2}$, we have

$$\nu_{\mathbf{X}_n}\left\{|F - \mathbb{E}_{\nu_{\mathbf{X}_n}}(F)| \geq \delta\right\} \leq 2e^{-\delta^2/2c\|F\|_{\mathrm{Lip}}^2} = 2e^{-n\delta^2/2\|F\|_{\mathrm{Lip}}^2}. \tag{12.1}$$

Now, let $f\colon \mathbb{R} \to \mathbb{R}$ be Lipschitz, and apply this with $F = f_{\mathrm{Tr}}$. Then we have

$$\mathbb{E}_{\nu_{\mathbf{X}_n}}(F) = \int \mathrm{Tr}\, f(\mathbf{X})\, \nu_{\mathbf{X}_n}(d\mathbf{X}) = \mathbb{E}(\,\mathrm{Tr}\, f(\mathbf{X}_n)) = n\mathbb{E}\left(\int f\, d\mu_{\mathbf{X}_n}\right)$$

where $\mu_{\mathbf{X}_n}$ is, as usual, the empirical distribution of eigenvalues. So, we have

$$\begin{aligned} \nu_{\mathbf{X}_n}\left\{\left|f_{\mathrm{Tr}} - n\int f\, d\mu_{\mathbf{X}_n}\right| \geq \delta\right\} &= \nu_{\mathbf{X}_n}\left\{\mathbf{X}\colon \left|\mathrm{Tr}\, f(\mathbf{X}) - n\mathbb{E}\left(\int f\, d\mu_{\mathbf{X}_n}\right)\right| \geq \delta\right\} \\ &= \nu_{\mathbf{X}_n}\left\{\mathbf{X}\colon \left|n\int f\, d\mu_{\mathbf{X}} - n\mathbb{E}\left(\int f\, d\mu_{\mathbf{X}_n}\right)\right| \geq \delta\right\} \\ &= \mathbb{P}\left(\left|\int f\, d\mu_{\mathbf{X}_n} - \mathbb{E}\left(\int f\, d\mu_{\mathbf{X}_n}\right)\right| \geq {!}n\right). \end{aligned}$$

Combining this with Equation 12.1, and letting $\epsilon = {!}n$, this gives

$$\mathbb{P}\left(\left|\int f\, d\mu_{\mathbf{X}_n} - \mathbb{E}\left(\int f\, d\mu_{\mathbf{X}_n}\right)\right| \geq \epsilon\right) \leq 2e^{-n^3\epsilon^2/2\|f_{\mathrm{Tr}}\|_{\mathrm{Lip}}^2}.$$

Now, from Corollary 12.2, $\|f_{\mathrm{Tr}}\|_{\mathrm{Lip}} \leq \sqrt{2n}\|f\|_{\mathrm{Lip}}$. Hence $-\frac{1}{\|f_{\mathrm{Tr}}\|_{\mathrm{Lip}}^2} \leq -\frac{1}{2n}\frac{1}{\|f\|_{\mathrm{Lip}}^2}$. We have thence proved the following theorem.

Theorem 12.3. *Let $\mathbf{X}_n$ be the Gaussian Wigner matrix of Section 9.1 (or, more generally, any Wigner matrix whose joint law of entries satisfies a LSI with constant $c = \frac{1}{n}$). Let $f \in \mathrm{Lip}(\mathbb{R})$. Then*

$$\mathbb{P}\left(\left|\int f\, d\mu_{\mathbf{X}_n} - \mathbb{E}\left(\int f\, d\mu_{\mathbf{X}_n}\right)\right| \geq \epsilon\right) \leq 2e^{-n^2\epsilon^2/4\|f\|_{\mathrm{Lip}}^2}. \tag{12.2}$$

Thus, the random variable $\int f\, d\mu_{\mathbf{X}_n}$ converges in probability to its mean. The rate of convergence is very fast: we have *normal concentration* (i.e. at least as fast as e^{-cn^2} for some $c > 0$). In this case, we can actually conclude almost sure convergence.

Corollary 12.4. *Let $\mathbf{X}_n$ satisfy Inequality 12.2. Let $f \in \mathrm{Lip}(\mathbb{R})$. Then as $n \to \infty$*

$$\int f\, d\mu_{\mathbf{X}_n} \to \mathbb{E}\left(\int f\, d\mu_{\mathbf{X}_n}\right) \quad a.s.$$

Proof. Fix $\epsilon > 0$, and note that

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left|\int f\, d\mu_{\mathbf{X}_n} - \mathbb{E}\left(\int f\, d\mu_{\mathbf{X}_n}\right)\right| \geq \epsilon\right) \leq 2\sum_{n=1}^{\infty} 2^{-n^2\epsilon^2/4\|f\|_{\mathrm{Lip}}^2} < \infty.$$

By the Borel-Cantelli lemma, it follows that

$$\mathbb{P}\left(\left|\int f\, d\mu_{\mathbf{X}_n} - \mathbb{E}\left(\int f\, d\mu_{\mathbf{X}_n}\right)\right| \geq \epsilon \;\; i.o.\right) = 0$$

That is, the event that $\left|\int f\, d\mu_{\mathbf{X}_n} - \mathbb{E}\left(\int f\, d\mu_{\mathbf{X}_n}\right)\right| < \epsilon$ for all sufficiently large n has probability 1. This shows that we have almost sure convergence. □

We would also like to have L^2 (or generally L^p) convergence, which would follow from the a.s. convergence using the dominated convergence theorem if there were a natural dominating bound. Instead, we will prove L^p-convergence directly from the normal concentration about the mean, using the following very handy result.

Proposition 12.5 (Layercake Representation)**.** *Let $X \geq 0$ be a random variable. Let κ be a positive Borel measure on $[0,\infty)$, and define $\phi(x) = \kappa([0,x])$. Then*

$$\mathbb{E}(\phi(X)) = \int_0^{\infty} \mathbb{P}(X \geq t)\, \kappa(dt).$$

Proof. Using positivity of integrands, we apply Fubini's theorem to the right-hand-side:

$$\int_0^{\infty} \mathbb{P}(X \geq t)\, \kappa(dt) = \int_0^{\infty}\int \mathbb{1}_{\{X\geq t\}}\, d\mathbb{P}\, \kappa(dt) = \int\int_0^{\infty} \mathbb{1}_{\{X\geq t\}}\, \kappa(dt)\, d\mathbb{P}.$$

The inside integral is

$$\int_0^{\infty} \mathbb{1}_{\{X\geq t\}}\, \kappa(dt) = \int_0^{X} \kappa(dt) = \phi(X),$$

proving the result. □

Corollary 12.6. *Let $\mathbf{X}_n$ satisfy Inequality 12.2. Let $f \in \mathrm{Lip}(\mathbb{R})$, and let $p \geq 1$. Then*

$$\lim_{n\to\infty} \mathbb{E}\left[\left|\int f\, d\mu_{\mathbf{X}_n} - \mathbb{E}\left(\int f\, d\mu_{\mathbf{X}_n}\right)\right|^p\right] = 0.$$

Proof. Let $\kappa_p(dt) = pt^{p-1}\, dt$ on $[0,\infty)$; then the corresponding cumulative function is $\phi_p(x) = \int_0^x d\kappa_p = \int_0^x pt^{p-1}\, dt = x^p$. Applying the Layercake representation (Proposition 12.5) to the random variable $X_n = \left|\int f\, d\mu_{\mathbf{X}_n} - \mathbb{E}\left(\int f\, d\mu_{\mathbf{X}_n}\right)\right|$ yields

$$\mathbb{E}(X_n^p) = \int_0^{\infty} \mathbb{P}(X_n \geq t) pt^{p-1}\, dt \leq 2p\int_0^{\infty} t^{p-1}e^{-n^2t^2/4\|f\|_{\mathrm{Lip}}^2}\, dt.$$

Make the change of variables $s = nt/2\|f\|_{\mathrm{Lip}}$; then the integral becomes

$$\int_0^\infty t^{p-1}e^{-n^2t^2/4\|f\|_{\mathrm{Lip}}^2}\,dt = \int_0^\infty \left(\frac{2\|f\|_{\mathrm{Lip}}s}{n}\right)^{p-1} e^{-s^2}\,\frac{2\|f\|_{\mathrm{Lip}}}{n}\,ds = \left(\frac{2\|f\|_{\mathrm{Lip}}}{n}\right)^p \int_0^\infty s^{p-1}e^{-s^2}\,ds.$$

The integral is some finite constant M_p, and so we have

$$\mathbb{E}(X_n^p) \le 2pM_p \cdot \left(\frac{2\|f\|_{\mathrm{Lip}}}{n}\right)^p \to 0$$

as $n \to \infty$, as required. □

12.3. **Return to Wigner's Theorem.** Now, for fixed $z \in \mathbb{C} \setminus \mathbb{R}$, take $f_z(t) = \frac{1}{t-z}$. Then f is bounded and C^1, and $f_z'(t) = -\frac{1}{(t-z)^2}$; thus

$$\|f_z\|_{\mathrm{Lip}} = \|f_z'\|_\infty = |\Im z|^{-2}.$$

By definition $\int f_z\,d\mu_{\mathbf{X}_n} = S_{\mu_{\mathbf{X}_n}}(z)$ is the Stieltjes transform. In this context, Corollary 12.6 with $p = 2$ gives us

$$\lim_{n\to\infty} \mathbb{E}\left[\left|S_{\mu_{\mathbf{X}_n}}(z) - \mathbb{E}S_{\mu_{\mathbf{X}_n}}(z)\right|^2\right] = 0.$$

It follows that

$$\begin{aligned}\left|\mathbb{E}(S_{\mu_{\mathbf{X}_n}}(z)^2) - \left(\mathbb{E}S_{\mu_{\mathbf{X}_n}}(z)\right)^2\right| &= \left|\mathbb{E}\left[\left(S_{\mu_{\mathbf{X}_n}}(z) - \mathbb{E}S_{\mu_{\mathbf{X}_n}}(z)\right)^2\right]\right| \\ &\le \mathbb{E}\left[\left|S_{\mu_{\mathbf{X}_n}}(z) - \mathbb{E}S_{\mu_{\mathbf{X}_n}}(z)\right|^2\right] \to 0 \text{ as } n \to \infty,\end{aligned}$$

which confirms Equation 9.8. This finally completes the argument in Section 9.1 demonstrating that the Stieltjes transform of $\mathbf{X}_n$ satisfies

$$\lim_{n\to\infty} \mathbb{E}S_{\mu_{\mathbf{X}_n}}(z) = S_{\sigma_1}(z)$$

and ergo the *averaged* empirical eigenvalue distribution converges to the semicircular law σ_1. (This is what we already proved in the general case of finite moments in Section 4.) But we can also combine this result immediately with Corollary 12.4: since we also know that $S_{\mu_{\mathbf{X}_n}}(z) \to \mathbb{E}S_{\mu_{\mathbf{X}_n}}(z)$ a.s. for each z, we conclude that for all $z \in \mathbb{C} \setminus \mathbb{R}$

$$\lim_{n\to\infty} S_{\mu_{\mathbf{X}_n}}(z) = S_{\sigma_1}(z) \quad a.s.$$

Now employing the Stieljtes Continuity Theorem 8.11, we conclude that $\mu_{\mathbf{X}_n} \to \sigma_1$ weakly a.s. This completes the proof of Wigner's theorem (for Gaussian Wigner matrices) in the almost sure convergence form.

13. Gaussian Wigner Matrices, and the Genus Expansion

We have now seen a complete proof of the strong Wigner Law, in the case that the (upper-triangular) entries of the matrix are real Gaussians. (As we verified in Section 5, the diagonal does not influence the limiting empirical eigenvalue distribution.) We will actually spend the rest of the course talking about this (and a related) specific matrix model. There are good reasons to believe that basically all fundamental limiting behavior of Wigner eigenvalues is universal (does not depend on the distribution of the entries), and so we may as well work in a model where finer calculations are possible. (Note: in many cases, whether or not the statistics are truly universal is still very much a front-line research question.)

13.1. GOE_n and GUE_n.

For $j \geq i \geq 1$, let $\{B_{jk}\}$ and $\{B'_{jk}\}$ be two families of independent $N(0,1)$ random variables (independent from each other as well). We define a **Gaussian Orthogonal Ensemble** GOE_n to be the sequence of Wigner matrices $\mathbf{X}_n$ with

$$[\mathbf{X}_n]_{jk} = [\mathbf{X}_n]_{kj} = n^{-1/2}\frac{1}{\sqrt{2}}B_{jk},\ 1 \leq j < k \leq n$$
$$[\mathbf{X}_n]_{jj} = n^{-1/2}B_{jj},\ 1 \leq j \leq n.$$

Except for the diagonal entries, this is the matrix we studied through the last three sections. It might seem natural to have the diagonal also consist of variance $1/n$ normals, but in fact this normalization is better: it follows (cf. the discussion at the beginning of Section 12.1) that the Hilbert Schmidt norm $\|\mathbf{X}_n\|_2^2 = \operatorname{Tr}[\mathbf{X}_n^2]$ exactly corresponds to the Euclidean norm of the $n(n+1)/2$-dimensional vector given by the upper-triangular entries.

Let us now move into the complex world, and consider a Hermitian version of the GOE_n. Let $\mathbf{Z}_n$ be the $n \times n$ complex matrix whose entries are

$$[\mathbf{Z}_n]_{jk} = \overline{[\mathbf{Z}_n]_{kj}} = n^{-1/2}\frac{1}{\sqrt{2}}(B_{jk} + iB'_{jk}),\ 1 \leq j < k \leq n$$
$$[\mathbf{Z}_n]_{jj} = n^{-1/2}B_{jj},\ 1 \leq j \leq n.$$

(Note: the i above is $i = \sqrt{-1}$.) The sequence $\mathbf{Z}_n$ is called a **Gaussian Unitary Ensemble** GUE_n. Since $\mathbf{Z}_n$ is a Hermitian $n \times n$ matrix, its real dimension is $n + 2 \cdot \frac{n(n-1)}{2} = n^2$ (n real entries on the diagonal, $\frac{n(n-1)}{2}$ complex strictly upper-triangular entries). Again, one may verify that normalizing the off-diagonal entries to scale with $\frac{1}{\sqrt{2}}$ of the diagonal entries gives an isometric correspondence between the Hilbert-Schmidt norm and the $\mathbb{R}^{n^2}$-Euclidean norm.

Although we have worked exclusively with real Wigner matrices thus far, Wigner's theorem applies equally well to complex Hermitian matrices (whose eigenvalues are real, after all). The combinatorial proof of Wigner's theorem requires a little modification to deal with the complex conjugations involved, but there are no substantive differences. In fact, we will give an extremely concise version of this combinatorial proof for the GUE_n in this section. In general, there are good reasons to work on the complex side: most things work more or less the same, but the complex structure gives some new simplifying tools.

Why "Gaussian Orthogonal Ensemble" and "Gaussian Unitary Ensemble"? To understand this, we need to write down the joint-law of entries. Let us begin with the GOE_n $\mathbf{X}_n$. Let V be a Borel subset of $\mathbb{R}^{n(n+1)/2}$, and for now assume B is a Cartesian product set: $V = \prod_{j \leq k} V_{jk}$. Let $\tilde{V}$ be the set V identified as a subset of the symmetrix $n \times n$ matrices. By the independence of the entries, we have

$$\mathbb{P}(\sqrt{n}\mathbf{X}_n \in \tilde{V}) = \prod_{j\le k} \mathbb{P}(\sqrt{n}[\mathbf{X}_n]_{jk} \in V_{jk}).$$

For $j < k$, $\sqrt{n}[\mathbf{X}_n]_{jk} = \frac{1}{\sqrt{2}}B_{jk}$ is a $N(0, \frac{1}{2})$ random variable; for the diagonal entries, $\sqrt{n}[\mathbf{X}_n]_{jj} = B_{jj}$ is a $N(0,1)$ random variable. Hence, the above product is

$$\prod_{j=1}^{n} \int_{V_{jj}} (2\pi)^{-1/2} e^{-x_{jj}^2/2}\, dx_{jj} \cdot \prod_{j<k} \int_{V_{jk}} \pi^{-1/2} e^{-x_{jk}^2}\, dx_{jk}.$$

Let $d\mathbf{x}$ denote the Lebesgue measure on $\mathbb{R}^{n(n+1)/2}$. Rearranging this product, we have

$$\mathbb{P}(\sqrt{n}\mathbf{X}_n \in \tilde{V}) = 2^{-n/2}\pi^{-n(n+1)/4} \int_V e^{-\frac{1}{2}\sum_{j=1}^n x_{jj}^2 - \sum_{j<k} x_{jk}^2}\, d\mathbf{x}.$$

Having verified this formula for product sets V, it naturally holds for all Borel sets. The Gaussian density on the right-hand-side may not look particularly friendly (having difference variances in different coordinates), but in fact this is extremely convenient. If we think of the variables as the entries of a symmetric matrix $\mathbf{X}$, then we have

$$\mathrm{Tr}\,\mathbf{X}^2 = \sum_{j,k} x_{jk}^2 = \sum_{j=1}^n x_{jj}^2 + 2\sum_{j<k} x_{jk}^2.$$

In light of this, let us write everything in terms of symmetric matrices (so we suppress the $\tilde{V}$ in favor of V being a set of symmetric matrices). Let $d\mathbf{X}$ denote the Lebesgue measure on the $n(n+1)/2$-dimensional space of symmetric real matrices. Hence, the law of $\mathbf{X}_n$ can be written as

$$\mathbb{P}(\sqrt{n}\mathbf{X}_n \in V) = 2^{-n/2}\pi^{-n(n+1)/2} \int_V e^{-\frac{1}{2}\mathrm{Tr}\,\mathbf{X}^2}\, d\mathbf{X}. \tag{13.1}$$

By making the change of variables $\mathbf{Y} = \mathbf{X}/\sqrt{n}$, we have $d\mathbf{Y} = (\sqrt{n})^{-n(n+1)/2} d\mathbf{X}$, and so

$$\mathbb{P}(\sqrt{n}\mathbf{X}_n \in V) = 2^{-n/2}\pi^{-n(n+1)/4} n^{n(n+1)/4} \int_{V/\sqrt{n}} e^{-\frac{1}{2}n\,\mathrm{Tr}\,\mathbf{Y}^2}\, d\mathbf{Y}.$$

Letting $W = V/\sqrt{n}$, this shows us that the joint law of entries $\nu_{\mathbf{X}_n}$ has a density with respect to Lebesgue measure,

$$\frac{d\nu_{\mathbf{X}_n}}{d\mathbf{X}} = C_n e^{-\frac{1}{2}n\,\mathrm{Tr}\,\mathbf{X}^2}$$

where $C_n = 2^{-n/2}\pi^{-n(n+1)/4}n^{n(n+1)/4}$ is a normalization coefficient. It will usually be more convenient to work with Equation 13.1, which gives the density of the joint law of entries $\nu_{\sqrt{n}\mathbf{X}_n}$.

Now, let $\mathbf{Q}$ be an orthogonal $n \times n$ matrix. If we use it to rotate $\mathbf{X}_n$, we can calculate from Equation 13.1 that

$$\mathbb{P}(\sqrt{n}\mathbf{Q}\mathbf{X}_n\mathbf{Q}^\top \in V) = \mathbb{P}(\sqrt{n}\mathbf{X}_n \in \mathbf{Q}^\top V\mathbf{Q}) = 2^{-n/2}\pi^{-n(n+1)/2} \int_{\mathbf{Q}^\top V\mathbf{Q}} e^{-\frac{1}{2}\mathrm{Tr}\,\mathbf{X}^2}\, d\mathbf{X}.$$

If we make the change of variables $\mathbf{Y} = \mathbf{Q}\mathbf{X}\mathbf{Q}^\top$, this is a linear transformation and so its Jacobian (derivative) is itself. In terms of the Hlibert-Schmidt norm (which is the Euclidean distance), we have $\|\mathbf{Y}\|_2^2 = \mathrm{Tr}\,\mathbf{Y}^2 = \mathrm{Tr}\,(\mathbf{Q}\mathbf{X}\mathbf{Q}^\top)^2 = \mathrm{Tr}\,(\mathbf{Q}\mathbf{X}\mathbf{Q}^\top\mathbf{Q}\mathbf{X}\mathbf{Q}^\top) = \mathrm{Tr}\,(\mathbf{Q}\mathbf{X}^2\mathbf{Q}^\top) = \mathrm{Tr}\,(\mathbf{Q}^\top\mathbf{Q}\mathbf{X}^2) = \mathrm{Tr}\,(\mathbf{X}^2) = \|\mathbf{X}\|_2^2$, where the penultimate equality uses the trace property $\mathrm{Tr}\,(AB) =$

$\mathrm{Tr}\,(BA)$. Thus, the map $\mathbf{X} \mapsto \mathbf{Y} = \mathbf{Q}\mathbf{X}\mathbf{Q}^\top$ is an isometry, and thus its Jacobian determinant is 1: i.e. $d\mathbf{X} = d\mathbf{Y}$. Also by the above calculation we have $\mathrm{Tr}\,\mathbf{X}^2 = \mathrm{Tr}\,\mathbf{Y}^2$. So, in total, this gives

$$\mathbb{P}(\sqrt{n}\mathbf{Q}\mathbf{X}_n\mathbf{Q}^\top \in V) = 2^{-n/2}\pi^{-n(n+1)/2}\int_V e^{-\frac{1}{2}\mathrm{Tr}\,\mathbf{Y}^2}\,d\mathbf{Y} = \mathbb{P}(\sqrt{n}\mathbf{X}_n \in V).$$

This is why $\mathbf{X}_n$ is called a Gaussian *orthogonal* ensemble: its joint law of entries is invariant under the natural (conjugation) action of the orthogonal group. That is: if we conjugate a GOE_n by a fixed orthogonal matrix, the resulting matrix is another GOE_n.

Turning now to the GUE_n, let us compute its joint law of entries. The calculation is much the same, in fact, with the modification that there are twice as many off-diagonal entries. A product Borel set looks like $V = \prod_j V_j \times \prod_{j<k} V_{jk} \times \prod_{j<k} V'_{jk}$, and independence again gives

$$\mathbb{P}(\sqrt{n}\mathbf{Z}_n \in V) = \prod_{j=1}^n \int_{V_j}(2\pi)^{-1/2}e^{-x_{jj}^2/2}\,dx_{jj}\cdot\prod_{j<k}\int_{V_{jk}}\pi^{-1/2}e^{-x_{jk}^2}\,dx_{jk}\int_{V'_{jk}}\pi^{-1/2}e^{-(x'_{jk})^2}\,dx'_{jk}.$$

Collecting terms as before, this becomes

$$2^{-n/2}\pi^{-n^2/2}\int_V e^{-\frac{1}{2}\sum_{j=1}^n x_{jj}^2 - \sum_{j<k}(x_{jk}^2+(x'_{jk})^2)}\,d\mathbf{x},$$

where this time $d\mathbf{x}$ denotes Lebesgue measure on the n^2-dimensional real vector space of Hermitian $n \times n$ matrices. Now, if we interpret the variables x_{jk} and x'_{jk} as the real and imaginary parts, then we have for any Hermitian matrix $\mathbf{Z} = \mathbf{Z}^* = \overline{\mathbf{Z}}^\top$

$$\mathrm{Tr}\,(\mathbf{Z}^2) = \mathrm{Tr}\,(\mathbf{Z}\mathbf{Z}^*) = \sum_{j,k}|[\mathbf{Z}]_{jk}|^2 = \sum_{j=1}^n x_{jj}^2 + 2\sum_{j<k}|x_{jk} + ix'_{jk}|^2$$

which then allows us to identify the law of $\mathbf{Z}_n$ as

$$\mathbb{P}(\sqrt{n}\mathbf{Z}_n \in V) = 2^{-n/2}\pi^{-n^2/2}\int_V e^{-\frac{1}{2}\mathrm{Tr}\,\mathbf{Z}^2}\,d\mathbf{Z} \tag{13.2}$$

where $d\mathbf{Z}$ denotes the Lebesuge measure on the n^2-dimensional real vector space of Hermitian $n \times n$ matrices.

An entirely analogous argument to the one above, based on formula 13.2, shows that if $\mathbf{U}$ is an $n \times n$ *unitary* matrix, then $\mathbf{U}\mathbf{Z}_n\mathbf{U}^*$ is also a GUE_n. This is the reason for the name Gaussian *unitary* ensemble.

13.2. **Covariances of GUE_n.** It might seem that the only difference between the GOE_n and the GUE_n is dimensional. This is true for the joint laws, which treat the two matrix-valued random variables simply as random vectors. However, taking the matrix structure into account, we see that they have somewhat difference covariances.

Let us work with the GUE_n. For convenience, denote $Z_{jk} = [\mathbf{Z}_n]_{jk}$. Fix (j_1, k_1) and (j_2, k_2); for the time being, suppose they are both strictly upper-triangular. Then

$$\begin{aligned}\mathbb{E}(Z_{j_1k_1}Z_{j_2k_2}) &= \mathbb{E}\left[(2n)^{-1/2}(B_{j_1k_1} + iB'_{j_1k_1})\cdot(2n)^{-1/2}(B_{j_2k_2} + iB'_{j_2k_2})\right]\\ &= \frac{1}{2n}\left(\mathbb{E}(B_{j_1k_1}B_{j_2k_2}) + i\mathbb{E}(B_{j_1k_1}B'_{j_2k_2}) + i\mathbb{E}(B'_{j_1k_1}B_{j_2k_2}) - \mathbb{E}(B'_{j_1k_1}B'_{j_2k_2})\right).\end{aligned}$$

Because all of the B_{jk} and B'_{jk} are independent, all four of these terms are 0 unless $(j,k) = (j',k')$. In this special case, the two middle terms are still 0 (since B_{jk} and $B'_{\ell m}$ are independent for all j,k,ℓ,m), and the two surviving terms give

$$\mathbb{E}(B_{j_1k_1}^2) - \mathbb{E}((B'_{j_1k_1})^2) = 1 - 1 = 0.$$

That is, the covariance of any two strictly upper-triangular entries is 0. (Note: in the case $(j_1,k_1) = (j_2,k_2)$ considered above, what we have is the at-first-hard-to-believe fact that if Z is a *complex* Gaussian random variable, then $\mathbb{E}(Z^2) = 0$. It is basically for this reason that the GUE_n is, in many contexts, easier to understand than the GOE_n.) Since $Z_{jk} = \overline{Z_{kj}}$, it follows that the covariance of any two strictly lower-triangular entries is 0.

Now, let us take the covariance of a diagonal entry with an off-diagonal entry.

$$\begin{aligned}\mathbb{E}(Z_{jk}Z_{mm}) &= \mathbb{E}\left[(2n)^{-1/2}(B_{jk} + iB'_{jk})\cdot n^{-1/2}B_{mm}\right] \\ &= \frac{1}{\sqrt{2}n}\left(\mathbb{E}(B_{jk}B_{mm}) + i\mathbb{E}(B'_{jk}B_{mm})\right).\end{aligned}$$

Since $j \neq k$, independence gives 0 for both expectations, and so again we get covariance 0. Similarly, for two diagonal entries,

$$\mathbb{E}(Z_{jj}Z_{kk}) = \mathbb{E}\left[n^{-1/2}B_{jj}\cdot n^{-1/2}B_{kk}\right] = \frac{1}{n}\mathbb{E}\left(B_{jj}B_{kk}\right) = \frac{1}{n}\delta_{jk}.$$

Of course we do get non-zero contributions from the variances of the diagonal entries, but independence gives 0 covariances for distinct diagonal entries.

Finally, let us consider the covariance of a strictly-upper-triangular entry $Z_{j_1k_1}$ with a strictly lower-triangular entry $Z_{j_2k_2}$.

$$\begin{aligned}\mathbb{E}(Z_{j_1k_1}Z_{j_2k_2}) &= \mathbb{E}\left[(2n)^{-1/2}(B_{j_1k_1} + iB'_{j_1k_1})\cdot(2n)^{-1/2}(B_{j_2k_2} - iB'_{j_2k_2})\right] \\ &= \frac{1}{2n}\left(\mathbb{E}(B_{j_1k_1}B_{j_2k_2}) - i\mathbb{E}(B_{j_1k_1}B'_{j_2k_2}) + i\mathbb{E}(B'_{j_1k_1}B_{j_2k_2}) + \mathbb{E}(B'_{j_1k_1}B'_{j_2k_2})\right).\end{aligned}$$

As noted before, the two middle terms are always 0. The first and last terms are 0 unless $(j_1,k_1) = (j_2,k_2)$, in which case the sum becomes $\frac{1}{n}$. That is

$$\mathbb{E}(Z_{j_1k_1}Z_{j_2k_2}) = \frac{1}{n}\delta_{j_1j_2}\delta_{k_1k_2}$$

in this case. We can then summarize all the above calculations together as follows: for all $1 \leq j,k,\ell,m \leq n$, we have

$$\mathbb{E}[Z_{jk}Z_{\ell m}] = \frac{1}{n}\delta_{jm}\delta_{k\ell}. \tag{13.3}$$

In other words: $\mathbb{E}(|Z_{jk}|^2) = \mathbb{E}(Z_{jk}Z_{jk}) = \frac{1}{n}$, and all other covariances are 0.

Remark 13.1. The calculation for covariances of the entries $X_{jk} = [\mathbf{X}_n]_{jk}$ of a GOE_n are much simpler, but the result is more complicated: because there is no distinction between $\mathbb{E}(X_{jk}^2)$ and $\mathbb{E}(|X_{jk}|^2)$ in this case, the covariances are

$$\mathbb{E}(X_{jk}X_{\ell m}) = \frac{1}{n}(\delta_{j\ell}\delta_{km} + \delta_{jm}\delta_{k\ell}). \tag{13.4}$$

The additional term presents serious challenges for the analysis in the next two sections, and demonstrates some substantial differences in the very fine structure of the GUE_n versus the GOE_n.

13.3. **Wick's Theorem.** Formula 13.3 expresses more than just the covariances; implicitly, it allows the direct calculation of *all mixed moments* in the Z_{jk}. This is yet another wonderful symmetry of Gaussian random variables. The result is known as *Wick's theorem*, although in the form it was stated by Wick (in the physics literature) it is unrecognizable. It is, however, the same statement below. First, we need a little bit of notation.

Definition 13.2. *Let m be even. A* **pairing** *of the set $\{1,\ldots,m\}$ is a collection of disjoint two element sets (pairs) $\{j_1,k_1\},\ldots,\{j_{m/2},k_{m/2}\}$ such that $\{j_1,\ldots,j_{m/2},k_1,\ldots,k_{m/2}\} = \{1,\ldots,m\}$. The set of all pairings is denoted $\mathscr{P}_2(m)$. If m is odd, $\mathscr{P}_2(m) = \varnothing$.*

Pairings may be thought of as special permutations: the pairing $\big\{\{j_1,k_1\},\ldots,\{j_{m/2},k_{m/2}\}\big\}$ can be identified with the permutation $(j_1,k_1)\cdots(j_{m/2},k_{m/2}) \in S_{m/2}$ (written in cycle-notation). This gives a bijection between $\mathscr{P}_2(m)$ and the set of *fixed-point-free involutions* in $S_{m/2}$. We usually denoted pairings by lower-case Greek letters π,σ,τ.

Theorem 13.3 (Wick)**.** *Let $B_1,\ldots,B_m$ be independent normal $N(0,1)$ random variables, and let $X_1,\ldots,X_m$ be linear functions of $B_1,\ldots,B_m$. Then*

$$\mathbb{E}(X_1\cdots X_m) = \sum_{\pi\in\mathscr{P}_2(m)} \prod_{\{j,k\}\in\pi} \mathbb{E}(X_jX_k). \tag{13.5}$$

In particular, when m is odd, $\mathbb{E}(X_1\cdots X_m) = 0$.

Remark 13.4. The precise condition in the theorem is that there exists a linear map $T\colon \mathbb{C}^m \to \mathbb{C}^m$ such that $(X_1,\ldots,X_m) = T(B_1,\ldots,B_m)$; complex coefficients are allowed. The theorem is sometimes stated instead for a "jointly Gaussian random vector $(X_1,\ldots,X_m)$", which is actually more restrictive: that is the requirement that the joint law of $(X_1,\ldots,X_m)$ has a density of the form $(2\pi)^{-m/2}(\det C)^{-1/2}e^{-\frac{1}{2}\langle \mathbf{x}C^{-1}\mathbf{x}\rangle}$ where C is a positive definite matrix (which turns out to be the covariance matrix of the entries). One can easily check that the random vector $(B_1,\ldots,B_m) = C^{-1/2}(X_1,\ldots,X_m)$ is a standard normal, and so this fits into our setting above (the linear map is $T = C^{1/2}$) which allows also for "degenerate Gaussians".

Proof. Suppose we have proved Wick's formula for a mixture of the i.i.d. normal variables $\mathbf{B}$:

$$\mathbb{E}(B_{i_1}\cdots B_{i_m}) = \sum_{\pi\in\mathscr{P}_2(m)} \prod_{\{a,b\}\in\pi} \mathbb{E}(B_{i_a}B_{i_b}). \tag{13.6}$$

(We do not assume that the i_a are all distinct.) Then since $(X_1,\ldots,X_m) = T(B_1,\ldots,B_m)$ for linear T, $\mathbb{E}(X_1\cdots X_m)$ is a linear combination of terms on the left-hand-side of (13.6). Similarly, the right-hand-side of (13.6) is multi-linear in the entries, and so with Wick sum $\sum_{\pi\in\mathscr{P}_2(m)}\prod_{\{i,j\}\in\pi}\mathbb{E}(X_iX_j)$ is *the same* linear combination of terms on the right-hand-side of (13.6). Hence, to prove formula (13.5), it suffices to prove formula (13.6).

We proceed by induction on m. We use Gaussian Integration by Parts (Theorem 9.1) to calculate

$$\begin{aligned}
\mathbb{E}\left[(B_{i_1}\cdots B_{i_{m-1}})\cdot B_{i_m}\right] &= \mathbb{E}\left[\partial_{B_{i_m}}(B_{i_1}\cdots B_{i_{m-1}})\right] \\
&= \sum_{a=1}^{m-1} \mathbb{E}\left[(\partial_{B_{i_m}}B_{i_a})\cdot(B_{i_1}\cdots B_{i_{a-1}}B_{i_{a+1}}\cdots B_{i_{m-1}})\right].
\end{aligned}$$

Of course, $\partial_{B_{i_m}} B_{i_a} = \delta_{i_m i_a}$, but since the B_i are indepenedent $N(0,1)$ random variables, this is also equal to $\mathbb{E}(B_{i_m} B_{i_a})$. Hence, we have

$$\mathbb{E}(B_{i_1}\cdots B_{i_m}) = \sum_{a=1}^{m-1} \mathbb{E}(B_{i_a}B_{i_m})\mathbb{E}(B_{i_1}\cdots B_{i_{a-1}}B_{i_{a+1}}\cdots B_{i_{m-1}}). \tag{13.7}$$

By the inductive hypothesis applied to the variables $B_{i_1},\ldots,B_{i_{m-1}}$, the inside expectations are sums over pairings. To ease notation, let's only consider the case $a = m-1$:

$$\mathbb{E}(B_{i_1}\cdots B_{i_{m-2}}) = \sum_{\pi\in\mathscr{P}_2(m-2)} \prod_{\{c,d\}\in\pi} \mathbb{E}(B_{i_c}B_{i_d}).$$

Given any $\pi \in \mathscr{P}_2(m-2)$, let π_{m-1} denote the pairing in $\mathscr{P}_2(m)$ that pairs $\{m-1,m\}$ along with all the pairings in π. Then we can write the $a = m-1$ terms in Equation 13.7 as

$$\mathbb{E}(B_{i_{m-1}}B_{i_m})\cdot \sum_{\pi\in\mathscr{P}_2(m-2)} \prod_{\{c,d\}\in\pi} \mathbb{E}(B_{i_c}B_{i_d}) = \sum_{\pi\in\mathscr{P}_2(m-2)} \prod_{\{c',d'\}\in\pi_{m-1}} \mathbb{E}(B_{i_{c'}}B_{i_{d'}}).$$

The other terms in Equation 13.7 are similar: given a pairing $\pi \in \mathscr{P}_2(m)$, suppose that m pairs with a; then we can decompose π as a pairing of the indices $\{1,\ldots,a-1,a+1,\ldots,m-1\}$ (which can be though of as living in $\mathscr{P}_2(m-2)$ together with the pair $\{a,m\}$; the product of covariances of the B_i over such a π is then equal to the ath term in Equation 13.7. Hence, we recover all the terms in Equation 13.6, proving the theorem. □

13.4. **The Genus Expansion.** Let us now return to our setup for Wigner's theorem. Recall, in the combinatorial proof of Section 4, our approach was to compute the matrix moments $\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,\mathbf{X}_n^k$ for fixed k, and let $n\to\infty$ to recover the moments of the limit (in expectation) of the empirical eigenvalue distribution. We will carry out the same computation here (with $\mathbf{Z}_n$), but in this case we can compute everything exactly. To begin, as in Section 4, we have

$$\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,[\mathbf{Z}_n^k] = \frac{1}{n}\sum_{1\le j_1,j_2,\ldots,j_k\le n} \mathbb{E}(Z_{j_1j_2}Z_{j_2j_3}\cdots Z_{j_kj_1}).$$

Now, the variables Z_{jk} are all (complex) linear combinations of the independent $N(0,1)$ variables B_{jk} and B'_{jk}; thus, we may apply Wick's theorem to calculate the moments in this sum:

$$\mathbb{E}(Z_{j_1j_2}Z_{j_2j_3}\cdots Z_{j_kj_1}) = \sum_{\pi\in\mathscr{P}_2(k)} \prod_{\{a,b\}\in\pi} \mathbb{E}(Z_{j_aj_{a+1}}Z_{j_bj_{b+1}})$$

where we let $i_{k+1} = i_1$. But these covariances were calculated in Equation 13.3:

$$\mathbb{E}(Z_{j_aj_{a+1}}Z_{j_bj_{b+1}}) = \frac{1}{n}\delta_{j_aj_{b+1}}\delta_{j_{a+1}j_b}.$$

There are $k/2$ pairs in the product, and so combining the last three equations, we therefore have

$$\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,[\mathbf{Z}_n^k] = \frac{1}{n^{k/2+1}} \sum_{\pi\in\mathscr{P}_2(k)} \sum_{1\le j_1,\ldots,j_k\le n} \prod_{\{a,b\}\in\pi} \delta_{j_aj_{b+1}}\delta_{j_{a+1}j_b}.$$

To understand the internal product, it is useful here to think of π as a permutation as discussed above. Hence $\{a,b\}\in\pi$ means that $\pi(a) = b$, and since these are involotions, also $\pi(b) = a$. Thus, we have

$$\prod_{\{a,b\}\in\pi} \delta_{j_aj_{b+1}}\delta_{i_{a+1}i_b} = \prod_{a=1}^{k} \delta_{j_aj_{\pi(a)+1}}.$$

Let us introduce the shift permutation $\gamma \in S_k$ given by the cycle $\gamma = (1\ 2\ \cdots\ k)$ (that is, $\gamma(a) = a + 1 \bmod k$). Then we have

$$\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,[\mathbf{Z}_n^k] = \frac{1}{n^{k/2+1}} \sum_{\pi\in\mathscr{P}_2(k)} \sum_{1\le j_1,\ldots,j_k\le n} \prod_{a=1}^{k} \delta_{j_a j_{\gamma\pi(a)}}.$$

To simplify the internal sum, think of the indices $\{j_1, \ldots, j_k\}$ as a function $\mathbf{j}\colon \{1, \ldots, k\} \to \{1, \ldots, n\}$. In this form, we can succinctly describe the product of delta functions:

$$\prod_{a=1}^{k} \delta_{j_a j_{\gamma\pi(a)}} \neq 0 \quad \text{iff} \quad \mathbf{j} \text{ is constant on the cycles of } \gamma\pi\text{, in which case it } = 1.$$

Thus, we have

$$\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,[\mathbf{Z}_n^k] = \frac{1}{n^{k/2+1}} \sum_{\pi\in\mathscr{P}_2(k)} \#\left\{\mathbf{j}\colon [k] \to [n] \colon \mathbf{j} \text{ is constant on the cycles of } \gamma\pi\right\}$$

where, for brevity, we write $[k] = \{1, \ldots, k\}$ and $[n] = \{1, \ldots, n\}$. This count is trivial: we must simply choose one value for each of the cycles in $\gamma\pi$ (with repeats allowed). For any permutation $\sigma \in S_k$, let $\#(\sigma)$ denote the number of cycles in σ. Hence, we have

$$\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,[\mathbf{Z}_n^k] = \frac{1}{n^{k/2+1}} \sum_{\pi\in\mathscr{P}_2(k)} n^{\#(\gamma\pi)} = \sum_{\pi\in\mathscr{P}_2(k)} n^{\#(\gamma\pi)-k/2-1}. \tag{13.8}$$

Equation 13.8 gives an exact value for the matrix moments of a GUE_n. It is known as the **genus expansion**. To understand why, first note that $\mathscr{P}_2(k) = \varnothing$ if k is odd, and so we take $k = 2m$ even. Then we have

$$\frac{1}{n}\mathbb{E}[\mathbf{Z}_n^{2m}] = \sum_{\pi\in\mathscr{P}_2(2m)} n^{\#(\gamma\pi)-m-1}. \tag{13.9}$$

Now, let us explore the exponent of n in the terms in this formula. There is a beautiful geometric way to visualize $\#(\gamma\pi)$. Draw a regular $2m$-gon, and label its vertices in cyclic order $v_1, v_2, \ldots, v_{2m}$. Its edges may be identified as (cyclically)-adjacent pairs of vertices: $e_1 = v_1v_2$, $e_2 = v_2v_3$, until $e_{2m} = v_{2m}v_1$. A pairing $\pi \in \mathscr{P}_2(2m)$ can now be used to glue the edges of the $2m$-gon together to form a compact surface. Note: by convention, we always identify edges in "tail-to-head" orientation. For example, if $\pi(1) = 3$, we identify e_1 with e_3 by gluing v_1 to v_4 (and ergo v_2 to v_3). This convention forces the resultant compact surface to be orientable.

Lemma 13.5. *Let S_π be the compact surface obtained by gluing the edges of a $2m$-gon according to the pairing $\pi \in \mathscr{P}_2(2m)$ as described above; let G_π be the image of the $2m$-gon in S_π. Then the number of distinct vertices in G_π is equal to $\#(\gamma\pi)$.*

Proof. Since e_i is glued to $e_{\pi(i)}$, by the "tail-to-head" rule this means that v_i is identified with $v_{\pi(i)+1}$ (with addition modulo $2m$); that is, v_i is identified with $v_{\gamma\pi(i)}$ for each $i \in [2m]$. Now, edge $e_{\gamma\pi(i)}$ is glued to $e_{\pi\gamma\pi(i)}$, and by the same argument, the vertex $v_{\gamma\pi(i)}$ (now the tail of the edge in question) gets identified with $v_{\gamma\pi\gamma\pi(i)}$. Continuing this way, we see that v_i is identifies with precisely those v_j for which $j = (\gamma\pi)^\ell(i)$ for some $\ell \in \mathbb{N}$. Thus, the cycles of $\gamma\pi$ count the number of distinct vertices after the gluing. □

This connection is very fortuitous, because it allows us to neatly describe the exponent $\#(\gamma\pi) - m - 1$. Consider the surface $S = S_\pi$ described above. The *Euler characteristic* $\chi(S)$ of S is a well-defined even integer, which (miraculously) can be defined as follows: if G is any imbedded polygonal complex in S, then

$$\chi(S) = V(G) - E(G) + F(G)$$

where $V(G)$ is the number of vertices in G, $E(G)$ is the number of edges in G, and $F(G)$ is the number of faces of G. What's more, $\chi(S)$ is related to another topological invariant of S: its *genus*. Any orientable compact surface is homeomorphic to a g-holed torus for some $g \geq 0$ (the $g = 0$ case is the sphere); this $g = g(S)$ is the genus of the surface. It is a theorem (due to Cauchy – which is why we name it after Euler) (?) that

$$\chi(S) = 2 - 2g(S).$$

Now, consider our imbedded complex G_π in S_π. It is a quotient of a $2m$-gon, which has only 1 face, therefore $F(G_\pi) = 1$. Since we identify edges in pairs, $E(G_\pi) = (2m)/2 = m$. Thus, by the above lemma,

$$2 - 2g(S_\pi) = \chi(S_\pi) = \#(\gamma\pi) - m + 1.$$

Thus: $\#(\gamma\pi) - m - 1 = -2g(S_\pi)$. Returning to Equation 13.9, we therefore have

$$\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,[\mathbf{Z}_n^{2m}] = \sum_{\pi\in\mathscr{P}_2(2m)} n^{-2g(S_\pi)} = \sum_{g\geq 0} \varepsilon_g(m)\frac{1}{n^{2g}}, \tag{13.10}$$

where

$$\varepsilon_g(m) = \#\{\text{genus-}g \text{ surfaces obtained by gluing pairs of edges in a } 2m\text{-gon}\}.$$

Since the genus of any surface is ≥ 0, this shows in particular that

$$\frac{1}{n}\mathbb{E}\,\mathrm{Tr}\,[\mathbf{Z}_n^{2m}] = \varepsilon_0(m) + O\left(\frac{1}{n^2}\right).$$

Now, from our general proof of Wigner's theorem, we know that $\varepsilon_0(m)$ must be the Catalan number C_m. To see why this is true from the above beautiful geometric description, we need a little more notation for pairings.

Definition 13.6. *Let $\pi \in \mathscr{P}_2(2m)$. Say that π has a* **crossing** *if there are pairs $\{j,k\}$ and $\{j',k'\}$ in π with $j < j' < k < k'$. If π has no crossings, call is a* **non-crossing pairing**. *The set of non-crossing pairings is denoted $NC_2(2m)$.*

Unlike generic pairings $\mathscr{P}_2(2m)$, the structure of $NC_2(2m)$ depends on the *order* relation in the set $\{1,\ldots,2m\}$. It is sometimes convenient to write $NC_2(I)$ for an ordered set I, allowing for changes of indices; for example, we need to consider non-crossing partitions of the set $\{1,2,\ldots,i-1,i+2,\ldots,2m\}$, which are in natural bijection with $NC_2(2m-2)$.

Exercise 13.6.1. *Show that the set $NC_2(2m)$ can be defined recursively as follows: $\pi \in NC_2(2m)$ if and only if there is an* adjacent *pair $\{i,i+1\}$ in π, and the pairing $\pi \setminus \{i,i+1\}$ is a non-crossing pairing of $\{1,\ldots,i-1,i+2,\ldots,2m\}$.*

Proposition 13.7. *The genus of S_π is 0 if and only if $\pi \in NC_2(2m)$.*

Proof. First, suppose that π has a crossing. As Figure 4 below demonstrates, this means that the surface S_π has an imbedded double-ring that cannot be imbedded in S^2. Thus, S_π must have genus ≥ 1.

On the other hand, suppose that π is non-crossing. Then by Exercise 13.6.1, there is an interval $\{a,a+1\}$ (addition modulo $2m$) in π, and $\pi \setminus \{i,i+1\}$ is in $NC_2(\{1,\ldots,i-1,i+2,\ldots,2m\})$.

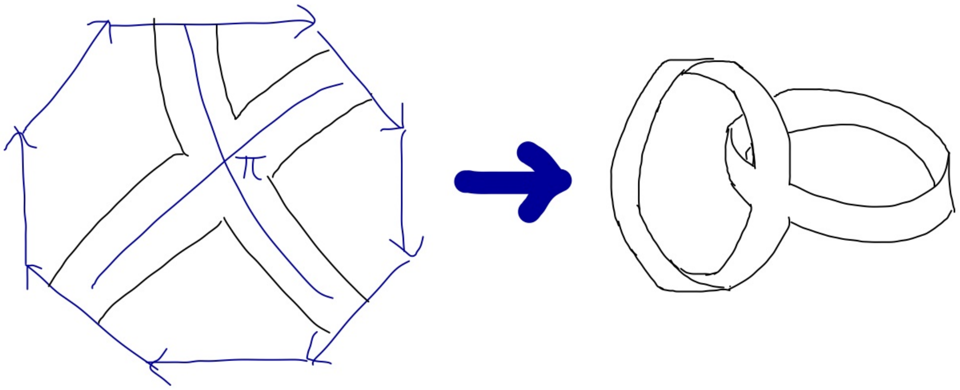


FIGURE 4. If π has a crossing, then the above surface (with boundary) is imbedded into S_π. Since this two-strip surface does not imbed in S^2, it follows the genus of S_π is ≥ 1.

As Figure 5 below shows, gluing along this interval pairing $\{i, i+1\}$ first, we reduce to the case of $\pi \setminus \{i, i+1\}$ gluing a $2(m-1)$-gon *within the plane*. Since $\pi \setminus \{i, i+1\}$ is also non-crossing, we proceed inductively until there are only two pairings left. It is easy two check that the two non-crossing pairings $\{\{1,2\},\{3,4\}\}$ and $\{\{1,4\},\{2,3\}\}$ of a square produce a sphere upon gluing. Thus, if π is non-crossing, $S_\pi = S^2$ which has genus 0. □

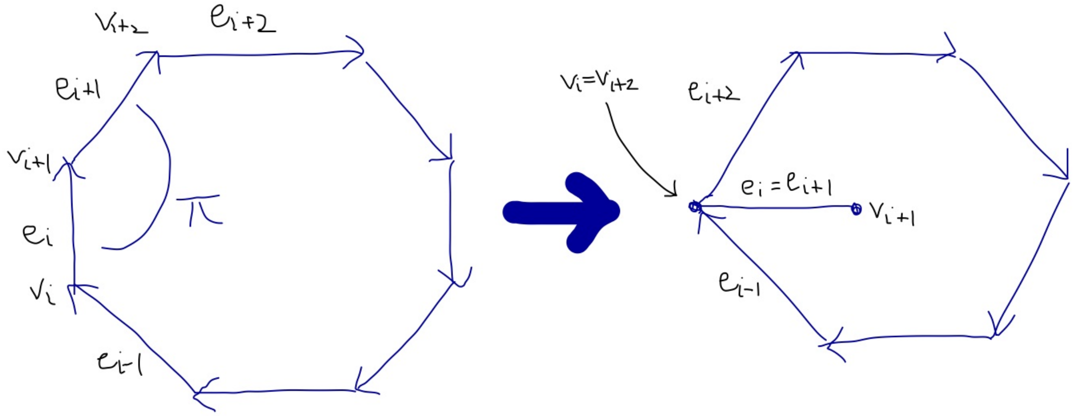


FIGURE 5. Given interval $\{i, i+1\} \in \pi$, the gluing can be done in the plane reducing the problem to size $2(m-1)$.

Thus, $\varepsilon_0(m)$ is equal to the number of non-crossing pairings $\#NC_2(2m)$. It is a standard result that non-crossing pairings are counted by Catalan numbers: for example, by decomposing $\pi \in NC_2(2m)$ by where 1 pairs, the non-crossing condition breaks π into two independent non-crossing partitions, producing the Catalan recurrence.

The genus expansion therefore provides yet another proof of Wigner's theorem (specifically for the GUE_n). Much more interesting are the higher-order terms in the expansion. There are no known formulas for $\varepsilon_g(m)$ when $g \geq 2$. The fact that these compact surfaces are "hiding inside" a GUE_n was (of course) discovered by Physicists, and there is much still to be understood in this direction. (For example, many other matrix integrals turn out to count interesting topological / combinatorial invariants.)

14. Joint Distribution of Eigenvalues of GOE_n and GUE_n

We have now seen that the joint law of entries of a GOE_n is given by

$$\mathbb{P}(\sqrt{n}\mathbf{X}_n \in B) = 2^{-n/2}\pi^{-n(n+1)/4}\int_B e^{-\frac{1}{2}\operatorname{Tr}(\mathbf{X}^2)}\,d\mathbf{X} \tag{14.1}$$

for all Borel subsets B of the space of $n \times n$ *symmetric* real matrices. Similarly, for a GUE_n matrix $\mathbf{Z}_n$,

$$\mathbb{P}(\sqrt{n}\mathbf{Z}_n \in B) = 2^{-n/2}\pi^{-n^2/2}\int_B e^{-\frac{1}{2}\operatorname{Tr}(\mathbf{Z}^2)}\,d\mathbf{Z} \tag{14.2}$$

for all Borel subsets B of the space of $n \times n$ *Hermitian* complex matrices.

Let us now diagonalize these matrices, starting with the GOE_n. By the spectral theorem, if $\mathbf{\Lambda}_n$ is the diagonal matrix with eigenvalues $\lambda_1(\mathbf{X}_n)$ through $\lambda_n(\mathbf{X}_n)$ (in increasing order, to avoid ambiguity), there is a (random) orthogonal matrix $\mathbf{Q}_n$ such that $\mathbf{X}_n = \mathbf{Q}_n^\top\mathbf{\Lambda}_n\mathbf{Q}_n$. Now, the matrix $\mathbf{Q}_n$ is *not unique*: each column (i.e. each eigenvector of $\mathbf{X}_n$) can still be scaled by ± 1 without affecting the result. If there are repeated eigenvalues, there is even more degeneracy. However, notice that the joint distribution of entries of $\mathbf{X}_n$ has a smooth density. Since the eigenvalues are continuous functions of the entries, it follows immediately that the eigenvalues of $\mathbf{X}_n$ are almost surely all distinct. (The event that two eigenvalues coincide has codimension ≥ 1 in the n-dimensional variety of eigenvalues; the existence of a density thus means this set is null.) So, if we avoid this null set, the scaling of the eigenvectors is the only undetermined quantity; we could do away with it by insisting that *the first non-zero entry of each column of* $\mathbf{Q}_n$ *is strictly positive*. Under this assumption, the map $(\mathbf{Q}, \mathbf{\Lambda}) \to \mathbf{Q}^\top\mathbf{\Lambda}\mathbf{Q}$ is a bijection (which is obviously smooth); so we can in principle compute the change of variables.

14.1. **Change of Variables on Lie Groups.** Let G be a Lie group. The Lie algebra $\mathfrak{g}$ of G is defined to be the tangent space at the identity. For the most relvant example, take $G = O(n)$, which is the set of matrices Q satisfying $Q^\top Q = I$. So $O(n)$ is a level set of the function $F(Q) = Q^\top Q$, which is clearly smooth. The tangent space is thus the kernel of dF_I, the differential at the identity. We calculate the differential as a linear map (acting on all matrices)

$$dF_I(T) = \lim_{t\to 0}\frac{1}{h}[F(I+hT) - F(I)] = \lim_{h\to 0}\frac{1}{h}[(I+hT)^\top(I+hT) - I] = T + T^\top.$$

Thus, the tangent space to $O(n)$ at the identity (which is denoted by $\mathfrak{o}(n)$) is equal to the space of matrices T with $T^\top + T = 0$; i.e. the *anti-symmetric matrices*. We denote this Lie algebra $\mathfrak{o}(n)$. It is a Lie algebra since it is closed under the bracket product $[T, S] = TS - ST$, as one can easily check; in general, the group operation on G translates into this bracket operation on $\mathfrak{g}$.

For smooth manifolds M and N, one defines smoothness of functions $S\colon M \to N$ in the usual local fashion. The derivative / differential of such a function at a point $\mathbf{x} \in M$ is a map $dS_\mathbf{x}\colon T_\mathbf{x}M \to T_{S(\mathbf{x})}N$ between the relevant tangent spaces. (In the special case that $N = \mathbb{R}$, dS is then a 1-form, the usual differential.) This map can be defined in local cordinates (in which case it is given by the usual derivative matrix). If $M = G$ happens to be a matrix Lie group, there is an easy global definition. If $T \in \mathfrak{g}$ is in the Lie algebra, then the matrix exponential e^T is in G. This exponential map can be used to describe the derivative based at the identity $I \in G$, as a linear map: for

$$dS_I(T) = \frac{d}{dt}S(e^{tT})\bigg|_{t=0}.$$

At $\mathbf{x} \in G$ not equal to the identity, we use the group invariance to extend this formula. In particular, it is an easy exercise that the tangent space $T_{\mathbf{x}}G$ is the translate of $T_I G = \mathfrak{g}$ by $\mathbf{x}$: $T_x G = \mathbf{x} \cdot \mathfrak{g}$. Hence, for $T \in T_{\mathbf{x}}G$, $\mathbf{x}^{-1}T \in \mathfrak{g}$ and so $e^{t\mathbf{x}^{-1}T} \in G$. In general, we have

$$dS_{\mathbf{x}}(T) = \frac{d}{dt} S(e^{t\mathbf{x}^{-1}T}\mathbf{x})\Big|_{t=0}. \tag{14.3}$$

We now move to the geometry of G. Let us specify an inner product on the Lie algebra $\mathfrak{g}$; in any matrix Lie group, the natural inner product is $\langle T_1, T_2 \rangle = \mathrm{Tr}\,(T_2^\top T_1)$. In particular, for $\mathfrak{g} = \mathfrak{o}(n)$ this becomes $\langle T_1, T_2 \rangle = -\mathrm{Tr}\,(T_1 T_2)$. We can then use the group structure to parallel translate this inner product everywhere on the group: that is, for two vectors $T_1, T_2 \in T_{\mathbf{x}}G$ for any $\mathbf{x} \in G$, we define $\langle T_1, T_2 \rangle_{\mathbf{x}} = \langle \mathbf{x}^{-1}T_1, \mathbf{x}^{-1}T_2 \rangle$. Thus, an inner product on $\mathfrak{g}$ extends to define a *Riemannian metric* on the manifold G (a smooth choice of an inner product at each point of G). This Riemannian metric is, by definition, left-invariant under the action of G.

In general, any Riemannian manifold M possesses a *volume measure* Vol_M determined by the metric. How it is defined is not particularly important to us. What is important is that on a Lie group G with its left-invariant Riemannian metric, the volume measure Vol_G is also left-invariant. That is: for any $\mathbf{x} \in G$, and any Borel subset $B \subseteq G$, $\mathrm{Vol}_G(\mathbf{x}B) = \mathrm{Vol}_G(B)$. But this pins down Vol_G, since any Lie group possesses a *unique* (up to scale) left-invariant measure, called the (left) *Haar* measure Haar_G. If G is compact, the Haar measure is a finite measure, and so it is natural to choose it to be a probability measure. This does not mean that Vol_G is a probability measure, however; all we can say is that $\mathrm{Vol}_G = \mathrm{Vol}_G(G) \cdot \mathrm{Haar}_G$.

Now, let M, N be Riemannian manifolds (of the same dimension), and let $S\colon M \to N$ be a smooth bijection. Then the change of variables (generalized from multivariate calculus) states that for any function $f \in L^1(N, \mathrm{Vol}_N)$,

$$\int_{S(M)} f(\mathbf{y})\,\mathrm{Vol}_N(d\mathbf{y}) = \int_M f(S(\mathbf{x}))|\det dS_{\mathbf{x}}|\,\mathrm{Vol}_M(d\mathbf{x}). \tag{14.4}$$

To be clear: $dS_{\mathbf{x}}\colon T_{\mathbf{x}}M \to T_{S(\mathbf{x})}N$ is the linear map described above. Since M and N are Riemannian manifolds, the tangent spaces $T_{\mathbf{x}}M$ and $T_{S(\mathbf{x})}N$ have inner-products defined on them, and so this determinant is well-defined: it is the volume of the image under $dS_{\mathbf{x}}$ of the unit box in $T_{\mathbf{x}}M$.

14.2. **Change of Variables for the Diagonalization Map.** Let $\mathrm{diag}(n)$ denote the linear space of diagonal $n \times n$ matrices, and let $\mathrm{diag}_<(n) \subset \mathrm{diag}(n)$ denote the open subset with strictle increasing diagonal entries $\lambda_1 < \cdots < \lambda_n$. As usual let $\mathscr{X}_n$ denote the linear space of symmetric $n \times n$ matrices. Now, consider the map

$$\begin{aligned} S\colon O(n) \times \mathrm{diag}_<(n) &\to \mathscr{X}_n \\ (\mathbf{Q}, \mathbf{\Lambda}) &\mapsto \mathbf{Q}^\top \mathbf{\Lambda} \mathbf{Q}. \end{aligned}$$

As noted above, this map is almost a bijection; it is surjective (by the spectral theorem) onto the open dense subset of symmetric matrices with distinct eigenvalues, but there is still freedom to choose a sign for each colum of $\mathbf{Q}$ without affecting the value of $O(n)$. Another way to say this is that *S descends to the quotient group $O_+(n) = O(n)/DO(n)$* where $DO(n)$ is the discrete group of *diagonal orthogonal matrices* (diagonal matrices with ±1s on the diagonal). This quotient map (also denoted S) is a bijection $S\colon O_+(n) \times \mathrm{diag}_<(n) \to \mathscr{X}_n$ onto the symmetric matrices with distinct eigenvalues. Quotienting by a discrete group does not change the Lie algebra; the only affect on our analysis is that $d\mathrm{Vol}_{O(n)} = 2^n d\mathrm{Vol}_{O_+(n)}$. The map S is clearly smooth, so we may apply the change of variables formula (Equation 14.5). Note that the measure $\mathrm{Vol}_{\mathscr{X}_n}$ is what we denoted by $d\mathbf{X}$ above. Hence, the change of variables formula (Equation 14.5) gives, for any $f \in L^1(\mathscr{X}_n)$

$$\int_{\mathscr{X}_n} f(\mathbf{X})\,d\mathbf{X} = 2^n \int_{O_+(n)\times \mathrm{diag}_<(n)} f(S(\mathbf{Q},\mathbf{\Lambda}))|\det dS_{\mathbf{Q},\mathbf{\Lambda}}|\,d\mathrm{Vol}_{O_+(n)\times\mathrm{diag}_<(n)}(\mathbf{Q},\mathbf{\Lambda}).$$

Of course, the volume measure of a product is the product of the volume measures. The volume measure $\mathrm{Vol}_{\mathrm{diag}_<(n)}$ on the open subset of the linear space $\mathrm{diag}(n)$ (with its usual inner-product) is just the Lebesgue measure $d\lambda_1\cdots d\lambda_n$ which we will denote $d\mathbf{\Lambda}$. The volume measure $\mathrm{Vol}_{O_+(n)}$ is the Haar measure (up to the scaling factor $\mathrm{Vol}(O_+(n))$). So we have (employing Fubini's theorem) for positive f

$$\int_{\mathscr{X}_n} f(\mathbf{X})\,d\mathbf{X} = 2^n\mathrm{Vol}(O_+(n)) \int_{O_+(n)}\int_{\mathrm{diag}_<(n)} f(S(\mathbf{Q},\mathbf{\Lambda}))|\det dS_{\mathbf{Q},\mathbf{\Lambda}}|\,d\mathbf{\Lambda}\,d\mathrm{Haar}_{O_+(n)}(\mathbf{Q}).$$

The factor $2^n\mathrm{Vol}(O_+(n))$ is equal to $\mathrm{Vol}(O(n))$. There is no harm, however, in integrating over the full group $O(n)$, since S is invariant under the action of $DO(n)$. This will scale the Haar measure by an additional 2^{-n}, so the result is

$$\int_{\mathscr{X}_n} f(\mathbf{X})\,d\mathbf{X} = c_n \int_{O(n)}\int_{\mathrm{diag}_<(n)} f(S(\mathbf{Q},\mathbf{\Lambda}))|\det dS_{\mathbf{Q},\mathbf{\Lambda}}|\,d\mathbf{\Lambda}\,d\mathrm{Haar}_{O(n)}(\mathbf{Q}) \tag{14.5}$$

where $c_n = 2^{-n}\mathrm{Vol}(O(n))$ is a constant we will largely ignore.

Let us now calculate the Jacobian determinant in Equation 14.5. Using Equation 14.3 for the $\mathbf{Q}$ direction (and the usual directional derivative for the linear $\mathbf{\Lambda}$ direction), we have for $T \in T_{\mathbf{Q}}O(n)$ and $D \in \mathrm{diag}(n)$

$$\begin{aligned} dS_{\mathbf{Q},\mathbf{\Lambda}}(T,D) = \frac{d}{dt}S(e^{t\mathbf{Q}^\top T}\mathbf{Q},\mathbf{\Lambda}+tD)\bigg|_{t=0} &= \frac{d}{dt}\left(e^{t\mathbf{Q}^\top T}\mathbf{Q}\right)^\top(\mathbf{\Lambda}+tD)e^{t\mathbf{Q}^\top T}\mathbf{Q}\bigg|_{t=0} \\ &= \mathbf{Q}^\top\,\frac{d}{dt}e^{t(\mathbf{Q}^\top T)^\top}(\mathbf{\Lambda}+tD)e^{t\mathbf{Q}^\top T}\bigg|_{t=0}\mathbf{Q}. \end{aligned}$$

Now, $T_{\mathbf{Q}}O(n) = \mathbf{Q}\cdot\mathfrak{o}(n)$; so for any $T \in T_{\mathbf{Q}}O(n)$, $\mathbf{Q}^\top T \in \mathfrak{o}(n)$. Hence, the matrix in the exponential is anti-symmetric, and we can rewrite the inside derivative as

$$\frac{d}{dt}e^{-t\mathbf{Q}^\top T}(\mathbf{\Lambda}+tD)e^{t\mathbf{Q}^\top T}\bigg|_{t=0} = -\mathbf{Q}^\top T\mathbf{\Lambda} + D + \mathbf{\Lambda}\mathbf{Q}^\top T = [\mathbf{\Lambda},\mathbf{Q}^\top T] + D.$$

Hence, we have

$$dS_{\mathbf{Q},\mathbf{\Lambda}}(T,D) = \mathbf{Q}^\top\left([\mathbf{\Lambda},\mathbf{Q}^\top T] + D\right)\mathbf{Q}. \tag{14.6}$$

This linear map is defined for $T \in T_{\mathbf{Q}}O(n) = \mathbf{Q}\cdot\mathfrak{o}(n)$. The form suggests that we make the following transformation: let $A_{\mathbf{Q},\mathbf{\Lambda}}\colon \mathfrak{o}(n)\times\mathrm{diag}(n) \to \mathscr{X}_n$ be the linear map

$$A_{\mathbf{Q},\mathbf{\Lambda}}(T,D) = \mathbf{Q}\cdot dS_{\mathbf{Q},\mathbf{\Lambda}}(\mathbf{Q}T,D)\cdot\mathbf{Q}^\top = [\mathbf{\Lambda},T] + D. \tag{14.7}$$

This transformation is isometric, and so in particular it preserves the determinant.

Lemma 14.1. $\det A_{\mathbf{Q},\mathbf{\Lambda}} = \det dS_{\mathbf{Q},\mathbf{\Lambda}}$.

Proof. We can write the relationship between $A_{\mathbf{Q},\mathbf{\Lambda}}$ and $dS_{\mathbf{Q},\mathbf{\Lambda}}$ as

$$A_{\mathbf{Q},\mathbf{\Lambda}} = \mathrm{Ad}_{\mathbf{Q}} \circ dS_{\mathbf{Q},\mathbf{\Lambda}} \circ (L_{\mathbf{Q}} \times \mathrm{id})$$

where $\mathrm{Ad}_{\mathbf{Q}}\mathbf{X} = \mathbf{Q}\mathbf{X}\mathbf{Q}^\top$ and $L_{\mathbf{Q}}T = \mathbf{Q}T$. The map $L_{\mathbf{Q}}\colon \mathfrak{o}(n) \to T_{\mathbf{Q}}O(n)$ is an isometry (by definition – the inner-product on $T_{\mathbf{Q}}O(n)$ is defined by translating the inner-product on $\mathfrak{o}(n)$ to $\mathbf{Q}$). The map $\mathrm{Ad}_{\mathbf{Q}}$ is an isometry of the common range space $\mathscr{X}_n$ (since it is equipped with the norm $\mathbf{X} \mapsto \sqrt{\mathrm{Tr}\,(\mathbf{X}^2)}$). Hence, the determinant is preserved. □

So, in order to complete the change of variables transformation of Equation 14.5,we need to calculate $\det A_{\mathbf{Q},\mathbf{\Lambda}}$. It turns out this is not very difficult. Let us fix an orthonormal basis for the domain $\mathfrak{o}(n) \times \mathrm{diag}(n)$. Let E_{ij} denote the matrix unit (with a 1 in the ij-entry and 0s elsewhere). For $1 \le i \le n$, denote $T_{ii} = (0, E_{ii}) \in \mathfrak{o}(n) \times \mathrm{diag}(n)$; and for $i < j$ let $T_{ij} = \frac{1}{\sqrt{2}}(E_{ij} - E_{ji}, 0) \in \mathfrak{o}(n)\times\mathrm{diag}(n)$. It is easy to verify that $\{T_{ij}\}_{1\le i\le j\le n}$ forms an orthonormal basis for $\mathfrak{o}(n)\times\mathrm{diag}(n)$. (The orthogonality between T_{ii} and T_{jk} for $j < k$ is automatic from the product structure; however, it is actually true that E_{ii} and $E_{jk} - E_{kj}$ are orthogonal in the trace inner-product; this reflects the fact that we could combine the product $\mathfrak{o}(n) \times \mathrm{diag}(n)$ into the set of $n \times n$ matrices with lower-triangular part the negative of the upper-triangular part, but arbitrary diagonal.)

Lemma 14.2. *The vectors* $\{A_{\mathbf{Q},\mathbf{\Lambda}}(T_{ij})\}_{1\le i\le j\le n}$ *form an orthogonal basis of* $\mathscr{X}_n$.

Proof. Let us begin with diagonal vectors T_{ii}. We have

$$A_{\mathbf{Q},\mathbf{\Lambda}}(T_{ii}) = [\mathbf{\Lambda}, 0] + E_{ii} = E_{ii}.$$

Since we have the general product formula $E_{ij}E_{k\ell} = \delta_{jk}E_{i\ell}$, distinct diagonal matrix units E_{ii} and E_{jj} actually have product 0, and hence are orthogonal. We also have the length of $A_{\mathbf{Q},\mathbf{\Lambda}}(T_{ii})$ is the length of E_{ii}, which is 1.

Now, for $i < j$, we have

$$A_{\mathbf{Q},\mathbf{\Lambda}}(T_{ij}) = \frac{1}{\sqrt{2}}[\mathbf{\Lambda}, E_{ij} - E_{ji}] + 0.$$

If we expand the diagonal matrix $\mathbf{\Lambda} = \sum_{a=1}^n \lambda_a E_{aa}$, we can evaluate this product

$$\begin{aligned}[\mathbf{\Lambda}, E_{ij} - E_{ji}] = \sum_{a=1}^n \lambda_a[E_{aa}, E_{ij} - E_{ji}] &= \sum_{a=1}^n \lambda_a[E_{aa}, E_{ij}] - \lambda_a[E_{aa}, E_{ji}] \\ &= \sum_{a=1}^n \lambda_a(E_{aa}E_{ij} - E_{ij}E_{aa}) - \lambda_a(E_{aa}E_{ji} - E_{ji}E_{aa}) \\ &= \sum_{a=1}^n \lambda_a(\delta_{ai}E_{aj} - \delta_{ja}E_{ia} - \delta_{aj}E_{ai} + \delta_{ia}E_{ja}) \\ &= \lambda_i E_{ij} - \lambda_j E_{ij} - \lambda_j E_{ji} + \lambda_i E_{ji} \\ &= (\lambda_i - \lambda_j)(E_{ij} + E_{ji}).\end{aligned}$$

Hence, $A_{\mathbf{Q},\mathbf{\Lambda}}(T_{ij}) = (\lambda_i - \lambda_j)\frac{1}{\sqrt{2}}(E_{ij} + E_{ji})$. It is again easy to calculate that, for distinct pairs (i, j) with $i < j$, these matrices are orthogonal in $\mathscr{X}_n$ (equipped with the trace inner-product). Similarly, T_{ii} is orthogonal from both $E_{ij} + E_{ji}$. This shows that all the vectors $\{A_{\mathbf{Q},\mathbf{\Lambda}}\}_{1\le i\le j\le n}$ are orthogonal. Since they number $n(n + 1)/2$ which is the dimension of $\mathscr{X}_n$, this completes the proof. □

Hence, $A_{\mathbf{Q},\mathbf{\Lambda}}$ preserves the right-angles of the unit cube, and so its determinant (the oriented volume of the image of the unit cube) is just the product of the lengths of the image vectors. As we saw above,

$$A_{\mathbf{Q},\mathbf{\Lambda}}(T_{ij}) = \begin{cases} E_{ii}, & i = j \\ (\lambda_i - \lambda_j)\frac{1}{\sqrt{2}}(E_{ij} + E_{ji}), & i < j \end{cases}.$$

The length of the diagonal images are all 1. The vectors $\frac{1}{\sqrt{2}}(E_{ij} + E_{ji})$ (with $i \neq j$) are normalized as well, and so the length of $A_{\mathbf{Q},\mathbf{\Lambda}}(T_{ij})$ is $|\lambda_i - \lambda_j|$. (Note: here we see that the map S is not even locally invertible at any matrix with a repeated eigenvalue; the differential dS has a non-trivial kernel at such a point.) So, employing Lemma 14.1, we have finally calculated

$$|\det dS_{\mathbf{Q},\mathbf{\Lambda}}| = |\det A_{\mathbf{Q},\mathbf{\Lambda}}| = \prod_{1 \le i < j \le n} |\lambda_i - \lambda_j|. \tag{14.8}$$

For a matrix $\mathbf{\Lambda} \in \mathrm{diag}(n)$ with $\mathbf{\Lambda}_{ii} = \lambda_i$, denote by $\Delta(\mathbf{\Lambda}) = \prod_{1 \le i < j \le n} (\lambda_j - \lambda_i)$. The quantity $\Delta(\mathbf{\Lambda})$ is called the **Vandermonde determinant** (for reasons we will highlight later). Equation 14.8 shows that the Jacobian determinant in the change of variables for the map S is $|\Delta(\mathbf{\Lambda})|$. In particular, the full change of variables formula (cf. Equation 14.5) is

$$\int_{\mathscr{X}_n} f(\mathbf{X})\, d\mathbf{X} = c_n \int_{O(n)} \int_{\mathrm{diag}_<(n)} f(\mathbf{Q}^\top \mathbf{\Lambda} \mathbf{Q}) |\Delta(\mathbf{\Lambda})|\, d\mathbf{\Lambda}\, d\mathrm{Haar}_{O(n)}(\mathbf{Q}). \tag{14.9}$$

14.3. **Joint Law of Eigenvalues and Eigenvectors for GOE_n and GUE_n.** Now, let B be any Borel subset of $\mathscr{X}_n$. Consider the positive function $f_B \in L^1(\mathbf{X}_n)$ given by

$$f_B(\mathbf{X}) = c_n' \mathbb{1}_B(\mathbf{X}) e^{-\frac{1}{2}\mathrm{Tr}\,(\mathbf{X}^2)}$$

where $c_n' = 2^{-n/2}\pi^{-n(n+1)/4}$. On the one hand, Equation 14.1 asserts that if $\mathbf{X}_n$ is a GOE_n, then

$$\mathbb{P}(\sqrt{n}\mathbf{X}_n \in B) = c_n' \int_B e^{-\frac{1}{2}\mathrm{Tr}\,(\mathbf{X}^2)}\, d\mathbf{X} = \int_{\mathscr{X}_n} f_B(\mathbf{X})\, d\mathbf{X}.$$

Now, applying the change of variables formula just developed, Equation 14.9, we can write

$$\int_{\mathscr{X}_n} f_B(\mathbf{X})\, d\mathbf{X} = c_n \int_{O(n)} \int_{\mathrm{diag}_<(n)} f_B(\mathbf{Q}^\top \mathbf{\Lambda} \mathbf{Q}) |\Delta(\mathbf{\Lambda})|\, d\mathbf{\Lambda}\, d\mathrm{Haar}_{O(n)}(\mathbf{Q}).$$

Since $\mathrm{Tr}\,[(\mathbf{Q}^\top \mathbf{\Lambda} \mathbf{Q})^2] = \mathrm{Tr}\,(\mathbf{\Lambda}^2)$, we therefore have

$$\mathbb{P}(\sqrt{n}\mathbf{X}_n \in B) = c_n c_n' \int_{O(n)} \int_{\mathrm{diag}_<(n)} \mathbb{1}\{\mathbf{Q}^\top \mathbf{\Lambda} \mathbf{Q} \in B\}\, e^{-\frac{1}{2}\mathrm{Tr}\,(\mathbf{\Lambda}^2)} |\Delta(\mathbf{\Lambda})|\, d\mathbf{\Lambda}\, d\mathrm{Haar}_{O(n)}(\mathbf{Q}).$$

Let us combine the two constants c_n, c_n' into a single constant which we will rename c_n. Now, in particular, consider a Borel set of the following form. Let $L \subseteq \mathrm{diag}_<(n)$ and $R \subseteq O(n)$ be Borel sets; set $B = R^\top L R$ (i.e. B is the set of all symmetric matrices of the form $Q^\top \Lambda Q$ for some $Q \in R$ and some $\Lambda \in L$). Then $\mathbb{1}\{\mathbf{Q}^\top \mathbf{\Lambda} \mathbf{Q} \in B\} = \mathbb{1}_L(\mathbf{\Lambda})\mathbb{1}_R(\mathbf{Q})$, and so the above formula separates

$$\mathbb{P}(\sqrt{n}\mathbf{X}_n \in R^\top L R) = c_n \int_R \int_L e^{-\frac{1}{2}\mathrm{Tr}\,(\mathbf{\Lambda}^2)} |\Delta(\mathbf{\Lambda})|\, d\mathbf{\Lambda}\, d\mathrm{Haar}_{O(n)}(\mathbf{Q}). \tag{14.10}$$

From this product formula, we have the following complete description of the joint law of eigenvalues and eigenvectors of a GOE_n.

Theorem 14.3. *Let $\mathbf{X}_n$ be a GOE_n, with (any) orthogonal diagonalization $\sqrt{n}\mathbf{X}_n = \mathbf{Q}_n^\top \mathbf{\Lambda}_n \mathbf{Q}_n$ where the entries of $\mathbf{\Lambda}_n$ increase from top to bottom. Then $\mathbf{Q}_n$ and $\mathbf{\Lambda}_n$ are independent. The distribution of $\mathbf{Q}_n$ is the Haar measure on the orthogonal group $O(n)$. The eigenvalue matrix $\mathbf{\Lambda}_n$, taking values (a.s.) in $\mathrm{diag}_<(n)$, has a law with probability density*

$$c_n e^{-\frac{1}{2}\,\mathrm{Tr}\,(\mathbf{\Lambda}^2)}|\Delta(\mathbf{\Lambda})| = c_n e^{-\frac{1}{2}(\lambda_1^2+\cdots+\lambda_n^2)} \prod_{1\le i<j\le n} (\lambda_j - \lambda_i)$$

for some constant c_n.

We can calculate the constant c_n in the theorem: tracking through the construction, we get

$$2^{-3n/2}\pi^{-n(n+1)/4}\mathrm{Vol}(O(n)),$$

but we then need to calculate the volume of $O(n)$. A better approach is to note that c_n is simply the normalization coefficient of the density of eigenvalues, and calculate it by evaluating the integral.

The situation for the GUE_n is quite similar. An analysis very analogous to the one above yields the appropriate change of variables formula. In this case, instead of quotienting by the discrete group $DO(n)$ of ± diagonal matrices, here we must mod out by the maximal torus of all diagonal matrices with complex diagonal entries of modulus 1. The analysis is the same, however; the only change (due basically to dimensional considerations) is that the Jacobian determinant becomes $\Delta(\mathbf{\Lambda})^2$. The result follows.

Theorem 14.4. *Let $\mathbf{Z}_n$ be a GUE_n, with (any) unitary diagonalization $\sqrt{n}\mathbf{Z}_n = \mathbf{U}_n^* \mathbf{\Lambda}_n \mathbf{U}_n$ where the entries of $\mathbf{\Lambda}_n$ increase from top to bottom. Then $\mathbf{U}_n$ and $\mathbf{\Lambda}_n$ are independent. The distribution of $\mathbf{U}_n$ is the Haar measure on the unitary group $U(n)$. The eigenvalue matrix $\mathbf{\Lambda}_n$, taking values (a.s.) in $\mathrm{diag}_<(n)$, has a law with probability density*

$$c_n e^{-\frac{1}{2}\,\mathrm{Tr}\,(\mathbf{\Lambda}^2)}\Delta(\mathbf{\Lambda})^2 = c_n e^{-\frac{1}{2}(\lambda_1^2+\cdots+\lambda_n^2)} \prod_{1\le i<j\le n} (\lambda_i - \lambda_j)^2$$

for some constant c_n.

Note: the constants (both called c_n) in Theorems 14.3 and 14.4 are not equal. In fact, it is the GUE_n case that has the much simpler normalization constant. In the next sections, we will evaluate the constant exactly. For the time being, it is convenient to factor out a $\sqrt{2\pi}$ from each variable (so the Gaussian part of the density at least is normalized). So, let us define the constant $C_n = (2\pi)^{-n/2}c_n$. The full statement is: for any Borel subset $L \subseteq \mathrm{diag}_<(n)$,

$$\mathbb{P}(\mathbf{\Lambda}_n \in L) = (2\pi)^{-n/2} C_n \int_L e^{-\frac{1}{2}(\lambda_1^2+\cdots+\lambda_n^2)} \prod_{1\le i<j\le n} (\lambda_j - \lambda_i)^2 \mathbb{1}_{\lambda_1\le\cdots\le\lambda_n}\, d\lambda_1\cdots d\lambda_n. \tag{14.11}$$

For many of the calculations that follow, it is convenient to drop the order condition on the eigenvalues. Let us define the **law of unordered eigenvalues** of a GUE_n: $\mathscr{P}_n$ is the probability measure defined on all of $\mathbb{R}^n$ by

$$\mathscr{P}_n(L) = \frac{(2\pi)^{-n/2}}{n!} C_n \int_L e^{-\frac{1}{2}(\lambda_1^2+\cdots+\lambda_n^2)} \Delta(\lambda_1,\ldots,\lambda_n)^2\, d\lambda_1\cdots d\lambda_n. \tag{14.12}$$

Here we are changing notation slightly and denoting $\Delta(\mathrm{diag}(\lambda_1,\ldots,\lambda_n)) = \Delta(\lambda_1,\ldots,\lambda_n)$. Of course, the order of the entries changes the value of Δ, but only by a sign; since it is squared in the density of $\mathscr{P}_n$, there is no ambiguity. It is actually for this reason that the normalization constant of the law of eigenvalues of a GUE_n is so much simpler than the constant for a GOE_n; in the latter case, if one extends to the law of unordered eigenvalues, the absolute-value signs must be added back onto the Vandermonde determinant term. For the GUE_n, on the other hand, the density is a *polynomial* times a Gaussian, and that results in much simpler calculations.

15. The Vandermonde Determinant, and Hermite Polynomials

The function $\Delta(\lambda_1,\ldots,\lambda_n) = \prod_{1\le i<j\le n}(\lambda_j-\lambda_i)$ plays a key role in the distribution of eigenvalues of Gaussian ensembles. As we mentioned in the previous section, Δ is called the *Vandermonde determinant*. The following proposition makes it clear why.

Proposition 15.1. *For any $\lambda_1,\ldots,\lambda_n\in\mathbb{R}$,*

$$\Delta(\lambda_1,\ldots,\lambda_n) = \prod_{1\le i<j\le n}(\lambda_j-\lambda_i) = \det\begin{bmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_n \\ \lambda_1^2 & \lambda_2^2 & \cdots & \lambda_n^2 \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_1^{n-1} & \lambda_2^{n-1} & \cdots & \lambda_n^{n-1}\end{bmatrix}.$$

Proof. Denote the matrix on the right-hand-side as V, with columns $V = [V_1,\ldots,V_n]$. The jth column is a vector-valued polynomial of λ_j, $V_j = \mathbf{p}(\lambda_j)$, where $\mathbf{p}$ is the same polynomial for each column: $\mathbf{p}(\lambda) = [1,\lambda,\ldots,\lambda^{n-1}]^\top$. Since the determinant of a matrix is a polynomial in its entries, it follows that $\det V$ is a polynomial in $\lambda_1,\ldots,\lambda_n$. Since the same polynomial $\mathbf{p}$ is used for each column, we see that if any $\lambda_i=\lambda_j$, V has a repeated column, and so its determinant vanishes. Hence, we can factor $\lambda_j-\lambda_i$ out of $\det V$ for each distinct pair i,j. It follows that Δ divides $\det V$.

Now, the degree of any term in the jth row is $j-1$. Using the expansion for the determinant

$$\det V = \sum_{\sigma\in S_n}(-1)^{|\sigma|}\prod_{i=1}^n V_{i\sigma(i)}, \tag{15.1}$$

the determinant is a sum of terms each of which is a product of n terms, one from each row of V; this shows that $\det V$ has degree $0+1+2+\cdots+n-1 = n(n-1)/2$. Since this is also the degree of Δ, we conclude that $\det V = c\Delta$ for some constant c.

To evaluate the constant c, we take the diagonal term in the determinant expansion (the term corresponding to $\sigma=\mathrm{id}$): this term is $1\cdot\lambda_2\cdot\lambda_3^2\cdots\lambda_n^{n-1}$. It is easy to check that the coefficient of this monomial in Δ is 1. Indeed, to get λ_n^{n-1} one must choose the λ_n term from every $(\lambda_n-\lambda_j)$ with $j<n$. This uses the $\lambda_n-\lambda_{n-1}$ term; the remaining terms involving λ_{n-1} are $(\lambda_{n-1}-\lambda_j)$ for $j<n-1$, and to ge tthe λ_{n-1}^{n-2} we must select the λ_{n-1} from each of these. Continuing this way, we see that, since each new factor of λ_i is chosen with a $+$ sign, and is uniquely acocunted for, the single term $\lambda_2\lambda_3^2\cdots\lambda_n^{n-1}$ has coefficient $+1$ in Δ. Hence $c=1$. □

The Vandermonde matrix is used in numerical analysis, for polynomial approximation; this is why it has the precise form given there. Reviewing the above proof, it is clear that this exact form is unnecessary; in order to follow the proof word-for-word, all that is requires is that there is a vector-valued polynomial $\mathbf{p} = [p_0,\ldots,p_{n-1}]^\top$ with $p_j(\lambda) = \lambda^j + O(\lambda^{j-1})$ such that the columns of V are given by $V_i = \mathbf{p}(\lambda_i)$. This fact is so important we record it as a separate corollary.

Corollary 15.2. *Let $p_0,\ldots,p_{n-1}$ be monic polynomials, with p_j of degree j. Set $\mathbf{p} = [p_0,\ldots,p_{n-1}]^\top$. Then for any $\lambda_1,\ldots,\lambda_n\in\mathbb{R}$,*

$$\det\begin{bmatrix}\mathbf{p}(\lambda_1) & \mathbf{p}(\lambda_2) & \cdots & \mathbf{p}(\lambda_n)\end{bmatrix} = \Delta(\lambda_1,\ldots,\lambda_n) = \prod_{1\le i<j\le n}(\lambda_j-\lambda_i).$$

In light of this incredible freedom to choose which polynomials to use in the *determinantal* interpretation of the quantity Δ that appears in the law of eigenvluaes 14.12, it will be convenient to use polynomials that are *orthogonal* with respect to the Gaussian measure that also appears in this law. These are the *Hermite polynomials*.

15.1. **Hermite polynomials.** For integers $n \geq 0$, we define

$$H_n(x) = (-1)^n e^{x^2/2} \frac{d^n}{dx^n} e^{-x^2/2}. \tag{15.2}$$

The functions H_n are, in fact, polynomials (easy to check by induction). They are called the **Hermite polynomials** (of unit variance), and they play a central role in Gaussian analysis. Here are the first few of them.

$$\begin{aligned}
H_0(x) &= 1 \\
H_1(x) &= x \\
H_2(x) &= x^2 - 1 \\
H_3(x) &= x^3 - 3x \\
H_4(x) &= x^4 - 6x^2 + 3 \\
H_5(x) &= x^5 - 10x^3 + 15x \\
H_6(x) &= x^6 - 15x^4 + 45x^2 - 15
\end{aligned}$$

Many of the important properties of the Hermite polynomials are evident from this list. By the definition of Equation 15.2, we have the following differential recursion:

$$\begin{aligned}
H_{n+1}(x) &= (-1)^{n+1} e^{x^2/2} \left(\frac{d^{n+1}}{dx^{n+1}} e^{-x^2/2} \right) \\
&= -e^{x^2/2} \frac{d}{dx} \left((-1)^n \frac{d^n}{dx^n} e^{-x^2/2} \right) \\
&= -e^{x^2/2} \frac{d}{dx} \left(e^{-x^2/2} H_n(x) \right) = -e^{x^2/2} \left(-x e^{-x^2/2} H_n(x) + e^{-x^2/2} H_n'(x) \right).
\end{aligned}$$

That is to say:

$$H_{n+1}(x) = x H_n(x) - H_n'(x). \tag{15.3}$$

Proposition 15.3. *The Hermite polynomials H_n satisfy the following properties.*

(a) *$H_n(x)$ is a monic polynomial of degree n; it is an even function if n is even and an odd function if n is odd.*

(b) *H_n are* orthogonal *with respect to the Gaussian measure $\gamma(dx) = (2\pi)^{-1/2} e^{-x^2/2}\, dx$:*

$$\int_{\mathbb{R}} H_n(x) H_m(x)\, \gamma(dx) = \delta_{nm}\, n!$$

Proof. Part (a) follows from induction and Equation 15.3: knowing that $H_n(x) = x^n + O(x^{n-1})$, we have $H_n'(x) = O(x^{n-1})$, and so $H_{n+1}(x) = x(x^n + O(x^{n-1})) + O(x^{n-1}) = x^{n+1} + O(x^n)$. Moreover, since xH_n and H_n' each have opposite parity to H_n, the even/odd behavior follows similarly.

To prove (b), we integrate by parts in the definition of H_n (Equation 15.2):

$$\begin{aligned}(2\pi)^{-1/2}\int_{\mathbb{R}} H_n(x)H_m(x)e^{-x^2/2}\,dx &= (2\pi)^{-1/2}\int_{\mathbb{R}} H_n(x)(-1)^m\frac{d^m}{dx^m}e^{-x^2/2}\,dx \\ &= (2\pi)^{-1/2}\int_{\mathbb{R}}\left(\frac{d^m}{dx^m}H_n(x)\right)e^{-x^2/2}\,dx.\end{aligned}$$

If $m > n$, since H_n is a degree n polynomial, the mth derivative inside is 0. If $m = n$, by part (a) we have $\frac{d^n}{dx^n}H_n(x) = n!$, and since γ is a probability measure, the integral is $n!$. By repeating the argument reversing the roles of n and m, we see the integral is 0 when $n > m$ as well. □

Proposition 15.3(b) is the real reason the Hermite polynomials are import in Gaussian analysis: they are **orthogonal polynomials** for the measure γ. In fact, parts (a) and (b) together show that if one starts with the vectors $\{1, x, x^2, x^3, \ldots\}$, all in $L^2(\gamma)$, and performs Gram-Schmidt orthogonalization, the resulting orthogonal vectors are the Hermite polynomials $H_n(x)$. This is summarized in the second statement of the following corollary.

Corollary 15.4. *Let $\langle f, g\rangle_\gamma$ denote $\int fg\,d\gamma$. Then $\langle x, H_n(x)^2\rangle_\gamma = 0$. Also, if f is a polynomial of degree $< n$, then $\langle f, H_n\rangle_\gamma = 0$.*

Proof. Since H_n is either an odd function or an even function, H_n^2 is an even function. Thus $H_n(x)^2e^{-x^2/2}$ is an even function. Since x is odd, it follows that $\int xH_n(x)^2e^{-x^2/2}\,dx = 0$, proving the first claim. For the second, Proposition 15.3(a) shows that $H_n(x) = x^n + a_{n,n-1}x^{n-1} + \cdots + a_{n,1}x + a_{n,0}$ for some coefficients $a_{n,k}$ with $k < n$. (In fact, we know that $a_{n,n-1} = a_{n,n-3} = \cdots = 0$.) We can express this together in the linear equation

$$\begin{bmatrix} H_0 \\ H_1 \\ H_2 \\ \vdots \\ H_n \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots & 0 \\ a_{1,0} & 1 & 0 & 0 & \cdots & 0 \\ a_{2,0} & a_{2,1} & 1 & 0 & \cdots & 0 \\ \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ a_{n,0} & a_{n,1} & a_{n,2} & a_{n,3} & \cdots & 1 \end{bmatrix} \begin{bmatrix} 1 \\ x \\ x^2 \\ \vdots \\ x^n \end{bmatrix}.$$

The matrix above is lower-triangular with 1s on the diagonal, so it is invertible (in fact, it is *easily* invertible by iterated back substitution; it is an order n operation to invert it, rather than the usual order n^3). Since the inverse of a lower-triangular matrix is lower-triangular, this means that we can express x^k as a linear combination of $H_0, H_1, \ldots, H_k$ for each k. Hence, any polynomial f of degree $< n$ is a linear combination of $H_0, H_1, \ldots, H_{n-1}$, and so by Proposition 15.3(b) f is orthogonal to H_n, as claimed. □

Taking this Gram-Schmidt approach, let us consider the polynomial $xH_n(x)$. By Equation 15.3, we have $xH_n(x) = H_{n+1}(x) + H_n'(x)$. However, since it is a polynomial of degree $n+1$, we know it can be expressed as a linear combination of the polynomials $H_0, H_1, \ldots, H_{n+1}$ (without any explicit differentiation). In fact, the requisite linear combination is quite elegant.

Proposition 15.5. *For any $n \geq 1$,*

$$xH_n(x) = H_{n+1}(x) + nH_{n-1}(x). \tag{15.4}$$

Proof. From the proof of Corollary 15.4, we see that $\{H_0, \ldots, H_n\}$ is a basis for the space of polynomials of degree $\leq n$; what's more, it is an orthogonal basis with respect to the inner-product

$\langle\,\cdot\,,\,\cdot\,\rangle_\gamma$ on that space. Taking the usual expansion, then, we have

$$xH_n(x) = \sum_{k=0}^{n+1} \langle xH_n, \hat{H}_k\rangle_\gamma \hat{H}_k$$

where $\hat{H}_k$ is the normalized Hermite polynomial $\hat{H}_k = H_k/\langle H_k, H_k\rangle_\gamma^{1/2}$. Well, note that $\langle xH_n, H_k\rangle_\gamma = \langle H_n, xH_k\rangle_\gamma$. If $k < n-1$, then xH_k is a polynomial of degree $< n$, and by Corollary 15.4 it is orthogonal to H_n. So we immediately have the reduction

$$xH_n(x) = \langle xH_n, \hat{H}_{n-1}\rangle \hat{H}_{n-1} + \langle xH_n, \hat{H}_n\rangle \hat{H}_n + \langle xH_n, \hat{H}_{n+1}\rangle \hat{H}_{n+1}.$$

The middle term is the same as $\langle x, H_n^2\rangle_\gamma$ which is equal to 0 by Corollary 15.4. Since we know the normalizing constants from Proposition 15.3(b), we therefore have

$$xH_n(x) = \frac{1}{(n-1)!}\langle xH_n, H_{n-1}\rangle_\gamma H_{n-1}(x) + \frac{1}{(n+1)!}\langle xH_n, H_{n+1}\rangle_\gamma H_{n+1}(x).$$

For the final term, note that since $H_n(x) = x^n + O(x^{n-1})$, we have $xH_n(x) = x^{n+1} + O(x^n)$, and hence $\langle xH_n, H_{n+1}\rangle_\gamma = \langle x^{n+1}, H_{n+1}\rangle_\gamma$ as all lower-order terms are orthogonal to H_{n+1}. The same argument shows that $\langle H_{n+1}, H_{n+1}\rangle_\gamma = \langle x^{n+1}, H_{n+1}\rangle_\gamma = (n+1)!$; so the coefficient of H_{n+1} is 1. By a similar argument, we have

$$\langle xH_n, H_{n-1}\rangle_\gamma = \langle H_n, xH_{n-1}\rangle_\gamma = \langle H_n, x^n\rangle_\gamma = \langle H_n, H_n\rangle_\gamma = n!$$

and so the coefficient of H_{n-1} is $n!/(n-1)! = n$, proving the claim. □

Corollary 15.6. *For $n \geq 1$, $H_n'(x) = nH_{n-1}(x)$. Furthermore, $-H_n''(x) + xH_n'(x) = nH_n(x)$.*

Proof. Equation 15.3 asserts that $H_{n+1}(x) = xH_n(x) - H_n'(x)$, so $H_n'(x) = xH_n(x) - H_{n+1}(x)$. Using Proposition 15.5 yields the first result. For the second, differentiate Equation 15.3:

$$H_n''(x) = \frac{d}{dx}H_n'(x) = \frac{d}{dx}\left(xH_n(x) - H_{n+1}(x)\right) = H_n(x) + xH_n'(x) - H_{n+1}'(x).$$

Using the first identity $H_{n+1}'(x) = (n+1)H_n(x)$ and simplifying yields the result. □

Remark 15.7. Recall the operator $L = \frac{d^2}{dx^2} - x\frac{d}{dx}$, which comes from Gaussian integration by parts of the energy form

$$\int f'(x)g'(x)\,\gamma(dx) = \int Lf(x)\,g(x)\,\gamma(dx).$$

Corollary 15.6 asserts that $LH_n = -nH_n$ – that is, H_n is an *eigenvector* for L with eigenvalue $-n$. Viewing the statement this way affords an alternate proof: for any n, m

$$\langle LH_n, H_m\rangle_\gamma = -\langle H_n', H_m'\rangle_\gamma = -\langle nH_{n-1}, mH_{m-1}\rangle_\gamma = -nm\delta_{nm}(n-1)!.$$

We can therefore rewrite this as

$$\langle LH_n, H_m\rangle_\gamma = -n\delta_{nm}n! = -n\langle H_n, H_m\rangle_\gamma.$$

This is enough to conclude that $LH_n = -nH_n$ (at least in L^2-sense), so long as we know that the orthogonal vectors $\{H_m\}_{m\geq 0}$ form a basis for $L^2(\gamma)$. This is our next result.

Proposition 15.8. *The normalized vectors $\{\hat{H}_n\}_{n\geq 0}$ form an orthonormal basis for $L^2(\mathbb{R}, \gamma)$.*

Proof. As we have established, $\{\hat{H}_n\}_{0\le n\le m}$ form a basis for the space of polynomials of degree $\le m$; hence, as a vector space, the space P of all polynomials is spanned by $\{\hat{H}_n\}_{n\ge 0}$. Since these vectors are orthonormal, to conclude the proof we need only show that $P \subset L^2(\mathbb{R},\gamma)$ is dense. So, suppose that $f \in L^2(\mathbb{R},\gamma)$ is orthogonal to P. This means in particular that $\langle f, x^n\rangle_\gamma = 0$ for all $n \ge 0$. Well, consider the Fourier transform of the function $fe^{-x^2/2}$:

$$\phi(\xi) = \int f(x)e^{-x^2/2}e^{ix\xi}\,dx.$$

Expanding $e^{ix\xi}$ as a power-series, we would like to use Fubini's theorem to exchange integrals; then we would have

$$\phi(\xi) = \int f(x)e^{-x^2/2}\sum_{n=0}^\infty \frac{(i\xi)^n}{n!}x^n\,dx = \sum_{n=0}^\infty \frac{(i\xi)^n}{n!}\int x^n f(x)e^{-x^2/2}\,dx.$$

This last integral is $\sqrt{2\pi}\langle f, x^n\rangle_\gamma = 0$ by assumption, and so we have $\phi(\xi) = 0$ for all ξ; it follows that the original function $fe^{-x^2/2}$ is 0 in $L^2(\mathbb{R})$, meaning that f is 0 in $L^2(\gamma)$ as claimed.

It remains to justify the application of Fubini's theorem. We need to show that for each ξ the function $F(x,n) = \frac{(i\xi)^n}{n!}x^n f(x)e^{-x^2/2}$ is in $L^1(\mathbb{R}\times\mathbb{N})$ (where the measure on $\mathbb{N}$ is counting measure). To prove this, note that

$$\int_{\mathbb{R}}\sum_n |F(x,n)|\,dx = \int_{\mathbb{R}}\sum_n \frac{|x\xi|^n}{n!}|f(x)|e^{-x^2/2}\,dx = \int_{\mathbb{R}} |f(x)|e^{-\frac{1}{2}x^2+|\xi||x|}\,dx.$$

Write the integrand as

$$|f(x)|e^{-\frac{1}{4}x^2}\cdot e^{-\frac{1}{4}x^2+|\xi||x|}.$$

By the Cauchy-Schwartz inequality, then, we have

$$\int_{\mathbb{R}}\sum_n |F(x,n)|\,dx \le \left(\int (|f(x)|e^{-\frac{1}{4}x^2})^2\,dx\right)^{1/2}\left(\int_{\mathbb{R}} (e^{-\frac{1}{4}x^2+|\xi||x|})^2\,dx\right)^{1/2}.$$

The first factor is $\int f(x)^2e^{-x^2/2}\,dx = \sqrt{2\pi}\|f\|_\gamma^2 < \infty$ by assumption. The second factor is also finite (by an elementary change of variables, it can be calculated exactly). This concludes the proof. □

15.2. **The Vandermonde determinant and the Hermite kernel.** Since the Hermite polynomial H_n is monic and has degree n, we can use Corollary 15.2 to express the Vandermonde determinant as

$$\Delta(\lambda_1,\ldots,\lambda_n) = \det\begin{bmatrix} H_0(\lambda_1) & H_0(\lambda_2) & \cdots & H_0(\lambda_n) \\ H_1(\lambda_1) & H_1(\lambda_2) & \cdots & H_1(\lambda_n) \\ \vdots & \vdots & \ddots & \vdots \\ H_{n-1}(\lambda_1) & H_{n-1}(\lambda_2) & \cdots & H_{n-1}(\lambda_n)\end{bmatrix} = \det[H_{i-1}(\lambda_j)]_{i,j=1}^n.$$

Why would we want to do so? Well, following Equation 14.12, we can then write the law $\mathscr{P}_n$ of unordered eigenvalues of a(n unscaled) GUE_n as

$$\mathscr{P}_n(L) = \frac{(2\pi)^{-n/2}}{n!}C_n\int_L e^{-\frac{1}{2}(\lambda_1^2+\cdots+\lambda_n^2)}(\det[H_{i-1}(\lambda_j)]_{i,j=1}^n)^2\,d\lambda_1\cdots d\lambda_n. \tag{15.5}$$

Since the polynomials $H_{i-1}(\lambda_j)$ are orthogonal with respect to the density $e^{-\frac{1}{2}\lambda_j^2}$, we will find many cancellations to simplify the form of this law. First, let us reinterpret a little further.

Definition 15.9. *The* **Hermite functions** *(also known as* **harmonic oscillator wave functions***)* Ψ_n *are defined by*

$$\Psi_n(\lambda) = (2\pi)^{-1/4}e^{-\frac{1}{4}\lambda^2}\hat{H}_n(\lambda) = (2\pi)^{-1/4}(n!)^{-1/2}e^{-\frac{1}{4}\lambda^2}H_n(\lambda).$$

The orthogonality relations for the Hermite polynomials with respect to the Gaussian measure γ (Proposition 15.3(b)) translate into orthogonality relations for the Hermite functions with respect to Lebesgue measure. Indeed:

$$\int \Psi_n(\lambda)\Psi_m(\lambda)\,d\lambda = (2\pi)^{-1/2}\frac{1}{\sqrt{n!m!}}\int e^{-\frac{1}{2}\lambda^2}H_n(\lambda)H_m(\lambda)\,d\lambda = \frac{1}{\sqrt{n!m!}}\int H_n(\lambda)H_m(\lambda)\,\gamma(d\lambda)$$

and by the aforementioned proposition this equals δ_{nm}.

Regarding Equation 15.5, by bringing terms inside the square of the determinant, we can express the density succinctly in terms of the Hermite functions Ψ_n. Indeed, consider the determinant

$$\det[\Psi_{i-1}(\lambda_j)]_{i,j=1}^n = \det\left[(2\pi)^{-1/4}((i-1)!)^{-1/2}e^{-\frac{1}{4}\lambda_j^2}H_{i-1}(\lambda_j)\right]_{i,j=1}^n.$$

If we (Laplace) expand this determinant (cf. Equation 15.1), each term in the sum over the symmetric group contains exactly one element from each row and column; hence we can factor out a common factor as follows:

$$\det[\Psi_{i-1}(\lambda_j)]_{i,j=1}^n = (2\pi)^{-n/4}\prod_{i=1}^n((i-1)!)^{-1/2}e^{-\frac{1}{4}(\lambda_1^2+\cdots+\lambda_n^2)}\det[H_{i-1}(\lambda_j)]_{i,j=1}^n.$$

Squaring, we see that the density of the law $\mathscr{P}_n$ can be written as

$$\begin{aligned}&\frac{(2\pi)^{-n/2}}{n!}e^{-\frac{1}{2}(\lambda_1^2+\cdots+\lambda_n^2)}(\det[H_{i-1}(\lambda_j)]_{i,j=1}^n)^2\\ =&\frac{1!2!\cdots(n-1)!}{n!}\left(\det[\Psi_{i-1}(\lambda_j)]_{i,j=1}^n\right)^2\end{aligned}$$

Thus, the the law of unordered eigenvalues is

$$\mathscr{P}_n(L) = \frac{1!2!\cdots(n-1)!}{n!}C_n\int_L\left(\det[\Psi_{i-1}(\lambda_j)]_{i,j=1}^n\right)^2\,d\lambda_1\cdots d\lambda_n.$$

Now, for this squared determinant, let $V_{ij} = \Psi_{i-1}(\lambda_j)$. Then we have

$$(\det V)^2 = \det V^\top \det V = \det(V^\top V)$$

and the entries of $V^\top V$ are

$$[V^\top V]_{ij} = \sum_{k=1}^n V_{ki}V_{jk} = \sum_{k=1}^n \Psi_{k-1}(\lambda_j)\Psi_{k-1}(\lambda_j).$$

Definition 15.10. *The* n*th* **Hermite kernel** *is the function* $K_n\colon \mathbb{R}^2 \to \mathbb{R}$

$$K_n(x,y) = \sum_{k=0}^{n-1}\Psi_k(x)\Psi_k(y).$$

So, the density of the law $\mathscr{P}_n$ is (up to the constant c_n the determinant of the Hermite kernel:

$$\mathscr{P}_n(L) = \frac{1!2!\cdots(n-1)!}{n!}C_n \int_L \det[K_n(\lambda_i,\lambda_j)]_{i,j=1}^n \, d\lambda_1\cdots d\lambda_n. \tag{15.6}$$

15.3. **Determinants of reproducing kernels.** Equation 15.6 is extremely useful, becuase of the following property of the Hermite kernels K_n.

Lemma 15.11. *For any n, the kernel K_n is L^2 in each variable, and it is a* **reproducing kernel**

$$\int K_n(x,u)K_n(u,y)\,du = K_n(x,y).$$

Proof. By definition $K_n(x,y)$ is a polynomial in (x,y) times $e^{-\frac{1}{4}(x^2+y^2)}$, which is easily seen to be in L^p of each variable for all p. For the reproducing kernel identity,

$$\begin{aligned}\int K_n(x,u)K_n(u,y)\,du &= \int \sum_{k=0}^{n-1} \Psi_k(x)\Psi_k(u)\cdot \sum_{\ell=0}^{n-1}\Psi_\ell(u)\Psi_\ell(y)\,du \\ &= \sum_{1\le k,\ell<n} \Psi_k(x)\Psi_\ell(y)\int \Psi_k(u)\Psi_\ell(u)\,du \\ &= \sum_{1\le k,\ell<n} \Psi_k(x)\Psi_\ell(y)\delta_{k\ell} = \sum_{k=0}^{n-1}\Psi_k(x)\Psi_k(y) = K_n(x,y),\end{aligned}$$

where the third equality is the orthogonality relation for the Hermite functions. □

This reproducing property has many wonderful consequences. For our purposes, the following lemma is the most interesting one.

Lemma 15.12. *Let $K\colon \mathbb{R}^2 \to \mathbb{R}$ be a reproducing kernel, where the diagonal $x \mapsto K(x,x)$ is integrable, and let $d = \int K(x,x)\,dx$. Then*

$$\int \det[K(\lambda_i,\lambda_j)]_{i,j=1}^n \, d\lambda_n = (d-n+1)\det[K(\lambda_i,\lambda_j)]_{i,j=1}^{n-1}.$$

Proof. We use the Laplace expansion of the determinant.

$$\det[K(\lambda_i,\lambda_j)]_{i,j=1}^n = \sum_{\sigma\in S_n} (-1)^{|\sigma|}K(\lambda_1,\lambda_{\sigma(1)})\cdots K(\lambda_n,\lambda_{\sigma(n)}).$$

Integrating against λ_n, let us reorder the sum according to where σ maps n:

$$\int \det[K(\lambda_i,\lambda_j)]_{i,j=1}^n \, d\lambda_n = \sum_{k=1}^n \sum_{\substack{\sigma\in S_n\\ \sigma(n)=k}} (-1)^{|\sigma|}\int K(\lambda_1,\lambda_{\sigma(1)})\cdots K(\lambda_n,\lambda_{\sigma(n)})\,d\lambda_n.$$

When $k=n$, then the variable λ_n only appears (twice) in the final K, and we have

$$\sum_{\substack{\sigma\in S_n\\ \sigma(n)=n}} (-1)^{|\sigma|}K(\lambda_1,\lambda_{\sigma(1)})\cdots K(\lambda_{n-1},\lambda_{\sigma(n-1)})\int K(\lambda_n,\lambda_n)\,d\lambda_n.$$

The integral is, by definition, d; the remaining sum can be reindexed by permutations in S_{n-1} since any permutation in S_n that fixes n is just a permutation in S_{n-1} together with this fixed point;

moreover, removing the fixed point does not affect $|\sigma|$, and so we have (from the Laplace expansion again)

$$\sum_{\substack{\sigma\in S_n\\ \sigma(n)=n}} (-1)^{|\sigma|} K(\lambda_1,\lambda_{\sigma(1)})\cdots K(\lambda_{n-1},\lambda_{\sigma(n-1)}) \int K(\lambda_n,\lambda_n)\,d\lambda_n = d\cdot \det[K(\lambda_i,\lambda_j)]_{i,j=1}^{n-1}.$$

For the remaining terms, if $\sigma(n)=k<n$, then we also have for some $j<n$ that $\sigma(j)=n$. Hence there are two terms inside the integral:

$$\sum_{\substack{\sigma\in S_n\\ \sigma(n)=k}} (-1)^{|\sigma|} K(\lambda_1,\lambda_{\sigma(1)})\cdots K(\lambda_{j-1},\lambda_{\sigma(j-1)})K(\lambda_{j+1},\lambda_{\sigma(j+1)})\cdots K(\lambda_{n-1},\lambda_{\sigma(n-1)})$$
$$\times \int K(\lambda_j,\lambda_n)K(\lambda_n,\lambda_k)\,d\lambda_n.$$

By the reproducing kernel property, the integral is simply $K(\lambda_j,\lambda_k)$. The remaining terms are then indexed by a permutation which sends j to k (rather than n), and fixed n; that is, setting $\hat{\sigma}=(n,k)\cdot\sigma$,

$$\sum_{\substack{\sigma\in S_n\\ \sigma(n)=n}} (-1)^{|\hat{\sigma}|} K(\lambda_1,\lambda_{\hat{\sigma}(1)})\cdots K(\lambda_{n-1},\lambda_{\hat{\sigma}(n-1)}).$$

Clearly $(-1)^{|\hat{\sigma}|}=-(-1)^{|\sigma|}$, and so reversing the Laplace expansion again, each one of these terms (as k ranges from 1 to $n-1$) is equal to $-\det[K(\lambda_i,\lambda_j)]_{i,j=1}^{n-1}$. Adding up yields the desired statement. □

By induction on the lemma, we then have the L^1 norm of the nth determinant is

$$\int \det[K(\lambda_i,\lambda_j)]_{i,j=1}^{n}\,d\lambda_1\cdots d\lambda_n = (d-n+1)(d-n+2)\cdots d.$$

Now, taking $K=K_n$ to be the nth Hermite kernel, since the Hermite functions Ψ_k are normalized in L^2, we have

$$d=\int K_n(x,x)\,dx = \int \sum_{k=0}^{n-1}\Psi_k(x)^2\,dx = n.$$

Hence, we have

$$\int \det[K_n(\lambda_i,\lambda_j)]_{i,j=1}^{n}\,d\lambda_1\cdots d\lambda_n = n!$$

Thus, taking $L=\mathbb{R}^n$ in Equation 15.6 for the law $\mathscr{P}_n$, we can finally evaluate the constant: we have

$$1=\mathscr{P}_n(\mathbb{R}^n)=\frac{1!2!\cdots(n-1)!}{n!}C_n\int \det[K_n(\lambda_i,\lambda_j)]_{i,j=1}^{n}\,d\lambda_1\cdots d\lambda_n = \frac{1!2!\cdots(n-1)!}{n!}C_n\cdot n!$$

and so we conclude that $C_n=\frac{1}{1!2!\cdots(n-1)!}$. In other words, the full and final form of the law of unordered eigenvalues of a GUE_n is

$$\mathscr{P}_n(L)=\frac{1}{n!}\int_L \det[K_n(\lambda_i,\lambda_j)]_{i,j=1}^{n}\,d\lambda_1\cdots d\lambda_n. \tag{15.7}$$

15.4. **The averaged empirical eigenvalue distribution.** The determinantal structure and reproducing kernels of the last section give us a lot more than just an elegant way to evaluate the normalization coefficient for the joint law of eigenvalues. In fact, we can use Equation 15.7, together with Lemma 15.12 to explicitly evaluate the averaged empirical eigenvalue distribution of a GUE_n. Let μ_n denote the empirical eigenvalue distribution, and let $\bar{\mu}_n$ denote its average: that is, for test functions f on $\mathbb{R}$:

$$\int f\,d\bar{\mu}_n = \mathbb{E}\left(\int f\,d\mu_n\right) = \frac{1}{n}\sum_{j=1}^n \mathbb{E}(f(\tilde{\lambda}_j))$$

where $\tilde{\lambda}_j$ are the (random) eigenvalues of the GUE_n. The law $\mathscr{P}_n$ (or more precisely its j-marginal) allows us to calculate this expectation. To save notation, let us consider the case $j=1$. Then we have

$$\begin{aligned}\mathbb{E}(f(\tilde{\lambda}_1)) &= \int_{\mathbb{R}^n} f(\lambda_1)\,d\mathscr{P}_n(\lambda_1,\ldots,\lambda_n)\\ &= \frac{1}{n!}\int_{\mathbb{R}^n} f(\lambda_1)\det[K_n(\lambda_i,\lambda_j)]_{i,j=1}^n\,d\lambda_1\cdots d\lambda_n.\end{aligned}$$

Integrating in turn over each of the variables from λ_n down to λ_2, induction on Lemma 15.12 (noting that $d=n$ for the kernel $K=K_n$) yields

$$\frac{1}{n!}\int_{\mathbb{R}^n} f(\lambda_1)\det[K_n(\lambda_i,\lambda_j)]_{i,j=1}^n\,d\lambda_1\cdots d\lambda_n = \frac{(n-1)!}{n!}\int_{\mathbb{R}} f(\lambda_1)\det[K_n(\lambda_i,\lambda_j)]_{i,j=1}^1\,d\lambda_1.$$

This determinant is, of course, just the single entry $K_n(\lambda_1,\lambda_1)$. Thus, we have

$$\mathbb{E}(f(\lambda_1)) = \frac{1}{n}\int_{\mathbb{R}} f(\lambda)K_n(\lambda,\lambda)\,d\lambda.$$

We have thus identified the marginal distribution of the lowest eigenvalue! Almost. This is the marginal of the law of *unordered* eigenvalues, so what we have here is just the average of all eigenvalues. This is born out by doing the calculation again for each $\tilde{\lambda}_j$, and doing the induction over all variables other than j. Thus, on averaging, we have

$$\int f\,d\bar{\mu}_n = \frac{1}{n}\int_{\mathbb{R}} f(\lambda)K_n(\lambda,\lambda)\,d\lambda. \tag{15.8}$$

This shows that $\bar{\mu}_n$ has a density: the diagonal of the kernel $\frac{1}{n}K_n$. Actually, it pays to return to the Hermite polynomials rather than the Hermite functions here. Note that $K_n(\lambda,\lambda)=\sum_{k=0}^{n-1}\Psi_k(\lambda)^2$, and we have

$$\Psi_k(\lambda) = (2\pi)^{-1/4}e^{-\frac{1}{4}\lambda^2}\hat{H}_k(\lambda).$$

Thus, we can rewrite Equation 15.8 as

$$\int f\,d\bar{\mu}_n = \frac{1}{n}\int_{\mathbb{R}} f(\lambda)(2\pi)^{-1/2}\sum_{k=0}^{n-1}\hat{H}_k(\lambda)^2e^{-\lambda^2/2}\,d\lambda$$

and so $\bar{\mu}_n$ has a density with respect to the Gaussian measure. We state this as a theorem.

Theorem 15.13. *Let $\mathbf{Z}_n$ be a GUE_n, and let μ_n be the empirical eigenvalue distribution of $\sqrt{n}\mathbf{Z}_n$. Then its average $\bar{\mu}_n$ has the density*

$$d\bar{\mu}_n = \frac{1}{n}\sum_{k=0}^{n-1}\hat{H}_k(\lambda)^2\,\gamma(d\lambda)$$

with respect to the Gaussian measure $\gamma(d\lambda) = (2\pi)^{-1/2}e^{-\frac{1}{2}\lambda^2}\,d\lambda$*. Here* $\hat{H}_k$ *are the* $L^2(\gamma)$*-normalized Hermite polynomials.*

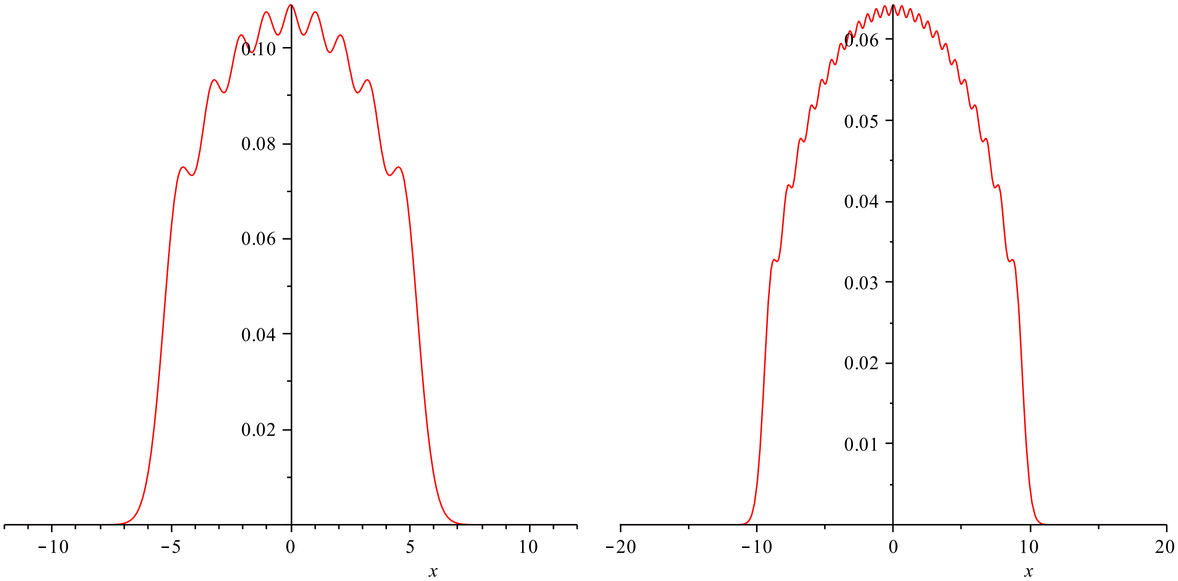


FIGURE 6. The averaged empirical eigenvalue distributions $\bar{\mu}_9$ and $\bar{\mu}_{25}$.

These measures have full support, but it is clear from the pictures that they are essentially 0 outside an interval very close to $[-2\sqrt{n}, 2\sqrt{n}]$. The empirical law of eigenvalues of $\mathbf{Z}_n$ (scaled) is, of course, the rescaling.

Exercise 15.13.1. *The density of* $\bar{\mu}_n$ *is* $\frac{1}{n}K_n(\lambda,\lambda)$*. Show that if* ν_n *is the empirical eigenvalue distribution of* $\mathbf{Z}_n$ *(scaled), then* $\bar{\nu}_n$ *has density* $n^{-1/2}K_n(n^{1/2}\lambda, n^{1/2}\lambda)$*.*

16. Fluctuations of the Largest Eigenvalue of GUE_n, and Hypercontractivity

We now have a completely explicit formula for the density of the averaged empirical eigenvalue distribution of a GUE_n (cf. Theorem 15.13). We will now use it to greatly improve our knowledge of the fluctuations of the largest eigenvalue in the GUE_n case. Recall we have shown that, in general (at least for finite-moment Wigner matrices), we know that the fluctiations of the largest eigenvalue around its mean 2 are at most $O(n^{-1/6})$ (cf. Theorem 6.2). The actual fluctuations are much smaller – they are $O(n^{-2/3})$. We will prove this for the GUE_n in two ways (one slightly sharper than the other).

16.1. L^p**-norms of Hermite polynomials.** The law $\bar{\mu}_n$ from Theorem 15.13 gives the explicit density of eigenvalues for the *unnormalized* GUE_n $\sqrt{n}\mathbf{Z}_n$. We are, of course, interested in the normalized $\mathbf{Z}_n$. One can write down the explicit density of eigenvalues of $\mathbf{Z}_n$ (cf. Exercise 15.13.1); this just amounts to the fact that, if we denote ν_n and $\bar{\nu}_n$ as the empirical eigenvalue distribution for $\mathbf{Z}_n$ (and its average), then for any test function f

$$\int_{\mathbb{R}} f\,d\bar{\nu}_n = \int_{\mathbb{R}} f\left(\frac{x}{\sqrt{n}}\right)\bar{\mu}(dx) = \int_{\mathbb{R}} f\left(\frac{x}{\sqrt{n}}\right)\frac{1}{n}\sum_{k=0}^{n-1}\hat{H}_k(x)^2\,\gamma(dx). \tag{16.1}$$

We are interested in the fluctuations of the largest eigenvalue $\lambda_n(\mathbf{Z}_n)$ around 2; that is:

$$\mathbb{P}(\lambda_n(\mathbf{Z}_n)\ge 2+\delta) = \mathbb{E}\left[\mathbb{1}_{[2+,\infty)}(\lambda_n(\mathbf{Z}_n))\right]$$

We can make the following fairly blunt estimate: of course

$$\mathbb{1}_{[2+,\infty)}(\lambda_n(\mathbf{Z}_n)) \le \sum_{j=1}^n \mathbb{1}_{[2+,\infty)}(\lambda_j(\mathbf{Z}_n))$$

and so

$$\begin{aligned}\mathbb{E}\left[\mathbb{1}_{[2+,\infty)}(\lambda_n(\mathbf{Z}_n))\right] \le \sum_{j=1}^n \mathbb{E}\left[\mathbb{1}_{[2+,\infty)}(\lambda_j(\mathbf{Z}_n))\right] &= \mathbb{E}\left[\int \mathbb{1}_{[2+,\infty)}(x)\sum_{j=1}^n \delta_{\lambda_j(\mathbf{Z}_n)}(dx)\right] \\ &= \mathbb{E}\left[n\int \mathbb{1}_{[2+,\infty)}\,d\nu_n\right] \\ &= n\int \mathbb{1}_{[2+,\infty)}\,d\bar{\nu}_n.\end{aligned}$$

Hence, using Equation 16.1, we can estimate

$$\begin{aligned}\mathbb{P}(\lambda_n(\mathbf{Z}_n)\ge 2+\delta) &\le n\int \mathbb{1}_{[2+,\infty)}\left(\frac{x}{\sqrt{n}}\right)\frac{1}{n}\sum_{k=0}^{n-1}\hat{H}_k(x)^2\,\gamma(dx) \\ &= \int_{(2+\delta)\sqrt{n}}^{\infty}\sum_{k=0}^{n-1}\hat{H}_k(x)^2\,\gamma(dx).\end{aligned}$$

Exchanging the (finite) sum and the integral, our estimate is

$$\mathbb{P}(\lambda_n(\mathbf{Z}_n)\ge 2+\delta) \le \sum_{k=0}^{n-1}\int_{(2+\delta)\sqrt{n}}^{\infty}\hat{H}_k^2\,d\gamma. \tag{16.2}$$

Remark 16.1. In fact, the blunt estimate we used above means that what we're really estimating here is the *sum* of the probabilities of *all* the eigenvalues being greater than $2+\delta$. Heuristically, we should expect this to be about n times as large as the desired probability; as we will see, this estimate does result in a slightly non-optimal bound (though a much better one than we proved in Theorem 6.2).

We must estimate the cut-off L^2-norm of the Hermite polynomials, therefore. One approach is to use Hölder's inequality: for any $p \ge 1$, with $\frac{1}{p}+\frac{1}{p'}=1$, we have

$$\int_{(2+\delta)\sqrt{n}}^{\infty} \hat{H}_k^2\, d\gamma = \int \mathbb{1}_{[(2+\delta)\sqrt{n},\infty)} \hat{H}_k^2\, d\gamma \le \left(\int (\mathbb{1}_{[(2+\delta)\sqrt{n},\infty)})^{p'}\, d\gamma\right)^{1/p'} \left(\int (\hat{H}_k^2)^p\, d\gamma\right)^{1/p}.$$

Since $\mathbb{1}_B^{p'} = \mathbb{1}_B$ for any set B, the first integral is just

$$\int (\mathbb{1}_{[(2+\delta)\sqrt{n},\infty)})^{p'}\, d\gamma = \gamma\Big([(2+\delta)\sqrt{n},\infty)\Big)$$

which we can estimate by the Gaussian density itself. Note, for $x \ge 1$, $e^{-x^2/2} \le xe^{-x^2/2}$, and so for $a \ge 1$

$$\int_a^\infty e^{-x^2/2}\, dx \le \int_a^\infty xe^{-x^2/2}\, dx = -e^{-x^2/2}\Big|_a^\infty = e^{-a^2/2}.$$

As such, since $a=(2+\delta)\sqrt{n} > 1$ for all n, we have

$$\gamma\Big([(2+\delta)\sqrt{n},\infty)\Big) = (2\pi)^{-1/2}\int_{(2+\delta)\sqrt{n}}^\infty e^{-x^2/2}\, dx \le (2\pi)^{-1/2} e^{-\frac{1}{2}(2+\delta)^2 n}.$$

We will drop the factor of $(2\pi)^{-1/2} \approx 0.399$ and make the equivalent upper-estimate of $e^{-\frac{1}{2}(2+\delta)^2 n}$; hence, Hölder's inequality gives us

$$\int_{(2+\delta)\sqrt{n}}^\infty \hat{H}_k^2\, d\gamma \le e^{-\frac{1}{2}(1-\frac{1}{p})(2+\delta)^2 n} \left(\int |\hat{H}_k|^{2p}\, d\gamma\right)^{1/p}. \tag{16.3}$$

This leaves us with the decidedly harder task of estimating the $L^{2p}(\gamma)$ norm of the (normalized) Hermite polynomials $\hat{H}_k$. If we take p to be an integer, we could conceivably expand the integral and evaluate it exactly as a sum and product of moments of the Gaussian. However, there is a powerful, important tool which we can use instead to give very sharp estimates with no work: this tool is called *hypercontractivity*.

16.2. **Hypercontractivity.** To begin this section, we remember a few constructs we have seen in Section 10. The *Ornstein-Uhlenbeck operator* $L = \frac{d^2}{dx^2} - x\frac{d}{dx}$ arises naturally in Gaussian integration by parts: for sufficiently smooth and integrable f, g,

$$\int f'g'\, d\gamma = -\int Lf \cdot g\, d\gamma.$$

The Ornstein-Uhlenbeck heat equation is the PDE $\partial_t u = Lu$. The unique solution with given initial condition $u_0 = f$ for some $f \in L^2(\gamma)$ is given by the Mehler kernel $P_t f$ (cf. Equation 10.3). In fact, the action of P_t on the Hermite polynomials is clear just from the equation: by Corollary 15.6, $LH_k = -kH_k$. Thus, if we define $u(t,x) = e^{-kt}H_k(x)$, we have

$$\frac{\partial}{\partial t}u(t,x) = -ke^{-kt}H_k(x); \quad Lu(t,x) = e^{-kt}LH_k(x) = -ke^{-kt}H_k(t,x).$$

These are equal, since since this $u(t,x)$ is continuous in t and $u(0,x) = H_k(x)$, we see that

$$P_t H_k = e^{-kt} H_k. \tag{16.4}$$

The semigroup P_t was introduced to aid in the proof of the *logarithmic Sobolev inequality*, which states that for f sufficiently smooth and non-negative in $L^2(\gamma)$,

$$\int f^2 \log f^2\,d\gamma - \int f^2\,d\gamma \cdot \log \int f^2\,d\gamma \le 2\int (f')^2\,d\gamma.$$

Before we proceed, let us state and prove an immediate corollary: the *L^q-log-Sobolev inequality*. We state it here in terms of a function that need not be ≥ 0.

Lemma 16.2. *Let $q > 2$, and let $u \in L^q(\gamma)$ be at least C^2. Then*

$$\int |u|^q \log |u|^q\,d\gamma - \int |u|^q\,d\gamma \cdot \log \int |u|^q\,d\gamma \le -\frac{q^2}{2(q-1)} \int (\mathrm{sgn}\ u)|u|^{q-1} Lu\,d\gamma$$

where $(\mathrm{sgn}\ u) = +1$ *if* $u \ge 0$ *and* $= -1$ *if* $u < 0$.

Proof. Let $f = |u|^{q/2}$. Then f sufficiently smooth (since $q/2 > 1$), non-negative, and in L^2. Thus, by the Gaussian log-Sobolev inequality, we have

$$\begin{aligned}\int |u|^q \log |u|^q\,d\gamma - \int |u|^q\,d\gamma \cdot \log \int |u|^q\,d\gamma &= \int f^2 \log f^2\,d\gamma - \int f^2\,d\gamma \cdot \log \int f^2\,d\gamma \\ &\le 2\int (f')^2\,d\gamma.\end{aligned}$$

Since $u \in C^2 \subset C^1$, $|u|$ is Lipschiptz and therefore is differentiable almost everywhere. On the set where it is differentiable, we have $f' = (|u|^{q/2})' = \frac{q}{2}|u|^{q/2-1}|u|'$. Hence

$$\int |u|^q \log |u|^q\,d\gamma - \int |u|^q\,d\gamma \cdot \log \int |u|^q\,d\gamma \le 2\int \left(\frac{q}{2}\right)^2 |u|^{q-2} \cdot (|u|')^2\,d\gamma.$$

Rewriting $|u|^{q-2}|u|' = \frac{1}{q-1}(|u|^{q-1})'$, the right-hand-side is

$$\frac{q^2}{2(q-1)} \int (|u|^{q-1})'|u|'\,d\gamma = -\frac{q^2}{2(q-1)} \int |u|^{q-1} L|u|\,d\gamma$$

where the euality follows from the definition of L (Gaussian integration by parts). Finally, on the set where $u \ge 0$ we have $|u| = u$ so $L|u| = Lu$; on the set where $u < 0$, $|u| = -u$ so $L|u| = -Lu$ there. All told, on the full-measure set where $|u|'$ is differentiable, $L|u| = (\mathrm{sgn}\ u)Lu$, completing the proof. □

In fact, the log-Sobolev inequality and its L^q-generalization are the “infinitesimal form” of a family of norm-estimates known as *hypercontractivity*. We already showed (in Section 10) that P_t is a contraction lf $L^2(\gamma)$ for all $t \ge 0$. In fact, P_t is very smoothing: it is actually a contraction from $L^2 \to L^q$ for sufficiently large t. We will prove a weak version of this theorem here that will suit our needs.

Theorem 16.3 (Hypercontractivity). *Let f be polynomial on $\mathbb{R}$, and let $q \ge 2$. Then*

$$\|P_t f\|_{L^q(\gamma)} \le \|f\|_{L^2(\gamma)} \quad \textit{for } \ t \ge \frac{1}{2}\log(q-1).$$

Remark 16.4. The full theorem has two parameters $1 \leq p \leq q < \infty$; the statement is that $\|P_t f\|_{L^q(\gamma)} \leq \|f\|_{L^p(\gamma)}$ for all $f \in L^p(\gamma)$, whenever $t \geq \frac{1}{2}\log\frac{q-1}{p-1}$. This theorem is due to E. Nelson, proved in this form in 1973. Nelson's proof did not involve the log-Sobolev inequality. Indeed, L. Gross *discovered* the log-Sobolev inequality as an equivalent form of hypercontractivity. The LSI has been discovered by at least two others around the same time (or slightly before); the reason Gross is (rightly) given much of the credit is his realization that it is equivalent to hypercontractivity, which is the source of much of its use in analysis and probability.

Proof. Let $t_c(q) = \frac{1}{2}\log(q-1)$ be the contraction time. It is enough to prove the theorem for $t = t_c(q)$, rather than for all $t \geq t_c(q)$. The reason is that P_t is a semigroup: that is $P_{t+s} = P_t P_s$, which follows directly from uniqueness of the solution of the OU-heat equation with a given initial condition. Thus, provided the theorem holds at $t = t_c(q)$, we have for any $t > t_c(q)$

$$\|P_t f\|_{L^q(\gamma)} = \|P_{t_c(q)} P_{t-t_c(q)} f\|_{L^q(\gamma)} \leq \|P_{t-t_c(q)} f\|_{L^2(\gamma)}.$$

By Corollary 10.8, this last quantity is $\leq \|f\|_{L^2(\gamma)}$ since $t - t_c(q) > 0$ and thus $P_{t-t_c(q)}$ is an $L^2(\gamma)$-contraction.

So, we must show that, for any polynomial f, $\|P_{t_c(q)} f\|_{L^q(\gamma)} \leq \|f\|_{L^2(\gamma)}$ for all $q \geq 2$. It is beneficial express this relationship in terms of varying t instead of varying q. So set $q_c(t) = 1 + e^{2t}$, so that $q_c(t_c(q)) = q$. Now set

$$\alpha(t) = \|P_t f\|_{L^{q_c(t)}(\gamma)}. \tag{16.5}$$

Since $t_c(0) = 2$, and therefore $\alpha(0) = \|P_0 f\|_{L^2(\gamma)} = \|f\|_{L^2(\gamma)}$, what we need to prove is that $\alpha(t) \leq \alpha(0)$ for $t \geq 0$. In fact, we will show that α is a non-increasing function of $t \geq 0$. We do so by finding its derivative. By definition

$$\alpha(t)^{q_c(t)} = \int |P_t f(x)|^{q_c(t)}\, \gamma(dx).$$

We would like to differentiate under the integral; to do so requires the use of the dominated convergence theorem (uniformly in t). First note that $P_t f$ is C^∞ since P_t has a smooth kernel; thus, since $q_c(t) \geq 2$, the integrand $|P_t f|^{q_c(t)}$ is at least C^2. Moreover, the following bound is useful: for any polynomial f of degree n, and for any $t_0 > 0$, there are constants $A, B, \alpha > 0$ such that for $0 \leq t \leq t_0$

$$\frac{\partial}{\partial t}|P_t f(x)|^{q_c(t)} \leq A|x|^{\alpha n} + B. \tag{16.6}$$

(The proof of Equation 16.6 is reserved to a lemma following this proof.) Since polynomials are in $L^1(\gamma)$, we have a uniform dominating function for the derivative on $(0, t_0)$, and hence it follows that

$$\frac{d}{dt}\alpha(t)^{q_c(t)} = \int \frac{\partial}{\partial t}|P_t f(x)|^{q_c(t)}\, \gamma(dx) \tag{16.7}$$

for $t < t_0$; since t_0 was chosen arbitrarily, this formula holds for all $t > 0$. We now calculate this partial derivative. For brevity, denote $u_t = P_t f$; note that $\partial_t u_t = L u_t$. By logarithmic differentiation

$$\begin{aligned}\frac{\partial}{\partial t}|u_t|^{q_c(t)} &= u_t^{q_c(t)}\left(q_c'(t)\log|u_t| + q_c(t)\frac{\partial}{\partial t}\log|u_t|\right)\\ &= q_c'(t)|u_t|^{q(t)}\log|u_t| + q_c(t)|u_t|^{q_c(t)}\cdot\frac{1}{u_t}\frac{\partial}{\partial t}u_t\\ &= q_c'(t)|u_t|^{q_c(t)}\log|u_t| + q_c(t)\frac{|u_t|^{q_c(t)}}{u_t}Lu_t.\end{aligned}$$

Combining this with Equation 16.7, we have

$$\frac{d}{dt}\alpha(t)^{q_c(t)} = q_c'(t)\int|u_t|^{q_c(t)}\log|u_t|\,d\gamma + q_c(t)\int(\text{sgn }u_t)|u_t|^{q_c(t)-1}Lu_t\,d\gamma \tag{16.8}$$

where $(\text{sgn }u_t) = 1$ if $u_t \ge 0$ and $= -1$ if $u_t < 0$. Now, we can calculate $\alpha'(t)$ using logarithmic differentiation on the outside: let $\beta(t) = \alpha(t)^{q_c(t)}$. We have shown that β is differentiable (and strictly positive); since $2 \le q_c(t) < \infty$, the function $\alpha(t) = \beta(t)^{1/q_c(t)}$ is also differentiable, and we have

$$\alpha'(t) = \frac{d}{dt}e^{\frac{1}{q_c(t)}\log\beta(t)} = \alpha(t)\left[-\frac{q_c'(t)}{q_c(t)^2}\log\beta(t) + \frac{1}{q_c(t)}\frac{\beta'(t)}{\beta(t)}\right].$$

Noting that $\alpha(t) = \beta(t)^{1/q_c(t)}$, multiplying both sides by $\alpha(t)^{q_c(t)-1}$ yields

$$(\alpha(t)^{q_c(t)-1})\cdot\alpha'(t) = -\frac{q_c'(t)}{q_c(t)^2}\beta(t)\log\beta(t) + \frac{1}{q_c(t)}\beta'(t). \tag{16.9}$$

By definition, $\beta(t) = \alpha(t)^{q_c(t)} = \int|u_t|^{q_c(t)}\,d\gamma$. Thus, combining Equations 16.8 and 16.9, we find that $(\alpha(t)^{q_c(t)-1})\cdot\alpha'(t)$ is equal to

$$-\frac{q_c'(t)}{q_c(t)^2}\int|u_t|^{q_c(t)}\,d\gamma\cdot\log\int|u_t|^{q_c(t)}\,d\gamma + \frac{q_c'(t)}{q_c(t)}\int|u_t|^{q_c(t)}\log|u_t|\,d\gamma + \int(\text{sgn }u_t)|u_t|^{q_c(t)-1}Lu_t\,d\gamma.$$

Writing the second term in terms of $\log|u_t|^{q_c(t)}$ (and so factoring out an additional $\frac{1}{q_c(t)}$), we can thence rewrite this to say that $(\alpha(t)^{q_c(t)-1})\cdot\alpha'(t)$ is equal to

$$\frac{q_c'(t)}{q_c(t)^2}\left[\int|u_t|^{q_c(t)}\log|u_t|^{q_c(t)}\,d\gamma - \int|u_t|^{q_c(t)}\,d\gamma\cdot\log\int|u_t|^{q_c(t)}\,d\gamma\right] + \int(\text{sgn }u_t)|u_t|^{q_c(t)-1}Lu_t\,d\gamma.$$

Finally, note that $q_c'(t) = \frac{d}{dt}(1 + e^{2t}) = 2e^{2t} = 2(q_c(t) - 1)$. The L^q-log-Sobolev inequality (cf. Lemma 16.2) asserts that this is ≤ 0. Since $\alpha(t)^{q_c(t)-1} > 0$, this proves that $\alpha'(t) \le 0$ for all $t > 0$, as desired. □

Lemma 16.5. *Let f be a polynomial of degree n, and let $t_0 > 0$. There are constants $A, B, \alpha > 0$ such that for $0 \le t \le t_0$*

$$\frac{\partial}{\partial t}|P_t f(x)|^{q_c(t)} \le A|x|^{\alpha n} + B. \tag{16.10}$$

Proof. Since f is a polynomial of degree n, we can expand it as a linear combination of Hermite polynomials $f = \sum_{k=0}^n a_k H_k$. From Equation 16.4, therefore

$$P_t f(x) = \sum_{k=0}^n a_k e^{-kt} H_k(x).$$

Therefore, we have

$$\beta(t,x) \equiv (P_t|f|(x))^{q_c(t)} = \left|\sum_{k=0}^{n} a_k e^{-kt} H_k(x)\right|^{q_c(t)}.$$

It is convenient to write this as

$$\beta(t,x) = \left[\left(\sum_{k=0}^{n} a_k e^{-kt} H_k(x)\right)^2\right]^{q_c(t)/2} = p(e^{-t},x)^{q_c(t)/2}.$$

The function $p(t,x)$ is a positive polynomial in both variables. Since the exponent $q_c(t)/2 = (1+e^{2t})/2 > 1$ for $t > 0$, this functon is differentiable for $t > 0$, and by logarithmic differentiation

$$\frac{\partial}{\partial t}\beta(t,x) = p(e^{-t},x)^{q_c(t)/2}\left[\frac{q_c'(t)}{2}\log p(e^{-t},x) + \frac{q_c(t)}{2}\frac{\partial_t p(e^{-t},x)}{p(e^{-t},x)}\right] \tag{16.11}$$

$$= \frac{q_c'(t)}{2} p(e^{-t},x)^{q_c(t)/2}\log p(e^{-t},x) - e^{-t}\frac{q_c(t)}{2} p(e^{-t},x)^{q_c(t)/2-1} p_{,1}(e^{-t},x), \tag{16.12}$$

where $p_{,1}$ denotes the partial derivative of $p(\,\cdot\,,x)$ in the first slot. Now, there are constants a_t and b_t so that $p(e^{-t},x) \le a_t x^n + b_t$, where the constants are continuous in t (indeed they can be taken as polynomials in e^{-t}). Let us consider the two terms in Equation 16.12 separately.

- The function $g_1(t,x) = p(e^{-t},x)^{q_c(t)/2}\log p(e^{-t},x) = \varphi_t(p(e^{-t},x))$ where $\varphi_t(x) = x^{q_c(t)/2}\log x$. Elementary calculus shows that $|\varphi_t(x)| \le \max\{1, x^{q_c(t)/2+1}\}$ for $x \ge 0$ (indeed, the cut-off can be taken as height $\frac{1}{eq}$ rather than the overestimate 1), and so
$$|g_1(t,x)| \le \max\left\{1, p(e^{-t},x)^{q_c(t)/2}\right\} \le \max\left\{1, (a_t x^n + b_t)^{q_c(t)/2+1}\right\}.$$
- The derivative $p_{,1}(e^{-t},x)$ satisfies a similar polynomial growth bound $|p_{,1}(e^{-t},x)| \le c_t x^n + d_t$ for constants c_t and d_t that are continuous in t. Hence
$$g_2(t,x) = p(e^{-t},x)^{q_c(t)/2-1} p_{,1}(e^{-t},x) \le (a_t x^n + b_t)^{q_c(t)/2-1}(c_t x^n + d_t).$$

Combining these with the terms in Equation 16.12, and noting that the coefficients $q_c'(t)/2$ and $e^{-t}q_c(t)/2$ are continuous in t, it follows that there are constants A_t and B_t, continuous in t, such that

$$\left|\frac{\partial}{\partial t}\beta(t,x)\right| \le A_t x^{(q_c(t)/2+1)n} + B_t.$$

Now, fix $t_0 > 0$. Because A_t and B_t are continuous in t, $A = \sup_{t\le t_0} A_t$ and $B = \sup_{t\le t_0} B_t$ are finite. Since $q_c(t) = 1 + e^{2t}$ is an increasing function of t, taking $\alpha = (q_c(t_0)/2 + 1)$ yields the desired inequality. □

Remark 16.6. The only place where the assumption that f is a polynomial came into play in the proof the the Hypercontractivity theorem is through Lemma 16.5 – we need it only to justify the differentiation inside the integral. One can avoid this by making appropriate approximations: first one can assume $f \ge 0$ since, in general, $|P_t f| \le P_t|f|$; in fact, by taking $|f| + \epsilon$ instead and letting $\epsilon \downarrow 0$ at the very end of the proof, it is sufficient to assume that $f \ge \epsilon$. In this case, the partial derivative computed above Equation 16.8 can be seen to be integrable for each t; then very careful limiting arguments (not using the dominated convergence theorem directly but instead using properties of the semigroup P_t to prove *uniform* integrability) show that one can differentiate inside the integral. The rest of the proof follows precisely as above (with no need to deal with the sgn term since all functions involved are strictly positive).

Alternatively, as we showed in Proposition 15.8, polynomials are dense in $L^2(\gamma)$; but this is not enough to extend the theorem as we proved. It is also necessary to show that one can choose a sequence of polynomials f_n converging to f in $L^2(\gamma)$ in such a way that $P_t f_n \to P_t f$ in $L^q(\gamma)$. This is true, but also requires quite a delicate argument.

16.3. **(Almost) optimal non-asymptotic bounds on the largest eigenvalue.** Returning now to Equations 16.2 and 16.3, we saw that the largest eigenvalue $\lambda_n(\mathbf{Z}_n)$ of a GUE_n satisfies the tail bound

$$\mathbb{P}(\lambda_n(\mathbf{Z}_n) \geq 2+\delta) \leq \sum_{k=0}^{n-1} e^{-\frac{1}{2}(1-\frac{1}{p})(2+\delta)^2 n} \left(\int |\hat{H}_k|^{2p}\, d\gamma \right)^{1/p}$$

for any $p \geq 1$. We will now use hypercontractivity to estimate this L^{2p}-norm. From Equation 16.4, $P_t \hat{H}_k = e^{-kt}\hat{H}_k$ for any $t \geq 0$. Thus

$$\int |\hat{H}_k|^{2p}\, d\gamma = \int |e^{kt} P_t \hat{H}_k|^{2p}\, d\gamma = e^{2pkt} \|P_t \hat{H}_k\|^{2p}_{L^{2p}(\gamma)}.$$

By the hypercontractivity theorem

$$\|P_t \hat{H}_k\|_{L^{2p}(\gamma)} \leq \|\hat{H}_k\|_{L^2(\gamma)} = 1$$

provided that $t \geq t_c(2p) = \frac{1}{2}\log(2p-1)$. Hence, we have

$$\mathbb{P}(\lambda_n(\mathbf{Z}_n) \geq 2+\delta) \leq e^{-\frac{1}{2}(1-\frac{1}{p})(2+\delta)^2 n} \sum_{k=0}^{n-1} e^{2kt} \quad \text{for } t \geq t_c(2p).$$

We now optimize over t and p. For fixed p, the right-hand-side increases (exponentially) with t, so we should take $t = t_c(2p)$ (the smallest t for which the estimate holds). So we have

$$\mathbb{P}(\lambda_n(\mathbf{Z}_n) \geq 2+\delta) \leq e^{-\frac{1}{2}(1-\frac{1}{p})(2+\delta)^2 n} \sum_{k=0}^{n-1} e^{2kt_c(2p)} = e^{-\frac{1}{2}(1-\frac{1}{p})(2+\delta)^2 n} \sum_{k=0}^{n-1} (2p-1)^k.$$

The geometric sum is $\frac{(2p-1)^n - 1}{2(p-1)} \leq \frac{1}{2(p-1)}(2p-1)^n$. Hence, we have the estimate

$$\mathbb{P}(\lambda_n(\mathbf{Z}_n) \geq 2+\delta) \leq \frac{1}{2(p-1)} e^{-\frac{1}{2}(1-\frac{1}{p})(2+\delta)^2 n + \log(2p-1)n} = \frac{1}{2(p-1)} e^{F_\delta(p)\cdot n} \tag{16.13}$$

where

$$F_\delta(p) = -\tfrac{1}{2}\left(1-\tfrac{1}{p}\right)(2+\delta)^2 + \log(2p-1).$$

We will optimize F_δ; this may not produce the optimal estimate for Equation 16.13, but it will give is a very tight bound as we will shortly see. Since $F_\delta(1) = 0$ while $\lim_{p\to\infty} F_\delta(p) = \infty$, the minimum value of the smooth function F_δ on $[1,\infty)$ is either 0 or a negative value achieved at a critical point. Critical points occur when

$$0 = F_\delta'(p) = -\frac{1}{2p^2}(2+\delta)^2 + \frac{2}{2p-1}$$

which is a quadratic equation with solutions

$$p_\pm(\delta) = \frac{1}{4}(2+\delta)\left[2+\delta \pm \sqrt{\delta^2+4\delta}\right].$$

One can easily check that $p_-(\delta) < 1$ for ¿0, and so we must choose $p_+(\delta)$. That is: we will use the estimate

$$\mathbb{P}(\lambda_n(\mathbf{Z}_n) \geq 2+\delta) \leq \frac{1}{2(p_+(\delta)-1)} e^{F_\delta(p_+(\delta))\cdot n}. \tag{16.14}$$

Lemma 16.7. *For all ¿0,*

$$\frac{1}{2(p_+(\delta)-1)} \leq \frac{1}{2}\delta^{-1/2}.$$

Proof. Take =ϵ^2, and do the Taylor expansion of the function p_+. The (Maple assisted) result is

$$p_+(\epsilon^2) = 1+\epsilon+\epsilon^2+O(\epsilon^3).$$

Hence $2(p_+(\delta)-1) \geq 2\delta^{1/2}$ for all sufficiently small δ; the statement for all δ is proven by calculating the derivative of $\epsilon \mapsto \epsilon^{-1}(p_+(\epsilon^2)-1)$ and showing that it is always positive. (This is a laborious, but elementary, task.) □

Lemma 16.8. *For all ¿0,*

$$F_\delta(p_+(\delta)) \leq -\frac{4}{3}\delta^{3/2}.$$

Proof. Again, do the Taylor expansion in the variable =ϵ^2. We have

$$F_{\epsilon^2}(p_+(\epsilon^2)) = -\frac{4}{3}\epsilon^3 - \frac{1}{10}\epsilon^5 + O(\epsilon^7).$$

This proves the result for all sufficiently small ϵ and thus sufficiently small δ; the statement for all δ can be proven by calculating the derivative of $\epsilon \mapsto \epsilon^{-3}F_{\epsilon^2}(p_+(\epsilon^2))$ and showing that it is always negative. (Again, this is laborious but elementary.) □

Combining Lemmas 16.7 and 16.8 with Equation 16.14, we finally arrive at the main theorem.

Theorem 16.9 (Ledoux's estimate)**.** *Let $\mathbf{Z}_n$ be a GUE_n, with largest eigenvalue $\lambda_n(\mathbf{Z}_n)$. Let ¿0. Then*

$$\mathbb{P}(\lambda_n(\mathbf{Z}_n) \geq 2+\delta) \leq \frac{1}{2}\delta^{-1/2}e^{-\frac{4}{3}\delta^{3/2}n}. \tag{16.15}$$

Theorem 16.9 shows that the order of fluctuations of the largest eigenvalue above its limit 2 is at most about $n^{-2/3}$. Indeed, fix $\alpha > 0$, and compute from Equation 16.15 that

$$\begin{aligned}\mathbb{P}(n^{2/3-\alpha}(\lambda_n(\mathbf{Z}_n)-2) \geq t) &= \mathbb{P}(\lambda_n \geq 2+n^{-2/3+\alpha}t) \\ &\leq \frac{1}{2}(n^{-2/3+\alpha}t)^{-1/2}e^{-\frac{4}{3}(n^{-2/3+\alpha}t)^{3/2}n} \\ &= \frac{1}{2}n^{2/3-\alpha/2}t^{-1/2}e^{-\frac{4}{3}(n^\alpha t)^{3/2}}\end{aligned}$$

and this tends to 0 as $n \to \infty$. With $\alpha = 0$, though, the right hands side is $\frac{1}{2}n^{2/3}t^{-1/2}e^{-\frac{4}{3}t^{3/2}}$, which blows up at a polynomial rate. (Note this is only an upper-bound.)

Theorem 16.9 thus dramatically improves the bound on the fluctuations of the largest eigenvalue above its mean from Theorem 6.2, where we showed the order of fluctuations is at most $n^{-1/6}$. Mind you: this result held for arbitrary Wigner matrices (with finite moments), and the present result is only for the GUE_n. It is widely conjectured that the $n^{-2/3}$ rate is universal as well; to date, this is only known for Wigner matrices whose entries have *symmetric* distributions (cf. the incredibly clever work of Soshnikov).

Exercise 16.9.1. *With the above estimates, show that in fact the fluctuations of $\lambda_n(\mathbf{Z}_n)$ above 2 are less than order $n^{-2/3}\log(n)^{2/3+\alpha}$ for any $\alpha > 0$.*

16.4. **The Harer-Zagier recursion, and optimal tail bounds.** Aside from the polynomial factor $\delta^{-1/2}$ in Equation 16.15, Theorem 16.9 provides the optimal rate of fluctuations of the largest eigenvalue of a GUE_n. To get rid of this polynomial factor, the best known technique is to use the moment method we used in Section 6. In this case, using the explicit density for $\bar{\mu}_n$, we can calculate the moments exactly. Indeed, they satisfy a recursion relation which was discovered by Harer and Zagier in 1986. It is best stated in terms of moments renormalized by Catalan numbers (as one might expect from the limiting distribution). To derive this recursion, we first develop an explicit formula for the (exponential) moment generating function of the density of eigenvalues.

Proposition 16.10. *Let $\bar{\nu}_n$ be the averaged empirical eigenvalue distribution of a (normalized) GUE_N. Then for any $z \in \mathbb{C}$,*

$$\int_{\mathbb{R}} e^{zt}\,\bar{\nu}_n(dt) = e^{z^2/2n}\sum_{k=0}^{n-1}\frac{C_k}{(2k)!}\frac{(n-1)\cdots(n-k)}{n^k}\,z^{2k}$$

where $C_k = \frac{1}{k+1}\binom{2k}{k}$ is the Catalan number.

Remark 16.11. Note that this proposition gives yet one more proof of Wigner's semicircle law for the GUE_n. As $n \to \infty$, the above function clearly converges (uniformly) to the power series $\sum_{k=0}^{\infty}\frac{C_k}{(2k)!}z^{2k}$, which is the (exponential) moment generating function of Wigner's law.

Proof. From Equation 15.8, we have

$$\int_{\mathbb{R}} e^{zt}\,\bar{\nu}_n(dt) = \int_{\mathbb{R}} e^{zt/\sqrt{n}}\,\bar{\mu}_n(dt) = \frac{1}{n}\int_{\mathbb{R}} e^{zt/\sqrt{n}}K_n(t,t)\,dt \tag{16.16}$$

where $K_n(t,t) = \sum_{k=0}^{n-1}\Psi_k(t)^2$ with $\Psi_k(t) = (2\pi)^{-1/4}e^{-\frac{1}{4}t^2}\hat{H}_k(t)$ the Hermite functions. In fact, one can write this sum only in terms of Ψ_n and Ψ_{n-1} and their derivatives, because of the following lemma.

Lemma 16.12. *For any $n \ge 1$, and any $x \ne y$,*

$$\sum_{k=0}^{n-1}\hat{H}_k(x)\hat{H}_k(y) = \sqrt{n}\frac{\hat{H}_n(x)\hat{H}_{n-1}(y) - \hat{H}_{n-1}(x)\hat{H}_n(y)}{x-y}. \tag{16.17}$$

The proof of the lemma can bee seen by integrating either side of the desired equality against the density $(x-y)e^{-x^2/2-y^2/2}$; using the orthogonality relations of the Hermite polynomials, the two integrals can be easily shown to be equal for any n. It then follows that the functions themselves are equal, since $\hat{H}_k$ form a basis for polynomials (so one can recover the polynomial functions from their inner-products with Gaussians; the corresponding integrals are sums of products of such inner-products).

Multiplying both sides of Equation 16.17 by $(2\pi)^{-1/2}e^{-\frac{1}{4}(x^2+y^2)}$, we see that the same relation holds with Ψ_k in place of $\hat{H}_k$. Hence, taking the limit as $y \to x$, one then has the following formula for the diagonal:

$$\frac{1}{\sqrt{n}}K_n(x,x) = \lim_{y\to x}\frac{\Psi_n(x)\Psi_{n-1}(y) - \Psi_{n-1}(x)\Psi_n(y)}{x-y} = \Psi_n'(x)\Psi_{n-1}(x) - \Psi_{n-1}'(x)\Psi_n(x). \tag{16.18}$$

Differentiating, this gives

$$\frac{d}{dx}\frac{1}{\sqrt{n}}K_n(x,x) = \Psi_n''(x)\Psi_{n-1}(x) - \Psi_{n-1}''(x)\Psi_n(x) \tag{16.19}$$

(because the cross-terms cancel). Now, the relation $L\hat{H}_n = -n\hat{H}_n$ (where $L = \frac{d^2}{dx^2} - x\frac{d}{dx}$) translates to a related second-order differential equation satisfied by Ψ_n; the reader can check that

$$\Psi_n''(x) = \left(-n - \frac{1}{2} + \frac{x^2}{4}\right)\Psi_n(x).$$

Combining with Equation 16.19, we see that the cross terms again cancel and all we are left with is

$$\frac{d}{dx}\frac{1}{\sqrt{n}}K_n(x,x) = -\Psi_n(x)\Psi_{n-1}(x). \tag{16.20}$$

Returning now to Equation 16.16, integrating by parts we then have

$$\int_{\mathbb{R}} e^{zt}\bar{\nu}_n(dt) = \frac{1}{n}\int_{\mathbb{R}} e^{zt/\sqrt{n}}K_n(t,t)\,dt = \frac{1}{z}\int_{\mathbb{R}} e^{zt/\sqrt{n}}\Psi_n(t)\Psi_{n-1}(t)\,dt. \tag{16.21}$$

To evaluate this integral, we use the handy relation $H_n' = (n-1)H_{n-1}$. As a result, the Taylor expansion of $H_n(t+\xi)$ (in ξ) is the usual Binomial expansion:

$$H_n(t+\xi) = \sum_{k=0}^{n}\binom{n}{k}H_{n-k}(t)\xi^k = \sum_{k=0}^{n}\binom{n}{k}H_k(t)\xi^{n-k}. \tag{16.22}$$

We use this in the moment generating function of the density $\Psi_n\Psi_{n-1}$ as follows: taking $\xi = z/\sqrt{n}$,

$$\begin{aligned}\int_{\mathbb{R}} e^{\xi t}\Psi_n(t)\Psi_{n-1}(t)\,dt &= (2\pi)^{-1/2}\int_{\mathbb{R}} e^{\xi t}e^{-\frac{1}{2}t^2}\hat{H}_n(t)\hat{H}_{n-1}(t)\,dt \\ &= (2\pi)^{-1/2}e^{\xi^2/2}\int_{\mathbb{R}} e^{-\frac{1}{2}(t-\xi)^2}\hat{H}_n(t)\hat{H}_{n-1}(t)\,dt \\ &= (2\pi)^{-1/2}e^{\xi^2/2}\int_{\mathbb{R}} e^{-\frac{1}{2}t^2}\hat{H}_n(t+\xi)\hat{H}_{n-1}(t+\xi)\,dt \\ &= \frac{1}{\sqrt{n!(n-1)!}}e^{\xi^2/2}\int_{\mathbb{R}} H_n(t+\xi)H_{n-1}(t+\xi)\,\gamma(dt)\end{aligned}$$

where, in making the substitution in the above integral, we assume that $\xi \in \mathbb{R}$ (the result for complex $z = \sqrt{n}\xi$ then follows by a standard analytic continuation argument). This has now become an integral against Gaussian measure $\gamma(dt) = (2\pi)^{-1/2}e^{-t^2/2}\,dt$. Substituting in the binomial expansion of Equation 16.22, and using the orthogonality relations of the H_k, we have

$$\begin{aligned}&\frac{\sqrt{n}}{n!}e^{\xi^2/2}\int_{\mathbb{R}}\sum_{k=0}^{n}\binom{n}{k}H_k(t)\xi^{n-k}\cdot\sum_{\ell=0}^{n-1}\binom{n-1}{\ell}H_\ell(t)\xi^{n-1-\ell}\,\gamma(dt) \\ = \;&\frac{\sqrt{n}}{n!}e^{\xi^2/2}\sum_{k=0}^{n}\sum_{\ell=0}^{n-1}\binom{n}{k}\binom{n-1}{\ell}\xi^{2n-k-\ell-1}\int_{\mathbb{R}} H_k(t)H_\ell(t)\,\gamma(dt) \\ = \;&\frac{\sqrt{n}}{n!}e^{\xi^2/2}\sum_{k=0}^{n-1}k!\binom{n}{k}\binom{n-1}{k}\xi^{2(n-k)-1}.\end{aligned}$$

Combining with Equation 16.21, substituting back $\xi = z/\sqrt{n}$, this gives

$$\int_{\mathbb{R}} e^{zt}\bar{\nu}_n(dt) = \frac{1}{z}\cdot e^{z^2/2n}\sum_{k=0}^{n-1}\sqrt{n}\frac{k!}{n!}\binom{n}{k}\binom{n-1}{k}\left(\frac{z}{\sqrt{n}}\right)^{2(n-k)-1}$$
$$= e^{z^2/2n}\sum_{k=0}^{n-1}\frac{k!}{n!}\binom{n}{k}\binom{n-1}{k}\left(\frac{z^2}{n}\right)^{n-k-1}.$$

Reindexing the sum $k\mapsto n-k-1$ and simplifying yields

$$e^{z^2/2n}\sum_{k=0}^{n-1}\frac{1}{(k+1)!}\binom{n-1}{k}\left(\frac{z^2}{n}\right)^{k}.$$

Since $\frac{C_k}{(2k)!} = \frac{1}{k!(k+1)!}$ and $\binom{n-1}{k} = \frac{(n-1)\cdots(n-k)}{k!}$, this proves the result.

□

Theorem 16.13 (Harer-Zagier recursion)**.** *The moment-generating function of the averaged empirical eigenvalue distribution $\bar{\nu}_n$ of a (normalized) GUE_n takes the form*

$$\int_{\mathbb{R}} e^{zt}\,\bar{\nu}_n(dt) = \sum_{k=0}^{\infty}\frac{C_k}{(2k)!}\,b_k^n\,z^{2k}$$

where $C_k = \frac{1}{k+1}\binom{2k}{k}$ is the Catalan number, and the coefficients b_k^n satisfy the recursion

$$b_{k+1}^n = b_k^n + \frac{k(k+1)}{4n^2}b_{k-1}^n.$$

The proof of Theorem 16.13 is a matter of expanding the Taylor series of $e^{z^2/2n}$ in Proposition 16.10 and comparing coefficients; the details are left to the generating-function-inclined reader. The relevant data we need from the recursion is the following.

Corollary 16.14. *The coefficients b_k^n in the Harer-Zagier recursion satisfy $b_k^n \le e^{k^3/2n^2}$.*

Proof. It is easy to check directly (from Proposition 16.10 and the definition of b_k^n in Theorem 16.13) that $b_1^n = 1$ while $b_1^n = \frac{1}{n}[\binom{n-1}{2}+1] > 0$. It follows from the recursion (which has positive coefficients) that $b_1^n > 0$ for all n. As such, the recursion again shows that $b_{k+1}^n \ge b_k^n$, and so we may estimate from the recursion that

$$b_{k+1}^n = b_k^n + \frac{k(k+1)}{4n^2}b_{k-1}^n \le \left(1+\frac{k(k+1)}{4n^2}\right)b_k^n \le \left(1+\frac{k^2}{2n^2}\right)b_k^n.$$

That is, if we define a_k^n by the recursion $a_{k+1}^n = (1+\frac{k^2}{2n^2})a_k^n$ with $a_0^n = b_0^n = 1$, then $b_k^n \le a_k^n$ for all k, n. By induction, we have

$$a_k^n = \left(1+\frac{1^2}{2n^2}\right)\left(1+\frac{2^2}{2n^2}\right)\cdots\left(1+\frac{(k-1)^2}{2n^2}\right) \le \left(1+\frac{k^2}{2n^2}\right)^k \le \left(e^{\frac{k^2}{2n^2}}\right)^k = e^{k^3/2n^2}.$$

□

Now, the definition of b_k^n in the exponential moment-generating function of $\bar{\nu}_n$ yields that the even moments $m_{2k}^n = \int t^{2k}\,d\bar{\nu}_n$ are given by $m_{2k}^n = C_k b_k^n$. Hence, by Corollary 16.14, the even moments are bounded by

$$m_{2k}^n \le C_k e^{k^3/2n^2}.$$

It is useful to have precise asymptotics for the Catalan numbers. Using Stirling's formula, one can check that

$$C_k \le \frac{4^k}{k^{3/2}\sqrt{\pi}}$$

where the ratio of the two sides tends to 1 as $k \to \infty$. Dropping the factor of $\pi^{-1/2} < 1$, we have

$$m_{2k}^n \le 4^k k^{-3/2} e^{k^3/2n^2}. \tag{16.23}$$

We can thence proceed with the same method of moments we used in Section 6. That is: for any positive integer k,

$$\mathbb{P}(\lambda_{\max} \ge 2+\epsilon) = \mathbb{P}(\lambda_{\max}^{2k} \ge (2+\epsilon)^{2k}) \le \mathbb{P}\left(\sum_{j=1}^n \lambda_j^{2k} \ge (2+\epsilon)^{2k}\right) \le \frac{1}{(2+\epsilon)^{2k}}\mathbb{E}\left(\sum_{j=1}^n \lambda_j^{2k}\right).$$

The expectation is precisely equal to (n times) the $2k$th moment of the averaged empirical eigenvalue distribution $\bar{\nu}_n$. Hence, utilizing Equation 16.23,

$$\mathbb{P}(\lambda_{\max} \ge 2+\epsilon) \le \frac{1}{(2+\epsilon)^{2k}}\, n \cdot m_{2k}^n \le \frac{n \cdot 4^k e^{k^3/2n^2}}{k^{3/2}(2+\epsilon)^{2k}} = \frac{n}{k^{3/2}}(1+\epsilon/2)^{-2k} e^{k^3/2n^2}.$$

Let us now restrict ϵ to be in $[0,1]$. Then we may make the estimate $1+\epsilon/2 \ge e^{\epsilon/3}$, and hence $(1+\epsilon/2)^{-2k} \le e^{-\frac{2}{3}k\epsilon}$. This gives us the estimate

$$\mathbb{P}(\lambda_{\max} \ge 2+\epsilon) \le \frac{n}{k^{3/2}} \exp\left\{\frac{k^3}{2n^2} - \frac{2}{3}k\epsilon\right\}. \tag{16.24}$$

We are free to choose k (to depend on n and ϵ). For a later-to-be-determined parameter $\alpha > 0$, set $k = \lfloor \alpha\sqrt{\epsilon}n \rfloor$. Then $\alpha\sqrt{\epsilon}n - 1 \le k \le \alpha\sqrt{\epsilon}n$, and so

$$\frac{k^3}{2n^2} \le \frac{\alpha^3\epsilon^{3/2}n^3}{2n^2} = \frac{\alpha^3}{2}\epsilon^{3/2}n$$

while

$$-\frac{2}{3}k\epsilon \le -\frac{2}{3}(\alpha\sqrt{\epsilon}n - 1)\epsilon = -\frac{2}{3}\alpha\epsilon^{3/2}n + \frac{2}{3}\epsilon$$

and

$$\frac{n}{k^{3/2}} \le \frac{n}{(\alpha\sqrt{\epsilon}n - 1)^{3/2}} = (\alpha\sqrt{\epsilon}n^{1/3} - n^{-2/3})^{-3/2}$$

where the last estimate only holds true if $\alpha\sqrt{\epsilon}n > 1$ (i.e. for sufficiently large n). For smaller n, we simply have the bone-headed estimate $\mathbb{P}(\lambda_{\max} \ge 2+\epsilon) \le 1$. Altogether, then, Equation 16.24 becomes

$$\begin{aligned}\mathbb{P}(\lambda_{\max} \ge 2+\epsilon) &\le (\alpha\sqrt{\epsilon}n^{1/3} - n^{-2/3})^{-3/2} \exp\left\{\frac{\alpha^3}{2}\epsilon^{3/2}n - \frac{2}{3}\alpha\epsilon^{3/2}n + \frac{2}{3}\epsilon\right\} \\ &\le 2(\alpha\sqrt{\epsilon}n^{1/3} - n^{-2/3})^{-3/2} \exp\left\{\left(\frac{1}{2}\alpha^3 - \frac{2}{3}\alpha\right)\epsilon^{3/2}n\right\}.\end{aligned}$$

In the final inequality, we used $e^{\frac{2}{3}\epsilon} \le e^{\frac{2}{3}} < 2$ since $\epsilon \le 1$. The pre-exponential factor is bounded for sufficiently large n, and so we simply want to choose $\alpha > 0$ so that the exponential rate is negative. The minimum occurs at $\alpha = \frac{2}{3}$; but for cleaner formulas we may simply pick $\alpha = 1$. Hence, we have proved that (for $0 < \epsilon \le 1$ and all sufficiently large n)

$$\mathbb{P}(\lambda_{\max} \ge 2+\epsilon) \le 2(\sqrt{\epsilon}n^{1/3} - n^{-2/3})^{-3/2} e^{-\frac{1}{6}\epsilon^{3/2}n}. \tag{16.25}$$

From here, we see that the fluctuations of $\lambda_{\max}$ above 2 are at most $O(n^{-2/3})$. Indeed, for any $t > 0$, letting $\epsilon = n^{-2/3}t$ (so take $t \le n^{2/3}$)

$$\begin{aligned}\mathbb{P}(n^{2/3}(\lambda_{\max} - 2) \ge t) &= \mathbb{P}(\lambda_{\max} \ge 2 + n^{-2/3}t) \\ &\le 2(n^{-1/3}\sqrt{t}n^{1/3} - n^{-2/3})^{-3/2}e^{-\frac{1}{6}n^{-1}t^{3/2}n} \\ &= 2(\sqrt{t} - n^{-2/3})^{-3/2}e^{-\frac{1}{6}t^{3/2}} \qquad (16.26)\end{aligned}$$

Recall that the estimate 16.25 held for $\sqrt{\epsilon}n > 1$; thus estimate 16.26 holds valid whenever $\sqrt{n^{-2/3}t}n > 1$: i.e. when $t > n^{-4/3}$ (as expected from the form of the inequality). For t of this order, the pre-exponential factor may be useless; in this case, we simply have the vacuous bound $\mathbb{P} \le 1$. But as long as $t \ge 4$, we have $\sqrt{t} - n^{-2/3} \ge \sqrt{t} - 1 \ge \frac{1}{2}\sqrt{t}$. Thus, for $t \ge 4$, $(\sqrt{t} - n^{-2/3})^{-3/2} \le 2^{3/2}t^{-3/4} \le 3t^{-3/4}$, and so we have the bound

$$\mathbb{P}(n^{2/3}(\lambda_{\max} - 2) \ge t) \le 6t^{-3/4}e^{-\frac{1}{6}t^{3/2}}, \quad 4 \le t \le n^{2/3}. \qquad (16.27)$$

It follows that if $\phi(n)$ is *any* function that tends to 0 as $n \to \infty$, then $n^{2/3}\phi(n)(\lambda_{\max} - 2)$ converges to 0 in probability; so we have a sharp bound for the rate of fluctuations. It turns out that inequality 16.27 not only describes the sharp rate of fluctuations of $\lambda_{\max}$, but also the precise tail bound. As the next (final) section outlines, the random variable $n^{2/3}(\lambda_n(\mathbf{Z}_n) - 2)$ has a limit distribution as $n \to \infty$, and this distribution has tail of order $t^{-a}e^{-ct^{3/2}}$ for some $a, c > 0$, as $t \to \infty$.

17. The Tracy-Widom Law and Level Spacing of Zeroes

Let $\mathbf{Z}_n$ be a (normalized) GUE_n. For the remainder of this section, let λ_n denote the largest eigenvalue $\lambda_n = \lambda_n(\mathbf{Z}_n)$. We have now seen (over and over) that $\lambda_n \to 2$ in probability as $n \to \infty$; moreover, the rate of fluctuations above 2 is $O(n^{-2/3})$. We now wish to study the random variable $n^{2/3}(\lambda_n - 2)$ as $n \to \infty$. Inequality 16.26 suggests that, if this random variable possesses a limiting distribution, it satisfies tail bounds of order $x^{-a}e^{-cx^{3/2}}$ for some $a, c > 0$. This is indeed the case. We state the complete theorem below in two parts. First, we need to introduce a new function.

Definition 17.1. *The* **Airy function** $\mathrm{Ai}\colon \mathbb{R} \to \mathbb{R}$ *is a solution* $u = \mathrm{Ai}$ *to the Airy ODE*

$$u''(x) = xu(x)$$

determined by the following asymptotics as $x \to \infty$*:*

$$\mathrm{Ai}(x) \sim \frac{1}{2}\pi^{-1/2}x^{-1/4}e^{-\frac{2}{3}x^{3/2}}.$$

One can represent the Airy function a little more concretely as a certain contour integral:

$$\mathrm{Ai}(x) = \frac{1}{2\pi i}\int_C e^{\frac{1}{3}\zeta^3 - x\zeta}\,d\zeta$$

where C is the contour given by the two rays $[0,\infty) \ni t \mapsto te^{\pm i\frac{\pi}{3}}$ in the plane. The **Airy kernel** is then defined to be

$$A(x,y) = \frac{\mathrm{Ai}(x)\mathrm{Ai}'(y) - \mathrm{Ai}'(x)\mathrm{Ai}(y)}{x-y} \tag{17.1}$$

for $x \neq y$ and defined by continuity on the diagonal: $A(x,x) = \mathrm{Ai}(x)\mathrm{Ai}''(x) - \mathrm{Ai}'(x)^2$.

Theorem 17.2. *The random variable* $n^{2/3}(\lambda_n - 2)$ *has a limit distribution as* $n \to \infty$*: its limiting cumulative distribution function is*

$$F_2(t) \equiv \lim_{n\to\infty} \mathbb{P}(n^{2/3}(\lambda_n - 2) \le t) = 1 + \sum_{k=1}^{\infty} \frac{(-1)^k}{k!}\int_t^\infty \cdots \int_t^\infty \det[A(x_i,x_j)]_{i,j=1}^n\,dx_1\cdots dx_k.$$

The alternating sum of determinants of Airy kernels is an example of a *Fredholm determinant*. Remarkable though this theorem is, it does not say a lot of computational worth about the cumulative distribution function F_2. The main theorem describing F_2 is as follows.

Theorem 17.3 (Tracy-Widom 1994)**.** *The cumulative distribution function* F_2 *from Theorem 17.2 satisfies*

$$F_2(t) = \exp\left\{-\int_t^\infty (x-t)q(x)^2\,dx\right\}$$

where q *is a solution of the* Painlevé II *equation* $q''(x) = xq(x) + 2q(x)^3$*, and* $q(x) \sim \mathrm{Ai}(x)$ *as* $x \to \infty$*.*

The proofs of Theorems 17.2 and 17.3 (and technology needed for them) would take roughly another full quarter to develop, so we cannot say much about them in the microscopic space remaining. Here we just give a glimpse of where the objects in Theorem 17.2 come from. Recall the Hermite kernel of Definition 15.10

$$K_n(x,y) = \sum_{k=0}^{n-1} \Psi_k(x)\Psi_k(y)$$

where Ψ_k are the Hermite functions. The Hermite kernel arose in a concise formulation of the joint law of eigenvalues of a *non-normalized* GUE_n; indeed, this (unordered) law $\mathscr{P}_n$ has as its density $\frac{1}{n!}\det[K_n(x_i,x_j)]_{i,j=1}^n$. Now, we have established that the right scale for structure in the largest eigenvalue of the *normalized* GUE_n is $n^{2/3}$; without the $\sqrt{n}$-rescaling, then, the fluctuations occur at order $n^{2/3}n^{-1/2} = n^{1/6}$, and the largest eigenvalue is of order $2\sqrt{n}$.

Proposition 17.4. *Let $A_n\colon \mathbb{R}^2 \to \mathbb{R}$ denote the kernel*

$$A_n(x,y) = n^{-1/6}K_n\left(2\sqrt{n} + n^{-1/6}x, 2\sqrt{n} + n^{-1/6}y\right).$$

Then $A_n \to A$ (the Airy kernel) as $n \to \infty$. The convergence is very strong: A_n and A both extend to analytic functions $\mathbb{C}^2 \to \mathbb{C}$, and the convergence $A_n \to A$ is uniform on compact subsets of $\mathbb{C}^2$; hence, all derivatives of A_n converge to the corresponding derivatives of A.

The proof of convergence in Proposition 17.4 is a very delicate piece of technical analysis. It is remarkable that the Airy kernel is "hiding inside" the Hermite kernel. Even more remarkable, however, is that other important kernels are also present at other scales.

Proposition 17.5. *Let $S_n(x,y) = n^{-1/2}K_n(n^{-1/2}x, n^{-1/2}y)$. Then $S_n \to S$ uniformly on bounded subsets of $\mathbb{R}^2$, where*

$$S(x,y) = \frac{1}{\pi}\frac{\sin(x-y)}{x-y}$$

is the **sine kernel***.*

Again, the proof of Proposition 17.5 is a very delicate technical proof. Structure at this scaling, however, tells us something not about the edge of the spectrum, but rather about the *spacing* of eigenvalues in "the bulk" (i.e. in a neighborhood of 0). The relevant theorem is:

Theorem 17.6 (Gaudin-Mehta 1960)**.** *Let $\lambda_1,\ldots,\lambda_n$ be the eigenvalues of a (normalized) GUE_n, and let V be a compact subset of $\mathbb{R}$. Then*

$$\lim_{n\to\infty}\mathbb{P}(n\lambda_1,\ldots,n\lambda_n \notin A) = 1 + \sum_{k=1}^\infty \frac{(-1)^k}{k!}\int_V\cdots\int_V \det[S(x_i,x_j)]_{i,j=1}^n\,dx_1\cdots dx_k.$$

The proof of Theorem 17.6 is not that difficult. Using the density $\frac{1}{n!}\det[K_n(x_i,x_j)]_{i,j=1}^n$ of the joint law $\mathscr{P}_n$ of eigenvalues, one can compute all the marginals, which are similarly given by determinants of K_n. Then (taking into account the $\sqrt{n}$-normalization) the probability we are asking about is $\mathscr{P}_n(A^c/\sqrt{n}\times\cdots\times A^c/\sqrt{n})$; appropriate determinantal expansion give a result for finite n mirroring the statement of Theorem 17.6, but involving the kernels $n^{-1/2}K_n(n^{-1/2}x, n^{-1/2}y)$. The dominated convergence theorem and Proposition 17.5 then yield the result.

Remark 17.7. Theorem 17.6 gives the spacing distribution of eigenvalues near 0: we see structure at the macroscopic scale n times the bulk scale (where the semicircle law appears). One can then ask for the spacing distribution near other points in the bulk: that is, suppose we blow up the eigenvalues near a point $t_0 \in (-2,2)$. The answer is of the same form, but involves the *expected density of eigenvalues* (i.e. the semiciricle law). It boils down to the following statement: if we define the shifted kernel $S_n^{t_0}$ to be

$$S_n^{t_0}(x,y) = n^{-1/2}K_n(t_0\sqrt{n} + n^{-1/2}x, t_0\sqrt{n} + n^{-1/2}y)$$

then the counterpart limit theorem to Proposition 17.5 is

$$\lim_{n\to\infty} S_n^{t_0}(x,y) = \frac{1}{\pi}\frac{\sin(s(t_0)(x-y))}{x-y} \equiv S^{t_0}(x,y)$$

where $s(t_0) = \frac{1}{2}\sqrt{4-t_0^2}$ is π times the semicircular density. Then the exact same proof outlines above shows that

$$\lim_{n\to\infty} \mathbb{P}(n\lambda_1,\ldots,n\lambda_n \notin t_0\sqrt{n}+V) = 1 + \sum_{k=1}^{\infty} \frac{(-1)^k}{k!} \int_V \cdots \int_V \det[S^{t_0}(x_i,x_j)]_{i,j=1}^n \, dx_1 \cdots dx_k.$$

This holds for any $|t_0| < 2$. When $|t_0| = 2$ the kernel $S^{t_0}(x,y) = 0$, which is to be expected: as we saw in Theorem 17.2, the fluctuations of the eigenvalues at the edge are of order $n^{-2/3}$, much smaller than the $n^{-1/2}$ in the bulk.

As with Theorem 17.2, Theorem 17.6 is a terrific theoretical tool, but it is lousy for trying to actually compute the limit spacing distribution. The relevant result (analogous to Theorem 17.3) in the bulk is as follows.

Theorem 17.8 (Jimbo-Miwa-Môri-Sato 1980)**.** *For fixed* $t > 0$*,*

$$\lim_{n\to\infty} \mathbb{P}(n\lambda_1,\ldots,n\lambda_n \notin (-t/2,t/2)) = 1 - F(t)$$

where F *is a cumulative probability distribution supported on* $[0,\infty)$*. It is given by*

$$1 - F(t) = \exp\left\{\int_0^t \frac{\sigma(x)}{x}\,dx\right\}$$

where σ *is a solution of the* Painlevé V *ODE*

$$(x\sigma''(x))^2 + 4(x\sigma'(x) - \sigma(x))(x\sigma'(x) - \sigma(x) + \sigma'(x)^2) = 0$$

and σ *has the expansion*

$$\sigma(x) = -\frac{x}{\pi} - \frac{x^2}{\pi^2} - \frac{x^3}{\pi^3} + O(x^4) \text{ as } x \downarrow 0.$$

Remark 17.9. The ODEs *Painlevé II* and *Painlevé V* are named after the late 19th / early 20th century mathematican and politician Paul Painlevé, who twice served as Prime Minster of (the Third Republic of) France (the first time in 1917 for 2 months, the second time in 1925 for 6 months). Earlier, around the turn of the century, he studied non-linear ODEs with the properties that the only movable singularities (singularities in solutions whose positions may depend on initial conditions) are poles (i.e. behave like $\frac{1}{(x-x_0)^n}$ for some natural number n in a neighborhood of the singularity x_0). All linear ODEs have this property, but non-linear equations can have essential movable singularities. In 1900, Painlevé painstakingly showed that all *second order* non-linear ODEs with this property (now called the Painlevé property) can be put into one of *fifty* canonical forms. In 1902, he went on to show that 44 of these equations could be transformed into other well-known differential equations. The 6 that remained after this enumeration are referred to as Painlevé I through VI.

The question of why two (unrelated) among the 6 unsolved Painlevé equations characterize behavior of GUE eigenvalues in the bulk and at the edge remains one of the deepest mysteries of this subject.