1

Introduction

This book is concerned with random matrices. Given the ubiquitous role that matrices play in mathematics and its application in the sciences and engineering, it seems natural that the evolution of probability theory would eventually pass through random matrices. The reality, however, has been more complicated (and interesting). Indeed, the study of random matrices, and in particular the properties of their eigenvalues, has emerged from the applications, first in data analysis (in the early days of statistical sciences, going back to Wishart [Wis28]), and later as statistical models for heavy-nuclei atoms, beginning with the seminal work of Wigner [Wig55]. Still motivated by physical applications, at the able hands of Wigner, Dyson, Mehta and co-workers, a mathematical theory of the spectrum of random matrices began to emerge in the early 1960s, and links with various branches of mathematics, including classical analysis and number theory, were established. While much progress was initially achieved using enumerative combinatorics, gradually, sophisticated and varied mathematical tools were introduced: Fredholm determinants (in the 1960s), diffusion processes (in the 1960s), integrable systems (in the 1980s and early 1990s), and the Riemann–Hilbert problem (in the 1990s) all made their appearance, as well as new tools such as the theory of free probability (in the 1990s). This wide array of tools, while attesting to the vitality of the field, presents, however, several formidable obstacles to the newcomer, and even to the expert probabilist. Indeed, while much of the recent research uses sophisticated probabilistic tools, it builds on layers of common knowledge that, in the aggregate, few people possess.

Our goal in this book is to present a rigorous introduction to the basic theory of random matrices that would be sufficiently self-contained to be accessible to graduate students in mathematics or related sciences who have mastered probability theory at the graduate level, but have not necessarily been exposed to advanced notions of functional analysis, algebra or geometry. With such readers in mind, we present some background material in the appendices, that novice and expert alike can consult; most material in the appendices is stated without proof, although the details of some specialized computations are provided.

Keeping in mind our stated emphasis on accessibility over generality, the book is essentially divided into two parts. In Chapters 2 and 3, we present a self-contained analysis of random matrices, quickly focusing on the Gaussian ensembles and culminating in the derivation of the gap probabilities at 0 and the Tracy–Widom law. These chapters can be read with very little background knowledge, and are particularly suitable for an introductory study. In the second part of the book, Chapters 4 and 5, we use more advanced techniques, requiring more extensive background, to emphasize and generalize certain aspects of the theory, and to introduce the theory of *free probability*.

So what is a random matrix, and what questions are we about to study? Throughout, let $\mathbb{F} = \mathbb{R}$ or $\mathbb{F} = \mathbb{C}$, and set $\beta = 1$ in the former case and $\beta = 2$ in the latter. (In Section 4.1, we will also consider the case $\mathbb{F} = \mathbb{H}$, the skew-field of quaternions, see Appendix E for definitions and details.) Let $\mathrm{Mat}_N(\mathbb{F})$ denote the space of N-by-N matrices with entries in $\mathbb{F}$, and let $\mathscr{H}_N^{(\beta)}$ denote the subset of self-adjoint matrices (i.e., real symmetric if $\beta = 1$ and Hermitian if $\beta = 2$). One can always consider the sets $\mathrm{Mat}_N(\mathbb{F})$ and $\mathscr{H}_N^{(\beta)}$, $\beta = 1,2$, as submanifolds of an appropriate Euclidean space, and equip it with the induced topology and (Borel) sigma-field.

Recall that a probability space is a triple $(\Omega, \mathscr{F}, P)$ so that $\mathscr{F}$ is a sigma-algebra of subsets of Ω and P is a probability measure on $(\Omega, \mathscr{F})$. In that setting, a *random matrix* X_N is a measurable map from $(\Omega, \mathscr{F})$ to $\mathrm{Mat}_N(\mathbb{F})$.

Our main interest is in the *eigenvalues* of random matrices. Recall that the eigenvalues of a matrix $H \in \mathrm{Mat}_N(\mathbb{F})$ are the roots of the characteristic polynomial $P_N(z) = \det(zI_N - H)$, with I_N the identity matrix. Therefore, on the (open) set where the eigenvalues are all simple, they are smooth functions of the entries of X_N (a more complete discussion can be found in Section 4.1).

We will be mostly concerned in this book with self-adjoint matrices $H \in \mathscr{H}_N^{(\beta)}$, $\beta = 1,2$, in which case the eigenvalues are all real and can be ordered. Thus, for $H \in \mathscr{H}_N^{(\beta)}$, we let $\lambda_1(H) \le \cdots \le \lambda_N(H)$ be the eigenvalues of H. A consequence of the perturbation theory of normal matrices (see Lemma A.4) is that the eigenvalues $\{\lambda_i(H)\}$ are continuous functions of H (this also follows from the Hoffman–Wielandt theorem, Theorem 2.1.19). In particular, if X_N is a random matrix then the eigenvalues $\{\lambda_i(X_N)\}$ are random variables.

We present now a guided tour of the book. We begin by considering *Wigner matrices* in Chapter 2. These are symmetric (or Hermitian) matrices X_N whose entries are independent and identically distributed, except for the symmetry constraints. For $x \in \mathbb{R}$, let δ_x denote the *Dirac* measure at x, that is, the unique probability measure satisfying $\int f d\delta_x = f(x)$ for all continuous functions on $\mathbb{R}$. Let $L_N = N^{-1}\sum_{i=1}^N \delta_{\lambda_i(X_N)}$ denote the *empirical measure* of the eigenvalues of X_N. Wigner's Theorem (Theorem 2.1.1) asserts that, under appropriate assumptions on the law of the entries, L_N converges (with respect to the weak convergence of measures) towards a deterministic probability measure, the *semicircle law*. We present in Chapter 2 several proofs of Wigner's Theorem. The first, in Section 2.1, involves a combinatorial machinery that is also exploited to yield central limit theorems and estimates on the spectral radius of X_N. After first introducing in Section 2.3 some useful estimates on the deviation between the empirical measure and its mean, we define in Section 2.4 the *Stieltjes transform* of measures and use it to give another quick proof of Wigner's Theorem.

Having discussed techniques valid for entries distributed according to general laws, we turn attention to special situations involving additional symmetry. The simplest of these concerns the *Gaussian ensembles*, the GOE and GUE, so named because their law is invariant under conjugation by orthogonal (resp., unitary) matrices. The latter extra symmetry is crucial in deriving in Section 2.5 an explicit joint distribution for the eigenvalues (thus effectively reducing consideration from a problem involving order of N^2 random variables, namely the matrix entries, to one involving only N variables). (The GSE, or Gaussian symplectic ensemble, also shares this property and is discussed briefly.) A large deviations principle for the empirical distribution, which leads to yet another proof of Wigner's Theorem, follows in Section 2.6.

The expression for the joint density of the eigenvalues in the Gaussian ensembles is the starting point for obtaining *local* information on the eigenvalues. This is the topic of Chapter 3. The bulk of the chapter deals with the GUE, because in that situation the eigenvalues form a *determinantal process*. This allows one to effectively represent the probability that no eigenvalues are present in a set as a *Fredholm determinant*, a notion that is particularly amenable to asymptotic analysis. Thus, after representing in Section 3.2 the joint density for the GUE in terms of a determinant involving appropriate orthogonal polynomials, the *Hermite* polynomials, we develop in Section 3.4 in an elementary way some aspects of the theory of Fredholm determinants. We then present in Section 3.5 the asymptotic analysis required in order to study the *gap probability at 0*, that is the probability that no eigenvalue is present in an interval around the origin. Relevant tools, such as the Laplace method, are developed along the way. Section 3.7 repeats this analysis for the edge of the spectrum, introducing along the way the method of steepest descent. The link with integrable systems and the *Painlevé equations* is established in Sections 3.6 and 3.8.

As mentioned before, the eigenvalues of the GUE are an example of a determinantal process. The other Gaussian ensembles (GOE and GSE) do not fall into this class, but they do enjoy a structure where certain Pfaffians replace determinants. This leads to a considerably more involved analysis, the details of which are provided in Section 3.9.

Chapter 4 is a hodge-podge of results whose common feature is that they all require new tools. We begin in Section 4.1 with a re-derivation of the joint law of the eigenvalues of the Gaussian ensemble, in a geometric framework based on Lie theory. We use this framework to derive the expressions for the joint distribution of eigenvalues of Wishart matrices, of random matrices from the various unitary groups and of matrices related to random projectors. Section 4.2 studies in some depth determinantal processes, including their construction, associated central limit theorems, convergence and ergodic properties. Section 4.3 studies what happens when in the GUE (or GOE), the Gaussian entries are replaced by Brownian motions. The powerful tools of stochastic analysis can then be brought to bear and lead to functional laws of large numbers, central limit theorems and large deviations. Section 4.4 consists of an in-depth treatment of concentration techniques and their application to random matrices; it is a generalization of the discussion in the short Section 2.3. Finally, in Section 4.5, we study a family of tri-diagonal matrices, parametrized by a parameter β, whose distribution of eigenvalues coincides with that of members of the Gaussian ensembles for $\beta = 1,2,4$. The study of the maximal eigenvalue for this family is linked to the spectrum of an appropriate random Schrödinger operator.

Chapter 5 is devoted to *free probability theory*, a probability theory for certain noncommutative variables, equipped with a notion of independence called free independence. Invented in the early 1990s, free probability theory has become a versatile tool for analyzing the laws of noncommutative polynomials in several random matrices, and of the limits of the empirical measure of eigenvalues of such polynomials. We develop the necessary preliminaries and definitions in Section 5.2, introduce free independence in Section 5.3, and discuss the link with random matrices in Section 5.4. We conclude the chapter with Section 5.5, in which we study the convergence of the spectral radius of noncommutative polynomials of random matrices.

Each chapter ends with bibliographical notes. These are not meant to be comprehensive, but rather guide the reader through the enormous literature and give some hint of recent developments. Although we have tried to represent accurately the historical development of the subject, we have necessarily omitted important references, misrepresented facts, or plainly erred. Our apologies to those authors whose work we have thus unintentionally slighted.

Of course, we have barely scratched the surface of the subject of random matrices. We mention now the most glaring omissions, together with references to some recent books that cover these topics. We have not discussed the theory of the Riemann–Hilbert problem and its relation to integrable systems, Painlevé equations, asymptotics of orthogonal polynomials and random matrices. The interested reader is referred to the books [FoIKN06], [Dei99] and [DeG09] for an in-depth treatment. We do not discuss the relation between asymptotics of random matrices and combinatorial problems – a good summary of these appears in [BaDS09]. We barely discuss applications of random matrices, and in particular do not review the recent increase in applications to statistics or communication theory – for a nice introduction to the latter we refer to [TuV04]. We have presented only a partial discussion of ensembles of matrices that possess explicit joint distribution of eigenvalues. For a more complete discussion, including also the case of non-Hermitian matrices that are not unitary, we refer the reader to [For05]. Finally, we have not discussed the link between random matrices and number theory; the interested reader should consult [KaS99] for a taste of that link. We further refer to the bibliographical notes for additional reading, less glaring omissions and references.

2

Real and complex Wigner matrices

2.1 Real Wigner matrices: traces, moments and combinatorics

We introduce in this section a basic model of random matrices. Nowhere do we attempt to provide the weakest assumptions or sharpest results available. We point out in the bibliographical notes (Section 2.7) some places where the interested reader can find finer results.

Start with two independent families of independent and identically distributed (i.i.d.) zero mean, real-valued random variables $\{Z_{i,j}\}_{1\le i<j}$ and $\{Y_i\}_{1\le i}$, such that $EZ_{1,2}^2 = 1$ and, for all integers $k \ge 1$,

$$r_k := \max\left(E|Z_{1,2}|^k, E|Y_1|^k\right) < \infty. \qquad (2.1.1)$$

Consider the (symmetric) $N \times N$ matrix X_N with entries

$$X_N(j,i) = X_N(i,j) = \begin{cases} Z_{i,j}/\sqrt{N}, & \text{if } i < j, \\ Y_i/\sqrt{N}, & \text{if } i = j. \end{cases} \qquad (2.1.2)$$

We call such a matrix a *Wigner matrix*, and if the random variables $Z_{i,j}$ and Y_i are Gaussian, we use the term *Gaussian Wigner matrix*. The case of Gaussian Wigner matrices in which $EY_1^2 = 2$ is of particular importance, and for reasons that will become clearer in Chapter 3, such matrices (rescaled by $\sqrt{N}$) are referred to as Gaussian orthogonal ensemble (GOE) matrices.

Let λ_i^N denote the (real) eigenvalues of X_N, with $\lambda_1^N \le \lambda_2^N \le \cdots \le \lambda_N^N$, and define the *empirical distribution* of the eigenvalues as the (random) probability measure on $\mathbb{R}$ defined by

$$L_N = \frac{1}{N}\sum_{i=1}^{N} \delta_{\lambda_i^N}.$$

Define the *semicircle distribution* (or *law*) as the probability distribution $\sigma(x)dx$

on $\mathbb{R}$ with density

$$\sigma(x) = \frac{1}{2\pi}\sqrt{4-x^2}\mathbf{1}_{|x|\le 2}\,. \tag{2.1.3}$$

The following theorem, contained in [Wig55], can be considered the starting point of random matrix theory (RMT).

Theorem 2.1.1 (Wigner) *For a Wigner matrix, the empirical measure L_N converges weakly, in probability, to the semicircle distribution.*

In greater detail, Theorem 2.1.1 asserts that for any $f \in C_b(\mathbb{R})$, and any $\varepsilon > 0$,

$$\lim_{N\to\infty} P(|\langle L_N, f\rangle - \langle \sigma, f\rangle| > \varepsilon) = 0\,.$$

Remark 2.1.2 The assumption (2.1.1) that $r_k < \infty$ for all k is not really needed. See Theorem 2.1.21 in Section 2.1.5.

We will see many proofs of Wigner's Theorem 2.1.1. In this section, we give a direct combinatorics-based proof, mimicking the original argument of Wigner. Before doing so, however, we need to discuss some properties of the semicircle distribution.

2.1.1 The semicircle distribution, Catalan numbers and Dyck paths

Define the moments $m_k := \langle \sigma, x^k\rangle$. Recall the Catalan numbers

$$C_k = \frac{\binom{2k}{k}}{k+1} = \frac{(2k)!}{(k+1)!k!}\,.$$

We now check that, for all integers $k \ge 1$,

$$m_{2k} = C_k\,, \quad m_{2k+1} = 0\,. \tag{2.1.4}$$

Indeed, $m_{2k+1} = 0$ by symmetry, while

$$\begin{aligned} m_{2k} &= \int_{-2}^{2} x^{2k}\sigma(x)dx = \frac{2\cdot 2^{2k}}{\pi}\int_{-\pi/2}^{\pi/2}\sin^{2k}(\theta)\cos^2(\theta)d\theta \\ &= \frac{2\cdot 2^{2k}}{\pi}\int_{-\pi/2}^{\pi/2}\sin^{2k}(\theta)d\theta - (2k+1)m_{2k}\,. \end{aligned}$$

Hence,

$$m_{2k} = \frac{2\cdot 2^{2k}}{\pi(2k+2)}\int_{-\pi/2}^{\pi/2}\sin^{2k}(\theta)d\theta = \frac{4(2k-1)}{2k+2}m_{2k-2}\,, \tag{2.1.5}$$

from which, together with $m_0 = 1$, one concludes (2.1.4).

The Catalan numbers possess many combinatorial interpretations. To introduce a first one, say that an integer-valued sequence $\{S_n\}_{0\le n\le \ell}$ is a *Bernoulli walk* of length ℓ if $S_0 = 0$ and $|S_{t+1} - S_t| = 1$ for $t \le \ell - 1$. Of particular relevance here is the fact that C_k counts the number of *Dyck paths* of length $2k$, that is, the number of nonnegative Bernoulli walks of length $2k$ that terminate at 0. Indeed, let β_k denote the number of such paths. A classical exercise in combinatorics is

Lemma 2.1.3 $\beta_k = C_k < 4^k$. *Further, the generating function* $\hat\beta(z) := 1 + \sum_{k=1}^{\infty} z^k \beta_k$ *satisfies, for* $|z| < 1/4$,

$$\hat\beta(z) = \frac{1 - \sqrt{1-4z}}{2z}\,. \tag{2.1.6}$$

Proof of Lemma 2.1.3 Let B_k denote the number of Bernoulli walks $\{S_n\}$ of length $2k$ that satisfy $S_{2k} = 0$, and let $\bar B_k$ denote the number of Bernoulli walks $\{S_n\}$ of length $2k$ that satisfy $S_{2k} = 0$ and $S_t < 0$ for some $t < 2k$. Then, $\beta_k = B_k - \bar B_k$. By reflection at the first hitting of -1, one sees that $\bar B_k$ equals the number of Bernoulli walks $\{S_n\}$ of length $2k$ that satisfy $S_{2k} = -2$. Hence,

$$\beta_k = B_k - \bar B_k = \binom{2k}{k} - \binom{2k}{k-1} = C_k\,.$$

Turning to the evaluation of $\hat\beta(z)$, considering the first return time to 0 of the Bernoulli walk $\{S_n\}$ gives the relation

$$\beta_k = \sum_{j=1}^{k} \beta_{k-j}\beta_{j-1}\,,\; k \ge 1\,, \tag{2.1.7}$$

with the convention that $\beta_0 = 1$. Because the number of Bernoulli walks of length $2k$ is bounded by 4^k, one has that $\beta_k \le 4^k$, and hence the function $\hat\beta(z)$ is well defined and analytic for $|z| < 1/4$. But, substituting (2.1.7),

$$\hat\beta(z) - 1 = \sum_{k=1}^{\infty} z^k \sum_{j=1}^{k} \beta_{k-j}\beta_{j-1} = z\sum_{k=0}^{\infty} z^k \sum_{j=0}^{k} \beta_{k-j}\beta_j\,,$$

while

$$\hat\beta(z)^2 = \sum_{k,k'=0}^{\infty} z^{k+k'}\beta_k\beta_{k'} = \sum_{q=0}^{\infty}\sum_{\ell=0}^{q} z^q \beta_{q-\ell}\beta_\ell\,.$$

Combining the last two equations, one sees that

$$\hat\beta(z) = 1 + z\hat\beta(z)^2\,,$$

from which (2.1.6) follows (using that $\hat{\beta}(0) = 1$ to choose the correct branch of the square-root). □

We note in passing that, expanding (2.1.6) in power series in z in a neighborhood of zero, one gets (for $|z| < 1/4$)

$$\hat{\beta}(z) = \frac{2\sum_{k=1}^{\infty} \frac{z^k (2k-2)!}{k!(k-1)!}}{2z} = \sum_{k=0}^{\infty} \frac{(2k)!}{k!(k+1)!} z^k = \sum_{k=0}^{\infty} z^k C_k ,$$

which provides an alternative proof of the fact that $\beta_k = C_k$.

Another useful interpretation of the Catalan numbers is that C_k counts the number of rooted planar trees with k edges. (A *rooted planar tree* is a planar graph with no cycles, with one distinguished vertex, and with a choice of ordering at each vertex; the ordering defines a way to "explore" the tree, starting at the root.) It is not hard to check that the Dyck paths of length $2k$ are in bijection with such rooted planar trees. See the proof of Lemma 2.1.6 in Section 2.1.3 for a formal construction of this bijection.

We note in closing that a third interpretation of the Catalan numbers, particularly useful in the context of Chapter 5, is that they count the *non-crossing partitions* of the ordered set $\mathscr{K}_k := \{1, 2, \ldots, k\}$.

Definition 2.1.4 A partition of the set $\mathscr{K}_k := \{1, 2, \ldots, k\}$ is called *crossing* if there exists a quadruple (a, b, c, d) with $1 \le a < b < c < d \le k$ such that a, c belong to one part while b, d belong to another part. A partition which is not crossing is a *non-crossing partition*.

Non-crossing partitions form a lattice with respect to refinement. A look at Figure 2.1.1 should explain the terminology "non-crossing": one puts the points $1, \ldots, k$ on the circle, and connects each point with the next member of its part (in cyclic order) by an internal path. Then, the partition is non-crossing if this can be achieved without arcs crossing each other.

It is not hard to check that C_k is indeed the number γ_k of non-crossing partitions of $\mathscr{K}_k$. To see that, let π be a non-crossing partition of $\mathscr{K}_k$ and let j denote the largest element connected to 1 (with $j = 1$ if the part containing 1 is the set $\{1\}$). Then, because π is non-crossing, it induces non-crossing partitions on the sets $\{1, \ldots, j-1\}$ and $\{j+1, \ldots, k\}$. Therefore, $\gamma_k = \sum_{j=1}^{k} \gamma_{k-j}\gamma_{j-1}$. With $\gamma_1 = 1$, and comparing with (2.1.7), one sees that $\beta_k = \gamma_k$.

Exercise 2.1.5 Prove that for $z \in \mathbb{C}$ such that $z \notin [-2, 2]$, the Stieltjes transform

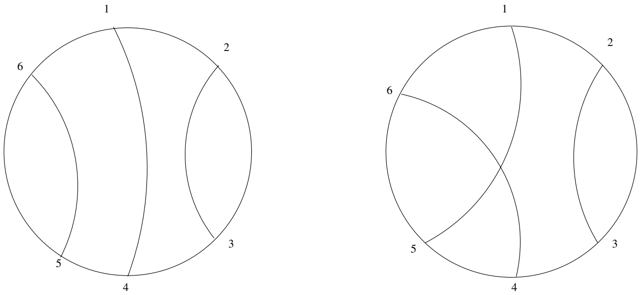

Fig. 2.1.1. Non-crossing (left, $(1,4),(2,3),(5,6)$) and crossing (right, $(1,5),(2,3),(4,6)$) partitions of the set $\mathscr{K}_6$.

$S(z)$ of the semicircle law (see Definition 2.4.1) equals

$$S(z) = \int \frac{1}{\lambda - z}\sigma(d\lambda) = \frac{-z+\sqrt{z^2-4}}{2z}.$$

Hint: Either use the residue theorem, or relate $S(z)$ to the generating function $\hat{\beta}(z)$, see Remark 2.4.2.

2.1.2 Proof #1 of Wigner's Theorem 2.1.1

Define the probability distribution $\bar{L}_N = EL_N$ by the relation $\langle \bar{L}_N, f\rangle = E\langle L_N, f\rangle$ for all $f \in C_b$, and set $m_k^N := \langle \bar{L}_N, x^k\rangle$. Theorem 2.1.1 follows from the following two lemmas.

Lemma 2.1.6 *For every $k \in \mathbb{N}$,*

$$\lim_{N\to\infty} m_k^N = m_k.$$

(See (2.1.4) for the definition of m_k.)

Lemma 2.1.7 *For every $k \in \mathbb{N}$ and $\varepsilon > 0$,*

$$\lim_{N\to\infty} P\left(\left|\langle L_N, x^k\rangle - \langle \bar{L}_N, x^k\rangle\right| > \varepsilon\right) = 0.$$

Indeed, assume that Lemmas 2.1.6 and 2.1.7 have been proved. To conclude the proof of Theorem 2.1.1, one needs to check that for any bounded continuous function f,

$$\lim_{N\to\infty} \langle L_N, f\rangle = \langle \sigma, f\rangle, \quad \text{in probability.} \tag{2.1.8}$$