CHAPTER 1

Introduction

1. What is random matrix theory?

A random matrix is a matrix whose entries are random variables. The eigenvalues and eigenvectors are then random too, and the main objective of the subject is to understand their distributions. This statement omits many other interesting aspects of random matrices, but is operationally useful to keep in mind. We start with examples.

(1) Let $X_1, \dots X_n$ be i.i.d $p \times 1$ random vectors having $N_p(0, \Sigma)$ distribution. Assume that Σ is unknown. Based on the data a natural estimate for Σ is the sample covariance matrix

$$S_n := \frac{1}{n} \sum_{k=1}^{n} X_k X_k^t.$$

Historically, this was the first random matrix to be studied,and goes by the name of *Wishart matrix*.

(2) Let $X = (X_{i,j})_{i,j \le n}$ where $X_{i,j}$, $i \le j$ are i.i.d real or complex valued random variables and $X_{i,j} = \overline{X}_{j,i}$. Then X is a Hermitian random matrix and hence has real eigenvalues. If we assume that $X_{i,j}$ have finite second moment, this matrix is called *Wigner matrix*.

Its origin lies in the study of heavy nucleii in Physics. Essentially, the behaviour of a nucleus is determined by a Hermitian operator (the Hamiltonian that appears in Schrodinger's equation). This opearator is a second order differential operator in about as many variables as the number of protons and neutrons and hence is beyond exact determination except in the simplest atoms. Eugene Wigner approached this problem by assuming that the exact details did not matter and replaced the Hermitian operator by a *random Hermitian matrix* of high dimensions. The eigenvalues of the original operator denote the energy levels and are of physical interest. By considering the eigenvalues of the random matrix, Wigner observed that statistically speaking, the

(3) Consider the matrix $A = (a_{i,j})_{i,j \le n}$ with i.i.d entries. There is less physical motivation for this model but probabilistically appears even simpler than the previous model as there is more independence. This is a false appearance, but we will come to that later!

(4) Patterned random matrices have come into fashion lately. For example, let X_i be i.i.d random variables and define the random *Toeplitz matrix* $T = \left(X_{|i-j|}\right)_{i,j \le n}$. One can also consider the asymmetric Toeplitz matrix. Many questions about the eigenvalues of these matrices are still open.

(5) Random unitary matrices.

(6) Random Schrodinger operators or random tridiagonal matrices.

2. Principal component analysis - a case for studying eigenvalues

We saw some situations in which random matrices arise naturally. But why study their eigenvalues. For Wigner matrices, we made the case that eigenvalues of the Hamiltonian are important in physics, and hence one must study eigenvalues of Wigner matrices which are supposed to model the Hamiltonian.

Here we make a case for studying the spectrum of the Wishart matrix which is more easy to understand for those of us physicsly challenged. Suppose $X_1, \ldots, X_n$ are $p \times 1$ vectors. For example, they could be vectors obtained by digitizing the photographs of employees in an office, in which case $n = 100$ and $p = 10000$ are not unreasonable values. Now presented with another vector Y which is one of the employees, we want a procedure to determine which of the X_is it is (for example, there is a door to a secure room where a photo is taken of anyone who enters the room, and the person is identified automatically). The obvious way to do it is to find the L^2 norm $\|Y - X_i\|_2$ for all $i \leq n$ and pick the value of i which minimizes the distance. As p is large, this involves a substantial amount of computation. Is there a more efficient way to do it?

There are many redundancies in the photograph. For example, if all employees have black hair, some of the co-ordinates have the same value in each of the X_is and hence is not helpful in distinguishing between individuals. Further, there are correlations. That is, if a few pixels (indicating the skin colour) are seen to be white, there is no need to check several other pixels which will probably be the same. How to use this redundancy in a systematic way to reduce computations?

We look for the unit vector $\alpha \in \mathbb{R}^p$ such that $\alpha^t X_1, \ldots, \alpha^t X_n$ have maximum variability. For siimplicity assume that $X_1 + \ldots + X_n = 0$. Then, the variance of the set $\alpha^t X_j$ is

$$\frac{1}{n} \sum_{j=1}^{n} (\alpha^t X_j)^2 = \alpha^t \left(\sum_{j=1}^{n} X_j X_j^t \right) \alpha = \alpha^t S_n \alpha$$

where S_n is the sample covariance matrix of X_js. But we know from linear algebra that the maximum of $\alpha^t S_n \alpha$ is the maximum eigenvalue of S_n and the maximizing α is the corresponding eigenvector. Thus we are led to eigenvalues and eigenvectors of S_n. In this problem, X_j are random, but it may be reasonable to suppose that X_js themselves (the employees) are samples from a larger population, say $N_p(0, \Sigma)$. If we knew Σ, we could use the first eigenvector of Σ, but if we do not know Σ, we would have to use the first eigenvector of S_n. The leads to the question of whether the first eigenvalue of S_n and of Σ are close to each other? If p is not small compared to n, one cannot expect such luck. More generally, by taking the top d eigenvectors, α_i, $i \leq d$, we reduce the dimension of vectors from p to d by replacing X_j by the vector $Y_j := (\alpha_1^t X_j, \ldots, \alpha_d^t X_j)$.

In any case, for now, this was just a motivation for looking into eigenvalues and eigenvectors of random matrices.

3. Some language and terminology and background

The space of probability measures: Let $\mathcal{P}(\mathbb{R})$ denote the space of Borel probability measures on $\mathbb{R}$. On $\mathcal{P}(\mathbb{R})$, define he Lévy metric

$$d(\mu, \nu) = \inf\{a > 0 : F_\mu(t - a) - a \leq F_\nu(t) \leq F_\mu(t + a) + a \; \forall t \in \mathbb{R}\}.$$

$\mathcal{P}(\mathbb{R})$ becomes a complete seperable metric space with the metric d. An important but easy fact is that $d(\mu_n, \mu) \to 0$ if and only if $\mu_n \to \mu$ in the sense of distribution (its importance is in that it shows weak convergence to be metrizable). Recall that *convergence in distribution*

or *convergence weakly* means that in terms of distribution functions, $F_{\mu_n}(x) \to F_\mu(x)$ for all x that are continuity points of F_μ.

ESD of a random or non-random matrix: Consider an $n \times n$ Hermitian matrix X with eigenvalues $\lambda_1, \ldots, \lambda_n$. The *empirical spectral distribution* (ESD) of X is the random measure $L_X := \sum_{k=1}^n \delta_{\lambda_k}$. If X is random, let $\overline{L}_X = \mathbf{E}[L_n]$ be the *expected ESD* of X. This means that $\overline{L}[a,b] = \mathbf{E}[L[a,b]] = \frac{1}{n}\mathbf{E}[\#\{k : \lambda_k \in [a,b]\}]$.

For a fixed matrix X, L_X is an element of $\mathcal{P}(\mathbb{R})$. If X is random, $\overline{L}_X$ is an element of $\mathcal{P}(\mathbb{R})$, while L_X is a random variable taking values in $\mathcal{P}(\mathbb{R})$ (that is, a measurable function with respect to the Borel sigma algebra on $\mathcal{P}(\mathbb{R})$).

Why do we talk about the empirical measures instead of eigenvalues directly? There are two advantages. Firstly, the eigenvalues of a matrix come without any special order, and L_X equally disregards the order and merely considers eigenvalues as a set (with appropriate multiplicities). Secondly, most often we study asymptotics of eigenvalues of a sequence of matrices X_n as the dimension n increases. If we think of eigenvalues as a vector $(\lambda_1, \ldots, \lambda_n)$, say by writing them in ascending order, then the space in which the vector takes values is $\mathbb{R}^n$ which changes with n. To talk of the limit of the vector becomes meaningless. But if we encode the eigenvalues by the ESD L_{X_n}, then they all take values in one space $\mathcal{P}(\mathbb{R})$ and we can talk about taking limits.

Exercise 1. Make sure you understand what the following statements mean.

1. $L_{X_n} \to \mu$ where X_n is a sequence of non-random matrices and $\mu \in \mathcal{P}(\mathbb{R})$.
2. $L_{X_n} \xrightarrow{P} \mu$ or $L_X \xrightarrow{a.s.} \mu$ where X_n is a sequence of random matrices and $\mu \in \mathcal{P}(\mathbb{R})$. Does this make sense if μ is itself a random probability measure?

4. Gaussian random variables

A standard normal random variable X is one that has density $(2\pi)^{-1/2}\exp\{-x^2/2\}$. We write $X \sim N(0,1)$. If X, Y are i.i.d $N(0,1)$, then the complex random variable $a := (X + iY)/\sqrt{2}$ is said to have standard complex Gaussian distribution. We write $a \sim CN(0,1)$. a has density $\pi^{-1}\exp\{-|z|^2\}$ on the complex plane.

We assume that you know all about multivariate normal distributions. Here is a quick recap of some facts, but stated for complex Gaussians which may be a tad unfamiliar. Let $a = (a_1, \ldots, a_n)^t$ where a_k are i.i.d $CN(0,1)$. If $Q_{m\times n}$ is a complex matrix and $u_{m\times 1}$ a complex vector, we say that $b = u + Qa$ has $CN_m(u, \Sigma)$ distribution, where $\Sigma = QQ^*$.

Exercise 2. Let a, b be as above.

1. Show that that distribution of b depends only on v and $\Sigma = QQ^*$.
2. Show that $\mathbf{E}[b_k] = u_k$ and $\mathbf{E}[(b_k - u_k)(b_\ell - u_\ell)] = 0$ while $\mathbf{E}[(b_k - u_k)\overline{(b_\ell - u_\ell)}] = \Sigma_{k,\ell}$.
3. If Q is nonsingular, show that b has density $\frac{1}{\pi^n \det(\Sigma)}\exp\{-(z-u)^*\Sigma^{-1}(z-u)\}$ on $\mathbb{C}^n$.
4. If $b \sim CN_m(u, \Sigma)$, find the distribution of $c := w + Rb$ where $w_{p\times 1}$ and $R_{p\times m}$.
5. The characteristic function of a $\mathbb{C}^m$-valued random vector c is the function $\varphi : \mathbb{C}^m \to \mathbb{C}$ defined as $\varphi(w) := \mathbf{E}[\exp\{i\Im\{w^*c\}\}]$. Show that if $u = 0$, then the characteristic function of b is $\varphi(w) = \exp\{-\frac{1}{4}w^*\Sigma w\}$.
6. If $b_{m\times 1}$ and $c_{n\times 1}$ are such that $(b^t, c^t)^t$ has $CN(u, \Sigma)$ we say that (b, c) has joint complex Gaussian distribution. Write

$$u = \begin{bmatrix} u_1 \\ u_2 \end{bmatrix}, \qquad \Sigma = \begin{bmatrix} A & B \\ B^* & C \end{bmatrix} \tag{1}$$

where the dimensions of u_i and A, B, C are self-evident. Then, show that $b \sim CN_m(u_1, A)$ and the conditional distribution of c given b is $CN(u_2 - B^*A^{-1}(b - u_1), C - B^*A^{-1}B)$.

(7) Suppose $X_{m\times 1}$ and $Y_{m\times 1}$ are real Gaussian vectors. Under what conditions is $X + iY$ have a complex Gaussian distribution?

Wick formula/Feynman diagram formula: Since the distribution of a real or complex Gaussian vector depends only on the mean vector and and covariance matrix, answers to all questions about the distribution must be presentable as a function of these parameters. Of course, in practice this may be impossible. One instance is the expectation of a product of Gaussians, and we show now that it can be written as a weighted sum over certain combinatorial objects. We first define two multilinear functions on matrices (the functions are linear in each column or each row). Let $\mathcal{S}_n$ denote the symmetric group on n letters. A *matching* of the set $[n]$ is a partition of $[n]$ into disjoint subsets of size two each. Let $\mathcal{M}_n$ denote the set of *matchings* of the set $[n]$ (it is nonempty if and only if n is even).

Definition 3. Let A be an $n \times n$ matrix. The *permanent of* A is defined as $\mathrm{per}(A) := \sum_{\pi\in\mathcal{S}_n} \prod_{i=1}^n a_{i,\pi_i}$. If A is symmetric, the *hafnian of* A is defined as $\mathrm{haf}(A) := \sum_{M\in\mathcal{M}_n} \prod a_{i,M_i}$. Here for each matching M, we take the product over all pairs in M, and each pair is taken only once.

Lemma 4. *Let $(b^t, c^t)^t$ be a complex Gaussian vector as in* (1)*. Then*

$$\mathbf{E}\left[\prod_{i=1}^{k} b_i \prod_{j=1}^{\ell} \overline{c}_\ell\right] = per(B).$$

In particular, if $b \sim CN(0,\Sigma)$ then $\mathbf{E}[|b_1|^2 \dots |b_m|^2] = per(\Sigma)$.

PROOF. It suffices to prove the second statement (why?). Thus, let $b \sim CN_m(0,\Sigma)$. Then, by exercise 2 we have its characteristic function

$$\mathbf{E}\left[\exp\left\{\frac{1}{2}w^*b - \frac{1}{2}b^*w\right\}\right] = \exp\left\{-\frac{1}{4}w^*\Sigma w\right\}.$$

Differentiate once with respect to $w_1, \dots, w_m$ and once with respect $\overline{w}_1, \dots, \overline{w}_m$ and then set $w = 0$. Differentiating under the expectation, on the left side we get $\frac{(-i)^m i^m}{2^{2m}}\mathbf{E}[|b_1 \dots b_k|^2]$. On the right side, expanding the exponential in series we get $\sum (k!)^{-1} 4^{-k} (w^*\Sigma w)^k$. Terms with $k < m$ vanish upon differentiation, while those with $k > m$ vanish when we set $w = 0$ (since at least one w_j factor remains after differentiating). Thus we only need to differentiate

$$(w^*\Sigma w)^m = \sum_{\substack{i_1,\dots,i_m \\ j_1,\dots,j_m}} \overline{w}_{i_1} w_{j_1} \dots \overline{w}_{i_m} w_{j_m} \sigma_{i_1,j_1} \dots \sigma_{i_m,j_m}.$$

Only those summands in which $\{i_1, \dots, i_m\}$ and $\{j_1, \dots, j_m\}$ are both permutations of $\{1, \dots, m\}$ survive the differentiation, and such a term contributes $\prod_k \sigma_{i_k,j_k}$. Thus, the right hand side finally reduces to

$$(m!)^{-1} 4^{-m} \sum_{\pi,\tau\in\mathcal{S}_m} \prod_{k=1}^{m} \sigma_{\pi_k,\tau_k} = m! 4^{-m} \sum_{\pi,\tau\in\mathcal{S}_m} \prod_{k=1}^{m} \sigma_{k,\tau\pi^{-1}(k)} = 4^{-m}\mathrm{per}(\Sigma)$$

since each permutation in $\mathcal{S}_m$ occurs $m!$ times as $\tau\pi^{-1}$. ∎

On similar lines (or can you think of another way without using characteristic functions?), prove the following Feynman diagram formula for real Gaussians.

Exercise 5. (1) Let $X \sim N_m(0,\Sigma)$. Then $\mathbf{E}[X_1X_2\dots X_m] = \mathrm{haf}(\Sigma)$. In particular, the expectation is zero if m is odd.
(2) For $X \sim N(0,1)$, we have $\mathbf{E}[X^{2m}] = (2m-1)(2m-3)\dots(3)(1)$, the number of matchings of the set $[2m]$.

The semicircle law: A probability distribution that arises frequently in random matrix theory and related subjects, but was never seen elsewhere in probability theory (as far as I know) is the *semicircular distribution* $\mu_{s.c}$ with density $\frac{1}{2\pi}\sqrt{4-x^2}$ on $[-2,2]$.

Exercise 6. Show that the odd moments of $\mu_{s.c}$ are zero and that the even moments are given by

$$\int x^{2n}\mu_{s.c}(dx) = \frac{1}{n+1}\binom{2n}{n}. \tag{2}$$

Catalan numbers: The number $C_n = \frac{1}{n+1}\binom{2n}{n}$ is called the n^{th} *Catalan number*. It has many combinatorial interpretations and arises frequently in mathematics. Here are some basic properties of Catalan numbers.

Exercise 7. (1) Show the recursion $C_{n+1} = \sum_{i=1}^{n} C_{i-1}C_{n-i}$ where the convention is that $C_0 = 1$.
(2) Show that the generating function of the Catalan numbers, $C(t) := \sum_{n=0}^{\infty} C_n t^n$ is satisfies $tC(t)^2 = C(t)+1$. Conclude that $C(t) = \frac{1}{2t}(1+\sqrt{1-4t})$. [**Note:** By Stirling's formula, estimate C_n and thus observe that $C(t)$ is indeed convergent on some neighbourhood of 0. This justifies all the manipulations in this exercise].

We show that Catalan numbers count various interesting sets of objects. The first is the set of *Dyck paths*.

Definition 8. If $X_1,\dots,X_n \in \{+1,-1\}$, let $S_k = X_1+\dots+X_k$. The sequence of lattice points $(0,0),(1,S_1),(2,S_2),\dots,(n,S_n)$ is called a "simple random walk path". A simple random walk path of length $2n$ is called a *bridge* if $S_{2n} = 0$. A simple random walk bridge is called a *Dyck path* of length $2n$ if $S_k \geq 0$ for all $k \leq 2n$.

Lemma 9. *The number of Dyck paths of length* $2n$ *is* C_n[1]

PROOF. Let A_q be the set of all sequences $X \in \{+1,-1\}^{2q+1}$ such that $\sum_i X_i = -1$ and such that $X_{2q+1} = -1$. Let B_q be the set of sequences X in A_q for which $S_j > -1$ for all $j \leq 2q$. Obviously, A_q is in one-to one correspondence with simple random walk bridges of length $2q$ (just pad a -1 at the end) and hence $|A_q| = \binom{2q}{q}$. Further, B_q is in bijection with the set of Dyck paths of length $2q$.

If $X,Y \in A_q$, define $X \sim Y$ if $(X_1,\dots,X_{2q})$ can be got by a cyclic permutation of $(Y_1,\dots,Y_q)$. This is an equivalence relationship and the equivalence classes all have size $q+1$ (since there are $q+1$ negative signs, and any of them can occur as the last one). We claim that exactly one path in each equivalence class belongs to B_q.

Indeed, fix $X \in A_q$, and consider the *first* index J such that $S_J = \min\{S_0,\dots,S_{2q}\}$. Obviously we must have $X_J = -1$. Consider the cyclic permute $Y = (X_{J+1},\dots,X_J)$. We leave it as an exercise to check that $Y \in B_q$ and that $Y' \notin B_q$ for any other cyclic shift of X. This shows that exactly one path in each equivalence class belongs to B_q and hence $|B_q| = (q+1)^{-1}|A_q| = C_q$. ∎

[1]The beautiful proof given here is due to Takács. An easy generalization is that if $X_i \geq -1$ are integers such that $S_n = -k$, then there are exactly k cyclic shifts of X for which $\min_{m<n} S_m > -k$. An interesting consequence is *Kemperman's formula*: If $X_i \geq -1$ are i.i.d integer valued random variables, then $\mathbf{P}(\tau_{-k} = n) = \frac{k}{n}\mathbf{P}(S_n = -k)$. Here τ_{-k} is the first hitting time of $-k$.

Exercise 10. In each of the following cases, show that the desired number is C_n by setting up a bijection with the set of Dyck paths. This is a small sample from Stanley's *Enumerative combinatorics*, where he gives sixty six such instances!

(1) The number of ways of writing n left braces "(" and n right braces ")" legitimately (so that when read from the left, the number of right braces never exceeds the number of left braces).
(2) A matching of the set $[2n]$ is a partition of this set into n pairwise disjoint two-element subsets. A matching is said to be non-crossing if there do not exist indices $i < j < k < \ell$ such that i is paired with k and j is paired with ℓ. The number of non-crossing matchings of $[2n]$ is C_n.
(3) $a_1, a_2, \ldots, a_n$ are elements in a group and they have no relationships among them. Consider all words of length $2n$ that use each a_k and a_k^{-1} exactly once (there are $(2n)!$ such words). The number of these words that reduce to identity is C_n.

Combine part (2) of exercise 6 with part (3) of exercise 10 to see that the $2n$ moment of the semicircle equals the number of non-crossing matchings of $[2n]$. Except for the phrase "non-crossing", this is identical to the combinatorial interpretation of Gaussian moments as given in part (2) of exercise 5. This analogy between the semicircle and Gaussian goes very deep as we shall see later.