Random Matrices and Communication Systems

Olivier Lévêque

EPFL, July 2012

Random matrices and communication systems: WEEK 2

In this lecture, we adopt the following (temporary) notations: small letters refer to scalar numbers and capital letters refer to scalar random variables.

Single antenna systems

Let us consider the additive white Gaussian noise (AWGN) channel:

$$Y_k = H_k\, X_k + Z_k$$

where $k \in \{1, \ldots, N\}$ is the time index and

- $(X_1, \ldots, X_N)$ is the input vector submitted to the average power constraint $\mathbb{E}\left(\frac{1}{N}\sum_{k=1}^N |X_k|^2\right) \leq P$;
- $(Z_1, \ldots, Z_N)$ is the noise vector, whose components are i.i.d.$\sim \mathcal{N}_{\mathbb{C}}(0, \sigma^2)$ random variables, independent of $X_1, \ldots, X_N$;
- $(Y_1, \ldots, Y_N)$ is the output vector;
- $H_1, \ldots, H_N$ are the fading coefficients (to be specified below).

The signal-to-noise ratio (SNR) of the system is defined as SNR $= P/\sigma^2$. In order to simplify notation, we will assume in the following that the noise variance $\sigma^2 = 1$, so that SNR $= P$.

The general question we would like to address in the present lecture is the following. Assume that two users wish to communicate over the above channel; what is then the maximum rate R at which communication can be established reliably? The answer depends of course on the specific model chosen for the fading coefficients H_k. We will review in the following various possible assumptions.

1 $H_k = h_k$, $k = 1, \ldots, N$ are deterministic coefficients

We start by considering the case of deterministic (complex-valued) fading coefficients, as an "appetizer" to the random case, which is the case of interest for this course.

In the deterministic case, the maximum rate at which one may possibly communicate over the time interval $[1, \ldots, N]$ is given by

$$\max R \leq \sup_{\substack{p_{X_1,\ldots,X_N} \\ \mathbb{E}(\frac{1}{N}\sum_{k=1}^N |X_k|^2) \leq P}} \frac{1}{N}\, I(X_1, \ldots, X_N; Y_1, \ldots, Y_N)$$

Let us compute

$$\begin{aligned} I(X_1, \ldots, X_N; Y_1, \ldots, Y_N) &= h(Y_1, \ldots, Y_N) - h(Y_1, \ldots, Y_N \,|\, X_1, \ldots, X_N) \\ &= h(h_1\, X_1 + Z_1, \ldots, h_N\, X_N + Z_N) - h(h_1\, X_1 + Z_1, \ldots, h_N\, X_N + Z_N \,|\, X_1, \ldots, X_N) \\ &= h(h_1\, X_1 + Z_1, \ldots, h_N\, X_N + Z_N) - h(Z_1, \ldots, Z_N \,|\, X_1, \ldots, X_N) \\ &= h(h_1\, X_1 + Z_1, \ldots, h_N\, X_N + Z_N) - h(Z_1, \ldots, Z_N) \end{aligned}$$

as $X_1, \ldots, X_N$ and $Z_1, \ldots, Z_N$ are independent by assumption. Using now the fact that for jointly continuous random variables $U_1, \ldots, U_N$,

$$h(U_1, \ldots, U_N) \leq \sum_{k=1}^N h(U_k)$$

with equality if and only if the U_k are independent, we obtain

$$I(X_1,\ldots,X_N;Y_1,\ldots,Y_N) \le \sum_{k=1}^N (h(h_k\, X_k + Z_k) - h(Z_k))$$

with equality if and only if the X_k are independent. Using then the fact

$$\sup_{p_U\,:\mathbb{E}(|U|^2)\le P} h(U) = \log(\pi e P)$$

where the supremum is attained for $U \sim \mathcal{N}_{\mathbb{C}}(0,P)$, we further obtain

$$I(X_1,\ldots,X_N;Y_1,\ldots,Y_N) \le \sum_{k=1}^N (\log(\pi e(P_k\,|h_k|^2+1)) - \log(\pi e)) = \sum_{k=1}^N \log(1+P_k\,|h_k|^2)$$

by taking $X_k \sim \mathcal{N}_{\mathbb{C}}(0,P_k)$ independent with $\frac{1}{N}\sum_{k=1}^N P_k \le P$ in order to meet the power constraint. This finally implies

$$\max R \le \sup_{\substack{P_1,\ldots,P_N\ge 0\\ \frac{1}{N}\sum_{k=1}^N P_k \le P}} \frac{1}{N}\sum_{k=1}^N \log(1+P_k\,|h_k|^2)$$

This optimization problem can be solved analytically; its solution is the well known “water-filling” solution, but let us not write this down explicitly at this stage. Also, without making any further assumption on the (arbitrary) sequence of fading coefficients h_k, we cannot conclude anything on the capacity of the channel in the large N limit.

1.1 $H_k \equiv h_0$ for all $k = 1,\ldots,N$

In this particular case, the above optimization problem is symmetric in $P_1,\ldots,P_N$ and has therefore the following simple solution:

$$\max R \le \sup_{\substack{P_1,\ldots,P_N\ge 0\\ \frac{1}{N}\sum_{k=1}^N P_k \le P}} \frac{1}{N}\sum_{k=1}^N \log(1+P_k\,|h_0|^2) = \log(1+P\,|h_0|^2)$$

Here, as h_0 is fixed, the above expression can also be shown to be equal to the capacity of the channel in the large N limit.

2 H_k, $k = 1,\ldots,N$ are random coefficients

In this section, we consider the H_k as random, in order to take into account the uncertainty about the fading coefficients. We should specify:

- how fast do these coefficients vary over time?

- among the receiver and the transmitter, who knows the realizations of these coefficients?

In the following, we assume that these coefficients have a given distribution and that this distribution is known to everyone. This is needed in order to be able to describe the statistics of the channel between X and Y. If even the distribution itself is not known, then the channel becomes an arbitrarily varying channel, which is out of the scope of the present course.

2.1 H_k are i.i.d. random variables (fast fading assumption)

This is in some sense an extreme assumption, which could be relaxed to "the coefficients H_k vary ergodically over time". By "ergodically", we actually mean that the empirical distribution of $H_1, \ldots, H_N$ converges to a given fixed distribution p_H (this holds in particular for an i.i.d. sequence, by the law of large numbers). But present in this assumption is also the fact that the coefficients H_k should vary relatively fast with respect to the duration of communication.

We need now to specify who knows the realizations of the coefficients H_k. In the following, we assume that that receiver is able to track (perfectly) the values of the H_k (by using pilot signals first, e.g.), but we make two different assumptions regarding the transmitter.

2.1.1 The transmitter knows the realizations of the coefficients H_k

This assumption is justified when feedback is easy to obtain at the transmitter. In this case, as everyone knows the channel realizations, it is as if these were actually deterministic, so the maximum rate achievable over the time interval $[1, \ldots, N]$ is bounded above by

$$\max R \leq \sup_{\substack{P_1,\ldots,P_N \geq 0 \\ \frac{1}{N}\sum_{k=1}^N P_k \leq P}} \frac{1}{N} \sum_{k=1}^N \log(1 + P_k\,|H_k|^2)$$

This leads to the definition of *ergodic capacity*:

$$C_{\text{erg}} = \lim_{N\to\infty} \sup_{\substack{P_1,\ldots,P_N \geq 0 \\ \frac{1}{N}\sum_{k=1}^N P_k \leq P}} \frac{1}{N} \sum_{k=1}^N \log(1 + P_k\,|H_k|^2) = \sup_{\substack{Q(\cdot)\geq 0 \\ \int_{\mathbb{C}} dh\, p_H(h)\, Q(h) \leq P}} \int_{\mathbb{C}} dh\, p_H(h)\, \log(1 + Q(h)\,|h|^2)$$

This can in turn be rewritten as

$$C_{\text{erg}} = \sup_{\substack{Q(\cdot)\geq 0 \\ \mathbb{E}_H(Q(H)) \leq P}} \mathbb{E}_H(\log(1 + Q(H)\,|H|^2))$$

The solution of this optimization problem is given by the water-filling solution:

$$C_{\text{erg}} = \mathbb{E}_H\left(\left(\log(\nu\,|H|^2)\right)^+\right) \quad \text{where } \nu \text{ satisfies} \quad \mathbb{E}_H\left(\left(\nu - \tfrac{1}{|H|^2}\right)^+\right) \leq P \qquad (1)$$

and $a^+ = \max(a, 0)$ denotes the positive part of $a \in \mathbb{R}$.

2.1.2 The transmitter does not know the realizations of the coefficients H_k

This assumption is justified when feedback is difficult, or even impossible, to obtain at the transmitter. In this case, the input vector $(X_1, \ldots, X_N)$ cannot be tuned according to the channel realizations $H_1, \ldots, H_N$, which we model mathematically by saying that $X_1, \ldots, X_N$ and $H_!, \ldots, H_N$ are independent.

As we assume on the other hand that the receiver knows the H_k, the channel between the transmitter and the receiver can be seen in this case as the channel with input $(X_1, \ldots, X_N)$ and output $(Y_1, \ldots, Y_N,\, H_1, \ldots, H_N)$; it is as if a genie were revealing the channel coefficients H_k to the receiver. So the mutual information of this channel is given by

$$\begin{aligned} & I(X_1, \ldots, X_N; Y_1, \ldots, Y_N,\, H_1, \ldots, H_N) \\ & \quad = \; I(X_1, \ldots, X_N; H_1, \ldots, H_N) + I(X_1, \ldots, X_N; Y_1, \ldots, Y_N \,|\, H_1 \ldots, H_N) \\ & \quad = \; 0 + h(Y_1, \ldots, Y_N \,|\, H_1, \ldots, H_N) - h(Y_1, \ldots, Y_N \,|\, X_1, \ldots, X_N,\, H_1, \ldots, H_N) \end{aligned}$$

where the chain rule was used in the first inequality and the independence of the H_k and X_k in the second inquality. As $Y_k = H_k\, X_K + Z_K$ we further obtain

$$\begin{aligned} & I(X_1, \ldots, X_N; Y_1, \ldots, Y_N,\, H_1, \ldots, H_N) \\ &= h(Y_1, \ldots, Y_N \,|\, H_1, \ldots, H_N) - h(Z_1, \ldots, Z_N) \\ &= \int_{\mathbb{C}^N} dh_1 \cdots dh_N\, p_{H_1,\ldots,H_N}(h_1, \ldots, h_N)\, h(h_1\, X_1 + Z_1, \ldots, h_N\, X_N + Z_N) - h(Z_1, \ldots, Z_N) \end{aligned}$$

where we have used the fact that the H_k, X_k, Z_k are independent. This can in turn be bouded above by

$$\begin{aligned} I(X_1, \ldots, X_N; Y_1, \ldots, Y_N,\, H_1, \ldots, H_N) &\leq \sum_{k=1}^{N} \int_{\mathbb{C}} dh_k\, p_{H_k}(h_k)\ (h(h_k\, X_k + Z_k) - h(Z_k)) \\ &\leq \sum_{k=1}^{N} \int_{\mathbb{C}} dh_k\, p_{H_k}(h_k)\ \log(1 + P_k\, |h_k|^2) \end{aligned}$$

by chosing $X_k \sim \mathcal{N}_{\mathbb{C}}(0, P_k)$ independent with $\frac{1}{N}\sum_{k=1}^{N} P_k \leq P$ (and notice that this choice of X_k maximzing the mutal information does *not* depend on the particular realizations of the fading coefficients H_k). The ergodic capacity of the channel is therefore given in this case by

$$\begin{aligned} C_{\text{erg}} &= \lim_{n\to\infty} \sup_{\substack{p_{X_1,\ldots,X_N} \\ \mathbb{E}\left(\frac{1}{N}\sum_{k=1}^{N}|X_k|^2\right)\leq P}} \frac{1}{N}\, I(X_1, \ldots, X_N; Y_1, \ldots, Y_N,\, H_1, \ldots, H_N) \\ &= \lim_{N\to\infty} \sup_{\substack{P_1,\ldots,P_N\geq 0 \\ \frac{1}{N}\sum_{k=1}^{N} P_k \leq P}} \frac{1}{N} \sum_{k=1}^{N} \int_{\mathbb{C}} dh_k\, p_{H_k}(h_k)\ \log(1 + P_k\, |h_k|^2) = \int_{\mathbb{C}} dh\, p_H(h)\ \log(1 + P\, |h|^2) \end{aligned}$$

because again of the symmetry of the optimization problem. This can in turn be rewritten as

$$C_{\text{erg}} = \mathbb{E}_H(\log(1 + P\, |H|^2)) \tag{2}$$

Remarks. - Comparing this expression with (1), we see the impact on the capacity of not knowing the channel coefficients at the transmitter.

- By Jensen's inequality, we obtain

$$C_{\text{erg}} = \mathbb{E}_H(\log(1 + P\, |H|^2)) \leq \log(1 + P\, \mathbb{E}_H(|H|^2))$$

so the ergodic capacity is less than or equal to the capacity of the channel with fixed and deterministic fading coefficient h_0 satisfying $|h_0|^2 = \mathbb{E}_H(|H|^2)$, which leads to the conclusion that fading degrades capacity in single antenna channels. We will see that the situation differs in multiple antenna channels.

2.2 $H_k \equiv H$ for all $k = 1, \ldots, N$ (slow fading assumption)

This is again an extreme assumption, which could be relaxed to "the variations of the coefficients H_k are sufficiently small over the duration of communication". We again assume that the receiver is able to track the fading coefficients and make two different assumptions on the transmitter.

2.2.1 The transmitter knows the realization of H

In this case, as H is fixed over time and known to everyone, it is as if we were in the fixed and determinisitc scenariio (see paragraph 1.1), so the capacity of the channel is given by

$$C = \log(1 + P\, |H|^2)$$

which is a random variable here, depending on the given realization of H.

2.2.2 The transmitter does not know the realization of H

In this case, let us moreover assume that the distribution of H admits a continuous pdf p_H whose support contains the point 0 (this is in particular verified for Rayleigh fading, that is, when $H \sim \mathcal{N}_{\mathbb{C}}(0,1)$). This is implying that whatever $\varepsilon > 0$, $\mathbb{P}(|H|^2 < \varepsilon) > 0$. As a consequence, whatever the rate chosen by the transmitter for communication (who does not know the value of H), there is always a non-zero probability that the chosen rate is above the actual capacity of the channel. Therefore, the capacity in this case is strictly speaking equal to zero. We therefore shift our attention to another performance measure: the *outage probability*, defined as, for a given target rate $R > 0$,

$$P_{\text{out}}(R) = \mathbb{P}_H(\log(1 + P\,|H|^2) < R)$$

The outage probability is a lower bound on the error probability achieved by any scheme on this channel (exactly like the capacity is an upper bound on the rate achieved by any scheme on a given channel). As long as the assumption on p_H made at the beginning of this paragraph is verified, the outage probability is always strictly positive.

Considering the high SNR regime (i.e. $P \to \infty$), this probability can still be made vanishingly small. First, observe that as $P \to \infty$,

$$C_{\text{erg}} = \mathbb{E}_H(\log(1 + P\,|H|^2)) \simeq \log P$$

So if one wants $P_{\text{out}}(R)$ to decrease to zero as $P \to \infty$, one should not choose the target rate R higher than $\log P$. Let us therefore choose $R = r \log P$, with $0 \le r \le 1$. In the case where $H \sim \mathcal{N}_{\mathbb{C}}(0,1)$, we obtain

$$\begin{aligned} P_{\text{out}}(r \log P) &= \mathbb{P}_H(\log(1 + P\,|H|^2) < r\,\log P) = \mathbb{P}_H\left(1 + P\,|H|^2 < P^r\right) \\ &= \mathbb{P}_H\left(|H|^2 < \frac{P^r - 1}{P}\right) \simeq \mathbb{P}_H\left(|H|^2 < P^{r-1}\right) \simeq P^{r-1} \end{aligned}$$

So the decay is polynomial in P. In addition, we observe the following tradeoff: the lower the target rate $r\,\log P$, the higher the speed of decrease to zero of the outage probability.

Random matrices and communication systems: WEEK 3

In this lecture and in the subsequent ones, we adopt the following notations: small letters refer to scalars (deterministic or random) and deterministic vectors, while capital letters refer to matrices (deterministic or random) and random vectors. In some cases, this rule will not be followed strictly, but what is actually meant will be clear from the context.

Summary of last lecture and single-letter characterization

- We saw first that when the fading coefficient h_0 is fixed over time and deterministic, the capacity of the channel is given by

$$C = \log(1 + P\,|h_0|^2)$$

This result can also be seen as the solution of the following single-letter characterization of the channel capacity (that remains valid in the more general context of multiple antenna systems):

$$C = \sup_{p_X\,:\,\mathbb{E}(|X|^2)\leq P} I(X;Y)$$

- Then, we saw that when the fading coefficients H_k are random and i.i.d. over time, known at both the transmitter and the receiver, the ergodic capacity of the channel is given by

$$C_{\text{erg}} = \sup_{\substack{Q(\cdot)\geq 0\\ \mathbb{E}_H(Q(H))\leq P}} \mathbb{E}_H(\log(1 + Q(H)\,|H|^2))$$

while when they are known at the receiver but not at the transmitter,

$$C_{\text{erg}} = \mathbb{E}_H(\log(1 + P\,|H|^2))$$

Again, in this second case (which we will mainly focus on in the following), this ergodic capacity expression may be found as the result of the more general single-letter characterization:

$$C_{\text{erg}} = \sup_{p_X\,:\,\mathbb{E}(|X|^2)\leq P} I(X;Y,H)$$

- Finally, when the fading coefficient H is random and fixed over time, known at both the transmitter and the receiver, the capacity of the channel is a random variable given by

$$C = \log(1 + P\,|H|^2)$$

while when H is not known at the transmitter, the capacity is equal to zero and the outage probability is given by

$$P_{\text{out}}(R) = \mathbb{P}_H(\log(1 + P\,|H|^2) < R)$$

for a target rate $R > 0$. Again, in this second case, this expression may be viewed as the result of the more general single-letter characterization:

$$P_{\text{out}}(R) = \inf_{p_X\,:\,\mathbb{E}(|X|^2)\leq P} \mathbb{P}_H(I(X;Y) < R)$$

(notice that in this case, $I(X;Y)$ is a random variable depending on the realization of H).

Preliminaries for this lecture

- An n-variate complex-valued random vector $X = (X_1, \ldots, X_n)$ is *(jointly) continuous* if it admits a joint pdf $p_X = p_{X_!,\ldots,X_n}$, i.e.

$$\mathbb{P}(X \in B) = \int_B dx\, p_X(x) \quad \forall B \subset \mathbb{C}^n \text{ Borel set}$$

Its *mean vector* is defined as $\mu = (\mathbb{E}(X_1), \ldots, \mathbb{E}(X_n))$ and its *covariance matrix* is defined as $(Q_X)_{jk} = \mathbb{E}(X_j \, \overline{X_k})$ (when they exist).

- Let X be a complex *Gaussian* random vector with mean 0 and positive definite covariance matrix Q_X (notation: $X \sim \mathcal{N}_{\mathbb{C}}(0, Q_X)$). This random vector admits the following joint pdf:

$$p_X(x) = \frac{1}{\pi^n \det(Q_X)} \, \exp\left(-x^* \, (Q_X)^{-1} \, x\right), \quad x \in \mathbb{C}^n$$

Notice that $\mathbb{E}(X_j) = 0$, $\mathbb{E}(X_j \, \overline{X_k}) = (Q_X)_{jk}$ and also that $X_1, \ldots, X_n$ are independent if and only if Q_X is diagonal.

- The *differential entropy* of a continuous random vector X is defined as

$$h(X) = -\int_{\mathbb{C}^n} dx \, p_X(x) \, \log(p_X(x))$$

One can check that

$$h(X) \leq \sum_{j=1}^n h(X_j)$$

with equality if and only if the X_j are independent, and also that

$$\sup_{p_X \,:\, \mathbb{E}(XX^*) = Q_X} h(X) = \log\det(\pi e \, Q_X)$$

is achieved by taking $X \sim \mathcal{N}_{\mathbb{C}}(0, Q_X)$.

Multiple antenna systems

We now consider the multiple antenna system (only one time-slot is considered here):

$$Y = H \, X + Z$$

where X, Y, Z are n-variate vectors and H is an $n \times n$ matrix. More precisely,

- X is the input vector submitted to the average power constraint $\mathbb{E}(\|X\|^2) \leq P$; (notice that $\mathbb{E}(\|X\|^2) = \mathbb{E}(X^* X) = \mathbb{E}(\mathrm{Tr}(X X^*)) = \mathrm{Tr}(Q_X)$)
- $Z \sim \mathcal{N}_{\mathbb{C}}(0, I)$ is the noise vector, independent of X;
- Y is the output vector;
- H is the channel fading matrix (to be specified below).

1 $H = H_0$ is deterministic and fixed over time

In this case, the single-letter characterization of the capacity reads

$$C = \sup_{p_X \,:\, \mathrm{Tr}(Q_X) \leq P} I(X;Y) = \sup_{p_X \,:\mathrm{Tr}(Q_X) \leq P} h(Y) - h(Y|X)$$

Notice that $h(Y|X) = h(H_0 \, X + Z \,|\, X) = h(Z)$, so

$$C = \left(\sup_{p_X \,:\, \mathrm{Tr}(Q_X) \leq P} h(H_0 \, X + Z) \right) - h(Z)$$

The above expression is maximized when $H_0\,X+Z$ is Gaussian, which happens when X itself is Gaussian. In this case, $H_0\,X+Z \sim \mathcal{N}_{\mathbb{C}}(0, I+H_0\,Q_X\,H_0^*)$, so

$$C = \left(\sup_{Q_X\,:\,\mathrm{Tr}(Q_X)\le P} \log\det(\pi e(I+H_0\,Q_X\,H_0^*))\right) - \log\det(\pi e I) = \sup_{Q_X\,:\,\mathrm{Tr}(Q_X)\le P} \log\det(I+H_0\,Q_X\,H_0^*)$$

In order to proceed further, we need the following inequality, whose proof is left as an exercise in the homework.

Hadamard's inequality. Let A be a positive semi-definite $n\times n$ matrix. Then $\det(A)\le\prod_{j=1}^n a_{jj}$.

A second ingredient is the *singular value decomposition* of H_0, stating that there exist unitary matrices U, V and $\Sigma = \mathrm{diag}(\sigma_1,\ldots,\sigma_n)$, with $\sigma_j\ge 0$ for all j, such that $H_0 = U\,\Sigma\,V^*$. Therefore,

$$\log\det(I+H_0\,Q_X\,H_0^*) = \log\det(I+U\,\Sigma\,V^*\,Q_X\,V\,\Sigma^*\,U^*) = \log\det(I+\Sigma\,V^*\,Q_X\,V\,\Sigma^*)$$

Let now $\widetilde{Q}_X = V^*\,Q_X\,V$. Notice that $\widetilde{Q}_X$ also satisfies the above constraints:

$$\widetilde{Q}_X \ge 0 \quad\text{and}\quad \mathrm{Tr}(\widetilde{Q}_X) = \mathrm{Tr}(Q_X)\le P$$

so

$$C = \sup_{Q_X\,:\,\mathrm{Tr}(Q_X)\le P} \log\det(I+H_0\,Q_X\,H_0^*) = \sup_{\widetilde{Q}_X\,:\,\mathrm{Tr}(\widetilde{Q}_x)\le P} \log\det(I+\Sigma\,\widetilde{Q}_X\,\Sigma^*)$$

Using now Hadamard's inequality, we obtain

$$\det(I+\Sigma\,\widetilde{Q}_X\,\Sigma^*) \le \prod_{j=1}^n \left(1+(\Sigma\,\widetilde{Q}_X\,\Sigma^*)_{jj}\right) = \prod_{j=1}^n \left(1+(\widetilde{Q}_X)_{jj}\,\sigma_j^2\right)$$

and the equality is met by taking $\widetilde{Q}_X$ diagonal, say $\widetilde{Q}_X = \mathrm{diag}(d_1,\ldots,d_n)$. The above expression for the capacity can therefore be rewritten as

$$C = \sup_{\substack{d_1,\ldots,d_n\ge 0\\ \sum_{j=1}^n d_j\le P}} \sum_{j=1}^n \log(1+d_j\,\sigma_j^2)$$

The solution of this optimization problem is again obtained via water-filling:

$$C = \sum_{j=1}^n \left(\log(\nu\,\sigma_j^2)\right)^+ \quad\text{where}\quad \sum_{j=1}^n \left(\nu-\frac{1}{\sigma_j^2}\right)^+ \le P$$

As an example, let us consider the following simple case: $(H_0)_{jk}=1$ for all j,k. In this case, $\sigma_1=n$, $\sigma_2=\ldots=\sigma_n=0$, so

$$C = \log(\nu n^2) \quad\text{such that}\quad \left(\nu-\frac{1}{n^2}\right)\le P$$

i.e. $\nu = P+\frac{1}{n^2}$ and $C=\log(1+P\,n^2)$.

2 H is random and varying ergodically over time (fast fading)

We assume now that the matrix H admits the pdf $p_H(\cdot)$ and that its realizations over time are i.i.d. (or ergodic), known at the receiver but not at the transmitter (so that X and H are independent). In this case, the single-letter characterization of the capacity reads:

$$C_{\mathrm{erg}} = \sup_{p_X\,:\,\mathrm{Tr}(Q_X)\le P} I(X;Y,H) = \sup_{p_X\,:\,\mathrm{Tr}(Q_X)\le P} I(X;H)+I(X;Y|H)$$

by the chain rule. Because of the independence of X and H, the first term is zero, so

$$C_{\text{erg}} = \sup_{p_X \,:\, \text{Tr}(Q_X) \le P} \int_{\mathbb{C}^{n^2}} dG \, p_H(G) \, I(X; Y | H = G)$$

Notice that for any fixed matrix G,

$$I(X; Y | H = G) = I(X; G\,X + Z) \le \log\det(I + G\,Q_X\,G^*)$$

and the equality is met when $X \sim \mathcal{N}_(0, Q_X)$ (which does not depend of the specific value of G). So

$$C_{\text{erg}} = \sup_{Q_X \ge 0 \,:\, \text{Tr}(Q_X) \le P} \int_{\mathbb{C}^{n^2}} dG \, p_H(G) \, \log\det(I + G\,Q_X\,G^*)$$

which can be rewritten as

$$C_{\text{erg}} = \sup_{Q_X \ge 0 \,:\, \text{Tr}(Q_X) \le P} \mathbb{E}_H(\log\det(I + H\,Q_X\,H^*))$$

Remark. It is *not* because the realizations of H are not known at the transmitter that the optimal Q_X should be a multiple of identity; it is indeed always possible to optimize over the *distribution* of H. Notice also that the solution is *not* the water-filling solution, as the singular values and vectors of H are not known at the transmitter.

Nevertheless, under symmetry conditions on the distribution of H, something more can be said on the optimal input covariance matrix Q_X. This is illustrated in the following lemmas, whose respective proofs are left as exercises in the homework.

Lemma 2.1. If h_{jk} are i.i.d. random variables, then the optimal input covariance matrix is of the form

$$Q_X = \frac{P}{n} \begin{pmatrix} 1 & c & \cdots & c & c \\ c & 1 & c & & c \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ c & & c & 1 & c \\ c & c & \cdots & c & 1 \end{pmatrix}$$

where $-\frac{1}{n-1} \le c \le 1$ is some real parameter.

Lemma 2.2. If h_{jk} are independent random variables such that $h_{jk} \sim -h_{jk}$ for all j, k, then the optimal input covariance matrix Q_X is diagonal.

Lemma 2.3. If h_{jk} are i.i.d random variables such that $h_{jk} \sim -h_{jk}$ for all j, k, then the optimal input covariance matrix $Q_X = \frac{P}{n}\, I$.

Notice that the third lemma is simply a combination of the first two. It holds true in particular when h_{jk} are i.i.d.$\sim \mathcal{N}_{\mathbb{C}}(0, 1)$ random variables, in which case

$$C_{\text{erg}} = \mathbb{E}_H \left(\log\det \left(I + \frac{P}{n}\, H H^* \right) \right)$$

We will analyze this expression further in the next lecture.

Random matrices and communication systems: WEEK 4

2bis H is random and varying ergodically over time (fast fading)

Let us consider again the expression found for the ergodic capacity, under the assumption that the fading coefficients h_{jk} are i.i.d.$\sim \mathcal{N}_{\mathbb{C}}(0,1)$ random variables:

$$C_{\text{erg}} = \mathbb{E}\left(\log\det\left(I + \frac{P}{n}\,HH^*\right)\right)$$

This expression can be computed explicitly via random matrix theory, which is the subject of this course. Before that, we will see below a relatively simple computation that will provide us with a lower bound on C_{erg}, that turns out to be asymptotically tight, either when $n \to \infty$ or $P \to \infty$.

As a preliminary, we need the *Paley-Zygmund inequality*: if $X \geq 0$ is a square-integrable random variable, then

$$\mathbb{P}(X > t) \geq \frac{(\mathbb{E}(X) - t)^2}{\mathbb{E}(X^2)} \quad \forall 0 \leq t \leq \mathbb{E}(X)$$

The proof of this fact is an application of Cauchy-Schwarz' inequality.

Let now $\lambda_1, \ldots, \lambda_n$ denote the eigenvalues of the (positive semi-definite) $n \times n$ matrix $\frac{1}{n}\,HH^*$. This allows us to write

$$C_{\text{erg}} = \mathbb{E}\left(\sum_{j=1}^n \log(1 + P\lambda_j)\right)$$

Let furthermore λ be one of the eigenvalues $\lambda_1, \ldots, \lambda_n$ picked uniformly at random. We obtain

$$C_{\text{erg}} = n\,\mathbb{E}(\log(1 + P\lambda)) \geq n\,\log(1 + Pt)\,\mathbb{P}(\lambda > t) \geq n\,\log(1 + Pt)\,\frac{(\mathbb{E}(\lambda) - t)^2}{\mathbb{E}(\lambda^2)}$$

for all $0 \leq t \leq \mathbb{E}(\lambda)$, by the above Paley-Zygmund inequality. Let us compute (these are by the way our first random matrix computations):

$$\mathbb{E}(\lambda) = \mathbb{E}\left(\frac{1}{n}\sum_{j=1}^n \lambda_j\right) = \mathbb{E}\left(\frac{1}{n}\text{Tr}\left(\frac{1}{n}\,HH^*\right)\right) = \frac{1}{n^2}\sum_{j,k=1}^n \mathbb{E}\left(|h_{jk}|^2\right) = \frac{1}{n^2}\,n^2 = 1$$

and

$$\mathbb{E}\left(\lambda^2\right) = \mathbb{E}\left(\frac{1}{n}\sum_{j=1}^n \lambda_j^2\right) = \mathbb{E}\left(\frac{1}{n}\text{Tr}\left(\left(\frac{1}{n}\,HH^*\right)^2\right)\right) = \frac{1}{n^3}\sum_{j,k,l,m=1}^n \mathbb{E}\left(h_{jk}\,\overline{h_{lk}}\,h_{lm}\overline{h_{jm}}\right)$$

Notice that because the h_{jk} are i.i.d. and $\mathbb{E}(h_{jk}) = 0$ for all j,k, it holds that $\mathbb{E}\left(h_{jk}\,\overline{h_{lk}}\,h_{lm}\overline{h_{jm}}\right) = 0$ unless $j = l$ or $k = m$. We therefore obtain

$$\begin{aligned}
\mathbb{E}\left(\lambda^2\right) &= \frac{1}{n^3}\left(\sum_{j,k=1}^n \mathbb{E}\left(|h_{jk}|^4\right) + \sum_{j\; k\neq m} \mathbb{E}\left(|h_{jk}|^2\,|h_{jm}|^2\right) + \sum_{j\neq l,\,k} \mathbb{E}\left(|h_{jk}|^2\,|h_{lk}|^2\right)\right) \\
&= \frac{1}{n^3}\,\left(2n^2 + n^2\,(n-1) + n^2\,(n-1)\right) = 2
\end{aligned}$$

This finally implies that for all $0 \leq t \leq 1$,

$$C_{\text{erg}} \geq n\,\log(1 + Pt)\,\frac{(1-t)^2}{2} = \frac{n}{8}\,\log(1 + P/2)$$

by choosing $t = 1/2$. This is to be compared with the case of a fixed deterministic matrix H_0 whose coefficients are all equal to 1, with corresponding capacity $C_0 = \log(1 + P\,n^2)$ (see last lecture). It holds that $\mathbb{E}(|h_{jk}|^2) = 1 = (H_0)_{jk}$ for all j,k, but notice that:

a) for fixed P and $n \to \infty$, $C_0 \simeq \log n$, while $C_{\text{erg}} \overset{\sim}{\geq} n$.

b) for fixed n and $P \to \infty$, $C_0 \simeq 1 \log P$, while $C_{\text{erg}} \overset{\sim}{\geq} n \log P$

So in multiple antenna systems, random (i.i.d.) fading actually improves the capacity, contrary to single antenna systems.

3 H is random and fixed over time (slow fading)

We assume now that the matrix H admits the continuous pdf $p_H(\cdot)$, whose support contains the all zero matrix, that H is fixed over time, and that its realization is known at the receiver, but not at the transmitter. In this case, the capacity of the channel is zero and the single-letter characterization of its outage probability reads:

$$P_{\text{out}}(R) = \inf_{p_X \,:\, \text{Tr}(Q_X) \leq P} \mathbb{P}_H(I(X;Y) < R)$$

where R is the target rate chosen by the transmitter. As we have already seen, for a any given H,

$$I(X;Y) \leq \log\det(I + H\,Q_X\,H^*)$$

and the equality is met when $X \sim \mathcal{N}_{\mathbb{C}}(0, Q_X)$ (that does not depend on the particular realization of H). So

$$P_{\text{out}}(R) = \inf_{Q_X \geq 0 \,:\, \text{Tr}(Q_X) \leq P} \mathbb{P}_H(\log\det(I + H\,Q_X\,H^*) < R)$$

Let us now consider the particular case where the matrix H has i.i.d entries $h_{jk} \sim \mathcal{N}_{\mathbb{C}}(0,1)$. In this case, H and HU share the same distribution, for any deterministic $n \times n$ unitary matrix U, so we may as well take Q_X diagonal in the above optimization problem. Therefore,

$$P_{\text{out}}(R) = \inf_{\substack{d_1,\ldots,d_n \geq 0 \\ \sum_{j=1}^n d_j \leq P}} \mathbb{P}_H(\log\det(I + HDH^*) < R)$$

where $D = \text{diag}(d_1, \ldots, d_n)$. Solving further this optimization problem turns out to be difficult. Notice first that the answer depends on the target rate R:

- if R is (sufficiently) small, then setting $D = \frac{P}{n} I$ achieves the minimum outage probability, as in this case, the law of large numbers plays for us: $\log\det(I + \frac{P}{n} HH^*)$ is highly likely to be around its average value and therefore to exceed R.

- if R is (sufficiently) large, then the law of large numbers plays on the contrary against us, and it is therefore better in this case to put “all our eggs in one basket”, that is, to use $D = \text{diag}(P, 0, \ldots, 0)$ and hope that the chosen channel is by chance a good one that will prevent the mutual information to fall below the target rate.

Besides, it has been conjectured that in general, the optimal matrix D should be of the form

$$D = \text{diag}\Big(\underbrace{\frac{P}{k}, \ldots, \frac{P}{k}}_{k \text{ times}}, 0, \ldots, 0\Big)$$

where $1 \leq k \leq n$ is some integer parameter. We will see in the course how to analyze further this outage probability, with the help of random matrix theory.

Random matrices and communication systems: WEEK 5

1 Wishart random matrices: joint distribution of the entries

1.1 Complex case

Let H be an $n \times m$ random matrix with i.i.d.$\sim \mathcal{N}_{\mathbb{C}}(0,1)$ entries and let W be the $n \times n$ matrix defined as $W = HH^*$. W is a positive semi-definite matrix, as $x^*Wx = \|H^*x\|^2 \geq 0$ for all $x \in \mathbb{C}^n$. So by the spectral theorem, there exist U $n \times n$ unitary (i.e. $UU^* = I$) and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ such that $\lambda_j \geq 0$ for all j and $W = U\Lambda U^*$. Ultimately, we are interested in computing the joint distribution $p(\lambda_1, \ldots, \lambda_n)$ of the eigenvalues of W (for fixed values of both n and m), as well as its marginals. This however requires a first step, namely to compute the joint distribution of the entries of W, which is the purpose of the present lecture.

Remark. If $m < n$, then $\text{rank}(W) \leq \text{rank}(H) \leq \min(m,n) = m < n$, which is saying the the matrix W is rank-deficient and admits at least $n - m$ zero eigenvalues. So in this case, the joint distribution of $\lambda_1, \ldots, \lambda_n$ is singular. Instead, we may as well consider the $m \times m$ matrix $\widetilde{W} = H^*H$, which has the same non-zero eigenvalues as W, and look for the joint distribution of its eigenvalues $\widetilde{\lambda}_1, \ldots, \widetilde{\lambda}_m$. In the following, we therefore restrict ourselves without loss of generality to the case where $m \geq n$.

Let us now describe the various steps that lead to the joint distribution of the entries of W.

Joint distribution of the entries of $\boldsymbol{H}$**.** This is an easy computation. By independence, we have

$$p_H(H) = \prod_{j,k=1}^{n,m} \frac{1}{\pi} \exp\left(-|h_{jk}|^2\right) = \frac{1}{\pi^{nm}} \exp\left(-\sum_{j,k=1}^{n,m} |h_{jk}|^2\right) = \frac{1}{\pi^{nm}} \exp\left(-\text{Tr}\,(HH^*)\right)$$

LQ **decomposition of** $\boldsymbol{H}$ **and Choleski decomposition of** $\boldsymbol{W}$**.** Let us recall the following fact:

Any $n \times m$ complex-valued matrix H may be decomposed into $H = LQ$, where L is an $n \times n$ lower-triangular matrix such that $l_{jj} \geq 0$ for all j and Q is an $n \times m$ matrix such that $QQ^* = I_n$ (i.e. Q is a submatrix of an $m \times m$ unitary matrix).

Consequently, $W = HH^* = LQQ^*L^* = LL^*$, which is the Cholesky decomposition of W.

The strategy now is, starting from the expression for $p_H(H)$, to compute $p_L(L)$ and then $p_W(W)$.

Joint distribution of the entries of $\boldsymbol{L}$**.** The relation $H = LQ$ can be seen as a change of variables $H \mapsto (L,Q)$. Let us first check how many free real parameters there are on each side of the equality. On the left-hand side, H has clearly $2nm$ free parameters. On the right-hand side,, there are n free real diagonal parameters in L and $2\frac{n(n-1)}{2} = n^2 - n$ free real off-diagonal one; this makes in total n^2 free parameters in L. Regarding Q, there are a priori $2m$ free parameters in each row, but the first row should have unit norm, the second row should also have unit norm *and* be orthogonal to the first one, and so on. This makes in total

$$(2m-1) + (2m-3) + \ldots + (2(m-n)+1) = m^2 - (m-n)^2 = 2mn - n^2 \quad \text{free real parameters in } Q$$

So the number of free real parameters coincide on each side. The joint distribution of L and Q may therefore be written as

$$p_{L,Q}(L,Q) = p_H(LQ)\,|J_1(L,Q)|$$

where $J_1(L,Q)$ is the Jacobian of the transformation $H \mapsto (L,Q)$. The computation of this Jacobian, that we shall skip here, gives

$$J_1(L,Q) = \prod_{j=1}^{n} l_{jj}^{2(m-j)+1}$$

Besides,

$$p_H(LQ) = \frac{1}{\pi^{nm}} \exp\left(-\mathrm{Tr}\left(LQQ^*L^*\right)\right) = \frac{1}{\pi^{nm}} \exp\left(-\mathrm{Tr}\left(LL^*\right)\right)$$

so finally, we obtain

$$P_{L,Q}(L,Q) = \frac{1}{\pi^{nm}} \exp\left(-\mathrm{Tr}\left(LL^*\right)\right) \prod_{j=1}^{n} l_{jj}^{2(m-j)+1} \, 1_{l_{jj}\geq 0}$$

As this expression does not depend explicitly on Q, this says two things: 1) the distribution of Q is uniform over the set of $n \times m$ complex matrices such that $QQ^* = I_n$ (we will come back to this below); 2) L and Q are actually independent! Let us now look more closely at the distribution of L:

$$\begin{aligned} p_L(L) &= c_{n,m} \exp\left(-\mathrm{Tr}\left(LL^*\right)\right) \prod_{j=1}^{n} l_{jj}^{2(m-j)+1} \, 1_{l_{jj}\geq 0} \\ &= c_{n,m} \prod_{j=1}^{n} \left(l_{jj}^{2(m-j)+1} \exp\left(-l_{jj}^2\right) \, 1_{l_{jj}\geq 0} \right) \prod_{k<j} \exp\left(-|l_{jk}|^2\right) \end{aligned}$$

Notice that $c_{n,m}$ above is not equal to $1/\pi^{nm}$, nor is the normalization constant in the uniform distribution $p_Q(Q)$ equal to 1 (the latter is actually equal to $1/V_{n,m}$, where $V_{n,m}$ is the volume of the set of $n \times m$ complex matrices Q such that $QQ^* = I_n$). What the above equality tells us is that:

1) all the entries of L are independent;

2) the off-diagonal entries l_{jk} are i.i.d.$\sim \mathbb{N}_{\mathbb{C}}(0,1)$ random variables;

3) the diagonal entry l_{jj} is a $\chi_{2(m-j+1)}$ random variable.

Remark. Even though we do not need it in the following, let us just make clear what we mean by "uniform" while talking about the distribution of Q. For every fixed $m \times m$ unitary matrix $U \in U(m)$, the matrices H and HU share the same distribution, so the same holds for LQ and LQU. By the independence of L and Q and the non-singularity of the distribution of L, this is saying that Q and QU share the same distribution, for every fixed $m \times m$ unitary matrix U. It is actually a fact that there is only one such distribution, that we call the uniform distribution over the set of $n \times m$ complex matrices Q such that $QQ^* = I_n$.

Joint distribution of the entries of W. We now consider the change of variables $L \mapsto W = LL^*$. Let us compute the number of free real parameters on each side. We have seen above that L contains n^2 free real parameters. The same holds for W, as W contains n diagonal free real parameters and $2\frac{n(n-1)}{2} = n^2 - n$ off-diagonal free real parameters (remembering that W is Hermitian). Considering the reverse transformation $W \mapsto L$ (just because it is easier), we obtain

$$p_L(L) = p_W\left(LL^*\right) \, |J_2(L)|$$

where $J_2(L)$ is the Jacobian of the transformation $W \mapsto L$. The computation of this Jacobian, that we shall again skip here, gives:

$$J_2(L) = 2^n \prod_{j=1}^{n} l_{jj}^{2(n-j)+1}$$

We therefore deduce that

$$p_W(LL^*) = \frac{P_L(L)}{|J_2(L)|} = c_{n,m} \exp\left(-\mathrm{Tr}\left(LL^*\right)\right) \prod_{j=1}^{n} l_{jj}^{2(m-n)}$$

where the constant $c_{n,m}$ differs from the previous one by a factor 2^n, but we keep the same notation for simplicity. Noticing that $LL^* = W$ and $\prod_{j=1}^n l_{jj}^2 = |\det(L)|^2 = \det(LL^*) = \det(W)$, we finally obtain the joint distribution of the entries of W:

$$p_W(W) = c_{n,m} \det(W)^{m-n} \exp(-\mathrm{Tr}(W)) \, 1_{\{W \geq 0\}}$$

Remark. In the case $n = m$, the above distribution reads

$$p_W(W) = c_{n,n} \exp(-\mathrm{Tr}(W)) \, 1_{\{W \geq 0\}}$$

which looks like a particularly simple expression: it indeed says on one hand that the diagonal entries w_{jj} are i.i.d. exponential random variables; on the other hand, the condition $W \geq 0$ induces lots of subtle constraints and dependencies between the off-diagonal entries.

1.2 Real case

The corresponding model in the real case is given by $W = HH^T$, where H is an $n \times m$ random matrix with i.i.d.$\sim \mathcal{N}_{\mathbb{R}}(0,1)$ entries. Without repeating the whole reasoning, let us briefly mention the corresponding result in this case. We again assume that $m \geq n$ without loss of generality. In this case, the joint distribution of the entries of H is given by

$$p_H(H) = \frac{1}{(2\pi)^{nm/2}} \exp\left(\frac{1}{2} \mathrm{Tr}\left(HH^T\right)\right)$$

and the joint distribution of the entries of W is given by

$$p_W(W) = c_{n,m} \det(W)^{\frac{m-n-1}{2}} \exp\left(-\frac{1}{2} \mathrm{Tr}(W)\right) 1_{\{W \geq 0\}}$$

Random matrices and communication systems: WEEK 6

1 Wishart random matrices: joint eigenvalue distribution

1.1 Real case

Recall first from previous lecture that if $W = HH^T$, where H is an $n \times m$ random matrix with i.i.d.$\sim \mathcal{N}_\mathbb{R}(0,1)$ entries and $m \geq n$, then the joint distribution of the entries of W is given by

$$p_W(W) = c_{n,m} \det(W)^{\frac{m-n-1}{2}} \exp\left(-\frac{1}{2}\operatorname{Tr}(W)\right) 1_{\{W \geq 0\}}$$

By the spectral theorem, the matrix W is orthogonally diagonalizable, that is, there exist V an $n \times n$ orthogonal matrix (i.e. $VV^T = I$) and $\Lambda = \operatorname{diag}(\lambda_1, \ldots, \lambda_n)$ such that $\lambda_j \geq 0$ for all $1 \leq j \leq n$ and $W = V\Lambda V^T$, i.e.

$$w_{jk} = \sum_{l=1}^n \lambda_l \, v_{jl} \, v_{kl}, \quad 1 \leq j,k \leq n$$

Again, this can be viewed as a change of variables; on the left-hand side, there are n diagonal free parameters w_{jj} and $\frac{n(n-1)}{2}$ off-diagonal free parameters w_{jk}, $j < k$ in the matrix W (the remaining off-diagonal parameters are fixed, as W is symmetric); on the right-hand side, there are n free parameters in the matrix Λ and $(n-1) + (n-2) + \ldots + 1 + 0 = \frac{n(n-1)}{2}$ free parameters in the matrix V. So the number of free parameters on both sides coincide. The joint distribution of Λ and V is then given by

$$p_{\Lambda,V}(\Lambda, V) = p_W\left(V\Lambda V^T\right) \, |J(\Lambda, V)|$$

where, according to the above formula,

$$\begin{aligned} p_W\left(V\Lambda V^T\right) &= c_{n,m} \det\left(V\Lambda V^T\right)^{\frac{m-n-1}{2}} \exp\left(-\frac{1}{2}\operatorname{Tr}\left(V\Lambda V^T\right)\right) \\ &= c_{n,m} \det(\Lambda)^{\frac{m-n-1}{2}} \exp\left(-\frac{1}{2}\operatorname{Tr}(\Lambda)\right) \end{aligned}$$

and $J(\Lambda, V)$ is the Jacobian of the transformation $W \mapsto (\Lambda, V)$. We now set out to compute this Jacobian. Let $N = \frac{n(n-1)}{2}$ and let us denote by $p_1, \ldots, p_N$ the N free parameters in the matrix V (leaving aside the explicit description of what these N parameters are: we will see in the following that this is actually not needed). The Jacobian is then given by

$$J(\Lambda, V) = \det \left(\begin{array}{c|c} \left\{\frac{\partial w_{jj}}{\partial \lambda_i}\right\}_{i,j} & \left\{\frac{\partial w_{jk}}{\partial \lambda_i}\right\}_{i,j<k} \\ \hline \left\{\frac{\partial w_{jj}}{\partial p_i}\right\}_{i,j} & \left\{\frac{\partial w_{jk}}{\partial p_i}\right\}_{i,j<k} \end{array}\right)$$

where the blocks in the above matrix are respectively of size $n \times n$, $n \times N$, $N \times n$ and $N \times N$. As $W = V\Lambda V^T$, the computation of the above partial derivatives gives, in matrix form:

$$\begin{aligned} \frac{\partial W}{\partial \lambda_i} &= V\Delta^{(i)}V^T, \quad \text{where } \Delta^{(i)}_{jk} = \delta_{ij}\,\delta_{ik} \\ \frac{\partial W}{\partial p_i} &= \frac{\partial V}{\partial p_i}\Lambda V^T + V\Lambda\frac{\partial V^T}{\partial p_i} \end{aligned}$$

Multiplying both these equations by V^T and V and the left-hand and right-hand side, we obtain

$$\begin{aligned} V^T\frac{\partial W}{\partial \lambda_i}V &= \Delta^{(i)} \\ V^T\frac{\partial W}{\partial p_i}V &= \left(V^T\frac{\partial V}{\partial p_i}\right)\Lambda + \Lambda\left(\frac{\partial V^T}{\partial p_i}V\right) \end{aligned}$$

Let now $S^{(i)} = V^T \frac{\partial V}{\partial p_i}$. As $V^T V = I$, we also obtain

$$V^T \frac{\partial V}{\partial p_i} + \frac{\partial V^T}{\partial p_i} V = 0, \quad \text{i.e.} \quad \frac{\partial V^T}{\partial p_i} V = -S^{(i)}$$

which allows us to rewrite $V^T \frac{\partial W}{\partial p_i} V = S^{(i)} \Lambda - \Lambda S^{(i)}$. Component-wise, the two equations for the derivatives with respect to λ_i and p_i therefore read:

$$\begin{cases} \displaystyle\sum_{l,m=1}^{n} \frac{\partial w_{lm}}{\partial \lambda_i} \, v_{lj} \, v_{mk} = \delta_{ij} \, \delta_{ik} \\ \displaystyle\sum_{l,m=1}^{n} \frac{\partial w_{lm}}{\partial p_i} \, v_{lj} \, v_{mk} = S^{(i)}_{jk} \, (\lambda_k - \lambda_j) \end{cases} \tag{1}$$

With the help of these formulas, let us now compute the Jacobian $J(\Lambda, V)$ when $V = I$. In this case, the above two formulas boil down to

$$\frac{\partial w_{jk}}{\partial \lambda_i} = \delta_{ij} \, \delta_{ik} \quad \text{and} \quad \frac{\partial w_{jk}}{\partial p_i} = S^{(i)}_{jk} \, (\lambda_k - \lambda_j)$$

so

$$\begin{aligned} J(\Lambda, V) &= \det \left(\begin{array}{c|c} I & 0 \\ \hline 0 & \left\{ S^{(i)}_{jk} \, (\lambda_k - \lambda_j) \right\}_{i,j<k} \end{array} \right) = \det \left(\left\{ S^{(i)}_{jk} \, (\lambda_k - \lambda_j) \right\}_{i,j<k} \right) \\ &= \prod_{j<k} (\lambda_k - \lambda_j) \det \left(\left\{ S^{(i)}_{jk} \right\}_{i,j<k} \right) = \prod_{j<k} (\lambda_k - \lambda_j) \, f(V) \end{aligned}$$

for some function f, as the matrix elements $S^{(i)}_{jk}$ possibly only depend on V. We claim that the same conclusion holds in the case where $V \neq I$. To this end, let us consider

$$\widetilde{J}(\Lambda, V) = \det \left(\left(\begin{array}{c|c} \left\{ \frac{\partial w_{ll}}{\partial \lambda_i} \right\}_{i,l} & \left\{ \frac{\partial w_{lm}}{\partial \lambda_i} \right\}_{i,l<m} \\ \hline \left\{ \frac{\partial w_{ll}}{\partial p_i} \right\}_{i,l} & \left\{ \frac{\partial w_{jk}}{\partial p_i} \right\}_{i,l<m} \end{array} \right) \left(\begin{array}{c|c} \left\{ v_{lj}^2 \right\}_{l,j} & \left\{ v_{lj} v_{lk} \right\}_{l,j<k} \\ \hline \left\{ 2 v_{lj} v_{mj} \right\}_{l<m,j} & \left\{ 2 v_{lj} v_{mk} \right\}_{l<m,j<k} \end{array} \right) \right)$$

Using the fact that $\det(AB) = \det(A) \det(B)$ and observing that the second term on the right-hand side only depends on V, we deduce that $\widetilde{J}(\Lambda, V) = J(\Lambda, V) \, g(V)$ for some function g. On the other hand, performing the matrix multiplication inside the determinant gives for matrix element $i, (jk)$ in the first n rows:

$$\sum_{l=1}^{n} \frac{\partial w_{ll}}{\partial \lambda_i} \, v_{lj} \, v_{lk} + 2 \sum_{l<m} \frac{\partial w_{lm}}{\partial \lambda_i} \, v_{lj} \, v_{mk} = \sum_{l,m=1}^{n} \frac{\partial w_{lm}}{\partial \lambda_i} \, v_{lj} \, v_{mk}$$

and likewise for the matrix element $i, (jk)$ in the last N rows:

$$\sum_{l=1}^{n} \frac{\partial w_{ll}}{\partial p_i} \, v_{lj} \, v_{lk} + 2 \sum_{l<m} \frac{\partial w_{lm}}{\partial p_i} \, v_{lj} \, v_{mk} = \sum_{l,m=1}^{n} \frac{\partial w_{lm}}{\partial p_i} \, v_{lj} \, v_{mk}$$

Using then equation (1), we obtain again

$$\widetilde{J}(\Lambda, V) = \det \left(\begin{array}{c|c} I & 0 \\ \hline 0 & \left\{ S^{(i)}_{jk} \, (\lambda_k - \lambda_j) \right\}_{i,j<k} \end{array} \right) = \prod_{j<k} (\lambda_k - \lambda_j) \, f(V)$$

which, together with the above observation that $\widetilde{J}(\Lambda, V) = J(\Lambda, V) \, g(V)$, proves the claim. What can be deduced so far from all these computations is that

$$p_{\Lambda,V}(\Lambda, V) = c_{n,m} \det(\Lambda)^{\frac{m-n-1}{2}} \exp\left(-\frac{1}{2} \operatorname{Tr}(\Lambda) \right) \prod_{j<k} |\lambda_k - \lambda_j| \, |f(V)|$$

for some function f. This is actually saying that $p_{\Lambda,V}(\Lambda, V) = p_\Lambda(\Lambda)\, p_V(V)$, so the eigenvalues and eigenvectors of W are independent! The joint distribution of the eigenvalues is given by

$$p(\lambda_1, \ldots, \lambda_n) = c_{n,m} \prod_{j=1}^{n} \left(\lambda_j^{\frac{m-n-1}{2}} \exp(-\lambda_j/2)\, 1_{\lambda_j \geq 0} \right) \prod_{j<k} |\lambda_k - \lambda_j|$$

where $c_{n,m}$ is the normalization constant, which can be computed explicitly; it differs from the constant in the expression for $p_W(W)$, but in order to keep notation simple, we do not change notation here.

The above distribution may also be rewritten in the following form:

$$p(\lambda_1, \ldots, \lambda_n) = c_{n,m} \exp\left(-\sum_{j=1}^{n} \left(\frac{\lambda_j}{2} - \frac{m-n-1}{2} \log(\lambda_j) \right) + \sum_{j<k} \log|\lambda_k - \lambda_j| \right) 1_{\lambda_1 \geq 0, \ldots, \lambda_n \geq 0}$$

and given the following interpretation: it represents the Gibbs distribution of a system of n particles in positions $\lambda_1, \ldots, \lambda_n$ evolving in a potential $U(\lambda) = \frac{\lambda}{2} - \frac{m-n-1}{2} \log(\lambda)$ and repelling each other. Two opposite forces operate here: on one hand, the particles would all like to be in the minimum of the potential, but as they repel each other, there is not enough room for them, so some are driven away from this minimum. The fact that eigenvalues repel each other is a common feature to most random matrix models (essentially because of the term resulting from the above Jacobian computation).

Now, what is the distribution of the eigenvectors? As already seen, for every fixed $n \times n$ orthogonal matrix $O \in O(n)$, the matrices H and OH share the same distribution. Therefore, so do the matrices W and OWO^T, which is saying that

$$V \Lambda V^T \quad \text{and} \quad (OV)\Lambda(OV)^T$$

also share the same distribution. By the independence of Λ and V (and the non-singularity of the distribution of Λ), this finally implies that V and OV share the same distribution, for every fixed $n \times n$ orthogonal matrix O. This in turn implies that the matrix V is distributed according to the Haar distribution on $O(n)$, which is the unique distribution on $O(n)$ being invariant under orthogonal transformations.

1.2 Complex case

We shall not repeat here the whole reasoning; we just mention the main steps of the computation. In this case, $W = HH^*$, where H is an $n \times m$ random matrix with i.i.d.$\sim \mathcal{N}_\mathbb{C}(0,1)$ entries and $m \geq n$. The joint distribution of the entries of W is given by

$$p_W(W) = c_{n,m} \det(W)^{m-n} \exp(-\text{Tr}(W))\, 1_{\{W \geq 0\}}$$

By the spectral theorem, the matrix W is unitarily diagonalizable, that is, there exist U an $n \times n$ unitary matrix (i.e. $UU^* = I$) and $\Lambda = \text{diag}(\lambda_1, \ldots, \lambda_n)$ such that $\lambda_j \geq 0$ for all $1 \leq j \leq n$ and $W = U\Lambda U^*$, i.e.

$$w_{jk} = \sum_{l=1}^{n} \lambda_l\, u_{jl}\, \overline{u_{kl}}, \quad 1 \leq j, k \leq n$$

On the left-hand side, there are n diagonal free real parameters and $2\, \frac{n(n-1)}{2} = n^2 - n$ off-diagonal free real parameters in the matrix W; on the right-hand side, there are n free real parameters in the matrix Λ and n^2 free real parameters in the matrix U. So here we see that there is a mismatch. The mismatch can be resolved by observing that in the complex case, an eigenvector rotated by $e^{i\phi}$ remains an eigenvector. So it is possible to set the first component of each eigenvector of W to be a real number, which reduces by n the numbers of free real parameters in U, so that the number of free real parameters on both sides coincide. The joint distribution of Λ and U is then given by

$$p_{\Lambda,U}(\Lambda, U) = p_W\left(U\Lambda U^*\right)\, |J(\Lambda, U)|$$

where

$$p_W\left(U\Lambda U^*\right) = c_{n,m}\det(\Lambda)^{m-n}\,\exp(-\mathrm{Tr}(\Lambda))$$

and the computation of the Jacobian gives

$$J(\Lambda, U) = \prod_{j<k}(\lambda_k - \lambda_j)^2\, f(U)$$

for some function f. Therefore, Λ and U are also independent in this case, and

$$p(\lambda_1, \ldots, \lambda_n) = c_{n,m}\prod_{j=1}^{n}\left(\lambda_j^{m-n}\,\exp(-\lambda_j)\,1_{\lambda_j \geq 0}\right)\prod_{j<k}(\lambda_k - \lambda_j)^2$$

where $c_{n,m}$ is the normalization constant, which differs from the previous $c_{n,m}$ and can be computed explicitly. Finally, similarly to the previous section, U is distributed according to the Haar distribution on $U(n)$, which is the unique distribution on $U(n)$ being invariant under unitary transformations.

Random matrices and communication systems: WEEK 7

1 Wishart random matrices: marginal eigenvalue distribution

We only consider here the complex case, as the analysis of the real case is sensibly more difficult than what is presented below. Let H be an $n \times n$ matrix with i.i.d.$\sim \mathcal{N}_{\mathbb{C}}(0,1)$ entries and $W = HH^*$. We have seen in the previous lecture that the joint distribution of the eigenvalues $\lambda_1, \ldots, \lambda_n$ of W is given by

$$p(\lambda_1, \ldots, \lambda_n) = c_n \prod_{j=1}^{n} e^{-\lambda_j} \prod_{j<k} (\lambda_k - \lambda_j)^2$$

where c_n is a normalization constant and where we have dropped the term "$1_{\lambda_j \geq 0}$" in the above expression in order to lighten the notation. Notice also that compared to the previous lecture, we consider here only the case $n = m$. This is meant to simplify the exposition in the sequel, but contrary to the above mentioned real case, the case $m \geq n$ can be handled with little extra effort.

Let us first give a possible reason for studying marginals of this joint distribution. The above expression allows to compute the expectation of a function of the eigenvalues $f(\lambda_1, \ldots, \lambda_n)$:

$$\mathbb{E}(f(\lambda_1, \ldots, \lambda_n)) = \int_0^\infty d\lambda_1 \cdots \int_0^\infty d\lambda_n \, p(\lambda_1, \ldots, \lambda_n) \, f(\lambda_1, \ldots, \lambda_n)$$

An example of such function is $f(\lambda_1, \ldots, \lambda_n) = \prod_{j=1}^n \lambda_j$ (which corresponds to $f(W) = \det(W)$).

If one wants now to compute the expectation of a funtion of the form $f(\lambda_1, \ldots, \lambda_n) = \sum_{j=1}^n g(\lambda_j)$, for some function g, it is of course possible to write

$$\mathbb{E}\left(\sum_{j=1}^n g(\lambda_j)\right) = \sum_{j=1}^n \int_0^\infty d\lambda_1 \cdots \int_0^\infty d\lambda_n \, p(\lambda_1, \ldots, \lambda_n) \, g(\lambda_j)$$

But notice that this may also be rewritten as

$$\mathbb{E}\left(\sum_{j=1}^n g(\lambda_j)\right) = \sum_{j=1}^n \int_0^\infty d\lambda_j \, p(\lambda_j) \, g(\lambda_j)$$

where

$$p(\lambda_j) = \int_0^\infty d\lambda_1 \cdots \int_0^\infty d\lambda_{j-1} \int_0^\infty d\lambda_{j+1} \cdots \int_0^\infty d\lambda_n \, p(\lambda_1, \ldots, \lambda_n)$$

are the first-order marginals of p. Notice in addition that the distribution $p(\lambda_1, \ldots, \lambda_n)$ is symmetric in any permutation of the λ's, so we may as well consider the eigenvalues $\lambda_1, \ldots \lambda_n$ as unordered. In this case, all the above marginals are the same, so the expression for the expectation boils down to

$$\mathbb{E}\left(\sum_{j=1}^n g(\lambda_j)\right) = n \int_0^\infty d\lambda \, p(\lambda) \, g(\lambda)$$

$p(\lambda)$ may be also interpreted here as the distribution of one of the eigenvalues $\lambda_1, \ldots, \lambda_n$ picked uniformly at random. An example where this formula applies is when $g(\lambda) = \log(\lambda)$, which corresponds to $f(\lambda_1, \ldots, \lambda_n) = \sum_{j=1}^n \log(\lambda_j)$, which corresponds in turn to $f(W) = \log\det(W)$.

Computation of the marginals. The first step for the computation of the marginal $p(\lambda)$ is to use Vandermonde's determinant formula:

$$\prod_{j<k} (\lambda_k - \lambda_j) = \det \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_n \\ \vdots & \vdots & & \vdots \\ \lambda_1^{n-1} & \lambda_2^{n-1} & \cdots & \lambda_n^{n-1} \end{pmatrix}$$

Next, we introduce *Laguerre polynomials*: these are defined as

$$L_k(\lambda) = \frac{1}{k!}\, e^{\lambda}\, \frac{d^k}{d\lambda^k}\left(e^{-\lambda}\, \lambda^k\right) \quad \text{where } k \in \mathbb{N},\ \lambda \geq 0$$

So $L_0(\lambda) = 1$, $L_1(\lambda) = 1 - \lambda$, $L_2(\lambda) = \frac{1}{2}(\lambda - 2)^2 - 1$, and so on. In general, L_k is a polynomial of degree k in λ that may be written as

$$L_k(\lambda) = \gamma_k\, \lambda^k + \text{ lower order terms}, \quad \text{where } \gamma_k = \frac{(-1)^k}{k!} \neq 0$$

One can check in addition that these polynomials satisfy in the following *orthogonality relations*:

$$\int_0^\infty d\lambda\, e^{-\lambda}\, L_k(\lambda)\, L_l(\lambda) = \delta_{kl} \tag{1}$$

We now use these Laguerre polynomials to rewrite the above determinant (using the basic rules for the determinant) as

$$\det \begin{pmatrix} 1 & 1 & \cdots & 1 \\ \lambda_1 & \lambda_2 & \cdots & \lambda_n \\ \vdots & \vdots & & \vdots \\ \lambda_1^{n-1} & \lambda_2^{n-1} & \cdots & \lambda_n^{n-1} \end{pmatrix} = \left(\prod_{j=1}^{n-1} \frac{1}{\gamma_j}\right) \det \begin{pmatrix} L_0(\lambda_1) & L_0(\lambda_2) & \cdots & L_0(\lambda_n) \\ L_1(\lambda_1) & L_1(\lambda_2) & \cdots & L_1(\lambda_n) \\ \vdots & \vdots & & \vdots \\ L_{n-1}(\lambda_1) & L_{n-1}(\lambda_2) & \cdots & L_{n-1}(\lambda_n) \end{pmatrix}$$

In order to simplify notation, let us rewrite the matrix on the right-hand side as $\{L_{j-1}(\lambda_k)\}_{j,k=1}^n$. Combining everything together, we obtain

$$p(\lambda_1, \ldots, \lambda_n) = c_n \prod_{j=1}^n e^{-\lambda_j} \left(\prod_{j=1}^{n-1} \frac{1}{\gamma_j}\right)^2 \det\left(\{L_{j-1}(\lambda_k)\}_{j,k=1}^n\right)^2$$

Noticing that the product of γ's is yet another constant, we may simply include it in the constant c_n. Also, the above expression may be transformed into

$$p(\lambda_1, \ldots, \lambda_n) = c_n \prod_{j=1}^n e^{-\lambda_j} \det\left(\left(\{L_{j-1}(\lambda_k)\}_{j,k=1}^n\right)^T \{L_{j-1}(\lambda_k)\}_{j,k=1}^n\right) = c_n \det\left(\{K(\lambda_j, \lambda_k\}_{j,k=1}^n\right)$$

where $K(\lambda, \mu) = e^{-\frac{\lambda+\mu}{2}} \sum_{l=0}^{n-1} L_l(\lambda)\, L_l(\mu)$.

Remarks. - Notice that even though we are talking about the eigenvalues of the complex-valued matrix W, this matrix is Hermitian (and even positive semi-definite), so its eigenvalues are real (we therefore only have a transpose matrix above, and not a complex-conjugate transpose).

- The above trick of writing $\det(A)^2 = \det(A^T A)$ is of course made possible because of the presence of the square in the expression for the joint eigenvalue distribution. In the real case, the square is missing, which makes things much more delicate (one can always write $|\det(A)| = \sqrt{\det(A^T A)}$, but the square root creates problems later).

The Kernel K has the following nice properties, that follow from the orthogonality relations (1) for the Laguerre polynomials.

Lemma 1.1. a) $K(\mu, \lambda) = K(\lambda, \mu)$ b) $\int_0^\infty d\lambda\, K(\lambda, \lambda) = n$ c) $\int_0^\infty d\mu\, K(\lambda, \mu)\, K(\mu, \nu) = K(\lambda, \nu)$

The last property above is known as the "self-reproducing property" of the Kernel K.

Proof. a) is obvious. b) $\int_0^\infty d\lambda\, K(\lambda, \lambda) = \sum_{l=0}^{n-1} \int_0^\infty d\lambda\, e^{-\lambda}\, L_l(\lambda)^2 = n.$

c) $\int_0^\infty d\mu\, K(\lambda, \mu)\, K(\mu, \nu) = e^{-\frac{\lambda+\nu}{2}}\, L_l(\lambda)\, L_m(\nu) \sum_{l,m=0}^{n-1} \int_0^\infty d\mu\, e^{-\mu}\, L_l(\mu) L_m(\mu) = K(\lambda, \nu).$ □

From these properties follows the remarkable fact below, known as *Mehta's lemma.*

Lemma 1.2. The m^{th} order marginal of the joint eigenvalue distribution $p(\lambda_1, \ldots, \lambda_n)$ is given by

$$p(\lambda_1, \ldots, \lambda_m) = \frac{(n-m)!}{n!} \det\left(\{K(\lambda_j, \lambda_k\}_{j,k=1}^m\right)$$

Proof. We first give the proof for the case $m = n-1$ (the general case follows then easily). We are interested in computing

$$p(\lambda_1, \ldots, \lambda_{n-1}) = c_n \int_0^\infty d\lambda_n \det\left(\{K(\lambda_j, \lambda_k\}_{j,k=1}^n\right)$$

In order to lighten the notation, let us write $A = \{a_{jk}\}_{j,k=1}^n = \{K(\lambda_j, \lambda_k)\}_{j,k=1}^n$. Using the expansion formula for the determinant, we obtain

$$p(\lambda_1, \ldots, \lambda_{n-1}) = c_n \int_0^\infty d\lambda_n \det(A) = c_n \sum_{l=1}^n (-1)^{n+l} \int_0^\infty d\lambda_n \, a_{nl} \det(A(n,l))$$

where $A(n,l)$ denotes the matrix A with n^{th} row and l^{th} column suppressed. This may be further rewritten as

$$p(\lambda_1, \ldots, \lambda_{n-1}) = c_n \left(\int_0^\infty d\lambda_n \, a_{nn}\right) \det(A(n,n)) + c_n \sum_{l=1}^{n-1} (-1)^{n+l} \int_0^\infty d\lambda_n \, a_{nl} \det(A(n,l)) \qquad (2)$$

where

$$\int_0^\infty d\lambda_n \, a_{nn} = \int_0^\infty d\lambda_n \, K(\lambda_n, \lambda_n) = n$$

by property b) of Lemma 1.1, Furthermore, we have for all $1 \leq l \leq n-1$,

$$\begin{aligned} & \int_0^\infty d\lambda_n \, a_{nl} \det(A(n,l)) \\ = & \int_0^\infty d\lambda_n \, a_{nl} \det \begin{pmatrix} a_{1,1} & \cdots & a_{1,l-1} & a_{1,l+1} & \cdots & a_{1,n-1} & a_{1,n} \\ \vdots & & \vdots & \vdots & & \vdots & \vdots \\ a_{n-1,1} & \cdots & a_{n-1,l-1} & a_{n-1,l+1} & \cdots & a_{n-1,n-1} & a_{n-1,n} \end{pmatrix} \\ = & \det \begin{pmatrix} a_{1,1} & \cdots & a_{1,l-1} & a_{1,l+1} & \cdots & a_{1,n-1} & \int_0^\infty d\lambda_n \, a_{1,n} \, a_{n,l} \\ \vdots & & \vdots & \vdots & & \vdots & \vdots \\ a_{n-1,1} & \cdots & a_{n-1,l-1} & a_{n-1,l+1} & \cdots & a_{n-1,n-1} & \int_0^\infty d\lambda_n \, a_{n-1,n} a_{n,l} \end{pmatrix} \end{aligned} \qquad (3)$$

Indeed, the integral can be brought inside the determinant for the following reason: writing down the full expansion formula for the determinant in (3), we see that each term only involves one occurrence of λ_n. We further have

$$\int_0^\infty d\lambda_n \, a_{jn} \, a_{nl} = \int_0^\infty d\lambda_n \, K(\lambda_j, \lambda_n) \, K(\lambda_n, \lambda_l) = K(\lambda_j, \lambda_l) = a_{jl}$$

by property c) of Lemma 1.1. So up to a column permutation, the matrix on the right-hand side of the above expression is $A(n,n)$, which leads to

$$\int_0^\infty d\lambda_n \, a_{nl} \det(A(n,l)) = (-1)^{n-l-1} \det(A(n,n))$$

Inserting this in equation (2) finally gives

$$\begin{aligned} p(\lambda_1, \ldots, \lambda_{n-1}) &= c_n \left(n \det(A(n,n)) + \sum_{l=1}^n (-1)^{2n-1} \det(A(n,n)) \right) \\ &= c_n \, (n - (n-1)) \det(A(n,n)) = c_n \det\left(\{K(\lambda_j, \lambda_k\}_{j,k=1}^{n-1}\right) \end{aligned}$$

which proves the claim for $m = n - 1$. A similar reasoning shows that

$$p(\lambda_1, \ldots, \lambda_{n-2}) = c_n \, (n - (n-2)) \det \left(\{K(\lambda_j, \lambda_k\}_{j,k=1}^{n-2} \right) = 2 \, c_n \det \left(\{K(\lambda_j, \lambda_k\}_{j,k=1}^{n-2} \right)$$

and more generally

$$p(\lambda_1, \ldots, \lambda_m) = (n-m)! \, c_n \det \left(\{K(\lambda_j, \lambda_k\}_{j,k=1}^{m} \right)$$

Finally, in order to compute the normalization constant, notice that for $m = 1$,

$$p(\lambda_1) = (n-1)! \, c_n \, K(\lambda_1, \lambda_1)$$

As we know that $\int_0^\infty d\lambda_1 \, p(\lambda_1) = 1$, this together with property b) of Lemma 1.1 implies that $n! \, c_n = 1$, i.e. $c_n = \frac{1}{n!}$, which completes the proof of the lemma. □

A word on the asymptotic analysis of the eigenvalue distribution. A particular instance of the above result is of course the case $m = 1$, which gives the first-order marginal:

$$p(\lambda) = \frac{1}{n} \, e^{-\lambda} \sum_{l=0}^{n-1} L_l(\lambda)^2$$

As mentioned above, this distribution represents the distribution of a "typical" eigenvalue of W in the "bulk" of the spectrum. Analyzing directly the behavior of $p(\lambda)$ for large values of n is not an easy task. First of all, let us mention that in order to obtain convergence, a rescaling is needed. Indeed:

$$\mathbb{E}(\lambda) = \mathbb{E}\left(\frac{1}{n} \sum_{j=1}^{n} \lambda_j \right) = \mathbb{E}\left(\frac{1}{n} \mathrm{Tr}(W) \right) = \frac{1}{n} \, \mathbb{E}(\mathrm{Tr}(HH^*)) = \frac{1}{n} \sum_{j,k=1}^{n} \mathbb{E}(|h_{jk}|^2) = n$$

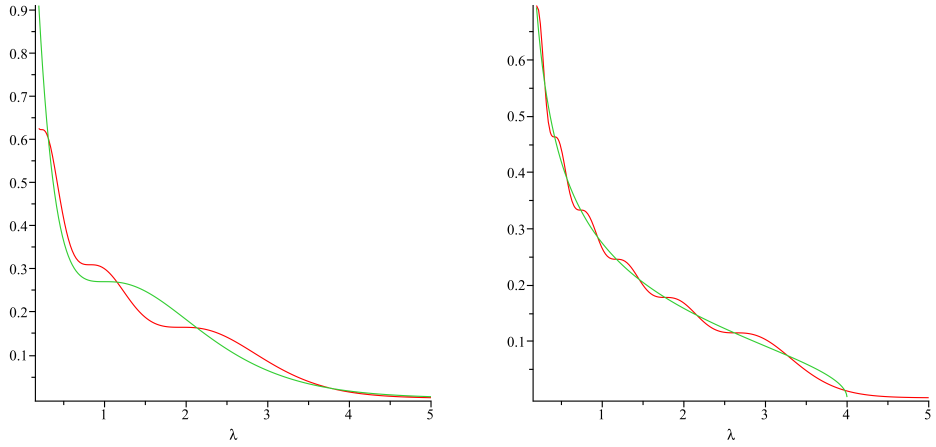
We should then rather consider $p^{(n)}(\lambda)$, the distribution of an eigenvalue of $\frac{1}{n} W$ picked uniformly at random:

$$p^{(n)}(\lambda) = n \, p(n\lambda) = e^{-n\lambda} \sum_{l=0}^{n-1} L_l(n\lambda)^2$$

Studying the asymptotic behavior of this expression as n gets large requires the knowledge of fine properties of Laguerre polynomials, that we skip here. The result gives (we are going to recover this result using a different approach later in the course):

$$\lim_{n \to \infty} p^{(n)}(\lambda) = \frac{1}{\pi} \sqrt{\frac{1}{\lambda} - \frac{1}{4}} \, 1_{0 < \lambda < 4}$$

which is illustrated on the figures below. On the left, $p^{(n)}(\lambda)$ is represented for $n = 2$ and $n = 4$, while on the right, it is represented for $n = 8$ and $n = \infty$ (spreading the rumor that for random matrices, $8 \sim \infty$):



Random matrices and communication systems: WEEK 8

The goal of this lecture is to introduce the notion of *asymptotic eigenvalue distribution*. In particular, we will see that this notion already appears for large *deterministic* matrices, such as the classical Toeplitz matrices. We first need a preliminary on a particular type of Toeplitz matrices, whose eigenvalues are particularly easy to compute; these are *circulant* matrices.

1 Circulant matrices

Let $c_0, c_1, \ldots, c_{n-1}$ be complex numbers and C be the $n \times n$ matrix defined as

$$C = \begin{pmatrix} c_0 & c_1 & \ddots & c_{n-2} & c_{n-1} \\ c_{n-1} & c_0 & c_1 & \ddots & c_{n-2} \\ \ddots & \ddots & \ddots & \ddots & \ddots \\ c_2 & \ddots & c_{n-1} & c_0 & c_1 \\ c_1 & c_2 & \ddots & c_{n-1} & c_0 \end{pmatrix}$$

Notation. $C = \mathrm{circ}(c_0, c_1, \ldots, c_{n-1})$.

The eigenvalues and eigenvectors C can be computed as follows.

Lemma 1.1. Let $\alpha \in \mathbb{C}$ be such that $\alpha^n = 1$ (i.e. α is an n^{th} root of unity). Then the vector u defined as

$$u = \begin{pmatrix} \alpha \\ \alpha^2 \\ \vdots \\ \alpha^n \end{pmatrix} \text{ is an eigenvector of } C, \text{ with corresponding eigenvalue } \lambda = \sum_{l=0}^{n-1} c_l \, \alpha^l$$

Proof. One has to check that $Cu = \lambda u$. Indeed, for all $1 \le j \le n$, we have

$$\begin{aligned} (Cu)_j &= \sum_{k=1}^{n} c_{jk} \, u_k = \sum_{k=1}^{j-1} c_{n-j+k} \, \alpha^k + \sum_{k=j}^{n} c_{k-j} \, \alpha^k = \sum_{l=n-j+1}^{n-1} c_l \, \alpha^{l+j-n} + \sum_{l=0}^{n-j} c_l \, \alpha^{l+j} \\ &\overset{(a)}{=} \sum_{l=0}^{n-1} c_l \, \alpha^{l+j} = \left(\sum_{l=0}^{n-1} c_l \, \alpha^l \right) \alpha^j = \lambda \, u_j \end{aligned}$$

where we have used the fact that $\alpha^n = 1$ in (a). □

Let now ($\alpha_k = \exp(2\pi i k/n)$, $k = 1, \ldots, n$) denote the n different roots of unity. One can check that the corresponding eigenvectors (u_k, $k = 1, \ldots, n$) are mutually orthogonal. As a consequence, we have the following proposition.

Proposition 1.2. For any values of $c_0, c_1, \ldots, c_{n-1}$, there exist $\Lambda = \mathrm{diag}(\lambda_1, \ldots, \lambda_n)$ and U $n \times n$ unitary with $C = U \Lambda U^*$. In addition, we have

$$\lambda_k = \sum_{l=0}^{n-1} c_l \, \exp(2\pi i l k/n) \quad \text{and} \quad u_{jk} = \frac{1}{\sqrt{n}} \, (u_k)_j = \frac{1}{\sqrt{n}} \, \exp(2\pi i j k/n)$$

Important consequences and remarks. - The above proposition says that all circulant matrices are *unitarily diagonalizable* (even though they may not be Hermitian).

- More than that, the matrix U of eigenvectors of C (called the Discrete Fourier Transform (DFT) matrix) does not depend on the values of the numbers $c_0, c_1, \ldots, c_{n-1}$, Therefore, *all the circulant matrices share the same set of eigenvectors.*

- As a consequence, if C is circulant, then C^* is, and $CC^* = C^*C$, i.e. C is normal (in accordance with the above proposition). Also, if C is circulant and invertible, then C^{-1} is circulant.

- More generally, if C_1, C_2 are circulant, then $C_1 + C_2$ and C_1C_2 are, with eigenvalues respectively given by $\lambda_k^{(sum)} = \lambda_k^{(1)} + \lambda_k^{(2)}$ and $\lambda_k^{(prod)} = \lambda_k^{(1)}\,\lambda_k^{(2)}$. Such nice rules for sums and products of matrices are the exception.

- Finally, only the eigenvalues of C depend on the values of $c_0, c_1, \ldots, c_{n-1}$ and this dependence is linear. This is in sharp contrast to arbitrary matrices for which the dependence of the eigenvalues on the entries is intricate in general.

2 Toeplitz matrices and the Grenander-Szegö theorem

Let l_0 be a fixed positive integer and let $(t_l,\ -l_0 \leq l \leq l_0)$ be given complex numbers. We consider now the $n \times n$ matrix $T^{(n)}$ defined as

$$(T^{(n)})_{jk} = \begin{cases} t_{k-j} & \text{if } |j-k| \leq l_0 \\ 0 & \text{if } |j-k| > l_0 \end{cases} \quad \text{that is:} \quad T^{(n)} = \begin{pmatrix} t_0 & t_1 & \ddots & t_{l_0} & 0 & \cdots & 0 \\ t_{-1} & t_0 & t_1 & \ddots & t_{l_0} & 0 & \vdots \\ \ddots & t_{-1} & t_0 & t_1 & \ddots & \ddots & 0 \\ t_{-l_0} & \ddots & \ddots & \ddots & \ddots & \ddots & t_{l_0} \\ 0 & \ddots & \ddots & t_{-1} & t_0 & t_1 & \ddots \\ \vdots & 0 & t_{-l_0} & \ddots & t_{-1} & t_0 & t_1 \\ 0 & \cdots & 0 & t_{-l_0} & \ddots & t_{-1} & t_0 \end{pmatrix}$$

Such a matrix is called a *finite-order Toeplitz matrix.* Let $\lambda_1^{(n)}, \ldots, \lambda_n^{(n)}$ denote its eigenvalues. Our aim in the following is to characterize the behavior of these eigenvalues as n gets large.

Remark 2.1. Contrary to circulant matrices, there is no general formula at finite n for the eigenvalues of $T^{(n)}$ in terms of the numbers t_l. Also, the matrix of eigenvectors of $T^{(n)}$ is not the DFT matrix.

Let now us define the following function:

$$g(x) = \sum_{l=-l_0}^{l_0} t_l\, e^{ilx}, \quad x \in [0, 2\pi]$$

g is a complex-valued, bounded and continuous function, and the numbers t_l are the Fourier coefficients of g:

$$t_l = \frac{1}{2\pi} \int_0^{2\pi} g(x)\, e^{-ilx}\, dx$$

The following lemma establishes a first connection between the eigenvalues of $T^{(n)}$ and the function g.

Lemma 2.2. For any $m \geq 0$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^{n} \left(\lambda_k^{(n)}\right)^m = \frac{1}{2\pi} \int_0^{2\pi} (g(x))^m\, dx$$

Proof. (sketch: the whole proof is left as an exercise in the homework) Let us consider the matrices (represented here in the case $l_0 = 1$ for simplicity)

$$T^{(n)} = \begin{pmatrix} t_0 & t_1 & 0 & \cdots & 0 \\ t_{-1} & t_0 & t_1 & 0 & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & 0 & t_{-1} & t_0 & t_1 \\ 0 & \cdots & 0 & t_{-1} & t_0 \end{pmatrix} \quad \text{and} \quad C^{(n)} = \begin{pmatrix} t_0 & t_1 & 0 & \cdots & t_{-1} \\ t_{-1} & t_0 & t_1 & 0 & \vdots \\ 0 & \ddots & \ddots & \ddots & 0 \\ \vdots & 0 & t_{-1} & t_0 & t_1 \\ t_1 & \cdots & 0 & t_{-1} & t_0 \end{pmatrix}$$

Let us denote by $\lambda_1^{(n)}, \ldots, \lambda_n^{(n)}$ and $\mu_1^{(n)}, \ldots, \mu_n^{(n)}$ the eigenvalues of $T^{(n)}$ and $C^{(n)}$, respectively. These two matrices can be shown to be *asymptotically equivalent*, which implies that for all $m \geq 0$,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^n \left(\lambda_k^{(n)}\right)^m = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^n \left(\mu_k^{(n)}\right)^m$$

Moreover, notice that $C^{(n)}$ is circulant, so by Section 1, its eigenvalues are given by (for $n \geq 2l_0 + 1$)

$$\mu_k^{(n)} = \sum_{l=-l_0}^{l_0} t_l \exp(2\pi i k l/n) = g(2\pi k/n)$$

Therefore,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^n \left(\lambda_k^{(n)}\right)^m = \lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^n (g(2\pi k/n))^m = \frac{1}{2\pi} \int_0^{2\pi} (g(x))^m \, dx$$

by definition of Riemann's integral of the continuous function g^m over the interval $[0, 2\pi]$. □

We need now the following assumption.

Assumption H. For all $-l_0 \leq l \leq l_0$, $t_{-l} = \overline{t_l}$.

This assumption implies both that $T^{(n)}$ is Hermitian for all values of n (so its eigenvalues $\lambda_1^{(n)}, \ldots, \lambda_n^{(n)}$ are real), and that the function g is real-valued. Moreover we have the following lemma, whose proof is again left as an exercise in the homework.

Lemma 2.3. Under Assumption H, let $a = \inf_{x\in[0,2\pi]} g(x)$ and $b = \sup_{x\in[0,2\pi]} g(x)$. Then for all $n \geq 1$ and all $1 \leq k \leq n$,

$$a \leq \lambda_k^{(n)} \leq b$$

This finally allows us to state the main theorem.

Theorem 2.4. (Grenander-Szegö, 1958 - Gray, 1972)
Under Assumption H, we have for any continuous function $f : [a, b] \to \mathbb{R}$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{k=1}^n f\left(\lambda_k^{(n)}\right) = \frac{1}{2\pi} \int_0^{2\pi} f(g(x)) \, dx$$

Proof. Notice first that because of Lemma 2.3, $\lambda_k^{(n)} \in [a, b]$ for all k, n, so the expression on the left-hand side is well defined (and so is the one on the right-hand side, as by definition, $a \leq g(x) \leq b$ for all $x \in [0, 2\pi]$). Next, we see that Lemma 2.2 proves the theorem for $f(y) = y^m$, for any $m \geq 0$. By linearity of the integral, this relation can be extended to any polynomial of the form $f(y) = P_m(y) = \sum_{j=1}^m c_j \, y^j$. The final step is made possible by Weierstrass' theorem, stating that any continuous function f on $[a, b]$ can be uniformly approached by a sequence of polynomials (P_m, $m \geq 0$), i.e.

$$\lim_{m\to\infty} \sup_{y\in[a,b]} |f(y) - P_m(y)| = 0$$

By the triangle inequality, we now have for any $m \geq 0$

$$\begin{aligned}
\left|\frac{1}{n}\sum_{k=1}^{n} f\left(\lambda_k^{(n)}\right) - \frac{1}{2\pi}\int_0^{2\pi} f(g(x))\,dx\right| &\leq \left|\frac{1}{n}\sum_{k=1}^{n} f\left(\lambda_k^{(n)}\right) - \frac{1}{n}\sum_{k=1}^{n} P_m\left(\lambda_k^{(n)}\right)\right| \\
&+ \left|\frac{1}{n}\sum_{k=1}^{n} P_m\left(\lambda_k^{(n)}\right) - \frac{1}{2\pi}\int_0^{2\pi} P_m(g(x))\,dx\right| + \left|\frac{1}{2\pi}\int_0^{2\pi} P_m(g(x))\,dx - \frac{1}{2\pi}\int_0^{2\pi} f(g(x))\,dx\right| \\
\leq\ & 2\sup_{y\in[a,b]}|f(y) - P_m(y)| + \left|\frac{1}{n}\sum_{k=1}^{n} P_m\left(\lambda_k^{(n)}\right) - \frac{1}{2\pi}\int_0^{2\pi} P_m(g(x))\,dx\right| \underset{n\to\infty}{\to} 2\sup_{y\in[a,b]}|f(y) - P_m(y)|
\end{aligned}$$

Taking finally the limit $m \to \infty$ allows to conclude. □

Remark 2.5. Without Assumption H, the theorem fails. Consider for example the sequence $(t_l,\, l \in \mathbb{Z})$ with $t_l = \delta_{l1}$ (i.e. $t_1 = 1$ and all other $t_l = 0$). Then

$$T^{(n)} = \begin{pmatrix} 0 & 1 & 0 & \cdots & 0 \\ 0 & 0 & 1 & 0 & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ \vdots & & 0 & 0 & 1 \\ 0 & \cdots & \cdots & 0 & 0 \end{pmatrix}$$

so all its eigenvalues $\lambda_k^{(n)} = 0$, but the function $g(x) = e^{itx}$ so in general

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} f\left(\lambda_k^{(n)}\right) = f(0) \neq \frac{1}{2\pi}\int_0^{2\pi} f\left(e^{itx}\right)\,dx$$

In order for this relation to hold, f should be a polynomial, or more generally, an analytic function on $\mathbb{C}$. The reason why the theorem fails in this case is that Weierstrass' theorem does not hold for continuous functions on $\mathbb{C}$.

Remark 2.6. On the other hand, the theorem can be generalized to Toeplitz matrices constructed from an infinite sequence $(t_l,\, l \in \mathbb{Z})$ satisfying Assumption H and the following additional condition:

$$\sum_{l\in\mathbb{Z}} |t_l| < \infty$$

Under these assumptions, it still holds that g is a real-valued, bounded and continuous function, but the proof of Lemma 2.2 becomes slightly more involved.

3 Asymptotic eigenvalue distribution

Theorem 2.4 may be rephrased as a weak convergence result for a sequence of distributions. To this end, let us consider the function $f_t(y) = 1_{y\leq t}$, where $t \in \mathbb{R}$. f_t is not a continuous function, but it can be approximated by a sequence of continuous functions (although not uniformly, which might sometimes create problems), implying that for almost all $t \in \mathbb{R}$ (details to follow below)

$$\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} f_t\left(\lambda_k^{(n)}\right) = \frac{1}{2\pi}\int_0^{2\pi} f_t(g(x))\,dx$$

The above line may be rephrased as

$$\lim_{n\to\infty}\frac{1}{n}\sharp\{1 \leq k \leq n \ :\ \lambda_k^{(n)} \leq t\} = \frac{1}{2\pi}\left|\{x \in [0, 2\pi] \ :\ g(x) \leq t\}\right|$$

where $\sharp A$ denotes the number of elements in the discrete set A and $|B|$ denotes the Lebesgue measure of $B \in \mathcal{B}(\mathbb{R})$. Let now

$$F_n(t) = \frac{1}{n} \sharp \{1 \leq k \leq n \; : \; \lambda_k^{(n)} \leq t\} \quad \text{and} \quad F(t) = \frac{1}{2\pi} \left| \{x \in [0, 2\pi] \; : \; g(x) \leq t\} \right|$$

These are cumulative distribution functions (or simply "distributions"): F_n is the distribution of an eigenvalue $\lambda^{(n)}$ picked uniformly at random among the eigenvalues $\lambda_1^{(n)}, \ldots, \lambda_n^{(n)}$ of $T^{(n)}$, while F is called the *asymptotic eigenvalue distribution* of the sequence of matrices $T^{(n)}$ [1]. The precise mathematical statement is now that the sequence F_n converges weakly towards F, i.e. that

$$\lim_{n \to \infty} F_n(t) = F(t) \quad \forall t \in \mathbb{R} \text{ continuity point of } F$$

The rigorous proof of this statement follows from Carleman's theorem and the following observation: the moments of F_n are given by

$$\int_{\mathbb{R}} y^m \, dF_n(y) = \mathbb{E}\left(\left(\lambda^{(n)} \right)^m \right) = \frac{1}{n} \sum_{k=1}^{n} \left(\lambda_k^{(n)} \right)^m \underset{n \to \infty}{\to} \frac{1}{2\pi} \int_0^{2\pi} (g(x))^m \, dx$$

for all $m \geq 0$, by Lemma 2.2. As g is a bounded function, these moments satisfy Carleman's condition, which implies the weak convergence of the sequence F_n towards the unique distribution F whose moments are given by the above expression. Notice that Assumption H is implicitly used here, as it ensures that the distributions are supported on the real line (for distributions on the complex plane, convergence of moments alone does not guarantee weak convergence of the corresponding distributions).

Example 3.1. Let $l_0 = 1$, $t_0 = 2$ and $t_1 = t_{-1} = -1$. In this case,

$$g(x) = 2 - e^{itx} - e^{-itx} = 2(1 - \cos(x)) = 4 \sin^2(x/2)$$

so $a = 0$ and $b = 4$ here. The corresponding asymptotic eigenvalue distribution F is given by

$$F(t) = \frac{1}{2\pi} \left| \{x \in [0, 2\pi] \; : \; g(x) \leq t \} \right.$$

An easy way to compute F is to observe that for any continuous function $f : [0, 4] \to \mathbb{R}$, we should have

$$\int_0^4 f(y) \, dF(y) = \frac{1}{2\pi} \int_0^{2\pi} f\left(4 \sin^2(x/2)\right) \, dx = \frac{1}{\pi} \int_0^{\pi} f\left(4 \sin^2(x/2)\right) \, dx$$

Performing the change of variable $y = 4 \sin^2(x/2)$, we obtain

$$dy = 4 \, \sin(x/2) \, \cos(x/2) \, dx \quad \text{i.e.} \quad dx = \frac{1}{\sqrt{y(4-y)}} \, dy$$

so finally,

$$\int_0^4 f(y) \, dF(y) = \int_0^4 f(y) \, \frac{1}{\pi} \, \frac{1}{\sqrt{y(4-y)}} \, dy$$

i.e. F admits the pdf

$$p(y) = \frac{1}{\pi} \, \frac{1}{\sqrt{y(4-y)}} \, 1_{]0,4[}(y)$$

[1] F may also be interpreted as the distribution of an eigenvalue of the infinite-dimensional operator $T^{(\infty)}$ picked uniformly at random, whatever that means...

Random matrices and communication systems: WEEK 9

In this lecture, we first give a quick reminder regarding distributions on the real line; we then recall the notion of weak convergence of sequences of distributions and finally move to various characterizations of this weak convergence, more particularly in terms of moments and Stieltjes transform.

1 Distributions without random variables

1.1 Distributions on the real line

Let $\mathcal{B}(\mathbb{R})$ be the *Borel σ-field on* $\mathbb{R}$, that is, the smallest σ-field on $\mathbb{R}$ that contains all the open sets in $\mathbb{R}$; its elements $B \in \mathcal{B}(\mathbb{R})$ are called the *Borel sets.*[1] Recall that a mapping $f : \mathbb{R} \to \mathbb{R}$ is said to be *Borel-measurable* if for all $B \in \mathcal{B}(\mathbb{R})$, $f^{-1}(B) = \{x \in \mathbb{R} \ : \ f(x) \in B\} \in \mathcal{B}(\mathbb{R})$. In particular, any continuous function is Borel-measurable.[2]

Definition 1.1. A (probability) distribution on $\mathbb{R}$ is a mapping $\mu : \mathcal{B}(\mathbb{R}) \to [0, 1]$ such that

$$\mu(\emptyset) = 0, \quad \mu(\mathbb{R}) = 1 \quad \text{and} \quad \text{if } (B_n,\, n \geq 1) \in \mathcal{B}(\mathbb{R}) \text{ are disjoint, then } \mu\left(\bigcup_{n\geq 1} B_n\right) = \sum_{n\geq 1} \mu(B_n)$$

Definition 1.2. The cumulative distribution function (cdf) associated to a distribution μ is the mapping $F_\mu : \mathbb{R} \to [0, 1]$ defined as

$$F_\mu(t) = \mu((-\infty, t]), \quad t \in \mathbb{R}$$

Fact. The knowledge of the cdf F_μ is equivalent to that of the distribution μ.

There are two well known particular classes of distributions.

Discrete distributions, for which there exists a countable set C such that $\mu(C) = 1$. In this case,

$$\mu(B) = \sum_{x\in B\cap C} \mu(\{x\}) \quad \forall B \in \mathcal{B}(\mathbb{R})$$

and F_μ is a step function.

Continuous distributions, for which there exists a probability density fiunction (pdf) $p_\mu : \mathbb{R} \to \mathbb{R}_+$ such that

$$\mu(B) = \int_B p_\mu(x)\, dx \quad \forall B \in \mathcal{B}(\mathbb{R})$$

In this case, F_μ is a continuous function.

1.2 Lebesgue's integral

The Lebesgue integral of a Borel-measurable function $f : \mathbb{R} \to \mathbb{R}$ with respect to a distribution μ is defined in three steps as follows.

1. First suppose f is of the form

$$f(x) = \sum_{j\geq 1} y_j\, 1_{B_j}(x), \quad \text{where } y_j \geq 0 \text{ and } B_j \in \mathcal{B}(\mathbb{R}) \tag{1}$$

Then the integral is defined as

$$\int_{\mathbb{R}} f(x)\, d\mu(x) = \sum_{j\geq 1} y_j\, \mu(B_j)$$

[1] If one is not familiar with this notion, one may think of $\mathcal{B}(\mathbb{R})$ as being simply the set of (nearly!) all subsets of $\mathbb{R}$.

[2] Again, one may simply consider that (nearly!) all functions are Borel-measurable.

2. Next, suppose that f is Borel-measurable and non-negative (i.e. $f(x) \geq 0$ for all $x \in \mathbb{R}$). Define then for $n \geq 1$

$$f_n(x) = \sum_{j \geq 1} \frac{j-1}{2^n} 1_{\left\{x \in \mathbb{R} \,:\, \frac{j-1}{2^n} \leq f(x) < \frac{j}{2^n}\right\}}(x), \quad x \in \mathbb{R}$$

Then one can check that for all $n \geq 1$ and $x \in \mathbb{R}$, $f_n(x) \leq f_{n+1}(x)$ as well as $\lim_{n\to\infty} f_n(x) = f(x)$. As f_n is of the form (1), we may define

$$\int_{\mathbb{R}} f(x)\, d\mu(x) = \lim_{n\to\infty} \int_{\mathbb{R}} f_n(x)\, d\mu(x) = \lim_{n\to\infty} \sum_{j\geq 1} \frac{j-1}{2^n}\, \mu\left(\left\{x \in \mathbb{R} \,:\, \tfrac{j-1}{2^n} \leq f(x) < \tfrac{j}{2^n}\right\}\right)$$

Notice that as $f_n \leq f_{n+1}$, the above sequence is increasing, so the limit always exists, but may take the value $+\infty$.

3. Assume now that f is any Borel-measurable function. In this case case, we say that the integral is well defined only if

$$\int_{\mathbb{R}} |f(x)|\, d\mu(x) < \infty \tag{2}$$

and we set

$$\int_{\mathbb{R}} f(x)\, d\mu(x) = \int_{\mathbb{R}} f^+(x)\, d\mu(x) - \int_{\mathbb{R}} f^-(x)\, d\mu(x)$$

where $f^+(x) = \max(f(x), 0)$ and $f^-(x) = \max(-f(x), 0)$.

It is worth noticing that condition (2) is satisfied for any distribution μ when f is Borel-measurable and *bounded*, as

$$\int_{\mathbb{R}} |f(x)|\, d\mu(x) \leq \sup_{x\in\mathbb{R}} |f(x)| \int_{\mathbb{R}} d\mu(x) = \sup_{x\in\mathbb{R}} |f(x)|\, \mu(\mathbb{R}) = \sup_{x\in\mathbb{R}} |f(x)| < \infty$$

by assumption. In particular, let us consider, for a given $t \in \mathbb{R}$, $f_t(x) = 1_{\{x \leq t\}}$: f_t is Borel-measurable and bounded, and

$$\int_{\mathbb{R}} f_t(x)\, d\mu(x) = \int_{-\infty}^{t} d\mu(x) = \mu((-\infty, t]) = F_\mu(t)$$

For discrete and continuous distributions, the Lebesgue integral simply reads:

For μ discrete, $\int_{\mathbb{R}} f(x)\, d\mu(x) = \sum_{x \in C} f(x)\, \mu(\{x\})$. **For μ continuous,** $\int_{\mathbb{R}} f(x)\, d\mu(x) = \int_{\mathbb{R}} f(x)\, p_\mu(x)\, dx$.

1.3 Objects associated to a distribution

The cdf is an example of object associated to a distribution (that moreover characterizes completely the distribution). Here are other examples.

Moments.

Definition 1.3. Let μ be a distribution on $\mathbb{R}$ and $k \geq 0$. If $\int_{\mathbb{R}} |x|^k\, d\mu(x) < \infty$, we then define the *moment of order k* associated to the distribution μ as

$$m_k = \int_{\mathbb{R}} x^k\, d\mu(x)$$

Here are some easy facts:

- As $f(x) = x^k$ is not a bounded function, the moment of order k of a distribution is not always well defined (with the exception of discrete distributions supported on a finite set: all their all moments are always finite).

- If μ has a finite moment of order k, then all its moments of lower order $l \leq k$ are also finite. In general, there is a limiting value k_0 below which all moments are finite and above which all moments are infinite (but k_0 may of course take the value ∞).

- If there exists $C > 0$ such that $\mu([-C, C]) = 1$ (we then say that μ is supported on a compact set), then all the moments of μ are finite and

$$|m_k| \leq \int_{\mathbb{R}} |x|^k \, d\mu(x) = \int_{-C}^{C} |x|^k \, d\mu(x) \leq C^k \int_{-C}^{C} d\mu(x) = C^k$$

There are of course other examples of distributions which are not supported on a compact set and whose moments are all finite (such as e.g. the Gaussian or the log-normal distributions).

In general, an important question is to decide whether a distribution is completely characterized by its moments (which can only possibly happen when all the moments of the distribution are finite). The answer is given by Carleman's theorem.

Theorem 1.4. (Carleman) Let μ be a distribution and $(m_k,\, k \geq 0)$ be the sequence of its moments. If in addition

$$\sum_{k \geq 1} (m_{2k})^{-\frac{1}{2k}} = \infty \tag{3}$$

then the distribution μ is the unique distribution with the sequence of moments $(m_k,\, k \geq 0)$.

Condition (3) is actually a condition on the growth of the moments m_k. It is satisfied in particular if $|m_k| \leq C^k$ for some $C > 0$ (which occurs e.g. for distributions supported on a compact set, as just seen above). Indeed, in this case,

$$m_{2k} \leq C^{2k} \quad \text{so} \quad (m_{2k})^{-\frac{1}{2k}} \geq \frac{1}{C}, \quad \text{so} \quad \sum_{k \geq 1} (m_{2k})^{-\frac{1}{2k}} = \infty$$

More generally, if $|m_k| \leq C \, \exp(k \log k)$, then condition (3) holds. This is the case for example for the Gaussian distribution (in which case $m_k \sim k!$), but not for the log-normal distribution (in which case $m_k \sim \exp(k^2)$)

Stieltjes (or Cauchy) transform.

Definition 1.5. Let μ be a distribution on $\mathbb{R}$ and $z \in \mathbb{C} \backslash \mathbb{R}$. The *Stieltjes transform of* μ is the mapping $g_\mu : \mathbb{C} \backslash \mathbb{R} \to \mathbb{C}$ defined as

$$g_\mu(z) = \int_{\mathbb{R}} \frac{1}{x - z} \, d\mu(z), \quad z \in \mathbb{C} \backslash \mathbb{R}$$

Notice that for $z \in \mathbb{C} \backslash \mathbb{R}$, the function $x \mapsto f_z(x) = \frac{1}{x-z}$ is bounded and continuous on the real line, so $g_\mu(z)$ is always well defined.

Basic properties.

- g_μ is analytic on $\mathbb{C} \backslash \mathbb{R}$
- $\text{Im } g_\mu(z) > 0$ for all $z \in \mathbb{C}$ such that $\text{Im } z > 0$
- $\lim_{v \to \infty} v \, |g_\mu(iv)| = 1$

Moreover, it turns out that any function g satisfying the above three properties is the Stieltjes transform of a distribution μ on $\mathbb{R}$. In addition, we have the following *inversion formula*:

If $a < b$ are continuity points of F_μ, then

$$F_\mu(b) - F_\mu(a) = \lim_{\varepsilon \downarrow 0} \frac{1}{\pi} \int_a^b \text{Im } g_\mu(u + i\varepsilon) \, du$$

In the case where μ is a continuous distribution with pdf p_μ, the above formula simplifies to

$$p_\mu(x) = \lim_{\varepsilon \downarrow 0} \frac{1}{\pi} \text{Im } g_\mu(x + i\varepsilon) \quad \forall x \in \mathbb{R}$$

2 Weak convergence of sequences of distributions

Definition 2.1. Let $(\mu_n,\, n \geq 1)$ be a sequence of distributions and μ be another distribution. The sequence μ_n is said to converge weakly to μ as n goes to infinity if

$$\lim_{n\to\infty} F_{\mu_n}(t) = F_\mu(t) \quad \forall t \in \mathbb{R} \text{ continuity point of } F_\mu$$

Notation. $\mu_n \underset{n\to\infty}{\Longrightarrow} \mu$.

So weak convergence of distributions means pointwise convergence of the corresponding cdfs, except in the points where the limiting cdf makes a jump[3]. This definition has several equivalents, among which the following, which is part of the so-called portmanteau theorem.

Proposition 2.2. $\mu_n \underset{n\to\infty}{\Longrightarrow} \mu$ if and only if for every bounded continuous function $f : \mathbb{R} \to \mathbb{R}$,

$$\lim_{n\to\infty} \int_{\mathbb{R}} f(x)\, d\mu_n(x) = \int_{\mathbb{R}} f(x)\, d\mu(x)$$

Checking either of these criteria in order to prove weak convergence is difficult in general. In the sequel, we propose other criteria that are easier to apply, in particular in random matrix theory.

2.1 Various characterizations of weak convergence

Via moments. The full version of Carleman's theorem is given below.

Theorem 2.3. (Carleman) Let $(\mu_n,\, n \geq 1)$ be a sequence of distributions and $(m_k,\, k \geq 0)$ be a sequence of numbers such that for all $k \geq 0$,

$$\lim_{n\to\infty} \int_{\mathbb{R}} x^k d\mu_n(x) = m_k$$

and such that the sequence $(m_k,\, k \geq 0)$ satisfies condition (3). Then $(m_k,\, k \geq 0)$ is a sequence of moments to which corresponds a unique distribution μ, and μ_n converges weakly to μ.

Via Stieltjes transform.

Theorem 2.4. Let $(\mu_n,\, n \geq 1)$ be a sequence of distributions and μ be another distribution. Then $\mu_n \underset{n\to\infty}{\Longrightarrow} \mu$ if and only if

$$\lim_{n\to\infty} g_{\mu_n}(z) = g_\mu(z)$$

for all $z \in \mathbb{C}_+ = \{z \in \mathbb{C} : \text{Im } z > 0\}$.

The above two criteria are very useful for (random) matrices, for the following reason. Let $A^{(n)}$ be an $n \times n$ Hermitian matrix, let $\lambda_1^{(n)}, \ldots, \lambda_n^{(n)}$ be its real eigenvalues and let μ_n be the distribution of one of these eigenvalues picked at random. Then

$$m_k^{(n)} = \int_{\mathbb{R}} x^k \, d\mu_n(x) = \frac{1}{n} \sum_{j=1}^n \left(\lambda_j^{(n)}\right)^k = \frac{1}{n} \operatorname{Tr}\left(\left(A^{(n)}\right)^k\right)$$

and

$$g_{\mu_n}(z) = \int_{\mathbb{R}} \frac{1}{x - z}\, d\mu_n(x) = \frac{1}{n} \sum_{j=1}^n \frac{1}{\lambda_j^{(n)} - z} = \frac{1}{n} \operatorname{Tr}\left(\left(A^{(n)} - zI\right)^{-1}\right)$$

In order to compute these quantities, one does actually not need to know what the eigenvalues $\lambda_j^{(n)}$ are!

[3] Notice indeed that asking for pointwise convergence of a sequence of functions towards a limiting discontinuous function in every point of $\mathbb{R}$ would be asking for too much.

Random matrices and communication systems: WEEK 10

In this lecture, we first introduce the Catalan numbers and then state and prove the Wigner theorem (a slightly modified version of the original theorem, actually).

1 Preliminary: the Catalan numbers

The Catalan numbers are defined as follows:

$$c_k = \frac{1}{k+1}\binom{2k}{k} = \frac{(2k)!}{k!(k+1)!}, \quad k \geq 0$$

So the sequence starts as $1, 1, 2, 5, 14, 42, ...$ These numbers have multiple combinatorial interpretations. Let us give simply one here. We consider paths that evolve in discrete time over the integer numbers. These paths go either up or down by one unit in one time step. We are interested in the number of paths that start from 0 and go back to 0 in $2k$ time steps, *without hitting the negative numbers* in the interval. Such paths are called *Dyck paths* and are illustrated on Figure 1 below.

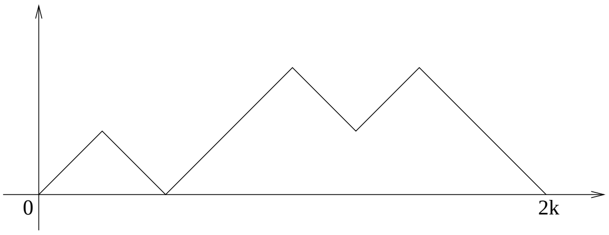


Figure 1. Dyck path

It turns out that for a given k, the number of Dyck paths of length $2k$ is equal to the Catalan number c_k. Here is the proof, also known as the *reflection principle.*

Proof. Let us first make the following trivial observation: the number of Dyck paths of length $2k$ is equal to the *total* number of paths from $(0, 0)$ to $(2k, 0)$, *minus* the number of paths from $(0, 0)$ to $(2k, 0)$ *that do hit the negative numbers* at least once in the interval.

For each of these paths hitting the negative numbers at least once, let us now define T as the first time the number -1 is hit by the path, as illustrated on Figure 2 below. From T onwards, we can draw a mirror path with respect to the horizontal axis with vertical coordinate -1, that necessarily lands in position -2 at time $2k$ (the mirror position of 0 with respect to -1).

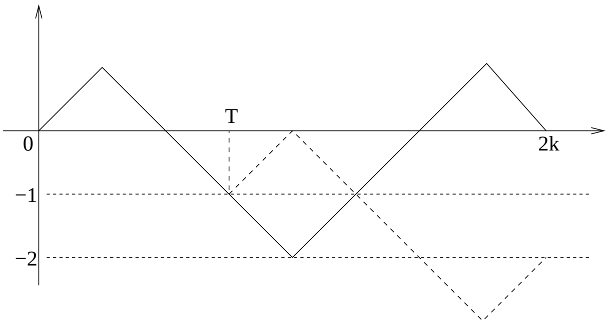


Figure 2. Reflection principle

Counting therefore the number of paths from $(0,0)$ to $(2k,0)$ hitting the negative numbers at least once is the same as counting the number of paths from $(0,0)$ to $(2k,-2)$ hitting the negative numbers at least once. But any such path *must* hit the negative numbers at some point, so the number we are computing is simply the *total* number of paths from $(0,0)$ to $(2k,-2)$.

Finally, we obtain that the number of Dyck paths of length $2k$ is equal to the total number of paths from $(0,0)$ to $(2k,0)$ minus the total number of paths from $(0,0)$ to $(2k,-2)$, which is equal to

$$\binom{2k}{k} - \binom{2k}{k-1} = \frac{(2k)!}{(k!)^2} - \frac{(2k)!}{(k-1)!\,(k+1)!} = \frac{(2k)!}{(k!)^2}\left(1 - \frac{k}{k+1}\right) = \frac{1}{k+1}\binom{2k}{k} = c_k$$

□

The Catalan numbers also have the following interpretation. Let μ be the distribution with pdf

$$p_\mu(x) = \frac{1}{\pi}\sqrt{\frac{1}{x} - \frac{1}{4}}\, 1_{\{0 < x < 4\}} \tag{1}$$

Then its moments are given by

$$m_k = \int_0^4 x^k p_\mu(x)\, dx = \frac{1}{k+1}\binom{2k}{k} = c_k$$

The proof is left as an exercise in the homework. Notice that as the distribution μ is compactly supported on the interval $[0,4]$, we know that

$$|m_k| \le 4^k \quad \text{(this can also be deduced directly from the definition of } c_k\text{)}$$

So the moments m_k satisfy Carleman's condition seen in the last lecture.

Remark 1.1. The above distribution μ is sometimes called the *quarter circle law.* Although this terminology is not really appropriate for such a distribution (drawing $p_\mu(x)$ as a function of x, we do not see a quarter circle...), here is the reason for this denomination. Say μ is the distribution of a (positive) random variable X. Then the distribution ν of $\sqrt{X}$ has the following pdf:

$$p_\nu(y) = p_\mu(y^2)\, 2y = \frac{1}{\pi}\sqrt{\frac{1}{y^2} - \frac{1}{4}}\, 2y\, 1_{\{0 < y < 2\}} = \frac{1}{\pi}\sqrt{4 - y^2}\, 1_{\{0 < y < 2\}} \tag{2}$$

which indeed has the form of a quarter circle.

2 Wigner's theorem

Let H be an $n \times n$ random matrix with i.i.d. complex-valued entries such that for all $1 \le j, l \le n$:

(i) $\mathbb{E}\left(|h_{jl}|^2\right) = 1$;

(ii) $\mathbb{E}\left(|h_{jl}|^k\right) < \infty$ for all $k \ge 0$;

(iii) $\mathbb{E}\left(h_{jl}^k\, \overline{h_{jl}}^{k'}\right) = 0$ if $k \neq k'$.

Notice that these last two assumptions are satisfied in particular when

(ii') the distribution of h_{jl} is compactly supported;

(iii') the distribution of h_{jl} is circularly symmetric.

Also, assumptions (i)-(iii) are satisfied when h_{jl} are i.i.d.$\sim \mathcal{N}_{\mathbb{C}}(0,1)$ random variables.

Let us now consider the (rescaled) Wishart random matrix $W^{(n)} = \frac{1}{n} HH^*$; this matrix is positive semi-definite, so it is unitarily diagonalizable and its eigenvalues $\lambda_1^{(n)}, \ldots, \lambda_n^{(n)}$ are non-negative. Let also

$$\mu_n = \frac{1}{n} \sum_{j=1}^n \delta_{\lambda_j^{(n)}} \quad \text{i.e.} \quad \mu_n(B) = \frac{1}{n} \sharp\{1 \leq j \leq n \ : \ \lambda_j^{(n)} \in B\}, \quad B \in \mathcal{B}(\mathbb{R})$$

μ_n is called the *empirical eigenvalue distribution* of the matrix $W^{(n)}$. Notice that it is a *random* distribution, because for each n, the eigenvalues $\lambda_1^{(n)}, \ldots, \lambda_n^{(n)}$ are random. For a given realization of the $\lambda^{(n)}$'s, μ_n may still be interpreted as the distribution of one of these eigenvalues picked uniformly at random.

Wigner's theorem is then the following.

Theorem 2.1. Under assumptions (i)-(iii), almost surely, the sequence $(\mu_n,\, n \geq 1)$ converges weakly towards the (deterministic) distribution μ, whose pdf is given by (1).

Before starting with the proof of this theorem, let us mention the following immediate corollary. Let $\sigma_1^{(n)}, \ldots, \sigma_n^{(n)}$ be the singular values of the matrix $H^{(n)} = \frac{1}{\sqrt{n}} H$. As $W^{(n)} = H^{(n)} \, (H^{(n)})^*$, it holds that $\sigma_j^{(n)} = \sqrt{\lambda_j^{(n)}}$. Let also

$$\nu_n = \frac{1}{n} \sum_{j=1}^n \delta_{\sigma_j^{(n)}}$$

The above theorem can now be rephrased as:

Corollary 2.2. Under assumptions (i)-(iii), almost surely, the sequence $(\nu_n,\, n \geq 1)$ converges weakly towards the (deterministic) distribution ν, whose pdf is given by (2).

Proof of Theorem 2.1. In order to prove the result, we will use Carleman's theorem, which requires us to show that almost surely,

$$m_k^{(n)} = \int_{\mathbb{R}} x^k d\mu_n(x) \underset{n \to \infty}{\to} c_k \quad \forall k \geq 0 \tag{3}$$

where c_k are the Catalan numbers, that is, the moments of the distribution μ. In the sequel, we show that

$$\left| \mathbb{E}\left(m_k^{(n)}\right) - c_k \right| = O\left(\frac{1}{n}\right) \quad \forall k \geq 0 \tag{4}$$

Using similar methods (involving slightly more combinatorics, though), it can also be shown that

$$\operatorname{Var}\left(m_k^{(n)}\right) = O\left(\frac{1}{n^2}\right) \quad \forall k \geq 0 \tag{5}$$

The last two lines imply (3). Indeed, for all $\varepsilon > 0$,

$$\begin{aligned} \sum_{n\geq 1} \mathbb{P}\left(\left|m_k^{(n)} - c_k\right| > \varepsilon\right) \quad &\leq \quad \frac{1}{\varepsilon^2} \sum_{n \geq 1} \mathbb{E}\left(\left(m_k^{(n)} - c_k\right)^2\right) = \frac{1}{\varepsilon^2} \sum_{n\geq 1} \mathbb{E}\left(\left(m_k^{(n)} - \mathbb{E}\left(m_k^{(n)}\right) + \mathbb{E}\left(m_k^{(n)}\right) - c_k\right)^2\right) \\ &\leq \quad \frac{2}{\varepsilon^2} \sum_{n \geq 1} \left(\operatorname{Var}\left(m_k^{(n)}\right) + \left(\mathbb{E}\left(m_k^{(n)}\right) - c_k\right)^2 \right) < \infty \end{aligned}$$

as both terms in the series are $O(1/n^2)$ by (4) and (5). The Borel-Cantelli lemma allows then to conclude that for all $\varepsilon > 0$,

$$\mathbb{P}\left(\left|m_k^{(n)} - c_k\right| > \varepsilon \text{ infinitely often}\right) = 0$$

which is saying that $m_k^{(n)}$ converges almost surely towards c_k as $n \to \infty$.

We now set out to prove (4). Let us first develop

$$\begin{aligned}
\mathbb{E}\left(m_k^{(n)}\right) &= \mathbb{E}\left(\int_{\mathbb{R}} x^k \, d\mu_n(x)\right) = \mathbb{E}\left(\frac{1}{n}\sum_{j=1}^{n}\left(\lambda_j^{(n)}\right)^k\right) = \mathbb{E}\left(\frac{1}{n}\mathrm{Tr}\left(\left(W^{(n)}\right)^k\right)\right) \\
&= \frac{1}{n^{k+1}}\,\mathbb{E}\left(\mathrm{Tr}\left((HH^*)^k\right)\right) = \frac{1}{n^{k+1}}\sum_{j_1,l_1,\ldots,j_k,l_k=1}^{n} \mathbb{E}\left(h_{j_1,l_1}\,\overline{h_{j_2,l_1}}\cdots h_{j_k,l_k}\,\overline{h_{j_1,l_k}}\right) \qquad (6)
\end{aligned}$$

Let us first look at the cases $k = 1$ and $k = 2$ for simplicity. For $k = 1$, we have

$$\mathbb{E}\left(m_1^{(n)}\right) = \frac{1}{n^2}\sum_{j,l=1}^{n}\mathbb{E}\left(|h_{jl}|^2\right) = 1$$

and for $k = 2$, because of assumption (iii) on the matrix entries h_{jl}, as well as the independence assumption, we have

$$\begin{aligned}
\mathbb{E}\left(m_2^{(n)}\right) &= \frac{1}{n^3}\sum_{j_1,j_2,l_1,l_2=1}^{n}\mathbb{E}\left(h_{j_1,l_1}\,\overline{h_{j_2,l_1}}h_{j_2,l_2}\,\overline{h_{j_1,l_2}}\right) \\
&= \frac{1}{n^3}\left(\sum_{j,l=1}^{n}\mathbb{E}\left(|h_{jl}|^4\right) + \sum_{j,l_1\neq l_2}\mathbb{E}\left(|h_{j,l_1}|^2\,|h_{j,l_2}|^2\right) + \sum_{j_1\neq j_2,l}\mathbb{E}\left(|h_{j_1,l}|^2\,|h_{j_2,l}|^2\right)\right) \\
&= \frac{1}{n^3}\left(O\left(n^2\right) + n^2\,(n-1) + n^2\,(n-1)\right) = 2 + O\left(\frac{1}{n}\right)
\end{aligned}$$

which proves the claim for these two cases. Notice that in the second case, there are a priori n^4 terms in the sum, but only $O\left(n^3\right)$ terms bring a non-zero contribution to the overall sum.

The remainder of the proof is left to the next lecture.

Remark 2.3. Notice that equation (5), which is not proven here, could seem a priori quite unexpected. It indeed says that

$$\mathrm{Var}\left(m_k^{(n)}\right) = \mathrm{Var}\left(\frac{1}{n}\sum_{j=1}^{n}\left(\lambda_j^{(n)}\right)^k\right) = O\left(\frac{1}{n^2}\right)$$

In case the $\lambda^{(n)}$'s were n i.i.d. random variables, one would rather expect this variance to be $O(1/n)$. But the eigenvalues of a random matrix are far from being i.i.d. in general, as already observed when computing their joint distribution in Lecture 6 in the Gaussian case. They are actually n random variables built from a matrix with $O(n^2)$ i.i.d. entries, which is in concordance with the fact that the above variance is $O(1/n^2)$.

Random matrices and communication systems: WEEK 11

1 End of the proof of Wigner's theorem

Let us first recall equation (6) from last lecture:

$$\mathbb{E}\left(m_k^{(n)}\right) = \frac{1}{n^{k+1}} \sum_{j_1,l_1,\ldots,j_k,l_k=1}^{n} \mathbb{E}\left(h_{j_1,l_1}\,\overline{h_{j_2,l_1}}\cdots h_{j_k,l_k}\,\overline{h_{j_1,l_k}}\right) \tag{1}$$

Our aim in the following is to deduce from there that

$$\left|\mathbb{E}\left(m_k^{(n)}\right) - c_k\right| = O\left(\frac{1}{n}\right) \tag{2}$$

where c_k is the k^{th} Catalan number. For a given $k \geq 0$, we need to find out which of the terms in (1) bring a non-negligible contribution to this expression in the large n limit. Observe first that to each term

$$\mathbb{E}\left(h_{j_1,l_1}\,\overline{h_{j_2,l_1}}\cdots h_{j_k,l_k}\,\overline{h_{j_1,l_k}}\right)$$

corresponds a sequence $(j_1, l_1, j_2, l_2, \ldots, j_k, l_k)$, or equivalently a directed bipartite graph from $\{1,\ldots,n\}$ to $\{1,\ldots,n\}$, defined pictorially as follows:

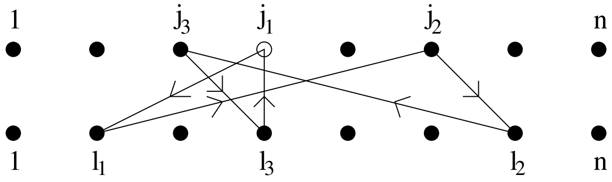


Figure 1. Directed bipartite graph associated to $\mathbb{E}\left(h_{j_1,l_1}\,\overline{h_{j_2,l_1}}\,h_{j_2,l_2}\,\overline{h_{j_3,l_2}}\,h_{j_3,l_3}\,\overline{h_{j_1,l_3}}\right)$

The vertex j_1 is called the *root* of the graph (it is both its starting and ending point). We will say that two sequences or graphs have the same *structure* if the order of appearance of new vertices in the sequence is the same. For example, the two sequences on the left-hand side below have the same structure, while the two on the right-hand side don't:

j_1	l_1	j_2	l_2	j_3	l_3	j_4	l_4
1	7	1	4	3	7	3	2
2	4	2	5	8	4	8	4

j_1	l_1	j_2	l_2	j_3	l_3	j_4	l_4
1	7	1	4	3	7	3	2
2	4	2	5	8	5	8	4

Because of assumption (iii) on the matrix entries h_{jl}, as well as the independence assumption, we see that in order for a given sequence to bring a non-zero contribution, it is necessary that whenever an edge from j to l appears in the graph (possibly a certain number of times), then it should also appear the *same* number of times in the opposite direction (that is, from l to j). A sequence or graph with this property is said to be *even*. Notice that an even graph with $2k$ edges can have at most $k+1$ vertices, as each edge in the graph is doubled.

The question is now: for an even graph with $2k$ edges, p vertices and a given structure, how many graphs with the same structure can we possibly have by permuting the positions of the vertices on each side? Clearly, there are at most n choices for each vertex, so the total number of choices is less than n^p.

Therefore,

$$\begin{aligned}
\mathbb{E}\left(m_k^{(n)}\right) &= \frac{1}{n^{k+1}} \sum_{(j_1,l_1,\ldots,j_k,l_k)\text{ even}} \mathbb{E}\left(h_{j_1,l_1}\,\overline{h_{j_2,l_1}}\cdots h_{j_k,l_k}\,\overline{h_{j_1,l_k}}\right) \\
&= \frac{1}{n^{k+1}} \sum_{p=2}^{k+1} \sum_{\substack{(j_1,l_1,\ldots,j_k,l_k) \\ \text{even with } p \text{ vertices}}} \mathbb{E}\left(h_{j_1,l_1}\,\overline{h_{j_2,l_1}}\cdots h_{j_k,l_k}\,\overline{h_{j_1,l_k}}\right) \\
&= \frac{1}{n^{k+1}} \sum_{\substack{(j_1,l_1,\ldots,j_k,l_k) \\ \text{even with } k+1 \text{ vertices}}} \mathbb{E}\left(h_{j_1,l_1}\,\overline{h_{j_2,l_1}}\cdots h_{j_k,l_k}\,\overline{h_{j_1,l_k}}\right) + O\left(\frac{1}{n}\right)
\end{aligned}$$

(Observe also that for a given p, each term is the sum is bounded because of assumption (ii).)

In conclusion, the only graphs that can possibly bring a non-negligible contribution are even graphs with $p = k+1$ vertices. In such graphs, each edge leads to a new vertex, so the resulting graph is actually a *tree*. For a tree with a given structure, there are order n^{k+1} different choices for the positions of the vertices. To be more precise, as the graph is bipartite, there are

$$n\,(n-1)\ldots(n-p_1+1)\,n\,(n-1)\cdots(n-p_2+1) \text{ choices} \tag{3}$$

where p_1, p_2 are the number of vertices on both sides of the graph, with $p_1 + p_2 = k+1$. In all cases, the above expression is of order $n^{p_1+p_2} = n^{k+1}$ as n grows large and k remains fixed. This factor n^{k+1} compensates therefore exactly with the $1/n^{k+1}$ factor in front of the sum. Notice that in such graphs, each edge appears exactly once in each direction of the original directed graph, so $\mathbb{E}\left(h_{j_1,l_1}\,\overline{h_{j_2,l_1}}\cdots h_{j_k,l_k}\,\overline{h_{j_1,l_k}}\right)$ is a product of $\mathbb{E}\left(|h_{jl}|^2\right) = 1$ by assumption (i), therefore the product itself is equal to 1.

The only question remaining is therefore: how many different structures of even graphs with $2k$ edges and $k+1$ vertices do there exist for a given k? In order to answer this question, let us make yet another identification: starting from the root j_1, explore the corresponding directed graph "following the arrows" and draw next to that a path that goes either up or down by one unit at each time step, according to the following rule:

$$\begin{cases}\text{if the current edge is a new edge, then go up by one unit} \\ \text{if the current edge has already been visited (in the other direction, necessarily), then go down by one unit}\end{cases}$$

The path being drawn is nothing but a Dyck path seen at the beginning of this lecture: it starts in 0, lands in 0 after $2k$ steps and cannot drop below zero in the meanwhile. Therefore, the number of different possible tree structures with $2k$ edges is equal to the number Dyck paths of length $2k$, that is, the Catalan number c_k. Gathering all the above observations together, we finally obtain (2):

$$\mathbb{E}\left(m_k^{(n)}\right) = c_k + O\left(\frac{1}{n}\right)$$

□

2 Largest eigenvalue

The results obtained in the previous lecture tell us something about the asymptotic distribution of a "typical" eigenvalue of the matrix $W^{(n)}$, that is, an eigenvalue picked uniformly at random. What can we say now on the *extreme eigenvalues* of such matrices, that is, the largest and the smallest one $\lambda_{\max}^{(n)}$ and $\lambda_{\min}^{(n)}$? It is first important to remember what Wigner's theorem actually says: the fact that almost surely, μ_n converges weakly towards μ means that for all $a < b$

$$\mu_n([a,b]) = \frac{1}{n}\,\sharp\{1 \le j \le n \;:\; \lambda_j^{(n)} \in [a,b]\} \underset{n\to\infty}{\to} \mu([a,b]) = \int_{a\vee 0}^{b\wedge 4} p_\mu(x)dx$$

Therefore, as soon as the interval $[a,b]$ has a non-empty intersection with the open interval $]0,4[$, the quantity on the right-hand side is strictly positive. This is saying in turn that the number of eigenvalues in this interval grows linearly in n as n grows to infinity. This applies in particular to the intervals $[0,\varepsilon]$ and $[4-\varepsilon,4]$, for any fixed $\varepsilon>0$, implying that almost surely, as n grows to infinity, both

$$\lim_{n\to\infty}\lambda_{\min}^{(n)}\leq\varepsilon \quad \text{and} \quad \lim_{n\to\infty}\lambda_{\max}^{(n)}\geq 4-\varepsilon$$

and therefore

$$\lim_{n\to\infty}\lambda_{\min}^{(n)}\leq 0 \quad \text{and} \quad \lim_{n\to\infty}\lambda_{\max}^{(n)}\geq 4$$

In the present case, this settles the limiting value of the smallest eigenvalue, as we know on the other hand that $\lambda_{\min}^{(n)}\geq 0$ for all n, because $W^{(n)}$ is positive semi-definite.

On the contrary, it is unclear whether $\lim_{n\to\infty}\lambda_{\max}^{(n)}=4$. Indeed, it could well be that one eigenvalue diverges from the interval $[0,4]$ in the large n limit: this would not affect the result of the Wigner theorem. Indeed, the weight of one eigenvalue in the distribution μ_n is equal to $1/n$, so an isolated eigenvalue cannot contribute to change the limiting distribution μ.

In order to study the asymptotic behavior of $\lambda_{\max}^{(n)}$, we will again use moments. This is made possible thanks to the following fact: for all $k\geq 1$,

$$\lambda_{\max}^{(n)}=\left((\lambda_{\max}^{(n)})^k\right)^{1/k}\leq\left(\sum_{j=1}^n(\lambda_j^{(n)})^k\right)^{1/k}=\left(\mathrm{Tr}\left((W^{(n)})^k\right)\right)^{1/k}$$

so

$$\lambda_{\max}^{(n)}\leq\lim_{k\to\infty}\left(\mathrm{Tr}\left((W^{(n)})^k\right)\right)^{1/k}$$

(and this last bound is actually known to be tight). We now set out to prove that under the following assumption:

$$h_{ik}=\exp(i\phi_{jk}) \quad \text{where } \phi_{jk} \text{ are i.i.d.} \sim\mathcal{U}([0,2\pi])\text{ random variables} \tag{4}$$

(which implies assumptions (i)-(iii) made in the last lecture), we have

$$\lim_{n\to\infty}\mathbb{E}\left(\lambda_{\max}^{(n)}\right)=4 \tag{5}$$

Using more refined techniques, one can prove the same result under weaker assumptions, as well as the almost sure version of the result: let us skip this.

Proof of equation (5). From the explanations above, it is clear that $\lim_{n\to\infty}\mathbb{E}\left(\lambda_{\max}^{(n)}\right)\geq 4$, so what remains to be proven is the upper bound

$$\lim_{n\to\infty}\mathbb{E}\left(\lambda_{\max}^{(n)}\right)\leq 4$$

Notice first that as $f(x)=x^{1/k}$ is concave, we obtain by Jensen's inequality[1],

$$\mathbb{E}\left(\lambda_{\max}^{(n)}\right)=\mathbb{E}\left(\lim_{k\to\infty}\left(\mathrm{Tr}\left((W^{(n)})^k\right)\right)^{1/k}\right)\leq\lim_{k\to\infty}\mathbb{E}\left(\mathrm{Tr}\left((W^{(n)})^k\right)\right)^{1/k}$$

As we have seen before,

$$\mathbb{E}\left(\mathrm{Tr}\left((W^{(n)})^k\right)\right)=\frac{1}{n^k}\,\mathbb{E}\left(\mathrm{Tr}\left((HH^*)^k\right)\right)=\frac{1}{n^k}\sum_{j_1,l_1,\ldots,j_k,l_k=1}^n\mathbb{E}\left(h_{j_1,l_1}\,\overline{h_{j_2,l_1}}\cdots h_{j_k,l_k}\,\overline{h_{j_1,l_k}}\right)$$

which is equation (1), up to a factor $1/n$. We now perform a similar analysis as before, but notice that we are interested in a different order of limits here: we first take $k\to\infty$ and then $n\to\infty$.

[1] and also by Fatou's lemma, which is needed here in order to interchange the limit and the expectation

Observe that as before, only the sequences $(j_1, l_1, \ldots, j_k, l_k)$ that correspond to even bipartite graphs bring a non-zero contribution to the sum, so

$$\begin{aligned}\mathbb{E}\left(\mathrm{Tr}\left((W^{(n)})^k\right)\right) &= \frac{1}{n^k} \sum_{(j_1,l_1,\ldots,j_k,l_k) \text{ even}} \mathbb{E}\left(h_{j_1,l_1}\,\overline{h_{j_2,l_1}} \cdots h_{j_k,l_k}\,\overline{h_{j_1,l_k}}\right) \\ &= \frac{1}{n^k} \sum_{p=2}^{k+1} \sum_{\substack{(j_1,l_1,\ldots,j_k,l_k) \\ \text{even with } p \text{ vertices}}} \mathbb{E}\left(h_{j_1,l_1}\,\overline{h_{j_2,l_1}} \cdots h_{j_k,l_k}\,\overline{h_{j_1,l_k}}\right)\end{aligned}$$

Notice also that for an even graph, the expression

$$\mathbb{E}\left(h_{j_1,l_1}\,\overline{h_{j_2,l_1}} \cdots h_{j_k,l_k}\,\overline{h_{j_1,l_k}}\right)$$

is the expectation of a product of even powers of $|h_{jl}|$, which are all equal to 1 here, because of assumption (4). So we simply have

$$\mathbb{E}\left(\mathrm{Tr}\left((W^{(n)})^k\right)\right) = \frac{1}{n^k} \sum_{p=2}^{k+1} N(k,p)$$

where $N(k,p)$ denotes the number of even graphs with $2k$ edges and p vertices (with $2 \leq p \leq k+1$). We have seen before that

$$N(k, k+1) \sim c_k\, n^{k+1}$$

as n gets large. It also holds that

$$\sum_{p=2}^{k+1} N(k,p) \leq c_k\, n^{k+1}$$

Indeed, when counting the number of graphs with $2k$ edges and $k+1$ vertices, we identified above c_k different structures for such graphs and slightly less than n^{k+1} graphs with a given structure, because of the constraint of having disjoint vertices on each side of the graph; see equation (3). Observe now that if we relax this constraint, we obtain all possible graphs with $2k$ edges and $k+1$ or less vertices. As there are at most n choices for each vertex, the total number of such graphs does not exceed $c_k\, n^{k+1}$.

This finally implies that

$$\mathbb{E}\left(\lambda_{\max}^{(n)}\right) \leq \lim_{k\to\infty} \left(\frac{1}{n^k}\, c_k\, n^{k+1}\right)^{1/k} = \lim_{k\to\infty} (nc_k)^{1/k} \leq 4, \quad \text{for all } n \geq 1$$

as we have already seen that $c_k \leq 4^k$ (and notice that as we consider first $k \to \infty$ and n fixed, the multiplicative factor n disappears in the large k limit). This completes the proof. □.

Random matrices and communication systems: WEEK 12

In this lecture, we reprove the theorem from last time using the Stieltjes transform method.

1 Marčenko-Pastur's theorem

Let H be an $n \times n$ random matrix with i.i.d. complex-valued entries such that for all $1 \leq j, l \leq n$:

(i) $\mathbb{E}(h_{jl}) = 0$, $\mathbb{E}\left(|h_{jl}|^2\right) = 1$;

(ii) the distribution of h_{jl} is compactly supported (this second assumption may be relaxed).

As last time, let us consider $W^{(n)} = \frac{1}{n}\, HH^*$, $\lambda_1^{(n)}, \ldots, \lambda_n^{(n)}$ its (non-negative) eigenvalues and the empirical distribution $\mu_n = \frac{1}{n}\sum_{j=1}^n \delta_{\lambda_j^{(n)}}$. A particular instance of Marčenko-Pastur's theorem is the following.

Theorem 1.1. Under assumptions (i) and (ii), almost surely, the sequence $(\mu_n,\, n \geq 1)$ converges weakly towards the quarter-circle law μ, whose pdf is given by

$$p_\mu(x) = \frac{1}{\pi}\sqrt{\frac{1}{x} - \frac{1}{4}}\, 1_{\{0<x<4\}}$$

As a preliminary to the proof of the theorem, let us consider, for a given $z \in \mathbb{C}\backslash\mathbb{R}$, the quadratic equation:

$$z\, g(z)^2 + z\, g(z) + 1 = 0 \tag{1}$$

This quadratic equation has two solutions

$$g_\pm(z) = -\frac{1}{2} \pm \sqrt{\frac{1}{4} - \frac{1}{z}}$$

It turns out that $g_+(z)$ is the Stieltjes transform of the above distribution μ. The proof is left as an exercise in the homework.

Proof of Theorem 1.1 (sketch).
The strategy for today is to use the characterization of weak convergence via Stieltjes transform: a sequence of distributions converges weakly towards a limiting distribution if the corresponding sequence of Stieltjes transforms converges pointwise on $\mathbb{C}_+ = \{z \in \mathbb{C} : \text{Im } z > 0\}$ towards the limiting Stieltjes transform. Thus, we will prove that almost surely, $g_n(z)$ converges in the large n limit towards a solution of equaton (1). As all the g_n are by definition Stieltjes transforms, but only one of the two solutions of this equation is a Stieltjes transform, a continuity argument allows then to conclude that g_n can only converge to g_+ and not to g_-.

The rule of the game is therefore now to try writing $g_n(z)$ on both sides of an equality sign. To this end, let us compute

$$g_n(z) = \int_{\mathbb{R}} \frac{1}{x-z}\, d\mu_n(x) = \frac{1}{n}\sum_{j=1}^n \frac{1}{\lambda_j^{(n)} - z} = \frac{1}{n}\,\text{Tr}\left(\left(W^{(n)} - zI\right)^{-1}\right)$$

First notice that $W^{(n)} = \frac{1}{n}\, HH^* = \frac{1}{n}\sum_{k=1}^n h_k\, h_k^*$, where h_k is the k^{th} column of H (and is therefore $n \times 1$). This way, $W^{(n)}$ is expressed as a sum of rank one $n \times n$ matrices. For a given $1 \leq k \leq n$, let us define

$$W_k^{(n)} = W^{(n)} - \frac{1}{n}\, h_k\, h_k^* = \frac{1}{n}\sum_{l=1,\, l\neq k}^n h_l\, h_l^*$$

as well as the resolvents

$$G^{(n)}(z) = \left(W^{(n)} - zI\right)^{-1} \quad \text{and} \quad G_k^{(n)}(z) = \left(W_k^{(n)} - zI\right)^{-1}$$

Notice that the object we are interested in is $g_n(z) = \frac{1}{n}\,\mathrm{Tr}\left(G^{(n)}(z)\right)$. Let us now prove the following two lemmas.

Lemma 1.2.

$$\frac{1}{n}\,h_k^*\,G^{(n)}(z)\,h_k = \frac{\frac{1}{n}\,h_k^*\,G_k^{(n)}(z)\,h_k}{1+\frac{1}{n}\,h_k^*\,G_k^{(n)}(z)\,h_k}$$

Proof. Let us compute

$$\begin{aligned} h_k^*\,G_k^{(n)}(z)\,\left(G^{(n)}(z)\right)^{-1} &= h_k^*\,G_k^{(n)}(z)\,\left(W^{(n)}-zI\right) = h_k^*\,G_k^{(n)}(z)\,\left(W_k^{(n)}-zI+\frac{1}{n}\,h_k\,h_k^*\right) \\ &= h_k^* + \frac{1}{n}\,h_k^*\,G_k^{(n)}(z)\,h_k\,h_k^* = \left(1+\frac{1}{n}\,h_k^*\,G_k^{(n)}(z)\,h_k\right)\,h_k^* \end{aligned}$$

Therefore,

$$h_k^*\,G_k^{(n)}(z) = \left(1+\frac{1}{n}\,h_k^*\,G_k^{(n)}(z)\,h_k\right)\,h_k^*\,G^{(n)}(z)$$

and

$$h_k^*\,G_k^{(n)}(z)\,h_k = \left(1+\frac{1}{n}\,h_k^*\,G_k^{(n)}(z)\,h_k\right)\,h_k^*\,G^{(n)}(z)\,h_k$$

which concludes the proof. □

Lemma 1.3.

$$g_n(z) = \frac{1}{n}\,\mathrm{Tr}\left(G^{(n)}(z)\right) = -\frac{1}{nz}\sum_{k=1}^n \frac{1}{1+\frac{1}{n}\,h_k^*\,G_k^{(n)}(z)\,h_k}$$

Proof. Let us compute

$$\begin{aligned} 1 &= \frac{1}{n}\,\mathrm{Tr}(I) = \frac{1}{n}\,\mathrm{Tr}\left(\left(W^{(n)}-zI\right)\,G^{(n)}(z)\right) = \frac{1}{n}\,\mathrm{Tr}\left(\frac{1}{n}\sum_{k=1}^n h_k\,h_k^*\,G^{(n)}(z) - z\,G^{(n)}(z)\right) \\ &= \frac{1}{n}\sum_{k=1}^n\left(\frac{1}{n}\,h_k^*\,G^{(n)}(z)\,h_k\right) - z\,g_n(z) = \frac{1}{n}\sum_{k=1}^n\left(\frac{\frac{1}{n}\,h_k^*\,G_k^{(n)}(z)\,h_k}{1+\frac{1}{n}\,h_k^*\,G_k^{(n)}(z)\,h_k}\right) - z\,g_n(z) \end{aligned}$$

by Lemma 1.2. So

$$z\,g_n(z) = \frac{1}{n}\sum_{k=1}^n\left(\frac{\frac{1}{n}\,h_k^*\,G_k^{(n)}(z)\,h_k}{1+\frac{1}{n}\,h_k^*\,G_k^{(n)}(z)\,h_k} - 1\right) = -\frac{1}{n}\sum_{k=1}^n \frac{1}{1+\frac{1}{n}\,h_k^*\,G_k^{(n)}(z)\,h_k}$$

which concludes the proof. □

Notice that so far, these formulas hold for *any* matrix of the form $W^{(n)} = \frac{1}{n}\,HH^*$, without any further assumption on the matrix H. On the contrary, the next lemma relies strongly on the assumptions (i) and (ii).

Lemma 1.4. Under assumptions (i) and (ii), for all $z \in \mathbb{C}\backslash\mathbb{R}$ and all $\varepsilon > 0$, there exists $C > 0$ independent of n such that

$$\mathbb{P}\left(\left|\frac{1}{n}\,h_k^*\,G_k^{(n)}\,h_k - \frac{1}{n}\mathrm{Tr}\left(G^{(n)}(z)\right)\right| \geq \varepsilon\right) \leq \frac{C}{n^2}$$

which implies by the Borel-Cantelli lemma that

$$\frac{1}{n}\,h_k^*\,G_k^{(n)}\,h_k - \frac{1}{n}\mathrm{Tr}\left(G_k^{(n)}(z)\right) \underset{n\to\infty}{\to} 0 \quad \text{almost surely}$$

We do not prove this lemma, but simply show in the following that for all $n \geq 1$,

$$\mathbb{E}\left(\frac{1}{n}\, h_k^*\, G_k^{(n)}\, h_k - \frac{1}{n}\mathrm{Tr}\left(G_k^{(n)}(z)\right)\right) = 0$$

which also requires the use of the assumptions made above:

$$\mathbb{E}\left(\frac{1}{n}\, h_k^*\, G_k^{(n)}\, h_k\right) = \frac{1}{n}\sum_{j,l=1}^{n} \mathbb{E}\left(\overline{h_{jk}}\, \left(W_k^{(n)} - zI\right)_{jl}^{-1}\, h_{lk}\right)$$

As the matrix $W_k^{(n)}$ does not "contain" h_k, the k^{th} column of H, it is independent of both $\overline{h_{jk}}$ and h_{lk}, so

$$\begin{aligned}
\mathbb{E}\left(\frac{1}{n}\, h_k^*\, G_k^{(n)}\, h_k\right) &= \frac{1}{n}\sum_{j,l=1}^{n} \mathbb{E}\left(\overline{h_{jk}}\, h_{lk}\right)\, \mathbb{E}\left(\left(W_k^{(n)} - zI\right)_{jl}^{-1}\right) \overset{(*)}{=} \frac{1}{n}\sum_{j=1}^{n} \mathbb{E}\left(\left(W_k^{(n)} - zI\right)_{jj}^{-1}\right) \\
&= \mathbb{E}\left(\frac{1}{n}\,\mathrm{Tr}\left(W_k^{(n)} - zI\right)\right) = \mathbb{E}\left(\frac{1}{n}\,\mathrm{Tr}\left(G_k^{(n)}(z)\right)\right)
\end{aligned}$$

where $(*)$ follows from the fact that $\mathbb{E}(\overline{h_{jk}}\, h_{lk}) = \delta_{jl}$, according to assumption (i) and the independence assumption.

The actual proof of Lemma 1.4 relies on the use of Chebychev's inequality with $\phi(x) = x^4$ and a similar analysis of the expectation (except that one must consider the moment of order 4 instead of the first order moment).

The last lemma is a technicality, which holds again for any matrix of the form $W^{(n)} = \frac{1}{n}\, HH^*$.

Lemma 1.5. For all $z \in \mathbb{C}\backslash\mathbb{R}$,

$$\left|\frac{1}{n}\mathrm{Tr}\left(G_k^{(n)}(z)\right) - \frac{1}{n}\mathrm{Tr}\left(G^{(n)}(z)\right)\right| \leq \frac{1}{n\,|\mathrm{Im}\ z|}$$

The proof of this lemma is still rather long for a technicality and is therefore omitted.

Gathering together the results of Lemmas 1.3, 1.4 and 1.5, we obtain that for large values of n,

$$\begin{aligned}
g_n(z) &= \frac{1}{n}\,\mathrm{Tr}\left(G^{(n)}(z)\right) = -\frac{1}{nz}\sum_{k=1}^{n} \frac{1}{1 + \frac{1}{n}\, h_k^*\, G_k^{(n)}(z)\, h_k} \\
&\simeq -\frac{1}{nz}\sum_{k=1}^{n} \frac{1}{1 + \frac{1}{n}\,\mathrm{Tr}\left(G_k^{(n)}(z)\right)} \simeq -\frac{1}{z}\,\frac{1}{1 + \frac{1}{n}\,\mathrm{Tr}\left(G^{(n)}(z)\right)} = -\frac{1}{z\,(1 + g_n(z))}
\end{aligned}$$

which may be rewritten as

$$z\, g_n(z)^2 + z\, g_n(z) + 1 \simeq 0$$

Taking some more precautions, we can conclude that $g_n(z)$ converges almost surely towards a solution of the quadratic equation (1), which should be chosen as g_+ for the reasons explained above. This "completes" the proof of the theorem. □

Random matrices and communication systems: WEEK 13

1 Capacity of multiple antenna channels

1.1 Finite-size anaylsis

Let us come back the multiple antenna channel considered in Lecture 3:

$$Y = H\,X + Z$$

where H is an $n \times n$ random channel matrix with i.i.d. $\mathcal{N}_{\mathbb{C}}(0,1)$ entries, varying ergodically over time, whose realizations are known at the receiver, but not at the transmitter. We have seen in Lecture 3 that the ergodic capacity of such a system is given by

$$C_{\text{erg}} = \mathbb{E}\left(\log\det\left(I + \frac{P}{n}\,HH^*\right)\right) = \mathbb{E}\left(\sum_{j=1}^{n} \log\left(1 + P\lambda_j^{(n)}\right)\right)$$

where the expectation is taken over all possible realizations of the random matrix H and $\lambda_1^{(n)}, \ldots, \lambda_n^{(n)}$ are the non-negative eigenvalues of the $n \times n$ Wishart matrix $W^{(n)} = \frac{1}{n}\,HH^*$. This may be further rewritten as

$$C_{\text{erg}} = n\,\mathbb{E}\left(\log\left(1 + P\lambda^{(n)}\right)\right)$$

where $\lambda^{(n)}$ is one of the eigenvalues $\lambda_1^{(n)}, \ldots, \lambda_n^{(n)}$ picked uniformly at random. We have seen in Lecture 7 that the distribution of $\lambda^{(n)}$ is given by

$$p^{(n)}(\lambda) = e^{-n\lambda} \sum_{l=0}^{n-1} L_l(n\lambda)^2$$

where the $L_l(\cdot)$ are the Laguerre polynomials. Therefore,

$$C_{\text{erg}} = n \int_0^\infty d\lambda\, p^{(n)}(\lambda)\, \log(1 + P\lambda) = n \sum_{l=0}^{n-1} \int_0^\infty d\lambda\, e^{-n\lambda}\, L_l(n\lambda)^2\, \log(1 + P\lambda)$$

1.2 Asymptotic analysis

In order to analyze the behavior of the ergodic capacity in the large n limit, let us rewrite it as

$$C_{\text{erg}}(n) = \mathbb{E}\left(\sum_{j=1}^{n} \log\left(1 + P\lambda_j^{(n)}\right)\right) = n\,\mathbb{E}\left(\int_{\mathbb{R}} \log(1 + Px)\, d\mu_n(x)\right)$$

where

$$\mu_n = \frac{1}{n}\sum_{j=1}^{n} \delta_{\lambda_j^{(n)}}$$

is the empirical eigenvalue distribution of the matrix $W^{(n)} = \frac{1}{n}\,HH^*$. We have seen in Lectures 10-12 that almost surely, μ_n converges weakly towards the limiting deterministic distribution μ whose pdf is given by

$$p_\mu(x) = \frac{1}{\pi}\sqrt{\frac{1}{x} - \frac{1}{4}}\; 1_{\{0<x<4\}}$$

This is saying that almost surely, for any bounded continuous function $f : \mathbb{R} \to \mathbb{R}$,

$$\lim_{n\to\infty} \int_{\mathbb{R}} f(x)\, d\mu_n(x) = \int_{\mathbb{R}} f(x) d\mu(x)$$

Taking expectations on both sides (which is OK here, thanks to the dominated convergence theorem), we obtain

$$\lim_{n\to\infty} \mathbb{E}\left(\int_{\mathbb{R}} f(x)\, d\mu_n(x) \right) = \int_{\mathbb{R}} f(x) d\mu(x)$$

for any bounded continuous function $f : \mathbb{R} \to \mathbb{R}$ (remember that μ is deterministic). We would like now to apply this to $f(x) = \log(1 + Px)$, which would allow us to conclude that

$$\lim_{n\infty} \frac{C_{\text{erg}}(n)}{n} = \lim_{n\to\infty} \mathbb{E}\left(\int_{\mathbb{R}} \log(1 + Px)\, d\mu_n(x) \right) = \int_0^4 \log(1 + Px)\, p_\mu(x)\, dx$$

therefore proving that $C_{\text{erg}}(n)$ is of order n for large values of n and at the same time providing an explicit expression for the multiplicative factor.

A first concern is that $f(x) = \log(1 + Px)$ is not defined for $x < -1/P$, but as both μ_n and μ are supported on $[0, \infty)$, this is not a problem. The other worry is that f is unbounded. To this end, let us define $f_M(x) = \min(f(x), M)$, which is bounded and continuous for any $M > 0$. This allows us to write, for fixed values of n and M

$$\begin{aligned}
\left| \mathbb{E}\left(\int_0^\infty f(x)\, d\mu_n(x) \right) - \int_0^\infty f(x) d\mu(x) \right| &\leq \left| \mathbb{E}\left(\int_0^\infty f(x)\, d\mu_n(x) \right) - \mathbb{E}\left(\int_0^\infty f_M(x)\, d\mu_n(x) \right) \right| \\
&+ \left| \mathbb{E}\left(\int_0^\infty f_M(x)\, d\mu_n(x) \right) - \int_0^\infty f_M(x)\, d\mu(x) \right| + \left| \int_0^\infty f_M(x)\, d\mu(x) - \int_0^\infty f_(x)\, d\mu(x) \right| \\
&\leq \ \mathbb{E}\left(\int_{x_M}^\infty f(x)\, d\mu_n(x) \right) + \left| \mathbb{E}\left(\int_0^\infty f_M(x)\, d\mu_n(x) \right) - \int_0^\infty f_M(x)\, d\mu(x) \right| + \int_{x_M}^\infty f(x)\, d\mu(x)
\end{aligned}$$

where $x_M = \inf\{x > 0 \,:\, f(x) \geq M\} = \frac{1}{P}\,(e^M - 1)$. By the weak convergence result above, we know that the term in the middle converges to zero for any value of M, so

$$\lim_{n\to\infty} \left| \mathbb{E}\left(\int_0^\infty f(x)\, d\mu_n(x) \right) - \int_0^\infty f(x) d\mu(x) \right| \leq \lim_{n\to\infty} \mathbb{E}\left(\int_{x_M}^\infty f(x)\, d\mu_n(x) \right) + \int_{x_M}^\infty f(x)\, d\mu(x)$$

for any $M > 0$. We also know that $\int_{x_M}^\infty f(x)\, d\mu(x) = 0$ for $x_M > 4$ (as μ is supported on $[0, 4]$), so there remains to prove that (also for $x_M > 4$)

$$\lim_{n\to\infty} \mathbb{E}\left(\int_{x_M}^\infty f(x)\, d\mu_n(x) \right) = 0$$

Notice that

$$\mathbb{E}\left(\int_{x_M}^\infty f(x)\, d\mu_n(x) \right) = \frac{1}{n} \sum_{j=1}^n \mathbb{E}\left(f(\lambda_j^{(n)})\, 1_{\{\lambda_j^{(n)} \geq x_M\}} \right) \leq \mathbb{E}\left(f(\lambda_{\max}^{(n)})\, 1_{\{\lambda_{\max}^{(n)} \geq x_M\}} \right)$$

As $f(x) = \log(1 + Px) \leq Px$, this is further bounded above by

$$P\, \mathbb{E}\left(\lambda_{\max}^{(n)}\, 1_{\{\lambda_{\max}^{(n)} \geq x_M\}} \right)$$

In Lecture 11, we have seen (under slightly different assumptions, though) that $\lim_{n\to\infty} \mathbb{E}(\lambda_{\max}^{(n)}) \leq 4$. Similar refined estimates on $\lambda_{\max}^{(n)}$ allow to conclude that the above expression converges to zero as $n \to \infty$ for $x_M > 4$.

2 Diversity-multiplexing tradeoff

Consider now the same scenario as above, except for the fact that H is fixed over time (see Lecture 4). In this case, the capacity of the multiple antenna channel is equal to zero, and the outage probability is given by

$$\mathbb{P}_{\text{out}}(R) = \inf_{Q \geq 0 \,:\, \text{Tr}(Q) \leq P} \mathbb{P}(\log\det(I + HQH^*) < R)$$

where again, the probability is taken over all possible realizations of the random matrix H. For a fixed value of n, we would like to characterize the behavior of this outage probability in the high SNR regime (that is, as P gets large), with the idea that it can be made vanishingly small in this regime. In the ergodic case, we have seen that for large P (see Lecture 3),

$$C_{\text{erg}} = \sup_{Q \geq 0 \,:\, \text{Tr}(Q) \leq P} \mathbb{E}(\log\det(I + HQH^*)) \simeq n \, \log P$$

So in order to obtain a small outage probability, the target rate R chosen in the above expression should not be higher than $n \log P$. Let us therefore choose $R = r \log P$, where $0 \leq r \leq n$: r is called the target *multiplexing gain*.

A priori, the analysis of the outage probability is particularly difficult, as the above minimization problem remains unsolved. But notice that

$$\mathbb{P}(\log\det(I + PHH^*) < r \, \log P) \leq \mathbb{P}_{\text{out}}(r \, \log P) \leq \mathbb{P}\left(\log\det\left(I + \frac{P}{n} HH^*\right) < r \, \log P\right)$$

Indeed, $Q = \frac{P}{n} I$ is a possible candidate for the minimization problem, which explains the inequality on the right-hand side. On the other hand, any matrix $Q \geq 0$ satisfying $\text{Tr}(Q) \leq P$ also satisfies $Q \leq PI$, which implies the inequality on the left-hand side. Observe now that as $P \to \infty$,

$$\begin{aligned} \mathbb{P}\left(\log\det\left(I + \frac{P}{n} HH^*\right) < r \, \log P\right) &\doteq \mathbb{P}(\log\det(I + PHH^*) < r \, \log(nP)) \\ = \quad \mathbb{P}(\log\det(I + PHH^*) < r\,(\log n + \log P)) &\doteq \mathbb{P}(\log\det(I + PHH^*) < r \, \log P) \end{aligned}$$

where the notation $f(P) \doteq g(P)$ stands for $\lim_{P\to\infty} \frac{\log(f(P))}{\log P} = \lim_{P\to\infty} \frac{\log(g(P))}{\log P}$. This together with the previous inequalities allows us to conclude that

$$\mathbb{P}_{\text{out}}(r \, \log P) \doteq \mathbb{P}(\log\det(I + PHH^*) < r \, \log P)$$

As mentioned above, by choosing the multiplexing gain r smaller than n, we expect the outage probability to converge to zero as P gets large. Our aim in the following is to discover at which speed, depending on r, does this probability converge to zero, namely to find the exponent $d(r)$ satisfying

$$P_{\text{out}}(r \, \log P) \doteq P^{-d(r)}$$

More formally, this exponent, also known as the *diversity order*, is defined as

$$d(r) = \lim_{P\to\infty} -\frac{\log(P_{\text{out}}(r \, \log P))}{\log P}$$

which the above analysis allows us to rewrite as

$$d(r) = \lim_{P\to\infty} -\frac{\log(\mathbb{P}(\log\det(I + PHH^*) < r \, \log P))}{\log P}$$

The computation of $d(r)$, which requires the knowledge of the joint eigenvalue distribution of the matrix HH^*, will be the subject of the next lecture.

Random matrices and communication systems: WEEK 14

1 Diversity-multiplexing tradeoff (cont'd)

Remember from last lecture that we are after computing the diversity order of a multiple antenna channel:

$$d(r) = \lim_{P\to\infty} -\frac{\log(P_{\text{out}}(r\,\log P))}{\log P} = \lim_{P\to\infty} -\frac{\log(\mathbb{P}(\log\det(I + PHH^*) < r\,\log P)))}{\log P}$$

where H is an $n \times n$ matrix with i.i.d.$\sim \mathcal{N}_{\mathbb{C}}(0,1)$ entries (and n is fixed). The above probability can be rewritten as

$$\mathbb{P}(\log\det(I + PHH^*) < r\,\log P) = \mathbb{P}\left(\sum_{j=1}^n \log(1 + P\lambda_j) < r\,\log P\right)$$

where $\lambda_1, \ldots, \lambda_n$ are the eigenvalues of the matrix HH^*, which are all non-negative. In Lecture 6, we have seen that the joint distribution of these eigenvalues is given by

$$p(\lambda_1, \ldots, \lambda_n) = c_n \prod_{j=1}^n e^{-\lambda_j} \prod_{j<k} (\lambda_k - \lambda_j)^2 \quad \text{for } \lambda_1, \ldots, \lambda_n \geq 0$$

where c_n is some positive constant. Using this, the above probability can be further rewritten as an n-fold integral:

$$\mathbb{P}(\log\det(I + PHH^*) < r\,\log P) = \int_{D_\lambda(r)} p(\lambda_1, \ldots, \lambda_n)\, d\lambda_1 \cdots d\lambda_n$$

where

$$D_\lambda(r) = \left\{ 0 \leq \lambda_1 \leq \ldots \leq \lambda_n \;:\; \sum_{j=1}^n \log(1 + P\lambda_j) < r\,\log P \right\}$$

(notice that the eigenvalues $\lambda_1, \ldots, \lambda_n$ are ordered in increasing order here). The explicit computation of this integral remains a challenge, because of the highly correlated nature of the eigenvalues. We will see below that taking the high SNR limit ($P \to \infty$) in the above expression allows to drastically simplify the analysis. To this end, let us make the change of variables:

$$\lambda_j = P^{-\alpha_j} = \exp(-\alpha_j\,\log P), \quad \text{so} \quad d\lambda_j = -(\log P)\,\exp(-\alpha_j\,\log P)\,d\alpha_j$$

This change of variable, even though depending on P, is perfectly valid for given value of P, and therefore also in the limit $P \to \infty$ (provided some care is taken here). This gives rise to the following expression for the above probability:

$$\mathbb{P}(\log\det(I + PHH^*) < r\,\log P) = \int_{D_\alpha(r)} q(\alpha_1, \ldots, \alpha_n)\, d\alpha_1 \cdots d\alpha_n$$

where

$$q(\alpha_1, \ldots, \alpha_n) = c_n\,\exp\left(-\sum_{j=1}^n P^{-\alpha_j}\right) \prod_{j<k} \left(P^{-\alpha_j} - P^{-\alpha_k}\right)^2 (\log P)^n\,\exp\left(-\sum_{j=1}^n \alpha_j\,\log P\right)$$

and

$$D_\alpha(r) = \left\{ \alpha_1 \geq \ldots \geq \alpha_n \ (\alpha_j \in \mathbb{R}) \;:\; \sum_{j=1}^n \log\left(1 + P^{1-\alpha_j}\right) < r\,\log P \right\}$$

So far, these are exact expressions. We will now make a series of approximations which are valid in the limit $P \to \infty$ (and which can all be rigorously justified by taking upper and lower bounds).

First observe that

$$\exp\left(-P^{-\alpha_j}\right)\ \begin{cases}\text{decays super-polynomially to zero} & \text{if } \alpha_j < 0\\ \text{tends to 1} & \text{if } \alpha_j \geq 0\end{cases}$$

so we may restrict the domain of integration $D_\alpha(r)$ to its positive part where $\alpha_1 \geq \ldots \geq \alpha_n \geq 0$. Next, observe that

$$\log\left(1+P^{1-\alpha_j}\right) \simeq \begin{cases}(1-\alpha_j)\,\log P & \text{if } \alpha_j \leq 1\\ 0 & \text{if } \alpha_j > 1\end{cases}$$

so $\log\left(1+P^{1-\alpha_j}\right) \simeq (1-\alpha_j)^+\,\log P$. We can therefore replace the domain of integration $D_\alpha(r)$ by

$$\widetilde{D}_\alpha(r) = \left\{\alpha_1 \geq \ldots \geq \alpha_n \geq 0\ :\ \sum_{j=1}^n (1-\alpha_j)^+ \leq r\right\}$$

and the above probability can be rewritten as

$$\mathbb{P}(\log\det(I+PHH^*) < r\,\log P) \doteq \int_{\widetilde{D}_\alpha(r)} \widetilde{q}(\alpha_1,\ldots,\alpha_n)\,d\alpha_1\cdots d\alpha_n$$

where

$$\widetilde{q}(\alpha_1,\ldots,\alpha_n) = c_n \prod_{j<k}\left(P^{-\alpha_j} - P^{-\alpha_k}\right)^2\,(\log P)^n\,\exp\left(-\sum_{j=1}^n \alpha_j\,\log P\right)$$

Furthermore, let us notice that $c_n(\log P)^n \doteq 1$, as $\lim_{P\to\infty}\frac{\log(c_n\,(\log P)^n)}{\log P} = 0$. Here comes now the "magic" trick: for $\alpha_1 > \ldots > \alpha_n$, we have

$$\prod_{j<k}\left(P^{-\alpha_j} - P^{-\alpha_k}\right)^2 \doteq \prod_{j<k} P^{-2\alpha_k} = \prod_{k=1}^n P^{-2(k-1)\,\alpha_k} = \exp\left(-\sum_{k=1}^n 2(k-1)\,\alpha_k\,\log P\right)$$

This implies that

$$\widetilde{q}(\alpha_1,\ldots,\alpha_n) \doteq \exp\left(-\sum_{j=1}^n (2j-1)\,\alpha_j\,\log P\right)$$

In this expression, we see that in the limit $P\to\infty$, the exponents α_j become so to speak "independent". Finally, we obtain

$$\mathbb{P}(\log\det(I+PHH^*) < r\,\log P) \doteq \int_{\widetilde{D}_\alpha(r)} \exp\left(-\sum_{j=1}^n (2j-1)\,\alpha_j\,\log P\right)\,d\alpha_1\cdots d\alpha_n$$

This expression can in turn be rewritten as

$$\mathbb{P}(\log\det(I+PHH^*) < r\,\log P) \doteq \int_{\widetilde{D}_\alpha(r)} P^{-f(\alpha)}\,d\alpha_1\cdots d\alpha_n$$

where

$$f(\alpha) = \sum_{j=1}^n (2j-1)\,\alpha_j$$

Using then Laplace's integration method, we obtain

$$\mathbb{P}(\log\det(I+PHH^*) < r\,\log P) \doteq P^{-d(r)}$$

where the diversity order $d(r)$ is given by

$$d(r) = \inf_{\widetilde{D}_\alpha(r)} f(\alpha) = \inf_{\substack{\alpha_1 \geq \ldots \geq \alpha_n \geq 0 \\ \sum_{j=1}^n (1-\alpha_j)^+ < r}} \sum_{j=1}^n (2j-1)\,\alpha_j$$

Doing this, we have therefore transformed the initial problem of evaluating an n-fold integral (in the limit $P \to \infty$) into a simple linear optimization probelem.

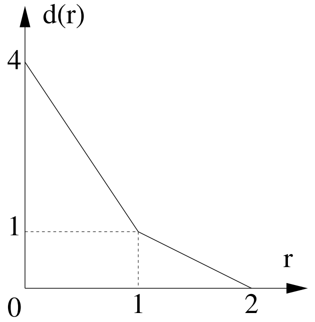
For $n = 2$, the problem reads

$$d(r) = \inf_{\substack{\alpha_1 \geq \alpha_2 \geq 0 \\ (1-\alpha_1)^+ + (1-\alpha_2)^+ \leq r}} \alpha_1 + 3\,\alpha_2$$

whose solution is given by

$$\begin{cases} 0 \leq r \leq 1: & \alpha_1 = 1,\, \alpha_2 = 1 - r,\, d(r) = 4 - 3r \\ 1 \leq r \leq 2: & \alpha_1 = 2 - r,\, \alpha_2 = 0,\, d(r) = 2 - r \end{cases}$$

For low multiplexing gain ($0 \leq r \leq 1$), outage occurs when both eigenvalues λ_1, λ_2 of HH^* are small (more precisely, $\lambda_1 \simeq P^{-1}$ and $\lambda_2 \simeq P^{r-1}$), while for higher multiplexing gain ($1 \leq r \leq 2$), outage occurs when only the smallest eigenvalue λ_2 is small (more precisely $\lambda_1 \simeq 1$ and $\lambda_2 \simeq P^{r-2}$). As expected, the diversity drops to zero for values of r larger than or equal to 2 (as in this case, the target rate is higher than the ergodic capacity). On the figure below, the diversity order is drawn as a function of the multiplexing gain r, which illustrates the tradefoff between diversity and multiplexing.



For general values of n, the curve $d(r)$ is the piecewise linear curve such that $d(k) = (n-k)^2$ at integer values of r (so $d(0) = n^2$ and $d(n) = 0$). Notice that the maximum diversity $d = n^2$ corresponding to $r = 0$ matches the number of independent random variables in the channel matrix H.