An Introduction to Random Matrices

Parsa Rangriz

Department of Statistics and Actuarial Science,
University of Waterloo,
Waterloo, ON, Canada, N2L 3G1

April 18, 2024

Contents

1 High Dimensional Probability

This chapter provides an overview of concentration inequalities, a significant subject in probability theory. We establish fundamental concentration inequalities such as Hoeffding's, Chernoff's, and Bernstein's inequalities. Additionally, the chapter aims to introduce two significant categories of probability distributions: sub-Gaussian and sub-exponential distributions. These distributions are foundational in the construction of Orlicz spaces, which are instrumental in analyzing high-dimensional probability phenomena. Concentration inequalities serve to quantify the extent to which a random variable X fluctuates around its mean μ. Typically, these inequalities provide bounds for the tails of the distribution of $X - \mu$, represented as

$$\mathrm{P}(|X - \mu| \geq t) \leq \text{small quantity as a function of t}$$

One of the most elementary concentration inequalities is used to derive tight bounds on the tails of the normal distribution. These bounds demonstrate an exponential decay with respect to the sample size N, which is notably superior to the linear decay observed in Chebyshev's inequality.

Theorem 1.1 (**Tails of the Gaussian random variable**). *Let $\phi(t)$ be the CDF of the standard Gaussian distribution. Then,*

$$\left(\frac{1}{t} - \frac{1}{t^3}\right) \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \leq \phi(t) \leq \frac{1}{t} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}$$

Proof. To obtain an upper bound on the tail

$$\mathbb{P}\{g \geq t\} = \frac{1}{\sqrt{2\pi}} \int_t^\infty e^{-x^2/2} dx$$

let us make the change of variables $x = t + y$. This gives

$$\mathbb{P}\{g \geq t\} = \frac{1}{\sqrt{2\pi}} \int_0^\infty e^{-t^2/2} e^{-ty} e^{-y^2/2} dy \leq \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \int_0^\infty e^{-ty} dy,$$

where we have used that $e^{-y^2/2} \leq 1$. Since the last integral equals $1/t$, the desired upper bound on the tail follows. The lower bound follows from the identity

$$\int_t^\infty \left(1 - 3x^{-4}\right) e^{-x^2/2} dx = \left(\frac{1}{t} - \frac{1}{t^3}\right) e^{-t^2/2}.$$

□

Let's begin with a straightforward concentration inequality that applies to the sums of identical and independently distributed (i.i.d.) symmetric Bernoulli random variables.

1.1 Hoeffding's Inequality

Theorem 1.2 (**Hoeffding's inequality for symmetric Bernoulli**). *Let $X_1, \cdots, X_N \overset{ind}{\sim} Unif\{\pm 1\}$, and let $a = (a_1, \cdots, a_N) \in \mathbb{R}^N$. Then for any $t \geq 0$, we have*

$$P\left(\sum_{i=1}^N a_i X_i \geq t\right) \leq \exp\left(-\frac{t^2}{2\|a\|_2^2}\right)$$

Proof. Without loss of generality assume that $\|a\|_2 = 1$. Using Chebyshev's inequality, one may write,

$$\mathrm{P}\left(\sum_{i=1}^N a_i X_i \geq t\right) \leq e^{-\lambda t} \,\mathrm{E} \exp\left(\lambda \sum_{i=1}^N a_i X_i\right)$$

Recall that the MGF of the sum is the product of the MGFs of the terms, thus

$$\mathrm{E} \exp\left(\lambda \sum_{i=1}^N a_i X_i\right) = \prod_{i=1}^N \mathrm{E} \exp(\lambda a_i X_i)$$

Since X_i takes values ± 1 with probabilities $\frac{1}{2}$ each, we have

$$\mathrm{E}\exp(\lambda a_i X_i) = \frac{\exp(\lambda a_i) + \exp(-\lambda a_i)}{2} = \cosh(\lambda a_i)$$

We can show that $\cosh(x) \leq \exp(x^2/2)$ for all $x \in \mathbb{R}$. This bounds leads that

$$\mathrm{E}\exp(\lambda a_i X_i) \leq \exp(\lambda^2 a_i^2/2)$$

Finally we obtain,

$$\mathrm{P}\left(\sum_{i=1}^N a_i X_i \geq t\right) \leq \exp(-\lambda t) \prod_{i=1}^N \exp(\lambda^2 a_i^2/2) = \exp\left(-\lambda t + \frac{\lambda^2}{2}\right)$$

This bound holds for arbitrary $\lambda > 0$. Choosing the optimized λ, we have

$$\mathrm{P}\left(\sum_{i=1}^N a_i X_i \geq t\right) \leq \exp(-t^2/2)$$

□

Hoeffding's inequality can be seen as a concentration counterpart to the central limit theorem. Essentially, the best we can hope for in a concentration inequality is that the tail behavior of $\sum a_i X_i$ mirrors that of the normal distribution. Remarkably, Hoeffding's tail bound achieves just that. When normalized with $\|a\|_2 = 1$. Hoeffding's inequality yields the tail $e^(- t^2/2)$, identical to the bound for the standard normal tail. This equivalence is significant as it ensures that the tails of sums exhibit the same exponential lightness as those of the normal distribution, despite the two distributions not being exponentially close in essence.

Corollary 1.3. *Let $X_1, \cdots, X_N \overset{ind}{\sim} Unif\{\pm 1\}$, and let $a = (a_1, \cdots, a_N) \in \mathbb{R}^N$. Then for any $t \geq 0$, we have*

$$P\left(\left|\sum_{i=1}^N a_i X_i \geq t\right|\right) \leq 2\exp\left(-\frac{t^2}{2\|a\|_2^2}\right)$$

Thus far, our exploration of concentration inequalities has been limited to Bernoulli random variables. Extending these findings to encompass a broader spectrum of distributions would be advantageous. At the very least, we anticipate that the normal distribution should be included in this broader class, given that concentration results are often viewed as quantitative manifestations of the central limit theorem.

1.2 Sub-Gaussian Random Variables

Let X be a random variable with general characteristics. The proposition below asserts that the attributes we previously discussed—sub-gaussian tail decay, the increase of moments, and the growth of the moment generating function—are interconnected and equivalent. The proof of this proposition is particularly valuable as it elucidates the method of converting one form of information regarding random variables into another.

Theorem 1.4 (**Sub-Gaussian properties**). *Let X be a random variable. Then the following properties are equivalent; the parameter $K_i > 0$ appearing in these properties differ from each other by at most an absolute constant factor.*

1. *The tails of X satisfy*
$$P(|X| \geq t) \leq 2\exp(-t^2/K_1^2), \qquad \forall t \geq 0$$
2. *The moments of X satisfy*
$$\|X\|_p = (\mathrm{E}\,|X|^p)^{1/p} \leq K_2\sqrt{p}, \qquad \forall p \geq 1$$

3. The MGF of X^2 satisfies

$$\mathrm{E}\exp(\lambda^2 X^2) \le \exp(K_3^2\lambda^2), \qquad \forall|\lambda| \le 1/K_3$$

4. The MGF of X^2 is bounded at some point, namely

$$\mathrm{E}\exp(X^2/K_4^2) \le 2$$

5. If $\mathrm{E}\,X = 0$, then the MGF of X satisfies

$$\mathrm{E}\exp(\lambda X) \le \exp(K_5^2\lambda^2) \qquad \forall\lambda \in \mathbb{R}$$

Proof. (1) $\to$ (2): By homogeneity and rescaling X to X/K_1 we can assume that $K_1 = 1$. Using layer cake representation, one may write,

$$\mathrm{E}\,|X|^p = \int_0^\infty \mathrm{P}(|X|^p \ge u)du = \int_0^\infty 2e^{-t^2}pt^{p-1}dt = p\Gamma(p/2) \le p(p/2)^{p/2}$$

where in the last step we use $\Gamma(x) \le x^x$.

(2) $\to$ (3): By homogeneity we may assume that $K_2 = 1$. Recalling the Taylor series expansion,

$$\mathrm{E}\exp(\lambda^2 X^2) = 1 + \sum_{p=1}^\infty \frac{\lambda^{2p}\,\mathrm{E}\,X^{2p}}{p!} \le 1 + \sum_{p=1}^\infty \frac{(2\lambda^2 p)^p}{(p/e)^p} = \frac{1}{1-2e\lambda^2}$$

where the property (2) guarantees that $\mathrm{E}\,X^{2p} \le (2p)^p$, while Stirling's approximation yields $p! \ge (p/e)^p$. To bound the geometric series we can use the numeric inequality $\frac{1}{1-x} \le e^{2x}$, which is valid for $x \in [0, 1/2]$.

(3) $\to$ (4): Trivial.

(4) $\to$ (1): As before, assume that $K_4 = 1$. Then,

$$\mathrm{P}(|X| \ge t) = \mathrm{P}(e^{X^2} \ge e^{t^2}) \le e^{-t^2}\,\mathrm{E}\,e^{X^2} \le 2e^{-t^2}$$

(3) $\to$ (5): Without loss of generality, take that $K_3 = 1$. Using the inequality $e^x \le x + e^{x^2}$ for all $x \in \mathbb{R}$, then

$$\mathrm{E}\,e^{\lambda X} \le \mathrm{E}(\lambda X + e^{\lambda^2 X^2}) = \mathrm{E}\,e^{\lambda^2 X^2} \le e^{\lambda^2}, \qquad \forall|\lambda| \le 1.$$

Now assume that $|\lambda| \ge 1$. Using the inequality $\lambda x \le \lambda^2 + x^2$,

$$\mathrm{E}\,e^{\lambda X} \le e^{\lambda^2/2}\,\mathrm{E}\,e^{X^2/2} \le e^{\lambda^2/2 + 1/2} \le e^{\lambda^2}$$

(5) $\to$ (1): We can assume that $K_5 = 1$. Using the idea from the proof of Hoeffding's inequality, then

$$\mathrm{P}(X \ge t) = e^{-\lambda t}\,\mathrm{E}\,e^{\lambda X} \le e^{-\lambda t + \lambda^2}$$

Optimizing λ, we conclude that $\mathrm{P}(|X| \ge t) \le 2e^{-t^2/4}$. □

Definition 1.5 (**Sub-Gaussian Random Variables**)**.** A random variable X that satisfies one of the equivalent sub-Gaussian properties is called a sub-Gaussian random variable. The sub-Gaussian norm of X, denoted $\|X\|_{\psi_2}$, is defined to be the smallest K_4 in property (4). In other words, we define

$$\|X\|_{\psi_2} = \inf\{t > 0 : \mathrm{E}\exp(X^2/t^2) \le 2\}$$

Remark 1.6. *Gaussian, Bernoulli, and Bounded random variables are some classical examples of sub-Gaussian random variables:*

After all the work we did in characterizing sub-gaussian distributions in the previous section, we can now easily extend Hoeffding's inequality to general sub-gaussian distributions.

Corollary 1.7 (**General Hoeffding's inequality**)**.** *Let $X_1, \cdots, X_N$ be independent mean-zero Gaussian random variables and let $a = (a_1, \cdots, a_N) \in \mathbb{R}^N$. Then for every $t \ge 0$, we have*

$$P\left(\left|\sum_{i=1}^N X_i\right| \ge t\right) \le 2\exp\left(\frac{-t^2}{2\sum_{i=1}^N \sigma_i^2}\right),$$

$$P\left(\left|\sum_{i=1}^N a_i X_i\right| \ge t\right) \le 2\exp\left(\frac{-t^2}{2\max_i \sigma_i^2 \|a\|_2^2}\right)$$

1.3 Sub-Exponential Random Variables

The sub-gaussian distribution class, while extensive and intuitive, excludes significant distributions with heavier tails than the Gaussian. In this section, we shift our focus to distributions featuring at least an exponential tail decay. We aim to establish an analogue of Hoeffding's inequality tailored to this class of distributions.

Theorem 1.8 (**Sub-exponential properties**). *Let X be a random variable. Then the following properties are equivalent; the parameter $K_i > 0$ appearing in these properties differ from each other by at most an absolute constant factor.*

1. *The tails of X satisfy*
$$P(|X| \geq t) \leq 2\exp(-t/K_1), \qquad \forall t \geq 0$$
2. *The moments of X satisfy*
$$\|X\|_p = (\mathrm{E}\,|X|^p)^{1/p} \leq K_2 p, \qquad \forall p \geq 1$$
3. *The MGF of X^2 satisfies*
$$\mathrm{E}\exp(\lambda|X|) \leq \exp(K_3\lambda), \qquad \forall|\lambda| \leq 1/K_3$$
4. *The MGF of $|X|$ is bounded at some point, namely*
$$\mathrm{E}\exp(X/K_4) \leq 2$$
5. *If $\mathrm{E}\,X = 0$, then the MGF of X satisfies*
$$\mathrm{E}\exp(\lambda X) \leq \exp(K_5^2\lambda^2) \qquad \forall|\lambda| \leq 1/K_5$$

Definition 1.9 (**Sub-exponential random variables**). A random variable X that satisfies one of the equivalent sub-exponential properties is called a sub-exponential random variable. The sub-exponential norm of X, denoted $\|X\|_{\psi_1}$, is defined to be the smallest K_4 in property (4). In other words, we define

$$\|X\|_{\psi_1} = \inf\{t > 0 : \mathrm{E}\exp(|X|/t) \leq 2\}$$

Sub-Gaussian and sub-exponential distributions have a close relationship. Firstly, it's evident that any sub-Gaussian distribution is also sub-exponential. Secondly, the square of a sub-Gaussian random variable is itself sub-exponential.

Lemma 1.10 (**Sub-exponential is sub-Gaussian squared**). *A random variable X is sub-Gaussian if and only if X^2 is sub-exponential. Moreover,*

$$\|X^2\|_{\psi_1} = \|X\|_{\psi_2}^2$$

Proof. This follows easily from the definition. Indeed, $\|X^2\|_{\psi_1}$ is the infimum of the numbers $K > 0$ satisfying $\mathrm{E}\exp(X^2/K) \leq 2$, while $\|X\|_{\psi_2}$ is the infimum of the numbers $L > 0$ satisfying $\mathrm{E}\exp(X^2/L^2) \leq 2$. So these two become the same definition with $K = L^2$. □

More generally, the product of two sub-Gaussian random variables is sub-exponential.

Lemma 1.11 (**The product of sub-Gaussians is sub-exponential**). *Let X and Y be sub-Gaussian random variables. Then XY is sub-exponential.*

Proof. Without loss of generality assume that $\|X\|_{\psi_2} = \|Y\|_{\psi_2} = 1$. The lemma claims that if

$$\mathrm{E}\exp(X^2) \leq 2, \qquad \mathrm{E}\exp(Y^2) \leq 2$$

then $\mathrm{E}\exp(|XY|) \leq 2$.

To prove this, one may use Young's inequality $ab \leq a^2/2 + b^2/2$. It yields,

$$\begin{aligned}\mathrm{E}\exp(|XY|) &\leq \mathrm{E}\exp\left(\frac{X^2}{2} + \frac{Y^2}{2}\right) \\ &= \mathrm{E}\left(\exp\left(\frac{X^2}{2}\right)\exp\left(\frac{Y^2}{2}\right)\right) \\ &\leq \frac{1}{2}\,\mathrm{E}(\exp(X^2) + \exp(Y^2)) = 2\end{aligned}$$

□

We are ready to state and prove a concentration inequality for sums of independent sub-exponential random variables.

Theorem 1.12 (**Bernstein's inequality**). *Let $X_1, \cdots, X_N$ be independent mean-zero sub-exponential random variables. Then, for every $t \geq 0$, we have*

$$P\left(\left|\sum_{i=1}^N X_i\right|\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{\sum_{i=1}^N K_i^2}, \frac{t}{\max_i K_i}\right)\right)$$

where $K_i = \inf\{t > 0 : \mathrm{E}\exp(|X_i|/t) \leq 2\}$.

Proof. Let's define $S = \sum_{i=1}^N X_i$. Using Chebyshev's inequality, one may write,

$$\mathrm{P}(S \geq t) \leq e^{-\lambda t}\prod_{i=1}^N \mathrm{E}\exp(\lambda X_i)$$

To bound the MGF of each term X_i, we use property (5) of sub-exponentials. It says that if λ is small enough that

$$|\lambda| \leq \frac{c}{\max_i K_i}$$

then

$$\mathrm{E}\exp(\lambda X_i) \leq \exp(C\lambda^2 K_i^2)$$

Thus we have

$$\mathrm{P}(S \geq t) \leq \exp(-\lambda t + C\lambda^2\sigma^2)$$

where $\sigma^2 = \sum_{i=1}^N K_i^2$. Now we minimize λ to find the optimal choice,

$$\mathrm{P}(S \geq t) \leq \exp\left(-\min\left(\frac{t^2}{4C\sigma^2}, \frac{ct}{2K_i}\right)\right)$$

Repeating the same argument for $-X_i$ instead of X_i, the proof is complete. □

Corollary 1.13. *Let $X_1, \cdots, X_N$ be independent mean-zero sub-exponential random variables, and let $a = (a_1, \cdots, a_N) \in \mathbb{R}^N$. Then, for every $t \geq 0$, we have*

$$P\left(\left|\sum_{i=1}^N a_i X_i\right| \geq t\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{K^2\|a\|_2^2}, \frac{t}{K\|a\|_\infty}\right)\right)$$

where $K = \max_i K_i$ and $K_i = \inf\{t > 0 : \mathrm{E}\exp(|X_i|/t) \leq 2\}$.

1.4 Concentration of the Norm

To conclude this chapter, we demonstrate the norm concentration property of a standard Gaussian random variable.

Theorem 1.14 (**Concentration of the norm**). *Let $X \sim N(0, I_n)$. Then for some $c > 0$, and for all $0 < t < c$,*

$$P(|\|X\|_2 - \sqrt{n}| \geq \sqrt{n}t) \leq 2e^{-cnt^2/2}$$

Proof. We know that $\|X\|_2^2 = \sum_{i=1}^n X_i^2$. Then, for $\lambda < \frac{1}{2}$, and $1 \leq i \leq n$, one may write,

$$\mathrm{E}\, e^{\lambda X_i^2} = \sqrt{\frac{1}{1-2\lambda}}$$

Hence,

$$\begin{aligned}
\mathrm{P}\left(\left|\sum_{i=1}^n \frac{1}{n}X_i^2 - 1\right| \geq t\right) &\leq 2e^{-\lambda t}\,\mathrm{E}\exp\left(\lambda\sum_{i=1}^n\left(\frac{1}{n}X_i^2 - 1\right)\right) \\
&\leq 2\exp\left(-\lambda t - \frac{1}{2}\log(1-2\lambda))\right) \\
&\leq 2\exp\left(-\frac{n}{2}t^2\left(\frac{1}{2} - \frac{t}{3}\right)\right)
\end{aligned}$$

where from line 2 to line 3 we optimized with $\lambda^* = \frac{t}{2(1+t)} < \frac{1}{2}$ and Taylor expanded $t - \log(1+t)$ to third order. Thus if $t \leq \frac{3}{2}$, the bound is acceptable.

This is a good concentration inequality for $\|X\|_2^2$, from which we are going to deduce a concentration inequality for $\|X\|_2$. By using the fact that for all $z \geq 0$, if $|z-1| \geq \delta$ then $|z^2 - 1| \geq \max(\delta, \delta^2)$, so it yields if $t \leq 1$ one may write,

$$\mathrm{P}\left(\left|\frac{1}{\sqrt{n}}\|X\|_2 - 1\right| \geq t\right) \leq \mathrm{P}\left(\left|\frac{1}{n}\|X\|_2^2 - 1\right| \geq t\right) \leq 2e^{-cnt^2/2}$$

□

2 Gaussian Concentration

The methodology for concentration inequalities we've explored thus far heavily depends on the independence of random variables. In the following sections, we will investigate alternative approaches to concentration that do not rely on independence. Specifically, we will illustrate how to derive the concentration of Gaussian Lipschitz functions and introduce Gaussian orthogonal ensemble matrices (GOE). Furthermore, we will delve into the concentration of the top eigenvalue of GOE matrices and study the concentration phenomena related to empirical spectral measures.

2.1 Gaussian Lipschitz Concentration

In our exploration of concentration phenomena, we turn our attention to the behavior of nonlinear functions $f(X)$ of random vectors X. While it's unreasonable to anticipate strong concentration for entirely arbitrary functions f, we may observe concentration if f exhibits limited oscillations. The introduction of Lipschitz functions at this juncture enables us to rigorously identify functions with restrained oscillatory behavior, thereby aiding in our analysis.

Definition 2.1 (**Lipschitz functions**). Let (X, d_X) and (Y, d_Y) be metric spaces. A function $f : X \to Y$ is called Lipschitz, if there exists $L \in \mathbb{R}$ such that

$$d_Y(f(u), f(v)) \leq L d_X(u, v), \qquad \forall u, v \in X$$

Also, $\inf L = K$ is called the Lipschitz norm of f and f is denoted K-Lipschitz.

Recall the Gaussian measure of a Borel set $A \in \mathbb{R}^n$ can be defined as

$$\gamma_n(A) = \mathrm{P}(X \in A) = \frac{1}{(2\pi)^{n/2}} \int_A e^{-\|x\|_2^2/2} dx$$

where $X \sim I(0, I_n)$. Then we have the following Gaussian concentration inequality.

Theorem 2.2 (**Gaussian Lipschitz concentration**). *Consider a random vector $X \sim N(0, I_n)$ and a K-Lipschitz function $f : \mathbb{R}^n \to \mathbb{R}$ (with respect to the Euclidean metric). Then,*

$$P(|f(X) - \mathrm{E}\, f(X)| \geq t) \leq 2\exp(-ct^2/K^2)$$

where $c > 0$ is a universal constant.

We will prove this theorem only for smooth functions. The idea behind what we will use here is called the *smart path method*.

Proof. Let X and Y denote two independent copies of this process, and for $0 \leq t \leq 1$, let

$$X_t = \cos(t)X + \sin(t)Y$$

Note that X_t is differentiable, thus

$$\dot{X}_t = -\sin(t)X + \cos(t)Y$$

In particular, for each, X_t and $\dot{X}_t$ are both mean zero and are identically distributed.

Using the Gaussian measure, the expected value can be written as

$$\begin{aligned}
\mathrm{E}\exp(\lambda(f(X) - \mathrm{E}\, f(X))) &= \int_{\mathbb{R}^n} \exp\left[\lambda\left(f(x) - \int_{\mathbb{R}^n} f(y) d\gamma_n(y)\right)\right] d\gamma_n(x) \\
&\leq \int_{\mathbb{R}^n \times \mathbb{R}^n} \exp\left[\lambda(f(x) - f(y))\right] d\gamma_n(x) d\gamma_n(y) \\
&= \int_{\mathbb{R}^n \times \mathbb{R}^n} \exp\left[\int_0^{\pi/2} \lambda \left\langle \nabla f(X_t), \dot{X}_t \right\rangle dt\right] d\gamma_n(x) d\gamma_n(y) \\
&\leq \fint_0^{\pi/2} \int_{\mathbb{R}^n \times \mathbb{R}^n} \exp\left[\lambda \frac{\pi}{2} \langle \nabla f(x_t), \dot{x}_t \rangle\right] d\gamma_n(x_t) d\gamma_n(\dot{x}_t) dt \\
&= \fint_0^{\pi/2} \mathrm{E}\left(\int_{\mathbb{R}^n} \exp\left[\lambda \frac{\pi}{2} \left\langle \nabla f(X_t), \dot{X}_t \right\rangle\right] d\gamma_n(\dot{x}_t) \middle| \dot{X}_t = \dot{x}_t\right) dt \\
&= \fint_0^{\pi/2} \mathrm{E}\exp\left[\frac{\lambda\pi}{2} \left\langle \nabla f(X_t), \dot{X}_t \right\rangle\right] dt \\
&= \fint_0^{\pi/2} \mathrm{E}\exp\left[\frac{\lambda\pi}{2} \|\nabla f(X_t)\|^2\right] dt \\
&\leq 2\exp\left(\frac{\lambda^2\pi^2}{4} K^2\right)
\end{aligned}$$

where in the second and fourth lines we used Jensen's inequality. Also in the sixth and seventh lines we used the fact that X_t and $\dot{X}_t$ are independent, and the MGF of the normal distribution, respectively.

Thus for $\lambda, t \geq 0$, we have

$$\mathrm{P}(|f(X) - \mathrm{E}\, f(X)| \geq t) \leq 2\exp\left(-\lambda t + \frac{\lambda^2\pi^2}{4} K^2\right)$$

Recall that $\frac{a}{2}t$ has Legendre transform $\sup_{\lambda \geq 0} \lambda t - \frac{a}{2}\lambda^2 = \frac{t^2}{2a}$, so that by minimizing the above term, one may have

$$\mathrm{P}(|f(X) - \mathrm{E}\, f(X)| \geq t) \leq 2\exp\left(-\frac{t^2}{\pi^2 K^2/2}\right)$$

□

Moreover, there is a remarkable partial generalization of the Gaussian Lipschitz concentration for random vectors $X = (X_1, \cdots, X_n)$ whose coordinates are independent and have arbitrary bounded distributions. By scaling, there is no loss of generality in assuming that $|X_i| \leq 1$, but we no longer require that the X i be uniformly distributed. However, we will not cover the proof of Talagrand's concentration!

Theorem 2.3 (**Talagrand's concentration inequality**). *Consider a random vector $X = (X_1, \cdots, X_n)$ whose coordinates are independent and satisfy $|X_i| \leq 1$ almost surely. Then,*

$$P(|f(X) - \mathrm{E}\, f(X)| \leq t) \leq 2\exp(-ct^2/K^2)$$

for any convex K-Lipschitz function $f : [0,1]^n \to \mathbb{R}$.

2.2 Gaussian Convex Lipschitz Concentration

Theorem 2.4 (**Convex Concentration**). *Let $X_1, \cdots, X_N$ be independent mean-zero sub-exponential random variables and let $f : \mathbb{R}^n \to \mathbb{R}$ convex and 1-Lipschitz. Then, for any $t \geq 0$,*

$$P(|f(X) - \mathrm{E}\, f(X)| > t) \leq 2\exp\left(-c\min\left(\frac{t}{\|\max_i |X_i|\|_{\psi_1}}, \frac{t^2}{\|\max_i |X_i|\|_{\psi_2}^2}\right)\right)$$

Proof. One may write,

$$X_i = X_i \mathbf{1}_{|X_i| \leq M} + X_i \mathbf{1}_{|X_i > M} =: Y_i + Z_i$$

with $M := 8\,\mathrm{E}\max_i |X_i|$, and let $Y = (Y_1, \cdots, Y_N)$, and $Z = (Z_1, \cdots, Z_N)$. By Lipschitz property of f,

$$\begin{aligned}\mathrm{P}(|f(X) - \mathrm{E}\, f(X)| \geq t) &\leq \mathrm{P}(|f(Y) - \mathrm{E}\, f(Y)| + |f(X) - f(Y)| + |\,\mathrm{E}\, f(Y) - \mathrm{E}\, f(X)| \geq t) \\ &\leq \mathrm{P}(|f(Y) - \mathrm{E}\, f(Y)| + \|Z\|_2 + \mathrm{E}\,\|Z\|_2 > t)\end{aligned}$$

and hence it suffices to bound the terms in the last line. Recall Talagrand's concentration inequality of the bounded random variables, we obtain,

$$\mathrm{P}(|f(Y) - \mathrm{E}\, f(Y)| > t) \leq 2\exp\left(-c\min\left(\frac{t}{\|\max_i |X_i|\|_{\psi_1}}, \frac{t^2}{\|\max_i |X_i|\|_{\psi_2}^2}\right)\right)$$

Furthermore, we know that

$$\|\|Z\|_2\|_{\psi_{1,2}} \leq C_1 \|\max_i |X_i|\|_{\psi_{1,2}}$$

Hence, for any $t \geq 0$,

$$\mathrm{P}(\|Z\|_2 \geq t) \leq 2\exp\left(-c\min\left(\frac{t}{\|\max_i |X_i|_{\psi_1}}, \frac{t^2}{\|\max_i |X_i|\|_{\psi_2}^2}\right)\right)$$

and also one may write,

$$\mathrm{E}\,\|Z\|_2 \leq C_{1,2} \|\max_i |X_i|\|_{\psi_{1,2}}$$

Define K be the minimum of $C_1\|\max_i |X_i|\|_{\psi_1}$ and $C_2\|\max_i |X_i|\|_{\psi_2}$. It yields,

$$\mathrm{P}(|f(X) - \mathrm{E}\, f(X)| \geq t) \leq \mathrm{P}(|f(Y) - \mathrm{E}\, f(Y)| + \|Z\|_2 \geq t - K$$

if $t \geq K$. Using subadditivity, we obtain

$$\mathrm{P}(|f(X) - \mathrm{E}\, f(X)| \geq t) \leq 4\exp\left(-\frac{(t-K)}{2K} \leq 4\exp(-\frac{t}{2cK}\right)$$

where the last step holds for $t \geq K + \delta$ for some $\delta > 0$. This bound extends trivially to any $t \geq 0$ by a suitable change of constants. □

2.3 Top Eigenvalue of the GOE Matrices

Next, we introduce the Gaussian Orthogonal Ensemble (GOE), a significant class of random matrices with broad applications in statistics, data science, and mathematical physics.

Definition 2.5 (**GOE matrix**). A matrix $A = (A_{ij}) \in M_N(\mathbb{R})$ is called the Gaussian orthogonal ensemble (GOE) if it is symmetric and its upper triangular entries are given by $A_{ij} \sim N(0, \frac{1+\delta_{ij}}{N})$, where the entries in the upper triangular part are independent. If A is a GOE matrix of dimension $N \times N$, we write $A \sim \mathrm{GOE}(N)$.

One question you might ask yourself is what can you say about the top-eigenvalue of $A \sim \mathrm{GOE}(N)$. To discuss the concentration for the top eigenvalue, we need the following important theorem.

Theorem 2.6 (**Borell's inequality**). *Let T be a compact metric space and $\{X_t\}_{t\in T}$ a centered Gaussian process. Suppose that the covariance kernel $K : T \times T \to \mathbb{R}$ given by $K(t,s) = \mathrm{E}\, X_t X_s$ is continuous and $\sup_{t\in T} X_t < \infty$ almost surely. Then $\mathrm{E}\sup_{t\in T} X_t < \infty$ and*

$$P\left(\left|\sup_{s\in T} X_s - \mathrm{E}\sup_{s\in T} X_s\right| > t\right) \le 2\exp(-t^2/2_T)$$

where $\sigma_T^2 = \sup_{t\in T} \mathrm{E}\, X_t^2$.

Proof. Step 1: T is finite, i.e., $T = \{1, \cdots, \tau\}$. Let K be $\tau \times \tau$ covariance matrix of X on T, then $K(t,s) = \mathrm{E}\, X_t X_s$, and

$$\sigma_T^2 = \max_{1\le t\le \tau} K(t,t) = \max_{1\le t\le \tau} \mathrm{E}\, X_t^2$$

Also, let Z be vector of independent standard Gaussian, i.e., $Z \sim N(0, I_n)$, and Σ such that $\Sigma^T\Sigma = K$. It yields $X \overset{d}{=} \Sigma Z$ and $\max_{1\le t\le\tau} X_t \overset{d}{=} \max_{1\le t\le\tau}(\Sigma Z)_t$. Consider the function $h(x) = \max_{1\le t\le\tau}(\Sigma x)_t$, then

$$\begin{aligned}
|h(x) - h(y)| &= \left|\max_{1\le t\le\tau}(Ax)_t - \max_{1\le t\le\tau}(Ay)_t\right| \\
&= \left|\max_{1\le t\le\tau}(e_t Ax) - \max_{1\le t\le\tau}(e_t Ay)\right| \\
&\le \max_{1\le t\le\tau} |e_t A(x-y)| \\
&\le \max_{1\le t\le\tau} |e_t A| \cdot |x-y|
\end{aligned}$$

where in the third and fourth lines we used the triangular and Cauchy-Schwarz inequalities. Moreover, one may write for all $t \in T$,

$$|e_t A|^2 = e_t^T A^T A e_t = e_t^T K e_t = K(t,t)$$

Hence, $|h(x) - h(y)| \le \sigma_T |x-y|$ and we showed that h is a Lipschitz function. So using the Lipschitz concentration,

$$\mathrm{P}(|h(X) - \mathrm{E}\, h(X)| \ge t) \le \exp(-t^2/2\sigma_T^2)$$

Since $h(X)$ and $\max_{1\le s\le\tau}(X_s)$ are equal in law, thus

$$\mathrm{P}\left(\left|\max_{1\le s\le\tau} X_t - \mathrm{E}\max_{1\le s\le\tau} X_t\right| \ge t\right) \le 2\exp(-t^2/2\sigma_T^2)$$

Step 2: General T. The idea is to use the compactness as the generalization of being finite. But, first we seek to show that $\mathrm{E}\sup_{t\in T} X_t$ is finite. By contradiction, assume $\mathrm{E}\sup t \in T X_t = \infty$ and choose $u_0 > 0$ such that

$$\exp(-u_0^2/\sigma_T^2) \le \frac{1}{4}, \quad \text{and,} \quad \mathrm{P}(\sup_{t\in T} X_t < u_0) \ge \frac{3}{4}$$

Choose n so that $\mathrm{E}\sup_{t\in T_n} f(X_t) > 2u_0$, then by Borell-Tsirelson-Ibragimov-Sudakov inequality,

$$\begin{aligned}\frac{1}{2} \geq 2\exp(-u_0^2/\sigma_T^2) &\geq \mathrm{P}\left(\left|\sup_{t\in T_n} X_t - \mathrm{E}\sup_{t\in T_n} X_t\right| > u_0\right) \\ &\geq \mathrm{P}\left(\mathrm{E}\sup_{t\in T_n} X_t - \sup_{t\in T} X_t > u_0\right) \\ &\geq \mathrm{P}(\sup_{t\in T} X_t < u_0) \geq \frac{3}{4}\end{aligned}$$

Then the contradiction results and it yields $\mathrm{E}\sup_{t\in T} X_t < \infty$.

Now, let T_n be a finite subset of T such that $T_n \subseteq T_{n-1}$ and T_n increases to a dense subset of T as T is compact. By separability,

$$\sup_{t\in T_n} X_t \overset{a.s.}{\to} \sup_{t\in T} X_t$$

Using the Monotone Convergence Theorem,

$$\mathrm{P}(\sup_{t\in T_n} X_t \geq u) \to \mathrm{P}(\sup_{t\in T} X_t \geq u)$$

and,

$$\mathrm{E}\sup_{t\in T_n} X_t \to \mathrm{E}\sup_{t\in T} X_t$$

As $\sigma_{T_n}^2 \uparrow \sigma_T^2$, the proof is complete. □

We can now apply Borell's inequality to discuss the top eigenvalue of a GOE matrix.

Theorem 2.7 (**Top eigenvalue of a GOE matrix**). *Suppose $A \sim GOE(N)$. Then,*

$$P(|\lambda_1(A) - \mathrm{E}\,\lambda_1(A)| > t) \leq 2\exp(-Nt^2/2)$$

Proof. Since A is Hermitian, we may write the eigenvalues of A as $\lambda_1(A) \leq \cdots \leq \lambda_N(A)$, so by the Courant-Fisher theorem,

$$\lambda_N(A) = \sup_{\|v\|_2=1} \langle v, Av\rangle$$

Define R_A as $R_A(v) = \langle v, Av\rangle$ for a unit vector v, and call this quantity the Rayleigh quotient. Note that $R_A(v)$ is a linear combination of the entries of A, and since A is a GOE matrix, $\{R_A(v)\}_{v\in S^{N-1}}$ is a centered Gaussian process. The covariance matrix can be computed to be

$$K(u,v) = \mathrm{E}\,R(u)R(v) = \mathrm{E}\langle u, Au\rangle\langle v, Av\rangle = \frac{1}{N}(\langle u, v\rangle^2 + \langle u, v\rangle)$$

Obviously the covariance kernel is continuous. We may now apply Borell's inequality,

$$\mathrm{P}(|\lambda_N(A) - E\lambda_N(A)| \geq t) = \mathrm{P}\left(\left|\sup_{v\in S^{N-1}} \langle v, Av\rangle - \mathrm{E}\sup_{v\in S^{N-1}} \langle v, Av\rangle\right| \geq t\right) \leq 2\exp(-Nt^2/2)$$

□

2.4 Empirical Spectral Measure

Consider a probability measure μ_N that is uniformly distributed on the sample $X_1, c\ldots, X_N$, that is,

$$\mu_N(X) = \frac{1}{N}, \qquad \forall i = 1, \cdots, n$$

Note that μ_N is a random measure. It is called the empirical measure. Now suppose the case that we study random matrices and their spectrum. For this reason, we define the empirical spectral measure as follows.

Definition 2.8 (**Empirical spectral measure**)**.** Let A be an $N \times N$ random matrix. Its empirical spectral measure is the random measure given by

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

where $\{\lambda_1, \cdots, \lambda_N\}$ is the spectrum of A.

Before we state the theorem, we need a metric for the space of probabilities of $\mathbb{R}$.

Definition 2.9 (**Wasserstein distance**)**.** Let $\mu, \nu \in \text{Prob}(\mathbb{R})$ be Radon probability measure. The Wasserstein distance between μ and ν is the quanitity,

$$d_W(\mu, \nu) = \sup_{f \in \mathcal{L}} \left| \int f d\mu - \int f d\nu \right|$$

where $\mathcal{L} = \{f \in C_b(\mathbb{R}) : \|f\|_\infty \leq \text{ and f is } K\text{-Lipschitz such that } K \leq 1\}$.

A consequence of Hoffman-Wielandt theorem (see Appendix C) is the following lemma.

Lemma 2.10. *Let $g : \mathbb{R}^n \to \mathbb{R}$ is 1-Lipschitz function. Then $A \mapsto g(\lambda_1(A), \cdots, \lambda_N(A))$ is 1-Lipschitz with resepct to the Frobenius norm. Moreover, if $f : \mathbb{R} \to \mathbb{R}$ is 1-Lipschitz function, then the map $A \mapsto \text{tr}(f(A))$ is 1-Lipschitz.*

The concentration of the Wasserstein distance between the empirical spectral measure and any arbitrary measure is attributed to several people.

Theorem 2.11 (**Ben Arous, Guionnet, Zeitouni**)**.** *For any μ, $N \geq 1$:*

$$P(d_W(\mu_N, \mu) - \mathrm{E}\, d_W(\mu_N, \mu) \geq t) \leq 2 \exp(-cN^2 t^2)$$

Proof. Let $f(A) = d_W(\mu_N(A), \mu)$. One may write,

$$\begin{aligned} |f(A) - f(B)| &= \sup_{g \in \mathcal{L}} \left(\frac{1}{N} \sum_{i=1}^{N} (g(\lambda_i(A)) - g(\lambda_i(B))) \right) \\ &= \frac{1}{N} \left\| \sum_{i=1}^{N} g(\lambda_i(A)) - g(\lambda_i(B)) \right\|_\infty \\ &\leq \frac{1}{N} \sum_{i=1}^{N} \|g(\lambda_i(A) - g(\lambda_i(B))\|_\infty \end{aligned}$$

Then f is $1/N$-Lipschitz. Furthermore, the map $A \mapsto \frac{1}{N} \text{tr}(f(A))$ is $1/N$-Lipschitz for all $f \in \mathcal{L}$. Using the Gaussian Lipschitz concentration,

$$\mathrm{P}(d_W(\mu_N, \mu) - \mathrm{E}\, d_W(\mu_N, \mu) \geq t) \leq 2 \exp(-cN^2 t^2)$$

□

2.5 Operator Norm

Recall the concentration of the norm for the Gaussian random variables. Now we use Borell's inequality, or other results from Gaussian concentration, to show that there is a concentration for the operator norm of a GOE matrix and this inequality tells us that our random variable, $\|X\|_{op}$ lives aorund some constant. Howover, this inequality does not tell anything about what that constant is.

Theorem 2.12 (**Norm concentration of the GOE matrices**)**.** *Suppose $A \sim GOE(N)$. Then,*

$$P(\|A\|_{op} > C(1 + t)) \leq \exp(-ct^2 N)$$

Proof. Step 1: Approximation. We know that the covering numbers of the unit Euclidean sphere S^{n-1} satisfy the following for any $\epsilon > 0$, Step 2: Concentration. Fix $x \in \mathcal{N}$ and $y \in \mathcal{M}$. Then the quadratic form

$$\langle Ax, y\rangle = \sum_{i=1}^{n}\sum_{j=1}^{m} A_{ij}x_i y_j$$

is a sum of independent random variables. According to Hoeffding's inequality, for all $t > 0$,

$$\mathrm{P}\left(\left|\sum_{i,j=1}^{N} A_{ij}x_i x_j\right| \geq t\right) \leq 2\exp\left(-\frac{t^2}{2\max_{i,j}\sigma_{i,j}^2\|a\|_2^2}\right)$$

where $\max_{i,j}\sigma_{i,j}^2 = \frac{2}{N}$, and $\|a\|_2 = 1$.

Step 3: Union bound. Suppose that the event $\max_{x\in\mathcal{N}}\langle Ax, x\rangle \geq u$ occurs. Then there exits $x \in \mathcal{N}$ such that $\langle Ax, y\rangle \geq u$. Thus the union bound yields,

$$\mathrm{P}\left(\max_{x\in N}\langle Ax, x\rangle \geq u\right) \leq \sum_{x\in\mathcal{N}} \mathrm{P}(\langle Ax, x\rangle \geq u) \leq 9^n \times \mathrm{P}\left(\langle Ax, y\rangle \geq u\right)$$

Therefore using the above results, one may write,

$$\mathrm{P}(\|A\|_{op} \geq u) \leq \mathrm{P}\left(\max_{x\in N}\langle Ax, x\rangle \geq 2u\right) \leq 2\exp(N(\log 9 - u^2))$$

Choose $u = C(1+t)$. Then $u^2 \geq C^2(1+t^2)$, and if the constant C is chose sufficiently large, finally,

$$\mathrm{P}(\|A\|_{op} > C(1+t)) \leq 2\exp(-ct^2N)$$

□

2.6 Hanson-Wright Inequality

Hanson-Wright inequality is a general concentration result for quadratic forms in sub-gaussian random variables. In this section we give a proof of Hanson-Wright inequality, whici is useful for the proof of the semicircle law in the next chapter.

Theorem 2.13 (**Hanson-Wright inequality**). *Let $X \sim N(0, I_N)$ be a centred Gaussian random vector. Let A be an $N \times N$ matrix. Then for every $t \geq 0$,*

$$P(|X^TAX - \mathrm{E}\,X^TAX| \geq t) \leq 2\exp\left(-c\min\left(\frac{t^2}{\|A\|_F^2}, \frac{t}{\|A\|_{op}^2}\right)\right)$$

Proof. Let $A = (a_{ij})_{i,j=1}^n$. By independence and zero mean of X_i, we can represent

$$\begin{aligned} X^\top AX - \mathbb{E}X^\top AX &= \sum_{i,j} a_{ij}X_iX_j - \sum_i a_{ii}\mathbb{E}X_i^2 \\ &= \sum_i a_{ii}\left(X_i^2 - \mathbb{E}X_i^2\right) + \sum_{i,j:i\neq j} a_{ij}X_iX_j. \end{aligned}$$

The problem reduces to estimating the diagonal and off-diagonal sums:

$$p \leq \mathbb{P}\left\{\sum_i a_{ii}\left(X_i^2 - \mathbb{E}X_i^2\right) > t/2\right\} + \mathbb{P}\left\{\sum_{i,j:i\neq j} a_{ij}X_iX_j > t/2\right\} =: p_1 + p_2.$$

Step 1: diagonal sum. Note that $X_i^2 - \mathbb{E}X_i^2$ are independent mean-zero subexponential random variables, and

$$\left\|X_i^2 - \mathbb{E}X_i^2\right\|_{\psi_1} \leq 2\left\|X_i^2\right\|_{\psi_1} \leq 4\left\|X_i\right\|_{\psi_2}^2 \leq 4K^2.$$

Then we can use a Bernstein-type inequality and obtain

$$p_1 \le \left[-c\min\left(\frac{t^2}{\sum_i a_{ii}^2}, \frac{t}{\max_i |a_{ii}|}\right)\right] \le \exp\left[-c\min\left(\frac{t^2}{\|A\|_{\mathrm{HS}}^2}, \frac{t}{\|A\|}\right)\right].$$

Step 2: decoupling. It remains to bound the off-diagonal sum

$$S := \sum_{i,j:i\neq j} a_{ij} X_i X_j.$$

The argument will be based on estimating the moment generating function of S by decoupling and reduction to normal random variables.

Let $\lambda > 0$ be a parameter whose value we will determine later. By Chebyshev's inequality, we have

$$p_2 = \mathbb{P}\{S > t/2\} = \mathbb{P}\{\lambda S > \lambda t/2\} \le \exp(-\lambda t/2)\mathbb{E}\exp(\lambda S).$$

Consider independent Bernoulli random variables $\delta_i \in \{0,1\}$ with $\mathbb{E}\delta_i = 1/2$. Since $\mathbb{E}\delta_i(1-\delta_j)$ equals $1/4$ for $i \neq j$ and 0 for $i = j$, we have

$$S = 4\mathbb{E}_\delta S_\delta, \quad \text{where} \quad S_\delta = \sum_{i,j} \delta_i(1-\delta_j) a_{ij} X_i X_j.$$

Here $\mathbb{E}_\delta$ denotes the expectation with respect to $\delta = (\delta_1, \dots, \delta_n)$. Jensen's inequality yields

$$\mathbb{E}\exp(\lambda S) \le \mathbb{E}_{X,\delta}\exp(4\lambda S_\delta)$$

where $E_{X,\delta}$ denotes expectation with respect to both X and δ. Consider the set of indices $\Lambda_\delta = \{i \in [n] : \delta_i = 1\}$ and express

$$S_\delta = \sum_{i\in\Lambda_\delta, j\in\Lambda_\delta^c} a_{ij} X_i X_j = \sum_{j\in\Lambda_\delta^c} X_j \left(\sum_{i\in\Lambda_\delta} a_{ij} X_i\right).$$

Now we condition on δ and $(X_i)_{i\in\Lambda_\delta}$. Then S_δ is a linear combination of meanzero sub-gaussian random variables $X_j, j \in \Lambda_\delta^c$, with fixed coefficients. It follows that the conditional distribution of S_δ is sub-gaussian, and its sub-gaussian norm is bounded by the ℓ_2-norm of the coefficient vector. Specifically,

$$\|S_\delta\|_{\psi_2} \le C\sigma_\delta \quad \text{where} \quad \sigma_\delta^2 := \sum_{j\in\Lambda_\delta^c} \left(\sum_{i\in\Lambda_\delta} a_{ij} X_i\right)^2.$$

Next, we use a standard estimate of the moment generating function of centered sub-gaussian random variables. It yields

$$\mathbb{E}_{(X_j)_{j\in\Lambda_\delta^c}} \exp(4\lambda S_\delta) \le \exp\left(C\lambda^2 \|S_\delta\|_{\psi_2}^2\right) \le \exp\left(C'\lambda^2\sigma_\delta^2\right).$$

Taking expectations of both sides with respect to $(X_i)_{i\in\Lambda_\delta}$, we obtain

$$\mathbb{E}_X \exp(4\lambda S_\delta) \le \mathbb{E}_X \exp\left(C'\lambda^2\sigma_\delta^2\right) =: E_\delta.$$

Recall that this estimate holds for every fixed δ. It remains to estimate E_δ.

Step 3: reduction to normal random variables. Consider $g = (g_1, \dots, g_n)$ where g_i are independent $N(0,1)$ random variables. The rotation invariance of normal distribution implies that for each fixed δ and X, we have

$$Z := \sum_{j\in\Lambda_\delta^c} g_j \left(\sum_{i\in\Lambda_\delta} a_{ij} X_i\right) \sim N\left(0, \sigma_\delta^2\right).$$

By the formula for the moment generating function of normal distribution, we have $\mathbb{E}_g \exp(sZ) = \exp\left(s^2\sigma_\delta^2/2\right)$. Comparing this with the formula defining E_δ, we find that the two expressions are somewhat similar. Choosing $s^2 = 2C'\lambda^2$, we can match the two expressions as follows:

$$E_\delta = \mathbb{E}_{X,g}\exp(C_1\lambda Z)$$

where $C_1 = \sqrt{2C'}$. Rearranging the terms, we can write $Z = \sum_{i\in\Lambda_\delta} X_i \left(\sum_{j\in\Lambda_\delta^c} a_{ij}g_j\right)$. Then we can bound the moment generating function of Z in the same way we bounded the moment generating function of S_δ in Step 2, only now relying on the sub-gaussian properties of $X_i, i \in \Lambda_\delta$. We obtain

$$E_\delta \le \mathbb{E}_g \exp\left[C_2\lambda^2 \sum_{i\in\Lambda_\delta}\left(\sum_{j\in\Lambda_\delta^c} a_{ij}g_j\right)^2\right].$$

To express this more compactly, let P_δ denotes the coordinate projection (restriction) of $\mathbb{R}^n$ onto $\mathbb{R}^{\Lambda_\delta}$, and define the matrix $A_\delta = P_\delta A\,(I - P_\delta)$. Then what we obtained

$$E_\delta \le \mathbb{E}_g \exp\left(C_2\lambda^2 \|A_\delta g\|_2^2\right).$$

Recall that this bound holds for each fixed δ. We have removed the original random variables X_i from the problem, so it now becomes a problem about normal random variables g_i.

Step 4: calculation for normal random variables. By the rotation invariance of the distribution of g, the random variable $\|A_\delta g\|_2^2$ is distributed identically with $\sum_i s_i^2 g_i^2$ where s_i denote the singular values of A_δ. Hence by independence,

$$E_\delta = \mathbb{E}_g \exp\left(C_2\lambda^2 \sum_i s_i^2 g_i^2\right) = \prod_i \mathbb{E}_g \exp\left(C_2\lambda^2 s_i^2 g_i^2\right).$$

Note that each g_i^2 has the chi-squared distribution with one degree of freedom, whose moment generating function is $\mathbb{E}\exp\left(tg_i^2\right) = (1-2t)^{-1/2}$ for $t < 1/2$. Therefore

$$E_\delta \le \prod_i \left(1 - 2C_2\lambda^2 s_i^2\right)^{-1/2} \qquad \text{provided} \qquad \max_i C_2\lambda^2 s_i^2 < 1/2.$$

Using the numeric inequality $(1-z)^{-1/2} \le e^z$ which is valid for all $0 \le z \le 1/2$, we can simplify this as follows:

$$E_\delta \le \prod_i \exp\left(C_3\lambda^2 s_i^2\right) = \exp\left(C_3\lambda^2 \sum_i s_i^2\right) \qquad \text{provided} \qquad \max_i C_3\lambda^2 s_i^2 < 1/2.$$

Since $\max_i s_i = \|A_\delta\| \le \|A\|$ and $\sum_i s_i^2 = \|A_\delta\|_{\mathrm{HS}}^2 \le \|A\|_{\mathrm{HS}}$, we have proved the following:

$$E_\delta \le \exp\left(C_3\lambda^2 \|A\|_{\mathrm{HS}}^2\right) \text{ for } \lambda \le c_0/\|A\|.$$

This is a uniform bound for all δ. Now we take expectation with respect to δ and we obtain the following estimate on the moment generating function of S :

$$\mathbb{E}\exp(\lambda S) \le \mathbb{E}_\delta E_\delta \le \exp\left(C_3\lambda^2 \|A\|_{\mathrm{HS}}^2\right) \qquad \text{for } \lambda \le c_0/\|A\|.$$

Step 5: conclusion. Putting this estimate into the exponential Chebyshev's inequality, we obtain

$$p_2 \le \exp\left(-\lambda t/2 + C_3\lambda^2 \|A\|_{\mathrm{HS}}^2\right) \qquad \text{for } \lambda \le c_0/\|A\|.$$

Optimizing over λ, we conclude that

$$p_2 \le \exp\left[-c\min\left(\frac{t^2}{\|A\|_{\mathrm{HS}}^2}, \frac{t}{\|A\|}\right)\right] =: p(A,t).$$

Now we combine with a similar estimate for p_1 and obtain

$$p = p_1 + p_2 \le 2p(A,t).$$

Repeating the argument for $-A$ instead of A, we get $\mathbb{P}\left\{X^\top AX - \mathbb{E}X^\top AX < -t\right\} \le 2p(A,t)$. Combining the two events, we obtain $\mathbb{P}\left\{\left|X^\top AX - \mathbb{E}X^\top AX\right| > t\right\} \le 4p(A,t)$. Finally, one can reduce the factor 4 to 2 by adjusting the constant c in $p(A,t)$. The proof is complete [4]. □

3 Wigner Matrices

3.1 Definition

In this section, we discuss the concept of random matrices and explore Wigner's semicircle law, as well as the Marchenko-Pastur distribution. We use two approaches, the method of moments and the Stieltjes transform, to demonstrate these significant results. First, we introduce a basic model of random matrices.

Definition 3.1 (**Wigner matrix**). Let $\{Z_{i,j}\}_{1\le i<j}$ and $\{Y_i\}_{1\le i}$ be two independent families of iid zero mean, real-valued random variables such that $\mathrm{E}\, Z_{1,2}^2 = 1$, and for all integers $k \ge 1$,

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

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

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

We call such a matrix a *Wigner matrix*, and if the random variables $Z_{i,j}$ and Y_i are Gaussian, we use the term *Gaussian Wigner matrix*. The case of Gaussian Wigner matrices in which $\mathrm{E}\, Y_1^2 = 2$ are referred to as *Gaussian orthogonal ensemble (GOE) matrices*.

Recall the concentration we provided in previous chapter, we have the following useful lemma.

Lemma 3.2. *Let $g : \mathbb{R}^N \to \mathbb{R}$ be a K-Lipschitz function. Then with X denoting the Hermitian matrix wit entries X_{ij}, the map*

$$\{X_{ij}\}_{1\le i\le j\le N} \mapsto g(\lambda_1(X), \cdots, \lambda_N(X))$$

is a $\sqrt{2}K$-Lipschitz function on $\mathbb{R}^{N^2}$. In particular, if f is a K-Lipschitz function on $\mathbb{R}$,

$$\{X_{ij}\}_{1\le i\le j\le N} \mapsto tr(f(X))$$

is a $\sqrt{2N}K$-Lipschitz function on $\mathbb{R}^{N(N+1)}$.

Theorem 3.3 (**Wigner matrix concentration**). *Let X_N be a Gaussian Wigner matrix. Then for any K-Lipschitz function f on $\mathbb{R}$, for any $\delta > 0$,*

$$P(|tr(f(X_N)) - \mathrm{E}\, tr(f(X_N))| \ge Nt) \le 2\exp(-Nt^2/4cK^2)$$

$$P(|f(\lambda_k^N) - \mathrm{E}\, f(\lambda_k^N)| \ge t) \le 2\exp(-t^2/4cK^2)$$

Proof. Define the function $G : \mathbb{R}^{N^2} \to \mathbb{R}$ as follows,

$$G(X_{N_{i,j}}, 1 \le i \le j \le N) = \mathrm{tr}(f(X_N))$$

From previous lemma, we know that if f is K-Lipschitz, G is also $\sqrt{2N}K$-Lipschitz. To see the next result, apply the same argument to the function

$$\tilde{G}(X_{N_{i,j}}, 1 \le i \le j \le N) = f(\lambda_k(X_N))$$

□

Before starting our essential theorems, first we need to get familiar with the concept of resolvent formalism.

3.2 Resolvent Formalism

Definition 3.4 (**Resolvent formalism**). Let A_N be $N \times N$ Hermitian. The resolvent of A_N is defined as follows,

$$R_{A_N}(z) = \frac{1}{zI_N - A_N}$$

Definition 3.5 (**Stieltjes transform**). Let μ be a positive, finite measure on the real line. The Stieltjes transform of μ is the function

$$S_\mu(z) := \int_{\mathbb{R}} \frac{\mu(dx)}{z-x}, \quad z \in \mathbb{C} \setminus \mathbb{R}$$

Observe that without loss of generality, if $z \in \mathbb{C}_+$, the map $x \in \mathbb{R} \mapsto 1/(z-x)$ is continuous and uniformly bounded, i.e.,

$$\frac{1}{|z-x|} = \frac{1}{\sqrt{(x-\mathrm{Re}(z))^2 + (\mathrm{Im}(z))^2}} \leq \frac{1}{|\mathrm{Im}(z)|} \implies |S_\mu(z)| \leq \mu(\mathbb{R})/|\mathrm{Im}(z)|$$

Theorem 3.6 (**Inversion formula of Stieltjes transforms**). *For any open interval I with neither endpoint on an atom of μ,*

$$\mu(I) = \lim_{\epsilon \to 0} \frac{1}{2\pi i} \int_I (S_\mu(\lambda + i\epsilon) - S_\mu(\lambda - i\epsilon)) d\lambda = \lim_{\epsilon \to 0} \frac{1}{2\pi i} \int_I \mathrm{Im}(S_\mu(\lambda + i\epsilon)) d\lambda$$

Proof. Note that since

$$\mathrm{Im}(S_\mu(i)) = \int \frac{1}{1+x^2} \mu(dx)$$

we have $S_\mu \equiv 0$ implies $\mu = 0$. Assume that S_μ does not vanish identically, then since

$$\lim_{y \uparrow \infty} y \mathrm{Im} S_\mu(iy) = \lim_{y \uparrow \infty} \int \frac{y^2}{x^2+y^2} \mu(dx) = \mu(\mathbb{R})$$

by bounded convergence, we may and will assume that $\mu(\mathbb{R}) = 1$, i.e., that μ is a probability measure.

Let X be distributed according to μ, and denote by C_ϵ a random variable, independent of X, Cauchy distributed with parameter ϵ, i.e., the law of C_ϵ has density $\epsilon dx/\pi(x^2+\epsilon^2)$. Then $\mathrm{Im}(S_\mu(\lambda+i\epsilon))/\pi$ is nothing but the density (with respect to Lebesgue measure) of the law of $X + C_\epsilon$ evaluated at $\lambda \in \mathbb{R}$. Then just a rewriting of the weak convergence of the law of $X + C_\epsilon$ so that of X as $\epsilon \to 0$. Thus the convergence of $\mu(I)$ is obtained. □

Theorem 3.7. *Let $\mu_n \in M_1(\mathbb{R})$ be a sequence of probability measures.*

1. *$\mu_n \to \mu$ weakly if and only if $S_{\mu_n}(z) \to S_\mu(z)$ for each $z \in \mathbb{C} \setminus \mathbb{R}$.*
2. *If μ is a probability measure, then if $S_{\mu_n}(z) \to S(z)$, then $\mu_n \to \mu$ weakly.*
3. *If μ_n are random and $S_{\mu_n}(z) \to S_\mu(z)$ for deterministic μ, then $\mu_n \to \mu$ weakly in probability.*

Proof. Part (1) is a restatement of the notion of weak convergence.

For part (2), let n_k be a subsequence on which μ_{n_k} converges to a probability measure μ. Because $x \mapsto 1/(z-x)$, for $z \in \mathbb{C} \setminus \mathbb{R}$ is continuous and decays to zero at infinity, one obtains the convergence $S_{\mu_{n_k}}(z) \to S_\mu(z)$ pointwise for such z. From the hypothesis, it follows that $S(z) = S_\mu(z)$. Since S_μ is unique, hence $\mu_n \to \mu$ weakly.

To see part (c), fix a sequence $z_i \to z_0$ in $\mathbb{C}\setminus\mathbb{R}$ with $z_i \neq z_0$, and define, for $\nu_1, \nu_2 \in M_1(\mathbb{R})$,

$$\rho(\nu_1, \nu_2) = \sum_i 2^{-i} |S_{\nu_1}(z_i) - S_{\nu_2}(z_i)|$$

Note that $\rho(\nu_n, \nu) \to 0$ implies that ν_n converges weakly to ν. Indeed, moving to a subsequence if necessary, ν_n converges vaguely to some probability measure θ, and thus $S_{\nu_n}(z_i) \to S_\theta(z_i)$ for each i. On the other hand, the uniform (in i, n) boundedness of $S_{\nu_n}(z_i)$ and $\rho(\nu_n, \nu) \to 0$ imply that $S_{\nu_n}(z_i) \to S_\nu(z_i)$. Thus, $S_\nu(z) = S_\theta(z)$ for all $z = z_i$ and hence, for all $z \in \mathbb{C}\setminus\mathbb{R}$ since the set $\{z_i\}$ possesses an accumulation point and S_ν, S_θ are analytic. By the inversion formula, it follows that $\nu = \theta$ and in particular θ is a probability measure and ν_n converges weakly to $\theta = \nu$. From the assumption of part (c) we have that $\rho(\mu_n, \mu) \to 0$, in probability, and thus μ_n converges weakly to μ in probability, as claimed. □

3.3 The Semicircle Law

The semicircle law is a fundamental result in random matrix theory, describing the distribution of eigenvalues of large random matrices with independent and identically distributed entries. Formulated by Eugene Wigner in the 1950s to model the behavior of complex atomic nuclei, it asserts that as the size of the matrix grows infinitely large, the distribution of eigenvalues converges to a semicircular shape in the complex plane. This elegant theorem has found applications in various fields, including physics, statistics, and computer science, providing insights into the behavior of complex systems and the properties of random matrices.

Definition 3.8 (**The semicircle law**). The semicircle law as the probability distribution $\mu_{sc}(x)dx$ on $\mathbb{R}$ can be defined with density

$$\mu_{sc}(x) = \frac{1}{2\pi}\sqrt{4-x^2}\mathbf{1}_{|x|\leq 2}$$

Theorem 3.9 (**Wigner's semicircle law**). *Let W be an $N \times N$ Wigner matrix. The empirical spectral measure μ_N converges weakly in probability to the semicircle law, μ_{sc} i.e.,*

$$\lim_{N\to\infty} P\left(\left|\int f\mu_N - \int f\mu_{sc}\right| > \epsilon\right) = 0, \qquad \forall f \in C_b(\mathbb{R})$$

In the following we will prove this important theorem using two approaches: Resolvent formalism and the method of moments.

Lemma 3.10 (**Stieltjes transform of the ESM**). *The Stieltjes transform of μ_N has the identity,*

$$S_N(z) = \int_{\mathbb{R}} \frac{1}{z-x} d\mu_N(x) = \frac{1}{N}\operatorname{tr}\left(R_W(z)\right)$$

where W be a $N \times N$ Wigner matrix.

Proof. Let $\lambda_1(W) \geq \cdots \lambda_N(W)$ be the eigenvalues of W. Let e_j be an eigenvector corresponding to λ_j. So,

$$We_j = \lambda_j e_j \implies (zI_N - W)e_j = (z-\lambda_j)e_j$$

Then one may write,

$$(zI_N - W)^{-1}e_j = (z-\lambda_j)^{-1}e_j$$

Hence,

$$\int_{\mathbb{R}} \frac{1}{z-x} d\mu_N(x) = \frac{1}{N}\operatorname{tr}(R_W(z)) = \frac{1}{N}\operatorname{tr}(zI_N - W)$$

□

Lemma 3.11 (**Stieltjes transform of the semicircle law**). *The Stieltjes transform of the semicircle law μ_{sc} is given by*

$$S_{sc}(z) = \frac{1}{2}\left(z - \sqrt{z^2-4}\right)$$

Proof. Fix $z \in \mathbb{C} \setminus \mathbb{R}$. Then,

$$S_{sc}(z) = \int_{\mathbb{R}} \frac{1}{z-x} d\mu_{sc}(x) = \frac{1}{2\pi}\int_{-2}^{2} \frac{1}{z-x}\sqrt{4-x^2}dx$$

Let $x = 2\cos y$. One may write,

$$S_{sc}(z) = \frac{1}{2\pi}\int_{\pi}^{0} \frac{1}{z-2\cos y}\sqrt{4-(2\cos y)^2}(-2\sin y dy)$$
$$\frac{1}{\pi}\int_0^{\pi} \frac{2}{z-2\cos y}\sin^2(y)dy$$

Let $\psi = e^{iy}$, then

$$\begin{aligned} S_{sc}(z) &= \frac{1}{\pi}\int_0^{2\pi} \frac{1}{z - 2\left(z\frac{e^{iy}+e^{-iy}}{2}\right)} \left(\frac{e^{iy}-e^{-iy}}{2i}\right)^2 dy \\ &= -\frac{1}{4\pi i}\oint_{|\psi|=1} \frac{(\psi^2-1)^2}{\psi^2(1+z\psi-\psi^2)} d\psi \end{aligned}$$

Thus by the Residue theorem,

$$\begin{aligned} S_{sc}(z) &= -\frac{1}{4\pi i}\oint_{|\psi|=1} \frac{(\psi^2-1)^2}{\psi^2(1+z\psi-\psi^2)} d\psi \\ &= \frac{1}{4\pi i}\left(2\pi i\left(z-\sqrt{z^2-4}\right)\right) \\ &= \frac{1}{2}\left(z-\sqrt{z^2-4}\right) \end{aligned}$$

□

Lemma 3.12. *If X is a centered Gaussian random variable, then for any function $f \in C^1(\mathbb{R})$, with polynomial growth of f and f',*

$$\mathrm{E}(Xf(X)) = \mathrm{E}(f'(X))\,\mathrm{E}(X^2)$$

Proof. Without loss of generality, suppose $\mathrm{E}\,X^2 = 1$. Easily by using the integration by parts, one may write,

$$\mathrm{E}(f'(X)) = \int_{-\infty}^{\infty} f'(x)e^{-x^2/2}dx = \int_{-\infty}^{\infty} xf(x)e^{-x^2/2}dx = \mathrm{E}(Xf(X))$$

□

Theorem 3.13 (**Wigner's semicircle law - Weak version**). *Let $W \sim GOE(N)$. Let $S_N(z)$ be the Stieltjes transform of the empirical spectral measure and $S_{sc}(z)$ be the Stieltjes transform of the semicircle law. Then $S_N(z)$ converges to $S_{sc}(z)$ in probability.*

Proof. Define the matrix Δ^{ij} as the symmetric $N \times N$ matrix satisfying

$$\Delta^{ij}_{kl} = \begin{cases} 1 & (i,j) \in \{(k,l),(l,k)\} \\ 0 & \text{otherwise} \end{cases}$$

According to the Woodbury matrix ideneity, $R_W(z) = z^{-1}(WR_W(z) + I)$ for all $z \in \mathbb{C} \setminus \mathbb{R}$, then

$$\frac{\partial}{\partial X_{ij}} R_W(z) = R_W(z)\Delta^{ij}R_W(z)$$

One concludes that

$$\mathrm{E}\,S_N(z) = \frac{1}{N}\mathrm{E}\,\mathrm{tr}(R_W(z)) = \frac{1}{z} + \frac{1}{zN}\mathrm{E}\,\mathrm{tr}(WR_W(z))$$

Since $W \sim \mathrm{GOE}(N)$, we have

$$\mathrm{E}\,W_{ij}^2 = \frac{1+\delta_{ij}}{N}$$

Then,

$$\begin{aligned} \mathrm{E}\,S_N(z) = \frac{1}{N}\mathrm{E}\,\mathrm{tr}(R_W(z)) &= \frac{1}{z} + \frac{1}{zN}\mathrm{E}\,\mathrm{tr}(WR_W(z)) \\ &= \frac{1}{z} + \frac{1}{zN^2}\mathrm{E}\left(\sum_{i\neq j} R_W(z)_{ii}R_W(z)_{jj} + R_W(z)_{ij}^2\right) + \frac{2}{zN^2}\sum_i \mathrm{E}\left(R_W(z)_{ii}^2\right) \\ &= \frac{1}{z} + \frac{1}{z}\left(\frac{1}{N}\mathrm{E}\left(\sum_i R_W(z)_{ii}\right)\right)^2 + O\left(\frac{1}{N}\right) \end{aligned}$$

where the last term is due to the boundedness of $\frac{1}{(z-x)^2}$ for a fixed $z \in \mathbb{C} \setminus \mathbb{R}$. Thus, one may write,

$$\mathrm{E}\, S_N(z) = \frac{1}{z} + \frac{1}{z}(\mathrm{E}\, S_N(z))^2 + O\left(\frac{1}{N}\right)$$

Moreover, we observed that $|S_N(z)|$ is $\frac{1}{|\mathrm{Im}(z)|}$-Lipschitz and W_{ij} has variance $\frac{1}{N}$. Then by Gaussian Lipschitz concentration,

$$\mathrm{P}(|S_N(z) - \mathrm{E}\, S_N(z)| \geq \epsilon) \leq 2\exp(-cN\epsilon^2(\mathrm{Im}(z))^2)$$

Or in other words,

$$S_N(z) = \mathrm{E}\, S_N(z) + o(1)$$

Therefore,

$$S_N(z) = \frac{1}{z} + \frac{1}{z}S_N(z)^2 + o(1)$$

For any limit point $s(z)$ of $S_N(z)$, the following holds,

$$s(z)(z - s(z)) - 1 = 0$$

Hence,

$$s(z) = \frac{1}{2}\left(z - \sqrt{z^2 - 4}\right) = S_{sc}(z)$$

which is the unique Stieltjes transform of the semicircle law. □

Now we aim to generalize the previous theorem.

Theorem 3.14 (**Wigner's semicircle law - Strong version**). *Let W be a Gaussian Wigner matrix. Let $S_N(z)$ be the Stieltjes transform of the empirical spectral measure and $S_{sc}(z)$ be the Stieltjes transform of the semicircle law. Then $S_N(z)$ converges to $S_{sc}(z)$ in probability.*

Proof. Step 1: Concentration. We observed that $|S_N(z)|$ is $1/|\mathcal{I}z|$-Lipschitz and W_{ij} has variance $1/N$. Then by Gaussian Lipschitz concentration,

$$\mathrm{P}\left(|S_N(z) - \mathrm{E}S_N(z)| \geq \epsilon\right) \leq 2\exp\left(-cN\epsilon^2(\mathcal{I}z)^2\right)$$

In other words,

$$S_N(z) = \mathrm{E}\, S_N(z) + o(1)$$

Step 2: Schur Complement Formula. Recall Schur complement formula (see Appendix D, Theorem 1),

$$S_N(z) = \frac{1}{N}\operatorname{tr}\left(R_W(z)\right) = \frac{1}{N}\sum_{i=1}^{N}\left(z - W_{ii} - \langle w_i, R_{W^{(i)}}(z)w_i\rangle\right)^{-1}$$

Step 3: Concentration with self-consistency. Note that w_i is independent of $W^{(i)}$, so we have that

$$\mathrm{E}\left(\langle w_i, R_{W^{(i)}}w_i\rangle \mid W^{(i)}\right) = \frac{1}{N}\operatorname{tr}\left(R_{W^{(i)}}\right)$$

Let $Z_i = \sqrt{N}w_i$. Since $Z_i \sim N(0, I_N)$, Using the Hanson-Wright inequality,

$$\mathrm{P}\left(\left|\frac{1}{N}\langle Z_i, R_{W^{(i)}}Z_i\rangle - \frac{1}{N}\operatorname{tr}\left(R_{W^{(i)}}\right)\right| \geq \frac{t}{N}\right) \leq 2\exp\left(-c\min\left(\frac{t^2}{\|R_{W^{(i)}}\|_F^2}, \frac{t}{\|R_{W^{(i)}}\|_{op}}\right)\right)$$

One can verify that (see Appendix A, Theorem 2 and Corollary 6),

$$\|R_{W^{(i)}}\|_{op} \leq \frac{1}{|\mathcal{I}z|}, \quad \text{and} \quad \|R_{W^{(i)}}\|_F \leq \frac{\sqrt{N}}{|\mathcal{I}z|}$$

By plugging these, one may verify that

$$\frac{1}{N}\left\langle Z_i, R_{W^{(i)}} Z_i\right\rangle = \frac{1}{N}\operatorname{tr}\left(R_{W^{(i)}}\right) + o(1)$$

Now, we use the Cauchy interlacing property (see Appendix D, Theorem 2),

$$\lambda_1(W) \geq \lambda_1\left(W^{(i)}\right) \geq \lambda_2(W) \geq \cdots \geq \lambda_{N-1}(W) \geq \lambda_{N-1}\left(W^{(i)}\right) \geq \lambda_N(W)$$

Then for all $z \in \mathbb{C} \setminus \mathbb{R}$,

$$\begin{aligned}
\frac{1}{N}\operatorname{tr}(R_W(z)) - \frac{1}{N}\operatorname{tr}(R_{W^{(i)}}) &= \frac{1}{N}\sum_{j=1}^{N-1}\left(\frac{1}{z-\lambda_j(W)} - \frac{1}{z-\lambda_j(W^{(i)})}\right) + \frac{1}{N}\frac{1}{z-\lambda_N(W)} \\
&= \frac{1}{N}\sum_{j=1}^{N-1}\frac{\lambda_j(W)-\lambda_j(W^{(i)})}{(z-\lambda_j(W))(z-\lambda_j(W^{(i)})} + \frac{1}{N}\frac{1}{z-\lambda_N(W)} \\
&\leq \frac{1}{N}\frac{1}{|\operatorname{Im}(z)|} + \frac{1}{N\cdot\operatorname{Im}(z)}\sum_{j=1}^{N-1}(\lambda_j(W)-\lambda_j(W^{(i)})) \\
&= \frac{1}{N}\frac{1}{|\operatorname{Im}(z)|}(1+\operatorname{tr}(W)-\operatorname{tr}(W^{(i)})-\lambda_N(W)) \\
&= \frac{1}{N}\frac{1}{|\operatorname{Im}(z)|}(1+W_{ii}-\lambda_N(W))
\end{aligned}$$

One may conclude that,

$$\frac{1}{N}\operatorname{tr}\left(R_{W^{(i)}}\right) = \frac{1}{N}\operatorname{tr}\left(R_W\right) + O\left(\frac{1}{N}\right)$$

Combining these results,

$$\begin{aligned}
\mathrm{E}\, S_N(z) &= \frac{1}{N}\sum_{i=1}^{N}\mathrm{E}(z - W_{ii} - \langle w_i, R_{W^{(i)}} w_i\rangle)^{-1} \\
&= \frac{1}{N}\sum_{i=1}^{N}\mathrm{E}(z - \langle w_i, R_{W^{(i)}} w_i\rangle + o(1))^{-1} \\
&= \mathrm{E}(z - \langle w_N, R_{W^{(N)}} w_N\rangle + o(1))^{-1} \\
&= (z - \mathrm{E}\, S_N(z) + o(1))^{-1} + o(1) \\
&= \frac{1}{z - \mathrm{E}\, S_N(z)} + o(1)
\end{aligned}$$

where in the second line, we used $W_{ii} = o(1)$. In the third line, we used the independence of each term of $\langle w_i, R_{W^{(i)}} w_i\rangle$. In the fourth line, we used the Gaussian Lipschitz concentration of $S_N(z)$ and the result of applying the Hanson-Wright inequality and the Cauchy interlacing property.

For any limit point $s(z)$ of $S_N(z)$, the following holds,

$$s(z)(z - s(z)) - 1 = 0$$

Hence,

$$s(z) = \frac{1}{2}\left(z - \sqrt{z^2-4}\right) = S_{sc}(z)$$

which is the unique Stieltjes transform of the semicircle law. □

3.4 The Marchenko-Pastur Law

The Marchenko-Pastur law, an essential law in random matrix theory, characterizes the spectral distribution of large random matrices (Wishart matrices) such as those arising in statistical signal processing and machine learning. Named after Gavril Marchenko and Leonid Pastur, who introduced it in the 1960s, this theorem describes the limiting behavior of eigenvalues when the matrix dimensions grow large compared to the number of its non-zero entries. With applications spanning fields like finance, telecommunications, and data analysis, the Marchenko-Pastur law offers valuable insights into the statistical properties of high-dimensional data and the behavior of complex systems.

Definition 3.15 (**The Marchenko-Pastur law**)**.** The Marchenko-Pastur law as the probability distribution $\mu_{MP}(x)dx$ on $\mathbb{R}$ can be defined with density

$$\mu_{MP} = \frac{\alpha}{2\pi x}\sqrt{(\lambda_+ - x)(x - \lambda_-)}\mathbf{1}_{x\in[\lambda_-,\lambda_+]}$$

where $\lambda_\pm = \alpha(1 \pm \alpha^{-1/2})^2$ for $\alpha \geq 1$.

Theorem 3.16 (**The Marchenko-Pastur law**)**.** *Let $(Y_l)_{l=1}^M$ be i.i.d. $N(0, I_N)$ with $M = \alpha N$, and let*

$$W_N = \frac{1}{M}\sum_{l=1}^{M} Y_l Y_l^T$$

be an $N \times N$ Wishart matrix. Then, for $\alpha > 1$, μ_N converges weakly almost surely to the Marchenko-Pastur law, μ_{MP}.

Now we will prove this theorem using the resolvent formalism. To do so, first we need to calculate the Stieltjes transform of the Marchenko-Pastur law.

Lemma 3.17 (**Stieltjes transform of the MP law**)**.** *The Stieltjes transform of the Marchenko-Pastur law μ_α is given by the following unique solution to the self-consistent equation,*

$$S_{MP}(z)\left(z - 1 + \frac{1}{\alpha} - \frac{z}{\alpha}S_{MP}(z)\right) = 1$$

Proof. See Lemma 3.11 [5] □

Theorem 3.18 (**The Marchenko-Pastur law**)**.** *Let $(Y_l)_{l=1}^M$ be i.i.d. $N(0, I_N)$ with $M = \alpha N$, and let*

$$W_N = \frac{1}{M}\sum_{l=1}^{M} Y_l Y_l^T$$

be an $N \times N$ Wishart matrix. Then, for $\alpha > 1$, μ_N converges weakly almost surely to the Marchenko-Pastur law, μ_{MP}. In other words, we say $S_N(z)$ converges to S_{MP}.

Proof. Step 1: Concentration. Since each $Y_l = (Y_{l_1}, \cdots, Y_{l_N})$ is a centered Gaussian random vectors thus each Y_{l_j} is sub-Gaussian. Moreover, the multiplication of any two sub-Gaussians, namely Y_{l_i} and Y_{l_j} is sub-exponential. Using the Cauchy-Schwarz inequality, we can show that the components of W_N a sub-exponential random variables.

Now using the Gaussian convex concentration (each element is sub-exponential) and Talagrand's inequality (see Chapter 2), since $f(z) = \frac{1}{z-x}$ is convex and Lipschitz, similar to the proof of the semicircle law, we concludes that

$$S_N(z) = \mathrm{E}\, S_N(z) + o(1)$$

Step 2: Shernman-Morisson identity. One may observe that,

$$R_{W_N}(z)W_N = \frac{1}{M}\sum_{l=1}^{M} R_{W_N}(z)Y_l Y_l^T$$

Define $W_N^l = W_N - \frac{1}{M} Y_l Y_l^T$, then

$$R_{W_N}(z) = (z - W_N)^{-1} = \left(z - W_N^l - \frac{1}{M} Y_l Y_l^T \right)^{-1}$$

Using the Shernman-Morisson formula (see Appendix E, Theorem 2),

$$\langle Y_l, R_{W_N}(z) Y_l \rangle = \frac{\left\langle Y_l, R_{W_N^l}(z) Y_l \right\rangle}{1 - \frac{1}{M} \left\langle Y_l, R_{W_N^l}(z) Y_l \right\rangle}$$

Step 3: Concentration with self-consistence. Using the Hanson-Wright inequality,

$$\mathrm{P}\left(\left| \left\langle Y_l, R_{W_N^l} Y_l \right\rangle - \mathrm{E} \left\langle Y_l, R_{W_N^l} Y_l \right\rangle \right| \geq t \right) \leq 2 \exp\left(-c \min\left(\frac{N^2 t^2}{\left\| R_{W_N^l} \right\|_F^2}, \frac{Nt}{\left\| R_{W_N^l} \right\|_{op}} \right) \right)$$

Similar to the Wigner's semicircle case, one can verify that

$$\left\| R_{W_N^l} \right\|_{op} \leq \frac{1}{|\mathcal{I}z|}, \quad \text{and} \quad \left\| R_{W_N^l} \right\|_F \leq \frac{\sqrt{N}}{|\mathcal{I}z|}$$

Having the above inequality and using the rotation invariance of W_N^l, we have the following concentration,

$$\left\langle Y_l, R_{W_N^l} Y_l \right\rangle = \frac{1}{M} \operatorname{tr} R_{W_N^l} + o(1)$$

Moreover, one may write,

$$\begin{aligned}
\frac{1}{M} \operatorname{tr}(R_{W_N}) - \frac{1}{M} \operatorname{tr}(R_{W_N^l}) &= \frac{1}{M} \sum_{j=1}^{M} \frac{\lambda_j(W_N) - \lambda_j(W_N^l)}{(z - \lambda_j(W_N))(z - \lambda_j(W_N^l))} \\
&\leq \frac{1}{M} \frac{1}{|\mathrm{Im}(z)|} \sum_{j=1}^{M} (\lambda_j(W_N) - \lambda_j(W_N^l)) \\
&= \frac{1}{M} \frac{1}{|\mathrm{Im}(z)|} \operatorname{tr}(W_N - W_N^l) \\
&= \frac{1}{M} \frac{1}{|\mathrm{Im}(z)|} \operatorname{tr}\left(\frac{1}{M} Y_l Y_l^T \right) \\
&= \frac{1}{\alpha^2 N^2} \frac{1}{|\mathrm{Im}(z)|} \sum_{i=1}^{N} Y_{l_i}^2 = O\left(\frac{1}{N} \right)
\end{aligned}$$

Therefore,

$$\frac{1}{M} \mathrm{E} \operatorname{tr} R_{W_N^l} = \frac{1}{M} \mathrm{E} \operatorname{tr} R_{W_N} + O\left(\frac{1}{N} \right)$$

Combining these results,

$$\frac{1}{M} \sum_{l=1}^{M} R_{W_N} Y_l Y_l^T = \frac{1}{1 - \frac{1}{\alpha} \mathrm{E}\, S_N(z) + o(1)} \frac{1}{M} \sum_{l=1}^{M} R_{W_N^l} Y_l Y_l^T$$

Taking expectation, one may write,

$$\begin{aligned}
\mathrm{E}\, R_{W_N} W_N &= \mathrm{E}\left[\frac{1}{M}\sum_{l=1}^{M} R_{W_N} Y_l Y_l^T\right] \\
&= \frac{1}{1-\frac{1}{\alpha}\mathrm{E}\, S_N(z)}\frac{1}{M}\sum_{l=1}^{M}\mathrm{E}\, R_{W_N^l} Y_l Y_l^T + o(1) \\
&= \frac{1}{1-\frac{1}{\alpha}\mathrm{E}\, S_N(z)}\frac{1}{M}\sum_{l=1}^{M}\mathrm{E}\, R_{W_N^l}\cdot \mathrm{E}\, Y_l Y_l^T + o(1) \\
&= \frac{1}{1-\frac{1}{\alpha}\mathrm{E}\, S_N(z)}\mathrm{E}\, R_{W_N^l} + o(1)
\end{aligned}$$

Also, we have that $-1 + zR_{W_N}(z) = R_{W_N}(z)W_N$. If we take an expectation and $\frac{1}{N}\,\mathrm{tr}$ we get,

$$1 - \frac{1}{N}\mathrm{E}\,\mathrm{tr}(zR_{W_N}) = \frac{1}{N}\mathrm{E}\,\mathrm{tr}(R_{W_N}W_N)$$

So, using the above concentration, for any limit point $s(z)$ of $S_N(z)$, the following holds,

$$-1 + zs(z) = \frac{1}{1-\frac{1}{\alpha}s(z)}s(z)$$

In other words,

$$s(z)\left(z - 1 + \frac{1}{\alpha} - \frac{z}{\alpha}s(z)\right) = 1$$

where the unique solution to the above self-consistence equation is the Marchenko-Pastur law. □

3.5 Method of Moments

3.5.1 The Semicircle Law

In this section, we give a direct combinatorics-based proof, mimicking the original argument of Wigner. Before doing so, however, we need to discuss some properties of the semicircle distribution.

Definition 3.19 (**Moment of the Semicircle Law**)**.** The moments of the semicircle law, m_k is defined as follows,

$$m_k := \int x^k d\mu_{sc}$$

Lemma 3.20 (**Moments and Catalan number**)**.** *For all $k \in Z$, the moments of the semicircle law is given by*

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

where $C_k = \frac{(2k)!}{(k+1)!k!}$ is the Catalan number.

Proof. According to the symmetry, one may write,

$$m_{2k+1} = \int_{-2}^{2} x^{2k}\mu_{sc}(x)dx = \frac{2\cdot 2^{2k}}{\pi}\int_{-\pi/2}^{\pi/2} \underbrace{\sin(\theta)}_{\text{odd function}}\ \underbrace{\sin^{2k}(\theta)\cos^2(\theta)}_{\text{even function}}\, d\theta = 0$$

Moreover,

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

Hence,

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

Since $m_0 = 1$, then one may deduct,

$$m_k = C_k = \frac{(2k)!}{(k+1)!k!}$$

□

In the following, we illustrate that the Catalan number counts the number of *Dyck paths* of length $2k$, that is, the number of nonnegative Bernoulli walks of length $2k$ that terminate at 0.

Lemma 3.21 (**Bernoulli walks and Catalan number**). *Let β_k be the number of Dyck paths. Then $\beta = C_k \leq 4^k$. Furthermore, for $|z| < 1/4$, the generating function $\hat{\beta}(z)$ is given by*

$$\hat{\beta}(z) = 1 + \sum_{k=1}^{\infty} z^k \beta_k = \frac{1 - \sqrt{1-4z}}{2z}$$

Proof. Let B_k be the number of Bernoulli walks $\{S_n\}$ of length $2k$ that satisfy $S_{2k} = 0$, and let $\bar{B}_k$ denote the number of Bernoulli walks $\{S_n\}$ of length $2k$ that satisfy $S_{2k} = 0$ and $S_t < 0$ for some $t < 2k$. Then,

$$\beta_k = B_k - \bar{B}_k$$

By reflection at the first hitting of -1, we know $\bar{B}_k$ equals the number of Bernoulli walks $\{S_n\}$ of length $2k$ that satisfy $S_{2k} = -2$. Hence,

$$\beta_k = B_k - \bar{B}_k = \binom{2k}{k} - \binom{2k}{k-1} = C_k$$

Moreover, considering the first return time to 0 of the Bernoulli walk $\{S_n\}$ gives the relation

$$\beta_k = \sum_{j=1}^{k} \beta_{k-j} \beta_{j-1}, \qquad k \geq 1$$

with the convention that $\beta_0 = 1$.

Since the number of Bernoulli walks of length $2k$ is bounded by 4^k, one has that $\beta_k \leq 4^k$. Then the function $\hat{\beta}(z)$ is well-defined and analytic for $|z| < 1/4$. Hence,

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

Moreover,

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

Using that $\hat{\beta}(0) = 1$ and combining the last two equations, the result will be deduced,

$$\hat{\beta}(z) = 1 + z\hat{\beta}(z)^2 \implies \hat{\beta}(z) = \frac{1 - \sqrt{1-4z}}{2z}$$

□

Definition 3.22. The expected value of the kth moment of the empirical spectral measure is defined by

$$m_k^N = \mathrm{E}\left[\int x^k d\mu_N(x)\right]$$

Lemma 3.23 (**Convergence of the moments of the ESM and the semicircle law**). *For every $k \in \mathbb{N}$,*

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

Proof. One may use the following identity,

$$\begin{aligned}
\mathrm{E}\left[\int x^k d\mu_N(x)\right] &= \frac{1}{N}\,\mathrm{E}\,\mathrm{tr}\,X_N^k \\
&= \frac{1}{N}\sum_{i_1,\cdots,i_k} \mathrm{E}\,X_{N_{i_1,i_2}} X_{N_{i_2,i_3}} \cdots X_{N_{i_{k-1},i_k}} X_{N_{i_k,i_1}} \\
&=: \frac{1}{N}\sum_{i_1,\cdots,i_k} \mathrm{E}\,T_{\mathbf{i}}^N
\end{aligned}$$

where we use the notation $\mathbf{i} = (i_1, \cdots, i_k)$.

Recall appnedix F, note that any k-tuple of integers $\mathbf{i}$ defines a closed word $w_{\mathbf{i}} = i_1 i_2 \cdots i_k i_1$ of length $k+1$. Then $\mathcal{W}(t_{\mathbf{i}}) = \mathcal{W}(w_{\mathbf{i}})$ which is the number of distinct integers in $\mathbf{i}$. Thus, one may write,

$$\mathrm{E}\,T_{\mathbf{i}}^N = \frac{1}{N^{k/2}} \prod_{e\in E_{w_{\mathbf{i}}}^c} \mathrm{E}(Z_{1,2}^{N_e^{w_{\mathbf{i}}}}) \prod_{e\in E_{w_{\mathbf{i}}}^s} \mathrm{E}(Y_1^{N_e^{w_{\mathbf{i}}}})$$

For all $e \in E_{w_{\mathbf{i}}}$, we know that $N_e^{w_{\mathbf{i}}} \geq 2$ and $\mathrm{E}\,T_{\mathbf{i}}^N \neq 0$. It yields $\mathcal{W}(t_{\mathbf{i}}) \leq k/2+1$. Also, the above equation shows that if $w_{\mathbf{i}} \sim w_{\mathbf{i}'}$ then $T_{w_{\mathbf{i}}} = T_{w_{\mathbf{i}'}}$. Furthermore, if $N \geq t$, there exists

$$C_{N,t} := N(N-1)(N-2)\cdots(N-t+1)$$

N-words such that are equivalent to a given N-words of weight t.

Moreover, one may define $\mathscr{W}_{k,t}$ denotes a set of representatives for equivalence classes of closed t-words w of length $k+1$ and weight t with $N_e^w \geq 2$ for each $e \in E_w$.

Using the two above equations, one may show

$$\mathrm{E}\left[\int x^k d\mu_N(x)\right] = \sum_{t=1}^{\lfloor k/2\rfloor+1} \frac{C_{N,t}}{N^{k/2+1}} \sum_{w\in\mathscr{W}_{k,t}} \prod_{e\in E_w^c} \mathrm{E}(Z_{1,2}^{N_e^w}) \prod_{e\in E_w^s} \mathrm{E}(Y_1^{N_e^w})$$

As we know, the cardinality of $\mathscr{W}_{k,t}$ is bounded by the number of closed $\mathscr{S}$-words of length $k+1$ when the cardinality of $\mathscr{S}$ is $t \leq k$, then $|\mathscr{W}_{k,t}| \leq t^k \leq k^k$. Theretofore, if k is odd,

$$\lim_{N\to\infty} \mathrm{E}\left[\int x^k d\mu_N(x)\right] = 0$$

while, for k even,

$$\lim_{N\to\infty} \mathrm{E}\left[\int x^k d\mu_N(x)\right] = \sum_{w\in\mathscr{W}_{k,k/2+1}} \prod_{e\in E_w^c} \mathrm{E}(Z_{1,2}^{N_e^w}) \prod_{e\in E_w^s} \mathrm{E}(Y_1^{N_e^w})$$

Using the definition of Wigner word (see appnedix F), it implies that E_w^s is empty for $w \in \mathscr{W}_{k,k/2+1}$, and thus, for k even,

$$\lim_{N\to\infty} \mathrm{E}\left[\int x^k d\mu_N(x)\right] = |\mathscr{W}_{k,k/2+1}|$$

Now we seek to show that $\mathscr{W}_{k,k/2+1} = C_{k/2}$. To start, let k be even. It is convenient to choose the set of representatives $\mathscr{W}_{k,k/2+1}$ such that each word $w = v_1 \cdots v_{k+1}$ in that set satisfies, for $i = 1, \ldots, k+1$, the condition that $\{v_1, \ldots, v_i\}$ is an interval in $\mathbb{Z}$ beginning at 1. Each element $w \in \mathscr{W}_{k,k/2+1}$ determines a path $v_1, v_2, \ldots, v_k, v_{k+1} = v_1$ of length k on the tree G_w. We refer to this path as the exploration process associated with w. Let $d(v, v')$ denote the distance between vertices v, v' on the tree G_w, i.e. the length of the shortest path on the tree beginning at v and terminating at v'. Setting $x_i = d(v_{i+1}, v_1)$, one sees that each word $w \in \mathscr{W}_{k,k/2+1}$ defines a Dyck path $D(w) = (x_1, x_2, \ldots, x_k)$ of length k. Conversely, given a

Dyck path $\mathbf{x} = (x_1, \ldots, x_k)$, one may construct a word $w = T(\mathbf{x}) \in \mathscr{W}_{k,k/2+1}$ by recursively constructing an increasing sequence $w_2, \ldots, w_k = w$ of words, as follows. For $i > 2$, if $x_{i-1} = x_{i-2} + 1$, then w_i is obtained by adjoining on the right of w_{i-1} the smallest positive integer not appearing in w_{i-1}. Otherwise, w_i is obtained by adjoining on the right of w_{i-1} the next-to-last letter of w_{i-1}. Note that for all i, G_{w_i} is a tree (because G_{w_2} is a tree and, inductively, at stage i, either a backtrack is added to the exploration process on $G_{w_{i-1}}$ or a leaf is added to $G_{w_{i-1}}$). Furthermore, the distance in G_{w_i} between first and last letters of w_i equals x_{i-1}, and therefore, $D(w) = (x_1, \ldots, x_k)$. With our choice of representatives, $T(D(w)) = w$, because each uptick in the Dyck path $D(w)$ starting at location $i-2$ corresponds to adjoinment on the right of w_{i-1} of a new letter, which is uniquely determined by supp w_{i-1}, whereas each downtick at location $i-2$ corresponds to the adjoinment of the next-to-last letter in w_{i-1}. This establishes a bijection between Dyck paths of length k and $\mathscr{W}_{k,k/2+1}$. Finally, Lemma 3.21 establishes that

$$\left|\mathscr{W}_{k,k/2+1}\right| = C_{k/2}.$$

□

Lemma 3.24 (**Concentration of the moments of the ESM**). *Any kth moment of the empirical spectral measure converges in probability to m_k^N. In other words, for every $k \in \mathbb{N}$ and $\epsilon > 0$,*

$$\lim_{N\to\infty} P\left(\left|\int x^k d\mu_N(x) - \mathrm{E}\left[\int x^k d\mu_N\right]\right| > \epsilon\right) = 0$$

Proof. One may write,

$$\mathrm{E}\left[\int x^k d\mu_N\right]^2 - \left(\mathrm{E}\left[\int x^k d\mu_N\right]\right)^2 = \frac{1}{N^2} \sum_{i_1, i_1' \cdots, i_k, i_k'} \mathrm{E}\, T^N_{\mathbf{i},\mathbf{i}'}$$

where

$$T^N_{\mathbf{i},\mathbf{i}'} = \mathrm{E}\, T^N_{\mathbf{i}} T^N_{\mathbf{i}'} - \mathrm{E}\, T^N_{\mathbf{i}} \,\mathrm{E}\, T^N_{\mathbf{i}'}$$

Recall the closed words $w_{\mathbf{i}}, w_{\mathbf{i}'}$ of length $k+1$, and define the two-word sentence $a_{\mathbf{i},\mathbf{i}'} = (w_{\mathbf{i}}, w_{\mathbf{i}'})$ (see appnedix F). Then in similar to what we did in previous lemma,

$$\begin{aligned}
\bar{T}^N_{\mathbf{i},\mathbf{i}'} = &\frac{1}{N^k}\left(\prod_{e\in E^c_{a_{i,i}}} E\left(Z_{1,2}^{N_e^a}\right) \prod_{e\in E^{\mathbf{s}}_{a_{i,i}}} E\left(Y_1^{N_e^a}\right)\right) \\
&- \frac{1}{N^k}\left(\prod_{e\in E^c_{w_{\mathbf{i}}}} E\left(Z_{1,2}^{N_e^{w_{\mathbf{i}}}}\right) \prod_{e\in E^{\mathbf{s}}_{w_{\mathbf{i}}}} E\left(Y_1^{N_e^{w_{\mathbf{i}}}}\right) \prod_{e\in E^c_{w_{\mathbf{i}'}}} E\left(Z_{1,2}^{N_e^{w_i}}\right) \prod_{e\in E^{\mathbf{s}}_{\mathbf{i}'}} E\left(Y_1^{N_e^{w_i}}\right)\right)
\end{aligned}$$

For all $e \in E_{a_{\mathbf{i},\mathbf{i}'}}$, we know that $N_e^{a_{\mathbf{i},\mathbf{i}'}} \geq 2$ and $\mathrm{E}\, T^N_{\mathbf{i},\mathbf{i}'} \neq 0$. Also if $E_{w_{\mathbf{i}}} \cap E_{w_{\mathbf{i}'}} \neq \emptyset$, then $\mathrm{E}\, T^N_{\mathbf{i},\mathbf{i}'} = 0$. Furthermore, the above equation shows that if $a_{\mathbf{i},\mathbf{i}'} \sim a_{\mathbf{j},\mathbf{j}'}$, then $\mathrm{E}\, T^N_{\mathbf{i},\mathbf{i}'} = \mathrm{E}\, T^N_{\mathbf{j},\mathbf{j}'}$. Finally, if $N \geq t$, then there are $C_{N,t}$ N-sentences such that are equivalent to a given N-sentence of weight t.

Moreover, one may define $\mathscr{W}^{(2)}_{k,t}$ denotes a set of representatives for equivalence classes of sentences a of weight t consisting of two closed t-words (w_1, w_2), each of length $k+1$, with $N_e^a \geq 2$ for each $e \in E_a$, and $E_{w_1} \cap E_{w_2} \neq \emptyset$.

Using the above equations, one may show

$$\begin{aligned}
&E\left(\left\langle L_N, x^k\right\rangle^2\right) - \left\langle \bar{L}_N, x^k\right\rangle^2 \\
&\quad = \sum_{t=1}^{2k} \frac{C_{N,t}}{N^{k+2}} \sum_{a=(w_1,w_2)\in\mathscr{W}^{(2)}_{k,t}} \left(\prod_{e\in E^c_a} E\left(Z_{1,2}^{N_e^a}\right) \prod_{e\in E^s_a} E\left(Y_1^{N_e^a}\right)\right. \\
&\qquad \left. - \prod_{e\in E^c_{w_1}} E\left(Z_{1,2}^{N_e^{w_1}}\right) \prod_{e\in E^s_{w_1}} E\left(Y_1^{N_e^{w_1}}\right) \prod_{e\in E^c_{w_2}} E\left(Z_{1,2}^{N_e^{w_2}}\right) \prod_{e\in E^s_{w_2}} E\left(Y_1^{N_e^{w_2}}\right)\right)
\end{aligned}$$

Now it suffices to check whether $\mathscr{W}_{k,t}^{(2)}$ is empty for $t \geq k+2$. However, we prove a stronger claim that $\mathscr{W}_{k,t}^{(2)}$ is empty for $t \geq k+1$. So if $a \in \mathscr{W}_{k,t}^{(2)}$, since $N_e^a \geq 2$ for $e \in E_a$ then G_a is a connected graph, with t vertices and at most k edges which is impossible when $t > k+1$. Now let $t = k+1$. Thus, G_a is a tree and each edge must be visited by the paths generated by a exactly twice. Since the path generated by w_1 in the tree G_a starts and end at the same vertex, it must visit each edge an even number of times. Therefore, the set of edges visited by w_1 is disjoint from the set of edges visited by w_2, contradiction the definition of $\mathscr{W}_{k,t}^{(2)}$. □

Theorem 3.25 (**Wigner's semicircle law**). *Let W be an $N \times N$ Wigner matrix. The empirical spectral measure μ_N converges weakly in probability to the semicircle law, μ_{sc} i.e.,*

$$\lim_{N\to\infty} P\left(\left|\int f d\mu_N - \int f d\mu_{sc}\right| > \epsilon\right) = 0, \qquad \forall f \in C_b(\mathbb{R})$$

Proof. Using Chebyshev's inequality, one may write,

$$\mathrm{P}\left(\int |x|^k \mathbf{1}_{|x|>B} d\mu_N > \epsilon\right) \leq \frac{1}{\epsilon}\,\mathrm{E}\left[\int |x|^k \mathbf{1}_{|x|>B} d\mu_N\right] \leq \frac{1}{\epsilon B^k}\,\mathrm{E}\left[\int x^{2k} d\mu_N\right]$$

According to the convergence of the moments of the ESM and the semicircle law,

$$\limsup_{N\to\infty} \mathrm{P}\left(\int |x|^k \mathbf{1}_{|x|>B} d\mu_N > \epsilon\right) \leq \frac{1}{\epsilon B^k}\int x^{2k} d\mu_{sc} \leq \frac{C_k}{\epsilon B^k} \leq \frac{4^k}{\epsilon B^k}$$

If we assume that f is supported on the interval $[-5,5]$ and set $B = 5$, then we have the following result,

$$\limsup_{N\to\infty} \mathrm{P}\left(\int |x|^k \mathbf{1}_{|x|>B} d\mu_N > \epsilon\right) = 0$$

Now fix f and $\delta > 0$. Using the Weierstrass approximation theorem, one may find a polynomial $Q_\delta(x) = \sum_{i=1}^{L} c_i x^i$ such that

$$\sup_{x:|x|\leq B} |Q_\delta(x) - f(x)| \leq \frac{\delta}{8}$$

Hence,

$$\begin{aligned}
\mathrm{P}\left(\left|\int f d\mu_N - \int f d\mu_{sc}\right| > \delta\right) &\leq \mathrm{P}\left(\left|\int Q_\delta d\mu_N - \mathrm{E}\left[\int Q_\delta d\mu_N\right]\right| > \delta/4\right) \\
&\quad + \mathrm{P}\left(\left|\mathrm{E}\left[\int Q_\delta d\mu_N\right] - \int Q_\delta d\mu_{sc}\right| > \delta/4\right) \\
&\quad + \mathrm{P}\left(\left|\int Q_\delta \mathbf{1}_{|x|>B} d\mu_N\right| > \delta/4\right) \\
&=: P_1 + P_2 + P_3
\end{aligned}$$

Using the concentration, Lemma 3.24, $P_1 \to 0$ as $N \to \infty$ and by an application of Lemma 3.23, $P_2 = 0$ for large N. Finally, using the above result, $P_3 \to 0$ as $N \to \infty$. Therefore,

$$\mathrm{P}\left(\left|\int f d\mu_N - \int f d\mu_{sc}\right| > \delta\right) \overset{N\to\infty}{\longrightarrow} 0$$

□

3.5.2 The Marchenko-Pastur Law

Theorem 3.26 (**The Marchenko-Pastur law**). *Let $(X_l)_{l=1}^M$ be i.i.d. $N(0, I_N)$ with $M = \alpha N$, and let*

$$W_N = Y_N Y_N^T$$

be an $N \times N$ Wishart matrix, where $Y_N = X_N/\sqrt{N}$. Then, for $\alpha > 1$, μ_N converges weakly almost surely to the Marchenko-Pastur law, μ_{MP}.

Proof. From the usual rules of matrix multiplication, we see that

$$\begin{aligned}\mathrm{E}\left[\int x^k d\mu_N\right] &= \frac{1}{N}\,\mathrm{E}\,\mathrm{tr}\,W_N^k = \frac{1}{N}\,\mathrm{E}\,\mathrm{tr}\left(Y_N Y_N^T\right)^k \\ &= \frac{1}{N}\sum_{\substack{i_1,\ldots,i_k \\ j_1,\ldots,j_k}} \mathrm{E}\,Y_N(i_1,j_1)\,Y_N(i_2,j_1)\,Y_N(i_2,j_2)\cdots Y_N(i_k,j_k)\,Y_N(i_1,i_k)\end{aligned}$$

where the row indices $i_1,\ldots,i_k$ take values in $\{1,\ldots,n\}$ and the column indices $j_1,\ldots,j_k$ take values in $\{1,\ldots,M\}$.

Because the entries of Y_n are independent, each factor in the product

$$Y_n(i_1,j_1)\,Y_N(i_2,j_1)\,Y_N(i_2,j_2)\,Y_N(i_3,j_2)\cdots Y_N(i_k,j_k)\,Y_N(i_1,i_k)$$

must appear at least twice for the expectation to be nonzero. We can think of each such product as a connected bipartite graph on the sets of vertices $\{i_1,\ldots,i_k\}$ and $\{j_1,\ldots,j_k\}$, where the total number of edges (with repetitions) is $2k$. Suppose N_i and N_j denote the number of distinct i indices and j indices, respectively. Since each edge needs to be traversed at least twice, there are at most $k+1$ distinct vertices, so $N_i+N_j\le k+1$. In particular, if $N_i+N_j=k+1$ there are k unique edges and the resulting graph is a tree. Such terms will become the dominant ones in the sum in the limit $N\to\infty$.

Indeed, let us show that all the terms with $N_i+N_j\le k$ contribute an amount that is $o(N)$ to the sum. To this end, define the weight vector $\mathbf{t_i}$ corresponding to the vector $\mathbf{i}=(i_1,\ldots,i_k)$, describing which entries of $\mathbf{i}$ are equal. One may associate a weight vector to $\mathbf{j}$. Because the entries of Y_N are identically distributed, it is easy to see that choices of $\mathbf{i}$ and $\mathbf{j}$ which generate the same weight vectors contribute the same amounts to the sum.

For a fixed weight vector $\mathbf{t_i}$ with N_i distinct entries (same N_i which gives the number of distinct row indices i_l), there are $N(N-1)\cdots(N-N_i+1)<N^{N_i}$ choices of $\mathbf{i}$ with weight vector $\mathbf{t_i}$. Similarly, there are $M(M-1)\cdot(M-N_j+1)<M^{N_j}$ choices of $\mathbf{j}$ corresponding to some fixed weight vector $\mathbf{t_j}$. Therefore, there are less than $N^{N_i}\cdot M^{N_j}<CN^{N_i+N_j}\le CN^k$ for $N_i+N_j\le k$, where C is a constant depending on k and α, but not N.

In addition, each term $\frac{1}{N}\,\mathrm{E}\,Y_N(i_1,j_1)\,Y_N(i_2,j_1)\cdots Y_N(i_k,j_k)\,Y_N(i_1,i_k)$ is $O\left(1/N^{k+1}\right)$ because of the scaling $Y_N=X_N/\sqrt{N}$ and the assumption that the moments of each entry $X_N(i,j)$ are finite. Therefore, the sum over all $\mathbf{i}$ and $\mathbf{j}$ corresponding to fixed weight vectors $\mathbf{t_i},\mathbf{t_j}$ is $o(N)$. Furthermore, the number of possible weight vectors $\mathbf{t_i}$ and $\mathbf{t_j}$ depends on k but not N, which means that the contribution of all terms in the assum is asymptotically 0 .

Therefore, we now focus on the terms corresponding to $\mathbf{i}$ and $\mathbf{j}$ with $N_i+N_j=k+1$. This is the case where the product

$$Y_N(i_1,j_1)\,Y_N(i_2,j_1)\,Y_N(i_2,j_2)\,Y_N(i_3,j_2)\cdots Y_N(i_k,j_k)\,Y_N(i_1,i_k)$$

contains exactly two copies of each distinct entry. Correspondingly, each edge in the path $i_1j_1\cdots i_kj_ki_1$ gets traversed twice, once in each direction. Because each entry of Y_N has variance $1/N$, we conclude that

$$\frac{1}{N}\,\mathrm{E}\,Y_N(i_1,j_1)\,Y_N(i_2,j_1)\,Y_N(i_2,j_2)\,Y_N(i_3,j_2)\cdots Y_N(i_k,j_k)\,Y_N(i_1,i_k)=\frac{1}{N^{k+1}}$$

for each choice of $\mathbf{i}$ and $\mathbf{j}$ with $N_i+N_j=k+1$. Again, fix two weight vectors $\mathbf{t_i}$ and $\mathbf{t_j}$ with N_i and N_j distinct entries. There are $N(N-1)\cdots(N-N_i+1)\cdot M(M-1)\cdots(M-N_j+1)$ corresponding choices for $\mathbf{i}$ and $\mathbf{j}$. Because $M=N\alpha$ and we are in the case $N_i+N_j=k+1$, the number of choices is equal to $N^{k+1}\alpha^{N_j}$.

From the last two observations, it follows that

$$\mathrm{E}\left[\int x^k d\mu_N\right]=\sum\alpha^{N_j},$$

where the sum is taken over all pairs $(\mathbf{t_i},\mathbf{t_j})$ with $N_i+N_j=k+1$ and distinct weight vectors. To continue, we proceed as in the moment method proof of the semicircle law. To each closed path $i_1j_1i_2j_2\cdots i_kj_ki_1$ we can associate a type sequence of length $2k$, whose j th term gives the number of free steps minus the number

of repetitive steps within the first j edge traversals. As before, every type sequence corresponding to a path where each edge gets traversed exactly twice starts at 1 , ends at 0 , and has consecutive terms differing by ± 1. Also note that the odd terms in a type sequence correspond to edges ending at a j vertex, whereas even terms correspond to edges terminating at an i vertex.

For a given type sequence, let l be the number of times there is a decrease by 1 going from an odd to an even term. Then $l = N_j$, the number of distinct j indices. Indeed, l counts the number of paths of the form $j_s i_{s+1}$ such that i_{s+1} has been visited once before, which by the condition that each edge is traversed exactly twice gives the number of distinct j s. Furthermore, pairs of weight vectors $(\mathbf{t_i}, \mathbf{t_j})$ correspond bijectively to type sequences. Thus, letting

$$\beta_k = \sum_{\substack{\text{dyck paths} \\ \text{of length } 2k}} \alpha^l,$$

we deduce

$$\mathrm{E}\left[\int x^k d\mu_N\right] = \beta_k.$$

The goal is to establish a recurrence relation between the β_k in order to compute the general term. To do this, associate to each type sequence of even length a second parameter $\bar{l}$ which counts the number of times there is a decrease by 1 going from an even to an odd term. Denote by $\gamma_k = \sum \alpha^{\bar{l}}$, where the sum is taken over type sequences of length $2k$.

Next, consider the (necessarily even) position $2j$ of the first occurrence of a zero in a type sequence of length $2k$. Then the elements beyond this index make up an arbitrary type sequence of length $2k - 2j$, with the first $2j$ terms forming a type sequence of length $2j$ with no zero occurring before the last position. By eliminating the first and last terms and subtracting 1 from each of the remaining elements, we see that such sequences are in bijection with arbitrary type sequences of length $2j-2$. Furthermore, if l counts the number of decreases from odd to even indices in the sequence of length $2j$, then $l - 1$ gives the number of decreases from even to odd indices in this new sequence of length $2j - 2$. Keeping in mind how β_k and γ_k were defined in terms of powers of α, we deduce

$$\beta_k = \alpha \sum_{j=1}^{k} \gamma_{j-1}\beta_{k-j} \qquad \gamma_k = \sum_{j=1}^{k} \beta_{k-j}\gamma_{j-1}.$$

Thus, $\beta_k = \alpha\gamma_k$ for $k \geq 1$, with $\beta_0 = \gamma_0 = 1$ in order for these recurrences to hold. Since we are primarily interested in the β_k, note that these identities imply

$$\beta_k = (\alpha - 1)\beta_{k-1} + \sum_{j=1}^{k} \beta_{k-j}\beta_{j-1}.$$

In particular, if $\hat{\beta}(x) := \sum_{k=0}^{\infty} \beta_k x^k$ is the generating function for the β_k, the previous identity leads to the following equality for $\hat{\beta}$:

$$\hat{\beta}(x) = 1 + x\hat{\beta}(x)^2 + (\alpha - 1)x\hat{\beta}(x).$$

The expected ESD $\bar{\mu}_N$ thus converges to a distribution whose moments are encoded by $\hat{\beta}$. This asymptotic density has a Stieltjes transform $s(z)$ which can be easily computed as $s(z) = -\hat{\beta}(1/z)/z$, a claim which follows directly from the definition of the Stieltjes transform. Upon solving the quadratic equation in $\hat{\beta}$ from earlier, we have:

$$s(z) = \frac{-z + (\alpha - 1) + \sqrt{z^2 - 2z(\alpha + 1) + (\alpha - 1)^2}}{2z}.$$

Upon inversion of the Stieltjes transform, we see that the limiting density is given by

$$f_\alpha(x) = \frac{\sqrt{(x - \lambda_-)(\lambda_+ - x)}}{2\pi x}\mathbf{1}_{x \in [\lambda_-, \lambda_+]},$$

where $\lambda_- = (1 - \sqrt{\alpha})^2$ and $\lambda_+ = (1 + \sqrt{\alpha})^2$, as before. We have thus far shown that $\mathrm{E}\,\mu_N \to f_\alpha$ deterministically. An argument similar to that used to prove lemma 3.25 shows that, in fact, the ESD of an arbitrary Wishart matrix converges to the Marcenko-Pastur distribution, and thus theorem is proven. □

A Singular Value Decomposition Theorem

Theorem A.1 (**Singular Value Decomposition (SVD)**). *Let A be an $m \times n$ matrix with real entries. Then A can be represented as*

$$A = U\Sigma V^T \iff A = \sum_{i=1}^{r} s_i u_i v_i^T$$

where $r = rank(A)$. Here the non-negative $s_i = s_i(A)$ are the singular values of A, the vectors $u_i \in \mathbb{R}^m$ are the left singular vectors of A, and the vector $v_i \in \mathbb{R}^n$ are the right singular vectors of A. Also, the singular values s_i are the square roots of the eigenvalues λ_i of both AA^T and A^TA.

Theorem A.2 (**Courant-Fisher min-max theorem**). *Let A be a symmetric matrix. Assume the eigenvalues $\lambda_i(A)$ are arranged in a non-decreasing order. Then,*

$$\lambda_i(A) = \max_{\dim E = i} \min_{x \in S(E)} \langle Ax, x \rangle \implies s_i(A) = \max_{\dim E = i} \min_{x \in S(E)} \|Ax\|_2$$

where the maximum is over all i-dimensional subspaces E of $\mathbb{R}^n$, the minimum is over all unit vectors $x \in E$, and $S(E)$ denotes the unit Euclidean sphere in the subspace E.

Definition A.3 (**Operator norm**). Let $A : l_2^n \to l_2^m$ be a matrix as a linear operator from l_2^n to l_2^m. Its operator norm, also called the spectral norm, is defined as

$$\|A\| = \max_{x \in \mathbb{R}^n \setminus \{0\}} \frac{\|Ax\|_2}{\|x\|_2} = \max_{x \in S^{n-1}} \|Ax\|_2 \iff \|A\| = \max_{x \in S^{n-1}, y \in S^{m-1}} \langle Ax, y \rangle$$

Corollary A.4 (**Largest singular value**). *Let A be a symmetric matrix. The operator norm of A equals the largest singular value of A,*

$$s_1(A) = \|A\|$$

Definition A.5 (**Frobenius or Hilbert-Schmidt norm**). The Frobenius norm of a matrix A with entries A_{ij} is defined as

$$\|A\|_F = \left(\sum_{i=1}^{m} \sum_{j=1}^{n} |A_{ij}|^2 \right)^{\frac{1}{2}}$$

Corollary A.6 (**Norm in terms of singular values**). *In terms of singular values, the Frobenius norm can be computed as*

$$\|A\|_F = \left(\sum_{i=1}^{r} s_i(A)^2 \right)^{\frac{1}{2}}$$

Now if we compare the norms,

$$\|A\| = \|s\|_\infty = s_1(A), \quad \|A\|_F = \|s\|_2 \implies \|A\| \leq \|A\|_F \leq \sqrt{r}\|A\|$$

Theorem A.7 (**Eckart-Young-Mirsky theorem**). *Let A be a matrix of rank r. Then A_k that has a given lower rank $k < r$, is the minimizer by truncating the singular value decomposition of A at the kth term,*

$$A_k = \sum_{i=1}^{k} s_i u_i v_i^T \iff \|A - A_k\| = \min_{rank(A') \leq k} \|A - A'\|$$

The matrix A_k is often called the best rank-k approximation of A.

Theorem A.8 (**Extreme singular values control the distortion**). *For a given matrix A there exists number M and m such that*

$$m\|x\|_2 \leq \|Ax\|_2 \leq M\|x\|_2, \quad \forall x \in \mathbb{R}^n$$

Then with applying $x - y$ we have,

$$s_n(A)\|x - y\|_2 \leq \|Ax - Ay\|_2 \leq s_1(A)\|x - y\|_2, \quad \forall x \in \mathbb{R}^n$$

B Packing and Covering Numbers

Definition B.1 (**ϵ-Net**). Let (T, d) be a metric space. Consider a subset $K \subset T$ and let $\epsilon > 0$. A subset $\mathcal{N} \subseteq K$ is called an ϵ-net of K if every point in K is within a distance ϵ of some point of $\mathcal{N}$, i.e.,

$$\forall x \in K, \quad \exists x_0 \in \mathcal{N} : \quad d(x, x_0) \leq \epsilon$$

Definition B.2 (**Covering number**). The smallest possible cardinality of an ϵ-net of K is called the covering number of K and is denoted $\mathcal{N}(K, d, \epsilon)$. Equivalently, $\mathcal{N}(K, d, \epsilon)$ is the smallest number of closed balls with centers in K and radii ϵ whose union covers K.

Definition B.3 (**Packing number**). A subset $\mathcal{N}$ of a metric space (T, d, ϵ) is ϵ-separated if $d(x, y) > \epsilon$ for all distinct points $x, y \in \mathcal{N}$. The largest possible cardinality of an ϵ-separated subset of a given set $K \subset T$ is called the packing number of K and is denoted $\mathcal{P}(K, d, \epsilon)$.

Theorem B.4 (**Equivalence of covering and packing numbers**). *For any set $K \subset T$ and any $\epsilon > 0$, we have*

$$\mathcal{P}(K, d, 2\epsilon) \leq \mathcal{N}(K, d, \epsilon) \leq \mathcal{P}(K, d, \epsilon)$$

Theorem B.5 (**Covering numbers and volume**). *Let K be a subset of $\mathbb{R}^n$ and $\epsilon > 0$. Then*

$$\frac{|K|}{|\epsilon B_2^n|} \leq \mathcal{N}(K, \epsilon) \leq \mathcal{P}(K, \epsilon) \leq \frac{|(K + \frac{\epsilon}{2} B_2^n|}{|\frac{\epsilon}{2} B_2^n|}$$

Corollary B.6 (**Covering number of the Euclidean ball**). *The covering numbers of the unit Euclidean ball B_2^n satisfy the following for any $\epsilon > 0$:*

$$\left(\frac{1}{\epsilon}\right)^n \leq \mathcal{N}(B_2^n, \epsilon) \leq \left(\frac{2}{\epsilon} + 1\right)^n$$

Theorem B.7 (**Computing the operator norm on a set**). *Let A be an $m \times n$ matrix and $\epsilon \in [0, 1]$. Then, for any ϵ-net $\mathcal{N}$ of the sphere S^{n-1}, we have*

$$\sup_{x \in \mathcal{N}} \|Ax\|_2 \leq \|A\| \leq \frac{1}{1 - \epsilon} \sup_{x \in \mathcal{N}} \|Ax\|_2$$

Corollary B.8 (**Computing the Euclidean norm**). *Let $x \in \mathbb{R}^n$ and let $\mathcal{N}$ be an ϵ-net of the sphere S^{n-1}. Then,*

$$\sup_{y \in \mathcal{N}} \langle x, y \rangle \leq \|x\|_2 \leq \frac{1}{1 - \epsilon} \sup_{y \in \mathcal{N}} \langle x, y \rangle$$

C Hoffman-Wielandt Theorem

Theorem C.1 (**Hoffman-Wielandt**). *Let A and B be $N \times N$ symmetric matrices, with eigenvalues $\lambda_1^A \leq \lambda_2^A \leq \cdots \leq \lambda_N^A$ and $\lambda_1^B \leq \lambda_2^B \leq \cdots \leq \lambda_N^B$. Then,*

$$\sum_{i=1}^{N} |\lambda_i^A - \lambda_i^B|^2 \leq tr(A - B)^2$$

Proof. Note that $\operatorname{tr} A^2 = \sum_i \left(\lambda_i^A\right)^2$ and $\operatorname{tr} B^2 = \sum_i \left(\lambda_i^B\right)^2$. Let U denote the matrix diagonalizing B written in the basis determined by A, and let D_A, D_B denote the diagonal matrices with diagonal elements λ_i^A, λ_i^B respectively. Then,

$$\operatorname{tr} AB = \operatorname{tr} D_A U D_B U^{\mathrm{T}} = \sum_{i,j} \lambda_i^A \lambda_j^B u_{ij}^2.$$

The last sum is linear in the coefficients $v_{ij} = u_{ij}^2$, and the orthogonality of U implies that

$$\sum_j v_{ij} = 1, \quad \sum_i v_{ij} = 1$$

Thus

$$\operatorname{tr} AB \leq \sup_{v_{ij}\geq 0:\Sigma_j v_{ij}=1,\Sigma_i v_{ij}=1} \sum_{i,j} \lambda_i^A \lambda_j^B v_{ij}.$$

But this is a maximization of a linear functional over the convex set of doubly stochastic matrices, and the maximum is obtained at the extreme points, which are well known to correspond to permutations The maximum among permutations is then easily checked to be $\sum_i \lambda_i^A \lambda_i^B$. Collecting these facts together implies Theorem C.1. Alternatively, one sees directly that a maximizing $V = \{v_{ij}\}$ in the above inequality is the identity matrix. Indeed, assume Without loss of generality that $v_{11} < 1$. We then construct a matrix $\bar{V} = \{\bar{v}_{ij}\}$ with $\bar{v}_{11} = 1$ and $\bar{v}_{ii} = v_{ii}$ for $i > 1$ such that $\bar{V}$ is also a maximizing matrix. Indeed, because $v_{11} < 1$, there exist a j and a k with $v_{1j} > 0$ and $v_{k1} > 0$. Set $v = \min(v_{1j}, v_{k1}) > 0$ and define $\bar{v}_{11} = v_{11} + v, \bar{v}_{kj} = v_{kj} + v$ and $\bar{v}_{1j} = v_{1j} - v, \bar{v}_{k1} = v_{k1} - v$, and $\bar{v}_{ab} = v_{ab}$ for all other pairs ab. Then,

$$\begin{aligned}\sum_{i,j} \lambda_i^A \lambda_j^B (\bar{v}_{ij} - v_{ij}) &= v\left(\lambda_1^A \lambda_1^B + \lambda_k^A \lambda_j^B - \lambda_k^A \lambda_1^B - \lambda_1^A \lambda_j^B\right) \\ &= v\left(\lambda_1^A - \lambda_k^A\right)\left(\lambda_1^B - \lambda_j^B\right) \geq 0.\end{aligned}$$

Thus, $\bar{V} = \{\bar{v}_{ij}\}$ satisfies the constraints, is also a maximum, and the number of zero elements in the first row and column of $\bar{V}$ is larger by 1 at least from the corresponding one for V. If $\bar{v}_{11} = 1$, the claims follows, while if $\bar{v}_{11} < 1$, one repeats this (at most $2N - 2$ times) to conclude. Proceeding in this manner with all diagonal elements of V, one sees that indeed the maximum of the right side of the first inequality is $\sum_i \lambda_i^A \lambda_i^B$, as claimed. □

D Schur Complement Formula

Theorem D.1 (**Schur complement formula**). *Let W be $N \times N$ Hermitian. Let w_i denote the ith column of W with the entry W_{ii} removed (i.e., w_i os an $N-1$-dimensional vector). Let $W^{(i)}$ denote the matrix obtained by erasing the ith column and row from W. Then, for every $z \in \mathbb{C} \setminus R$,*

$$(zI - W)_{ii}^{-1} = (z - W_{ii} - \langle w_i, R_{W^{(i)}}(z) w_i \rangle)^{-1}$$

Proof. From Cramer's rule,

$$(zI_N - W)_{ii}^{-1} = \frac{\det(zI_{N-1} - W^{(i)})}{\det(zI - W)}$$

write next

$$zI_N - W = \begin{pmatrix} zI_{N-1} - W^{(N)} & w_N \\ w_N^T & z - W_{NN} \end{pmatrix}$$

Then one may write,

$$\det(zI_N - W) = \det(zI_{N-1} - W^{(N)}) \det(z - W_{NN} - \langle w_N, R_{W^{(N)}} w_N \rangle)$$

The last formula holds in the same manner with $W^{(i)}$, w_i, and W_{ii} replacing $W^{(N)}$, w_N, and W_{NN} respectively. □

Theorem D.2 (**Cauchy's interlace theorem**). *Let A_N be an $N \times N$ Hermitian matrix. Write,*

$$A_N = \begin{pmatrix} B_{N-1} & X \\ X^T & a_{NN} \end{pmatrix}$$

where B_{N-1} is the top left $(N-1) \times (N-1)$ principle submatrix of A_n, X is the rightmost $(N-1)$ column vector of A_N, X^T is the complex conjugate transpose of X,and a_{NN} is the bottom rightmost entry of A_N. Let $\alpha_1 \geq \cdots \geq \alpha_N$ be the eigenvalues of A_N, and let $\beta_1 \geq \cdots \geq \beta_{N-1}$ be the eigenvalues of B_{N-1}. Then,

$$\alpha_k \geq \beta_k \geq \alpha_{k+1}$$

Proof. Let u_i, for $i = 1, \ldots, n$, and v_j, for $j = 1, \ldots, n-1$ be the eigenvectors of A_n and B_{n-1}, respectively. Note that $u_i^T u_j = \delta_{ij}$, where δ_{ij} is the Kronecker delta, because A_n is Hermitian. Similarly, $v_i^T v_j = \delta_{ij}$. Let

$$w_i = \begin{bmatrix} v_i \\ 0 \end{bmatrix}.$$

Let $1 \leq k \leq n-1$. Denote the span of $u_k, \ldots, u_n$ by S_1 and the span of $w_1, \ldots, w_k$ by S_2. Note that $\dim(S_1) = n - k + 1$ and $\dim(S_2) = k$. By the elementary formula

$$\dim(S_1 \cap S_2) = \dim(S_1) + \dim(S_2) - \dim(S_1 + S_2),$$

we know that $\dim(S_1 \cap S_2) > 0$, since $\dim(S_1 + S_2) \leq n$. Thus, there exists $y \in S_1 \cap S_2$ such that $y^T y = 1$ and $\alpha_k \leq y^T A_n y \leq \beta_k$. This is because y is a linear combination of the eigenvectors $u_k, \ldots, u_n$ of A_n and a linear combination of the eigenvectors $w_1, \ldots, w_k$ of B_{n-1}. Thus,

$$\alpha_k \leq y^T A_n y \leq \alpha_n.$$

Also,

$$y^T A_n y = y^T B_{n-1} y, \quad \beta_1 \leq y^T B_{n-1} y \leq \beta_k$$

imply

$$\beta_1 \leq y^T A_n y \leq \beta_k.$$

Therefore,

$$\alpha_k \leq y^T A_n y \leq \beta_k.$$

Now we will consider

$$-A_n = \begin{bmatrix} -B_{n-1} & -X \\ -X^T & -a_{nn} \end{bmatrix}.$$

Then, the eigenvalues of $-A_n$ are $\tilde{\alpha}_1 \geq \ldots \geq \tilde{\alpha}_n$, where $\tilde{\alpha}_i = -\alpha_i$. Similarly, the eigenvalues of $-B_{n-1}$ are $\tilde{\beta}_1 \geq \ldots \geq \tilde{\beta}_n$, where $\tilde{\beta}_i = -\beta_i$. The eigenvectors are still u_i and v_j for $-A_n$ and $-B_{n-1}$, respectively. Then, let $\tilde{S}_1$ be the span of $u_1, \ldots, u_{k+1}$, and let $\tilde{S}_2$ be the span of $w_k, \ldots, w_n$. As before, there must exist some $z \in \tilde{S}_1 \cap \tilde{S}_2$ such that $z^T z = 1$ and

$$\tilde{\alpha}_1 \geq z^T (-A_n) z \geq \tilde{\alpha}_{k+1}$$

which means $z^T A_n z \leq -\tilde{\alpha}_{k+1} = \alpha_{k+1}$. Similarly,

$$z^T (-A_n) z = z^T (-B_{n-1}) z$$

which leads to the inequality

$$\tilde{\beta}_k \geq z^T (-B_n) z \geq \tilde{\beta}_n.$$

Thus, $z^T B_{n-1} z \geq -\tilde{\beta}_k = \beta_k$. Therefore, $\beta_k \leq z^T A_n z \leq \alpha_{k+1}$. Putting everything together, we have

$$\alpha_k \leq \beta_k \leq \alpha_{k+1}$$

□

E Woodbury Matrix Identity

Theorem E.1 (**Woodbury matrix identity**). *Let A be an invertible $N \times N$ matrix and let U, C, V be $N \times K$, $K \times K$, and $K \times N$ invertible matrices, respectively. The Woodbury matrix identity is*

$$(A + UCV)^{-1} = A^{-1} - A^{-1}U(C^{-1} + VA^{-1}U)^{-1}VA^{-1}$$

Proof. The formula can be proven by checking that $(A+UCV)$ times its alleged inverse on the right side of the Woodbury identity gives the identity matrix:

$$\begin{aligned}
&(A+UCV)\left[A^{-1}-A^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}\right]\\
=&\left\{I-U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}\right\}+\left\{UCVA^{-1}-UCVA^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}\right\}\\
=&\left\{I+UCVA^{-1}\right\}-\left\{U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}+UCVA^{-1}U\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}\right\}\\
=&I+UCVA^{-1}-\left(U+UCVA^{-1}U\right)\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}\\
=&I+UCVA^{-1}-UC\left(C^{-1}+VA^{-1}U\right)\left(C^{-1}+VA^{-1}U\right)^{-1}VA^{-1}\\
=&I+UCVA^{-1}-UCVA^{-1}\\
=&I.
\end{aligned}$$

□

Theorem E.2 (**Sherman-Morisson formula**). *Let A be an invertible $N\times N$ matrix. Let $t\in\mathbb{R}$ and $v\in\mathbb{R}^N$ we have that*

$$(A+tvv^T)^{-1}v=\frac{A^{-1}v}{1+t\langle v,A^{-1}v\rangle}$$

Proof. ($\Leftarrow$) To prove that the backward direction $1+v^\top A^{-1}u\neq 0\Rightarrow A+uv^\top$ is invertible with inverse given as above) is true, we verify the properties of the inverse. A matrix Y (in this case the right-hand side of the Sherman-Morrison formula) is the inverse of a matrix X (in this case $A+uv^\top$) if and only if $XY=YX=I$.

We first verify that the right hand side (Y) satisfies $XY=I$.

$$\begin{aligned}
XY&=\left(A+uv^\top\right)\left(A^{-1}-\frac{A^{-1}uv^\top A^{-1}}{1+v^\top A^{-1}u}\right)\\
&=AA^{-1}+uv^\top A^{-1}-\frac{AA^{-1}uv^\top A^{-1}+uv^\top A^{-1}uv^\top A^{-1}}{1+v^\top A^{-1}u}\\
&=I+uv^\top A^{-1}-\frac{uv^\top A^{-1}+uv^\top A^{-1}uv^\top A^{-1}}{1+v^\top A^{-1}u}\\
&=I+uv^\top A^{-1}-\frac{u\left(1+v^\top A^{-1}u\right)v^\top A^{-1}}{1+v^\top A^{-1}u}\\
&=I+uv^\top A^{-1}-uv^\top A^{-1}
\end{aligned}$$

To end the proof of this direction, we need to show that $YX=I$ in a similar way as above:

$$YX=\left(A^{-1}-\frac{A^{-1}uv^\top A^{-1}}{1+v^\top A^{-1}u}\right)\left(A+uv^\top\right)=I.$$

(In fact, the last step can be avoided since for square matrices X and Y, $XY=I$ is equivalent to $YX=I$.)

($\Rightarrow$) Reciprocally, if $1+v^\top A^{-1}u=0$, then via the matrix determinant lemma,

$$\det\left(A+uv^\top\right)=\left(1+v^\top A^{-1}u\right)\det(A)=0$$

So $\left(A+uv^\top\right)$ is not invertible. □

F $\mathscr{S}$-words and $\mathscr{S}$-sentences

Definition F.1 (**$\mathscr{S}$-words**). Given a set $\mathscr{S}$, an $\mathscr{S}$-letter s is simply an element of $\mathscr{S}$. An $\mathscr{S}$-word w is a finite sequence of letters $s_1\cdots s_n$ at least one letter long. An $\mathscr{S}$-word w is closed if its first and last letters are the same. Two $\mathscr{S}$-words w_1,w_2 are called equivalent, denoted $w_1\sim w_2$, if there is a bijection on $\mathscr{S}$ that maps one into the other.

Definition F.2 (**Length and weight of an $\mathscr{S}$-word**). For any $\mathscr{S}$-word $w = s_1 \cdots s_k$, we use $\mathcal{L}(w) = k$ to denote the length of w, define the weight $\mathcal{W}(w)$ as the number of distinct elements of the set $\{s_1, \cdots, s_k\}$ and the support of w, denoted $\text{supp}(w)$, as the set of letters appearing in w.

Definition F.3 (**Graph associated with an $\mathscr{S}$-word**). Given a word $w = s_1 \cdots s_k$, we let $G_w = (V_w, E_w)$ be the graph with set of vertices $V_w = \text{supp}(w)$ and undirected edges $E_w = \{\{s_i, s_{i+1}\}, i = 1, \cdots, k-1\}$. We define the set of self edges as $E_w^s = \{e \in E_w : e = \{u, u\}, u \in V_w\}$ and the set of connecting edges as $E_w^c = E_w \setminus E_w^s$. Also, for $e \in E_w$ we use N_e^w to denote the number of times this path traverses the edge e.

Remark F.4. *The graph G_w is connected since the word w defines a path connecting all the vertices of G_w which further starts and terminate at the same vertex if the word is closed.*

Lemma F.5 (**Equivalence of words and graphs**). *Equivalence words generate the same graph G_w and the same passage-count N_e^w.*

Definition F.6 (**Wigner word**). A closed word w of length $k + 1 \geq 1$ is called a Wigner word if either $k = 0$ or k is even and w is equivalent to an element of $\mathscr{W}_{k,k/2+1}$.

Lemma F.7 (**Wigner words and tree graphs**). *If $w \in \mathscr{W}_{k,k/2+1}$, then G_w is a tree*

Proof. Indeed, G_w is a connected graph with $|V_w| = k/2 + 1$, hence $|E_w| \geq k/2$, while the condition $N_e^w \geq 2$ for each $e \in E_w$ implies that $|E_w| \leq k/2$. Thus, $|E_w| = |V_w| - 1$, implying that G_w is a tree. □

Definition F.8 (**$\mathscr{S}$-sentences**). Given a set $\mathscr{S}$, an $\mathscr{S}$-sentence a is a finite sequence of $\mathscr{S}$-words $w_1, \cdots w_n$, at least one word long. Two $\mathscr{S}$-sentences a_1, a_2 are called equivalent, denoted $a_1 \sim a_2$ if there is a bijection on $\mathscr{S}$ that maps one into the other.

Definition F.9 (**Length and weigth of an $\mathscr{S}$-sentence**). For a sentence $a = (w_1, w_2, \cdots, w_n)$, we define the support as $\text{supp}(a) = \cup_{i=1}^n \text{supp}(w_i)$, and the weight $\mathcal{L}(a)$ as the cardinality of $\text{supp}(a)$.

Definition F.10 (**Graph associated with an $\mathscr{S}$-sentence**). Given a sentence $a = (w_1, \cdots, w_k)$, with $w_i = s_1^i s_2^i \cdots s_{l(w_i)}^i$, we set $G_a = (V_a, E_a)$ to be the graph with set of vertices $V_a = \text{supp}(a)$, and undirected edges

$$E_a = \{\{s_j^i, s_{j+1}^i\}, j = 1, \cdots, l(w_i) - 1, i = 1, \cdots, k\}$$

We define the set of self edges as $E_a^s = \{e \in E_a : e = \{u, u\}, u \in V_a\}$ and the set of connecting edges as $E_a^c = E_a \setminus E_a^s$.

Remark F.11. *The graph associated with a sentence may be disconnected.*

Definition F.12. The sentence a defines k paths in the graph G_a. For $e \in E_a$, we use N_e^a to denote the number of times the union of these paths traverses the edge e.

Lemma F.13. *Equivalent sentences generate the same graphs G_a and the same passage-count N_e^a.*

References

[1] A. Jagannath, *STAT 946: Mathematics of Data Science*, The University of Waterloo, Winter 2021.

[2] R. Vershynin, *High-Dimensional Probability*, Cambridge University Press, 2018.

[3] S. Mei, *STAT 260: Mean Field Asymptotics in Statistical Learning*, UC Berkeley, Spring 2021

[4] M. Rudelson & R. Vershynin, *Hanson-Wright Inequality and Sub-Gaussian Concentration*, arXiv:1306.2872

[5] Z. Bai & J. W. Silverstein, *Spectral Analysis of Large Dimensional Random Matrices*, Springer, 2010.

[6] C. Bordenave, P. Caputo, & D. Chafai, *Spectrum of non-Hermitian heavy tailed random matrices*. Communications in mathematical physics, 307(2), 2011.