553.810 Probability Theory III
Random Matrix Theory

Presented by: Joshua Agterberg

Contents

1 Setting **1**

2 Stieljes Transforms and Properties **2**

2.1 Properties and Definitions 2

2.2 Connections to Random Matrices 4

3 Marcenko-Pastur Law **4**

3.1 Low dimensions quickly 4

3.2 High dimensions 5

3.3 Proof Overview 9

4 Wigner Semicircle Law **10**

1 Setting

Random matrix theory is quite broad. We will mostly ignore measurability issues, and instead focus on very nice results in RMT with connections to high dimensional data analysis.

In general, "high-dimensional data analysis" refers to data analysis in high-dimensionality regimes, i.e. when you have (possibly more) features than observations. Classical asymptotic statistical theory involves fixing the dimension p of the data and sending the number of observations n to infinity. However, the primary reason we study asymptotics in the classical setting is that we typically would observe large n, small p, so that the asymptotics would essentially "kick in." However, in the current world of data science, we may not have an extremely large n and small p, meaning that asymptotics hasn't yet kicked in. So how do we understand this?

Well, instead of sending $n \to \infty$, we instead send both n and p to infinity, and assume something about the ratio p/n. For example, we may still assume that $p/n \to 0$, but p is allowed to grow with n. However, in typical high-dimensional data analysis problems, $p = O(n)$, so that p/n stays bounded as $n \to \infty$.

To be specific, we will assume that we either observe vectors $X_1, ..., X_n \in \mathbb{R}^p$ with $p/n \to c < \infty$, or random matrices $W_1, ..., W_n \in \mathbb{R}^{n\times n}$ with the entries W_{ij} iid F, where the matrices are symmetric. A basic starting point are the spectral properties of theses matrices; in particular, we will be focusing on the study of the eigenvalues. Since all of our matrices are symmetric, we know that all their eigenvalues are real, but what can we say more quantitatively about that?

For example, we may interested in the following:

- If we order the eigenvalues of W_i and $\mathbb{E}W_i = 0$ for all i, is there an appropiate normalization such that the empirical cdf of the eigenvalues of W_i converges to a fixed distribution as $n \to \infty$?
- If $p/n \to c < \infty$, what does the empirical distribution of the eigenvalues of the empirical covariance matrix look like (again after suitable normalization)?
- What about the top eigenvalues of theses matrices?

We will see that the first two questions are related to the Marcenko-Pastur Law and the Wigner Semicircicle Law, and the last question has to do with the Tracy-Widom Distribution. I will likely only be able to prove the first two results, but be advised that there are lots more results in random matrix theory that are distinct from the above results, such as concentration inequalities for matrices (Lu and Peng, 2012; Rudelson and Vershynin, 2010), limiting laws for the top eigenvalue (Karoui, 2007; Erdos et al., 2013; Wang, 2012; Knowles and Yin, 2017), properties of kernel random matrices (Karoui, 2010), and many others. In particular, the upcoming seminar relates to the BBP phase transition, as described in Baik et al. (2005) and further examined in Knowles and Yin (2017). Note that all of these results have to do with eigenvalues as opposed to eigenvector– there are a wide variety of results on eigenvectors as well.

2 Stieljes Transforms and Properties

There are two common ways to prove distributional results for eigenvalues. The first combinatorial bounds for moments and traces of the matrix in question, I call this "moment bashing." See Anderson et al. (2010) for some of these arguments. I will not be using this method for proofs, as they typically do not have any fundamentally new ideas contained within them outside of combinatorial bounds. Instead, I will be using what is referred to as the "Stieljes Transform method."

2.1 Properties and Definitions

First, we need a definition.

Definition 1 (Stieljes Transform)**.** Let ρ be a measure, and let $\mathbb{C}^+ := \{z \in \mathbb{C} : Im(Z) > 0\}$. Define $\mathcal{F} := \{f : \mathbb{C}^+ \to \mathbb{C}\}$, and $\mathcal{M} := \{\text{Measures on } \mathbb{R}\}$. The Stieljes transform $s_\rho(\cdot) : \mathcal{M} \to \mathcal{F}$ is defined as

$$s_\rho(z) = \int_{\mathbb{R}} (x - z)^{-1} d\rho(x),$$

for $z \in \mathbb{C}^+$ where it is defined.

We first explore some properties of s_ρ. For now, we will assume ρ is a probability measure.

Proposition 1 (Silverstein (2009))**.** *For a probability measure ρ the function s_ρ satisfies the following:*

1. *s_ρ is an analytic funciton on $\mathbb{C}^+$.*
2. *$Im(s_\rho) > 0$.*
3. *$|s_\rho(z)| \leq \frac{1}{Im(z)}$.*
4. *For continuity points of the CDF F_ρ of ρ,*

$$\rho([a, b]) = \frac{1}{\pi} \lim_{\eta \to 0+} \int_a^b Im(s_\rho(\xi + i\eta)) d\xi.$$

Proof. Proof of 1: Immediate, since it is an integral over $\mathbb{R}$ and there are no singularities.

Proof of 2: Let z be such that $Im(z) > 0$. Then for all $x \in \mathbb{R}$, $Im(\frac{1}{x-z}) = Im(\frac{x+z}{|x-z|} = Im(z/|x-z|) > 0$ by assumption. Then since the integrand satisfies $Im((x-z)^{-1}) > 0$ a.e. ρ, the result is proven.

Proof of 3: Fix $z \in \mathbb{C}^+$. Let ρ be the probability measure with mass at $x = Re(z)$. Then $s_{\rho(z)} = \frac{1}{Imz}$. The result follows by considering any other measure and noting that the absolute value of the integrand is maximized at $x = Re(z)$ (since it is the orthogonal projection of z to the real line, so that the denominator is minimized).

Proof of 4: Note that

$$\begin{aligned}
\frac{1}{\pi}\lim_{\eta\to 0^+}\int_a^b Im\left(\int_{\mathbb{R}} \frac{1}{x-\xi-i\eta} d\rho(x)\right) d\xi &= \frac{1}{\pi}\lim_{\eta\to 0^+}\int_a^b \int_{\mathbb{R}} Im\left(\frac{1}{x-\xi-i\eta}\right) d\rho(x) d\xi \\
&= \frac{1}{\pi}\lim_{\eta\to 0^+}\int_a^b \int_{\mathbb{R}} \frac{\eta}{(x-\xi)^2+\eta^2} d\rho(x) d\xi \\
&= \frac{1}{\pi}\lim_{\eta\to 0^+}\int_{\mathbb{R}} \int_a^b \frac{\eta}{(x-\xi)^2+\eta^2} d\xi d\rho(x) \\
&= \frac{1}{\pi}\lim_{\eta\to 0^+}\int_{\mathbb{R}} \left[\arctan\left(\frac{b-x}{\eta}\right) - \arctan\left(\frac{a-x}{\eta}\right)\right] d\rho(x) \\
&= \int_{\mathbb{R}} I_{[a,b]} d\rho(x) \\
&= \rho([a,b]).
\end{aligned}$$

□

Remark 1. Part 4 above can also be proven neatly, following Anderson et al. (2010). Assume X is distributed according μ, and define the random variable C_η independent of X, by a Cauchy distribution with parameter η, i.e. the density of C_η is given by

$$\frac{\eta}{\pi(x^2+\eta^2)}.$$

Then the convergence in 4 above is the rewriting of the convergence in distribution of $X + C_\eta$ to X as $\eta \to 0$.

Corollary 1. *If P and Q are probability measures, $s_P = s_Q$ if and only if $P = Q$.*

Proof. The fact follows by considering continuity points of the CDF and noting the inversion formula (point 4 above). □

We also have the following useful property for weak convergence.

Proposition 2 (Silverstein (2009))**.** *Let $S \subset C^+$ be a countable subset with a limit point in $\mathbb{C}^+$. Then a sequence of probability measures ρ_n converges vaguely to a probabilitiy measure ρ if and only if $s_{\rho_n}(z) \to s_\rho(z)$ for all $z \in S$.*

Proof. Suppose $\rho_n \to \rho$. Then by the Portmanteau Lemma, $\mathbb{E}_{\rho_n} f(X) \to \mathbb{E}_\rho f(X)$ for all continuous bounded functions f. In particular, for all $z \in S$, the function $f(x) = (x-z)^{-1}$ is a continuous bounded function (consider its real and imaginary parts separately). This shows that $s_{\rho_n}(z) \to s_\rho(z)$ for all $z \in S$.

Suppose $s_{\rho_n}(z) \to s_\rho(z)$ for all $z \in S$. Then from complex analysis, since S is countable and has a limit point, s_{ρ_n} is completley determined by the values it takes on S. The result follows by considering all continuity points of the corresponding CDFs and the inversion formula. □

If one wants to prove results about empirical distributions, one can use Proposition 2 with ρ_n as the empirical measures and ρ as a limiting measure, and "reduce" the problem to only worrying about s_{ρ_n}. Before we consider random matrices, we need to make sure that we have a probability measure.

Example 1 (Geronimo and Hill (2003))**.** Let $P_n := \delta_n$, the dirac measure at n. Then $s_{P_n} = (n-z)^{-1}$ for all n and z, so that $\lim s_{P_n}(z) \equiv 0$, which is not a Stieljes transform of any probability measure P.

Much like in the general case concerning weak convergence of probability measures, we need to find a condition where we can ensure convergence of $P_n \to P$ in distribution (where mass does not "escape to infinity"). Actually, the very readable Geronimo and Hill (2003) gives the condition that $\lim_{y\to\infty} iyS_\rho(iy) = -1$ as necessary and sufficient. One can easily check that this is satisfied for probability measures, so if the limiting measure/Stieljes transform is known, then the problem is much simpler. Clearly, the previous example doesn't satisfy this.

2.2 Connections to Random Matrices

Suppose we have a Hermitian $p \times p$ matrix A generated by some fixed distribution on $\mathbb{R}^p$ with n observations, where $p \asymp n$, but $p/n \to c \in (0,1)$. We let $\hat{F}_n(A)$ denote the empirical distribution function of the eigenvalues $\lambda_1(A) \le \lambda_2(A) \le ... \le \lambda_p(A)$ of A (which are all real). In other words,

$$\hat{F}_n(A)(x) := \frac{1}{p}\sum_{i=1}^{p} \mathbb{I}\{\lambda_i(A) \le x\},$$

which counts the number of eigenvalues less than or equal to x. This is a well-defined measure. Furthermore,

$$\begin{aligned} s_{\hat{F}_n}(z) &= \int_{\mathbb{R}} \frac{1}{x-z} d\hat{F}_n(x) \\ &= \frac{1}{p}\sum_{i=1}^{p} \frac{1}{\lambda_i(A) - z} \\ &= \frac{1}{p} Tr(A - zI)^{-1} \end{aligned}$$

since the eigenvalues of $(A - zI)^{-1}$ are given by the values in the summand. Hence, by Proposition 2, it suffices to show that if A_n are a sequence of $n \times n$ random matrices that $\frac{1}{p}Tr(A - zI)^{-1} \to s_\rho(z)$ almost surely to show convergence of the distribution of the eigenvalues.

3 Marcenko-Pastur Law

3.1 Low dimensions quickly

Suppose $\{X_i\}_{i=1}^n$ are a sequence of mean-zero random variables taking values in $\mathbb{R}^p$ with p fixed and covariance Σ. Define

$$\hat{\Sigma} := \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top.$$

By the Law of Large Numbers, $||\hat{\Sigma} - \Sigma||_2 \to 0$ almost surely. In addition, by Weyl's inequality,

$$|\lambda_i(\hat{\Sigma}) - \lambda_i(\Sigma)| \le ||\hat{\Sigma} - \Sigma||_2,$$

so that the eigenvalues of $\hat{\Sigma}$ are tending to $\lambda_i(\Sigma)$. We could have also used the continuous mapping theorem since the eigenvalues are continuous functions of the entries of the matrix. Regardless, we see that the eigenvalues of $\hat{\Sigma}$ are tending towards the eigenvalues of Σ almost surely. In particular, there is no randomness in the limit – if we picked an eigenvalue at random, it would just be a uniform distribution on the eigenvalues of Σ. In particular, when $\Sigma = I_p$, all the eigenvalues are just 1, so the random eigenvalue we'd pick would just be 1. This is contrast to the high-dimensional regime, in which if we were to randomly pick an eigenvalue, we'd actually get a distribution.

3.2 High dimensions

Primary source: RMT.

If you are concerned with the measure-theoretic aspects we are considering here, we can actually define the double-array X_{ik} and assume that the rows are iid, where we slowly "reveal" the upper left-hand rectangle of this array at a rate with $p \asymp n$.

We first state the Marcenko-Pastur Theorem.

Theorem 1. *Let $X_1, ..., X_n \in \mathbb{R}^p$ be random variables whose entries are independent identically distributed random variables with mean 0 and variance $\sigma^2 < \infty$. Define $\mathbf{Y_n} := \frac{1}{n}\mathbf{X}^\top \mathbf{X} \in \mathbb{R}^{p\times p}$, where each row of $\mathbf{X}$ is the $\mathbb{R}^p$ vector X_i. Let $\lambda_1, ...\lambda_p$ be the eigenvalues of $\mathbf{Y_n}$, and denote by $\hat{F}_n$ their empirical distribution. Assume $p/n \to c \in (0, \infty)$. Then $\hat{F}_n \to \mu$ where μ is a probability measure of the form*

$$\mu(A) = \begin{cases} (1 - \frac{1}{c})1_{0\in A} + \nu_{1/c}(A) & c > 1 \\ \nu_c(A) & 0 \leq c \leq 1 \end{cases}$$

where $d\nu_c(x) := \frac{1}{2\pi\sigma^2}\frac{\sqrt{(\lambda_+(c)-x)(x-\lambda_-(c))}}{cx}$ for $x \in [\lambda_-, \lambda_+]$ with $\lambda_\pm = \sigma^2(1 \pm c)^2$.

Let's take a little to examine the statement of this theorem. First, if $c > 1$, it is clear that there has to be point mass at zero, since the matrix is rank $n < p$. It follows that the fraction of zero eigenvalues should be approximately $1 - \frac{1}{c} \approx \frac{p-n}{p}$, so that there are $p - n$ zeros for each matrix (as expected). For $p < n$ (the more interesting case), we see that the pdf is compactly supported on the interval above, and approximately looks peaked towards the lower half. when $c = 1$, the interval is simply $[0, 4\sigma^2]$, and the density looks like a "quarter circle."

Note that one can weaken the statement that the coordinates have the same variance and are iid. We can actually assume that each X_i has a variance Σ_p, and then we neecd to assume that $\frac{1}{p}Tr\Sigma_p \to \gamma < \infty$; or, in other words, each different eigenvalue is $O(1)$. Note that in the iid case with constant variance, $Tr\Sigma_p = p\sigma^2$, so that $\gamma = \sigma^2$.

What are the data-analytic consequences for this theorem? For a fixed data set, one might want to normalize and examine a histogram of the eigenvalues of the covariance matrix; if this looks approximately Marcenko-Pastur, one may need to be careful about what sort of techniques one is using, since classical statistical methods may fail in this regime.

Loosely the proof will be as follows. Define s_n as the Stieljes transform for $\hat{F}_n$. First, we show that $s_n - \mathbb{E}s_n \to 0$ a.s. and pointwise. Then we show that $\mathbb{E}s_n \to s_\nu$ where ν is the measure defined above, again pointwise.

We will need a series of small lemmas. We only prove in the case $c \leq 1$, and in the case that $\sigma^2 = 1$ (the first is not without loss of generality, but the analysis is similar, and the second is without loss of generality).

Lemma 1 (Stieljes Transform of the M-P Law)**.** *The Stieljes Transform of the M-P law above is given by*

$$s_\nu(z) = \frac{z + 1 - c - \sqrt{(z - 1 - c^2)^2 - 4c}}{2z}.$$

Proof. I will use the inversion formula to prove this. Note that since in the inversion formula we have an integral, by the Dominated Convergence Theorem we only need to show that the imaginary part of the Stieljes transform is equal to the pdf on any interval $[a, b]$ in the support. To wit, we have

$$
\begin{aligned}
&\frac{1}{\pi}\lim_{y\to 0} Im\left(\frac{x+iy-c-\sqrt{(x+iy-1-c^2)^2-4c}}{2(x+iy}\right)\\
&\qquad = \frac{1}{2\pi}\lim_{y\to 0} Im\left(\frac{x+iy-c-\sqrt{(x+iy-1-c^2)^2-4c}}{x+iy}\right)\\
&\qquad = \frac{1}{2\pi}\lim_{y\to 0} Im\left(\frac{(x-iy)(x+iy-c-\sqrt{(x+iy-1-c^2)^2-4c)}}{x^2+y^2}\right)\\
&\qquad = \frac{1}{2\pi}\lim_{y\to 0} Im\left(\frac{x^2+y^2-cx+iyc-(x-iy)\sqrt{(x+iy-1-c^2)^2-4c}}{x^2+y^2}\right)\\
&\qquad = \frac{1}{2\pi}\lim_{y\to 0} Im\left(1+\frac{-cx+iyc}{x^2+y^2}-\frac{(x-iy)\sqrt{(x+iy-1-c^2)^2-4c}}{x^2+y^2}\right)\\
&\qquad = \frac{1}{2\pi}\lim_{y\to 0} Im\left(-\frac{(x-iy)\sqrt{(x+iy-1-c^2)^2-4c}}{x^2+y^2}\right)\\
&\qquad = \frac{1}{2\pi}\lim_{y\to 0} Im\left(-\frac{(x-iy)\sqrt{c^4-2c^2x+2c^2+x^2-2x-1-2ic^2y+i2xy-y^2-2iy}}{x^2+y^2}\right)\\
&\qquad = \frac{1}{2\pi} Im\left(-\frac{\sqrt{c^4-2c^2x+2c^2+x^2-2x-1}}{x}\right)\\
&\qquad = \frac{1}{2\pi}\frac{\sqrt{-c^4+2c^2x-2c^2-x^2+2x+1}}{x}
\end{aligned}
$$

which factorizes to be the pdf. □

This next lemma shows how we will characterize the Stieljes transform of the MP distribution. There are only two possible solutions, and the only one with $Im(z) > 0$ is the one with the Stieljes Transform desired.

Lemma 2. *Consider the quadratic equation $zS^2 + (z-1+c)S + c = 0$, where $S = s(z)$. Then the unique solution is $s(z) = cs_\nu(z)$, where $s_\nu(z)$ is the Stieljes Transform of the Marcenko-Pastur distribution with $c = 1$.*

Now that we have the Stieljes transform desired and a characterization in terms of the quadratic equation, the goal will be to write the formula for the $s_n(z)$ in terms of the equation above plus terms of order $o(1)$. Now we do some manipulations.

We first recall the Stieljes transform of the original matrix:

$$
s_n(z) = \frac{1}{p}\mathrm{Tr}\left((\mathbf{Y_n} - zI)^{-1}\right),
$$

where $\mathbf{Y_n} = \frac{1}{n}\sum_{i=1}^{n} X_i X_i^\top$, which are all $p \times p$ rank-one matrices. Define

$$
\begin{aligned}
\mathbf{Y_n^{(i)}} &:= \mathbf{Y_n} - \frac{1}{n}X_i X_i^\top\\
&= \frac{1}{n}\sum_{j\neq i} X_j X_j^\top,
\end{aligned}
$$

and the corresponding resolvents $\mathbf{G_n} = (\mathbf{Y_n} - zI_p)^{-1}$, and $\mathbf{G_n^{(k)}} = \left(\mathbf{Y_n^{(k)}} - zI_p\right)^{-1}$.

Recall we are interested in $\frac{1}{p}\mathrm{Tr}(\mathbf{G_n}(z))$. We can derive the following facts.

Lemma 3.

$$
X_k^\top \mathbf{G_n}(z) X_k = \frac{X_k^\top \mathbf{G_n^{(k)}}(z) X_k}{1 + \frac{1}{n} X_k^\top \mathbf{G_n^{(k)}}(z) X_k}
$$

Proof. First, note that

$$\begin{aligned}
X_k^\top \mathbf{G}_{\mathbf{n}}^{(\mathbf{k})}(z)\Big(\mathbf{G}_{\mathbf{n}}(z)\Big)^{-1} &= X_k^\top \mathbf{G}_{\mathbf{n}}^{(\mathbf{k})}(z)\Big(\mathbf{Y}_{\mathbf{n}} - zI_p\Big) \\
&= X_k^\top \mathbf{G}_{\mathbf{n}}^{(\mathbf{k})}(z)\Big(\mathbf{Y}_{\mathbf{n}}^{(\mathbf{k})} - zI_p + \frac{1}{n}X_kX_k^\top\Big) \\
&= X_k^\top \mathbf{G}_{\mathbf{n}}^{(\mathbf{k})}(z)\Big(\mathbf{Y}_{\mathbf{n}}^{(\mathbf{k})} - zI_p\Big) + X_k^\top \mathbf{G}_{\mathbf{n}}^{(\mathbf{k})}(z)\Big(\frac{1}{n}X_kX_k^\top\Big) \\
&= X_k^\top + X_k^\top \mathbf{G}_{\mathbf{n}}^{(\mathbf{k})}(z)\Big(\frac{1}{n}X_kX_k^\top\Big) \\
&= \Big[1 + \frac{1}{n}X_k^\top \mathbf{G}_{\mathbf{n}}^{(\mathbf{k})}(z)X_k)\Big]X_k^\top.
\end{aligned}$$

Rearranging the resulting equation yields

$$X_k^\top \mathbf{G}_{\mathbf{n}}^{(\mathbf{k})}(z) = \Big[1 + \frac{1}{n}X_k^\top \mathbf{G}_{\mathbf{n}}^{(\mathbf{k})}(z)X_k)\Big]X_k^\top \mathbf{G}_{\mathbf{n}}(z)$$

and multiplying on the left by the vector X_k yields

$$X_k^\top \mathbf{G}_{\mathbf{n}}^{(\mathbf{k})}(z)X_k = \Big[1 + \frac{1}{n}X_k^\top \mathbf{G}_{\mathbf{n}}^{(\mathbf{k})}(z)X_k)\Big]X_k^\top \mathbf{G}_{\mathbf{n}}(z)X_k.$$

Rearranging gives the result. □

Lemma 4.

$$\frac{1}{p}\mathrm{Tr}(\mathbf{G}_{\mathbf{n}}(z)) = -\frac{1}{nz}\sum_{k=1}^{n}\frac{1}{1 + \frac{1}{n}X_k^\top \mathbf{G}_{\mathbf{n}}^{(\mathbf{k})}(z)X_k} + o(1)$$

Proof. Simply compute

$$\begin{aligned}
1 &= \frac{1}{p}\mathrm{Tr}(I_p) \\
&= \frac{1}{p}\mathrm{Tr}\Big(\Big(\mathbf{Y}_{\mathbf{n}} - zI\Big)\mathbf{G}_{\mathbf{n}}(z)\Big) \\
&= \frac{1}{p}\mathrm{Tr}\Big(\frac{1}{n}\sum_{k=1}^{n} X_kX_k^\top \mathbf{G}_{\mathbf{n}}(z) - z\mathbf{G}^{\mathbf{n}}(z)\Big) \\
&= \frac{1}{p}\mathrm{Tr}\Big(\frac{1}{n}\sum_{k=1}^{n} X_kX_k^\top \mathbf{G}_{\mathbf{n}}(z)\Big) - \frac{z}{p}\mathrm{Tr}\Big(\mathbf{G}_{\mathbf{n}}(z)\Big) \\
&= \frac{1}{p}\frac{1}{n}\sum_{k=1}^{n}\mathrm{Tr}\Big(X_kX_k^\top \mathbf{G}^{\mathbf{n}}(z)\Big) - \frac{z}{p}\mathrm{Tr}\Big(\mathbf{G}_{\mathbf{n}}(z)\Big) \\
&= \frac{1}{p}\frac{1}{n}\sum_{k=1}^{n}\mathrm{Tr}\Big(X_k^\top \mathbf{G}_{\mathbf{n}}(z)X_k\Big) - \frac{z}{p}\mathrm{Tr}\Big(\mathbf{G}_{\mathbf{n}}(z)\Big) \\
&= \frac{1}{p}\frac{1}{n}\sum_{k=1}^{n}\Big(X_k^\top \mathbf{G}_{\mathbf{n}}(z)X_k\Big) - \frac{z}{p}\mathrm{Tr}\Big(\mathbf{G}_{\mathbf{n}}(z)\Big) \\
&= \frac{1}{p}\frac{1}{n}\sum_{k=1}^{n}\Big(\frac{X_k^\top \mathbf{G}_{\mathbf{n}}^{(\mathbf{k})}(z)X_k}{1 + \frac{1}{n}X_k^\top \mathbf{G}_{\mathbf{n}}^{(\mathbf{k})}(z)X_k}\Big) - \frac{z}{p}\mathrm{Tr}\Big(\mathbf{G}_{\mathbf{n}}(z)\Big).
\end{aligned}$$

Rearrange the equation to obtain

$$\begin{aligned}
\frac{z}{p}\mathrm{Tr}\Bigg(\mathbf{G_n}(z)\Bigg) &= \frac{1}{p}\frac{1}{n}\sum_{k=1}^{n}\Bigg(\frac{X_k^\top \mathbf{G_n^{(k)}}(z)X_k}{1+\frac{1}{n}X_k^\top \mathbf{G_n^{(k)}}(z)X_k}\Bigg) - 1 \\
&= \frac{1}{p}\frac{1}{n}\sum_{k=1}^{n}\Bigg(\frac{X_k^\top \mathbf{G_n^{(k)}}(z)X_k}{1+\frac{1}{n}X_k^\top \mathbf{G_n^{(k)}}(z)X_k}\Bigg) - \frac{1}{n}\frac{1}{p}\sum_{k=1}^{n} p \\
&= \frac{1}{p}\frac{1}{n}\sum_{k=1}^{n}\Bigg(\frac{X_k^\top \mathbf{G_n^{(k)}}(z)X_k}{1+\frac{1}{n}X_k^\top \mathbf{G_n^{(k)}}(z)X_k} - p\Bigg).
\end{aligned}$$

Rearrange another time by multiplying through by p and modifying the p factor, we obtain

$$\begin{aligned}
z\mathrm{Tr}\Bigg(\mathbf{G_n}(z)\Bigg) &= \frac{p}{n}\sum_{k=1}^{n}\Bigg(\frac{\frac{1}{p}X_k^\top \mathbf{G_n^{(k)}}(z)X_k}{1+\frac{1}{n}X_k^\top \mathbf{G_n^{(k)}}(z)X_k} - 1\Bigg) \\
&= -\frac{p}{n}\sum_{k=1}^{n}\Bigg(\frac{1+\frac{p-n}{np}X_k^\top \mathbf{G_n^{(k)}}(z)X_k}{1-\frac{1}{n}X_k^\top \mathbf{G_n^{(k)}}(z)X_k}\Bigg) \\
&= -\frac{p}{n}\sum_{k=1}^{n}\Bigg(\frac{1}{1-\frac{1}{n}X_k^\top \mathbf{G_n^{(k)}}(z)X_k}\Bigg) + \frac{p}{n}\frac{p-n}{np}\sum_{k=1}^{n}\frac{X_k^\top \mathbf{G_n^{(k)}}(z)X_k}{1-\frac{1}{n}X_k^\top \mathbf{G_n^{(k)}}(z)X_k} \\
&= -\frac{p}{n}\sum_{k=1}^{n}\Bigg(\frac{1}{1-\frac{1}{n}X_k^\top \mathbf{G_n^{(k)}}(z)X_k}\Bigg) + o(1)
\end{aligned}$$

where the $o(1)$ term comes from the fact that $p/n \to c$ and $p-n = o(np)$ and $n, p \to \infty$. □

From here on, we assume $p/n \to 1$, since our $o(1)$ term above will make not much difference. The result holds for any p and n, but does not really reveal the heart of the problem.

Also, the notes I used as a main reference assume $p = n$.

Lemma 5.

$$\mathbb{P}\Bigg[|\frac{1}{n}X_k^\top \mathbf{G_n^{(k)}}X_k - \frac{1}{n}\mathrm{Tr}(\mathbf{G_n^{(k)}}(z))| \geq \varepsilon\Bigg] \leq \frac{C}{n^2}.$$

Hence, by the Borel-Cantelli Lemma, $\frac{1}{n}X_k^\top \mathbf{G_n^{(k)}}X_k - \frac{1}{n}\mathrm{Tr}(\mathbf{G_n^{(k)}}(z)) \to 0$ *almost surely.*

Note: proving this lemma requires a fourth moment Markov inequality and the fact that

$$\mathbb{E}\Bigg(\frac{1}{n}X_k^\top \mathbf{G_n^{(k)}}X_k - \frac{1}{n}\mathrm{Tr}(\mathbf{G_n^{(k)}}(z))\Bigg) = 0. \tag{1}$$

I will skip the fourth moment markov inequality to avoid calculating fourth moments of big random variables.

Showing that 1 holds is much easier. Note that since $\mathbf{Y_n^{(k)}}$ does not include the k-th vector X_k, it is independent of it, so we have

$$\begin{aligned}
\mathbb{E}\frac{1}{n}X_k^\top \mathbf{G_n^{(k)}}X_k &= \frac{1}{n}\sum_{j,l=1}^{n}\mathbb{E}\Bigg(X_k(j)X_k(l)\Bigg)\mathbb{E}\Bigg(\Bigg(\mathbf{Y_n^{(k)}} - zI\Bigg)^{-1}_{jl}\Bigg) \\
&= \frac{1}{n}\sum_{j=1}^{n}\mathbb{E}\Bigg(\Bigg(\mathbf{Y_n^{(k)}} - zI\Bigg)^{-1}_{jj}\Bigg) \\
&= \mathbb{E}\Bigg(\frac{1}{n}\mathrm{Tr}(\mathbf{G_n^{(k)}}z)\Bigg).
\end{aligned}$$

This next lemma is a bit of a technicality.

Lemma 6. *For all $z \in \mathbb{C} \setminus \mathbb{R}$,*

$$\left|\frac{1}{n}\text{Tr}\bigg(\mathbf{G_n^{(k)}}(z)\bigg) - \frac{1}{n}\text{Tr}\bigg(\mathbf{G_n}(z)\bigg)\right| \leq \frac{1}{n|Im(z)|}.$$

Proof. First, note that $\mathbf{Y_n} = \sum_{i=1}^n X_i X_i^\top$ and $\mathbf{Y_n^{(k)}} = \sum_{i=1}^n X_i X_i^\top - X_k X_k^\top$, or in the perhaps more revealing way as

$$\mathbf{Y_n} = \mathbf{Y_n^{(k)}} + X_k X_k^\top;$$

that is $\mathbf{Y_n}$ is a rank-one perturbation of $\mathbf{Y_n^{(k)}}$. Hence, by the Cauchy-Interlacing inequalities,

$$\lambda_k(\mathbf{Y_n^{(k)}}) \leq \lambda_{k+1}(\mathbf{Y_n}) \leq \lambda_{k+2}(\mathbf{Y_n^{(k)}});$$
$$\lambda_k(\mathbf{Y_n}) \leq \lambda_{k+1}(\mathbf{Y_n^{(k)}}) \leq \lambda_{k+2}(\mathbf{Y_n})$$

for $1 \leq k \leq n-2$. Now, recall $\mathbf{G_n}$ has the eigenvalues $\frac{1}{\lambda_i(\mathbf{Y_n})-z}$, so that

$$\lambda_{k+2}(\mathbf{G_n^{(k)}}) \leq \lambda_{k+1}(\mathbf{G_n}) \leq \lambda_k(\mathbf{G_n^{(k)}});$$
$$\lambda_{k+2}(\mathbf{G_n}) \leq \lambda_{k+1}(\mathbf{G_n^{(k)}}) \leq \lambda_k(\mathbf{G_n}).$$

Summing up over $k = 1, ..., n$ gives that

$$\left|\text{Tr}\bigg(\mathbf{G_n^{(k)}}(z)\bigg) - \text{Tr}\bigg(\mathbf{G_n}(z)\bigg)\right| \leq \left|\frac{1}{\lambda_1(\mathbf{Y_n}) - z}\right|,$$

which is a Stieljes transform of just point mass at one point. Hence, the result follows by multiplying by $\frac{1}{n}$ and applying Proposition 1, part 3. □

Proof Idea of Marcenko-Pastur Theorem. We see that

$$\begin{aligned}
s_n(z) :&= \frac{1}{n}\text{Tr}(\mathbf{G_n}(z)) \\
&= -\frac{1}{nz}\sum_{k=1}^n \frac{1}{1 + \frac{1}{n}X_k^\top \mathbf{G_n^{(k)}}(z) X_k} + o(1) \\
&= -\frac{1}{nz}\sum_{k=1}^n \frac{1}{1 + \frac{1}{n}\text{Tr}\bigg(\mathbf{G_n^{(k)}}(z)\bigg)} + o(1) \\
&= -\frac{1}{z}\frac{1}{1 + \frac{1}{n}\text{Tr}\bigg(\mathbf{G_n}(z)\bigg)} + o(1) \\
:&= -\frac{1}{z}\frac{1}{1 + s_n(z)} + o(1).
\end{aligned}$$

We can rewrite this as

$$zs_n^2 + zs_n + 1 = o(1),$$

so that s_n is tending towards the solution with positive imaginary part almost surely. Note we would still have to take care of the $o(1)$'s above, but that will be the focus of the next section. □

3.3 Proof Overview

Lots of the lemmas were mostly technical, so it seemed relevant to give the proof overview here.

Set $s_n(z) := \frac{1}{n}\text{Tr}(\mathbf{G_n}(z))$, the Stieljes transform of the empirical eigenvalue distribution, and $s_\nu(z)$ as our limiting Stieljes Transform.

- We first characterized the Stieljes transform we wanted exactly ($s_\nu(z)$) and showed it is a solution to a quadratic (or some other pde perhaps)
- We then used a bit of matrix analysis and real analysis to rewrite $s_n(z)$ in terms we could understand
- After rewriting, we showed we could write $s_n(z)$ on either side of an equation with high probability using essentially Markov bounds
- Assuming we had this equation, we showed that $s_n(z)$ approximately satisfies the quadratic above, yielding to almost sure convergence of $s_n(z)$ to the Stieljes transform of our limiting distribution.

This technique is common for Stieljes-transform proofs. Another common method of proof is showing that first the distribution is moment-determined, and then showing that all the moments converge; this was the original method given by Wigner in his proof of the Semicircle Law. This involves lots of combinatorial approximations to show that certain numbers converge; these can be quite tedious. We mostly avoid this with the Stieljes transform method (except for our one particular bound).

4 Wigner Semicircle Law

We begin with a definition.

Definition 2 (Wigner matrix)**.** A Wigner matrix A is a random Hermitian (symmetric) matrix with the entries a_{ij} are mean zero and variance 1 iid for $i \neq j$, and the entries on the diagonal are real-valued, independent, and iid with mean zero and variance 1. Furthermore, all entries are uniformly bounded by some K.

Theorem 2. *Suppose W_n is a sequence of $n \times n$ Wigner matrices, and let μ_n denote the empirical spectral distribution of $\frac{1}{\sqrt{n}} W_n$. Then $\mu_n \to \mu$ weakly almost surely where μ is the measure associated to the semicircle law*

$$d\mu = \frac{1}{2\pi}\sqrt{4 - x^2} \qquad |x| \leq 2$$

where $d\mu$ is the radon-nikodym derivative (pdf) of μ with respect to Lebesgue measure.

The proof is almost exactly the same as the previous one. We will assume that the entries are bounded; i.e. $|\sqrt{n} W_{ij}| \leq C$ for some $C > 0$. This is not strictly required, but involves a short truncation argument (i.e. we can replace the matrices with their truncated versions almost surely). We will prove that the convergence of $\mu_n \to \mu$ happens in probability, since higher moments are more difficult to bound.

First the proof shows that we have a similar characterization of the semicircle law in terms of some ODE, plust some $\delta_n(z)$ term. The rest of the proof proceeds first by showing that $\delta_n(z) \to 0$ for any fixed $z \in \mathbb{C} \setminus \mathbb{R}$. Since in the first section I focused on showing that we could get such a characterization, in this section, I will show that $\delta_n(z)$ (the analogue of our $o(1)$ term in the previous proof) actually does go to zero.

The proof follows Anderson et al. (2010).

Proof. Define $s_n(z)$ similarly as before (the Stieljes Transform of μ_n), and $\mathbf{G_n}(z) := (W_n - zI)^{-1}$. Then by Cramer's rule,

$$(W_n - zI)^{-1}(i,i) = \frac{\det(W^{(i)} - zI_{n-1})}{\det(W - zI)},$$

where $W^{(i)}$ is the principal submatrix. Furthermore,

$$W - zI_n = \begin{pmatrix} W^{(n)} - zI_{n-1} & w_n \\ w_n^\top & W(n,n) - z \end{pmatrix}.$$

By a classic property of determinants, we see that

$$\det(W - zI_n) = \det(W^{(n)} - zI_{n-1}) \det(W(n,n) - z - w_n^\top (W^{(n)} - zI_{n-1})^{-1} w_n).$$

Plugging this in above, we see that

$$\begin{aligned}(W_n - zI)^{-1}(n,n) &= \frac{\det(W^{(i)} - zI_{n-1})}{\det(W - zI)} \\ &= \frac{\det(W^i - zI_{n-1})}{\det(W^{(n)} - zI_{n-1}) \det(W(n,n) - z - w_n^\top (W^{(n)} - zI_{n-1})^{-1} w_n)} \\ &= \frac{1}{\det(W(n,n) - z - w_n^\top (W^{(n)} - zI_{n-1})^{-1} w_n)} \\ &= \frac{1}{W(n,n) - z - w_n^\top (W^{(n)} - zI_{n-1})^{-1} w_n},\end{aligned}$$

since it is just a number. Hence, we have a characterization for the diagonals of $(W_n - zI)^{-1}$ since we only need to replace w_n above with w_i and $W^{(n)}$ above with $W^{(i)}$ to get the characterization for any i.

We will now completely switch notation from above, but it shouldn't make much a difference.

Let α_k k'th column of W and $\tilde{\alpha}_k$ be the $n-1$ dimensional vector obtained from α_k by erasing the entry $\alpha_k(k)$, and let $W_n^{(i)}$ the matrix W_n with the i'th row and column removed. Using the above characterization, we see that

$$\begin{aligned}\frac{1}{n} s_n(z) &= \frac{1}{n} \sum_{i=1}^n \frac{1}{-z - \tilde{\alpha}_i^\top (W_n^{(i)} - zI_{n-1}) \tilde{\alpha}_i} \\ &= -\frac{1}{z + \frac{1}{n} s_n(z)} - \delta_n(z),\end{aligned}$$

where

$$\delta_n(z) := \frac{1}{n} \sum_{i=1}^n \frac{\varepsilon_{i,n}}{(-z - n^{-1} s_n(z) + \varepsilon_{i,n})(-z - n^{-1} s_n(z))}$$

and

$$\varepsilon_{i,n} = n^{-1} s_n(z) - \tilde{\alpha}_i^\top (W_n^{(i)} - zI_{n-1}) \tilde{\alpha}_i.$$

First, note that term in the denominator $-z - n^{-1} s_n(z)$ has modulus at least $Im(z)$ by a similar argument to the proof of Lemma 6. Hence, if $\sup_{i \le n} |\varepsilon_{i,n}| \to 0$ in probability, then it will also be true that $\delta_i \to 0$ in probability.

Let the matrix $\bar{W}_n^{(i)}$ be the same matrix as $W_n^{(i)}$ except with the $i'th$ row and column set to zero (instead of removed). Then their eigenvalues coincide, so again using similar logic to the proof of Lemma 6,

$$\left| \frac{1}{n} \mathrm{Tr}(\bar{W}_n^{(i)} - zI_n)^{-1} - \frac{1}{n} \mathrm{Tr}(W_n^{(i)} - zI_{n-1})^{-1} \right| \le \frac{1}{nIm(z)}.$$

Now, let $\lambda_1^{(i)} \le ... \lambda_n^{(i)}$ be the eigenvalues of $\bar{W}_n^{(i)}$, and let $\lambda_1 \le ... \le \lambda_n$ be the eigenvalues of W_n. Then note that $(\bar{W}_n^{(i)} - zI_n)^{-1}$ is a principle submatrix of $\mathbf{G_n}$, so

$$\begin{aligned}\frac{1}{n} \left| \mathrm{Tr}\mathbf{G_n} - \mathrm{Tr}(\bar{W}_n^{(i)} - zI_n)^{-1} \right| &\le \frac{1}{nIm(z)^2} \sum_{k=1}^n |\lambda_k^{(i)} - \lambda_k| \\ &\le \frac{1}{Im(z)^2} \left(\frac{1}{n^2} \sum_{k=1}^n |\lambda_k^{(i)} - \lambda_k|^2 \right)^{1/2} \\ &\le \frac{1}{Im(z)^2} \left(\frac{2}{n^2} \sum_{k=1}^n W_n(i,k)^2 \right)^{1/2}\end{aligned}$$

by the Hoffman-Wielandt inequality. Since all of the $W_n(i,k)$ are bounded by $C/\sqrt{n}$, this term tends to zero. Hence, we only need to worry that $\sup|\bar{\varepsilon}_{i,n}|$ tends to zero in probability, where

$$\begin{aligned}
\bar{\varepsilon}_{i,n} &= \tilde{\alpha}_i^\top B_n^{(i)}(z)\tilde{\alpha}_i - \frac{1}{n}\mathrm{Tr}B_n^{(i)}(z) \\
&= \frac{1}{n}\sum_{k=1}^{n-1}\left(\left[\sqrt{n}\tilde{\alpha}_i(k)\right]^2 - 1\right)B_n^{(i)}(z)(k,k) + \sum_{k,k'=1,k\neq k'}^{n-1}\tilde{\alpha}_i(k)\tilde{\alpha}_i(k')B_n^{(i)}(z)(k,k') \\
:&= \bar{\varepsilon}_{i,n}^{diag} + \bar{\varepsilon}_{i,n}^{off}
\end{aligned}$$

and $B_n^{(i)}(z) = (W_n^{(i)} - zI_{n-1})^{-1}$.

Note that $\tilde{\alpha}_i$ is independent of $B_n^{(i)}(z)$ (since it is the missing row/column) and has zero-mean and variance $1/n$, so by conditioning ofn $\mathcal{F}_{i,n}$, the σ-field generated by W_n^i, $\mathbb{E}\bar{\varepsilon}_{i,n} = 0$. In addition, since

$$\frac{1}{n}\mathrm{Tr}(B_n^{(i)}(z)^2) \leq \frac{1}{Im(z)^2},$$

and the random variables $\sqrt{n}\alpha_i(k)$ are uniformly bounded, we have that $\mathbb{E}|\bar{\varepsilon}_{i,n}^{diag}|^4 \leq c_1 n^{-2}$ for some constant c_1 dependent on z and C, the bounding constant. Through a similar argument, we can see $\mathbb{E}|\bar{\varepsilon}_{i,n}^{off}|^4 \leq c_2 n^{-2}$ for some c_2 dependent again on z and C. From this, we see that $\sup_{i\leq n}|\varepsilon_{i,n}(z)| \to 0$ by Chebyshev's inequality and a union bound. □

References

Notes from: http://ipgold.epfl.ch/ leveque/matrix/.

Greg W Anderson, Alice Guionnet, and Ofer Zeitouni. *An introduction to random matrices*, volume 118. Cambridge university press, 2010.

Jinho Baik, Gerard Ben Arous, and Sandrine Peche. Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices. *The Annals of Probability*, 33(5):1643–1697, September 2005. ISSN 0091-1798, 2168-894X. doi: 10.1214/009117905000000233.

Laszlo Erdos, Antti Knowles, Horng-Tzer Yau, and Jun Yin. Spectral statistics of Erdos–Renyi graphs I: Local semicircle law. *The Annals of Probability*, 41(3B):2279–2375, May 2013. ISSN 0091-1798, 2168-894X. doi: 10.1214/11-AOP734.

Jeffrey S. Geronimo and Theodore P. Hill. Necessary and sufficient condition that the limit of Stieltjes transforms is a Stieltjes transform. *Journal of Approximation Theory*, 121(1):54–60, March 2003. ISSN 00219045. doi: 10.1016/S0021-9045(02)00042-4.

Noureddine El Karoui. Tracy–Widom limit for the largest eigenvalue of a large class of complex sample covariance matrices. *The Annals of Probability*, 35(2):663–714, March 2007. ISSN 0091-1798, 2168-894X. doi: 10.1214/009117906000000917.

Noureddine El Karoui. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, February 2010. ISSN 0090-5364, 2168-8966. doi: 10.1214/08-AOS648.

Antti Knowles and Jun Yin. Anisotropic local laws for random matrices. *Probability Theory and Related Fields*, 169(1):257–352, October 2017. ISSN 1432-2064. doi: 10.1007/s00440-016-0730-4.

Linyuan Lu and Xing Peng. Spectra of edge-independent random graphs. *arXiv:1204.6207 [math]*, April 2012.

Mark Rudelson and Roman Vershynin. Non-asymptotic theory of random matrices: extreme singular values. *arXiv:1003.2990 [math]*, March 2010.

Jack W. Silverstein. THE STIELTJES TRANSFORM AND ITS ROLE IN EIGENVALUE BEHAVIOR OF LARGE DIMENSIONAL RANDOM MATRICES. In *Lecture Notes Series, Institute for Mathematical Sciences, National University of Singapore*, volume 18, pages 1–25. WORLD SCIENTIFIC, July 2009. ISBN 978-981-4273-11-4 978-981-4273-12-1. doi: 10.1142/9789814273121_0001.

Ke Wang. RANDOM COVARIANCE MATRICES: UNIVERSALITY OF LOCAL STATISTICS OF EIGENVALUES UP TO THE EDGE. *Random Matrices: Theory and Applications*, 01(01):1150005, January 2012. ISSN 2010-3263, 2010-3271. doi: 10.1142/S2010326311500055.