Introduction to Probability Theory

Max Simchowitz

February 25, 2014

1 An Introduction to Probability Theory

1.1

In probability theory, we are given a set Ω of outcomes ω, and try to consider the probabilities of subsets $A \subset \Omega$, which we call events. To this end, we define a probability measure $Pr(\cdot)$ or $\mu(\cdot)$ (we will use each notation interchangably) that is a function from these sets $A \subset \Omega \to [0, 1]$. As a technical complicate, μ or Pr may only be defined on *some* of the subsets of Ω, but this will be adressed a bit later on. Intuitively, $Pr(A)$ is the probability that the event A will happen. It makes sense then to come up with the following rules about $Pr(\cdot)$.

Definition 1.1. Probability Axioms

1. First, we expect that some outcome will occur. Thus, we want $Pr(\Omega) = 1$.
2. Second, if an event A_1 is contained in another event A_2 - for example, the set of outcomes comprising the event of seeing two consecutives heads when flipping a coin is contained in the set of outcomes comprising the event that you see at least one heads - then, we expect that A_2 is at least as probable than A_1; that is $Pr(A_2) \geq Pr(A_1)$; this is called monotonicity.
3. Nonnegativity: no events can have a negative probability. Formally, $\mu(A) \geq 0$ for subset $A \subset \Omega$.
4. Next, we expect that the probability of something happening is just one minus the probability it doesn't happen; that is, $Pr(A) = 1 - Pr(A^c)$.
5. Finally, we expect that if we have pairwise disjoint events $A_1, \ldots, A_n$, then the probability of at least one of them happening is sum of the probabilities that each happen: that is $Pr(A_1 \cup \cdots \cup A_n) = Pr(A_1) + \cdots + Pr(A_n)$. This is known as finite additivity.
6. Unfortunately, finite additivity is a little weak for our purposes; often we will want to consider unions of sets with tend to the whole space, that is $\lim_{n\to\infty} \bigcup_{i=1}^{n} A_i = \Omega$, or intersections of sets which go to the emptyset. For this, we will need to strengthen finite additivity to $\sigma-$additivity, which says that if we have a countable collection $\{A_\alpha\}$ of pairwise disjoint events, then $Pr(\bigcup_\alpha A_\alpha) = \sum_\alpha Pr(A_\alpha)$.

Remark. Note that some of these axioms are in fact rendundant. Clearly 5 follows from 6. Moreover, from countable addivity, we have that $1 = Pr(\Omega) = Pr(A) + Pr(A^c)$, which yields axiom 3, and if $A \subset B$, then $B = A \cup (B/A)$, which gives $Pr(B) = Pr(A) + Pr(B/A) \geq Pr(A)$ by additivity and then nonnegativity.

Remark. If μ is a function from the subsets of Ω to $[0, \infty]$ that satisfies the third and sixth rules (from these, we get the second rule for free), then we call μ a measure. From rule three, we see that if $\mu(\Omega) = 0$, then $\mu = 0$. If $\mu(\Omega) = c$, then by the previous remark, $\frac{1}{c}\mu$ is a probability measure. If $\mu(\Omega) = \infty$, then there is no systematic way to make μ into a probability measure. However, if there are a countable collection of sets $\{A_\alpha\}$of finite measure for which $\Omega = \bigcup_\alpha A_\alpha$ and $\mu(A_\alpha) < \infty$, the μ is called $\sigma-$finite. For such measures, we can often "weight" these measures appropriately to recover a probability measure. More on this later.

You might ask: why do we define probabilities for events rather than outcomes? Consider the following scenario:

Example 1.1. Let Ω be the space of all infinite sequences of coin flips $(\omega_1, \omega_2, \dots)$ of a fair coin. Now, the event (or outcome) that we observe an sequence of heads or tails $\omega = (\omega_1, \omega_2, \dots)$ is contained in the event A^n_ω that we observed the same sequence first for the first n flips, but possibly other values in the following flips.

From elementary probability theory, we would like to define the probability of any sequence of n coin flips as 2^{-n}, and by monotonicity, $Pr(\omega) \leq Pr(A^n_\omega) = 2^{-n}$ for all n. Taking $n \to \infty$, we see that $Pr(\omega) = 0$. Thus, if we only knew the probabilities of the outcomes, we could never conclude that any *events* themselves had possitive probability. Moreover, the space of all tails of outcomes $\Omega^n = (\omega_n, \omega_{n+1}, \dots)$ is uncountable, since it has the cardinality of $2^{\mathbb{Z}}$. Thus, if we only knew probabilities of outcomes, we could never find the probability over the whole space using our additivity axiom, since we aren't allowed to take uncountable sums.

One technical problem that we may run into, which we will try to skip over in this exposition, is that it may be impossible to define a function $Pr(\cdot) : \Omega \to [0, 1]$ on all subsets of Ω. For our purposes, it suffices to define $Pr(\cdot)$ on a collections of subsets, known as a σ-algebra. We will denote our σ-algebra $\mathcal{F}$, and call sets $A \in \mathcal{F}$ *measurable*. Intuitively, $\mathcal{F}$ are the sets whose probabilities we can reasonably evaluate. A σ algebra has the following axioms:

Definition 1.2. Axioms for a σ-algebra

1. $\Omega \in \mathcal{F}$. This makes sense, since we know that $Pr(\Omega) = 1$.
2. If $A \in \mathcal{F}$, then $A^c \in \mathcal{F}$. This is also what we would expect, since $Pr(A^c) = 1 - Pr(A)$.
3. Given a countable collection of sets $\{A_\alpha\}$, $\bigcup_\alpha A_\alpha \in F$.

The following theorem gives us a convenient way to talk about σ-algebras.

Theorem 1.1. *Given a collection $\mathcal{E}$ of subsets $E \subset \Omega$, there is a unique, minimal σ-Algebra containing all sets $E \in \mathcal{E}$, denoted $\sigma(E)$. This is called the σ-algebra generated by $\mathcal{E}$*

When we do need a $\sigma-$Algebra on an uncountable space, it will generably the *Borel σ-Algebra*:

Definition 1.3. Borel σ-Algebra: Let Ω be a topological space (e.g. $\mathbb{R}^d$) and let $\mathcal{U}$ be the set of all open sets in Ω. The Borel σ algebra $\mathcal{B}(\Omega)$ is defined to be the smallest σ algebra containing $\mathcal{U}$, that is $\sigma(\mathcal{U})$

Remark. Most "reasonable sets" are Borel sets. For example, all open, closed, compact, and countable sets are Borel sets, including their complements. On the real line, half open intervals are also Borel sets, and in fact generate the Borel σ algebra (exercise).

On countable spaces, it makes sense to use the discrete σ-algebra:

Definition 1.4. Discrete σ Algebra The discrete σ algebra on a space Ω is the collection of all subsets of Ω.

Remark. Note that if Ω is countable, the axioms of a σ algebra tell use that the discrete σ algebra is precisely the σ algebra containing all elements of Ω. For you topologists out there, the discrete σ algebra is just the Borel σ on Ω where Ω is given the discrete topology.

We now arrive at the technical definition of probability space:

Definition 1.5. A probability space is a triple $(\Omega, \mathcal{F}, Pr)$, where Ω is an underlying space, $\mathcal{F}$ is a $\sigma-$algebra of events of Ω, and $Pr(\cdot)$ is a function from $\mathcal{F}$ to $[0,1]$, satisfying the axioms of a probability measure.

Probability measures also satisfy another property, called countable subbaditivity

Proposition 1.2. *Let $(\Omega, \mathcal{F}, Pr)$ be a probability space and $\{A_n\}$ be a collection of sets of $\mathcal{F}$. Then,*

$$Pr(\bigcup_n A_n) \leq \sum_n Pr(A_n) \tag{1}$$

Sketch. Define the sets $B_1 = A_1$, $B_n = A_n - \bigcup_{1 \leq k \leq n-1} A_k$. Now, note that the sets B_n are disjoint, but $\bigcup_{n \geq 1} A_n = \bigcup_{n \geq 1} B_n$. Apply countable additivity, and use monoticity to note that $Pr(A_n) \geq Pr(B_n)$. □

From this proposition follows a classic lemma.

Lemma 1.3. ***Borel Cantelli.*** *Let $(\Omega, \mathcal{F}, Pr)$ be a probability space. Let $A_1, A_2, \cdots \in \mathcal{F}$. If $\sum_{n=1}^{\infty} Pr(A_n) < \infty$, then with probability 1, only finitely many A_n will occur.*

Proof. The probability that A_n occur infinitely often is just $\bigcap_{N=1}^{\infty} \bigcup_{n=N}^{\infty} A_n$. By monotonicity, this probability is bounded abound by $Pr(\bigcup_{n=N}^{\infty} A_n)$ for all N, which, by countable subbaditivity, is in turn bounded above by $\sum_{n \geq N} Pr(A_n)$. Taking the infinum over N proves the lemma. □

Another general probabilistic concept is the notion of "Almost Sure" properties. In probability theory, we couldn't care less about events with zero probability. Thus, if $(\Omega, \mathcal{F}, Pr)$ is a probability space, we say an event $A \in \mathcal{F}$ occurs *almost surely* if $Pr(A) = 1$ and does not occur almost surely if $Pr(A) = 0$. We will frequently abbreviate almost surely with the letters *a.s.*. There are some other general properties of probability measures, and we list them below:

Proposition 1.4. ***Properties of Probability Measures***

1. *Set Differences: If $A \subset B$ then $Pr(B/A) = Pr(B) - Pr(A)$.*
2. *Continuity from Below: If $A_1 \subset A_2 \subset \ldots$, then $\lim_{i \to \infty} Pr(A_i) = Pr(\bigcup_i A_i)$*
3. *Continuity from Above: If $A_1 \supset A_2 \supset \ldots$, then $\lim_{i \to \infty} Pr(A_i) = Pr(\bigcap_i A_i)$.*

Example 1.2. Here are some common probability spaces

1. **Discrete Probability Spaces** Let T be a countable set, and let p be a function on T which is nonnegative and such that $\sum_{t\in T} p(t) = 1$. Then $(T, \mathcal{F}, p)$ is a probability space, where $\mathcal{F}$ is the discrete σ algebra. For example, the function $p(x) = \frac{x^n e^{-\lambda}}{x!}\mathbb{1}_{x\in\mathbb{N}}$ gives the probability measure for the Poisson Distribution.

2. **The Lebesgue Measure**. We can defined a probability measure on the interval $([0,1], \mathcal{B}([0,1])$ by taking the probability of any open subinterval (recall that this is an interval of the form (a,b), $[0,b)$, $(a,1]$, since we are in the subspace topology of $[0,1]$) to be its length. Note that the measure of a single point is zero, by continuity from above, and thus the measure of a countable number of points is also zero. For example, $\mathbb{Q}\cap[0,1]$, which is indeed dense in $[0,1]$, is of measure zero. Moreover, this observation implies that open, closed, and half open half closed intervals all have the same measure. Using the rules of a probability measure, we can extend this measure to all Borel Sets, since for each σ algebra operation, there is a corresponding operation for the probability measure.

   In probability, this measure is known as the Lebesgue measure. In real analysis, there is a σ algebra which strictly contains $\mathcal{B}([0,1])$ on which the Lebesgue measure can be well defined, but some of the additional sets in this large σ algebra often pathological and not worth our time. What we will borrow from real analysis is that we can define the Lebsgue measure on $\mathcal{B}(\mathbb{R})$ by taking the measure of an open interval to be its length, and extending in a similar fashion. We can even define the Lebesgue measure on $\mathcal{B}(\mathbb{R}^d)$ by considering the volumes of open rectangles. It is easy to see that the Lebesgue measure is σ finite.

3. **Probability Densities** Here is a classic example of how to reconstruct a probability measure from a measure. Let $f : \mathbb{R} \to \mathbb{R}$ be a nonnegative Riemann integrable function that $\int_{\mathbb{R}} f(x)dx = 1$. Then we can define a probability measure as in the Lebegue measure case by taking the measure of an open interval (a,b) to be $\int_a^b f(x)$, and extending similarly. In this case $f(x)$ is called a probability density. For example, you may have seen the case when $f(x) = \frac{1}{\sqrt{2\pi}}\exp(-x^2/2)$, which is called the Gaussian density. We can event extend this to function $f : \mathbb{R}^d \to \mathbb{R}_+$ for which $\int_{\mathbb{R}^d} f(x)dx = 1$. These are called multivariate densities.

1.2 Construction of the Integral

So far, we have defined probabilities as taking place on abstract spaces Ω, which have no clear mathematical properties (other than, perhaps, cardinality). To consider probabilities of numerical events, we define functions $X : \Omega \to \Omega'$, where Ω' is often a "numerical space" - generally a metric space - like $\mathbb{Z}$, $\mathbb{R}$, or $\mathbb{C}$. To make this rigorous, we first define random variables in the most general setting

Definition 1.6. Let $(\Omega_1, \mathcal{F}, Pr(\cdot))$ be a probability space, and let $(\Omega_2, \mathcal{G})$ be a set equipped with a $\sigma-$algebra. A random variable X is function from $\Omega_1 \to \Omega_2$ that preserves measurable sets in the following sense:

$$X^{-1}(A) = \{\omega : X(\omega) \in A\} \in \mathcal{F} \quad \forall A \in \mathcal{G} \tag{2}$$

If instead $(Omega_1, \mathcal{F}, \mu)$ is a general measure space, then we call X a measurable function.

A random variable is therefore a deterministic function of elements $\omega \in \Omega$, whose random-ness arises from the randomness of Ω. In English, 2 formula says that the probability of seeing certain numerical realizations of a random variable is just the probability of the event which leads to those possible realizations. Random variables also allow us to define a probability distribution on Ω_2 by "pulling back" to Ω_2

Definition 1.7. Let $X : (\Omega_1, \mathcal{F}, Pr) \to (\Omega_2, \mathcal{G})$ be a random variable. The pushforward measure induced by X is defined by

$$Pr_X(A) \triangleq Pr(X \in A) \triangleq Pr(X^{-1}(A)) \quad \forall A \in \mathcal{G} \tag{3}$$

Exercise 1.1. Prove that $(\Omega_2, \mathcal{G}, Pr_X)$ is a probabilty space.

Next, we need a notion of averaging, called the expectation. We know from calculus that the "average" of a some piece-wise continuous function $f(x)$ over some finite interval $[A, B]$ is just $\frac{1}{B-A}\int_A^B f(x)dx$. This example will help us generalize this notion of an average to a more general setting. In the following exposition, the integral or expectation we define can be generalized to arbitrary measure spaces, and we will do so without comment. Construction of the integral in this manner is called the Lebegues integral, and provides a general framework for defining a notion of integration on arbitrary measure spaces.

In the above example, we can think of dx as being an infinitesmal part of the probabilty measure on $[A, B]$ which says that the measure of a set $V \subset [A, B]$ is just $\frac{1}{B-A}\int_A^B \mathbb{1}_{x\in V}dx$, where $\mathbb{1}_{x\in V}$ is the function which takes the value of 1 if $x \in V$, and 0 otherwise. There is one key difference between Riemann integration and the general, Lebesgue integral, which we are about to introduce. Instead of partitioning the domain into "rectangles" (which may make little sense for an arbitrary probability space Ω), Lebesgue integration partitions on the range. To see this, let f be a Riemann integrable function on an interval - say $[0, 1]$, and let g_k be the function

$$g_k(x) = \sum_{j=4^{-k}}^{4^k} \frac{j}{2^k} I(f(x) \in [\frac{j}{2^k}, \frac{j+1}{2^k}]) \tag{4}$$

We can view f and g as random variables from $([0, 1], \mathcal{B}([0, 1]) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ where the domain is given the Lebesgue measure. Moreover, you can verify that $g_k(x) \to f(x)$ as $k \to \infty$. But rather than approximating f by partitioning the domain into little rectangles, (which only works for continuous functions), f is approximated by a finite combination of indicator functions, or a simple functions.this case, the intervals $[\frac{j}{2^k}, \frac{j+1}{2^k}]$ are all elements of $\mathcal{B}[0, 1]$, so we can write

$$\int g_k = \sum_{j=4^{-k}}^{4^k} \frac{j}{2^k} P(x : f(x) \in \frac{j}{2^k}, \frac{j+1}{2^k}]) \tag{5}$$

We can then define the integral of f as the limiting value of the g_k; in fact, we hope that $\lim \int g_k = \int \lim g_k = \int f$. However, there are couple precations to take. First, as in Riemann integration, if $f^- = \max(-f, 0)$ and $f^+ = \max(f, 0)$ may both integrate to infinity, so the integral of $f = f^+ - f^-$ may be undefined. Second, we must ensure that the integral is well defined, in the sense that if g_k and h_k are two sequence of simple functions approximating f, then $\lim \int g_k = \lim \int h_k$. The following definition of the integral ensures that it is air tight to these concerns

Definition 1.8. Let $X : (\Omega, Pr, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})$ be a random variable. We will interchangably refer to the expectation or the integral of X as $\mathbb{E}[X]$ or $\int X dP$. It is defined as follows:

1. Suppose X is a simple function, that is, we can write $X = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}$ for $a_i \in \mathbb{R}$ and $A_i \in \mathcal{F}$. Then $\mathbb{E}[X] = \sum_{i=1}^{n} a_i Pr(A_i)$.
2. Suppose that $X \geq 0$ almost surely. Then $\mathbb{E}[X] = \sup_{0 \leq Y \leq X \text{ a.s.}} \mathbb{E}[Y]$.
3. If X is real valued, we write $X^+ = \max(X, 0)$ and $X^- = \max(-X, 0)$. Note then that X^+ and X^- are also random variables $(\Omega, Pr, \mathcal{F}) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})$ (you can check that they satisfy the measurability condition). We define $\mathbb{E}[X] = \mathbb{E}[X^+] - \mathbb{E}[X^-]$. Note that that this quantity may be equal to $\infty - \infty$, and hence ill defined.

If X is real vector valued, we can defined $\mathbb{E}[X]$ as the vector whose entries are the expectations of the components of X. If X is complex-vector valued, identify X with a real valued random variable of twice its dimension. $\int X d\mu$, integral of a measurable function X on a general measure space $(\Omega, \mu, \mathcal{F})$, is defined analogously. /

The following two theorems give conditions under which taking limits commutes with taking integrals. In the following, two theorems, let X, $\{X_n\}_{n \geq 0}$ and Y be random variables from $(\Omega, \mathcal{F}, Pr)$ and image space $(\mathbb{R}, \mathcal{B}(\mathbb{R})$. Then

Theorem 1.5. ***Monotone Convergence Theorem*** *Suppose that $0 \leq X_n \leq X_{n+1}$ almost surely, and suppose that $X_n \to X$ almost surely (this means that the set of $P(\omega \in \Omega : \lim_{n \to \infty} X_n(\omega) \neq X(\omega)) = 0$). Then $\lim_{n \to \infty} \mathbb{E}[X_n] = \mathbb{E}[X]$. The analogous statement is true for integrals on general measure spaces.*

Remark. An application of the monotone convergence theorem shows that the expectation of the functions g_k in Equation 4 do in fact convergence to expectation of their limiting function f. In fact, this theorem alleviates us of the burden of showing that any approximating sequence of simple random variables $g_k \leq f$ *actually attains* the supremum over simple functions less than f. It suffices to do our computation for one monotonically increasing sequence, and this will suffice. Another interesting fact is that the monotone convergence thereom implies countable additivity axiom for probability measures. Indeed, let $\{A_k\}_{k \geq 1}$ be a sequence of disjoint sets and let $A = \bigcup_{k \geq 1} A_k$. If we define $X_n = \sum_{k \leq n} \mathbb{1}_{A_k}$, then $\lim_{n \to \infty} X_n = \mathbb{1}_A$. By finite additivity, $\mathbb{E}[X_n] = \sum_{1 \leq k \leq n} Pr(A_k)$, and by monotone convergence, it follows that $\lim_{n \to \infty} \mathbb{E}[X_n] = \lim_{n \to \infty} \sum_{1 \leq k \leq n} Pr(A_k) = \mathbb{E}[X] = Pr(A)$.

Theorem 1.6. ***Dominated Convergence Theorem*** *Suppose that $X_n \to X$ and $X_n \leq Y$ almost surely, where $\mathbb{E}[Y] < \infty$. Then $\lim_{n \to \infty} \mathbb{E}[|X_n - X|] = 0$, and in particular, $\lim_{n \to \infty} \mathbb{E}[X_n] = \mathbb{E}[X]$. The analogous statement is true for integrals on general measure spaces.*

Remark. There are cases in which lim

We also have the following basic facts about expectations:

Proposition 1.7. *The expectation satisfies the following properties:*

1. *Linearity: $\mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$ for $a, b \in \mathbb{R}$ and measurable X and Y (assuming the quantity on the right is well defined)*

2. *Monotonicity: If $X \leq Y$ almost surely, than $\mathbb{E}[X] < \mathbb{E}[Y]$*

3. *If $X = 0$ almost surely, than $\mathbb{E}[X] = 0$. Equivalenty, if $X = Y$ almost surely, then $\mathbb{E}[X] = \mathbb{E}[Y]$*

Each statement is also true for integrals on general measure spaces.

Remark. One of the cooler applications of the dominated convergence and linearity is proving a generalization of Liebniz's rule to expecations. That is, X be a random variable, and let $f(x,t)$ be function a determinsitic function of x and t such that with derivative $f'(x,t) = \frac{d}{dt}F(x,t)$ is integrable. Write $F(t) = \mathbb{E}[f(X,t)]$. Then

$$\begin{aligned} F'(t) = \frac{d}{dt}F(t) &= \lim_{u\to 0} \frac{1}{u}\left(\mathbb{E}[f(X,t+u)] - \mathbb{E}[f(X,t)]\right) \\ &= \lim_{u\to 0}\left(\mathbb{E}[\frac{1}{u}\{f(X,t+u) - f(X,t)\}]\right) \quad \text{by linearity} \\ &= \lim_{u\to 0}\left(\mathbb{E}[f'(X,t)]\right) \quad \text{by dominated convergnece} \end{aligned} \tag{6}$$

This comes in handy when understanding characteristic functions and the moment matching method.

One last fact about random variables that we shall need is the notion of independence.

Definition 1.9. Two random variables $X, Y : (\Omega, \mathcal{F}, Pr) \to (\Omega_2, \mathcal{G})$ are said to be indepdendent if, for all $A, B \in \mathcal{G}$, $Pr(X \in A, Y \in B) = Pr(\{\omega : X(\omega) \in A\} \cap \{\omega : Y(\omega) \in B\}) = Pr(X \in A) \cdot Pr(Y \in B)$.

Remark. **Integration in Discrete and Lebesgues Spaces** In a discrete measure space on a countable set T, the integral of a function $f : T \to \mathbb{R}$ is reduces to to be $\int f(t) = \sum_{t\in T: f(t)\geq 0} f(t) - \sum_{t:f(t)\leq 0} f(t)$, whenever the quantity is defined. In $(\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$, you can show that the Lebesgue integral of a Riemann integral function f is equal to its Riemann integral. Thus, we often write the Lebesgue integral of f simply as $\int_{\mathbb{R}^d} f dx$. This makes sense, since all Riemann integral fucntions are Lebesgue integrable, but the converse is false. For example, a Riemann integrable function must be continuous Lebesgue-almost everywhere (that is, outside a set with Lebesgue measure zero). However, the indicator function I_q of the rationals is discontinuous everywhere. Nevertheless, the rationals are countable, and thus have Lebesgue measure zero, so by 1.7, the Lebesgue integral is well defined and in fact zero.

Example 1.3. Using Expectations to define Probability Measures In a previous example, we saw that we could use a nonnegative Riemann integrable function f that integrates to 1 in order to define a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ by taking $Pr((a,b)) = \int_a^b f(x)dx$. From the previous remark, this can be generalized to any Lebesgue measurable function that satifies the nonnegativity and normalization conditions. In fact, this can be generalized further still to an arbitrary measurable nonnegative function integrating to 1 on an arbitrary measure space. In particular, if $(\Omega, Pr, \mathcal{F})$ is a probability space and X is an almost surely nonnegative real valued random variable such that $\mathbb{E}[X] = 1$, then we can define a probability measure Pr' on $(\Omega, \mathcal{F})$ by setting $Pr'(A) = \mathbb{E}[\mathbb{1}_A X]$ for all $A \in \mathcal{F}$. Similarly, we can define an arbitrary measure μ' as $\mu(A) = \int \mathbb{1}_A d\mu$. Measures μ' defined in this way are said to be absolutely continuous with respect to μ, in the sense that $\mu'(A) = 0$ if $\mu(A) = 0$. In this case write $\mu' << \mu$. The Radon Nikodyn theorem shows that all absolutely continuous measures arise in this way.

Remark. You will often hear the operations $\mathbb{E}[\cdot]$ or $\int \cdot d\mu$ referred to as linear functionals. This is because they can be thought as linear transformation from the space of measurable functions to the reals. Expectations can be taken with respect to different measures, by weighting the measure by another measurable function, as seen in the previous example. If we fix a measure μ and consider integrating with respect to all the measures μ' absolutely continuous with respect to μ, we get a very large family of linear transformations on the space of measure functions. These functionals are said to inhabit the dual space to linear functions, as motivated by the following example: In a measure space with finite cardinality n and the discrete sigma algebra, the measurable functions are just real vectors in $v \in \mathbb{R}^n = V$, and all possible integrals are given by $\langle w, v \rangle$ for $w \in V^*$.

1.3 Notions of Convergence

Discussing the convergence of a random variable X is a bit more delicate than considering the convergence sequence of finite, or even countably infinite, numbers. The first is that X is really a function on a potentially uncountable space. The second is that, unlike in sequences, we want converge to capture *probabilities* properties of X.

Example 1.4. Let $\Omega = \{\omega_1, \ldots, \omega_6\}$. We can identify random variables on this space with all length-6 sequences, such that $X_i = X(\omega_i)$. However, this is probabilistically misleeding. Indeed, let $X(\omega_i) = 1 \ \forall 1 \leq i \leq 6$, and let

$$\widetilde{X}(\omega) = \begin{cases} 1 & \omega = \omega_i \quad \forall 1 \leq i \leq 5 \\ 0 & \omega = \omega_6 \end{cases} \tag{7}$$

Now set $Pr(X = \omega_1) = \cdots = Pr(\omega_5) = 1/5$, while $Pr(\omega_6) = 0$. Now, $\widetilde{X}$ and X correspond to two different sequences. But the probability that these random variables differ, that is $Pr(\widetilde{X} \neq X)$, is just $Pr(\omega_6) = 0$; that is from, a probabilistic perspective, these random variables cannot be told apart.

When it comes to convergence, we want to preserve the same intuition. That is, since we what $\widetilde{X}$ and X to be, for all intents and purposes equal, we also want an sequence of random variables X_n that converges to $\widetilde{X}$ to also converge to X. A strong type of convergence that has this property is called almost sure convergence:

Definition 1.10. Almost Sure Convergence A sequence of random variables $(X_n)_{n\geq 0}$ is said to converge to a random variable almost surely X if $Pr(A) = 0$, where $A \triangleq \{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) \neq X(\omega)\}$. Here, we right $X_n \overset{a.s.}{\to} X_n$.

Almost sure convergence is *almost* pointwise; it lets us neglect a couple points here and there that have zero probability. Almost sure convergence is almost purely probabilistic: we don't even need X_n to take numerical values for it to makes sense, just have some notion of convergence in the image space (for example, X can map to a topological space). If we suppose X_n takes values inside some normed spaces, we can formulate a weaker notion of convergence:

Definition 1.11. Convergence in Probability A sequence of random variables $(X_n)_{n\geq 0}$ is said to converge to a random variable X in probability if, for any $\epsilon > 0$, $\lim_{n\to\infty} Pr(\|X - X_n\| > \epsilon) = 0$

Remark. Almost sure covergence is just convergence in probability with $\epsilon = 0$. We see immediately that almost sure convergence is stronger than convergence in probability, and clever counterexamples show that it is in general a strictly stronger condition.

Almost sure convergence and convergence in probability let us formulate the Strong and Weak Laws of Large Numbers, two theorems which capture the probabilistic intuition that the average of random variables corresponds to its expectation:

Theorem 1.8. ***Law of Large Numbers*** *Let $(X_n)_{n\geq 1} : (\Omega, \mathcal{F}, Pr) \to (\mathbb{R}, \mathcal{B}(\mathbb{R})$ be a sequence of independent random variables that are identically distributed: that is, for all $A \in \mathcal{B}(\mathbb{R})$, $Pr(X_n \in A) = Pr(X_m \in A)$ for all $m, n \geq 0$. Define $\overline{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_n$, and write $\mathbb{E}[X]$ for $\mathbb{E}[X_n]$ for all n (note the sequence has identical expecation). Then*

1. *$\overline{X}_n \to \mathbb{E}[X]$ in probability. This is the weak law of large numbers*
2. *$\mathbb{X}_n \to \mathbb{E}[X]$ almost surely. This is the strong law of large numbers.*

Unfortunately, almost sure convergence and convergence in probability are not very "numerical" as can be seen from the following example:

Example 1.5. Let X_n take the value of 0 with probability $1 - 2^{-n}$ and take the value of 4^n with probability 2^{-n}. At once, we see that X_n convergences to X both almost surely and, consequently, in probability. However, $\mathbb{E}[X_n] = 4^n(2^{-n})$ grows to be much larger than 0. This is an example where the "decay at infinity" is too slow, or even gets worse over time.

Convergence in a norm L_p, for $p \in [1, \infty)$ ensures that what we observed in Example 1.5 can not happen

Definition 1.12. Convergence in L^p: A sequence of random variables $(X_n)_{n\geq 0}$ taking values in a space with norm $\|\cdot\|$ is said to converge to a random variable X in L^P for some $p \in [1, \infty)$ if $\lim_{n\to\infty} \mathbb{E}[\|X - X_n\|^p] = 0$. We write $X_n \overset{L_p}{\to} X$.

We have seen that convergence almost everywhere does not imply convergence in L^p, and the conversely convergence in L^p does not imply convergence almost everywhere. However, the following proposition shows that convergence in L^p does imply convergence in probability:

Proposition 1.9. ***Markov's Inequality*** *Let X be any nonnegative random variable and let $t > 0$. Then*

$$Pr(X > t) \leq \frac{\mathbb{E}[X]}{t} \tag{8}$$

Proof.

$$Pr(X > t) \quad = \quad \mathbb{E}[\mathbb{1}_{X>t}] \leq \mathbb{E}[\frac{X}{t}\mathbb{1}_{X>t}] \leq \mathbb{E}[\frac{X}{t}] = \frac{\mathbb{E}[X]}{t}$$

□

Both the semiricle law and the Marcheno Pastur law are results about the weakest form of convergence, convergence in distribution. It is straightforward to show that convergence in probability (and thus the two stronger modes of convergence) imply this type of convergence. Before we continue, we must first define a distribution function:

Definition 1.13. Distribution Function Let X be a real valued random variable. Its distribution function $F_X(t) : \mathbb{R} \to [0,1]$ is defined as

$$F_X(t) \triangleq Pr(X \leq t) \tag{9}$$

If X is a random variable taking values in $\mathbb{R}^d$, then $F_X(t) : \mathbb{R}^d \to [0,1]$ is defined as

$$F_X(t) \triangleq Pr(\bigcup_{1 \leq i \leq n} \{X_i \leq t_i\}) \tag{10}$$

If X takes values in $\mathbb{C}^d$, we define the distribution function by identifying $\mathbb{C}^d \simeq \mathbb{R}^{2d}$. When X has a distribution function F, we write $X \sim F$.

The following theorem motivates the study of distribution functions:

Theorem 1.10. *Let X be a Borel measurable random variable taking values in a $\mathbb{R}^d$. Then $F_X(t)$ uniquely determines $Pr(X \in A)$ for all $A \in \mathcal{B}(\mathbb{R}^d)$.*

Definition 1.14. A sequence of real valued random variables $(X_n)_{n \geq 1}$ is said to converge in distribution to a random variable X if $F_{X_n}(t) \to F_X(t)$ at all points for which $F_X(t)$ is continuous.

1.4 Characteristic Functions and Moment Matching

It turns out that there is another function which completely characterizes the distribution of a random variable:

Definition 1.15. Characteristic Function For a real valued random variable X, its characteristic function $\varphi_X(t) : \mathbb{R} \to \mathbb{C}$ is defined by

$$\varphi_X(t) = \mathbb{E}[\exp(itX)] = \mathbb{E}[\cos(X)] + i\mathbb{E}[\sin X] \tag{11}$$

For a real vector valued random variable X taking values in $\mathbb{R}^d$, its characteristic function $\varphi_X(t) : \mathbb{R}^D \to \mathbb{C}$ is defined by

$$\varphi_X(t) = \mathbb{E}[\exp(i\langle t, X\rangle)] = \mathbb{E}[\cos(\langle t, X\rangle)] + i\mathbb{E}[\sin(\langle t, X\rangle)] \tag{12}$$

The characteristic function can be thought of as the Fourier Transform of the random variable. In particular, if X is a real valued random variable, and if $Pr(|X| > N)$ decays suitably fast then we can switch derivatives with expectations and compute

$$\frac{d^n}{d^n t}\psi_X(t) = \frac{d}{dt}\mathbb{E}[\exp(itX)] = i^n\mathbb{E}[X^n \exp(itX)] \tag{13}$$

evaluating the expression at $t = 0$ shows that

$$\mathbb{E}[X^n] = (-i)^n \frac{d^n}{d^n t}\psi_X(t)\big|_{t=0} \tag{14}$$

One sufficient condition that lets us switch expectations and and taking n derivatives is that $E[X^{2n}]$ is finite. This fact will come in handy for developing the Moment-Matching method for convergence in distribution. First, we cite nice result from Probability Theory which says which characterizes the relationship betwen distribution functions and characteristic functions:

Theorem 1.11. *There exists a bijection between characteristic functions of real-vector-valued random variables and distribution functions, with explicit formulae to derive one from the other. Moreover, for random variables $X : \Omega \to \mathbb{R}^d$, the following are equivalent:*

1. $X_n \to X$ *in distribution*
2. $\varphi_{X_n}(t) \to \varphi_X(t)$ *for all* $t \in \mathbb{R}^d$
3. $\mathbb{E}[f(X_n)] \to \mathbb{E}[f(X)]$ *for any bounded, continuous function* f

1.11 gives us a new way to compute convergence in distribution. Now, consider the power series of X

$$p_X(t) \triangleq \sum_{k \geq 0} t^k \mathbb{E}[X^k] \tag{15}$$

If $p_X(t)$ converges absolutely in some range $(-\epsilon, \epsilon)$, then we have that

$$\varphi_X(z) = p_X(iz) \forall z \in \mathbb{C} : \|z\| < \epsilon \tag{16}$$

and, by analytic continuation, the two functions agree on their entire domain. Thus, if the moments of a random variable form an absolutely convergent power series, then $\psi_t(X)$ is uniquely determined by the moments of X. This proves the following theorem:

Theorem 1.12. ***Moment Matching Method*** *Let X be a real valued random variable with moments $M_0, M_1, \ldots,$ and define $p_t(x) = \sum_{k \geq 0} t^k M_k$. If $p_t(x)$ defines an absolutely convergent power series in some small interval contaning the origin, then for any sequence of random variables (X_n), the following are equivalent*

1. $X_n \xrightarrow{D} X$
2. $\lim_{n \to \infty} \mathbb{E}[X_n^k] = M_k$ *for all* $k \geq 0$.

Remark. From complex analysis, we know that the power series $p_t(x) = \sum_{k \geq 0} t^k M_k$ converges in some neighborhood about the origin as long as $\limsup_j (M_j)^{1/j}$ is finite. This is one way to check that the conditions for the moment matching method.

Remark. Here is an intuitive way to think about characteristic functions in terms of the Fourier transform of a compactly supported function. We saw about that if X is a random variables such that $\mathbb{E}[X^k] < \infty$ for all k and $\limsup_{k \to \infty} (\mathbb{E}[X^k]^{1/k} < \infty$, then we can think of X as being almost compactly supported, in the sense that its momenets are small enough that we can neglect the values X takes outside a large enough region of length C. From Fourier theory, we therefore expect X to be determined by its countable sequence of Fourier modes, and this intuition motivates the moment matching. It should not be strange that a function taking uncountably many values is specifying by a countable set; indeed, an continuous function on the reals is uniquely determined by its values on the rationals.

2 Conditional Expectation and Martingales

To prove the Marchenko Pastur Law, we will rely on the notion of conditional expectations. Intuitively, conditioning a random variable $X : (\mathcal{F}, \Omega \to (\mathcal{F}_1, \Omega_1)$ on another random variable Y removes all the "randomness" in X that cannot be attributed to Y. Here are two concrete examples of conditional expectation:

Example 2.1. Discrete Distributions Let $\Omega = T_1 \times T_2$ where T_1 and T_2 are countable sets. If endow Ω with the discrete sigma-algebra, then there is a function $p : T_1 \times T_2 \to [0, 1]$ such that, for any subset $A \subset \Omega$,

$$Pr(A) = \sum_{(t_1,t_2)\in A} p(t_1, t_2) \tag{17}$$

Now, consider a random variable $X : \Omega \to \mathbb{R}$ which is Borel measurable. Then

$$\mathbb{E}[X] = \sum_{(t_1,t_2)\in T_1\times T_2} X(t_1, t_2)Pr(t_1, t_2) \tag{18}$$

assuming the sum converges.

Let Y be a random variable such that, for any $t_2, t_2' \in T_2$, $Y(t_1, t_2) = Y(t_1, t_2')$, but for any $t_1 \neq t_1'$, $Y(t_1, t_2) = Y(t_1', t_2)$. This means that if we know the value of Y at some point $x \in \mathbb{R}$, we know the first coordinate of $Y^{-1}(x)$. It is easy to check then that the smallest σ algebra $\mathcal{F}_Y$ generated by Y consists of all sets $\{t_1\} \times T_2$: this is just measure-theoretic jargon for the statement that Y is entirely determined by the sets $\{t_1\} \times T_2$, and each set in the collection is necessary to specify the behavior of Y. Thus, if some magic oracle were to tell you everything you possibly could about a realization x of $Y(\omega)$, you still could not pinpoint the value of $X(\omega)$, as the second coordinate of $Y^{-1}(x)$ in unspecified.

The conditional expectation of X given Y is then obtained by averaging over these potential values of X. More formaly,

$$\mathbb{E}[X|Y = A] = \begin{cases} \frac{1}{Pr(Y^{-1}(A))}\mathbb{E}[X\mathbb{1}_A] & Pr(Y = A) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{19}$$

Setting $\mathbb{E}[X|Y = A] = 0$ if $Pr(A) = 0$ is just a convention, since in probability, we negelect measure zero events.

In particular, suppose $Y^{-1}(x) = \{t_x\} \times T_2$

$$\mathbb{E}[X|Y = x] = \begin{cases} \frac{1}{\sum_{t_2\in T} Pr(t_x,t_2)} \sum_{t_2\in T_2} X(t_x, t_2)Pr(t_x, t_2) & Pr(Y = A) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{20}$$

Example 2.2. Let X and Y be continuous, real valued random variables with a joint density $p(x, y)$ on $\mathbb{R}^2$. That is, for some subset $U \subset \mathbb{R}^2$, $Pr((X, Y) \in U) = \int_U p(x, y)dxdy$. Then

$$\mathbb{E}[X|Y = y] = \int_{\mathbb{R}} xp(x, y)dx \tag{21}$$

The previous examples motivates the following defintion:

Definition 2.1. Let X be a random variable $(\Omega, \mathcal{F}, Pr) \to (\mathbb{R}^d, \mathcal{B}(\mathbb{R}^d))$. If $\mathcal{H}$ is σ-subalgebra of $\mathcal{F}$, then $\mathbb{E}[X|\mathcal{H}]$ is an $\mathcal{H}$ measurable random variable such that for any $A \in \mathcal{H}$,

$$\mathbb{E}[X\mathbb{1}_A] = \mathbb{E}[\mathbb{E}[X|H]\mathbb{1}_A] \tag{22}$$

If $Y : \Omega \to \mathbb{R}^d$ is another random variable such that $\sigma(Y) \subset \mathcal{F}$, we define $E[X|Y] = E[X|\sigma(Y)]$. A standard theorem in probability theory guarantees that $\mathbb{E}[X|\mathcal{H}]$ exists, and is unique up to a set of measure zero.

Remark. Note that the above definition is *descriptive*, not *constructive*. The usual way to find a conditional expectation is to make a reasonable guess, and then verify that it satisfies Definition 2.1. As in the above example, the conditional expectation is pretty apparent for most elementary applications, and, in the case of continuous and discrete distributions, it reduces to the definitions of conditional expectation you might have seen in non-measure theoretic statistics and probability courses.

Conditional expectations will come in handy in proving the Marchenko pastur because it will let us "build up" an $n \times n$ random sample covariance S from its (random) $k \times k$ submatrices, and treat the remaining indices as deterministic (that is, setting them to their expectations). For any $n \times n$ square matrix S, Let $S_{(k)}$ be the $k \times k$ upper left square subsmatrix of S, which itself is a random variable. Since an observation of $S_{(j)}$ determines an observation of $S_{(k)}, k \leq j$, we see that

$$\sigma(S_{(1)}) \subset \sigma(S_{(2)}) \cdots \subset \sigma(S_{(n)}) \tag{23}$$

This is an example of *filtration*, an ordered set of nondecreasing σ-Algebras $\{\mathcal{F}_t\}$. When we take a random variable X with finite expectation and create a collection of random variables X_t by conditioning on a filtration: $X_t = \mathbb{E}[X|\mathcal{F}_t]$, we get a *martingale*; a random variable such that $\mathbb{E}[X_t|X_s] = X_s$ for any $s \leq t$. We often refer to X_t as X *at time* t. We will also write $\mathbb{E}_t[X] = \mathbb{E}[X|\mathcal{F}_t]$.

Martingales are used very frequently in probability theory because knowledge of a martingale at times t constrains the behavior of the martinagle at time $s \geq t$. In fact, if are given a sequence of martingales $X_1, X_2, \ldots$ such that $X_j - X_{j+1}$ is bounded, then we get the following useful inequality:

Lemma 2.1. *Let $\{X_k\}$ be a complex martingale sequence with respect to a filtration $\{\mathcal{F}_k\}$, and let $Y_k = X_k - X_{k-1}$. Then for any $p > 1$, there exists a constant K_p depending only on p such that*

$$\mathbb{E}\left|\sum_{Y_k}\right|^p \leq K_p \mathbb{E}\left(\sum |Y_k|^2\right)^{p/2} \tag{24}$$

And for $p \geq 2$, there exists a constant C_p such that

$$\mathbb{E}\left|\sum_{Y_k}\right|^p \leq C_p\left(\sum \mathbb{E}_{k-1}|X_k|^2 + \mathbb{E}\sum |X_k|^p\right) \tag{25}$$

Remark. The key step in proving many of these matingale inequalities is a form of Jensen's inequality for conditional expectations. Recall the classical formulation of Jensen's inequality: Let $f : \mathcal{X} \to \mathbb{R}$ be a convex function, that is, one for which $f(\theta x + (1-\theta)y) \leq \theta f(x) + (1-\theta)f(y)$

for $\theta \in [0,1]$ and all $x, y \in \mathcal{X}$. Finite induction shows that for any $x_1, \ldots, x_K \in \mathcal{X}$ and a vector $(\theta_1, \ldots, \theta_K)$ on the $K-1$ simplex - i.e. $\sum_{i=1}^K \theta_i = 1$, and $\theta_i \geq 0$ - then $f(\sum_{i=1}^K \theta_i x_i) \leq \sum_i \theta_i f(x_i)$. If we take the limit where $K \to \infty$ and apply the definition of integration, the equality generalizes further to the statement $f(\mathbb{E}[X]) \leq \mathbb{E}[f(X)]$. It can be shown, in fact, that Jensen's inequality generalizes further still to conditional expectations; that is for a convex (measurable) function f,

$$f(\mathbb{E}[X|Y]) \leq \mathbb{E}[f(X)|Y] \tag{26}$$

Since the entire space Ω is always $Y-$measurable and $f(X)$ is a random variable, $\mathbb{E}[f(X)] = \mathbb{E}[\mathbb{E}[f(X)|Y]]$. Thus, Jensen's inequality gives

$$\mathbb{E}[f(\mathbb{E}[X|Y])] \leq \mathbb{E}[f(X)] \tag{27}$$

We can make this more concrete. For $p \geq 1$, the function $x \mapsto (x-a)^p$ is convex (note that the second derivative is nonnegative, which is a sufficient condition for convexity). In particular if we set $a = \mathbb{E}[X] = \mathbb{E}[\mathbb{E}[X]|Y]$ and $p = 2$, we get

$$\mathrm{Var}[\mathbb{E}[X|Y]] \leq \mathrm{Var}[X] \tag{28}$$

That is, conditional expectation does not increase the variance. In many cases, conditioning strictly decreases the variance, as one would intuitively expect. For a fun (but entirely optional) exercise, try to find necessary and sufficient conditions under which conditioning one discrete random variable on another discrete random variable results in a strict decreases in variance.

3 A Detour into Matrices

Here we develop some matrix identities which we shall need for subsequent calculations. Given a Hermitian Matrix A, the goal is to described difference between the trace A^{-1} and the trace of A_k^{-1}, the inverse of A with its k^{th} row and column removed. Though the computations will follow from elementary linear algebra, they are not entirely obvious and are hard to see without a couple well known matrix decompositions.

Let $A = (a_{ij})$ and write A_{ij} for the cofactor of a_{ij} (that is, the determinant of the matrix A with row i and column j deleted, multiplied by $(-1)^{i+j}$). We define the adjugate matrix of A by $A^a = (A_{ij})^T$. We begin with the following theorem

Theorem 3.1. ***Cramer's Rule*** *Let A be a $n \times n$ matrix. Then*

$$A^{-1} = \frac{1}{\det(A)} A^a \tag{29}$$

$$\tag{30}$$

Proof. Recall the identity

$$\det(A) = \sum_{i=1}^n a_{ij} A_{ij} \tag{31}$$

From the formula above, we see that

$$(AA^a)_{ii} = \det(A)$$

while, for $i \neq j$.

$$(AA^a)_{ij} = \sum_{j=1}^{n} a_{ik} A_{jk}$$

which is the determinant of a matrix in which one of the rows appear twice, and hence is zero. It follows that

$$(AA^a)_{ii} = \det(A) I_n \tag{32}$$

Inverting proves the theorem. □

Another useful matrix computation which will come in Handy is Hua's holding method:

$$\begin{bmatrix} I & 0 \\ -CA^{-1} & I \end{bmatrix} \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \begin{bmatrix} A & B \\ 0 & D - CA^{-1}B \end{bmatrix} \tag{33}$$

By multiplicativity of the determinant, we have the following result:

Theorem 3.2. *If A is a square nonsingular matrix, then*

$$\det \begin{bmatrix} A & B \\ C & D \end{bmatrix} = \det(A) \det(D - CA^{-1}B) \tag{34}$$

As we shall see, the trace becomes very important in random matrix theory. The following gives us a nice formula for the trace. For an $n \times n$ makes A, let A_k be the matrix resulting from deletering the $k - th$ row and column from A the .. yield the following theorem

Theorem 3.3. *Suppose both A and A_k are nonsingular for all $k = 1, \ldots, n$. Write $A^{-1} = [a^{kl}]$, and define a_{kk} to be the K^{th} diagonal entry of A, α_k' the row vector obtained by deleting a_{kk} from the k^{th} row of A, and β_k the column vector obtaining by deltering a_{kk} from the k^{th} column of A. Then the following hold:*

$$a^{kk} = \frac{1}{a_k k - \alpha_k' A_k^{-1} \beta_k} \tag{35}$$

and consequently

$$Tr(A) = \sum_{k=1}^{n} \frac{1}{a_k k - \alpha_k' A_k^{-1} \beta_k} \tag{36}$$

Lemma 3.4. *Let Σ be a positive definite matrix partitioned as $\begin{bmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{bmatrix}$ Define $\Sigma_{22.1} = \Sigma_{22} - \Sigma_{21}\Sigma_{11}\Sigma_{12}$. Then*

$$\Sigma^{-1} = \begin{bmatrix} \Sigma_{11}^{-1} + \Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22.1}^{-1}\Sigma_{21}\Sigma_{11} & -\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22.1}^{-1} \\ -\Sigma_{22.1}^{-1}\Sigma_{21}\Sigma_{11}^{-1} & \Sigma_{22.1}^{-1} \end{bmatrix} \tag{37}$$

With this identity, we can proove the following theorem:

Theorem 3.5. *Let A be a Hermitian positive definite matrix, and again let A_k be the matrix obtained by deleting the k^th row and the k^{th} column (which we assume to be nonsingular). Then*

$$tr(A^{-1}) - tr(A_k^{-1}) = \frac{1 + \alpha_k A_k^{-2} \alpha_k}{a_k k - a_{kk} - -\alpha_k * A_k^{-1} \alpha_k} \tag{38}$$

Proof. By permuting the rows and columns, we may assume that $k = n$. Partition A as in the previous lemma with $\Sigma_{11} = A_n$, $\Sigma 12 = \alpha_n = \Sigma_{21}^*$. and $\Sigma_{22} = a_k k$. We see at once that

$$\mathrm{tr}(A^{-1}) = \mathrm{tr}(\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22.1}^{-1}\Sigma_{21}\Sigma_{11}^{-1}) + \mathrm{tr}(\Sigma_{22.1}^{-1}) \tag{39}$$

We first note that

$$\Sigma_{22.1} = a_{nn} - \alpha_n^* A_n \alpha_n \triangleq c_n \tag{40}$$

So by pulling out a constant term, we have

$$\Sigma_{11}^{-1}\Sigma_{12}\Sigma_{22.1}^{-1}\Sigma_{21}\Sigma_{11}^{-1} = \frac{1}{c_n}(\alpha_n(A_n)^{-1})^*(\alpha_n(A_n)^{-1}) \tag{41}$$

where we use the fact that for a hermitian matrix, $(A^{-1})^* = A^{-1}$. Now, the trace is invariant to cyclic permutations, so in fact

$$\mathrm{tr}(\alpha_n(A_n)^{-1})^*(\alpha_n(A_n)^{-1}) = \mathrm{tr}(\alpha_n'(A_n)^{-2}\alpha_n) \tag{42}$$

Plugging back into Equation 39 concludes the proof. □

4 Key concepts in random matrix theory

4.1 The Empiricial Spectral Distribution

In general, if one observes a finite set of values $\lambda_1, \ldots, \lambda_n \subset \Lambda$, then we can define a probability measure over Λ by taking

$$P(S) = \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}_{x_i \in S}$$

for all $S \subset \Lambda$. This is known as an empirical measure, since it is based entirely on the observed values and the assumption that the underlying probability distribution over Λ agrees with ones observations. If $\Lambda = \mathbb{R}$, then the empirical measure gives rise to an empirical distribution, by taking

$$F(x) = \frac{1}{n}\#\{i \leq n : \lambda_i \leq x\} \tag{43}$$

In particular, if A is a self adjoint matrix, it has real eigenvalues, so we can take $\{\lambda_i\}$ to be the eigenvalues of A and define the empirical spectral distribution function or ESD of A, which we denote $F^A(x)$. In the case where A is arbitrary and can have imaginary eigenvectors, we can define a two dimensional distribution function over its eignvectors. This won't be important since the Semicircular and Marchenko Pastur laws only concern matrix with real spectra.

A crucial question in random matrix theory is: under what conditions to the ESDs F^{A_n} of a sequence of matrix A_n converge, in the sense of distributions, to a limiting spectral distribution F? And, can we find an explcity formula for F?. Note that F is continuous, convergence in distribution simply means pointwise convergence, and this task seems rather innocuous. Unfortunately, there are two problems. First, the set of eigenvalues of a matrix is a very complicated function of the matrix itself. Indeed, in dimensions greater than five, the eigenvalues are solutions of polynomials of degree for five or greater, and hence no closed form exists. The second, and perhaps more subtle point, is that if A_n are random matrices, than F^{A_n} is itself a random distribution. "Pointwise" convergence is therefore ill-defined, unless we stipulate that this convergence is in probability, almost surely, etc... Our results will therefore be concerned with showing that these random functions F^{A_n} converge ...

4.2 The Stieljes Transform

Above, we have introduced the characteristic function as a standard too for establishing results about convergence in distribution. In random matrix theory, the more natural function to look at is the Stiltjes transform, as will be defined shortly. In the following exposition, we emphasize that the Stiljes transform shares many of the same properties of the characteristic function, and hence serves a very similar purpose as a proof tool.

Definition 4.1. Let X be a random variable with distribution function F. The Stiltjes transform of F is the function

$$s_F(z) = \int_{\mathbb{R}} \frac{dF(x)}{x-z} = \mathbb{E}[\frac{1}{X-Z}] \quad z \in \mathbb{C}/\mathbb{R} \tag{44}$$

Write $z = u+iv$ for $u, v \in \mathbb{R}$. We see at once that $s_F(z) \leq \frac{1}{|v|}$, so this value is well defined. Like the characteristic function, the Stieljes transform can be inverted and its pointwise convergence is equivalent to convergence in distribution. To prove the inversion, we need a classic result in probability theory known as Slutsky's theorem:

Theorem 4.1. ***Slutsky's Theorem*** *Let X_n and Y_n be a sequence of random variables such that $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} c$ where c is a constant. Then $X_n + Y_n \xrightarrow{D} X + c$.*

Theorem 4.2. ***Inversion Formula*** *Let F be a distribution and let I be an open interval (a, b) such that F is continuous at a, b. Then*

$$\begin{aligned} F(b) - F(a) &= \lim_{t\to 0} \frac{1}{\pi} \int_I \frac{S_F(x+it) - S_F(x-it)}{2i} \\ &= \lim_{t\to 0} \frac{1}{\pi} \int_I Im(S_F(x+it)) \end{aligned} \tag{45}$$

Proof. Write $z = u + iv$. Note that

$$\begin{aligned} \mathrm{Im}(S_F(x+it))/\pi &= \frac{1}{\pi} \int_{\mathbb{R}} \frac{dF(x)}{x-z} \\ &= \frac{1}{\pi} \int_{\mathbb{R}} \frac{dF(x)}{(x-u)^2 + v^2} \end{aligned} \tag{46}$$

If we let C_t be a continuous following random variable on the interval $(-\infty, \infty)$ with density $\frac{t}{\pi(x^2+t^2)}$ (this is known as a Cauchy random variable), then we see that $\mathrm{Im}(S_F(x+it))$ has the

density of $X + C_t$, where X has distribution function F. A routine computation shows that $\lim_{t\to\infty} Pr(|C_t| > \epsilon) = 0$ for any $\epsilon > 0$, that is, $C \xrightarrow{P} 0$. By Slutsky's theorem, $\lim_{t\to 0} X + C_t \xrightarrow{D} X$, and since the endpoints (a, b) of I are points of continuity of F, we see that

$$\lim_{t\to 0} \frac{1}{\pi} \int_I \mathrm{Im}(S_F(x + it))dx = \lim t \to 0\{F_{X+C_t}(b) - F_{X+C_t}(a)\} = F(b) - F(a) \tag{47}$$

as needed. □

Remark. Note that if X is a continuous random variable with density f, then we can compute the density at a point $x \in \mathbb{R}$ by evaluatng $F'(x)$. Thus, we can recover a density from the Stiltjes transform. Moreover, we have the following crucial theorem.

Theorem 4.3. *Let X_n be a sequence of random variables with distribution F_n, and X be a random variable distributed according to F. Then*

1. *If $X_n \xrightarrow{D} F$ then $S_{F_n}(z)$ converges to $S_F(z)$ for all $z \in \mathbb{C}/\mathbb{R}$*
2. *If F_n are random distributions (e.g. ESDs), and for each $z \in \mathbb{C}/\mathbb{R}$, $S_{F(z)}$ converges in probability to a deterministic limit $S(z)$ which is the Stiltjes transform of a distribution function F, then $F_n \xrightarrow{D} F$*

Proof. The first point follows from Theorem **??**, since the real and imaginary parts of the Stiltjes integrand are bounded, continuous functions. The proof of the second statement, while succinct, relies on a bit of analytic machinery and is omitted. However, the argument is essentially the same as in the proof of the equivalence of convergence in distribution and pointwise convergence of characteristic functions. The main idea is to use a result known as Helly's theorem to extract a convergent subsequence of distributions, and then apply the inversion formula to demonstrate that all convergent subsequences have the same limit. Indeed, this type of argument works is standard for when you have an invertible transformation of a random variable obtained by integrating a bounded and continuous function. □

The Stiltjes transform $s_n(z)$ of the ESD of a diagonalizable matrix $n \times n$ matrix $A = U\Sigma U^*$ takes a rather nice form:

$$\begin{aligned} s_n(z) = & \int \frac{1}{x - z}]dF^{M/\sqrt{n}}(x) \\ = & \frac{1}{n}\sum_{i=1}^{n}(\lambda_i/\sqrt{n} - z)^{-1} \\ = & \frac{1}{n}\mathrm{tr}(\Lambda\sqrt{n} - zI)^{-1} \\ = & \frac{1}{n}\mathrm{tr}(M\sqrt{n} - zI)^{-1} \end{aligned} \tag{48}$$

where the last step follows from conjugation by U. We can rewrite the above identity as a formal power series in z:

$$s_n(z) = -\frac{1}{n}\sum_{k=0}^{\infty} \frac{\mathrm{tr}(M^k)}{z^{k+1}} \tag{49}$$

which converges as long as z is large enough. Note the similarity between the power series in the Stiltjes transform and the cofficents in the characteristic function power series. If A is Hermitian, then using the indentity in Equation 36, we can also express the empirical spectral density as

$$s_n(z) = \frac{1}{n}\sum_{k=1}^{n} \frac{1}{a_k k - z - \alpha_k^*(A_k - zI)^{-1}\alpha_k} \tag{50}$$

where the notation is just as in Equation 36: $a_{kk} = A_{kk}$, A_k is A with the k^{th} row and columns removed, and α_k is the k^{th} column of A with the k^{th} entry removed.
We will aso need this following, quite uninspiring, technical lemma about the Stiljes transform later on

Lemma 4.4. *Let y be real and greater than zero, let F be a distribution function on $\mathbb{R}^+$, and let $s(z)$ be the stieltjes transform of F. Then, the quantity $y + z - 1 + yzs(z)$ lies in the upper half plane.*

Proof. Write $z = u + iv$. We compute

$$\begin{aligned} \mathrm{Im}(y + z - 1 + yzs(z)) &= \mathrm{Im}\{z - 1 + \int_0^\infty \frac{yxdF(x)}{x - z}\} \\ &= \mathrm{Im}\{v + \int_0^\infty \frac{yx(x - \overline{z})}{(x - z)(x - \overline{z})}\} \\ &= v\{1 + \int_0^\infty \frac{yx}{|x - z|}\} > 0 \end{aligned} \tag{51}$$

□

5 Now, for the proof

5.1 Set Up and Motivation

We define the covariance between two random complex-valued random variables as

$$\mathrm{Cov}(X, Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))^*] \tag{52}$$

In the case where X and Y are real valued, positive correlation means that X tends to above its average when Y is also above its average, and negative correlations means that X tends to be above its average when Y is below is its average, and vice versay.For two vectors $X, Y \in \mathbb{R}^n$ where each entry has mean 0, we define the covariance between two vectors:

$$\mathrm{Cov}(X, Y)_{ij} = \mathbb{E}[(X_i - \mathbb{E}[X_i])Y_j] = (\mathbb{E}[(X - \mathbb{E}[X])(Y - \mathbb{E}[Y])^*])_{ij} \tag{53}$$

Often, we are concerned with correlations between the entries of X, so our goal is to understand the properties of the covariance matrix of X, given by $\mathrm{Cov}(X, X)$. For example, the entries of X could be levels of gene expression or phases of a discrete time Fourier transform, and our goal is to see how these different variables interact. We will see further motivation for considering the eigenspectrum of XX^* in the subsequent talks on PCA, but to whet your pallete, the following remark should give some prelimary motivation

Remark. Let $C = \mathrm{Cov}(X, X)$. Since conjugation is an $\mathbb{R}$-linear operation, it computes with the trace we see that

$$C^* = (\mathbb{E}[X - \mathbb{E}[X](X^* - \mathbb{E}[X])])^* = (\mathbb{E}[(X - \mathbb{E}[X])(X^* - \mathbb{E}[X])]) = C \tag{54}$$

Consequently, C is hermitian, and can be diagonalized as $UCU^* = \Sigma$, where Σ is a diagonal matrix of real eigenvalues. Again, using the linearity of expectation, we can verify that the entries of Σ are nnnegative. Talk about more...

In real world applications, we don't actually observe the covariance matrix; we only see realizations of the random variables XX^T and we would like to know how this object, and its properties, are distributed. Now, our random variables can have very wild distributions, but we hope that if we do some averaging, we will start to see some regularity.

To this end, let X be a random vector in $\mathbb{C}^p$, and suppose we have observations $x_1, \ldots, x_n$ of X. For convenient notation, let M be the matrix whose columns are $x_1, \ldots, x_n$.

We define the following quantities:

Definition 5.1. The sample mean of X is defined as the average over all observations $x_1, \ldots, x_n$ of X:

$$\overline{x}_n = \frac{1}{n} \sum_i x_i \tag{55}$$

Definition 5.2. In high dimensional spectral analysis, the sample covariance matrix of X is defined by

$$S = \frac{1}{n} \sum_{i=1}^{n} x_k x_k^* = \frac{1}{n} \sum_{k=1}^{n} (x_k x_k^*) = \frac{MM^*}{n} \tag{56}$$

Remark. In statistics, the sample covariance matrix is defined by

$$\begin{aligned} S &= \frac{1}{n-1} \sum_{i=1}^{n} x_k x_k^* = \frac{1}{n} \sum_{k=1}^{n} (x_k - \overline{x}_n)(x_k - \overline{x}_n)^* \\ &= \frac{1}{n}(MM^* - \overline{x}_n \overline{x}_n^*) \end{aligned}$$

Later on, we will see a lemma which states that in the limit of n large, adding the rank one matrix $\overline{x}_n \overline{x}_n^*$ does not change the limiting spectral distribution.

In spectral analysis, we assume that the number of observations n grow at the space pace as the dimension p of X, that is, $n/p \to y \in (0, \infty)$.

5.2 Perterbation Inequalities

In this section, we will use π to denote a finite permutation on n elements. We will need the following theorem, which sqeezes the Frobenius distance between two $n \times p$ matrices $\mathbf{A}$ and $\mathbf{B}$ by the l_2 distance between their singular values:

Theorem 5.1. *(i) Let* $\mathbf{A}$ *and* $\mathbf{B}$ *be two* $n \times n$ *normal matrices with eigenvalues* λ_k *and* δ_k*, respectively, for* $k = 1, \ldots, n$*. Then,*

$$\min_{\pi \in S_n} \sum_{k=1}^{n} |\lambda_k - \delta_{\pi(k)}|^2 \leq tr[(\mathbf{A} - \mathbf{B})(\mathbf{A} - \mathbf{B})^*] \leq \min_{\pi \in S_n} \sum_{k=1}^{n} |\lambda_k - \delta_{\pi(k)}|^2 \tag{57}$$

(ii) If $\mathbf{A}$ *and* $\mathbf{B}$ *are two* $n \times p$ *matrices where* λ_k *and* δ_k *are the singular values, then (i) holds as well. Moreover, if the eigenvalues are arranged in decreasing order, we have the following bound:*

$$\sum_{k=1}^{\nu} |\lambda_k - \delta_{\pi(k)}|^2 \leq tr[(\mathbf{A} - \mathbf{B})(\mathbf{A} - \mathbf{B})^*] \tag{58}$$

where $\nu = \min(n, p)$

The proof of this theorem is long, tedious, and beyond the scope of this exposition. However, it will come in handy proving that, if A is a matrix and A' a peterbation, then their eigenvalues will still be relatively close. To make the notion of "closeness" more rigorous, we introduce the *Levy Distance*.

Definition 5.3. Given two, two-dimensional distribution functions F and G, we define the *Levy Distance* between them as

$$L(F, G) = \inf\{\epsilon : F(x - \epsilon, y - \epsilon) - \epsilon \leq F(x + \epsilon, y + \epsilon) + \epsilon\} \quad \forall (x, y) \in \mathbb{R}^2 \tag{59}$$

If F and G are one dimensional distribution functions, we can extend them to two-dimensdional distribution functions as follows:

$$\tilde{F}(x, y) = \begin{cases} F(x) & \text{if } y \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{60}$$

and

$$\tilde{G}(x, y) = \begin{cases} G(x) & \text{if } y \geq 0 \\ 0 & \text{otherwise} \end{cases} \tag{61}$$

Remark. Recall that if F is a continuous distribution function corresponding to a real valued random variable X, then the real valued random variables X_n are said to converge in distribution to X if their corresponding distributions F_n converge pointwise to F. Since distributions are monotone, we see that convergence in Levi-Distance of $F_n \to F$ implies a pointwise convergence. Thus, when we impose slightly stronger assumptions about our random variables (i.e. truncating and removing the mean) than in the general statement, it will suffice to use a perturbation inequality to show that these assumptions do not affect the Levy Distance too strongly. Of course, this will be made more precise in the following subsection.

Since distribution functions take values between 0 and 1, the Levy distance is always less than or equal to 1. This observation let's us prove the following result:

Theorem 5.2. *Let $\{\lambda_k\}$ and $\{\delta_k\}$, $k = 1, 2, \ldots n$ be two sets of complex numbers and let their empiricial distributions be F and $\overline{F}$, respectively. Then, for any $\alpha > 0$, we have*

$$L(F, \overline{F})^{\alpha+1} \leq \min_{\pi \in S_n} \frac{1}{n} \sum_{k=1}^{n} \|\delta_k - \lambda_{\pi(k)}\|^{\alpha} \tag{62}$$

where F and $\overline{F}$ are regarded as two dimensional distributions for the real and imaginary components of λ_k and δ_k.

Proof. Let $d = \frac{1}{n} \sum_{k=1}^{n} \|\delta_k - \lambda_{\pi(k)}\|^{\alpha}$. By the remark above, we may assume that $d < 1$. Now, take ϵ for which $1 > \epsilon^{\alpha+1} > d$. We now define the sets $A(x, y)$ and $B(x, y)$ as

$$\begin{aligned} A(x, y) &= \{k \leq n; \mathrm{Re}(\lambda_k) \leq x, \mathrm{Im}(\lambda_k) \leq y\} \\ B(x, y) &= \{k \leq n; \mathrm{Re}(\delta_k) \leq x + \epsilon, \mathrm{Im}(\lambda_k) \leq y + \epsilon\} \end{aligned} \tag{63}$$

We then have that

$$\begin{aligned} F(x, y) - \overline{F}(x + \epsilon, y + \epsilon) &\leq \frac{1}{n} |A(x, y) - B(x, y)| \\ &\leq \frac{1}{n\epsilon^{\alpha}} \sum_{k=1}^{n} |\lambda_k - \delta_k|^{\alpha} \\ &\leq \epsilon \end{aligned} \tag{64}$$

In the first inequality, we note that the elements in $A(x, y) - B(x, y)$ are precisely the ones which contribute to $F(x, y)$ but may not contrribute $\overline{F}(x, y)$. The second inequality relies upon the observation that if $k \in A(x, y) - B(x, y)$, we must have that $|\lambda_k - \delta_k| \leq \epsilon$ The same argument shows that $\overline{F}(x - \epsilon, y - \epsilon) - F(x, y) \leq \epsilon$, which gives us the desired result. □

We will need the following corrolary for our proof:

Corollary. *Let $\mathbf{A}$ and $\mathbf{B}$ be two $p \times n$ matrices, and denote the ESDs of $\mathbf{S} = \mathbf{A}\mathbf{A}^*$ and $\overline{\mathbf{S}} = \mathbf{B}\mathbf{B}^*$ be denoted by $\mathbf{F}^{\mathbf{S}}$ and $\mathbf{F}^{S}$ respectively. Then,*

$$L^4(F^{\mathbf{S}}, F^{S}) \leq \frac{2}{p^2} \{tr(\mathbf{A}\mathbf{A}^* + \mathbf{B}\mathbf{B}^*)\} \{tr\,[(\mathbf{A} - \mathbf{B})(\mathbf{A} - \mathbf{B})^*]\} \tag{65}$$

Proof. Let λ_k and δ_k be the singular values of $\mathbf{A}$ and $\mathbf{B}$, respectively. Then λ_k and δ_k are the squares of the eigenvalues of $\mathbf{S}$ and $\overline{\mathbf{S}}$. From **??** and **??**, we set $\alpha = 1$ to see

$$\begin{aligned} L^2(F^{\mathbf{S}}, F^{S}) &\leq \frac{1}{p} \sum_{k=1}^{p} |\lambda_k^2 - \delta_k^2| \\ &\leq \frac{1}{p} (2 \sum_{k=1}^{p} (\lambda_k + \delta_k)^2)^{1/2} \frac{1}{p} (\sum_{k=1}^{p} |\lambda_k + \delta_k|^2)^{1/2} \\ &\leq \frac{1}{p} (\sum_{k=1}^{p} (\lambda_k + \delta_k)^2)^{1/2}) (\frac{1}{p} \sum_{k=1}^{p} |\lambda_k + \delta_k|^2)^{1/2} \\ &= \{\frac{2}{p} \mathrm{tr}(\mathbf{A}\mathbf{A}^* + \mathbf{B}\mathbf{B}^*)\}^{1/2} (\frac{1}{p} \sum_{k=1}^{p} |\lambda_k + \delta_k|^2)^{1/2} \\ &\leq \{\frac{2}{p} \mathrm{tr}(\mathbf{A}\mathbf{A}^* + \mathbf{B}\mathbf{B}^*)\}^{1/2} \{\frac{1}{p} \mathrm{tr}\,[(\mathbf{A} - \mathbf{B})(\mathbf{A} - \mathbf{B}))^*]\}^{1/2} \end{aligned} \tag{66}$$

□

5.3 Truncation Assumptions

Proposition 5.3. *We may assume, without loss of generality, that*

1. $|X_i| < \eta_n \sqrt{n}$ *for all* $1 \leq i \leq p(n)$, *and all* $n > 0$.
2. $\mathbb{E}[x_{ij}] = 0$ *and* $Var(x_{ij}) = 1$.

Remark. With the right lemmas, the proof of this proposition is rather straightforward. Unfortunately, the lemmas themselves are a bit clunky and technical. Nevertheless, truncation is very helpful whenever you can show it works. Indeed, both proofs of the MP law given in Bai and Silverstein both make use of the above truncation assumptions. For a broader discussion of how truncation can be used to prove probability-theoretic asymptotics, see [Tao].

Proof. The assumption that $\mathrm{Var}(x_{ij}) = 1$ is left as part of one of the exercises on this weeks homework. The rest of the propsition is proven essentially verbatin from Bai and Silverstein, with minor modifcations (we don't assume that the random variables are iid). Assuming that the variance scaling holds, define

$$\begin{aligned}
\hat{x}_{ij} &= x_{ij}\mathbb{1}(|x_{ij}| \leq C) \qquad \tilde{x}_{ij} = \hat{x}_{ij} - \mathbb{E}[x_{ij}] \\
\hat{x}_i &= (\hat{x}_{i1}, \ldots, x_{ip})^T \qquad \widetilde{x}_i = (\widetilde{x}_{i1}, \ldots, \widetilde{x}_{ip})^T \\
\hat{S}_n &= \frac{1}{n}\sum_{i=1}^{n} \hat{x}_i \hat{x}_i^* = \frac{1}{n}\hat{X}\hat{X}^* \\
\tilde{S}_n &= \frac{1}{n}\sum_{i=1}^{n} \tilde{x}_i \tilde{x}_i^* = \frac{1}{n}\tilde{X}\tilde{X}^*
\end{aligned}$$

Write the ESDs of $\hat{S}^n$ and $\tilde{S}^n$ as $F^{\hat{S}_n}$ and $F^{\tilde{S}_n}$. We have that

$$\begin{aligned}
L^4(F^S, F^{\hat{S}_n}) &\leq \left(\frac{2}{np}\sum_{i,j}(|x_{ij}^2| + |\hat{x}_{ij}^2|)\right)\left(\frac{1}{np}\sum_{i,j}(|x_{ij} - \hat{x}_{ij}|^2)\right) \\
&\leq \left(\frac{4}{np}\sum_{i,j}(|x_{ij}^2| + |\hat{x}_{ij}^2|)\right)\left(\frac{1}{np}\sum_{i,j}(|x_{ij}^2|\mathbb{1}_{x_{ij}>C})\right) \qquad (67) \\
&\overset{a.s}{\to} 4\mathbb{E}[|x_{ij}^2|\mathbb{1}(x_{ij} > C)] \\
&\leq 4Pr(x_{ij} > C) \quad \text{Cauchy Schwartz}
\end{aligned}$$

which can be made arbitrarily small by letting $C \to 0$. Now, by theorem

$$\|F^{\hat{S}_n} - F^{\tilde{S}_n}\| \leq \frac{1}{p}\mathrm{rank}(\mathbb{E}[\hat{X}]) \leq \frac{1}{p}\mathrm{rank}(n^2 \overline{x}_n \overline{x}_n^*) \qquad (68)$$

□

In what follows, we will work with these asumptions. Moreover, we will use the notation $y_n = p(n)/n$.

5.4 Getting the right Stieljes Transform

6 Proof Part 2: Showing the Necessary Convergences

6.1 Proof of Claim 1

For some $n \times n$ Wishart matrix M, let $\mathcal{F}_k$ be the sigma algebra generated by the entries $\{x_{ij} : i, j > k\}$, and let $\mathbb{E}_k[\cdot] = \mathbb{E}[\cdot|\mathcal{F}_k]$. Also, let α_k denote the k^th column of M with α removed, and let M_k denote the $(n-1) \times (n-1)$ matrix with the k^{th} row and columns removed Note then that $\mathbb{E}_k[M]$ and $\mathbb{E}_{k-1}[M-1]$ differ only on the k^{th} row and column of M, so in fact

$$\mathbb{E}_k[\text{tr}(M_k - zI)^{-1})] \quad = \quad \mathbb{E}_{k-1}[\text{tr}(M_k - zI)^{-1})] \tag{69}$$

Thus,

$$\begin{aligned}
\gamma_k &= \frac{1}{n}\left(\mathbb{E}_{k-1}\text{tr}(M_n - zI)^{-1} - \mathbb{E}_k(\text{tr}(M_n - zI)^{-1}\right) \\
\gamma_k &= \frac{1}{n}(\mathbb{E}_{k-1}\left[\text{tr}(M_n - zI)^{-1} - (M_k - zI)^{-1})\right] \\
&\quad - \mathbb{E}_k\left[\text{tr}(M_n - zI)^{-1} - \text{tr}(M_k - zI)^{-1})\right] \\
&= \frac{1}{n}[\mathbb{E}_{k-1} - \mathbb{E}_k]\left[\frac{1 + \alpha_k^*(M_k - zI_p)^{-2}\alpha_k}{-z - \alpha_k^*(M_k - zI_{n-1})^{-1}\alpha_k}\right]
\end{aligned} \tag{70}$$

By Cauchy Schwartz,

$$\begin{aligned}
\left|\alpha_k^*(M_{nk} - zI_{m-1})^{-2}\alpha_k\right| &\leq 1 + \|(W_k - zI_{n-1})\alpha_k\|_2^2 \\
&= 1 + x_k^*\left((M_{nk} - uI_p)^2 + v^2 I_p\right)^{-1} x_k
\end{aligned} \tag{71}$$

Applying an orthongal basis we can diagonalize $S_{nk}^{-1} = U\Sigma U_k^U$and write $y_k = Ux_k$, which gives

$$x_k^*\left((M_{nk} - uI_p)^2 + v^2 I_p\right)^{-1} x_k = \sum_{k=1}^{p}(\frac{1}{(\Lambda_{kk} - u)^2 + v^2})^2\|y_k\| \tag{72}$$

while

$$\begin{aligned}
1 + x_k^*(S_{nk} - ZI_p)^{-1}x_k = & \quad 1 + \sum_{k=1}^{p}\frac{\|y_k\|^2}{\Lambda_{kk} - uI_p - ivI_p} \\
= & \quad 1 + \sum_{k=1}^{p}\frac{\|y_k\|^2(\Lambda_{kk} - u + iv)}{\Lambda_{kk} - u + iv} \\
= & \quad 1 + \sum_{k=1}^{p}\frac{\|y_k\|^2(\Lambda_{kk} - u + iv)}{(\Lambda_{kk} - u)^2 + v^2}
\end{aligned} \tag{73}$$

Taking the quotient of Equation 72 with the imaginary part of Equation 73 gives the bound

$$\left|\frac{x_k^*(S_{nk} - zI_p)^{-2}x_k}{1 + x_k^*(S_{nk} - zI_p)^{-1}x_k}\right| \leq \frac{1}{v} \tag{74}$$

Now $\{y_k\}$ is a sequence of matringale differences, so by ... with $p = 4$, we see that

$$\mathbb{E}[s_n(z) - \mathbb{E}s_n(z)] \quad \leq \quad \frac{K_4}{p^4}\mathbb{E}\left(\sum_{k=1}^{n} |y_k|^2\right) \quad \leq \quad \frac{4K_4 n^2}{v^4 p^2} = O(n^{-1}) \tag{75}$$

recalling that n/p tends to a constant. Now, by Markov's Inequality,

$$P(\|s_n - \mathbb{E}s_n(z)\|^2 \geq n^{-1/2}) \leq n^{1/2}\|s_n - \mathbb{E}s_n(z)\|^2 = O(n^{-3/2}) \tag{76}$$

so that

$$\sum_{n\geq 0} P(\|s_n - \mathbb{E}s_n(z)\|^4 \geq n^{-1/2}) < \infty \tag{77}$$

It follows from the Borel Cantelli Lemma, applied to the events $E_n = \{\omega : \|s_n - \mathbb{E}s_n(z)\| \geq n^{-1/8}\}$, that $s_n \to \mathbb{E}s_n(z)$ almost surely (that is, the probability of the set of ω for which the two disagree for more than finitely many n is zero).

6.2 Proof of Claim 2: Turning the Stieljes Crank

As Professor Rigollet in the ORFE department remarked "The results in random matrix theory are very beautiful, but to get them, you just apply the Stieljes Transform and turn the crank". In this part of the proof, we begin turning the crank. The key techniques here are inverting some generating functions, beging the right roots, and confirming that the asymptotics are what we want. Define

$$\lambda_k = \frac{1}{n + \overline{\alpha_k \alpha_k^*} - 1 - \frac{1}{n^2}\alpha_k^T X_k^*(\frac{1}{n}X_k X_k^* - zI_{p-1})^{-1}X_k\overline{\alpha}_k} \tag{78}$$

$$\epsilon_k = \delta_k + y_n + y_n z\mathbb{E}[s_n(z)] \tag{79}$$

Then

$$s_n(z) = \frac{1}{p}\sum_{i=1}^{p}\frac{1}{\lambda_k + 1 - z} \tag{80}$$

With a litle bit of algebra, we can invert this expression to get

$$\mathbb{E}[s_n(z)] = \frac{1}{1 - z - y_n - y_n z\mathbb{E}s_n(z)} + \delta_n \tag{81}$$

where

$$\delta_n = -\frac{1}{p}\sum_{k=1}^{p}\mathbb{E}\left[\frac{1}{(1 - z - y_n - y_n z\mathbb{E}s_n(z))(1 - z - y_n - y_n z\mathbb{E}s_n(z) + \epsilon_k)}\right] \tag{82}$$

With the quadratic formula, we can solve Equation **??** for $\mathbb{E}[s_n(z)]$, giving us two complex roots:

$$s_1(z) = \frac{1}{2y_n z}(1 - z - y_n + y_n z\delta_n + \sqrt{(1 - z - y_n - y_n z\delta_n)^2})$$

$$s_2(z) = \frac{1}{2y_n z}(1 - z - y_n + y_n z\delta_n - \sqrt{(1 - z - y_n - y_n z\delta_n)^2})$$

The Marchenko Pastur law will follow from the following proposition

Proposition 6.1. *For all $n > 0$ and all $z \in \mathbb{C}$*

$$\mathbb{E}s_n(z) = s_1(z) \tag{83}$$

and, as $n \to \infty$

$$\delta_n \to \tag{84}$$

Proof. Write $z = u + iv$. If $v \to \infty$, then we see that $\mathbb{E}s_n(z) \to 0$ (PROOVE) and then $\delta_n \to 0$. Recall from complex analysis that the square root function can be defined analytically in half of the complex plane. Thus, there is a neighborhood containing all large v for which $\mathbb{E}s_n = s_1(z)$. By continuity of s_1 and s_2, and continuity of $\mathbb{E}s_n(z)$, there is a $z_0 \in \mathbb{C}$ (WHY) such that $s_1(z_0) = s_2(z_0)$ which means that the quantity inside the square root vanishes at that z_0. It follows that, for that z_0

$$\mathbb{E}s_n(z_0) = s_1(z_0) = \frac{1 - z_0 - y_n + y_n z_0 \delta_n}{2y_n z_0} \tag{85}$$

Using the solution δ_n in terms of $\mathbb{E}s_n(z_0)$ from Equation 82, we see that

$$\mathbb{E}s_n(z_0) = \frac{1 - z_0 - y_n}{y_n z_0} + \frac{1}{y_n + z_0 - 1 + y_n z_0 \mathbb{E}s_n(z_0)} \tag{86}$$

By the lemma (which one), the second term lies in lower half plane. We can solve a quadratic equation implied by 86 and choose the correct root using Lemma 4.4. It follows that

$$y_n + z_0 - 1 + y_n z_0 \mathbb{E}[s_n(z_0)] = \sqrt{y_n z_0} \tag{87}$$

The contradiction now follows from a classic matrix identity that XX^* and X^*X have the same set of nonzero eigenvectors. It follows that the ESD of X^*X is obtained by EXPLAIN. This gives us the identify

$$s_n(z) = y_n^{-1} \underline{s}_n(z) - \frac{1 - 1/y_n}{z} \tag{88}$$

taking the Expectation of this equation shows that

$$y_n - 1 + y_n z_0 \mathbb{E}[s_n(z_0)] = z_0 \mathbb{E}[\underline{s}_n(z_0)] \tag{89}$$

which we subsitute into Equation 87:

$$1 + \mathbb{E}s_n(z_0) = \sqrt{y}/\sqrt{z_0} \tag{90}$$

Recall that z_0 lies in the upper half plane, so that $1/\sqrt{z_0}$ has a negative imaginary component. But the left hand side has a positive imaginary component, which yields a contradiction. It follows that $\mathbb{E}s_n(z) = s_1(z)$, as needed.

Lets now show that $\delta_n \to 0$. Here we will make finally use of our truncation assumptions.δ_n is the expectation of sum of elements of the form $\mathbb{E}_b \frac{b}{a(a+b)}$. From the elementary identity

$$\frac{b}{a(a+b)} = \frac{b^2}{a^2(a+b)} - \frac{b}{a^2} \tag{91}$$

We can write

$$\delta_n = J_1 + J_2 \tag{92}$$

Where,

$$J_1 = -\frac{1}{p}\sum_{k=1}^{p}\left(\frac{\mathbb{E}[\epsilon_k]}{(1-z-y_n-y_n z\mathbb{E}[z_n])^2}\right) \tag{93}$$

$$J_2 = \frac{1}{p}\sum_{k=1}^{p}\left(\frac{\mathbb{E}[\epsilon_k]}{(1-z-y_n-y_n z\mathbb{E}[z_n])((1-z-y_n-y_n z\mathbb{E}[z_n])+\epsilon_k)}\right) \tag{94}$$

To show $J_1 \to 0$, it suffices to show that $\mathbb{E}_k \to 0$. We compute

$$|\mathbb{E}\epsilon_k| = \left|-\frac{1}{n^2}\mathbb{E}[\mathrm{tr}]\right| \tag{95}$$

□

As a consequence,

7 Finally..Some Random Matrix Theory

7.1 Random Matrices

7.2 Stiljes Transform

7.3 An Introduction to Free Probability

Unfortunately, the non-commutativity of matrix multiplication introduces some technical complications. To get around them, we will appeal to the theory of Free Probability. First, we will need to define a $\mathbb{C}*$ring algebra, which basically a generalization of the algebraic properties of complex matrices. More formally,

Definition 7.1. A $\mathbb{C}*$ algebra A is $\mathbb{C}$ vector space with the following additional structure:

1. There is an associative binary operation, which we call multiplication, from $A \times A \to A$ write as $X \cdot Y$ or XY. Addition distributes over multiplication, and multiplication commutes with scalar multiplication.

2. There is an operation $* : A \to A$ with the following properties for all $x, y \in A$.

   (a) $(x^*)^* = x$
   (b) $(x+y)^* = x^* + y^*$
   (c) $(xy)^* = y^* x^*$
   (d) $(cx) = \bar{c}x^*$ for all $c \in \mathbb{C}$

Just as with matrices, we say X is *self-adjoint* if $X = X^*$ and if $X^*X = XX^*$, we say that X is *normal*

We are now ready to prove the Marchenko Pastur Law. We begin by noting that each matrix W_n can be written as the sum of rank one matrices: $W_n = (r_i^s r_j^s$