Random Matrix Theory Notes

Written By Robert Ferydouni

Based On Math 233 Lectures By Torsten Ehrhardt, UCSC Spring 2022

1 Review of Probability Theory

1.1 Basics

A **probability space** is a triple $(\Omega, \mathcal{F}, P)$ where Ω is a nonempty set that is the sample space (the atomic events), $\mathcal{F}$ is the σ-algebra of measurable events, and $P : \mathcal{F} \to [0, 1]$ is a probability measure (though in general we can also have complex valued probability measures). $\mathcal{F} \in \mathcal{P}(\Omega)$ satisfies

- $A \in \mathcal{F} \implies \Omega/A = A^c \in \mathcal{F}$
- $A, B \in \mathcal{F} \implies A \cup B \in \mathcal{F}$
- $A_n \in \mathcal{F}$ for all $n \in \mathbb{N} \implies \bigcup_{n\in\mathbb{N}} A_n \in \mathcal{F}$

The second point is redundant for a σ-algebra. The first two points together alone make $\mathcal{F}$ an algebra. It can be shown by the above $\mathcal{F}$ is closed under finite and countable complements as well. One can also take any $A \in \mathcal{F}$ and see $\Omega = A \cup A^c \in \mathcal{F}$ so that $\varnothing = \Omega^c \in \mathcal{F}$ as well. P on $(\Omega, \mathcal{F})$ (domain is the σ-algebra $\mathcal{F}$) is a probability measure if

- $P(\Omega) = 1$ (shows $P(\varnothing) = 0$, which agrees with fact P is a measure)
- $P(A \cup B) = P(A) + P(B)$ for $A, B \in \mathcal{F}$ disjoint
- $P\left(\bigcup_{n\in\mathbb{N}} A_n\right) = \sum_{n\in\mathbb{N}} P(A_n)$ where each $A_n \in \mathcal{F}$ and pairwise disjoint

We will refer to the second property as additivity and the third property as σ-additivity. Again we note the second property is redundant.

Regarding language, we make the following comparison in the kinds of questions we ask.

Measure Theory	**Probability Theory**
P(A)	$P(\omega \in A)$ for an ”event” A
$A \cap B$	A and B, $(\omega \in A) and (\omega \in B)$
$A \cup B$	A or B, $(\omega \in A)$ or $(\omega \in B)$

Example 1.1.1. *Tossing a Die: Let* $\Omega = \{1, ..., 6\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ *and* $P(A) = \frac{|A|}{6}$

Example 1.1.2. N *Coin Tosses: Let* $\Omega = \{0, 1\}^N$, $\mathcal{F} = \mathcal{P}(\Omega)$ *and* $P(A) = \frac{|A|}{2^N}$

Example 1.1.3. *Let* $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}(\mathbb{R})$, *the Borel* σ*-algebra, the smallest* σ*-algebra containing open, closed and* h*-intervals.*

We remark that the smallest σ-algebra containing some $S \in \Omega$ is the intersection of all σ-algebras containing S. We also refer to this as the σ-algebra generated by S, and denote it by $\sigma(S)$.

1.2 Extension of Probability Spaces

We say $(\Omega, \mathcal{F}, P)$ has extension $(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P})$ if there is a surjective projection map $\pi : \widehat{\Omega} \to \Omega$ that is measurable. I.e., $A \in \mathcal{F} \implies \pi^{-1}(A) \in \widehat{F}$. We also require $P(A) = \widehat{P}(\pi^{-1}(A))$.

Example 1.2.1. *Recall the probability triple $(\Omega, \mathcal{F}, P)$ where $\Omega = \{1, ..., 6\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ and $P(A) = \frac{|A|}{6}$. Consider the extension $(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P})$ where $\widehat{\Omega} = \{1, ..., 6\} \times \{1, ..., 6\}$, $\widehat{\mathcal{F}} = \mathcal{P}(\widehat{\Omega})$ and $\widehat{P}(B) = \frac{|B|}{36}$. We then have the projection map $\pi(n, m) \mapsto n$.*

All probabilistic notions should not depend on the extension of the probability space. For example, for some $A \in \mathcal{F}$ and $B \in \widehat{\mathcal{F}}$, we no longer need to distinguish between $P(A)$ and $\widehat{P}(B) = \widehat{P}(\pi^{-1}(A))$, or between $A = \varnothing, \Omega$ and $B = \varnothing, \widehat{\Omega}$. However we must distinguish between $|A|$ and $|B|$ as cardinality is not a probabilistic notion.

We say $\omega \in A$ holds **surely** (respectively, **almost surely**) if $A = \Omega$ (respectively, $P(\omega \in A) = 1$). We say $\omega \in A$ holds **never** (respectively, **almost never**) if $A = \varnothing$ (respectively, $P(\omega \in A) = 0$). The notions of "almost surely" and "almost never" are respectively known as "almost everywhere" and "almost nowhere" in more general measure theory.

Properties of Probability Measures:

- $P(A^c) = 1 - P(A)$. Pf. Write $\Omega = A \cup A^c$ an apply the additive property of probability measures.
- $P(A) \leqslant P(B)$. Pf. Write $B = A \cup (B/A)$ so that by additivity $P(B) = P(A) + P(B/A) \geqslant P(A)$.
- $P(A \cup B) \leqslant P(A) + P(B)$. Pf. Write $A \cup B = A \cup (B/A)$ so that by additivity $P(A \cup B) = P(A) + P(B/A) \leqslant P(A) + P(B)$ by the last part.
- $P\left(\bigcup_{n \in \mathbb{N}} A_n\right) \leqslant \sum_{n \in \mathbb{N}} P(A_n)$. Pf. Write $\bigcup_{n \in \mathbb{N}} A_n = A_1 \cup (A_2/A_1) \cup (A_3/(A_1 \cup A_2)) \cdots$, which is a disjoint union, and apply σ-additivity.

Theorem 1.2.1. *Caratheordory Extension: Let $\mathcal{F}_0$ be an algebra on Ω and $\mathcal{F} = \sigma(\mathcal{F}_0)$. Then any probability measure on $\mathcal{F}_0$ uniquely extends to a probability measure on $\mathcal{F}$.*

We give a sketch of the proof of this theorem after the following propositions and lemmas. Note that a "probability measure" μ_0 on $\mathcal{F}_0$ is a premeasure on this algebra, i.e., $\mu_0\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{n \in \mathbb{N}} \mu_0(A_n)$ if $\bigcup_{n \in \mathbb{N}} A_n \in \mathcal{F}_0$, where each $A_n \in \mathcal{F}_0$.

For the following proposition, we say for a sequence of sets $\{A_n\}_n \in \mathbb{N}$ that $A_n \nearrow A$ if $A = \bigcup_{n \in \mathbb{N}} A_n$ and $A_n \subseteq A_{n+1}$ for each $n \in \mathbb{N}$. We say for a sequence of sets $\{A_n\}_n \in \mathbb{N}$ that $A_n \searrow A$ if $A = \bigcap_{n \in \mathbb{N}} A_n$ and $A_{n+1} \subseteq A_n$ for each $n \in \mathbb{N}$

Proposition 1.2.1. *Let μ be a σ-additive measure on a σ-algebra $\mathcal{F}$. (a) If $A_n \nearrow A$ then $\mu(A_n) \to \mu(A)$ as $n \to \infty$. (b) If $A_n \searrow A$ and $\mu(A_1) < \infty$ then $\mu(A_n) \to \mu(A)$ as $n \to \infty$*

Proof. We only prove (a) as the second part is similar. Since $A_n \subseteq A_{n+1}$ for each n, write $A = A_1 + \cup (A_2 \cup A_1) \cup (A_3/A_2) \cup \cdots$, which is a disjoint union. Thus by σ-additivity $\mu(A) = \mu\left(\bigcup_{n \in \mathbb{N}} A_n\right) = \sum_{n \in \mathbb{N}} \mu(A_{n+1}/A_n) = \lim_{j \to \infty} \sum_{n=1}^{j} \mu(A_{n+1}/A_n) = \lim_{j \to \infty} \mu(A_j)$, where the last inequality is by σ-additivity and since $A_n \subseteq A_{n+1}$ for all n. □

Proposition 1.2.2. *Let μ be an additive measure on a σ-algebra $\mathcal{F}$. If (a) $A_n \nearrow A$ implies $\mu(A_n) \to \mu(A)$ as $n \to \infty$ (b) $A_n \searrow A$ implies $\mu(A_n) \to \mu(A)$ as $n \to \infty$, then μ is σ-additive on $\mathcal{F}$.*

Proof. Let $B = \bigcup_{n\in\mathbb{N}} B_n$ where $\{B_n\}_{n\in\mathbb{N}}$ are pairwise disjoint. We want to show $\mu(B) = \sum_{n\in\mathbb{N}} \mu(B_n)$. Define $A_n = B_1 \cup \cdots \cup B_n$. Then we see $A_n \nearrow B$ so that by assumption $\mu(A_n) \to \mu(B)$ as $n \to \infty$. But we also note $\mu(A_n) = \mu(B_1) + \cdots + \mu(B_n) \to \sum_{n\in\mathbb{N}} \mu(B_n)$ as $n \to \infty$, where the first equality is by additivity of μ. □

We are now ready to sketch the proof of the existence of the extension in Caratheordory's Theorem.

Proof. Let μ_0 be a premeasure on $\mathcal{F}_0$. Step 1. Define $\mathcal{G} = \{G \subseteq \Omega \mid \text{ there exists } \{G_n\}_{n\in\mathbb{N}} \text{ such that each } G_n \in \mathcal{F}_0 \text{ and } G_n \nearrow G\}$. We will extend μ_0 to $\mathcal{G}$ by defining $\mu_0(G) = \lim_{n\to\infty} \mu_0(G_n)$. One can show that μ_0 for $G_1, G_2 \in \mathcal{G}$ satisfies $\mu_0(G_1 \cup G_2) + \mu_0(G_1 \cup G_2) = \mu_0(G_1) + \mu_0(G_2)$.

Step 2. Define the outer measure μ^* on $P(\Omega)$ by $\mu^*(A) = \inf\{\mu_0(G) \mid A \subseteq G \text{ and } G \in \mathcal{G}\}$. One can show $\mu^*(G) = \mu(G)$ for all $G \in \mathcal{G}$. One can also show $\mu^*(A_1 \cap A_2) + \mu^*(A_1 \cup A_2) \leqslant \mu^*(A_1) + \mu^*(A_2)$ for all $A_1, A_2 \in \mathcal{P}(\Omega)$.

Step 3. Look at $\mathcal{H} = \{H \in \Omega \mid \mu^*(H) + \mu^*(G/H) = \mu^*(\Omega) = 1\}$. Note $\mathcal{F}_0 \subseteq \mathcal{G} \subseteq \mathcal{H} \subseteq \mathcal{P}(\Omega)$. One can show $\mathcal{H}$ is a σ-algebra and that $\mu := \mu^* |_{\mathcal{H}}$ is additive and σ-additive. □

Remark: The uniqueness part of the statement requires the monotone class theorem.

Theorem 1.2.2. *Let $(\Omega, \mathcal{F}, \mu)$ be a probability space where $\mathcal{F} = \sigma(\mathcal{F}_0)$ for an algebra $\mathcal{F}_0$. Then for all $\varepsilon > 0$ and for all $A \in \mathcal{F}$, there exists $B \in \mathcal{F}_0$ such that $\mu(A \Delta B) < \varepsilon$ (where $A\Delta B = A/B \cup B/A$)*

Proof. Let $\mathcal{G}, \mathcal{H}$ and μ^* as in the proof of the existence of the extension in Caratheordory's Theorem. Given $A \in \mathcal{F}$, we have $\mu(A) = \mu^*(A) = \inf\{\mu(G) \mid A \subseteq G \text{ and } G \in \mathcal{G}\}$. By properites of the infimum, there exists $G \in \mathcal{G}$ such that $\mu(A) \leqslant \mu(G) \leqslant \mu(A) + \varepsilon$. Since $A \subseteq G$, $\mu(G/A) = \mu(G) - \mu(A) \leqslant \varepsilon$ so that $\mu(G\Delta A) < \varepsilon$. Now by definition of $G \in \mathcal{G}$ there exists a sequence $\{G_n\}_{n\in\mathbb{N}}$ where each $G_n \in \mathcal{F}_0$ and $G_n \subseteq G$, such that $G_n \nearrow G$. I.e., $\mu(G_n) \to \mu(G)$ as $n \to \infty$. Thus we find can find an $n \in \mathbb{N}$ such that $\mu(G/G_n) = \mu(G) - \mu(G_n) < \varepsilon$, so that $\mu(G\Delta G_n) < \varepsilon$. Letting $B = G_n$ we have $\mu(A\Delta B) \leqslant \mu(A\Delta G) + \mu(G\Delta B) < 2\varepsilon$. □

Theorem 1.2.3. *Let $(\Omega_1, \mathcal{F}_1, P_1)$ and $(\Omega_2, \mathcal{F}_2, P_2)$ be two probability spaces. Then we have the unique product space $(\Omega, \mathcal{F}, P)$ where $\Omega = \Omega_1 \times \Omega_2$, $\mathcal{F} = \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$ (the σ-algebra generated by subsets of $\mathcal{F}_1 \times \mathcal{F}_2$) and $P((F_1, F_2)) = P(F_1)P(F_2)$ for all $(F_1, F_2) \in \mathcal{F}_1 \times \mathcal{F}_2$.*

Proof. Define $\mathcal{F}_0 = \{\bigcup_{i=1}^N F_i^{(1)} \times F_i^{(2)} \mid F_i^{(1)} \in F_1, F_i^{(2)} \in F_2\}$. Clearly $\mathcal{F}_0$ is closed under finite unions by definition. One can also show it is closed under complements so that $\mathcal{F}_0$ is an algebra. Without loss of generality we can assume the unions $\bigcup_{i=1}^N F_i^{(1)} \times F_i^{(2)}$ are disjoint. Thus define P on $\mathcal{F}_0$ by $P\left(\bigcup_{i=1}^N F_i^{(1)} \times F_i^{(2)}\right) = \sum_{i=1}^N P_1\left(F_i^{(1)}\right) P_2\left(F_i^{(2)}\right)$. It can be shown P is a premeasure on $\mathcal{F}_0$. Now apply Caratheordory's Theorem. □

Theorem 1.2.4. *Let $(\Omega_n, \mathcal{F}_n, P_n)$ be probability spaces for each $n \in \mathbb{N}$. Then we have the unique product space $(\Omega, \mathcal{F}, P)$ where $\Omega = \Pi_{n\in\mathbb{N}}\Omega_n$, $\mathcal{F} = \sigma\left(\Pi_{n\in\mathbb{N}}\mathcal{F}_n\right)$ and*

$$P((F_1, F_2, ..., F_n, \Omega, \Omega, ...)) = P(F_1) \cdots P(F_n) \tag{1}$$

where $(F_1, F_2, ..., F_n, \Omega_{n+1}, \Omega_{n+2}, ...) \in \Pi_{n\in\mathbb{N}}\mathcal{F}_n$

Remark: Note we require an element of $\Pi_{n\in\mathbb{N}}\mathcal{F}_n$ to have all but finitely many components equal to Ω_i, for appropriate $i \in \mathbb{N}$, to have the expression in (1) make sense.

Proof. Define $\mathcal{F}_{00} = \{(F_1, F_2, ..., F_n, \Omega_{n+1}, \Omega_{n+2}, ...) \mid n \in \mathbb{N}$ and $F_i \in \mathcal{F}_i$ for each $i \in \mathbb{N}\}$. Now define $\mathcal{F}_0$ to be the set of finite disjoint unions of elements of $\mathcal{F}_{00}$. One can show $\mathcal{F}_0$ is an algebra. Note $(F_1, F_2, ..., F_n, \Omega_{n+1}, \Omega_{n+2}, ...)$ is a union of $2^n - 1$ sets of this kind. Define P on $\mathcal{F}_{00}$ by (1) in the statement of the theorem, and extend to $\mathcal{F}_0$ by looking at finite sums as in the case for the product space of two probability spaces. One can show this extended version of P is a premeasure on $\mathcal{F}_0$. Then apply Caratheodory's Theorem. □

Remark: $(F_1, F_2, ...) = \Pi_{n\in\mathbb{N}} F_n \in \mathcal{F} :- \sigma(\mathcal{F}_0)$ where each $F_n \in \mathcal{F}_n$ since

$$\Pi_{n\in\mathbb{N}} F_n = \bigcap_{n\in\mathbb{N}} (F_1, F_2, ..., F_n, \Omega_{n+1}, \Omega_{n+2}, ...)$$

This implies $\mathcal{F} = \sigma(\{\Pi_{n\in\mathbb{N}} F_n \mid F_n \in \mathcal{F}_n\})$.

1.3 Random Variables

Given a measurable space $(R, \mathcal{R})$, a **random variable** X is a measurable map from an underlying probability space $(\Omega, \mathcal{F}, P)$ to $(R, \mathcal{R})$. In other words, a function $X : \Omega \to R$ such that $X^{-1}(A) \in \mathcal{F}$ for all $A \in \mathcal{R}$. Random variables are invariant under the extensions of probability spaces, in the sense that if $(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P})$ is an extension of $(\Omega, \mathcal{F}, P)$ with $\pi : (\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P}) \to (\Omega, \mathcal{F}, P)$ the projection map, then the map $X \circ \pi : (\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P}) \to (R, \mathcal{R})$ is a surjection. We define the **distribution** (or **law**) of a random variable X to be the function $\mu_X : \mathcal{R} \to [0, 1]$ defined by $\mu_X(A) = P(X^{-1}(A))$. Note μ_X is a probability measure and makes $(R, \mathcal{R}, \mu_X)$ into a probability space. We now discuss the construction of random variables.

Example 1.3.1. *For a probability space $(R, \mathcal{R}, \mu)$ one can define $(\Omega, \mathcal{F}, P) = (R, \mathcal{R}, \mu)$. Then one can show that the identity map $X : \Omega \to R$ is a random variable with $\mu_X = \mu$.*

Example 1.3.2. *Let $(\Omega, \mathcal{F}, P)$ and $(R, \mathcal{R}, \mu)$ be two probability spaces. We want a probability space $(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P})$ such that $\pi : (\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P}) \to (\Omega, \mathcal{F}, P)$ is the projection map and there exists a random variable $X : (\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P}) \to (R, \mathcal{R})$, whose law agrees with the measure on $(R, \mathcal{R})$.*

Take $\widehat{\Omega} = \Omega \times R$, $\widehat{\mathcal{F}} = \sigma(\mathcal{F} \times \mathcal{R})$ and $\widehat{P}$ the product measure of P and μ. Then $\pi : \Omega \times R \to R$ is given by $\pi((\omega, r)) = \omega$ and we let $X((\omega, r)) = r$. Then for all $A \in \mathcal{F}$ we have $\pi^{-1}(A) = A \times R$ so that $\widehat{P}(\pi^{-1}(A)) = P(A)\mu(R) = P(A)$. Additionally, for all $B \in \mathcal{R}$ we have $X^{-1}(B) = \Omega \times B$ so that $\widehat{P}(X^{-1}(B)) = P(\Omega)\mu(B) = \mu(B)$. This creates a new random variable such that $\mu_X(B) = \widehat{P}(X^{-1}(B)) = \mu(B)$.

Example 1.3.3. *Let $(\Omega, \mathcal{F}, P)$ and $(R_n, \mathcal{R}_n, \mu_n)$ be probability spaces for each $n \in \mathbb{N}$. We want a probability space $(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P})$ such that $\pi : (\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P}) \to (\Omega, \mathcal{F}, P)$ is the projection map and there exists a random variable $X_n : (\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P}) \to (R_n, \mathcal{R}_n)$, whose law agrees with the measure on $(R_n, \mathcal{R}_n)$ for each $n \in \mathbb{N}$.*

Take $\widehat{\Omega} = \Omega \times R_1 \times R_2 \times \cdots$, $\widehat{\mathcal{F}} = \sigma(\mathcal{F} \times R_1 \times R_2 \times \cdots)$ and $\widehat{P}$ the product measure of P and μ_n for all $n \in \mathbb{N}$. Thus $P(A \times B_1 \times B_2 \times \cdots \times B_m \times R_{m+1} \times R_{m+2} \times \cdots) = P(A)\mu_1(B_1) \cdots \mu_n(B_m)$ for all $A \in \mathcal{F}$ and $B_n \in \mathcal{R}_n$, for all $n \in \mathbb{N}$. Then $\pi : \Omega \times R \to R$ is given by $\pi((\omega, r_1, ...)) = \omega$ and we let $X_n((\omega, r_1, ...)) = r_n$. Then for all $A \in \mathcal{F}$ we have $\pi^{-1}(A) = A \times R_1 \times R_2 \times \cdots$, so that $\widehat{P}(\pi^{-1}(A)) = P(A)\mu_1(R_1)\mu_2(R_2) \cdots = P(A)$. Additionally, for all $B_n \in \mathcal{R}_n$, for all $n \in \mathbb{N}$, we have $X_n^{-1}(B_n) = \Omega \times R_1 \times \cdots \times R_{n-1} \times B_n \times R_{n+1} \cdots$, so that $\widehat{P}(X_n^{-1}(B)) = P(\Omega)\mu_1(R_1) \cdots \mu_{n-1}(R_{n-1})\mu_n(B)\mu_{n+1}(R_{n+1}) \times \cdots = \mu_n(B)$

Example 1.3.4. *Let* $X : (\Omega, \mathcal{F}, P) \to (R, \mathcal{R})$ *be a random variable. If* $(S, \mathcal{S})$ *is a measure space and* $f : R \to S$ *is a measurable function, then* $Y = f(X) : (\Omega, \mathcal{F}, P) \to (S, \mathcal{S})$ *is a random variable. This can be generalized to function of several variables. If* X_1, X_2 *are random variables, then* $Y = X_1 + X_2$ *are random variables (as long as* X_1 *and* X_2 *take values in a space where addition is defined). If* $\{X_n\}_{n \in \mathbb{N}}$ *is a sequence of random variables with each* X_n *taking values in* $\{0, 1\}$*, then* $Y_n = \frac{X_1 + \cdots + X_n}{n}$ *is a random variable.*

Example 1.3.5. *A discrete, real and complex random variable is one with values in the measurable spaces* $(\mathbb{Z}, \mathcal{P}(\mathbb{Z}))$*,* $(\mathbb{R}, \mathcal{P}(\mathbb{R}))$ *and* $(\mathbb{C}, \mathcal{P}(\mathbb{C}))$ *respectively. In each of these cases, we are interested in asking* $P(x = n)$*,* $P(X \in (a, b])$ *and* $P(|X - c| < \varepsilon)$ *respectively, for some* $n \in \mathbb{N}$*,* $a, b, \varepsilon \in \mathbb{R}$ *and* $c \in \mathbb{C}$*.*

Example 1.3.6. *Random vectors (special cases of random matrices)* $X = (X_1, ..., X_n)$ *are vectors whose entries are random variables. We are typically concerned with the case the values of* X *are in* $\mathbb{R}^n$ *or* $\mathbb{C}^n$*. More generally, we can construct random sequences* $X = (X_1, X_2, ...)$*. There are also notions of random functions and random measures (related to "point processes") that we will not discuss further here.*

1.4 Real Random Variables

As stated above, a real random variable is a random variable $X : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$. We define the probability distribution function of $X : \mathbb{R} \to [0, 1]$ to be the function $F_X(t) = \mu_X((-\infty, t]) = P(X \leqslant t)$. Note the law defined earlier is a measure, whereas the probability distribution function is a function on $\mathbb{R}$ by evaluating the law on a particular subset of $\mathbb{R}$. We make the following remarks.

- F_x is monotonically increasing.
- $P(X \in (a, b]) = \mu_X((a, b]) = F_X(b) - F_X(a)$
- $\lim_{t \to -\infty} F_X(t) = 0$ and $\lim_{t \to \infty} F_X(t) = 1$
- Continuity From The Right: $\lim_{t \to a^+} F_X(t) = F_X(a)$. Pf. Let $\{t_n\}_{n \in \mathbb{N}} \subseteq \mathbb{R}$ be an arbitrary non increasing sequence such that $t_n \to a^+$. Thus $(a, t_n] \searrow \varnothing$ so that $\bigcap_{n \in \mathbb{N}} (a, t_n] = \varnothing$. Then $\mu_X((a_n, t_n]) \to \mu_X(\varnothing) = 0$, so that $F_X(t_n) - F_X(a) = \mu_X((a, t_n]) \to 0$ as $t_n \to a^+$. Hence $F_X(t_n) \to F_X(a)$ as $t_n \in a^+$.
- $P(X = a) = \mu_X(\{a\}) = F_X(a) - \lim_{t \to a^-} F_X(t)$. Pf. Observe $F_X(a) - \lim_{t \to a^-} F_X(t) = \lim_{t \to a^-} (F_X(a) - F_X(t)) = \lim_{t \to a^-} \mu_X((t, a])$. We can find $\{t_n\}_{n \in \mathbb{N}}$ non decreasing such that $t_n \to a^-$. Thus $(t_n, a] \nearrow \{a\}$ so that $\bigcap_{n \in \mathbb{N}} (t_n, a] = \{a\}$. Thus $\mu_X((t_n, a]) \to \mu_X(\{a\})$.

We want to reverse this process and obtain a random variable from a function that has some of these properties. The following theorem answers this question.

Theorem 1.4.1. *Let* $F : \mathbb{R} \to [0, 1]$ *be monotonically increasing, right continuous,* $\lim_{t \to -\infty} F_X(t) = 0$ *and* $\lim_{t \to \infty} F_X(t) = 1$*. Then there exists a random variable* X *such that its probability distribution function is* $F_X = F$*.*

Before sketching the proof of this theorem, we first prove a preliminary lemma.

Lemma 1.4.1. *Let* F *be as in the theorem statement and* $\mathcal{F}_0$ *the algebra of finite disjoint unions of* h*-intervals on* $\mathbb{R} \cup \{\infty\}$ *with premasure* $\mu\left(\bigcup_{n=1}^N (a_n, b_n]\right) = \sum_{n=1}^N F(b_n) - F(a_n)$*. For all* $\varepsilon > 0$ *and for all* $A \in \mathcal{F}_0$*, there exists* $B \in \mathcal{F}_0$ *such that the closure of* B *is contained in* A *and* $\mu(A) \leqslant \mu(B) + \varepsilon$*.*

Before proving this lemma, we note $\mathcal{F}_0$ is not contained in $\mathbb{R}$, and that the proof of the theorem before the lemma actually constructs a random variable $\widehat{X} \in (\mathbb{R} \cup \{\infty\}, \mathcal{B}(\mathbb{R} \cup \{\infty\}))$. However, we can simply treat this space and $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ as the same by taking $F(\infty) = 1$.

Proof. Without loss of generality let A be the union of $\{(a_n, b_n]\}_{n=1,\dots,N}$. Let B be the union of the intervals $\{[a_n + \delta, b_n]\}_{n=1,\dots,k}$ for sufficiently small δ. Then the closure of B is contained in A and $\mu(A) - \mu(B) = \mu(A/B) = \sum_{n=1}^{k} \mu((a_n, a_n + \delta]) = \sum_{n=1}^{k} F(a_n + \delta) - F(a_n) \to 0$ as $\delta \to 0$ by the right continuity of F. In other words, for all $\varepsilon > 0$ there exists $\delta > 0$, which is equivalent to finding a $B \in \mathcal{F}_0$, such that $\mu(A) - \mu(B) < \varepsilon$. □

We now sketch the proof of the theorem.

Proof. We want to show there exists a probability measure μ on $\mathbb{R}$ such that $F(t) = \mu((-\infty, t])$. Consider the algebra $\mathcal{F}_0$ on $\mathbb{R} \cup \{\infty\}$ given by finite disjoint unions of h-intervals with premeasure $\mu\left(\bigcup_{n=1}^{k} (a_n, b_n]\right) = \sum_{n=1}^{k} F(b_n) - F(a_n)$. To show μ is a premeasure we show it is σ-additive (showing $\mu(\varnothing) = 0$ is left as an exercise). To do this, let $\{A_n\}_{n \in \mathbb{N}}$ be a sequence of sets from $\mathcal{F}_0$. By our previous proposition, we instead show that if $A_n \searrow \varnothing$ where each $A_n \in \mathcal{F}_0$, then $\mu(A_n) \to 0$.

Case 1: Assume each $A_n \subseteq [-M, M]$ for some $M \in \mathbb{R}_{\geqslant 0}$. For each n, for all $\varepsilon > 0$, we have by our lemma there exists B_n such that the closure of B_n is contained in A_n and $\mu(A_n) \leqslant \mu(B_n) + \frac{\varepsilon}{2^n}$. Denote the closure of B_n by $\overline{B}_n$. Then $\{\overline{B}_n\}_{n \in \mathbb{N}}$ is a sequence of compact sets such that $\overline{B}_n \searrow \varnothing$, so that $\bigcap_{n \in \mathbb{N}} \overline{B}_n = \varnothing$. Writing $\bigcap_{n \in \mathbb{N}} \overline{B}_n = \overline{B}_1 \cap (\overline{B}_2 \cap \overline{B}_1) \cap (\overline{B}_3 \cap \overline{B}_2 \cap \overline{B}_1) \cap \cdots$, we have by the compactness of each $\overline{B}_n$ that there exists $N \in \mathbb{N}$ such that $\overline{B}_1 \cap \cdots \cap \overline{B}_N = \varnothing$. Thus for all $n \geqslant N$ we have $A_n = A_n / (\overline{B}_1 \cap \cdots \cap \overline{B}_N) = \bigcup_{i=1}^{N} A_n / B_i \subseteq \bigcup_{i=1}^{N} A_i / \overline{B}_i$. Since μ is additive, one can show it is monotonic, so that $\mu(A_n) \leqslant \varepsilon\left(\frac{1}{2} + \cdots + \frac{1}{2^N}\right) = \varepsilon$. Hence, for all $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that $\mu(A_n) < \varepsilon$ for all $n \geqslant N$.

Case 2: Let each $A_n \in \mathcal{F}_0$ be as assumed before Case 1. Given $\varepsilon > 0$, we have by definition of $lim_{t \to -\infty} F_X(t) = 0$ and $lim_{t \to \infty} F_X(t) = 1$ there exists $M \in \mathbb{R}$ large enough so that $F(-M) < \varepsilon$ and $F(M) > 1 - \varepsilon$. Then for each $n \in \mathbb{N}$ we have

$$\mu(A_n) = \mu(A_n \cap (\mathbb{R} \cup \{\infty\})) = \mu(A_n \cap ((-\infty, -M] \cup (-M, M] \cup (M, \infty]))$$

$$\leqslant \mu(A_n \cap (-M, M]) + \mu(A_n \cap (-\infty, -M]) + \mu(A_n \cap (M, \infty])$$

$$\leqslant \mu(A_n \cap (-M, M]) + \mu((-\infty, -M]) + \mu((M, \infty])$$

$$\leqslant \mu(A_n \cap (-M, M]) + \mu((-\infty, -M]) + \mu((M, \infty]) = \mu(A_n \cap (-M, M]) + F(-M) + (1 - F(M))$$

$$\leqslant \mu(A_n \cap (-M, M]) + \varepsilon + \varepsilon$$

where the first term can be made arbitrarily small by Case 1.

The above shows μ is a premeasure, so by Caratheodory's Theorem we extend μ to a measure on $\sigma(\mathcal{F}_0) = \mathcal{B}(\mathbb{R} \cup \{\infty\})$. By our construction, $F(t) = \mu((-\infty, t])$ (where now μ has domain $\mathcal{B}(\mathbb{R} \cup \{\infty\})$. We also notice $\mu(\{\infty\}) = 0$ since $\mu(\mathbb{R}) = 1$. □

Remark: If $A \in \mathbb{R}$ and X is a real random variable, then $P(X \in A) = \mu_X(A) = \int_A d\mu_X = \int_{\mathbb{R}} \chi_A \, d\mu_X$.

Let μ be a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ and $F(t) = \mu((-\infty, t])$. Then we can uniquely decomposed as $\mu = \mu_d + \mu_{sc} + \mu_{ac}$, a sum of three specific probability measures (d, ac and sc are acronyms for discrete, absolutely continuous and singularly continuous respectively).

- We have $\mu_d(A) = \sum_i p_i \chi_A(x_i) = \sum_i p_i \delta_{x_i}(A)$, where each $x_i \in \mathbb{R}$, and $p_i = P(x_i \in A)$ is the probability of x_i occurring in the event A. Each $p_i \in [0,1]$ and $\sum_i p_i \in [0,1]$. $\delta_x(A) = 1$ if $x \in A$ and 0 if $x \notin A$ (the Dirac delta measure). The probability distribution function takes the form $F_d(x) = \sum_i p_i \chi_{[x_i,\infty)}(x)$. The collection of values for i can be countable or finite.

- We have μ_{ac} is absolutely continuous with respect to the Lebesgue measure m. In this case $d\mu_{ac} = \rho dm$ for some $\rho \in L^1(\mathbb{R})$ and $\rho \geqslant 0$. In other words $\mu_{ac}(A) = \int_A \rho dm$.

- μ_{sc} satisfies $\frac{d\mu_{sc}}{dm} = 0$, meaning there exists B such that $m(B) = 0$ and $\mu_{sc}(\mathbb{R}/B) = 0$. In other words, μ_{sc} lives on a set of Lebesgue measure zero (μ_d also satiosfies this property). The corresponding probability distribution function $F_{sc}(t) = \mu_{sc}((-\infty, t])$ can be shown to be continuous.

A **discrete random variable** X is one whose law has decomposition that satisfies $\mu_{ac} = \mu_{sc} = 0$. For some finite or countable set $S = \{x_n\}$ we have $\mu_X(\mathbb{R}/S) = 0$ and $P(X \in A) = \sum_{x_n \in A} P(X = x_n)$. Below we give some examples of well known discrete random variables. We typically define these random variables in terms of their laws as they are more relevant to computation.

Example 1.4.1. *Dirac Delta Distribution: The constant random variable $X = x_0 \in \mathbb{R}$ has law δ_{x_0}, also known as the Dirac delta distribution. We have $X = x_0$ almost surely, i.e., $P(X = x_0) = 1$. X is seen to be discrete intuitively since it takes values in the singleton set $\{x_0\}$.*

Example 1.4.2. *Uniform Distribution: Let $R = \{x_1, ..., x_r\} \subseteq \mathbb{R}$. We let X be such that $P(X = x_i) = \frac{1}{r}$ for each $i = 1, ..., r$. The associated distribution $\mu_X(x) = \sum_{i=1}^{r} \frac{\delta_{x_i}(x)}{r}$.*

Example 1.4.3. *Bernoulli Distribution With Parameter p: Let $X \in \{0,1\}$. We let $P(X = 0) = p$ and $P(X = 1) = 1 - p$.*

Example 1.4.4. *Geometric Distribution With Parameter p: Let $X \in \{0,1,2,...\}$. We let $p_k = P(X = k) = p(1-p)^k$, the probability of the "first hit" (independent Bernoulli trials until one gets the desired result for the first time). We see $\sum_{k \in \mathbb{N}} p_k = p\frac{1}{1-(1-p)} = 1$.*

Example 1.4.5. *Binomial Distribution With Parameter p: Let $X \in \{0,1,2,...,N\}$. We let $p_k = \binom{N}{k} p^k (1-p)^{N-k}$. This gives the probability of how many times we get $X = 0$ from N independent Bernoulli trials.*

Example 1.4.6. *Poisson Distribution With Parameter $\lambda \geqslant 0$: Let $X \in \{0,1,2,...\}$. We let $p_k = \frac{e^{-\lambda}\lambda^k}{k!}$. We immediately observe $\sum_{k \in \mathbb{N}_0} p_k = 1$. This is the probability of asking how many points land in $[0,1]$ after dropping them uniformly and independently on $\mathbb{R}$, with "density" λ.*

We now show $P(X_N = k) \to P(X = k)$ as $N \to \infty$, for each $k \in \mathbb{N}_0$, where X_N has a binomial distribution with parameter $\frac{\lambda}{N}$ and X has a Poisson distribution with parameter λ. Intuitively, drop N points on an interval of length $\frac{N}{\lambda}$. The probability of k points landing in $[0,1]$ has a binomial distribution with parameter $p = \frac{1}{N/\lambda} = \frac{\lambda}{N}$. Thus for fixed λ and k

$$p_k^{(N)} = \binom{N}{k} \left(\frac{\lambda}{N}\right)^k \left(1 - \frac{\lambda}{N}\right)^{N-k}$$

$$= \frac{1}{k!} \cdot \frac{N(N-1)\cdots(N-k+1)}{N^k} \cdot \lambda^k \left(\left(1 - \frac{\lambda}{N}\right)^N\right)^{1-k/N} \to \frac{1}{k!}(1)\left(e^{-\lambda}\right)$$

as $N \to \infty$.

A **continuous random variable** X is one whose law has decomposition that satisfies $\mu_d = \mu_{sc} = 0$. We have in this case μ_X is absolutley continuous with respect to the Lebesgue measure and $d\mu_X = \rho_X dm$ for some $\rho_X \in L^1(\mathbb{R})$ such that $\rho_X \geqslant 0$. Thus $\int_{\mathbb{R}} \rho_X dm = \mu_X(\mathbb{R}) = 1$ and $P(X \leqslant s) = F_X(s) = \mu_X((-\infty, s]) = \int_{-\infty}^{s} \rho_X dm$. We have that the probability density function of X is defined, and is given by ρ_X (exactly the Radon-Nikodym derivative). Below we give some examples of well known continuous random variables. We typically define these random variables in terms of their probability density functions as they are more relevant to computation.

Example 1.4.7. *Uniform Distribution on* $[a, b]$*: We define* $\rho(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \textit{otherwise} \end{cases}$.

Example 1.4.8. *Real Valued Normal/Gaussian Distribution With Mean* μ *and Variance* $\sigma^2 > 0$, $N(\mu, \sigma^2)$*: We define* $\rho(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$.

Example 1.4.9. *Cauchy Distribution With Parameter* γ*: We define* $\rho(x) = \frac{\gamma}{\pi(\gamma^2+x^2)}$. *The tails of the distribution decay too slow for us to define a notion of mean and variance, as opposed to a Gaussian distribution.*

Example 1.4.10. *Exponential Distribution With Parameter* $\lambda > 0$*: We define* $\rho(x) = \begin{cases} \frac{1}{\lambda} e^{-\lambda x} & x \geqslant 0 \\ 0 & x < 0 \end{cases}$.

Example 1.4.11. *Gamma Distribution With Parameter* $p, \lambda > 0$ *(Generalization of Exponential): We define* $\rho(x) = \begin{cases} \frac{1}{\Gamma(p)\lambda^p} x^{p-1} e^{-\lambda x} & x > 0 \\ 0 & x \leqslant 0 \end{cases}$.

The following theorem gives Skorokhod's construction of a random variable (in our case real valued) given a prescribed distribution function F.

Theorem 1.4.2. *Let* $F : \mathbb{R} \to [0, 1]$ *be non-decreasing and satisfy* $\lim_{t \to -\infty} F(t) = 0$, $\lim_{t \to \infty} F(t) = 1$ *and* $\lim_{t \to s^+} F(t) = s$. *Take* $(\Omega, \mathcal{F}, P) = ([0, 1], \mathcal{B}([0, 1], m)$. *Then then there exists a random variable* $X : ([0, 1], \mathcal{B}([0, 1], m) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ *such that* $F_X = F$.

Proof. (Sketch) If F is continuous and strictly increasing we let $X(t) = F^{-1}(t)$. In the more general case of F (where one can break down into the special case) we let $X(s) = \inf\{x \in \mathbb{R} \mid F(x) \geqslant s\}$ (infimum is achieved in the set by the right continuity of F) where $t \in [0, 1]$. One can show X is measurable. Observe $F_X(t) = \mu_X((-\infty, t]) = P(X^{-1}((-\infty, t])) = P(X \leqslant t) = m(\{s \in [0, 1] \mid X(s) \leqslant t\}) = m(\{s \in [0, 1] \mid F(t) \geqslant s\}) = m([0, F(t)]) = F(t)$. □

1.5 Random Vectors

We can think of a random vector either as a measurable function $X : (\Omega, \mathcal{F}, P) \to (\mathbb{R}^n, \mathcal{B}(\mathbb{R}))$ where $X(\omega) = (X_1(\omega), ..., X_n(\omega))$, or as a collection of n random variables where $X_k : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ for $k = 1, ..., n$ and $X = (X_1, ..., X_n)$. X is measurable if and only if each X_k is measurable. A random vector X gives rise to a probability measure $\mu_X(A) :- P(X^{-1}(A))$ for all $A \in \mathcal{B}(\mathbb{R})$, similarly to the one dimensional case. Its probability distribution function is given by $F_X(t_1, ..., t_n) = P(X_1 \leqslant t_1, ..., X_n \leqslant t_n) = \mu_X((-\infty, t_1] \times \cdots \times (-\infty, t_n])$. This is also known as the joint probability distribution of $X_1, ..., X_n$. We have that $F_{X_k}(t_k) = P(X_k \leqslant t_k) = \mu_{X_k}((-\infty, t_k]) = \mu_X(\mathbb{R} \times \cdots \times \mathbb{R} \times (-\infty, t_k] \times \mathbb{R} \times \cdots \times \mathbb{R})$ and $F_{X_k}(t_k) = \lim_{t_1 \to \infty} \cdots \lim_{t_{k-1} \to \infty} \lim_{t_{k+1} \to \infty} \cdots \lim_{t_n \to \infty} F_X(t_1, ..., t_n)$ for each $k = 1, ..., n$ (the second expression is also known as the marginal distribution function of X_k).

The classification/decomposition of probability measures on $\mathbb{R}^n$ is more complicated than that on $\mathbb{R}$. However, there is still a notion of absolutely continuous probability measures, where the probability density function ρ_X is defined, for $\rho_X \in L^1(\mathbb{R}^n)$ non-negative. The probability distribution function of X is given by $F(t_1, ..., t_n) = \int_{-\infty}^{t_1} \cdots \int_{-\infty}^{t_n} \rho_X(s_1, ..., s_n) ds_n \cdots ds_1$, where each integral is with respect to the Lebesgue measure.

Example 1.5.1. *Multivariate Normal/Gaussian Distribution* $N(\mu, \Sigma)$*: In this case* $\mu \in \mathbb{R}^n$ *and* Σ *is an* $n \times n$ *real symmetric (so that it is diagonalizable) and strictly positive definite matrix. I.e.,* $\Sigma^T = \Sigma$ *and* $x^T \Sigma x > 0$ *for all* $x \in \mathbb{R}^n/\{0\}$*. Let* $t = (t_1, ..., t_n)^T$*. We define*

$$\rho(t_1, ..., t_n) = \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}(t-\mu)^T \Sigma^{-1}(t-\mu)\right)$$

$$= \frac{1}{(2\pi)^{n/2}\sqrt{\det \Sigma}} \exp\left(-\frac{1}{2}\sum_{j,k=1}^{n}(t_j - u_j)(t_k - \mu_k)(\Sigma^{-1})_{jk}\right)$$

where $(\Sigma^{-1})_{jk}$ *denotes the* j, k *entry of* Σ^{-1}*. In the special case* $\Sigma = \begin{pmatrix} \sigma_1^2 & & \\ & \ddots & \\ & & \sigma_n^2 \end{pmatrix}$ *(with non-diagonal entries equal to zero) and each* $\sigma_k^2 > 0$ *we have*

$$\rho(t_1, ..., t_n) = \Pi_{k=1}^n \frac{1}{\sqrt{2\pi\sigma_k^2}} \exp\left(-\frac{(t_k - \mu_k)}{2\sigma_k^2}\right)$$

In the more general case of Σ*, we have* $\Sigma = U^T D U$ *where* D *is a diagonal matrix and* U *is a real orthogonal matrix. If one defines* $Y = UX$ *where* X *is a multivariate Gaussian distribution with parameters* $\mu \in \mathbb{R}^n$ *and* Σ*, then* Y *is a multivariate Gaussian distribution with parameters* $U\mu$ *and* D*.*

We give some properties of the joint probability distribution function $F_X(t_1, ..., t_n) = P(X_1 \leqslant t_1, ..., X_n \leqslant t_n)$.

- $\lim_{t_1 \to \infty} \cdots \lim_{t_n \to \infty} F_X(t_1, ..., t_n) = 1$
- $\lim_{t_k \to -\infty} F_X(t_1, ..., t_n) = 0$ for each $k = 1, ..., n$
- $F_n(s_1, ..., s_n) \to F(t_1, ..., t_n)$ whenever $s_k \to t_k^+$ for each $k = 1, ..., n$.

$F_X(t_1, ..., t_n)$ is not only non-decreasing but satisfies something stronger. This stronger property is required in addition to the above for a prescribed function F to give a random vector, while we do not need this extra property in the one dimensional case. For the case $n = 2$ this property says $0 \leqslant P(X_1 \in (a_1, b_1], X_2 \in (a_2, b_2]) = F(b_1, b_2) - F(a_1, b_2) - F(b_1, a_2) + F(a_1, a_2)$ for all $a_1, b_1, a_2, b_2 \in \mathbb{R}$ such that $a_1 < b_1, a_2 < b_2$. At a high level, this says we can separate the rectangles representing this quantity accordingly. For an arbitrary $n \in \mathbb{N}$, we define the operator

$$\Delta_{a_k, b_k}^k : \Phi(t_1, ..., t_n) \to \Psi(t_1, ..., t_{k-1}, t_{k+1}, ..., t_n)$$

given by $\Phi(t_1, ..., t_{k-1}, b_k, t_{k+1}, ..., t_n) - \Phi(t_1, ..., t_{k-1}, a_k, t_{k+1}, ..., t_n)$. In this case, the stronger property says $0 \leqslant P(X_1 \in (a_1, b_1], ..., X_n \in (a_n, b_n]) = \Delta_{a_1, b_1}^1 \cdots \Delta_{a_n, b_n}^n F_X$ (this is a sum difference of 2^n terms).

Theorem 1.5.1. *Let* $F : \mathbb{R}^n \to [0, 1]$ *have the above properties. Then there exists a random vector* X *with probability distribution function* $F_X = F$*.*

The proof defines $P\left(\Pi_{k=1}^n (a_k, b_k]\right) = \Delta_{a_1, b_1}^1 \cdots \Delta_{a_n, b_n}^n F$.

1.6 Independence

Given a probability space $(\Omega, \mathcal{F}, P)$, two events $A_1, A_2 \in \mathcal{F}$ are said to be **independent** if $P(A_1 \cap A_2) = P(A_1)P(A_2)$. Let J be an index set. A collection of events $\{A_j\}_{j\in J}$, where each $A_j \in \mathcal{F}$, are called **mutually independent** if for all $k \in \mathbb{N}$, for all j_i where $i = 1, ..., k$,

$$P(A_{j_1} \cap \cdots \cap A_{j_k}) = P(A_{j_1}) \cdots P(A_{j_k})$$

A collection of events $\{A_j\}_{j\in J}$ is said to be k**-independent** if the above equality holds for fixed $k \in \mathbb{N}$.

Example 1.6.1. *Two Coin Tossings: Take* $\Omega = \{0,1\} \times \{0,1\}$, $\mathcal{F} = \mathcal{P}(\Omega)$ *and* $P(A) = \frac{|A|}{4}$. *Consider the events* $A = \{(0,0),(0,1)\}$, $B = \{(0,1),(1,1)\}$ *and* $C = \{(0,0),(1,1)\}$. *Then* $\{A, B, C\}$ *is* 2*-independent but not mutually independent since* $|A \cap B| = |A \cap C| = |B \cap C| = 1$ *but* $|A \cap B \cap C| = 0$.

We say two σ-algebras $\mathcal{F}_1, \mathcal{F}_2 \subseteq \mathcal{F}$, where $(\Omega, \mathcal{F}, P)$ is a probability space, are **mutually independent** if $P(A_1 \cap A_2) = P(A_1)P(A_2)$ for all $A_1 \in \mathcal{F}_1$ and for all $A_2 \in \mathcal{F}_2$.

Example 1.6.2. *Take* $\Omega = \mathbb{R}^2$, $\mathcal{F} = \mathcal{B}(\mathbb{R}^2)$, $\mathcal{F}_1 = \{A_1 \times \mathbb{R} \mid A_1 \in \mathcal{B}(\mathbb{R})\}$ *and* $\mathcal{F}_2 = \{\mathbb{R} \times A_2 \mid A_2 \in \mathcal{B}(\mathbb{R})\}$. *Then* $\mathcal{F}_1$ *and* $\mathcal{F}_2$ *are mutually independent if for all* $A_1 \in \mathcal{F}_1$ *and for all* $A_2 \in \mathcal{F}_2$, $P(A_1 \cap A_2) = P(A_1 \times \mathbb{R})P(\mathbb{R} \times A_2) = P_1(A_1)P_2(A_2)$ *where* P_1 *and* P_2 *are restrictions of* P *to* $\mathcal{F}_1$ *and* $\mathcal{F}_2$ *respectively. In other words, if* P *is the product measure of some* P_1 *and* P_2.

Similarly to the above conditions, a collection of σ-algebras $\{\mathcal{F}_\omega\}_{\omega\in\Omega}$ is mutually independent. Let $X_1, ..., X_n$ be random variables such that each $X_k : (\Omega, \mathcal{F}, P) \to (R, \mathcal{R}_k)$. We say the random variables $X_1, ..., X_n$ are mutually independent if the sigma algebras $\{\mathcal{F}_1, ..., \mathcal{F}_n\}$ are mutually independent, where each $\mathcal{F}_k = \{X_k^{-1}(A_k) \mid A_k \in \mathcal{R}_k\}$. In other words, for every choice of $A_1 \in \mathcal{R}_1, ..., A_n \in \mathcal{R}_n$ we have $P(X_1^{-1}(A_1) \cap \cdots \cap X_n^{-1}(A_n)) = \Pi_{k=1}^n P(X_k^{-1}(A_k))$. In other words, $P(X_1 \in A_1 \text{ and } \cdots \text{ and } X_n \in A_n) = \Pi_{k=1}^n P(X_k \in A_k)$.

Theorem 1.6.1. *A finite number of real random variables* $X_1, ..., X_n$ *are mutually independent if and only if* $F_X(t_1, ..., t_n) = \Pi_{k=1}^n F_{X_k}(t_k)$ *(the joint probability distribution is a product of the marginal probability distributions).*

Proof. Observe $F_X(t_1, ..., t_n) = P(X_1 \leqslant t_1, ..., X_n \leqslant t_n)$ and $F_{X_k}(t_k) = P(X_k \leqslant t_k)$, where $X_k \leqslant t_k$ if and only if $X_k \in (-\infty, t_k]$. Thus for the "only if" direction we take $A_k = (-\infty, t_k]$ in the definition. For the "if" direction, we sketch the proof in the case $n = 2$. Assume $F_X(t_1, t_2) = F_{X_1}(t_1)F_{X_2}(t_2)$. In other words, $P(X_1 \leqslant t_1, X_2 \leqslant t_2) = P(X_1 \leqslant t_1)P(X_2 \leqslant t_2)$.

Step 1. Let $(a_1, b_1], (a_2, b_2] \subseteq \mathbb{R}$. We want to show $P(X_1 \in (a_1, b_1], X_2 \in (a_2, b_2]) = P(X_1 \in (a_1, b_1])P(X_2 \in (a_2, b_2])$. But we see

$$\begin{aligned} P(X_1 \in (a_1, b_1], X_2 \in (a_2, b_2]) &= F_X(b_1, b_2) - F_X(a_1, b_2) - F_X(b_1, a_2) + F_X(a_1, a_2) \\ &= F_{X_1}(b_1)F_{X_2}(b_2) - F_{X_1}(a_1)F_{X_2}(b_2) - F_{X_1}(b_1)F_{X_2}(a_2) + F_{X_1}(a_1)F_{X_2}(a_2) \\ &= (F_{X_1}(b_1) - F_{X_1}(a_1))(F_{X_2}(b_2) - F_{X_2}(a_2)) = P(X_1 \in (a_1, b_1])P(X_2 \in (a_2, b_2]) \end{aligned}$$

Step 2. Let $A_1, A_2 \in \mathcal{F}_0$ where $\mathcal{F}_0$ is the algebra of finite disjoint unions of h-intervals. One can show $P(X_1 \in A_1, X_2 \in A_2) = P(X_1 \in A_1)P(X_2 \in A_2)$.

Step 3. Extend to $A_1 \in \mathcal{F} = \sigma(\mathcal{F}_0) = \mathcal{B}(\mathbb{R})$ for a fixed $A_2 \in \mathcal{F}_0$. If $P(X_2 \in A_2) = 0$ then $P(X_1 \in A_1, X_2 \in A_2) \leqslant P(X_2 \in A_2) = 0$ so that $P(X_1 \in A_1, X_2 \in A_2) = 0$ and we are done with Step 3. If $P(X_2 \in A_2) \neq 0$, then define

$$\mu(A_1) :- \frac{P(X_1 \in A_1, X_2 \in A_2)}{P(X_2 \in A_2)} \text{ and } \tilde{\mu}(A_1) = P(X_1 \in A_1)$$

Then μ and $(\tilde{\mu})$ are probability measures and $\mu(A_1) = (\tilde{\mu})(A_1)$ for all $A_1 \in \mathcal{F}_0$. Then $\mu(A_1) = (\tilde{\mu})(A_1)$ for all $A_1 \in \mathcal{F}$ by the uniqueness of Caratheodory's Theorem.

Step 4. Similarly extend to $A_1, A_2 \in \mathcal{F}$ for a fixed $A_1 \in \mathcal{F}$. □

Corollary 1.6.1. *Let X_1, X_2 be real random variables and $X = (X_1, X_2)$. (a) If X_1 and X_2 are both absolutely continuous and independent, then X is absolutely continuous and $\rho_X(t_1, t_2) = \rho_{X_1}(t_1)\rho_{X_2}(t_2)$. (b) If X is absolutely continuous with probability density function $\rho_X(t_1, t_2) = \rho_1(t_1)\rho_2(t_2)$ (up to re-normalization constants), then X_1 and X_2 are both absolutely continuous and independent, with probability density functions $\rho_1(t_1)$ and $\rho_2(t_2)$ respectively.*

Proof. Relies on the fact $F_X(t_1, t_2) = \int_{-\infty}^{t_1}\int_{-\infty}^{t_2} \rho_X(s_1, s_2)ds_2 ds_1$, $F_{X_1} = \int_{-\infty}^{t_1} \rho_{X_1}(s_1)ds_1$ and $F_{X_2} = \int_{-\infty}^{t_2} \rho_{X_2}(s_2)ds_2$. □

Example 1.6.3. *If $X = (X_1, ..., X_n)$ has a multivariable normal distribution $N(\mu, \Sigma)$, then $X_1, ..., X_n$ are independent if and only if Σ is a diagonal matrix. This is because* $\exp\left(-\frac{1}{2}\sum_{j,k=1}^{n}(t_j - \mu_j)(t_k - \mu_k)\left(\Sigma^{-1}\right)_{j,k}\right)$ *can be split into an n product if and only if $\left(\Sigma^{-1}\right)_{j,k} = 0$ for each $j \neq k$, which is if and only if Σ^{-1} is diagonal, which is if and only if Σ is diagonal.*

Proposition 1.6.1. *If $X_1, ..., X_n$ are independent real random variables and $f_k : \mathbb{R} \to \mathbb{R}$ is measurable for $k = 1, ..., n$, then $f_k(X_k)$ are independent real random variables.*

Proof. Observe

$$P(f_1(X_1) \in A_1, ..., f_n(X_n) \in A_n) = P(X_1 \in f_1^{-1}(A_1), ..., X_n \in f_n^{-1}(A_n))$$

$$= \Pi_{k=1}^{n} P(X_k \in f^{-1}(A_k)) = \Pi_{k=1}^{n} P(f_k(X_k) \in A_k)$$

where the second equality is by the independence of $X_1, ..., X_n$. Each f_k needs to be measurable so that $P(X_k \in f^{-1}(A_k))$ makes sense. □

1.7 Expectation of a Real Random Variable

Given a real random variable $X : (\Omega, \mathcal{F}, P) \to (\mathbb{R}, \mathcal{B}(\mathbb{R}), \mu_X)$ one can (sometimes) define the **expectation** of X as $E(X) = \int_\Omega X(\omega)dP(\omega)$ (provided $X(\omega)$ is integrable on $(\Omega, \mathcal{F}, P)$). Some remarks are noted below.

- $E(X)$ can also be $\pm\infty$ or indeterminate if X is not integrable.
- $E(X)$ is independent of the extension of a probability space. I.e., if $(\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P})$ is an extension of $(\Omega, \mathcal{F}, P)$ with projection map $\pi : (\widehat{\Omega}, \widehat{\mathcal{F}}, \widehat{P}) \to (\Omega, \mathcal{F}, P)$, then $\int_\Omega X(\omega)dP(\omega) = \int_{\widehat{\Omega}}(X \circ \pi)(\widehat{\omega})d\widehat{P}(\widehat{\omega})$.

  This is shown with an approximation, starting with the case X is a simple function. I.e., of the form $X(\omega) = \sum_{k=1}^{n} a_k \chi_{B_k}(\omega)$. The result follows from linearity after noting that for the case X is a characteristic function χ_B, we have to show $\int_\Omega \chi_B(\omega)dP(\omega) = \int_{\widehat{\Omega}}(\chi_B \circ \pi)(\widehat{\omega})d\widehat{P}(\widehat{\omega})$. Since $\chi_B \circ \pi = \chi_{\pi^{-1}(B)}$ this amounts to showing $P(B) = \widehat{P}(\pi^{-1}(B))$, which we already showed to be true via Caratheordory's Theorem.
- If X is a real random variable and $f : \mathbb{R} \to \mathbb{R}$ is a measurable function, then $f(X)$ is a random variable such that $E(f(X)) = \int_\Omega f(X(\omega))dP(\omega)$.

Proposition 1.7.1. *Let X be a real random variable with law μ_X and $f : \mathbb{R} \to \mathbb{R}$ a measurable function. Then $E(f(X)) = \int_{\mathbb{R}} f(X)d\mu_X$ and $E(X) = \int_{\mathbb{R}} Xd\mu_X$*

Proof. Recall $E(f(X)) = \int_{\Omega} f(X(\omega))dP(\omega)$. The result is shown with an approximation, starting with the case f is a simple function, i.e., of the form $f = \sum_{k=1}^{n} a_k \chi_{B_k}$. The result follows from linearity after noting that for the case f is a characteristic function χ_B, we have to show $\int_{\Omega} \chi_B(X(\omega))dP(\omega) = \int_{\mathbb{R}} \chi_B(x)d\mu_X(x)$. But observe $\int_{\Omega} \chi_B(X(\omega))dP(\omega) = \int_{\Omega} \chi_{X^{-1}(B)}(\omega)dP(\omega) = P(X^{-1}(B))$ and $\int_{\mathbb{R}} \chi_B(x)d\mu_X(x) = \mu_X(B) = P(X^{-1}(B))$. The second equality follows by taking f to be the identity function. □

This proposition implies two special cases. If X is an absolutely continuous random variable with density ρ_X, then $E(f(X)) = \int_{\mathbb{R}} f(t)\rho_X(t)dt$. If X is a discrete random variable and $p_n = P(X = \lambda_n)$, (where n can be over a countable or finite set) we have $E(f(X)) = \sum_n p_n f(\lambda_n)$ (this is because $\mu_X = \sum_n p_n \delta_{\lambda_n}$).

The k^{th} **moment** of a random variable X is $m_k := E(X^k)$ (assuming it exists) where $k \in \mathbb{N}$. The **variance** of X is $Var(X) = \sigma^2(X) = E((X - E(X))^2)$. One can show $E(X_1 + X_2) = E(X_1) + E(X_2)$, $E(\lambda X) = \lambda E(X)$ and $E(\lambda) = \lambda$ for $\lambda \in \mathbb{R}$, where the first two follow from the linearity of integration (assuming all the expectations exist and the scaling and addition operation make sense for the values of the random variables). With these properties we observe $Var(X) = E((X - E(X))^2) = E(X^2 - 2XE(X) + E(X)^2) = E(X^2) = E(2XE(X)) + E(X)^2 = E(X^2) - 2E(X)E(X) + E(X)^2 = E(X^2) - E(X)^2$. Since $Var(X) \geqslant 0$ we have $E(X)^2 \leqslant E(X^2)$.

Example 1.7.1. *Recall a Cauchy random variable with $\gamma = 1$ has density $\rho(x) = \frac{1}{\pi(1+x^2)}$. We see $\int_{\mathbb{R}} \frac{x}{\pi(1+x^2)}dx$ does not exist so that $E(X)$ does not exist by the previous proposition.*

Example 1.7.2. *If X is a Bernoulli random variable, we have $P(X = 0) = 1 - p$ and $P(X = 1) = p$. Then $E(X) = 0(1 - p) + 1(p) = p$ and $Var(X) = (1 - p)p$.*

Example 1.7.3. *If X is a Poisson distribution with parameter λ, we have $P(X = n) = \frac{e^{-\lambda}\lambda^n}{n!}$ for $n = 0, 1, 2, ...$ Then $E(X) = \sum_{n\in\mathbb{N}_0} n \cdot \frac{e^{-\lambda}\lambda^n}{n!} = \lambda e^{-\lambda} \sum_{n\in\mathbb{N}_0} n \cdot \frac{\lambda^{n-1}}{n!} = \lambda e^{-\lambda} \sum_{n\in\mathbb{N}_0} \frac{\lambda^{n-1}}{(n-1)!} = \lambda e^{-\lambda} e^{\lambda} = \lambda$ and $Var(X) = \lambda$.*

Example 1.7.4. *Recall an exponential random variable has density $\rho(x) = \begin{cases} \frac{1}{\lambda}e^{-\lambda x} & x \geqslant 0 \\ 0 & x < 0 \end{cases}$. One can show $E(X) = \lambda = Var(X)$.*

Example 1.7.5. *Recall a normal random variable with parameters μ and σ^2 has density $\rho(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. Then $E(X) = \mu$ and $Var(X) = \sigma^2$ as the name suggests.*

It follows from the properties of expectation and variance that if $Y = aX + b$ for scalars a, b, then $E(Y) = aE(X) + b$ and $Var(Y) = a^2E(X)$. One can also show that if $Var(X) = 0$ then $X = E(X)$ almost surely.

Proposition 1.7.2. *If $X_1, ..., X_n$ are random variables, $X = (X_1, ..., X_n)$ and $Y = f(X_1, ..., X_n)$ is a measurable function, then*

$$E(Y) = \int_{\mathbb{R}^n} f(t_1, ..., t_n)d\mu_X(t_1, ..., t_n) = \underbrace{\int_{\mathbb{R}} \cdots \int_{\mathbb{R}}}_{n\text{-}times} f(t_1, ..., t_n)\rho_X(t_1, ..., t_n)dt_1 \cdots dt_n$$

If $(X_1, ..., X_n)$ is a random vector then one defines the **covariance matrix** as the $n \times n$ matrix with $\mathrm{Cov}(X_i, X_j) :- E((X_i - E(X_i))(X_j - E(X_j)))$ in the (i,j) entry.

Proposition 1.7.3. *Chebyshev's Inequality: Assume* $X \geqslant 0$ *is a real random variable and* $E(X)$ *exists. Then for all* $\varepsilon > 0$ *we have* $P(X \geqslant \varepsilon) \leqslant E(X)/\varepsilon$.

Proof. Observe

$$E(X) = \int_{\mathbb{R}} t d\mu_x(t) = \int_{\mathbb{R}} t\chi_{[\varepsilon,\infty)} + t\chi_{(-\infty,\varepsilon)} d\mu_X(t) \geqslant \int_{\mathbb{R}} \varepsilon\chi_{[\varepsilon,\infty)} + 0 d\mu_X(t)$$

$$\geqslant \varepsilon \int_{\mathbb{R}} \chi_{[\varepsilon,\infty)} d\mu_X(t) = \varepsilon P(X \geqslant \varepsilon)$$

□

Corollary 1.7.1. *Assume* X *is a real random variable,* $E(X)$ *exists and* $Var(X) < \infty$. *Then* $P(|X - E(X)| \geqslant \varepsilon) \leqslant Var(X)/\varepsilon^2$.

Proof. Apply Chebyshev's Inequality with the random variable $|X - E(X)|^2$ and tail bounds by ε^2, along with the fact $E(|X - E(X)|^2)/\varepsilon^2 = Var(X)/\varepsilon^2$. □

Proposition 1.7.4. *Jensen's Inequality: Let* $X : \Omega \to I$ *be a random variable where* $I \subseteq \mathbb{R}$ *is open. Let* $\varphi : I \to \mathbb{R}$ *be a convex function, i.e.,* $\varphi(tx + (1-t)y) \leqslant t\varphi(x) + (1-t)\varphi(y)$ *for all* $t \in [0,1]$ *and* $x, y \in I$. *If* $E(|X|), E(|\varphi(X)|) < \infty$ *then* $\varphi(E(X)) \leqslant E(\varphi(X))$.

Proof. Define $\Delta_{u,v}\varphi :- \frac{\varphi(u)-\varphi(v)}{u-v}$. By the convexity of φ we have that $\Delta_{u,v}\varphi$ is monotonically increasing in v for fixed u. Define $D_u^{\pm}\varphi = \lim_{v \to u^{\pm}} \Delta_{u,v}\varphi$ (one sided derivatives). The monotonicity of $\Delta_{u,v}\varphi$ implies $D_u^{-}\varphi \leqslant D_u^{+}\varphi$. Moreover $D_u^{\pm}\varphi$ monotonically increases in u. In particular for fixed $x_0 \in I$ and arbitrary $x \in I$, $\varphi(x) \geqslant \varphi(x_0) + (x - x_0)m$ for any $m \in [D_u^{-}\varphi(x_0), D_u^{+}\varphi(x_0)]$. Taking x in the previous inequality to be the random variable X and $x_0 = E(X)$, we have $\varphi(X) \geqslant \varphi(E(X)) + (X - E(X))m$. Taking expectations on both sides (which preserves the inequality by properties of integration) we obtain $E(\varphi(X)) \geqslant \varphi(E(X) + E(X - E(X))m = \varphi(E(X)) + (E(X) - E(X))m = \varphi(E(X))$ □

Proposition 1.7.5. *If* $X_1, ..., X_n$ *are independent random variables and* $f_1, ..., f_n$ *are measurable functions such that* $E(f_k(X_k))$ *exists for* $k = 1, ..., n$, *then* $E(f_1(X_1) \cdots f_n(X_n)) = \Pi_{k=1}^{n} E(f_k(X_k))$.

Proof. Observe

$$E(f_1(X_1) \cdots f_n(X_n)) = \int_{\mathbb{R}^n} f_1(t_1) \cdots f_n(t_n) d\mu_X(t_1, ..., t_n)$$

$$= \underbrace{\int_{\mathbb{R}} \cdots \int_{\mathbb{R}}}_{n\text{-times}} f(t_1) \cdots f(t_n) d\mu_1(t_1) \cdots d\mu_n(t_n) = \Pi_{k=1}^{n} \int_{\mathbb{R}} f_k(t_k) d\mu_{X_k}(t_k)$$

where the second equality is by independence, which allows μ_X to be expressed as a product measure. □

1.8 Convergence of Real Random Variables

Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables and X a random variable defined on $(\Omega, \mathcal{F}, P)$. We say $X_n \to X$ **almost surely** as $n \to \infty$ if $P(\{\omega \mid X_n(\omega) \to X(\omega)\}) = 1$. Note $\{X_n(\omega)\}$ will be a sequence of real numbers and $\{\omega \mid X_n(\omega) \to X(\omega)\} \subseteq \Omega$ but may not be the entire set Ω. We say $X_n \to X$ **in probability** if for all $\varepsilon > 0$, $P(|X_n - X| \geqslant \varepsilon) \to 0$ as $n \to \infty$. We say $X_n \to X$ in **r-mean** if $E(|X_n - X|^r) \to 0$ as $n \to \infty$. We say $X_n \to X$ **in distribution** if for all $f \in C_0^\infty(\mathbb{R})$ (the set of compactly supported C^∞ functions on $\mathbb{R}$) if $\int_{\mathbb{R}} f d\mu_{X_n} \to \int_{\mathbb{R}} f d\mu_X$ as $n \to \infty$. We make the following remarks.

- Convergence in distribution only depends on μ_{X_n} and μ_X, and not whether or not the X_n's and/or X are independent. The same if true for convergence in r-mean and in probability if X is constant.
- Convergence in probability is equivalent to the measure theory notion of convergence in measure
- Convergence in r-mean is equivalent to convergence in $L^r(\Omega, \mathcal{F}, P)$.

Let $C = \{\omega \mid X_n(\omega) \to X(\omega)\}$. To make sense of $P(C)$ so we can answer the question of almost sure convergence, we want $C \in \mathcal{F}$. Note $X_n(\omega) \to X(\omega)$ if and only if for all $m \in \mathbb{N}$ there exists $N \in \mathbb{N}$ such that for all $n \geqslant N$ we have $|X_n(\omega) - X(\omega)| < \frac{1}{m}$. Thus $C = \bigcap_{m=1}^\infty \bigcup_{N=1}^\infty \bigcap_{n=N}^\infty \{\omega \mid |X_n(\omega) - X(\omega)| < \frac{1}{m}\}$. Thus $C \in \mathcal{F}$ if $\{\omega \mid |X_n(\omega) - X(\omega)| < \frac{1}{m}\} \in \mathcal{F}$ (since $\mathcal{F}$ is closed under σ-algebra operations). It suffices to show $\{\omega \mid |X_n(\omega) - c| < \varepsilon\}$ and $\{\omega \mid |X(\omega) - c| < \varepsilon\}$ for all $\varepsilon > 0$, $c \in \mathbb{R}$ and $n \in \mathbb{N}$.

Proposition 1.8.1. *(Strong Law of Large Numbers): Let $\{X_k\}_{k\in\mathbb{N}}$ be a sequence of independent real random variables and $S_n = X_1 + \cdots + X_n$. If each $E(X_k) < \infty$ and $E(|X_k|^4) \leqslant C \in \mathbb{R}$ then $\frac{S_n - E(S_n)}{n} \to 0$ almost surely.*

Before giving the proof, we remark that if $\{X_n\}_{n\in\mathbb{N}}$ were independent and identically distributed (i.i.d) with their distribution being that of a random variable X, then $\frac{S_n}{n} \to E(X)$ almost surely since we see $E(X) = \frac{1}{n}E(X_1 + \cdots + X_n)$ (where each $E(X_k) = E(X)$ for $1 \leqslant k \leqslant n$).

Proof. Without loss of generality assume each $E(X_k) = 0$ (otherwise we work with $\tilde{X}_k = X_k - E(X_k)$). Then observe

$$E(S_n^4) = E((X_1 + \cdots + X_n)^4)$$
$$= E(X_1^4 + \cdots + X_n^4) + 6E(X_1^2X_2^2 + X_1^2X_3^2 + \cdots + X_{n-1}^2X_n^2) + E(X_1X_2\cdots X_n) + E(X_1^3X_2 + \cdots)$$

The last two terms will vanish because by independence they will turn into a product of expectations, some being of the form $E(X_k) = 0$. Thus

$$E(S_n^4) = E(X_1^4 + \cdots + X_n^4) + 6E(X_1^2X_2^2 + X_1^2X_3^2 + \cdots + X_{n-1}^2X_n^2)$$
$$= E(X_1^4) + \cdots + E(X_n^4) + 6E(X_1^2)E(X_2^2) + 6E(X_1^2)E(X_3^2) + \cdots + 6E(X_{n-1}^2)E(X_n^2)$$
$$\leqslant nC + \frac{6n(n-1)}{2}C = 3n^2C$$

where the second equality is since each $E(X_j^2X_k^2) = E(X_j^2)E(X_k^2)$ by independence, and the inequality is since each $E(X_k^2)^2 \leqslant E(X_k^4) \leqslant C$ by Jensen's inequality taking $\varphi(t) = t^2$. The above implies

$E\left(\frac{S_n^4}{n^4}\right) \leqslant \frac{3C}{n^2}$. Observe $\sum_{n=1}^{\infty} \frac{S_n^4}{n^4} \geqslant 0$ and $E\left(\sum_{n=1}^{\infty} \frac{S_n^4}{n^4}\right) = \sum_{n=1}^{\infty} E\left(\frac{S_n^4}{n^4}\right) \leqslant \sum_{n=1}^{\infty} \frac{3C}{n^2} < \infty$. Thus $P\left(\sum_{n=1}^{\infty} \frac{S_n^4}{n^4} = \infty\right) = 0$ so that $\sum_{n=1}^{\infty} \frac{S_n^4}{n^4} < \infty$ almost surely. Therefore $\frac{S_n^4}{n^4} \to 0$ almost surely. which implies $\frac{S_n}{n} \to 0$ almost surely. This along with the fact $E(S_n) = 0$ (since each $E(X_k) = 0$ by assumption) gives the result. □

Proposition 1.8.2. *(Kolomogorov's Strong Law of Large Numbers) Let $\{X_k\}_{k \to \mathbb{N}}$ be a sequence of i.i.d real random variables, each with distribution corresponding to a random variable X, and $S_n = X_1 + \cdots + X_n$. (a) If $E(|X|) < \infty$ then $\frac{S_n}{n} \to E(X)$ almost surely. (b) If $\frac{S_n}{n} \to C$ then $E(|X|) < \infty$ and $E(X) = C$. (c) If $E(|X|) = \infty$ then $\limsup_{n\to\infty} \frac{|S_n|}{n} \to \infty$ almost surely.*

Proposition 1.8.3. *(Weak Law of Large Numbers): Let $\{X_k\}_{k\in\mathbb{N}}$ be a sequence of independent real random variables and $S_n = X_1 + \cdots + X_n$. If each $E(X_k) < \infty$ and $Var(X_k) \leqslant C \in \mathbb{R}$ (bounded by C uniformly for all $k \in \mathbb{N}$) then $\frac{S_n - E(S_n)}{n} \to 0$ in probability. In other words, for all $\varepsilon > 0$ we have $P\left(\left|\frac{S_n - E(S_n)}{n}\right| \geqslant \varepsilon\right) \to 0$ as $n \to \infty$.*

Proof. Without loss of generality assume each $E(X_k) = 0$ so that $E(S_n) = 0$. By our corollary of Chebyshev's inequality we have

$$P\left(\left|\frac{S_n - E(S_n)}{n}\right| \geqslant \varepsilon\right) \leqslant \frac{1}{\varepsilon^2} Var\left(\frac{S_n - E(S_n)}{n}\right)$$

$$= \frac{1}{\varepsilon^2} E\left(\frac{S_n^2}{n^2}\right) = \frac{1}{\varepsilon^2 n^2} E\left(S_n^2\right) = \frac{1}{\varepsilon^2 n^2} E((X_1 + \cdots + X_n)^2)$$

$$= \frac{1}{\varepsilon^2 n^2} E(X_1^2 + \cdots + X_n^2) = \frac{1}{\varepsilon^2 n^2} Var(X_1^2) + \cdots + Var(X_n^2)$$

$$\leqslant \frac{1}{\varepsilon^2 n^2} \cdot nC = \frac{C}{\varepsilon^2 n} \to 0$$

as $n \to \infty$. The second equality is since $E(Sn) = 0$. The fifth equality is since after expanding and applying linearity, the expectation will have terms of the form $E(X_k X_j) = E(X_k)E(X_j) = 0$, leaving terms only of the form $E(X_k^2)$. □

Theorem 1.8.1. *Convergence almost surely implies convergence in probability. Convergence in r-mean implies convergence in probability. Convergence in probability implies convergence in distribution.*

Proof. We first show convergence almost surely implies convergence in probability. Assume $X_n \to X$ almost surely. Let $A_n^m = \{\omega \in \Omega : |X_n(\omega) - X(\omega)| < \frac{1}{m}\}$. Thus $P(X_n \to X) = P\left(\bigcap_{m\in\mathbb{N}} \bigcup_{N\in\mathbb{N}} \bigcap_{n\geqslant N} A_n^m\right)$. Therefore, $P(X_n \to X) = 1$ if and only if $P\left(\bigcup_{m\in\mathbb{N}} \bigcap_{N\in\mathbb{N}} \bigcup_{n\geqslant N} A_n^m\right) = 0$, which is if an only if for all $m \in \mathbb{N}$

$$P\left(\bigcup_{N\in\mathbb{N}} \bigcap_{n\geqslant N} A_n^m\right) = 1 \iff P\left(\bigcap_{N\in\mathbb{N}} \bigcup_{n\geqslant N} (A_n^m)^c\right) = 0$$

Define $B_N^m = \bigcup_{n\geqslant N} (A_n^m)^c$ so that the sequence $\{B_N^m\}$ is descending in N. By continuity of measures we have the above holds if and only if

$$\lim_{N\to\infty} P(B_N^m) = 0 \iff \lim_{N\to\infty} P\left(\bigcup_{n\geqslant N} (A_n^m)^c\right) = 0$$

which implies (in one direction)

$$\lim_{N\to\infty} P((A_N^m)^c) = 0 \iff \lim_{N\to\infty} P\left(\{\omega \in \Omega : |X_N(\omega) - X(\omega)| \geqslant \frac{1}{m}\}\right) = 0$$

Since $m \in \mathbb{N}$ was arbitrary we have $X_n \to X$ in probability as well, by definition.

We now show convergence in r-mean implies convergence in probability. Assume $X_n \to X$ in r-mean. Thus $E(|X_n - X|^r) \to 0$ as $n \to \infty$. By Chebyshev's inequality we have for all $\varepsilon \geqslant 0$ that $P(|X_n - X|^r \geqslant \varepsilon^r) \leqslant \frac{1}{\varepsilon^r} E(|X_n - X|^r) \to 0$. Since $\varepsilon > 0$ was arbitrary we have $X_n \to X$ in probability as well, by definition.

Now we show convergence in probability implies convergence in distribution. Let f be a compactly supported continuous function on $\mathbb{R}$ (so f could also be an arbitrary compactly supported infinitely differentiable function on $\mathbb{R}$), with support $[-M, M]$. Hence f is uniformly continuous on $[-M, M]$ and $|f(x)| < C \in \mathbb{R}$ for all $x \in [-M, M]$. Thus $\varepsilon > 0$, there exists $\delta > 0$ such that $|f(x) - f(y)| < \varepsilon$ whenever $|x - y| < \delta$, for all $x, y \in [-M, M]$. Then

$$\left|\int f d\mu_{X_n} - \int f d\mu_X\right| = |E(f(X_n)) - E(f(X))| \leqslant E(|f(X_n) - f(X)|)$$

$$\leqslant E(|f(X_n) - f(X)|\chi_{|X_n - X| < \delta}) + E(|f(X_n) - f(X)|\chi_{|X_n - X| \geqslant \delta})$$

$$= \int_\Omega |f(X_n) - f(X_n)|\chi_{|X_n - X| < \delta} dP + \int_\Omega |f(X_n) - f(X_n)|\chi_{|X_n - X| \geqslant \delta} dP$$

$$\leqslant \varepsilon \int_\Omega \chi_{|X_n - X| < \delta} dP + 2C \int_\Omega \chi_{|X_n - X| \geqslant \delta} dP$$

$$\leqslant \varepsilon + 2CP(|X_n - X| \geqslant \delta)$$

where the first inequality is by linearity and Jensen's inequality for $\varphi(t) = |t|$. The second term after the last inequality converges to 0 as $n \to \infty$ by assumption of $X_n \to X$ converging in probability. Thus

$$\limsup_{n\to\infty} \left|\int f d\mu_{X_n} - \int f d\mu_X\right| \leqslant \limsup_{n\to\infty} \varepsilon + 2CP(|X_n - X| \geqslant \delta)$$

Since $2CP(|X_n - X| \geqslant \delta) \to 0$ as $n \to \infty$ as $n \to \infty$ by assumption of $X_n \to X$ in probability, we have

$$0 \leqslant \limsup_{n\to\infty} \left|\int f d\mu_{X_n} - \int f d\mu_X\right| \leqslant \varepsilon \to 0$$

as $\varepsilon \to 0$ (which can be done since $\varepsilon > 0$ was arbitrary). By definition $X_n \to X$ in distribution as well. □

Theorem 1.8.2. *Let $\{X_n\}_{n\in\mathbb{N}}$ and X be random variables with distribution functions F_n (for each X_n) and F respectively, and laws μ_{X_n} (for each X_n) and μ_X respectively. Then the following are equivalent:*

1. *$\int f d\mu_{X_n} \to \int f d\mu_X$ for all continuous compactly supported functions f on $\mathbb{R}$.*
2. *$\int f d\mu_{X_n} \to \int f d\mu_X$ for all continuous bounded functions f on $\mathbb{R}$.*
3. *$F_n(t) \to F(t)$ as $n \to \infty$ at each t for which $F(t)$ is continuous.*
4. *$\int \chi_{(a,b]} d\mu_{X_n} \to \int \chi_{(a,b]} d\mu_X$ for all $a, b \in \mathbb{R}$ such that $a < b$ and F is continuous at a and b.*

Proof. (Sketch) (2) implying (1) is trivial. For (3) implying (4) consider the fact $F_n(t) = \int \chi_{(-\infty,t]} d\mu_{X_n}$ and $F(t) = \int \chi_{(-\infty,t]} d\mu_X$. For (4) implying (3) use the fact $\int \chi_{(a,b]} d\mu_{X_n} = F_n(b) - F_n(a)$ and $\int \chi_{(a,b]} d\mu_X = F(b) - F(a)$. Using the equivalence of (3) and (4) we show that (3) and (4) imply (2). We have by assumption and linearity that for step functions φ of the form $\varphi = \sum a_k \chi_{(a_k,b_k]}$ that $\int \varphi d\mu_{X_n} \to \int \varphi d\mu_X$. For any $\varepsilon > 0$ we can find $a, b \in \mathbb{R}$ such that $a < b$, $F(a) < \varepsilon$, $F(b) > 1 - \varepsilon$ and F is continuous at a and b. Let f be a continuous bounded function on $\mathbb{R}$. Then for the given ε we can approximate f on $[a, b]$ by a step function φ so that $|f(t) - \varphi(t)| < \varepsilon$ for all $t \in [a, b]$ (note φ depends on ε). Then

$$\left|\int f d\mu_{X_n} - \int f d\mu_X\right| = \left|\int f(d\mu_{X_n} - d\mu_X)\right|$$

$$\leqslant \left|\int_a^b f(d\mu_{X_n} - d\mu_X)\right| + \left|\int_b^\infty f(d\mu_{X_n} - d\mu_X)\right| + \left|\int_{-\infty}^a f(d\mu_{X_n} - d\mu_X)\right|$$

$$\leqslant \left|\int_a^b f - \varphi(d\mu_{X_n} - d\mu_X)\right| + \left|\int_a^b \varphi(d\mu_{X_n} - d\mu_X)\right| + \left|\int_b^\infty f(d\mu_{X_n} - d\mu_X)\right| + \left|\int_{-\infty}^a f(d\mu_{X_n} - d\mu_X)\right|$$

$$\leqslant 2\varepsilon + \left|\int_a^b \varphi(d\mu_{X_n} - d\mu_X)\right| + M((1 - F_n(b)) + (1 - F(b)) + F_n(a) + F(a))$$

where we had $1 - F(b), F(a) < \varepsilon$ and we have $1 - F_n(b), F_n(a) < \varepsilon$ since $F_n(b) \to F(b)$ and $F_n(a) \to F(a)$, and the second term can be arbitrarily small by (4) and linearity given φ is a linear combination of characteristic functions. For (1) implying (4) we approximate $\chi_{(a,b]}$ by a sequence of continuous functions. □

We now ask, does $X_n \to X$ for some mode of convergence imply $f(X_n) \to f(X)$ for a measurable function f? This is true if $X_n \to X$ almost surely. It is not true in general if $X_n \to X$ in r-mean. It is true if $X_n \to X$ in probability and f is continuous. It is also true if $X_n \to X$ in distribution and f is continuous (proof uses Skorkhod's representation theorem).

1.9 Characteristic Function of a Real Random Variable

If X is a real random variable we define the **characteristic function** of X as $\varphi_X(t) = \int_{\mathbb{R}} e^{itx} d\mu_X(x) = E(e^{itX}) = \int_\Omega e^{itX(\omega)} dP(\omega)$. We note that the characteristic function is the Fourier transform of the measure μ_X. If X is an absolutely continuous random variable with density $\rho_X(x)$, then the characteristic function of X is given by $\varphi_X(t) = \int_{\mathbb{R}} e^{itx} \rho_X(x) dx$, which is exactly the Fourier transform of the density function. Some properties of characteristic functions are now given.

- $\varphi_X(0) = 1$ and $|\varphi_X(t)| \leqslant 1$.
- $E(e^{itX}) = E(\cos tX) + iE(\sin tX)$.
- $\varphi_X(t)$ is well defined and uniformly continuous on $\mathbb{R}$.
- $\overline{\varphi_X(t)} = \varphi_X(-t) = \varphi_{-X}(t)$ where the over line denotes complex conjugation.
- If $Y = aX + b$ for $a, b \in \mathbb{R}$ then $\varphi_Y(t) = e^{itb} \varphi_X(ta)$. Proof: $\varphi_Y(t) = E(e^{itY}) = E(e^{it(aX+b)}) = e^{itb} E(e^{itaX}) = e^{itb} \varphi_X(ta)$.
- If $X_1, ..., X_n$ are independent real random variables and $S = X_1 + \cdots + X_n$, then $\varphi_S(t) = \varphi_{X_1}(t) \cdots \varphi_{X_n}(t)$. Proof: $\varphi_S(t) = E(e^{it(X_1 + \cdots + X_n)}) = E(e^{itX_1} \cdots e^{itX_n}) = E(e^{itX_1}) \cdots E(e^{itX_n}) = \varphi_{X_1}(t) \cdots \varphi_{X_n}(t)$ where the third equality is by independence.

Example 1.9.1. *Let* $X = a$ *be the constant function so that* X *has a Dirac distribution. Then* $\varphi_X(t) = e^{ita}$.

Example 1.9.2. *Let* X *be a Bernoulli random variable with parameter* p *so that* $P(X = 0) = 1 - p$ *and* $P(X = 1) = p$. *Then* $\varphi_X(t) = 1 - p + pe^{it}$.

Example 1.9.3. *Let* X *be a Poisson random variable with parameter* λ *so that* $P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$ *for all* $k \in \mathbb{N}_0$. *Then* $\varphi_X(t) = e^{\lambda(e^{it}-1)}$.

Example 1.9.4. *Let* X *have a uniform distribution on* $[a, b]$ *so that* $\rho(x) = \begin{cases} \frac{1}{b-a} & x \in [a, b] \\ 0 & \textit{otherwise} \end{cases}$. *Then* $\varphi_X(t) = \int_a^b \frac{e^{itx}}{b-a} dx = \frac{e^{itb} - e^{ita}}{b-a}$.

Example 1.9.5. *Let* X *have an exponential distribution with parameter* $\lambda > 0$ *so that* $\rho(x) = \begin{cases} \frac{1}{\lambda} e^{-\lambda x} & x \geqslant 0 \\ 0 & x < 0 \end{cases}$. *Then* $\varphi_X(t) = \frac{1}{1 - i\lambda t}$.

Example 1.9.6. *Let* X *have a normal distribution with mean* μ *and variance* σ^2 *so that* $\rho(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$. *Then* $\varphi_X(t) = e^{it\mu - \frac{1}{2}t^2\sigma^2}$.

Example 1.9.7. *Let* X *have a Cauchy distribution with parameter* $\gamma = 1$ *so that* $\rho(x) = \frac{1}{\pi(1+x^2)}$. *Then* $\varphi_X(t) = \frac{1}{\pi} \int_{\mathbb{R}} \frac{e^{itx}}{1+x^2} dx = e^{-|t|}$.

Example 1.9.8. *Consider* n *independent Bernoulli trials where each* X_k *satisfies* $P(X_k = 0) = 1 - p$ *and* $P(X_k = 1) = p$. *Then* $S = X_1 + \cdots + X_n$ *has a binomial distribution with parameters* n *and* p. *Recall* $\varphi_{X_k}(t) = 1 - p + pe^{it}$ *so that by our properties of the characteristic function* $\rho_S(t) = (1 - p + pe^{it})^n = \sum_{k=0}^{n} \binom{n}{k} p^k (1-p)^{n-k} e^{itk}$.

Example 1.9.9. *Let* X *and* Y *be independent normal random variables with mean* $\mu = 0$ *and variance* $\sigma^2 = 1$, *and* $S = \frac{X+Y}{\sqrt{2}}$. *Recall* $\varphi_X(t) = \varphi_Y(t) = e^{-t^2/2}$ *so that by our properties of the characteristic function* $\varphi_{X/\sqrt{2}} = e^{-(t/\sqrt{2})^2/2} = e^{-t^2/4}$, *which implies* $\varphi_S(t) = \varphi_{X/\sqrt{2}}(t)\varphi_{Y/\sqrt{2}}(t) = e^{-t^2/4}e^{-t^2/4} = e^{-t^2/2}$. *We also note* $E(S) = E\left(\frac{X}{\sqrt{2}}\right) + E\left(\frac{Y}{\sqrt{2}}\right) = \frac{1}{\sqrt{2}}E(X) + \frac{1}{\sqrt{2}}E(Y) = 0 + 0 = 0$ *and* $\mathrm{Var}(S) = \mathrm{Var}\left(\frac{X+Y}{\sqrt{2}}\right) = \mathrm{Var}\left(\frac{X}{\sqrt{2}}\right) + \mathrm{Var}\left(\frac{Y}{\sqrt{2}}\right) = \frac{1}{2}\mathrm{Var}(X) + \frac{1}{2}\mathrm{Var}(Y) = 1$, *where the second equality holds by independence. Thus* S *is a normal random variable with mean* 0 *and variance* 1 *as well.*

Example 1.9.10. *Let* X *and* Y *be independent Cauchy random variables with parameter* $\gamma = 1$, *and* $S = \frac{X+Y}{2}$. *Recall* $\varphi_X(t) = e^{-|t|}$ *so that by our properties of the characteristic function* $\varphi_{X/2}(t) = e^{-|t|/2}$, *which gives* $\varphi_S(t) = \varphi_{X/2}(t)\varphi_{Y/2}(t) = e^{-|t|/2}e^{-|t|/2} = e^{-|t|}$.

Theorem 1.9.1. *(Uniqueness of Characteristic Function) If* X *and* Y *are real random variables, then* $\varphi_X = \varphi_Y$ *holds if and only if* $\mu_X = \mu_Y$ *holds, which is if and only if* $F_X = F_Y$ *holds.*

This theorem follows from the following version of the inversion theorem for characteristic functions.

Theorem 1.9.2. *(Inversion of Characteristic Function) Let* X *be a real random variable with distribution function* $F = F_X$ *and characteristic function* $\varphi = \varphi_X$. *Then for all* $a, b \in \mathbb{R}$

$$\frac{1}{2}\left(\lim_{x \to b^+} F(x) + \lim_{x \to b^-} F(x)\right) - \frac{1}{2}\left(\lim_{x \to a^+} F(x) + \lim_{x \to a^-} F(x)\right)$$
$$= \lim_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-itb} - e^{-ita}}{-it} \varphi(t) dt$$

Before giving a proof of the above theorem, we make some notes. First recall that for all $t \in \mathbb{R}$ we have $\lim_{x\to t^+} F(x) = F(t)$ and $\lim_{x \to t^-} F(x) = F(t) - P(X = t)$. Second, knowing $\frac{1}{2}(\lim_{x\to t+} F(x) + \lim_{x\to t^-} F(x))$ completely determines $F(t)$. Lastly, taking inverse Fourier transforms would give the density of X (assuming it is absolutely continuous). To obatin a version of the distribution function, we have the factor $\frac{1}{-it}$ in the integrand above.

Proof. (Sketch) Since $\left|e^{ix} - 1\right| \leqslant |x|$ we observe

$$\left|\frac{e^{-itb} - e^{-ita}}{t}\right| = \left|\frac{e^{-it(b-a)} - 1}{t}\right| \leqslant |b - a|$$

Therefore, since $|\varphi(t)| \leqslant 1$,

$$\left|\int_{-T}^{T} \frac{e^{-itb} - e^{-ita}}{t} \varphi(t) dt\right| \leqslant 2T(b - a)$$

Thus by Fubini's theorem

$$\begin{aligned}\frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-itb} - e^{-ita}}{-it} \varphi(t) dt &= \frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-itb} - e^{-ita}}{-it} \int_{-\infty}^{\infty} e^{itx} d\mu_X(x) dt \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-T}^{T} \frac{e^{-itb} - e^{-ita}}{-it} e^{itx} dt d\mu_X(x) \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\int_{-T}^{T} \frac{e^{-it(x-b)} - e^{-it(x-a)}}{-it} dt d\mu_X(x)\end{aligned}$$

Splitting the inner integral into two integrals from $-T$ to 0 and 0 to T, and making the substitution $t \to -t$ gives the above

$$= \frac{1}{\pi}\int_{-\infty}^{\infty}\int_{0}^{T} \frac{\sin t(x-b)}{t} - \frac{\sin t(x-a)}{t} dt d\mu_X(x) \tag{1}$$

Define $H(a, b, x, T) = \int_0^T \frac{\sin t(x-b)}{t} - \frac{\sin t(x-a)}{t} dt$. One can show using the fact

$$\int_0^{\infty} \frac{\sin \zeta t}{t} dt = \begin{cases} \frac{\pi}{2} & \zeta > 0 \\ 0 & \zeta = 0 \\ \frac{-\pi}{2} & \zeta < 0 \end{cases} \tag{2}$$

that

$$\lim_{T\to\infty} H(a, b, x, T) = \begin{cases} 0 & x < a \\ \frac{\pi}{2} & x = a \\ \pi & a < x < b \\ \frac{\pi}{2} & x = b \\ 0 & x > 0 \end{cases}$$

pointwisely in x. Additionally, since we see from (2) that $\left|\int_0^T \frac{\sin \zeta t}{t} dt\right| \leqslant \frac{\pi}{2}$, we have by the dominated convergence theorem and (1) that

$$\lim_{T\to\infty} \frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-itb} - e^{-ita}}{-it} \varphi(t) dt = \frac{1}{\pi}\int_{-\infty}^{\infty} \lim_{T\to\infty} H(a, b, x, T) d\mu_X(x)$$

$$= \frac{1}{2}P(X = a) + P(a < x < b) + \frac{1}{2}P(X = b)$$

$$= \frac{1}{2}\left(\lim_{x\to b^+} F(x) + \lim_{x\to b^-} F(x)\right) - \frac{1}{2}\left(\lim_{x\to a^+} F(x) + \lim_{x\to a^-} F(x)\right)$$

□

Corollary 1.9.1. *If* $\int_{\mathbb{R}} |\varphi_X(t)|\, dt < \infty$, *then* X *is absolutely continuous and the density of* X *is given by*

$$F'(x) = \rho_X(x) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi_X(t)dt$$

Proof. (Sketch) By the inversion formula we have

$$\frac{F(x+h) - F(x)}{h} = \lim_{T\to\infty} \frac{1}{2\pi}\int_{-T}^{T} \frac{e^{-it(x+h)} - e^{itx}}{-ith}\varphi(t)dt$$

$$= \lim_{T\to\infty} \frac{1}{2\pi}\int_{-T}^{T} e^{-itx}\frac{1 - e^{ith}}{-ith}\varphi(t)dt$$

Since $\varphi(t) \in L^1$ by assumption the above integrand is integrable on $\mathbb{R}$, so that the above can be written as

$$= \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\frac{1 - e^{ith}}{-ith}\varphi(t)dt$$

Since $\frac{1-e^{ith}}{-ith} \to 1$ pointwisely in t as $h \to 0$, we have by the dominated convergence theorem

$$\lim_{h\to 0} \frac{F(x+h) - F(x)}{h} = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-itx}\varphi(t)dt$$

What remains is to show $F'(x) = \rho_X(x)$ and that X is absolutely continuous. □

If $X_1, ..., X_n$ are random variables and $X = (X_1, ..., X_n)$ we define $\varphi_X(t)(t_1, ..., t_n) = E\left(e^{i\sum_{i=1}^n t_i X_i}\right)$.

Theorem 1.9.3. *Let* X *and* Y *be random vectors. Then* $\varphi_X = \varphi_Y$ *if and only if* X *and* Y *have identical distributions.*

Corollary 1.9.2. *Let* $X = (X_1, ..., X_n)$. *Then* $X_1, ..., X_n$ *are mutually independent if and only if* $\varphi_X(t_1, ..., t_n) = \varphi_{X_1}(t_1) \cdots \varphi_{X_n}(t_n)$.

Proof. The "only if" direction follows from definition as an expectation of products of exponentials will become a product of expectations of exponentials by independence. For the "if" direction, one can define independent random variables $\tilde{X}_1, ..., \tilde{X}_n$ such that $\tilde{X}_k$ and X_k have identical distributions. This implies from the single variable case of the previous theorem that $\varphi_{X_k} = \varphi_{\tilde{X}_k}$. Let $\tilde{X} = (\tilde{X}_1, ..., \tilde{X}_n)$. Then by our assumption $\varphi_{\tilde{X}} = \varphi_{\tilde{X}_1} \cdots \varphi_{\tilde{X}_n} = \varphi_{X_1} \cdots \varphi_{X_n} = \varphi_X$. By the previous theorem, X and $\tilde{X}$ have identical distributions. Since $\tilde{X}$ has independent components, its distribution can be expressed as a product measure. Thus so can the distribution of X, which implies the components of X are mutually independent. □

We note an important relationship between the moments and characteristic function of a real random variable X. Informally, observe

$$\varphi_X(t) = E\left(e^{itX}\right) = E\left(\sum_{k\in\mathbb{N}} \frac{(it)^k X^k}{k!}\right) = \sum_{k\in\mathbb{N}} \frac{(it)^k}{k!} E\left(X^k\right)$$

Therefore, $\varphi_X^{(k)}(0) = i^k E(X^k)$, i.e., the k^{th} derivative of the characteristic function evaluated at 0 gives the k^{th} moment of X (up to a factor of i^k). We make this observation precise with the following theorem.

Theorem 1.9.4. *Let X be a real valued random variable and $\varphi(t)$ its characteristic function. (a) If $E(|X|^n) < \infty$ for some $n \in \mathbb{N}$, then $\varphi^{(k)}(t)$ exists for $k = 0, ...n$,*

$$\varphi^{(k)}(0) = \int_{\mathbb{R}} (ix)^k e^{itx} d\mu_X(x) = i^k E(X^k)$$

and

$$\varphi(t) = \sum_{k=0}^{n} \frac{(it)^k}{k!} E\left(X^k\right) + o(t^n) \text{ as } t \to 0$$

(b) If $\varphi^{(2n)}(0)$ exists, then $E(X^{2n}) < \infty$ (so that $E(|X|^{2n}) < \infty$.

Proof. (Sketch) Existence of $\varphi^{(k)}$ is done by induction on k using the expression $\frac{\varphi^{(k)}(t+\Delta t)-\varphi^{(k)}(t)}{\Delta t}$. Assume the statement in (a) holds for some k. Then

$$\frac{\varphi^{(k)}(t+\Delta t) - \varphi^{(k)}(t)}{\Delta t} = \int_{\mathbb{R}} (ix)^k \left(\frac{e^{i\Delta tx} - 1}{\Delta t}\right) e^{itx} d\mu_X(x)$$

$$= \int_{\mathbb{R}} \frac{e^{i\Delta tx} - 1}{i\Delta tx} e^{itx} (ix)^{k+1} d\mu_X(x) = \int_{\mathbb{R}} \frac{e^{i\Delta tx} - 1}{i\Delta tx} e^{itx} \left(i\frac{x}{|x|}\right)^{k+1} |x|^{k+1} d\mu_X(x)$$

The term $\frac{e^{i\Delta tx}-1}{i\Delta tx} \to 0$ pointwisely in x as $\Delta t \to 0$. Both exponential terms and $\left(i\frac{x}{|x|}\right)^{k+1}$ in the integrand are bounded. We also have $\int_{\mathbb{R}} |x|^n d\mu_X(x) < \infty$. Thus we can apply the dominated convergence theorem to obtain the first equation in (a). For showing the second equation in (a) let

$$\psi_n(t) = \varphi(t) - \sum_{k=0}^{n} \frac{(it)^k E(X^k)}{k!} = \int_{\mathbb{R}} e^{itx} - \sum_{k=0}^{n} \frac{(itx)^k}{k!} d\mu_X(x)$$

and let $h(tx) = e^{itx} - \sum_{k=0}^{n} \frac{(itx)^k}{k!}$. We show $\psi_n(t) = o(t^n)$ by showing $\frac{\psi_n(t)}{t^n} \to 0$ as $t \to 0$. Note

$$\frac{\psi_n(t)}{t^n} = \int_{\mathbb{R}} \frac{h(tx)}{t^n x^n} x^n d\mu_X(x)$$

Since $\int_{\mathbb{R}} |x|^n d\mu_X(x) < \infty$, it suffices to show $\frac{h(tx)}{t^n x^n} \to 0$ pointwisely in x as $t \to 0$. Substituting $s = tx$ (so $s \to 0$ as $t \to 0$)

$$\frac{h(s)}{s^n} = \frac{e^{is} - \sum_{k=0}^{n} \frac{(is)^k}{k!}}{s^n} = \frac{\sum_{k=n+1}^{\infty} \frac{(is)^k}{k!}}{s^n} = \frac{(is)^{n+1} \sum_{k=n+1}^{\infty} \frac{(is)^{k-n-1}}{k!}}{s^n}$$

$$= i^{n+1} s \sum_{k=n+1}^{\infty} \frac{(is)^{k-n-1}}{k!} = i^{n+1} s \sum_{k=0}^{\infty} \frac{(is)^k}{k!} = O(s)$$

as $s \to 0$ for each fixed n, since the sum will be convergent. □

The **moment generating function** of a real random variable X is given by $M(t) = E(e^{tX}) = \int_{\mathbb{R}} e^{tx} d\mu_X(x)$. It is the Laplace transform of the measure μ_X (the Laplace transform has a $-t$ instead t but we will see this transformation does not matter given $M_X(t)$ is defined on an interval $|t| < h$). Some properties are noted below.

- $M^{(k)}(0) = E(X^k)$ for each $k \in \mathbb{N}_0$. Proof: Expand exponential term in the definition by its Taylor series.
- The moment generating function may not be defined. We say a random variable X has a moment generating function if $M(t)$ exists on some interval $|t| < h$, where $h > 0$.
- If $M(t)$ exists on $|t| < h$ then it has a convergent Taylor series. In particular, $M(t) = \sum_{k \in \mathbb{N}_0} \frac{E(X^k)}{k!} t^k = \sum_{k \in \mathbb{N}_0} \frac{m_k}{k!} t^k$ for all $|t| < h$.
- If $E(X^k)$ exists for all $k \in \mathbb{N}_0$ and $\limsup_{k \to \infty} \frac{E(|X|^k)^{1/k}}{k} < \infty$ then $M(t)$ exists for all $|t| < h$ (if the limit supremum is finite then it actually equals $\frac{1}{eh}$). The proof uses Sterling's formula.
- If X and Y are independent random variables then $M_{X+Y}(t) = M_X(t)M_Y(t)$ (the existence of $M_X(t)$ and $M_Y(t)$ imply the existence of $M_{X+Y}(t)$). It also follows imeadiatley from the definition $M_{\lambda X}(t) = M_X(\lambda t)$ for all $\lambda \in \mathbb{R}$.

We now begin a discussion of cumulants and moments. Assume $E(|X|^n) < \infty$ for some $n \in \mathbb{N}$. Then recall $\varphi_X(t)$ is n times continuously differentiable and $\varphi_X(0) = 1$. Thus $\log \varphi_X(t)$ is well defined and n times differentiable on $|t| < \varepsilon$ for some $\varepsilon > 0$. Note then by our previous theorem

$$\log \varphi_X(t) = \log E(e^{itX}) = \sum_{k=1}^{n} \frac{(it)^k}{k!} c_k + o(t^n)$$

as $t \to 0$, for some $c_1, ..., c_n$. In this case, the coefficients c_k are determined uniquely, and are the cumulants we are interested in. However, we want to obtain c_k for all $k \in \mathbb{N}$, and to do this we need the existence of the moment generating function for the existence of a converging infinite series. If $M(t)$ exists for all $|t| < h$ then

$$\log M_X(t) = \log E(e^{tX}) = \sum_{k \in \mathbb{N}} \frac{t^k}{k!} c_k$$

The coefficient c_k is called the k^{th} **cumulant**. There is a one to one correspondence between the k^{th} moment m_k and the k^{th} cumulant c_k. Note $M_X(t) = E(e^{tX}) = 1 + \sum_{k \in \mathbb{N}} \frac{m_k}{k!} t^k$ while also (by exponentiating both sides of the expression for $\log M_X(t)$)

$$E(e^{tX}) = \exp\left(\sum_{k \in \mathbb{N}} \frac{t^k}{k!} c_k \right) = \exp\left(c_1 t + \frac{c_2 t^2}{2} + \frac{c_3 t^3}{6} + \cdots \right)$$

$$= 1 + \left(c_1 t + \frac{c_2 t^2}{2} + \frac{c_3 t^3}{6} + \cdots \right) + \frac{1}{2}\left(c_1 t + \frac{c_2 t^2}{2} + \frac{c_3 t^3}{6} + \cdots \right)^2 + \frac{1}{6}\left(c_1 t + \frac{c_2 t^2}{2} + \frac{c_3 t^3}{6} + \cdots \right)^3 + \cdots$$

Thus by comparing the two expressions for $E(e^{tX})$ we see as example $m_1 = c_1$, $\frac{m_2}{2} = \frac{c_2}{2} + \frac{c_1^2}{2}$ and $\frac{m_3}{6} = \frac{c_3}{6} + \frac{1}{2}(2c_1c_2) + \frac{1}{6}c_1^3$. Solving for the first two cumulants gives $c_1 = m_1 = E(X)$, $c_2 = m_2 - m_1^2 = E(X) - E(X^2) = Var(X)$. The third and fourth cumulants also have interpretations as well. One can show $c_3 = m_3 - 3m_1m_2 + 2m_1^2$ and $c_4 = m_4 - 3m_2^2 - 4m_1m_3 + 12m_1^2m_2 - 6m_1^4$. If we let $E(X) = \mu$ and $\sqrt{Var(X)} = \sigma$ (typically referred to as standard deviation), then $\gamma_1 = \frac{c_3}{c_2^{3/2}} = \frac{E((X-\mu)^3)}{\sigma^3}$ and $\gamma_2 = \frac{c_4}{c_2^2} = \frac{E((X-\mu)^4)}{\sigma^4} - 3$ are respectively known as **skewness** and **coefficient of excess** of the law of X.

If X and Y are independent random variables and all the moments exist, thinking of the cumulants as a function of the random variables, $c_k(X+Y) = c_k(X) + c_k(Y)$ for all $k \in \mathbb{N}$. For the cases $k = 1,2$ we get the additivity of expectation and variance we are familiar with for sums of independent random variables. To show this for all $k \in \mathbb{N}$, we recall for such X and Y that $M_{X+Y}(t) = M_X(t)M_Y(t)$ so that $\log M_{X+Y}(t) = \log M_X(t) + \log M_Y(t)$. Thus $\log M_{X+Y}(t) = \sum_{k\in\mathbb{N}} \frac{c_k(X+Y)t^k}{k!}$ and $\log M_X + \log M_Y(t) = \sum_{k\in\mathbb{N}} \frac{c_k(X)t^k}{k!} + \sum_{k\in\mathbb{N}} \frac{c_k(Y)t^k}{k!} = \sum_{k\in\mathbb{N}} \frac{(c_k(X)+c_k(Y))t^k}{k!}$ implies $c_k(X+Y) = c_k(X) + c_k(Y)$ as desired. We also have $c_k(\lambda X) = \lambda^k x_k(X)$ for each $k \in \mathbb{N}$.

1.10 Central Limit Theorem

The following theorems show the "universality" of the Gaussian distribution.

Theorem 1.10.1. *(Central Limit Theorem) Let $X_1, X_2, ...$ be a sequence of i.i.d random variables with mean $\mu < \infty$ and variance $\sigma^2 \in (0,\infty)$. Let $S_n = X_1 + \cdots + X_n$ and $Z_n = \frac{S_n - n\mu}{\sqrt{n}\sigma}$. Then Z_n converges in distribution to a Gaussian random variable with mean 0 and variance 1.*

Before giving a sketch of the proof, we note $E(S_n) = E(X_1) + \cdots E(X_n) = n\mu$ and $Var(S_n) = Var(X_1) + \cdots + Var(X_n) = n\sigma^2$. Recall $\frac{S_n - n\mu}{n} \to 0$ almost surely as $n \to \infty$ by Kolomogorov's strong law of large numbers. The reason we consider a factor of $\sqrt{n}$ instead of n in the denominator of Z_n is since we observe $Var\left(\frac{S_n-n\mu}{n}\right) = \frac{Var(S_n)}{n^2} = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n} \to 0$ as $n \to \infty$. For the Z_n defined as in the statement of the theorem, note $E(Z_n) = 0$ and $Var(Z_n) = Var\left(\frac{S_n-n\mu}{\sqrt{n}\sigma}\right) = \frac{Var(S_n-n\mu)}{n\sigma^2} = \frac{Var(S_n)}{n\sigma^2} = \frac{n\sigma^2}{n\sigma^2} = 1$.

Proof. Let N be a Gaussian random variable with mean 0 and variance 1. We equivalently show $\varphi_{Z_n}(t) \to \varphi_N$. Let X denote a random variable with the same distribution as each X_k, so that each $\varphi_{X_k} = \varphi_X$. By independence $\varphi_{S_n}(t) = \varphi_{X_1}(t) \cdots \varphi_{X_n}(t) = \varphi_X(t)^n$. Then we observe

$$\varphi_{Z_n} = E\left(e^{it\frac{S_n-n\mu}{\sqrt{n}\sigma}}\right) = e^{\frac{-it\sqrt{n}\mu}{\sigma}} E\left(e^{it\frac{S_n}{\sqrt{n}\sigma}}\right) = e^{\frac{-it\sqrt{n}\mu}{\sigma}} \varphi_{S_n}\left(\frac{t}{\sqrt{n}\sigma}\right) = e^{\frac{-it\sqrt{n}\mu}{\sigma}} \varphi_X\left(\frac{t}{\sqrt{n}\sigma}\right)^n$$

Since $E(X^2) < \infty$ (by finite variance) we have

$$\varphi_X(s) = 1 + i\mu s - \frac{\sigma^2}{2}s^2 + o(s^2)$$

as $s \to 0$. Thus for fixed t and $s = \frac{t}{\sqrt{n}\sigma}$ we see

$$\begin{aligned}
\varphi_{Z_n} &= \left(e^{\frac{-it\mu}{\sqrt{n}\sigma}}\left(1 + i\mu\frac{t}{\sqrt{n}} - \frac{\sigma^2}{2}\frac{t^2}{n} + o\left(\frac{1}{n}\right)\right)\right)^n \\
&= \left(\left(1 + \frac{it\mu}{\sqrt{n}\sigma} - \frac{t^2\mu^2}{2n\sigma^2}\right)\left(1 + \frac{it\mu}{\sqrt{n}\sigma} - \frac{t^2\sigma^2\mu^2}{2n\sigma^2}\right) + o\left(\frac{1}{n}\right)\right)^n \\
&= \left(1 - \frac{t^2}{2n}o\left(\frac{1}{n}\right)\right) \to e^{-t^2/2}
\end{aligned}$$

as $n \to \infty$. But $e^{-t^2/2}$ is exactly the characteristic function of N. □

Proof. We now give an alternate proof of the central limit theorem. Let N be a Gaussian random variable with mean 0 and variance 1. We equivalently show $\varphi_{Z_n}(t) \to \varphi_N$. Let X denote a random variable with the same distribution as each X_k, so that each $\varphi_{X_k} = \varphi_X$. We rewrite $Z_n = \frac{1}{\sqrt{n}}(Y_1 + \cdots + Y_n)$ where each $Y_k = \frac{X_k - \mu}{\sigma}$. Note the Y_k are i.i.d and each Y_k has mean 0 and variance 1. Let Y denote a random variable with the same distribution as each Y_k, so that each $\varphi_{Y_k} = \varphi_Y$. Since $E(Y^2) < \infty$ (by finite variance) we have

$$\varphi_Y(t) = 1 - \frac{t^2}{2} + o(t^2)$$

as $t \to 0$. Then by independence if the Y_k we have

$$\varphi_{Z_n}(t) = \left(\varphi_Y\left(\frac{t}{\sqrt{n}}\right)\right)^n = \left(1 - \frac{t^2}{2n} + o\left(\frac{t^2}{n}\right)\right)^n \to e^{-t^2/2}$$

as $n \to \infty$. But $e^{-t^2/2}$ is exactly the characteristic function of N. □

Theorem 1.10.2. *(Central Limit Theorem of Lindeberg-Levy-Feller) Let $X_1, X_2, ...$ be a sequence of independent random variables, where each $E(X_k) = \mu_k$ and each $Var(X_k) = \sigma_n^2$. Let $S_n = X_1 + \cdots + X_n$ and $\tilde{S}_n = \sigma_1^2 + \cdots + \sigma_n^2$. If $\frac{1}{\tilde{S}_n} \max\{\sigma_1^2, ..., \sigma_n^2\} \to 0$ as $n \to \infty$ and for all $\varepsilon > 0$ we have $\frac{1}{\tilde{S}_n} \sum_{k=1}^n E\left(|X_k - \mu_k|^2 \chi_{|X_k - \mu_k| \geqslant \varepsilon \tilde{S}_n}\right) \to 0$ as $n \to \infty$, then $\frac{S_n - E(S_n)}{\tilde{S}_n} = \frac{\sum_{k=1}^n X_k - \mu_k}{\tilde{S}_n}$ converges in distribution to a Gaussian random variable with mean 0 and variance 1.*

1.11 Convergence of Characteristic Functions and Moments

Our main goal in the question of convergence of characteristic function is to prove the following theorem.

Theorem 1.11.1. *$X_n \to X$ in distribution if and only if $\varphi_{X_n}(t) \to \varphi_X(t)$ pointwisely for all $t \in \mathbb{R}$.*

For the "only if" direction, we have by assumption $\int f d\mu_{X_n} \to \int f d\mu_X$ for each bounded continuous function on $\mathbb{R}$. Thus we can take $f(x) = e^{itx}$ for fixed t and we would have $\varphi_{X_n}(t) \to \varphi_X(t)$ for each t. The "if" direction is more complicated and requires the following theory.

Theorem 1.11.2. *Let $\{X_n\}_{n \in \mathbb{N}}$ be a sequence of random variables. If for a function $\varphi(t)$ we have $\varphi_{X_n}(t) \to \varphi(t)$ for each $t \in \mathbb{R}$ and $\varphi(t)$ is continuous at $t = 0$, then $\varphi(t)$ is the characteristic function of some random variable X (so that $\varphi(t) = \varphi_X(t)$) and $X_n \to X$ in distribution.*

A sequence of random variables $\{X_n\}_{n \in \mathbb{N}}$ is called **tight** if $P(|X_n| \geqslant a) \to 0$ uniformly in n as $a \to \infty$. I.e.,

$$\lim_{a \to \infty} \left(\sup_{n \in \mathbb{N}} P(|X_n| \geqslant a)\right) = 0$$

Intuitively, the "probability mass" does not go to $\pm\infty$.

Example 1.11.1. *Let each X_n have a uniform distribution on $[-n, n]$. Then the density function is given by $\rho_{X_n} = \frac{1}{2n}\chi_{[-n,n]}$. Then one can show $F_{X_n}(t) \to \frac{1}{2}$ as $n \to \infty$ for each fixed t. However $\frac{1}{2}$ cannot be a distribution function for some random variable since $\lim_{t \to \infty} \frac{1}{2} \neq 1$ and $\lim_{t \to -\infty} \frac{1}{2} \neq 0$.*

Note $P(|X_n| \geqslant a) = 2\left(\frac{n-a}{2n}\right) = 1 - \frac{a}{n}$ for $n > a$. Thus $\sup_{n \in \mathbb{N}} P(|X_n| \geqslant a) = 1$, whose limit as $a \to \infty$ is $1 \neq 0$. Thus the sequence $\{X_n\}_{n \in \mathbb{N}}$ is not tight.

Theorem 1.11.3. *(Helly's Selection Principle) Let $\{X_n\}_{n\in\mathbb{N}}$ be a tight sequence of real random variables. Then there exists a subsequence $\{X_{n_k}\}_{k\in\mathbb{N}}$ and a real random variable X such that $X_{n_k} \to X$ in distribution.*

Proof. (Sketch) Let $F_n = F_{X_n}$ be the distribution function of each X_n for simplicity of notation. Choose any countable dense subset $S \subseteq \mathbb{R}$ (such as $\mathbb{Q}$). Since $F_n(s) \in [0,1]$ for each $s \in S$, the sequence $F_n(s)$ has an accumulation point. By selecting subsequences of subsequences, and then making a diagonal argument, one can show there exists $n_1 < n_2 < \cdots$ such that for all $s \in S$, $F_{n_k}(s)$ converges as $k \to \infty$. In particular, we have the nested subsequences of natural numbers $\{k\}_{k\in\mathbb{N}} \supseteq \{n_k^{(1)}\}_{k\in\mathbb{N}} \supseteq \{n_k^{(2)}\}_{k\in\mathbb{N}} \supseteq \cdots$ and we choose to index by the sequence $\{n_k^{(k)}\}_{k\in\mathbb{N}}$. Then we define $\tilde{F}(s) = \lim_{k\to\infty} F_{n_k}(s)$ for each $s \in S$. $\tilde{F}$ will be monotonically increasing. We can extend this uniquely to a function $\tilde{F}$ on all $\mathbb{R}$. We then define $F(s) = \lim_{t\to s^+} \tilde{F}(t)$ and we would have F is a redefined version of $\tilde{F}$ such that $F(s) = \lim_{t\to s^+} \tilde{F}(t)$ and $\lim_{t\to s^-} F(t)$ exists.

Using monotonicity of $\tilde{F}$ one can show that $F : \mathbb{R} \to [0,1]$ is monotonically increasing, $F(x) = \lim_{t\to s^+} F(t)$ and $F(x) = \lim_{k\to\infty} F_{n_k}(x)$ for all $x \in \mathbb{R}$ (except when F has a jump discontinuity at x). By the tightness of the sequence $\{X_n\}$ we have $\lim_{x\to-\infty} F(x) = 0$ and $\lim_{x\to\infty} F(x) = 1$. This is because for all $\varepsilon > 0$ there exists sufficiently large $a \in \mathbb{R}$ such that $(1 - F_n(a)) + F_n(-a) = P(|X_n| \geqslant a) < \varepsilon$ for all $n \in \mathbb{N}$. Since the bound is uniform on n, we have by taking the limit as $n \to \infty$ that $(1 - F(a)) + F(-a) < \varepsilon$ which implies $0 \leqslant 1 - F(a), F_n(-a) < \varepsilon$. These properties imply by earlier results that F is the distribution function of some random variable X so that $F_{n_k}(x) \to F(x) = F_X(x)$ at every $x \in \mathbb{R}$ where $F_X(s)$ is continuous. Thus by earlier results $X_{n_k} \to X$ is distribution. □

Lemma 1.11.1. *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of random variables. If $\varphi_{X_n}(t) \to \varphi(t)$ for all $t \in \mathbb{R}$, for some φ that is continuous at $t = 0$, then $\{X_n\}_{n\in\mathbb{N}}$ is tight.*

Proof. (Sketch) We have in general that if $f(t)$ is a continuous function at $t = 0$, then

$$\lim_{b\to 0} \frac{1}{b} \int_0^b f(t)dt = 0 \tag{1}$$

We first show $P\left(|X| \geqslant \frac{1}{b}\right) \leqslant \frac{k}{b} \int_0^b 1 - \operatorname{Re} \varphi_X(t)dt$ for some $k > 0$. Observe

$$\frac{1}{b}\int_0^b 1 - \operatorname{Re}\varphi_X(t)dt = \frac{1}{b}\int_0^b \int_{-\infty}^{\infty} 1 - \cos(xt)d\mu_X(x)dt$$

$$\int_{-\infty}^{\infty} \frac{1}{b}\int_0^b 1 - \cos(xt)dtd\mu_X(x) = \int_{-\infty}^{\infty} 1 - \frac{\sin bx}{bx} d\mu_X(x)$$

$$\geqslant k\int_{|bx|\geqslant 1} d\mu_X = kP(b|X| \geqslant 1)$$

Then, the result can be shown using the fact

$$0 \leqslant P\left(|X_n| \geqslant \frac{1}{b}\right) \leqslant \frac{k}{b}\int_0^b 1 - \operatorname{Re}\varphi_{X_n}(t)dt \to 0$$

as $b \to 0$ by (1), and finding bounds on $\sup_{n\in\mathbb{N}} P(|X_n| \geqslant \frac{1}{b})$ for different cases of n. □

Now we are ready to go back to showing the "if" direction in the proof of our first theorem of this subsection.

Proof. Since our assumption is the hypothesis of the previous lemma, we have that the sequence $\{X_n\}$ is tight. By Helly's Selection Principle there exists a subsequence $\{X_{n_k}\}_{k\in\mathbb{N}}$ such that $X_{n_k} \to X$ in distribution as $k \to \infty$. To show $X_n \to X$ in distribution, assume this does not hold for the sake of contradiction. Then there exists a continuous function f on $\mathbb{R}$ such that $f(X_n)$ does not converge to $f(X)$ in distribution. Thus there exists a subsequence $\{f(X_{n'_k})\}_{k\in\mathbb{N}}$ and $\varepsilon > 0$ such that $\left|f(X_{n'_k}) - f(X)\right| \geqslant \varepsilon$ for all $k \in \mathbb{N}$. By Helly's Selection Principle applied to $\{X_{n_k}\}_{k\in\mathbb{N}}$ there exists a subsequence $\{X_{n''_k}\}_{k\in\mathbb{N}}$ such that $X_{n''_k} \to Y$ in distribution as $k \to \infty$, for some random variable Y. Thus

$$|f(X) - f(Y)| \geqslant \varepsilon \tag{1}$$

Now since $X_{n_k} \to X$ in distribution as $k \to \infty$ we have $\varphi_{X_{n_k}} \to \varphi_X$ as $k \to \infty$. Similarly, since $X_{n''_k} \to Y$ in distribution as $k \to \infty$ we have $\varphi_{X_{n''_k}} \to \varphi_Y$ as $k \to \infty$. Since $\varphi_{X_n} \to \varphi_X$ by assumption, we have $\varphi_X = \varphi_Y$. Thus X must have identical distribution to Y, which contradicts (1). □

We now discuss the method of moments. The central question is, does $X_n \to X$ in distribution if and only if $E(X_n^k) \to E(X^k)$ as $n \to \infty$, for all $k \in \mathbb{N}$? If $X_n \to X$ in distribution, we know $X_n^k \to X^k$ in distribution for all $k \in \mathbb{N}$. However, we may not always apply expectations to this and expect $E(X_n^k) \to E(X^k)$ to hold as well, as the following example illustrates.

Example 1.11.2. *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of real random variables such that $P(X = 0) = 1 - \frac{1}{n}$ and $P(X = n) = \frac{1}{n}$ for each $n \in \mathbb{N}$. Let X be a random variable such that $P(X = 0) = 1$. Then $X_n \to X$ in distribution. However, we notice $E(X_n) = 0\left(1 - \frac{1}{n}\right) + n\left(\frac{1}{n}\right) = 1$ and $E(X) = 0(1) = 0$, so that $E(X_n)$ does not converge to $E(X)$.*

Proposition 1.11.1. *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of real random variables such that $\sup_{n\in\mathbb{N}} E(|X_n|^k) < \infty$ for some $k \in \mathbb{N}$ and $X_n \to X$ in distribution. Then $E(X_n^j) \to E(X^j)$ for all $j \in \mathbb{N}$ such that $1 \leqslant j < k$.*

Theorem 1.11.4. *Let $\{X_n\}_{n\in\mathbb{N}}$ be a sequence of real random variables such that all moments exist, and $E(X_n^k) \to E(X^k)$ as $n \to \infty$, for each $k \in \mathbb{N}$. Then $X_n \to X$ in distribution, for some real random variable X, if the moments uniquely determine X (i.e., if X and Y are random variables such that $E(X^k) = E(Y^k)$ for all $k \in \mathbb{N}$, then X and Y have the same distribution).*

Proof. (Sketch) First, it can be shown that the convergence of $E(X_n^{2k})$ implies X_n^{2k} is uniformly integrable and tight. Uniform integrability means $\sup_n E\left(|X_n|^{2k} \chi_{|X_n|\geqslant a}\right) \to 0$ as $a \to \infty$. Now assume X_n does not converge to X in distribution for the sake of contradiction. Then there exists a continuous function f on $\mathbb{R}$ such that $E(f(X_n))$ does not converge to $E(f(X))$. Thus there exists a subsequence $\{n_i\}_{i\in\mathbb{N}} \subseteq \{n\}_{n\in\mathbb{N}}$ such that there exists $\varepsilon > 0$ such that for all $N \in \mathbb{N}$ there exists $i \geqslant N$ such that $|E(f(X_{n_i})) - E(f(X))| \geqslant \varepsilon$. Since $\{X_n\}$ is a tight sequence we have that $\{X_{n_i}\}_{i\in\mathbb{N}}$ is also a tight sequence. By Helly's selection principle there exists a random variable Y and a subsequence $\{n'_i\}_{i\in\mathbb{N}} \subseteq \{n_i\}_{i\in\mathbb{N}}$ such that $X_{n'_i} \to Y$ in distribution.

One can show using the fact $E(|X|^k)^2 \leqslant E(|X|^{2k})$ and a Cauchy estimate that $\{X_n\}_{n\in\mathbb{N}}$ is uniformly integrable as well. Hence $\sup_{n\in\mathbb{N}} E(|X_n|^k) < \infty$ so that $E(X_{n'_i}^k) \to E(Y^k)$ by the previous proposition. However recall by assumption $E(X_n^k) \to E(X^k)$ so that $E(X_{n'_i}^k) \to E(X^k)$. By the uniqueness of limits we must have $E(X^k) = E(Y^k)$ for all $k \in \mathbb{N}$. By our assumption that the moments uniquely determine a random variable, X and Y have the same distribution. Since we showed $X_{n'_i} \to Y$ we have $E(f(X_{n'_i})) \to E(f(Y))$. Since we had $|E(f(X_{n_i})) - E(f(X))| \geqslant \varepsilon$ we can also make $|E(f(Y)) - E(f(X))| \geqslant \varepsilon$, which would contradict the fact $E(f(X)) = E(f(Y))$, which must hold since X and Y have the same distribution. □

Another question, known as "the moment problem," is given a sequence of real numbers $\{m_k\}_{k=1}^{\infty}$ (taking $m_0 = 1$), does there exist a probability measure μ that is the distribution of some random variable X such that $m_k = E(X^k) = \int x^k d\mu(x)$, for all $k \geqslant 1$? If the measure is on $\mathbb{R}$ the problem is known as the Hamburger problem. If the measure is on $[0, \infty)$ the problem is known as the Stieltjes problem. If the measure is on $[-1, 1]$ or $[0, 1]$ the problem is known as the Hausdorff problem.

Theorem 1.11.5. *(Carleman) If* $\sum_{k \in \mathbb{N}} m_{2k}^{-1/2k} = \infty$, *then the moment problem is unique if it is solvable.*

Note that $m_{2k} \geqslant m_k^2$ which gives us a bound for m_k when k is odd. We also note that the condition in the hypothesis makes sure m_k does not "grow too fast." A similar condition is given in the hypothesis below.

Theorem 1.11.6. *If* X *has a moment generating function* $M_X(t) = E(e^{tX})$ *for* $|t| < h$, *then the moments determine* X *uniquely.*

Proof. Recall $M_X(t)$ existing on $|t| < h$ implies $M_X(t) = \sum_{k \in \mathbb{N}} \frac{m_k}{k!} t^k$ has a radius of convergence given by h, so that

$$\limsup_{k \to \infty} \frac{e m_k^{1/k}}{k} \leqslant \frac{1}{h}$$

If Y is another random variable with the same moments, then $M_Y(t) = M_X(t)$ for all $|t| < h$. Using analyticity, one can show this can be extended to $z = t + is$ for all $|t| < h$ and $s \in \mathbb{R}$, so that $M_Y(t + is) = M_X(t + is)$. Taking $t = 0$ we have $M_Y(is) = M_X(is)$ which gives $\varphi_Y(s) = \varphi_X(s)$ for each $s \in \mathbb{R}$. Hence X has the same distribution as Y. □

Note the condition that $\limsup_{k \to \infty} \frac{e m_k^{1/k}}{k} \leqslant \frac{1}{h} < \infty$ implies the condition in the hypothesis of Carleman's theorem, but not vice versa.

2 Introduction to Random Matrix Theory, Wigner Matrices and Wigner's Semicirle Law

2.1 Introduction to Random Matrix Theory

A **random matrix** is an $N \times N$ matrix (can also be $N \times M$ but we will not consider those here) whose entries are random variables. It will also have a joint probability distribution which depends on N and the entries. There are many classes of random matrices. For example, we can consider real symmetric, complex self adjoint, or quaternion self-dual random matrices (each of these matrices will have real eigenvalues). We can also consider real or complex random matrices in general (which will have complex eigenvalues). We will typically be interested in the eigenvalues (sometimes eigenvectors) of these matrices.

For example, let $A = (A_{jk})_{j,k=1}^{N}$ be a real symmetric random matrix with eigenvalues $\lambda_1, ..., \lambda_N$. Since the eigenvalues are real we can order them by $\lambda_1 \leqslant \cdots \leqslant \lambda_N$. If they are unordered, every permutation has some likelihood. If we were to pick an arbitrary eigenvalue λ_{EV} (the superscript to signify it depends on N) from $\{\lambda_1, ..., \lambda_N\}$, two questions one may ask is what is the random variable λ_{EV}'s probability distribution, and does it have a notion of convergence as $N \to \infty$. For the first question, one can show by the random nature of the selection of λ_{EV} from the set of N elements $\{\lambda_1, ..., \lambda_N\}$ that $E((\lambda_{EV})^k) = \frac{1}{N} E(\lambda_1^k + \cdots + \lambda_N^k)$. If we let P_{EV} denote the distribution of λ_{EV}, then $\int x^k dP_{EV} = E((\lambda_{EV})^k) = \frac{1}{N} E(\lambda_1^k + \cdots + \lambda_N^k) = \frac{1}{N} E(\operatorname{Tr} A^k)$. One can say more generally for certain continuous and bounded functions $\int f(x)dP_{EV} = E(f(\lambda_{EV})) = \frac{1}{N}E(f(\lambda_1^k) + \cdots + f(\lambda_N^k)) = \frac{1}{N}E(\operatorname{Tr} f(A))$. To define what $f(A)$ means, we first begin with the fact that if A is diagonalizable (in the above example the property of being real symmetric implies this) then for an invertible matrix S we have

$$A = S \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{pmatrix} S^{-1}.$$

Then

$$f(A) = S \begin{pmatrix} f(\lambda_1) & & \\ & \ddots & \\ & & f(\lambda_N) \end{pmatrix} S^{-1}.$$

In both equations we have that the middle matrix is a diagonal matrix with zeros on the non diagonal entries. We will see further below that the analysis of $\frac{1}{N}\operatorname{Tr}(A^k)$ or $\frac{1}{N}\operatorname{Tr} f(A)$ helps with showing asymptotic properties of λ_{EV} as $N \to \infty$.

We consider another approach to the matrix A to address the question of whether the random variable λ_{EV} has a notion of convergence as $N \to \infty$. We define the **empirical eigenvalue measure** by

$$\mu_{EV}(A) = \frac{1}{N}\sum_{j=1}^{N} \delta_{\lambda_j}(A)$$

where recall $\delta_{\lambda_j}(A) = \begin{cases} 1 & \lambda_j \in A \\ 0 & \lambda_j \in A \end{cases}$. The measure μ_{EV} is what is called a random measure, which is a measure that is also a random variable. To address our question, we ask whether μ_{EV} converges to some random measure σ as $N \to \infty$. We define the empirical eigenvalue distribution function by $F_{EV}(t) = \mu_{EV}((-\infty, t]) = \frac{1}{N}\sum_{j=1}^{N} \delta_{\lambda_j}((-\infty, t]) = \frac{1}{N}\left|\{\lambda_j : \lambda_j \leqslant t\}\right|$ (where the absolute value symbols here denote the cardinality of the set). We say $\mu_{EV} \to \sigma$ weakly, almost surely (respectively weakly, in probability and weakly, in distribution) if for all continuous and bounded functions f on $\mathbb{R}$ we have $\int f d\mu_{EV} \to \int f d\sigma$ almost surely (respectively in probability and in distribution) as $N \to \infty$. Note $\int f d\mu_{EV}$ is a random variable. We say $\mu_{EV} \to \sigma$ converges weakly in moments and almost surely (respectively weakly in moments and in probability, and weakly in moments and in distribution) if $\int x^k d\mu_{EV} \to \int x^k d\sigma$ almost surely (respectively in probability, and in distribution) as $N \to \infty$. Note $\int x^k d\mu_{EV}$ is a random variable. Similarly to the previous question regarding A, we can obtain expressions in terms of the trace of A by noting $\int f d\mu_{EV} = \frac{1}{N}\sum_{j=1}^{N} \int f d\delta_{\lambda_j} = \frac{1}{N}\sum_{j=1}^{N} f(\lambda_j) = \frac{1}{N}\operatorname{Tr} f(A)$. Similarly $\int x^k d\mu_{EV} = \frac{1}{N}\operatorname{Tr}(A^k)$. Expressing these two integrals in terms of the trace of A and the trace of powers of A is another way to see these integral expressions are random variables. This is since the trace is a function of the eigenvalues, which are random variables in this setting. We summarize these equalities in the following proposition.

Proposition 2.1.1. *Let $A = (A_{jk})_{j,k=1}^{N}$ be a real symmetric random matrix with eigenvalues $\{\lambda_1, ..., \lambda_N\}$.*
(a) If λ_{EV} stands for an arbitrarily chosen eigenvalue from $\{\lambda_1, ..., \lambda_N\}$, then

$$E(f(\lambda_{EV})) = \int f(t)dP_{EV}(t) = \frac{1}{N}E(\operatorname{Tr} f(A))$$

$$E((\lambda_{EV})^k) = \int t^k dP_{EV}(t) = \frac{1}{N}E(\operatorname{Tr} A^k)$$

for a continuous and bounded function f*, if the expectations exist. (b) If* μ_{EV} *is the empirical eigenvalue measure, then*

$$\int f(t)d\mu_{EV}(t) = \frac{1}{N}\operatorname{Tr} f(A)$$

$$\int t^k d\mu_{EV} = \frac{1}{N}\operatorname{Tr} A^k$$

2.2 Wigner Random Matrices and Wigner's Semicircle Law

A **Wigner random matrix** is an $N \times N$ random matrix $A = (A_{jk})_{j,k=1}^{N}$ where each A_{jk} is a real random variable such that $A_{jk} = A_{kj}$, and $\{A_{jk}\}$ are i.i.d for $1 \leqslant j \leqslant k \leqslant N$ with distribution given by the distribution of a reference random variable X, where $E(X) = 0$ and $E(X^2) = 1$. The first condition on the real random variables A_{jk} implies A is real symmetric, and the third condition implies $Var(X) = 1$.

Given a Wigner random matrix A we consider the case $k = 1$ in the previous proposition.

$$\frac{1}{N}E(\operatorname{Tr} A) = \frac{1}{N}E\left(\sum_{j=1}^{N} A_{jj}\right) = \frac{1}{N} \cdot NE(X) = 0$$

For the case $k = 2$, let $B = A^2$. Then $B_{jk} = \sum_{i=1}^{N} A_{ji}A_{ik}$ so that $B_{kk} = \sum_{j=1}^{N} A_{kj}A_{jk} = \sum_{j=1}^{N} A_{jk}^2$. Hence

$$\frac{1}{N}E(\operatorname{Tr} A^2) = \frac{1}{N}E\left(\sum_{j,k=1}^{N} A_{jk}^2\right) = \frac{1}{N}E\left(\sum_{j=1}^{N} \left(A_{j1}^2 + \cdots + A_{jN}^2\right)\right)$$

$$= \frac{1}{N}E\left(\left(A_{11}^2 + \cdots + A_{1N}^2\right) + \cdots + \left(A_{N1}^2 + \cdots + A_{NN}^2\right)\right)$$

$$= \frac{1}{N}\left(\left(E(A_{11}^2) + \cdots + E(A_{1N}^2)\right) + \cdots + \left(E(A_{N1}^2) + \cdots + E(A_{NN}^2)\right)\right)$$

$$= \frac{1}{N}\left(\left(E(X^2) + \cdots + E(X^2)\right) + \cdots + \left(E(X^2) + \cdots + E(X^2)\right)\right) = \frac{1}{N}\left(N \cdot E(X^2) + \cdots + N \cdot E(X^2)\right)$$

$$= \frac{1}{N}\left(N^2 \cdot E(X^2)\right) = \frac{N^2}{N}E\left(X^2\right) = N \cdot E\left(X^2\right) = N$$

In order to make sure $\frac{1}{N}E(\operatorname{Tr} A^2)$ does not diverge when considering its asymptotics, we need to consider some scaling factor of $\frac{1}{\sqrt{N}}$. Let $\tilde{A} = \frac{1}{\sqrt{N}}A$. Then we have $\tilde{A}_{jk} = \frac{1}{\sqrt{N}}A_{jk}$ which implies $\tilde{A}_{jk}$ has the same distribution as $\frac{1}{\sqrt{N}}X$, $E(\tilde{A}_{jk}) = 0$ and $E(\tilde{A}_{jk}^2) = \frac{1}{N}$. Regarding the eigenvalues $\tilde{\lambda}_1, ..., \tilde{\lambda}_N$ of $\tilde{A}$ we have that each $\tilde{\lambda}_j = \frac{1}{\sqrt{N}}\lambda_j$ so that $\tilde{\lambda}_{EV} = \frac{1}{\sqrt{N}}\lambda_{EV}$, which implies $d\tilde{P}_{EV}(x) = \sqrt{N}dP_{EV}(\sqrt{N}x)$. Going back to the above computation, this scaling gives $\frac{1}{N}E(\operatorname{Tr}\tilde{A}^2) = 1$ so that $\int t^k d\tilde{P}_{EV}(t) = 1$. Then we can hope that $\tilde{P}_{EV}$ converges to some limit.

Actually, Wigner's semicircle law states that under certain conditions $\tilde{P}_{EV}$ converges to a semicircle measure, which is defined as having density $d\sigma = \chi_{[-2,2]}(x) \cdot \frac{1}{2\pi}\sqrt{4 - x^2}dx$. In particular, λ_{EV} converges in distribution to a random variable with semicircle distribution. I.e., $\int f dP_{EV} \to \int f d\sigma$ as $N \to \infty$ for all continuous and bounded functions f. A stronger version of this law states that the empirical eigenvalue measure converges weakly, almost surely to σ. In other words, $\int f d\mu_{EV} \to \int f d\sigma$ almost surely as $N \to \infty$, for all continuous and bounded functions f.

In the theorem below, we make use the following result, which is lef as an exercise. Show

$$\sigma_k :- \int_{\mathbb{R}} t^k d\sigma = \begin{cases} 0 & k \text{ is odd} \\ c_{k/2} & k \text{ is even} \end{cases}$$

where $c_{k/2} = \frac{1}{k/2+1}\binom{k}{k/2}$ are the even indexed Catalan numbers.

Theorem 2.2.1. *Let* $A = (A_{jk})_{j,k=1}^N$ *be an* $N \times N$ *Wigner matrix with each* A_{jk} *having the same distribution as a random variable* X *with* $E(X) = 0$ *and* $E(X^2) = 1$*. Assume that* $E(|X|^k) < \infty$ *for all* $k \geqslant 3$ *as well. Let* $\tilde{A} = \frac{1}{\sqrt{N}}A$ *and* $Y_N = \frac{1}{N}\operatorname{Tr}\tilde{A}^k$*. Then* $E(Y_N) \to \sigma_k$ *as* $N \to \infty$ *and* $Y_n \to \sigma_k$ *almost surely as* $N \to \infty$*, for each* $k \geqslant 1$*.*

Proof. (Sketch) First we show the convergence in expectation, then obtain a bound for the variance of Y_N, then show the convergence almost surely. Note $E(Y_N) = E\left(\frac{1}{N}\operatorname{Tr}\tilde{A}^k\right) = \frac{1}{N^{1+k/2}}E(\operatorname{Tr}A^k)$ and that

$$\operatorname{Tr}A^k = \sum_{\{i_1,\dots,i_k\}\in\{1,\dots,N\}} A_{i_1i_2}A_{i_2i_3}\cdots A_{i_ki_1}$$

We give a graph theoretical interpretation. Consider a graph with vertices $\{1,\dots,N\}$ and edges in $\{1,\dots,N\}^2$. Then $i = (i_1,\dots,i_k) \in \{1,\dots,N\}^k$ corresponds to a rooted directed cycle. For example, $(1,2,3,3,2,4)$ is such a cycle, which has edges in cycle form $(1,2),(2,3),(3,3),(3,2),(2,4),(4,1)$. A rooted graph means we keep track of i_1, the root. The corresponding skeleton of the graph is the undirected graph obtained from the cycle by removing all multiplicities, where we denote the edges of the skeleton by $e_1,\dots,e_j$.

Let $P(i) = E(A_{i_1i_2}\cdots A_{i_ki_1}) = E\left(\Pi_{m=1}^j (A_{e_m})^{n_m}\right) = \Pi_{m=1}^j E((A_{e_m})^{n_m})$, where each n_m are the edge multiplicites in the cycle i. If $n_m = 1$ then $E(A_{e_m}) = 0$ so that $P(i) = 0$. To see this, let i have an edge $\alpha \to \beta$ that is traversed precisely once. Then $A_{i_1i_2}\cdots A_{i_ki_1} = A_{\alpha\beta}(\text{remaining terms})$ so that $E(A_{i_1i_2}\cdots A_{i_ki_1}) = E(A_{\alpha\beta})E(\text{remaining terms}) = 0E(\text{remaining terms}) = 0$. Thus every edge in i should be traversed at least twice in our consideration, i.e.,

$$\frac{1}{N}E(\operatorname{Tr}\tilde{A}^k) = \frac{1}{N^{1+k/2}} \sum_{i \text{ such that each edge is traversed at least twice}} P(i) \tag{1}$$

Actually, we consider i where each edge is traversed precisely two times and where the skeleton is a tree with $\frac{k}{2}$ vertices. In all other cases, i is negligible and the skeleton has strictly less than $\frac{k}{2}$ vertices. We use the following lemma.

Lemma 2.2.1. *Let* $i = (i,\dots,i_k)$ *be a* k*-cycle such that each edge is traversed at least twice. Then the skeleton has at most* $\frac{k}{2}$ *edges and at most* $\frac{k}{2}+1$ *vertices. We have equality if and only if each edge is traversed precisely twice and if the skeleton is a tree.*

The proof uses the fact that the skeleton of i is a connected graph. If the skeleton is not a tree we can remove edges from loops while not removing vertices. If it is a tree then the vertices, v, and edges, e, satisfy $v = e+1$. We remark that if k is odd then equality cannot hold, so that the skeleton has always strictly less than $\frac{k}{2}+1$ vertices.

We can decompose the sum in (1) into a sum of the form

$$\frac{1}{N^{1+k/2}} \sum_{S \text{ is a skeleton}} \sum_{i \text{ has skeleton } S} P(i) \tag{2}$$

Let S be a skeleton with l vertices (so $l \leqslant \frac{k}{2}+1$). We must now answer how many cycles correspond to a given skeleton S. We first consider the case S is a tree and every edge is traversed twice. Thus we have $l = \frac{k}{2}+1$ vertices, we can choose $i_1, ..., i_l$ different in the cycle i so that for i_1 we have N possible choices of edges, for i_2 we have $N-1$ choices and so on. Thus we have

$$N(N-1)(N-2)\cdots(N-l+1) = N^l\left(1\left(1-\frac{1}{N}\right)\left(1-\frac{2}{N}\right)\cdots\left(1-\frac{l+1}{N}\right)\right)$$
$$= N^l\left(1+o\left(\frac{1}{N}\right)\right)$$

is this the number of cycles corresponding to a given skeleton. Thus we can write (2) in the form

$$\frac{1}{N^{1+k/2}}\left(\sum_{\text{S is a skeleton that is also a tree}} N^{1+k/2}\left(1+o\left(\frac{1}{N}\right)\right)+\sum_{\text{all other skeletons}}\cdots\right)$$
$$= (\text{number of rooted trees})\left(1+o\left(\frac{1}{N}\right)\right)+\text{error terms}$$

This holds if k is even. For the case S is not a tree, we have $l < \frac{k}{2}+1$ so that $l \leqslant \frac{k}{2}+1-\frac{1}{2}$. A counting argument gives us that we can write (2) in the form

$$N^l\left(1+O\left(\frac{1}{\sqrt{N}}\right)\right)+\text{error terms}$$

The error terms in our sum is of the form

$$\frac{1}{N^{1+k/2}}\sum_{\text{other skeletons}}\cdots = \frac{1}{N^{1+k/2}}\cdot N^{1+k/2-1/2}\cdot C = \frac{C}{\sqrt{N}}$$

The constant comes from the fact each $|E((A_{e_m})^{n_m})| = |E(X^{n_m})| \leqslant E\left(|X^{n_m}|\right) \leqslant E\left(\max\{1, |X|^k\}\right) \leqslant C$ for some constant C.

Given we can rewrite (2) in the ways above, it suffices to show the number of rooted trees equals σ_k. If k is odd then the number of rooted trees is 0 agreeing with σ_k when k is odd. For the case k is even, the number of rooted trees equals the number of paths on a lattice from $(0,0)$ to $\left(\frac{k}{2},\frac{k}{2}\right)$ that do not cross the line $y=x$. This equals the number of all paths minus the number of paths that cross the line $y=x$. This equals

$$\binom{k/2+k/2}{k/2}-\binom{k}{k/2+1} = \binom{k}{k/2}-\frac{k!}{(k/2+1)!(k/2-1)!} = \binom{k}{k/2}\left(1-\frac{k/2}{k/2+1}\right)$$
$$= \binom{k}{k/2}\frac{1}{k/2+1} = c_{k/2}$$

which shows the convergence in expectation in the theorem statement.

Now we show $Var(Y_N) = O\left(\frac{1}{N^2}\right)$. Observe

$$Var(Y_N) = E(Y_N^2)-E(Y_N)^2 = \frac{1}{N^{2+k}}\left(E((\operatorname{Tr}A^k)^2)-E(\operatorname{Tr}A^k)^2\right)$$

Let $i, i' \in \{1, ..., N\}^k$ be arbitrary. Recall they are of the form $i = (i_1, ..., i_k)$ and $i' = (i'_1, ..., i'_k)$. We then recall $\operatorname{Tr}A^k = \sum_i A_{i_1i_2}\cdots A_{i_ki_1}$ so that $(\operatorname{Tr}A^k)^2 = \sum_{i,i'} A_{i_1i_2}\cdots A_{i_ki_1}A_{i'_1i'_2}\cdots A_{i'_ki'_1}$. Thus

$E((\operatorname{Tr} A^k)^2) = \sum_{i,i'} P(i,i')$ where $P(i,i') = E(A_{i_1 i_2} \cdots A_{i_k i_1} A_{i'_1 i'_2} \cdots A_{i'_k i'_1})$. Additionally, recalling the fact $E(\operatorname{Tr} A^k) = \sum_i P(i)$ we have that $E(\operatorname{Tr} A^k)^2 = \sum_{i,i'} P(i)P(i')$. Hence going back to the above expression for $Var(Y_n)$ we have

$$Var(Y_N) = \frac{1}{N^{2+k}} \sum_{i,i'} P(i,i') - P(i)P(i') \tag{3}$$

where the sum is over all cycles $i, i \in \{1, ..., N\}^k$.

We now look at the skeletons of $i \cup i'$. In the case there is one edge which is traversed only once, either by i and i' but not both, or in the case the skeletons for i and i' are disjoint (no common edges) we have $P(i,i') = P(i)P(i')$ so that $P(i,i') - P(i)P(i') = 0$. Then the sum in (3) can be taken over cycles i and i' that have at least one edge in common and every edge in $i \cup i'$ is traversed at least twice. After counting the number of vertices, we can show $Var(Y_N) = O\left(\frac{1}{N^2}\right)$.

For the last step, we use the fact $E(Y_N) \to \sigma_k$ as $N \to \infty$ and $Var(Y_N) = O\left(\frac{1}{N^2}\right)$. The second condition implies $\sum_{N \in \mathbb{N}} Var(Y_N) < \infty$. By Chebyshev's inequality, we have for all $\varepsilon > 0$ that $P(|Y_N - E(Y_N)| \geqslant \varepsilon) \leqslant \frac{Var(Y_N)}{\varepsilon^2}$, which implies for arbitrarily fixed ε we have $\sum_{N \in \mathbb{N}} P(|Y_N - E(Y_N)| \geqslant \varepsilon) \leqslant \sum_{N \in \mathbb{N}} \frac{1}{\varepsilon^2} Var(Y_N) \leqslant \infty$. We use the Borel-Cantelli lemma, which is stated and proved below.

Lemma 2.2.2. *(Borel-Cantelli) If $\{E_n\}_{n \in \mathbb{N}}$ is a sequence of events and $\sum_{n \in \mathbb{N}} P(E_n) < \infty$, then $P\left(\bigcap_{n=1}^\infty \bigcup_{m=n}^\infty E_m\right) = 0$.*

Proof.

$$P\left(\bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} E_m\right) \leqslant P\left(\bigcup_{m=n}^{\infty} E_m\right) \leqslant \sum_{n=m}^{\infty} P(E_m) \to 0$$

as $n \to \infty$ by assumption of $\sum_{n \in \mathbb{N}} P(E_n) < \infty$. □

Now by the Borel-Cantelli lemma we have $P\left(\limsup_{N \to \infty} |Y_N - E(Y_N)| \geqslant \varepsilon\right) = 0$. We now prove a secondary lemma.

Lemma 2.2.3. *If $X \geqslant 0$ is a random variable such that $P(X \geqslant \varepsilon) = 0$ for all $\varepsilon > 0$, then $P(X > 0) = 0$.*

Proof. Observe

$$P(X > 0) = P\left(\text{ there exists } m \in \mathbb{N} \text{ such that } X > \frac{1}{m}\right) = P\left(\bigcup_{m \in \mathbb{N}} \{X > \frac{1}{m}\}\right)$$

$$\leqslant \sum_{m \in \mathbb{N}} P\left(X > \frac{1}{m}\right) = 0$$

□

By the secondary lemma we have

$$P\left(\limsup_{N \to \infty} |Y_N - E(Y_N)| > 0\right) = 0 \implies P\left(\lim_{N \to \infty} |Y_N - E(Y_N)| = 0\right) = 1$$

$$\implies Y_N - E(Y_N) \to 0$$

almost surely. Since $E(Y_N) - \sigma_k \to 0$ as $N \to \infty$, we have $Y_N - \sigma_k \to 0$ almost surely. □

We note $E(Y_N) = E\left(\frac{1}{N}\operatorname{Tr}\tilde{A}^k\right) = E\left(\frac{\tilde{\lambda}_1^k + \cdots + \tilde{\lambda}_N^k}{N}\right) = E(\tilde{\lambda}_{EV}^k)$. Thus $E(Y_N)$ is the k^{th} moment of $\tilde{\lambda}_{EV}$. Since $E(Y_N) \to \sigma_k$ we have that the moments of $\tilde{\lambda}_{EV}$ converges to the moments of a random variable with semicircle distribution. Since the σ_k satisfies the Carleman condition, $\tilde{\lambda}_{EV}$ converges in distribution to a random variable with semicircle distribution.

We also note Y_N is the "moment" of the empirical eigenvalue measure $\mu_{EV} = \frac{1}{N}\sum_{j=1}\delta_{\tilde{\lambda}_j}$. The theorem implies that $\int x^k d\mu_{EV} = Y_N \to \int x^k d\sigma$ almost surely, for all $k \geqslant 1$. Thus by linearity we have $\int p(x) d\mu_{EV} \to \int p(x) d\sigma$ for each polynomial $p(x)$.

A variant of the semicircle theorem is given below.

Theorem 2.2.2. *Let $A = (A_{jk})_{j,k=1}^N$ be an $N \times N$ Wigner matrix with each A_{jk} having the same distribution as a random variable X with $E(X) = 0$ and $E(X^2) = 1$. Then the empirical eigenvalue measure of A converges weakly, almost surely to a random variable with semicircle distribution. In other words, for all bounded and continuous functions on $\mathbb{R}$ we have $\int f(x) d\mu_{EV} \to \int f(x) d\sigma$ almost surely.*

Theorem 2.2.3. *Let $A = (A_{jk})_{j,k=1}^N$ be an $N \times N$ Wigner matrix with each A_{jk} having the same distribution as a random variable X with $E(X) = 0$ and $E(X^2) = 1$. Assume that $E(|X|^k) < \infty$ for all $k \geqslant 3$ as well. Let $\tilde{A} = \frac{1}{\sqrt{N}}A$ and $Y_N = \frac{1}{N}\operatorname{Tr}\tilde{A}^k$. Also let $Z_N = N\left(Y_N - E(Y_N)\right)$. Then Z_N converges in distribution to a random variable with Gaussian distribution with mean 0 and variance σ_k^2.*

One can show

$$\sigma_k^2 = k^2 C_{(k-1)/2}^2 + \frac{k^2}{2} C_{k/2}^2 (E(X^4) - 2) + \sum_{r \geqslant 3} \frac{2k^2}{r} \left(\sum_{k_i \geqslant 0 \text{ and } k_1 + \cdots k_r = (k-2)/2} C_{k_1} + \cdots + C_{k_r} \right)^2$$

The above theorem is shown by evaluating the asymptotics of $E(Z_N^l)$ as $N \to \infty$, for each $l \geqslant 1$. When $l = 1$ we have $E(Z_N) = 0$. When $l = 2$ we have $E(Z_N^2) = N^2 Var(Y_N) = N^2\left(\frac{\sigma_k^2}{N^2}\left(1 + O\left(\frac{1}{N}\right)\right)\right) = \sigma_k^2 + O\left(\frac{1}{N}\right)$. When $l \geqslant 3$ we have $E(Z_N^l) = \frac{1}{N^{kl/2}} \sum_{l \text{ cycles } i = (i_1, \ldots, i_l)} E(G(i_1) \cdots G(i_l))$, where $G(i) = A_{i_1 i_2} \cdots A_{i_N i_1} - E(A_{i_1 i_2} \cdots A_{i_N i_1})$. One has to show $E(Z_N^l)$ converges to

$$\frac{1}{\sqrt{2\pi\sigma_k}} \int x^l e^{-x^2/(2\sigma_k)} dx = \begin{cases} 0 & l \text{ is odd} \\ \sigma_k^l (l-1)!! & l \text{ is even} \end{cases}$$

where $(l-1)!! = (l-1)(l-1)\cdots(3)(1)$.

3 Wishart Random Matrices and the Marchenko-Pateur Law

Consider a matrix $A = (A_{jk})$ where $j = 1, \ldots, N$ and $k = 1, \ldots, M$, with i.d.d random variable entries. Let each A_{jk} have the same distribution as that of a reference random variable X, $E(X) = 0$ and $E(X^2) = 1$. A **Wishart matrix** is the matrix $B = \frac{1}{N}A^T A$. Note B is an $M \times M$ symmetric and positive definite matrix. Recall this means $B = B^T$ and $y^T B y \geqslant 0$ for all $y \in \mathbb{R}^M$ (this second condition implies all eigenvalues will be nonegative). In the case $M < N$, then the rank of B is less than or equal to N, so that B has $N - M$ eigenvalues that have value 0. We avoid this by instead considering $\tilde{B} = \frac{1}{N}AA^T$, so that $\tilde{B}$ and B have the same nonzero eigenvalues. Consider a randomly chosen eigenvalue λ_{EV} from the eigenvalues of $B = \frac{1}{N}A^T A$. We will consider the "double scaling limit," i.e., the asymptotics of λ_{EV} as $N, M \to \infty$, under the condition $\frac{M}{N} \to \eta$ for some $\eta \in (0, 1]$. Before giving the next theorem that answers this question, we define the **Marchenko-Pasteur distribution** with parameter η to the be the distribution of a random variable that has density $\rho_\eta(x) = \frac{1}{2\pi x}\chi_{[a,b]}(x)\sqrt{(b-x)(x-a)}$, where $a = (1-\sqrt{\eta})^2$ and $b = (1+\sqrt{\eta})^2$.

Theorem 3.0.1. *(Marchenko-Pasteur) Let* A *and* X *be as above. Assume* $E(|X|^k) < \infty$ *for all* $k \geqslant \mathbb{N}$. *Then the* λ_{EV} *converges in distribution to a random variable with Marchenko-Pasteur distribution with parameter* $\eta \in (0,1]$ *in the limit* $M, N \to \infty$, *where* $\frac{M}{N} \to \eta$.

Theorem 3.0.2. *Let* A *and* X *be as above. Assume* $E(|X|^k) < \infty$ *for all* $k \geqslant \mathbb{N}$. *Then* $\frac{1}{N}\operatorname{Tr} B^k \to c_k$ *in expectation and almost surely, for* $k \geqslant 0$, *where* c_k *are the Catalan numbers.*

We note that $c_k = \frac{1}{2\pi}\int_{\mathbb{R}} x^{2k}\sqrt{4-x^2}dx = \frac{1}{\pi}\int_0^\infty x^{2k}\sqrt{4-x^2}dx$. If we set $y = x^2$ then $c_k = \frac{1}{2\pi}\int_{\mathbb{R}} y^k\sqrt{\frac{4-y}{y}}dy$.

The importance of Wishart matrices arise in statistics. Consider a situation where one wants to observe N quantities and has M samples for each quantity. One can ask the question: is there a correlation between the N quantities? Let each quantity be represented by the random variables $\tilde{X}_1, ..., \tilde{X}_N$. If $j \in \{1, ..., N\}$ and $k \in \{1, ..., M\}$, then one can consider the random vectors $(X_{1,k}, ..., X_{N,k}) :- (\tilde{X}_1, ..., \tilde{X}_N)$ for each $k = 1, ..., M$. Thus each $X_{j,k}$ has the same distribution as $\tilde{X}_j$, for each $j = 1, ..., N$. We can construct the matrix

$$\begin{pmatrix} X_{1,1} & \cdot & \cdot & \cdot & X_{N,1} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ X_{1,M} & \cdot & \cdot & \cdot & X_{N,M} \end{pmatrix}$$

Then we consider the $N \times N$ matrix $B = \frac{1}{M}X^TX$, where each $B_{ij} = \frac{1}{M}\sum_{k=1}^M X^T_{i,k}X_{k,j} = \frac{1}{M}\sum_{k=1}^M X_{k,i}X_{k,j}$. For each fixed N, taking $M \to \infty$ gives each $B_{i,j} \to E(\tilde{X}_i\tilde{X}_j)$ almost surely by the law of large numbers. Then the matrix B will converge to the covariance matrix (briefly discussed in our probability theory review) of the random variables $\tilde{X}_1, ..., \tilde{X}_N$ almost surely as $M \to \infty$. This matrix gives us information as to whether these random variables are correlated or not. A problem occurs when N is also simultaneously large (which corresponds to having many parameters to consider in our model). If the $\tilde{X}_1, ..., \tilde{X}_N$ are independent, then each entry $X_{k,j}$ in X is independent. Then B will be a Wishart random matrix, so that a randomly chosen eigenvalue converges in distribution to a random variable with Marchenko-Pasteur distribution. If the randomly chosen eigenvalue from B does not have this convergence, it indicates the random variables $\tilde{X}_1, ..., \tilde{X}_N$ may not be independent.

4 The Gaussian Matrix Ensembles

4.1 The Gaussian Matrix Ensembles: The Matrix Entries

Here we introduce the **Gaussian Orthogonal Ensemble**, **Gaussian Unitary Ensemble** and **Gaussian Symplectic Ensemble**, which we will abbreviate by GOE, GUE and GSE respectively. An $N \times N$ matrix $A = (A_{jk})_{j,k=1}^N$ from the GOE will satisfy $A = A^T$, where A^T denotes the transpose of A, and each entry $A_{jk} = A_{kj}$ and is real valued. An $N \times N$ matrix $A = (A_{jk})_{j,k=1}^N$ from the GUE will satisfy $A = A^*$, where A^* denotes the conjugate transpose of A, and each entry $A_{jk} = \overline{A}_{kj}$ and is complex valued. An $N \times N$ matrix $A = (A_{jk})_{j,k=1}^N$ from the GSE will satisfy $A = A^+$, where A^+ denotes the dual of A, and each entry $A_{jk} = A^+_{kj}$ and is quaternion valued. Note in each case, the conditions imposed on A imply A will have N real eigenvalues.

To deal with quaternion valued matrices, we briefly discuss the quaternions and some related notions. We define the quaternions as numbers of the form $a = a_0 + ia_1 + ja_2 + ka_3$ where

$a_0, a_1, a_2, a_3 \in \mathbb{R}$. We define addition of quaternions component wise. We define multiplication by first defining $i^2 = j^2 = k^2 = -1$, $ij = k, jk = i, ki = j$ and $ji = -k, kj = -i, ik = -j$, then by extending with the distributive law. We let the q-conjugate of a be $a^+ = a_0 - ia_1 - ja_2 - ka_3$. One can then show $a^+ a = |a|^2 = a_0^2 + a_1^2 + a_2^2 + a_3^2$ so that $a^{-1} = \frac{a^+}{|a|^2}$. We can represent quaternions by 2×2 matrices by identifying $1, i, j, k$ with

$$\begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} i & 0 \\ 0 & i \end{pmatrix}, \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & i \\ i & 0 \end{pmatrix}$$

respectively. Then we can identify a and a^+ with

$$\begin{pmatrix} a_0 + ia_1 & a_2 + ia_3 \\ -a_2 + ia_3 & a_0 - ia_1 \end{pmatrix} \text{ and } \begin{pmatrix} a_0 - ia_1 & -a_2 - ia_3 \\ a_2 - ia_3 & a_0 + ia_1 \end{pmatrix}$$

respectively. Note the matrix representing a^+ is the conjugate transpose of that of a. One can show that a complex 2×2 matrix, a, is a quaternion if and only if

$$\overline{a} = \begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix} a \begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$$

For a matrix A from the GOE, we take the entries to be independent random variables with Gaussian distribution. The diagonal entries A_{jj} have $N(0, 1)$ distribution, and the below diagonal entries (and thus the above diagonal entries as well by symmetry) have $N(0, \frac{1}{2})$ distribution. For a matrix A from the GUE, we also take the entries to be independent random variables with Gaussian distribution. The diagonal entries A_{jj} must be real random variables as they equal their complex conjugate, and have $N(0, 1)$ distribution. The below diagonal entries (which will determine the above diagonal entries by the fact A equals its conjugate transpose) have real and imaginary parts that are random variables with $N(0, \frac{1}{2})$ distribution. A matrix from the GSE is defined similarly, but we instead have two extra components to consider for the entries than in the GUE case.

4.2 The Gaussian Matrix Ensembles: The Joint Probability Density Function for the Matrix Entries

Let $A = (A_{jk})_{j,k=1}^{M}$ be from the GOE. Then the joint density function for its entries, given they are independent, will be

$$\exp\left(-\frac{1}{2}\sum_{j=1}^{N} A_{jj}^2 - \sum_{j<k} A_{jk}^2\right) \Pi_{j=1}^{N} dA_{jj} \Pi_{j<k=1}^{N} dA_{jk}$$

But note

$$\mathrm{Tr}(A^2) = -\frac{1}{2}\sum_{1\leqslant j,k\leqslant N} A_{jk}^2 = -\frac{1}{2}\sum_{j=1}^{N} A_{jj}^2 + -\frac{1}{2}\sum_{1\leqslant j<k\leqslant N} A_{jk}^2 + -\frac{1}{2}\sum_{1\leqslant k<j\leqslant N} A_{jk}^2$$

$$= -\frac{1}{2}\sum_{j=1}^{N} A_{jj}^2 - \sum_{1\leqslant j<k\leqslant N} A_{jk}^2$$

so that the joint density function of A (i.e., that of its entries) is given by $\exp\left(-\frac{1}{2}\operatorname{Tr}(A^2)\right) dA$. The term dA comes from the product measure of the matrix entries and their independence. Similarly to the GOE case, a matrix A from the GUE will have density

$$\exp\left(-\frac{1}{2}\sum_{j=1}^{N} A_{jj}^2 - \sum_{j<k}\left((\operatorname{Re} A_{jk})^2 + (\operatorname{Im} A_{jk}^2)\right)\right) \Pi_{j=1}^{N} dA_{jj}\Pi_{j<k} d(\operatorname{Re} A_{jk})d(\operatorname{Im} A_{jk})$$

and this will also be equal to $\exp\left(-\frac{1}{2}\operatorname{Tr}(A^2)\right)$. The GSE case is similar.

4.3 Invariance Under Transformations for the Gaussian Ensembles

An orthogonal transformation is a transformation $A \mapsto O^TAO$ where O is a real matrix such that $O^TO = I$. A unitary transformation is a transformation $A \mapsto U^*AU$ where U is a complex matrix such that $U^*U = I$. A symplectic transformation is a transformation $A \mapsto S^+AS$ where S is a quaternion matrix such that $S^+S = I$.

The probability distribution measure on the GOE is invariant under orthogonal transformations. That is, if K is a subset of the real symmetric matrices, then $P(A \in K) = P(O^TAO \in K)$. This means, using the above expressions for the density, at a high level using symbolic notation

$$\int_K \exp\left(-\frac{1}{2}\operatorname{Tr}(A^2)\right) dA = \int_K \exp\left(-\frac{1}{2}\operatorname{Tr}((O^TAO)^2)\right) d(O^TAO)$$

Since $\operatorname{Tr}((O^TAO)^2) = \operatorname{Tr}(O^TAOO^TAO) = \operatorname{Tr}(O^TAAO) = \operatorname{Tr}(O^TA^2O) = \operatorname{Tr}(A^2O^TO) = \operatorname{Tr}(A^2O^{-1}O) = \operatorname{Tr}(A^2)$, we have the above becomes

$$\int_K \exp\left(-\frac{1}{2}\operatorname{Tr}(A^2)\right) d(O^TAO)$$

Thus to show the invariance of the probability distribution for the GOE under orthogonal transformations, one can show $dA = d(O^TAO)$, which we develop in the following lemmas and the theorem that follows.

Lemma 4.3.1. *Let* $X = (X_1, ..., X_m)^T$ *be a random vector whose components are i.i.d with distribution* $N(0, 1)$*, and* $Y = (X_1, ..., X_m)^T$ *where* $Y = TX$ *for a fixed* $m \times m$ *real matrix* T*. Then* Y *is also a random vector whose components are i.i.d with distribution* $N(0, 1)$ *if and only if* $T^TT = I$.

Proof. Consider the joint characteristic function of X given by

$$\varphi_X(t_1, ..., t_m) = E\left(e^{i\sum_{i=1}^m t_iX_i}\right) = E\left(e^{it_1X_1}\right)\cdots E\left(e^{it_mX_m}\right)$$

$$= \frac{1}{\sqrt{2\pi}}e^{-t_1^2/2}\cdots\frac{1}{\sqrt{2\pi}}e^{-t_m^2/2} = \frac{1}{(2\pi)^{m/2}}e^{-\|t\|^2/2}$$

where $t = (t_1, ..., t_m)$. Note Y is also a random vector whose components are i.i.d with distribution $N(0, 1)$ if and only if $\varphi_X = \varphi_Y$. Observe

$$\varphi_Y(t_1, ..., t_m) = E\left(e^{i\sum_{i=1}^m t_iY_i}\right) = E\left(e^{i<t,Y>}\right) = E\left(e^{i<t,TX>}\right)$$

$$= E\left(e^{i<T^Tt,X>}\right) = E\left(e^{i\sum_{i=1}^m (T^Tt)_iX_i}\right) = E\left(e^{i(T^Tt)_1X_1}\right)\cdots E\left(e^{i(T^Tt)_mX_m}\right)$$

$$= \frac{1}{\sqrt{2\pi}} e^{-(T^\mathsf{T} t)_1^2/2} \cdots \frac{1}{\sqrt{2\pi}} e^{-(T^\mathsf{T} t)_m^2/2} = \frac{1}{(2\pi)^{m/2}} e^{-\|T^\mathsf{T} t\|^2/2}$$

where the fourth equality is since for real matrices, the adjoint, i.e. the conjugate transpose, is just the transpose. Thus we must require $\|t\|^2 = \|T^\mathsf{T} t\|^2$ for all $t \in \mathbb{R}^m$. In other words, we must require $< t, t >=< t, TT^\mathsf{T} t >$ for all $t \in \mathbb{R}^m$, which is if and only if $< t, (TT^\mathsf{T} - I)t >= 0$ for all $t \in \mathbb{R}^m$. This is if and only if $TT^\mathsf{T} - I = 0$, i.e., $TT^\mathsf{T} = I$. □

We say a $N \times N$ matrix O is an **elementary orthogonal matrix** if

$$O = P^{-1} \begin{pmatrix} \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix} & 0 \\ 0 & I_{N-2} \end{pmatrix} P$$

where P is a permutation matrix, I_{N-2} is the $N-2 \times N-2$ identity matrix, $\alpha^2 + \beta^2 = 1$ and the zeros in the matrix indicate all remaining entries besides the diagonal blocks are 0.

Lemma 4.3.2. *Each orthogonal matrix is the product of elementary orthogonal matrices and a diagonal matrix with diagonal entries* ±1.

Lemma 4.3.3. *Let* X *be the matrix*

$$X = \begin{pmatrix} X_1 & \frac{1}{\sqrt{2}} X_{12} \\ \frac{1}{\sqrt{2}} X_{12} & X_2 \end{pmatrix}.$$

where X_1, X_2, X_{12} *are i.i.d random variables with distribution* $N(0,1)$. *Let* O *be the matrix*

$$O = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}$$

where $\alpha^2 + \beta^2 = 1$, *and* $Y = O^\mathsf{T} X O$. *Then* Y *can be written in the form*

$$Y = \begin{pmatrix} Y_1 & \frac{1}{\sqrt{2}} Y_{12} \\ \frac{1}{\sqrt{2}} Y_{12} & Y_2 \end{pmatrix}$$

and Y_1, Y_2, Y_{12} *are i.i.d random variables with distribution* $N(0,1)$.

Proof. Observe $Y^\mathsf{T} = (O^\mathsf{T} X O)^\mathsf{T} = O^\mathsf{T} X^\mathsf{T} (O^\mathsf{T})^\mathsf{T} = O^\mathsf{T} X^\mathsf{T} O = O^\mathsf{T} X O$ since X is symmetric, so that Y is also symmetric. Adding a factor of $\frac{1}{\sqrt{2}}$ gives us the above form of Y. The rest of the proof relies on the fact

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_{12} \end{pmatrix} = \begin{pmatrix} \alpha^2 & \beta^2 & \sqrt{2}\alpha\beta \\ \beta^2 & \alpha^2 & -\sqrt{2}\alpha\beta \\ -\sqrt{2}\alpha\beta & \sqrt{2}\alpha\beta & \alpha^2 - \beta^2 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_{12} \end{pmatrix}$$

which can be shown through computation. Letting the above transformation matrix be denoted by T, one can show $T^\mathsf{T} T = I$ so that the result follows from the previous lemma. □

Theorem 4.3.1. *The Gaussian Orthogonal Ensemble is invariant under orthogonal transformations* $X \to O^\mathsf{T} X O$.

Proof. (Sketch) We instead show invariance under transformations by elementary orthogonal matrices and diagonal matrices by the second of the previous lemmas, and then apply the previous lemma. Without loss of generality, one can take $P = I$ in the definition of elementary orthogonal matrices, so that X is of the form

$$X = \begin{pmatrix} \begin{pmatrix} X_1 & \frac{1}{\sqrt{2}}X_{12} \\ \frac{1}{\sqrt{2}}X_{12} & X_2 \end{pmatrix} & \begin{pmatrix} \frac{1}{\sqrt{2}}X_{13} & \frac{1}{\sqrt{2}}X_{14} & \cdots \\ \frac{1}{\sqrt{2}}X_{23} & \frac{1}{\sqrt{2}}X_{24} & \cdots \end{pmatrix} \\ \begin{pmatrix} \frac{1}{\sqrt{2}}X_{13} & \frac{1}{\sqrt{2}}X_{23} \\ \frac{1}{\sqrt{2}}X_{14} & \frac{1}{\sqrt{2}}X_{24} \\ \vdots & \vdots \end{pmatrix} & \text{unaffected block} \end{pmatrix}$$

Then by the above theory, we consider

$$Y = \begin{pmatrix} \begin{pmatrix} \alpha & \beta \\ -\beta & \alpha \end{pmatrix} & 0 \\ 0 & I_{N-2} \end{pmatrix} X \begin{pmatrix} \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix} & 0 \\ 0 & I_{N-2} \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} Y_1 & \frac{1}{\sqrt{2}}Y_{12} \\ \frac{1}{\sqrt{2}}Y_{12} & Y_2 \end{pmatrix} & \begin{pmatrix} \frac{1}{\sqrt{2}}Y_{13} & \frac{1}{\sqrt{2}}Y_{14} & \cdots \\ \frac{1}{\sqrt{2}}Y_{23} & \frac{1}{\sqrt{2}}Y_{24} & \cdots \end{pmatrix} \\ \begin{pmatrix} \frac{1}{\sqrt{2}}Y_{13} & \frac{1}{\sqrt{2}}Y_{23} \\ \frac{1}{\sqrt{2}}Y_{14} & \frac{1}{\sqrt{2}}Y_{24} \\ \vdots & \vdots \end{pmatrix} & \text{unaffected block} \end{pmatrix}$$

We note that in the above we have

$$\begin{pmatrix} Y_{1m} \\ Y_{2m} \end{pmatrix} = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix} \begin{pmatrix} X_{1m} \\ X_{2m} \end{pmatrix}$$

for each $m = 1, ..., N$. Thus using this fact and the matrix from the proof of the previous lemma, we have

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_{12} \\ Y_{13} \\ Y_{23} \\ \vdots \\ Y_{1n} \\ Y_{2n} \end{pmatrix} = \begin{pmatrix} \begin{pmatrix} \alpha^2 & \beta^2 & \sqrt{2}\alpha\beta \\ \beta^2 & \alpha^2 & -\sqrt{2}\alpha\beta \\ -\sqrt{2}\alpha\beta & \sqrt{2}\alpha\beta & \alpha^2 - \beta^2 \end{pmatrix} & & & \\ & \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix} & & \\ & & \ddots & \\ & & & \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_{12} \\ X_{13} \\ X_{23} \\ \vdots \\ X_{1n} \\ X_{2n} \end{pmatrix}$$

with zeros in the remaining entries in the above matrix transformation. Since each block is an orthogonal matrix, the whole matrix is orthogonal. Thus, since the vector containing X_1, X_2, X_{12}, X_{13}, $X_{23}, ..., X_{1n}, X_{2n}$ are i.i.d with distribution $N(0, 1)$, we have that the vector containing Y_1, Y_2, Y_{12}, Y_{13}, $Y_{23}, ..., Y_{1n}, Y_{2n}$ are i.i.d with distribution $N(0, 1)$. Since we disregard the unaffected block, we have by symmetry that the remaining entries in Y are also i.i.d with distribution $N(0, 1)$, hence showing the result. □

Consider random matrices with a probability measure invariant under orthogonal transformations and having independent entries, but had diagonal entries with distribution $N(c_1, \frac{1}{c_2})$ and non-diagonal entries with distribution $N(0, \frac{1}{2c_2})$. Note if $c_2 = 1$ and $c_1 = 0$ we just have a matrices from the GOE. However in his more general case, we would have a density function other than $\exp\left(-\frac{1}{2}\operatorname{Tr}(A^2)\right)$. The following theorem gives the density for this more general class of matrices.

Theorem 4.3.2. *Let* μ *be a probability measure on the set of* $N \times N$ *real symmetric matrices such that (1)* μ *is invariant under orthogonal transformations and (2) all entries* $A_{j,k}$ *are indepedent where* $j \leqslant k$ *for a given matrix* $A = (A_{jk})_{j,k=1}^N$*. Then* $d\mu = \exp\left(-c_2 \operatorname{Tr}(A^2) - c_1 \operatorname{Tr} A - c_0\right) dA$*, where* $c_2 > 0$ *and* $c_1, c_0 \in \mathbb{R}$.

Proof. (Sketch) We assume μ is an absolutely continuous measure so that $d\mu = f(A)dA$ for some function A. We already noted dA is invariant under orthogonal transformations so that $f(A)$ must be invariant under orthogonal transformations, so that $d\mu$ is as well. In other words, we need to show (1) $f(A) = f(O^T AO)$ for all orthogonal matrices O and real symmetric matrices A, and (2) $f(A) = f((A_{jk})_{j\leqslant k})$ should be of the form $\prod_{j\leqslant k} f_{jk}(A_{jk})$. One can diagonalize A so that $f(A)$ is a function depending only on the eigenvalues of A.

The diagonal form of A can also be multiplied by a permutation matrix, which are orthogonal, to permute the eigenvalues. In other words, $f(A)$ is symmetric in the eigenvalues. The Lagrange-Newton formula says a symmetric function in $\lambda_1, ..., \lambda_N$ is a function in $s_j = \lambda_1^j + \cdots + \lambda_N^j$ for $j = 1, ..., N$, where we note each $s_j = \operatorname{Tr}(A^j)$. Therefore, we can write $f(A) = \tilde{f}(s_1, ..., s_N) = \tilde{f}(\operatorname{Tr} A, \operatorname{Tr}(A^2), ..., \operatorname{Tr}(A^N))$. One can show f will not depend on $\operatorname{Tr}(A^j)$ for $j \geqslant 3$. To see this, observe for example that $\operatorname{Tr}(A^3) = A_{11}^3 + A_{11}A_{12}^2 + \cdots$, so that the trace has terms that are products of two different entries, and we cannot get rid of this dependence. However, $\operatorname{Tr} A$ and $\operatorname{Tr}(A^2)$ has no such terms. Thus $f(A) = \tilde{f}(\operatorname{Tr} A, \operatorname{Tr}(A^2))$. Using the independence of the entries in A one can solve this equation for f and show $f(A) = \tilde{f}(\operatorname{Tr} A, \operatorname{Tr}(A^2)) = \exp\left(-c_2 \operatorname{Tr}(A^2) - c_1 \operatorname{Tr} A - c_0\right)$. □

We now give some of the analogous lemmas needed to show a probability measure on the GUE is invariant under unitary transformations $A \to U^*AU$.

We say an $N \times N$ matrix U is an **elementary unitary matrix** if

$$U = P^{-1} \begin{pmatrix} \begin{pmatrix} a & c \\ b & d \end{pmatrix} & 0 \\ 0 & I_{N-2} \end{pmatrix} P$$

where P is a permutation matrix, I_{N-2} is the $N-2 \times N-2$ identity matrix. In the 2×2 case (particularly in the first block on the matrix in between P^{-1} and P), we can say

$$\begin{pmatrix} a & c \\ b & d \end{pmatrix} = \begin{pmatrix} J_1 & 0 \\ 0 & J_2 \end{pmatrix} \begin{pmatrix} \alpha & \gamma \\ \beta & \delta \end{pmatrix}$$

where $|J_1| = |J_2| = 1$, $\alpha, \beta \geqslant 0$ and $\alpha^2 + \beta^2 = \alpha\overline{\gamma} + \beta\overline{\delta} = |\gamma| + |\delta| = 1$.

Lemma 4.3.4. *Each unitary matrix is the product of elementary unitary matrices and a diagonal matrix with diagonal entries of modulus* 1.

By the above lemma, showing invariance under unitary transformation reduces to showing invariance under transformations with diagonal and elementary unitary matrices.

Lemma 4.3.5. *Let* X *be the matrix*

$$X = \begin{pmatrix} X_1 & \frac{X_{12}+i\tilde{X}_{12}}{\sqrt{2}} \\ \frac{X_{12}-i\tilde{X}_{12}}{\sqrt{2}} & X_2 \end{pmatrix}.$$

where $X_1, X_2, X_{12}, \tilde{X}$ *are i.i.d random variables with distribution* $N(0, 1)$*. Let* O *be the matrix*

$$O = \begin{pmatrix} \alpha & -\beta \\ \beta & \alpha \end{pmatrix}$$

where $\alpha^2 + \beta^2 = 1$, *and* $Y = O^T X O$. *Then* Y *can be written in the form*

$$Y = \begin{pmatrix} Y_1 & \frac{Y_{12}+i\tilde{Y}_{12}}{\sqrt{2}} \\ \frac{Y_{12}-i\tilde{Y}_{12}}{\sqrt{2}} & Y_2 \end{pmatrix}$$

and $Y_1, Y_2, Y_{12}, \tilde{Y}_{12}$ *are i.i.d random variables with distribution* $N(0,1)$.

Proof. Similarly to the GOE case, the proof relies on the fact

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_{12} \\ \tilde{Y}_{12} \end{pmatrix} = \begin{pmatrix} \alpha^2 & \beta^2 & \sqrt{2}\alpha\beta & 0 \\ \beta^2 & \alpha^2 & -\sqrt{2}\alpha\beta & 0 \\ -\sqrt{2}\alpha\beta & \sqrt{2}\alpha\beta & \sqrt{\alpha^2-\beta^2} & 0 \\ 0 & 0 & 0 & \alpha^2+\beta^2 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_{12} \\ \tilde{X}_{12} \end{pmatrix}$$

which can be shown by computation □

Lemma 4.3.6. *If* $X, \tilde{X}$ *are i.i.d random variables with distribution* $N(0,1)$ *and* $Y + \tilde{Y} = (\cos\theta + i\sin\theta)(X + i\tilde{X})$, *then* $Y, \tilde{Y}$ *are also i.i.d random variables with distribution* $N(0,1)$.

Proof. One uses the fact

$$\begin{pmatrix} Y \\ \tilde{Y} \end{pmatrix} = \begin{pmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{pmatrix} \begin{pmatrix} X \\ \tilde{X} \end{pmatrix}$$

□

We now introduce some of the physical motivation for analyzing the GOE and GSE. Quantum mechanics tells us that a system can be represented by a self adjoint operator on a Hilbert space, known as the Hamiltonian, given by $H = H^*$. The real eigenvalues of H describe the energy levels of the system. Since H could have infinite components, we replace it with an $N \times N$ random matrix A, for N large, and then look at the behavior of the eigenvalues as $N \to \infty$. There is a notion of symmetry for this system by the fact we require $A = A^*$, but there could be other notions of symmetry. An example is **time reversal symmetry**, also known as time reversal invariance. The time reversal operator can be written as $T = KC$, where C is complex conjugation (i.e, the vector $x = (x_j) \to Cx = (\overline{x}_j))$ and K is a unitary matrix that is either symmetric or antisymmetric. In other words, $K = \pm K^T$ and $K^*K = I$. This holds if and only if $\overline{K}K = \pm I$. Time reversing a system A is given by the transformation $A \to A^+ :- KA^TK^{-1} = TA^*T^{-1}$. Note this is the dual of a matrix discussed earlier. A system being time reversal invariant means $A = A^+$. By unitary transformations, one can transform K into two ways given these two cases. In one way, without loss of generality, we can take $K = I$. Then being time reversal invariant gives $A = KA^TK^{-1} = A^T$. This means A belongs to the family of real symmetric matrices, hence reason to consider the GOE. In another way, without loss of generality, we can take K to be a block diagonal matrix with diagonal blocks equal to $\begin{pmatrix} 0 & 1 \\ -1 & 0 \end{pmatrix}$. Then the condition $A = KA^TK^{-1}$ implies A belongs to the family of $2N \times 2N$ matrices such that $A = A^+ = A^*$, i.e., quaternion valued self dual matrices, hence reason to consider the GSE.

4.4 Joint Eigenvalue Density Function for the Gaussian Ensembles

Now consider a matrix $A = (A_{jk})_{j,k=1}^N$ from the GOE, GUE or GSE, and $\lambda_1, ..., \lambda_N$ its eigenvalues. We to answer the question what the joint probability density function is for the eigenvalues. First consider the case A is from the GOE. By the symmetry of A, A will have $\frac{N(N+1)}{2}$ independent parameters for its entries. Since we can say

$$A = O^T \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{pmatrix} O,$$

where the matrix between O^T and O is diagonal and has zeros in the remaining entries, we want to obtain the $\frac{N(N+1)}{2}$ parameters for A in terms of the entries of the O and diagonal matrix above. There are N eigenvalues so that gives N parameters, and the remaining $\frac{N(N+1)}{2} - N = \frac{N(N-1)}{2}$ can come from O after appropriate transformation.

Let $\Lambda \subseteq \mathbb{R}^N$, $\lambda = (\lambda_1, ..., \lambda_N)$, and $\theta = (\theta_1, ..., \theta_{N(N-1)/2})$ be the vector containing the remaining $\frac{N(N-1)}{2}$ parameters. The joint eigenvalue density function $\rho(\lambda_1, ..., \lambda_N)$ will satisfy

$$P((\lambda_1, ..., \lambda_N) \in \Lambda) = \int_\Lambda \rho(\lambda_1, ..., \lambda_N) d\lambda$$

But note, recalling the density for a matrix A from the GOE,

$$P((\lambda_1, ..., \lambda_N) \in \Lambda) = \int_{\{A:\text{ the eigenvalues of } A \text{ are contained in } \Lambda\}} \exp\left(-\frac{1}{2}\operatorname{Tr}(A^2)\right) dA$$

Now consider the function $(A_{jk})_{j \leqslant k} \to (\lambda, \theta)$, i.e., our parametrization of the matrix entries. This change of variables yields the Jacobian $J(\lambda, \theta) = \det\begin{pmatrix} \frac{\partial A_{jk}}{\partial \lambda_i} & \frac{\partial A_{jk}}{\partial \theta_l} \end{pmatrix}$, where $i = 1, ..., N$ and $l = 1, ..., \frac{N(N-1)}{2}$. Thus, letting R denote the space of all possible values for θ, the above integral becomes

$$= \int_\Lambda \int_R \exp\left(-\frac{1}{2}(\lambda_1^2 + \cdots + \lambda_N^2)\right) |J(\lambda, \theta)|\, d\theta d\lambda = \int_\Lambda \exp\left(-\frac{1}{2}(\lambda_1^2 + \cdots + \lambda_N^2)\right) \int_R |J(\lambda, \theta)|\, d\theta d\lambda$$

Thus by equating our two expressions for $P((\lambda_1, ..., \lambda_N) \in \Lambda)$ we must have

$$\rho(\lambda_1, ..., \lambda_N) = \exp\left(-\frac{1}{2}(\lambda_1^2 + \cdots + \lambda_N^2)\right) \int_R |J(\lambda, \theta)|\, d\theta$$

Example 4.4.1. *Consider a* 2×2 *matrix* $A = O^T \Lambda O$ *from the GOE, where*

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \text{ and } O = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix}$$

for $\theta \in (-\pi, \pi)$. *Then one can show by computation*

$$\begin{pmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \end{pmatrix} :- A = \begin{pmatrix} \lambda_1 \cos^2\theta + \lambda_2 \sin^2\theta & (\lambda_2 - \lambda_1)\cos\theta\sin\theta \\ (\lambda_2 - \lambda_1)\cos\theta\sin\theta & \lambda_1 \sin^2\theta + \lambda_2 \cos^2\theta \end{pmatrix}.$$

Then the Jacobian matrix is given by

$$\begin{pmatrix} \frac{\partial A_{11}}{\partial \lambda_1} & \frac{\partial A_{11}}{\partial \lambda_2} & \frac{\partial A_{11}}{\partial \theta} \\ \frac{\partial A_{22}}{\partial \lambda_1} & \frac{\partial A_{22}}{\partial \lambda_2} & \frac{\partial A_{22}}{\partial \theta} \\ \frac{\partial A_{12}}{\partial \lambda_1} & \frac{\partial A_{12}}{\partial \lambda_2} & \frac{\partial A_{12}}{\partial \theta} \end{pmatrix} = \begin{pmatrix} \cos^2\theta & \sin^2\theta & 2(\lambda_2 - \lambda_1)\cos\theta\sin\theta \\ \sin^2\theta & \cos^2\theta & 2(\lambda_1 - \lambda_2)\cos\theta\sin\theta \\ -\cos\theta\sin\theta & \cos\theta\sin\theta & (\lambda_2 - \lambda_1)(\cos^2\theta - \sin^2\theta) \end{pmatrix}.$$

Thus $|J(\lambda, \theta)| = |\lambda_1 - \lambda_2| \cdot f(\theta)$ *for some function* f *depending only on* θ. *Therefore* $\int |J(\lambda, \theta)|\, d\theta = C\,|\lambda_1 - \lambda_2|$ *for some constant* C. *Hence, the joint eigenvalue density function is given by* $\rho(\lambda_1, \lambda_2) = C \exp\left(-\frac{1}{2}(\lambda_1^2 + \lambda_2^2)\right) |\lambda_1 - \lambda_2|$.

Example 4.4.2. *Consider a* 2×2 *matrix* $A = U^* \Lambda U$ *from the GUE, where*

$$\Lambda = \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \text{ and } U = \begin{pmatrix} \cos\theta & -e^{i\varphi}\sin\theta \\ e^{i\varphi}\sin\theta & \cos\theta \end{pmatrix}$$

for $\theta \in (0, \pi/2)$ *and* $\varphi \in [0, 2\pi]$. *One can follow similar computations as in the previous example and obtain* $|J(\lambda,\theta)|^2 = |\lambda_1 - \lambda_2| \cdot f(\theta)$ *for some function* f *depending only on* θ. *Therefore* $\int |J(\lambda,\theta)| \, d\theta = C\,|\lambda_1 - \lambda_2|^2$ *for some constant* C. *Hence, the joint eigenvalue density function is given by* $\rho(\lambda_1, \lambda_2) = C \exp\left(-\frac{1}{2}(\lambda_1^2 + \lambda_2^2)\right) |\lambda_1 - \lambda_2|^2$.

We now generalize these results in the below theorem.

Theorem 4.4.1. *The joint probability density function for the eigenvalues* $\lambda_1, ..., \lambda_N$ *of a matrix* A *from the GOE, GUE and GSE is given by*

$$\rho(\lambda_1, ..., \lambda_N) = C_{N,\beta} \exp\left(-\frac{\beta}{2}\sum_{j=1}^{N} \lambda_j^2\right) \Pi_{j<k} \left|\lambda_j - \lambda_k\right|^{\beta}$$

where $\beta = 1$ *in the case of the GOE,* $\beta = 2$ *in the case of the GUE and* $\beta = 4$ *in the case of the GSE.*

Note the value of β is the number of components an entry for A has (real, complex or quaternion). Such models occur in a so called "log-gas" system, where the logarithm of the above density is the potential energy for the system. That is, we have an expression of the form $V(\lambda_1, ..., \lambda_N) = \frac{1}{2}\sum_{j=1}^{N} \lambda_j^2 - \beta \sum_{j<k} \log\left|\lambda_j - \lambda_k\right|$. The term $\left|\lambda_j - \lambda_k\right|$ indicates there is a repelling phenomenon occurring between the particles in the system, and β indicates the strength of such a phenomenon. We now give sketch of the proof of the above theorem just for the GOE case.

Proof. Our sketch of the proof will focus on calculating the Jacobian in the GOE case. Recall the setting where we have

$$A = O^T \Lambda O = O^T \begin{pmatrix} \lambda_1 & & \\ & \ddots & \\ & & \lambda_N \end{pmatrix} O$$

where $O^T O = I$, $i = 1, ..., N$ and $l = 1, ..., \frac{N(N-1)}{2}$. Let E_i be the diagonal matrix such that there is a 1 in the i^{th} diagonal entry and zeros elsewhere. Then $\frac{\partial A}{\partial \lambda_i} = O^T E_i O$. We also have

$$\frac{\partial A}{\partial \theta} = \frac{\partial O^T}{\partial \theta}\Lambda O + O^T \Lambda \frac{\partial O}{\partial \theta} = \frac{\partial O^T}{\partial \theta} O O^T \Lambda O + O^T \Lambda \frac{\partial O}{\partial \theta}$$

Using the fact $\frac{\partial O^T}{\partial \theta} O + O^T \frac{\partial O}{\partial \theta}$ is the zero matrix, which implies $\frac{\partial O^T}{\partial \theta} O = -O^T \frac{\partial O}{\partial \theta}$, we have that the above becomes

$$= -O^T \frac{\partial O}{\partial \theta} O^T \Lambda O + O^T \Lambda \frac{\partial O}{\partial \theta}$$

Hence, letting $C_l = \frac{\partial O}{\partial \theta_l} O^T$ we have the equalities

$$O \frac{\partial A}{\partial \lambda_i} O^T = E_i \text{ and } O \frac{\partial A}{\partial \theta_l} O^T = \Lambda C_l - C_l \Lambda$$

for each i and l. Our Jacobian is thus given by (using the fact the determinant of an orthogonal matrix is ± 1)

$$J(\lambda, \theta) = \det\begin{pmatrix} \frac{\partial A_{jk}}{\partial \lambda_i} & \frac{\partial A_{jk}}{\partial \theta_l} \end{pmatrix} = \det\begin{pmatrix} (E_i)_{jk} & (\Lambda C_l - C_l \Lambda)_{jk} \end{pmatrix}$$
$$= \det\begin{pmatrix} \delta_{i,j}\delta_{i,k} & (\lambda_j - \lambda_k)(C_l)_{jk} \end{pmatrix} = \Pi_{j<k} \left|\lambda_j - \lambda_k\right| f(\theta)$$

where f is a function depending only on θ. □

4.5 Further Random Matrix Ensembles

One can define the circular random matrix ensembles, given by the circular orthogonal ensemble, circular unitary ensemble and the circular symplectic ensemble, abbreviated by COE, CUE and CSE respectively. The CUE, denoted $U(N)$, is the set of all $N \times N$ complex unitary matrices. This set is also a compact group (a group under matrix multiplication) that is a subset of $\mathbb{R}^{2N^2}$. One can define a probability distribution measure μ on $U(N)$, called the Haar measure, that is invariant under the group multiplication. That is,

$$\begin{aligned}\mu(\{A \in S\}) &= \mu(\{UA \in S\}) = \mu(\{AU \in S\}) \\ &= \mu(\{U^*AU \in S\})\end{aligned}$$

for all $U \in U(N)$. The above can still be shown despite the fact $U(N)$ is not a commutative group. The eigenvalues of a unitary matrix $A \in U(N)$ lie on the unit circle and can be denoted by $\{e^{i\theta_1}, ..., e^{i\theta_N}\}$. The joint density function for the eigenvalues will be $\rho(\theta_1, ..., \theta_N) = C_N \Pi_{j \leqslant k} \left|e^{i\theta_j} - e^{i\theta_k}\right|^2$, for some normalization constant C_N.

Unlike the CUE, the COE will not be the group of $N \times N$ real orthogonal matrices $O(N)$. We instead take it to be the set of complex unitary matrices that are also symmetric. Each such A in this set can be written, not uniquely, in the form $A = U^TU$, where U is a unitary matrix. By looking at what other matrices we can use to express A with in the non-uniqueness, one can identify the COE with $U(N)/O(N)$, where and U_1 and U_2 are in the same equivalence class if and only if $U_1U_2^{-1} \in O(N)$. If $\pi : U(N) \to U(N)/O(N)$ is corresponding projection map and μ is the Haar measure on $U(N)$, then the measure on the COE is $\mu^* = \mu \circ \pi^{-1}$. Similarly to the Haar measure on $U(N)$, μ^* is the unique measure on the set of complex unitary and symmetric matrices such that it is invariant under orthogonal transformations. If we consider $A = U^TU$, then making the orthogonal transformation $A = O^TAO$ gives us another matrix in the COE, so that we can write it as $O^TAO = V^TV$ for some unitary matrix V. This will imply $V = UO$, so that orthogonal transformations for A correspond to multiplying U by an orthogonal matrix. Again, the eigenvalues of a matrix from the COE lie on the unit circle and can be denoted by $\{e^{i\theta_1}, ..., e^{i\theta_N}\}$. The joint density function for the eigenvalues will be $\rho(\theta_1, ..., \theta_N) = C_N \Pi_{j \leqslant k} \left|e^{i\theta_j} - e^{i\theta_k}\right|$, for some normalization constant C_N.

The CSE is the set of quaternion self dual unitary matrices. Each such $N \times N$ matrix can be identified with a $2N \times 2N$ complex Hermitian matrix, and be written, not uniquely, in the form $A = U^+U$, where U is a unitary matrix. By looking at what other matrices we can use to express A with in the non-uniqueness, one can identify the CSE with $U(2N)/Sp(2N)$, where $Sp(2N)$ is the group of $N \times N$ self-dual quaternion matrices ($2N \times 2N$ complex Hermitian matrices) under matrix multiplication. The probability measure on this set will be invariant under self dual transformations, and the joint density function for the eigenvalues will be $\rho(\theta_1, ..., \theta_N) = C_N \Pi_{j \leqslant k} \left|e^{i\theta_j} - e^{i\theta_k}\right|^4$, for some normalization constant C_N.

Lets look at the matrix groups $O(N)$ and $Sp(2N)$ further. Let $B \in O(N)$. The eigenvalues of B can be 1, -1 or come in pairs $e^{\pm i\theta}$. If N is odd then 1 is always an eigenvalue. We also have $(\det B)^2 = 1$. Thus $\det B = 1$ or $\det B = -1$. Let $O^+(N)$ be those real orthogonal matrices with determinant 1. This group is actually a connected set with respect to the topology on $\mathbb{R}^{N^2}$. The eigenvalues of a matrix from $O^+(2N)$ are of the form $\{e^{\pm i\theta_j}\}_{j=1,...,N}$. Their joint density function is given by $\rho(\theta_1, ..., \theta_N) = C_N \Pi_{j \leqslant k} \left(\cos\theta_j - \cos\theta_k\right)^2$, for some normalization constant C_N. The eigenvalues of a matrix from $O^+(2N+1)$ are of the form $\{e^{\pm i\theta_j}\}_{j=1,...,N} \cup \{1\}$. Their joint density function is given by $\rho(\theta_1, ..., \theta_N) = C_N \Pi_{j=1}^N (1 - \cos\theta_j) \Pi_{j<k} \left(\cos\theta_j - \cos\theta_k\right)^2$, for some normalization constant C_N. The eigenvalues of a matrix from $Sp(2N)$ are of the form $\{e^{\pm i\theta_j}\}_{j=1,...,N}$. Their joint

density function is given by $\rho(\theta_1,...,\theta_N) = C_N \Pi_{j=1}^N (\sin\theta_j)^2 \Pi_{j<k} \left|e^{i\theta_j} - e^{i\theta_k}\right|^2 \cdot \left|e^{i\theta_j} - e^{-i\theta_k}\right|^2$, for some normalization constant C_N. Note when $\theta_j = \pm 1$ we have the first product is 0, which gives us the repelling effect we saw when discussing the joint density functions of the eigenvalues for matrices form the Gaussian ensembles.

We now discuss Laguerre and Jacobi ensembles, which are motivated by the theory of orthogonal polynomials, which we define in the next paragraph. The Laguerre Unitary Ensemble, LUE, is the set of all $N \times N$ positive self adjoint matrices, with a probability measure invariant under unitary transformations. If A is from the LUE and has eigenvalues $x_1,...,x_N$ (which will all be positive), then their joint density function is given by $\rho(x_1,...,x_N) = C_N \Pi_{j=1}^N w(x_j) \Pi_{j<k} \left|x_j - x_k\right|^2$, where $w(x) = x^\alpha e^{-x/2}$ for $x > 0$ and $\alpha > -1$ is a weight function, and C_N is some normalization constant. Note $w(x)$ gives the density for a standard normal Gaussian random variable when $\alpha = 0$ as a special case. The Jacobi Unitary Ensemble, JUE, is the set of all $N \times N$ self adjoint matrices A such that $0 \leqslant A \leqslant I$ or $-I \leqslant A \leqslant I$. If $x_1,...,x_N$ are its eigenvalues, then each $x_j \in [0,1]$ or each $x_j \in [-1,1]$ respectively. Their joint density function is given by $\rho(x_1,...,x_N) = C_N \Pi_{j=1}^N w(x_j) \Pi_{j<k} \left|x_j - x_k\right|^2$, where $w(x) = x^\alpha(1-x)^\beta$ or $w(x) = (1+x)^\alpha(1-x)^\beta$ for $\alpha, \beta > 1$ in each respective case, and C_N is some normalization constant.

Let $I \subseteq \mathbb{R}$ be an interval, and $w(x) \geqslant 0$ a weight function on I. Then a set of orthogonal polynomials is a sequence of polynomials $\{p_n(x)\}_{n\in\mathbb{N}_0}$ such that $deg(p_n) = n$ and $< p_n, p_m >_w = \delta_{n,m}$ (the Kronecker delta function; we could also have $\lambda_n \cdot \delta_{n,m}$ for some $\lambda \neq 0$), where $< p_n, p_m >_w := \int_I p_n(x)\overline{p_m(x)} w(x) dx$. If $w \in L^1(I)$ and $w(x) > 0$ almost everywhere, the orthogonal polynomials exist. We also have the recursion relation $p_{n+1}(x) = (x - \alpha_n)p_n(x) + \beta_n p_{n-1}(x)$ for some constants α_n, β_n. We can define a "kernel"

$$K_n(x,y) = \sqrt{w(x)w(y)} \sum_{j=0}^{n-1} p_j(x)p_j(y)$$

The Christoffel-Darboux formula say this equals a square root term, times a constant depending on n, times $\frac{p_n(x)p_{n-1}(y) - p_n(y)p_{n-1}(x)}{x-1}$. The connection to our unitary ensembles is given by the following theorem.

Theorem 4.5.1. *Consider a unitary ensembles such that the joint density function for the eigenvalues of a matrix from the ensemble is of the form* $\rho(x_1,...,x_N) = C_N \Pi_{j=1}^N w(x_j) \Pi_{j<k} \left|x_j - x_k\right|^2$. *Then* $\rho(x_1,...,x_N) = \frac{1}{N!} \det\left(K_N(x_j,x_k)\right)_{j,k=1}^N$.

We can define the n-term correlation function for an $N \times N$ random matrix , for $1 \leqslant n \leqslant N$, by

$$R_N^{(n)}(x_1,...,x_n) = \underbrace{\int_I \cdots \int_I}_{N-n \text{ times}} \rho(x_1,...,X_N) dx_{n+1} \cdots dx_N.$$

Then we have

$$R_N^{(n)} = \frac{N!}{(N-n)!} \det\left(K_N(x_j,x_k)\right)_{j,k=1}^N.$$