Probability And Stochastic Processes Notes

Liam Flaherty

Professor Bates

NCSU: MA746-001

August 24, 2023

Contents

1 Probability Triples

1.1 Definitions

Definition 1.1. Sample Space Ω: any set containing outcomes (e.g. heads/tails, 1:6, etc.).

Definition 1.2. σ-Algebra: A collection of sets $\mathcal{F}$ from a non-empty set Ω is a sigma-algebra provided $\mathcal{F}$ is closed under countable union and complements. That is, $\mathcal{F}$ is a sigma-algebra if whenever $A_1, A_2, \cdots \in \mathcal{F}$ we have $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$ and $A_1^C \in \mathcal{F}$.

Example 1.1: For $\Omega = \{1, 2, 3\}$, the set $\mathcal{F} = \{\emptyset, \{1\}, \{2, 3\}, \Omega\}$ is a sigma algebra.

Non-example 1.1: For $\Omega = \{1, 2, 3\}$, the set $\mathcal{F} = \{\emptyset, \{1, 2\}, \{2, 3\}, \Omega\}$ isn't a sigma algebra.

Example 1.2: For Ω_∞ denoting all possible sequences of a coin flipped infinitely many times, and where A_H denotes all sequences of flips which begin with a "head" (analogously, A_{TH} denotes all sequences of flips which begin with first a "tail" and then a "head", etc.), the set $\mathcal{F} = \{\emptyset, A_H, A_T, A_{HH}, A_{HT}, A_{HT} \cup A_T, A_{HH} \cup A_T, \Omega\}$ is a sigma algebra (the generating sets are A_{HH}, A_{HT}, and A_T).

Non-example 1.2: Where $\Omega = \mathbb{Z}$, $\mathcal{F} = \{A \in \mathbb{Z} : |A| < \infty \text{ or } |A^c| < \infty\}$ is an algebra (Definition 1.9, Page 5) but not a sigma-algebra. To see this, consider a sequence of sets $A_i = \{i, i+1\}$ for $i \in \mathbb{N}$. Since each A_i is finite, each A_i is in $\mathcal{F}$. If $\mathcal{F}$ was to be a sigma-algebra, then the countable union of these sets, namely $\mathbb{N}$, must also be in $\mathcal{F}$. But both $\mathbb{N}$ and it's complement $-\mathbb{N} \cup \{0\}$ are infinite, so aren't in $\mathcal{F}$.

Definition 1.3. Event Space $\mathcal{F}$: a σ-algebra consisting of unions, intersections, and complements from elements in the sample space.

Example 1.3: If our sample space is $\{1, 2, 3\}$, an event space could be the following: $\{\emptyset, \{1, 2, 3\}, \{1\}, \{2\}, \{3\}, \{1, 2\}, \{1, 3\}, \{2, 3\}\}$. We denote such sigma-algebras 2^Ω to indicate it is the **power set**. Since the power set is the set of all subsets, it is the largest possible sigma algebra on any finite Ω (compare this example to Example 1.1, for instance).

Definition 1.4. Measurable Space (X, Σ): A set X (for example a sample space) along with a sigma-algebra Σ on the set.

Definition 1.5. Measure μ: In the context of a measure space (X, Σ), a measure $\mu : \Sigma \to \mathbb{R}$ is a function from the sigma-algebra to the real line such that $\mu(\emptyset) = 0$ and μ is countably additive, i.e. for all disjoint $A_1, A_2, \cdots \in \Sigma$, $\mu\left(\biguplus_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mu(A_i) \geq 0$.

Definition 1.6. Measure Space (X, Σ, μ): A measurable space along with a measure acting on the space.

Definition 1.7. Probability Measure $\mathbb{P}$: A probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a function on a sigma-algebra $\mathcal{F}$ of the sample space Ω such that $\mathbb{P}(\Omega) = 1$ and $\mathbb{P}$ is countably additive, i.e. for disjoint A_i's, $\mathbb{P}\left(\biguplus_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$. This is a specific case of a general measure. Note that $\mathbb{P}(\emptyset) = 0$ as a consequence of the other two conditions.

Example 1.4: Probability measures are not unique. Suppose our measure space is $\Omega = \{1, 2, 3\}$ and $\mathcal{F} = 2^\Omega$. One probability measure is $\mathbb{P}(A) = \frac{|A|}{|\Omega|}$, the **uniform measure**. A different measure on the same space is $\widetilde{\mathbb{P}}$ where $\widetilde{\mathbb{P}}$ assigns probability $\frac{1}{2}$ to the event $\{1\}$ and probability $\frac{1}{4}$ to the events $\{2\}$ and $\{3\}$ (specifying the probability of all the singletons in a finite power set suffices to completely determine the measure).

Example 1.5: Suppose our sample space is $[0, 1]$ and our sigma-algebra is $\mathbb{B}([0,1])$ (a **borel set**, which is formed by starting with all the closed intervals in $[0, 1]$, and adding in all other sets necessary for a sigma-algebra). One probability measure is the **Lebesgue Measure**, $\mathbb{P}([a,b]) = b - a$. Another probability measure is $\widetilde{\mathbb{P}}([a,b]) = b^2 - a^2$.

Definition 1.8. Probability Space $(\Omega, \mathcal{F}, \mathbb{P})$: A triple consisting of a sample space Ω, an event space $\mathcal{F}$, and a probability measure $\mathbb{P}$ acting on the measurable space $(\Omega, \mathcal{F})$.

Definition 1.9. Algebra: A collection of sets $\mathcal{A}$ from a non-empty set Ω is an algebra provided $\mathcal{A}$ is closed under finite unions and complements. That is, $\mathcal{A}$ is an algebra if whenever $A_1, A_2, \ldots, A_n \in \mathcal{A}$ we have $\bigcup_{i=1}^{n} A_i \in \mathcal{A}$ and $A_1^C \in \mathcal{A}$.

Example 1.6: Any sigma-algebra (Definition 1.2, Page 4) is automatically an algebra since if sets are closed under countable union, they are of course also closed under finite union.

Non-example 1.3: Where $\Omega = \{1, 2, 3, 4\}$, the set $\mathcal{A} = \{\{1, 2\}, \{2, 3\}\}$ is not an algebra since $\{1, 2\} \cup \{2, 3\} = \{1, 2, 3\} \notin \mathcal{A}$.

Example 1.7: Where $\Omega = \mathbb{Z}$, $\mathcal{F} = \{A \in \mathbb{Z} : A \text{ or } A^c \text{ is countable}\}$ is an algebra.

Non-example 1.4: Where $\Omega = \mathbb{Z}$, $\mathcal{F} = \{A \in \mathbb{Z} : A \text{ is countable}\}$ is not an algebra.

Definition 1.10. σ-Algebra (Generated By An Event A, $\sigma(A)$): It is trivial to see that the intersection of sigma algebras is itself a sigma-algebra. So we can define $\sigma(A)$ to be the intersection of all sigma-algebras containing A (in this sense, it is the smallest such set). Constructively, this means we start with the sets in A, and allow for countably many unions, intersections, and complements until we run out of ability to add more.

Example 1.8: Consider the sample space Ω_∞ from Example 1.2 and the event A_H along with the event A_{HT} (i.e. all sequences of coin flips that start with a head and then a tail). If $A = \{A_H, A_{HT}\}$, what is $\sigma(A)$? We start with $\{\emptyset, \Omega_\infty, A_H, A_{HT}\} \subseteq \sigma(A)$, the two generating elements along with the empty and full set which are included by default.

The complement of A_H is simply A_T. The complement of A_{HT} is everything that doesn't start with a head and then a tail, so everything that either starts with a tail, or starts with back-to-back heads; $A_{HT}^C = A_{HH} \cup A_T$. So after adding the initial complements, we have $\{\emptyset, \Omega_\infty, A_H, A_{HT}, A_T, A_{HH} \cup A_T\} \subseteq \sigma(A)$.

The union of $\emptyset, \Omega_\infty, A_H,$ and A_T with all elements to their right are already in the set. The union of A_{HT} with A_T is a new element $A_{HT} \cup A_T$. The complement of this new element is A_{HH}, whose pairwise union with each of the other seven elements are already in the set. So, $\sigma(A) = \{\emptyset, \Omega_\infty, A_H, A_{HT}, A_{HH} \cup A_T, A_T, A_{HT} \cup A_T, A_{HH}\}$.

Example 1.9: The **Borel sigma-algebra** $\mathbb{B}(\mathbb{R})$ is generated by intervals on the real line (it starts with closed intervals, and adds in everything needed to be a sigma-algebra).

An interesting element in $\mathbb{B}([0,1])$ is the **Cantor Set**. Label $C_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$, label $C_2 = [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9}, 1]$, and so on (each new C_k removes the middle third of all parts of C_{k-1}). The Cantor Set is $C = \bigcap_{k=1}^{\infty} C_k$.

Since there are 2^k disjoint segments in each C_k, each segment of length $\frac{1}{3^k}$, under the Lebesgue measure $\mathbb{P}$, $\mathbb{P}(C_k) = \left(\frac{2}{3}\right)^k$. Then as $C_1 \supseteq C_2 \supseteq \cdots$, by continuity from above (Theorem 1.1, Page 7), $\mathbb{P}(C) = \mathbb{P}\left(\bigcap_{k=1}^{\infty} C_k\right) = \lim_{k\to\infty} \mathbb{P}(C_k) = \lim_{k\to\infty} \left(\frac{2}{3}\right)^k = 0$.

What is interesting about this set is that it has zero probability despite having uncountably many points. To see this, imagine there was an enumeration of points $c_1, c_2, \cdots \in C$. Let K_1 be the portion of C_1 that doesn't contain c_1 (so, K_1 is either $[0, \frac{1}{3}]$ or $[\frac{2}{3}, 1]$), K_2 be the portion of $K_1 \cap C_2$ that doesn't contain c_2 (if $c_2 \notin K_1$, pick either section), and so on. Then $K_1 \supseteq K_2 \supseteq \cdots$ and $c_1 \notin K_1, c_2 \notin K_2, \ldots$. Due to the nesting of the non-empty K_n, there must be some element $y \in \bigcap_{n=1}^{\infty} K_n \subset C$. But due to the construction of the K_i, y is not in the list $c_1, c_2, \ldots$; there cannot be an enumeration of the points of C.

Definition 1.11. Resolved Sets: Suppose we are given a measure space $(\Omega, \mathcal{F})$ and an outcome $\omega \in \omega$. The sets in the event space $\mathcal{F}$ which are resolved by some level of information are those sets $A \in \mathcal{F}$ that either definitely contain or definitely don't contain ω. For this reason, it may be helpful to informally think of sigma-algebras as "information".

Example 1.10: Let $\Omega = \Omega_3$ denote all the possible outcomes of three coin flips. Suppose that someone performs the coin flips, and you are interested in their outcome $\omega \in \Omega_2$. If the person tells you the value of the first flip, you are not able to fully know ω, but you can narrow down the possibilities.

Of the sets in 2^{Ω}, $\emptyset$ and Ω are always resolved (Ω is definitely in Ω, and $\emptyset$ is definitely not in $\emptyset$). With the additional information given, the sets $A_H = \{\omega_{HH}, \omega_{HT}\}$ and $A_T = \{\omega_{TH}, \omega_{TT}\}$ are also resolved (e.g., if they tell you the first flip is a head, then ω is definitely in A_H and definitely not in A_T). All together, the sets that are resolved by the information form a sigma-algebra $\mathcal{F}_1 = \{\emptyset, \Omega, A_H, A_T\}$.

Definition 1.12. Filtration: Where Ω is a sample space, where T is some fixed positive number, and where $\mathcal{F}_t$ is a sigma-algebra for all $t \in [0, T]$, then if $\mathcal{F}_s \subseteq \mathcal{F}_t$ whenever $s \leq t$, we say the collection $\mathcal{F}_t$ for $t \in [0, T]$ is a filtration.

Example 1.11: Suppose the person in Example 1.10 now reveals the first two flips of ω. Then the sets A_{HH}, A_{HT}, A_{TH}, and A_{TT} are also resolved, and we get the sigma-algebra $\mathcal{F}_2$ of all these unions and complements. Then $\{\emptyset, \Omega\} = \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2$ is a filtration– as we get further along, we get more and more information about ω.

1.2 Theorems And Examples

Theorem 1.1. Properties of Probability Measures: All probability triples $(\Omega, \mathcal{F}, \mathbb{P})$ satisfy the following:

1. Monotonicity, if $A, B \in \mathcal{F}$ with $A \subseteq B$, then $\mathbb{P}(A) \leq \mathbb{P}(B)$

2. Subadditivity, if $A_1, A_2, \cdots \subset \mathcal{F}$, then $\mathbb{P}\left(\bigcup_i A_i\right) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$

3. Continuity from below and above:

   (a) From below, $A_1 \subseteq A_2 \subseteq \cdots \implies \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \lim_{n \to \infty} \mathbb{P}(A_n)$

   (b) From above, $B_1 \supseteq B_2 \supseteq \cdots \implies \mathbb{P}\left(\bigcap_{i=1}^{\infty} B_i\right) = \lim_{n \to \infty} \mathbb{P}(B_n)$

Proof. For 1, $\mathbb{P}(B) = \mathbb{P}(B \cap A^c) + \mathbb{P}(A)$ by countable additivity, and $\mathbb{P}(B \cap A^c) \geq 0$.

For 2, we first make the union disjoint, then apply countable additivity, then apply monotonicity. Call $A_1' = A_1$, $A_2' = A_2 \cap (A_1')^C$, etc. Then $\mathbb{P}(\bigcup_i A_i) = \mathbb{P}(\biguplus_i A_i') = \sum_{i=1}^{\infty} \mathbb{P}(A_i')$ by countable additivity. By construction, $A_i' \subseteq A_i$ for every i, so by 1, we conclude $\sum_{i=1}^{\infty} \mathbb{P}(A_i') \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$.

For 3a, we use the same construction as for 2 and the fact that $\bigcup_{i=1}^{n} A_i' = A_n'$ for all n. Then $\mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = \mathbb{P}\left(\biguplus_{i=1}^{\infty} A_i'\right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i') = \lim_{n \to \infty} \sum_{i=1}^{n} \mathbb{P}(A_i') = \lim_{n \to \infty} \mathbb{P}\left(\biguplus_{i=1}^{n} A_i'\right) = \lim_{n \to \infty} \mathbb{P}(A_n)$. An analogous process is used for 3b. ∎

Theorem 1.2. Caratheodory's Extension Theorem: Let $\mathcal{A}$ be an algebra, and assume $\mathbb{P} : \mathcal{A} \to [0, 1]$ satisfies the requirements for a probability measure. Then there exists a unique $\widetilde{\mathbb{P}} : \sigma(\mathcal{A}) \to [0, 1]$ such that $\widetilde{\mathbb{P}}(A) = \mathbb{P}(A)$ for all $A \in \mathcal{A}$.

Theorem 1.3. Uniqueness of CDF: Where $\Omega = \mathbb{R}$, and $\mathcal{F} = \mathbb{B}(\mathbb{R})$, define a new function, $F : \mathbb{R} \to [0, 1]$ given by $F(x) = \mathbb{P}((-\infty, x])$ that fulfills the following:

1. Monotone Increasing: if $a \leq b$ then $F(a) \leq F(b)$

2. Right Continuous: if $x_n \searrow x$ (i.e. $x_1 > x_2 > \cdots$ and $\lim_{n \to \infty} x_n = x$), then $F(x_n) \searrow F(x)$

   (essentially, continuity from below, Theorem 1.1, with $\bigcap_{n=1}^{\infty} (-\infty, x_n] = (-\infty, x]$).

3. Limits at $\pm\infty$: if $x_n \searrow -\infty$, then $F(x_n) = 0$ and if $x_n \nearrow \infty$, then $F(x_n) = 1$

Then there exists a unique probability measure $\mathbb{P}$ such that $\mathbb{P}([a, b]) = \mathbb{P}\left((-\infty, b)\right) - \mathbb{P}\left((-\infty, a)\right) = F(b) - F(a)$ for all $a, b \in \mathbb{R}$. We call F the **Cumulative Distribution Function of $\mathbb{P}$**.

1.3   Problems

Problem 1.1) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Show that for any $A, B \in \mathcal{F}$, we have $\mathbb{P}(A \cup B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B)$.

Identity the union of A and B as everything in A that's not in B (notationally, $(A \cap B^c)$), along with everything in B that's not in A (notationally, $(B \cap A^c)$), along with everything shared between A and B (notationally, $A \cap B$). This is a disjoint union, so by the probability axioms:

$$\mathbb{P}(A \cup B) = \mathbb{P}(A \cap B^c) + \mathbb{P}(B \cap A^c) + \mathbb{P}(A \cap B) \tag{1.1}$$

Now see A can be written as everything in A that's not in B along with the shared elements of A and B (notationally, $A = (A \cap B^c) \cup (A \cap B)$). For the same reasoning, $B = (B \cap A^c) \cup (A \cap B)$. Both of these are disjoint unions, so again by the probability axioms:

$$\mathbb{P}(A) = \mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B) \implies \mathbb{P}(A \cap B^c) = \mathbb{P}(A) - \mathbb{P}(A \cap B) \tag{1.2}$$

$$\mathbb{P}(B) = \mathbb{P}(B \cap A^c) + \mathbb{P}(A \cap B) \implies \mathbb{P}(B \cap A^c) = \mathbb{P}(B) - \mathbb{P}(A \cap B) \tag{1.3}$$

Plugging in Equation 1.2 and 1.3 to Equation 1.1, we reach our conclusion:

$$\mathbb{P}(A \cup B) = [\mathbb{P}(A) - \mathbb{P}(A \cap B)] + [\mathbb{P}(B) - \mathbb{P}(A \cap B)] + \mathbb{P}(A \cap B) \tag{1.4}$$

$$= \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B) \tag{1.5}$$

Problem 1.2) Let $\mathcal{F}$ be the collection of subsets $A \subset \mathbb{R}$ that are either countable or cocountable (meaning A^c is countable). Define $\mathbb{P} : \mathcal{F} \to [0,1]$ by $\mathbb{P}(A) = \begin{cases} 0, & A \text{ is countable} \\ 1, & A \text{ is cocountable} \end{cases}$. Note that a subset of $\mathbb{R}$ cannot be both countable and cocountable, so this map is well-defined.

a. Show that $\mathcal{F}$ is a σ-algebra.

We just check the axioms, starting with closure of complement. Let A be a generic element of $\mathcal{F}$. If A is uncountable, then it is in $\mathcal{F}$ because A^c is countable, and we are immediately done as A^c is in $\mathcal{F}$ since A^c is countable. So instead suppose A is countable. Then A^c is in $\mathcal{F}$ since $(A^c)^c = A$ is countable.

Now let $A_1, A_2, \cdots \in \mathcal{F}$ be a generic (possible countably infinite) collection of sets in $\mathcal{A}$. If all these sets are countable, then so is their union. To see this, label the elements of $A_1 = \{a_{1_1}, a_{1_2}, \dots\}$, the elements of $A_2 = \{a_{2_1}, a_{2_2}, \dots\}$, etc. (this is what it means to be countable). Then the union of the sets is seen to be countable through a diagonal argument (first a_{1_1}, then a_{1_2}, then a_{2_1}, then a_{1_3}, etc.); the union of countably many countable sets is countable. So instead suppose at least one of the sets in the union is uncountable, call it A_k. Then the complement of the union is $\left(\bigcup\limits_{i=1} A_i \right)^c = \bigcap\limits_{i=1} A_i^c$ from DeMorgan. Since A_k is a member of the intersection and is countable, so is the intersection (and therefore the complement of the union). So $\mathcal{F}$ is closed under countable union and we see it is a σ-algebra.

b. Show that $\mathbb{P}$ is a probability measure.

Recall $\mathbb{P}$ is a probability measure if $\mathbb{P}(\Omega) = 1$ and if it is countably additive, that is if $\mathbb{P}(\biguplus\limits_{i=1} A_i) = \sum\limits_{i=1} \mathbb{P}(A_i)$ for a countable collection of $A_i \in \Omega$. Since $\mathbb{R}$ is uncountable (recall Cantor's diagonalization argument, that if $f(n)$ was a listing of numbers from $n \in \mathbb{N}$, with each $f(n)$ having a decimal expansion $0.a_{n_1} a_{n_2} a_{n_3} \dots$, such a collection would necessarily exclude real numbers in $(0,1)$ since the decimal expansion $b = 0.b_1 b_2 \dots$ where $b_n = 2$ if $a_{n_n} = 1$ and $b_n = 1$ if $a_{n_n} \neq 1$ is not contained in the $f(n)$), and has a countable complement (namely, the null set), $\mathbb{P}(\mathbb{R}) = 1$.

Now let $\{A_i\}_{i \in \mathbb{N}}$ be a countable collection of disjoint subsets of $\mathcal{F}$. If the A_i are all countable, then so to is their union (see part a above), and thus $\mathbb{P}(\biguplus\limits_{i=1} A_i) = 0 = \sum\limits_{i=1} 0 = \sum\limits_{i=1}^{\infty} \mathbb{P}(A_i)$. If one A_i is cocountable, call it A_k, then by construction of $\mathcal{F}$, the complement of A_k is countable. Since the union is disjoint, this means every other set in the union is countable (because they must reside in the complement of A_k). So $\mathbb{P}(\biguplus\limits_{i=1} A_i) = 1 = 1 + 0 = \mathbb{P}(A_j) + \sum\limits_{i=1, i \neq j} \mathbb{P}(A_i) = \sum\limits_{i=1} \mathbb{P}(A_i)$.

Problem 1.3) Let $\Omega = \{1, 2, 3, 4\}$ and $\mathcal{F} = 2^\Omega$ (this is notation for the power set of Ω, which is the σ-algebra consisting of all subsets of Ω.)

a. Give an example of a collection $\mathcal{A} \subset \mathcal{F}$ and a map $\mathbb{P} : \mathcal{A} \to [0, 1]$ such that:

  1. $\sigma(\mathcal{A}) = \mathcal{F}$

  2. $\Omega \in \mathcal{A}$ and $\mathbb{P}(\Omega) = 1$

  3. **Whenever $A_1, \ldots, A_n \in \mathcal{A}$ are disjoint and $\cup_{i=1}^n A_i$ belongs to $\mathcal{A}$, we have**
     $\mathbb{P}(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mathbb{P}(A_i)$

and yet there is no probability measure $\widetilde{\mathbb{P}} : \mathcal{F} \to [0, 1]$ such that $\widetilde{\mathbb{P}}(A) = \mathbb{P}(A)$ for all $A \in \mathcal{A}$. This shows that the existence part of Caratheodory's Extension Theorem fails without the assumption that $\mathcal{A}$ is an algebra.

Consider the collection $\mathcal{A} = \{\{1, 2, 3, 4\}, \{2, 3, 4\}, \{1, 3, 4\}, \{1, 2, 4\}, \{1, 2, 3\}\}$ and the probability measure $\mathbb{P}(A) = 1$ for all $A \in \mathcal{A}$.

The smallest sigma algebra containing $\mathcal{A}$ is $\mathcal{F}$ since the complement of the triples gives the singletons, after which all the possible unions and intersections can be generated. The probability measure as stated is valid, since the full sample space has probability 1 and since each of the five elements of $\mathcal{A}$ are not disjoint to begin with.

Now imagine there was such a probability measure $\widetilde{\mathbb{P}}$ on $\sigma(\mathcal{A})$. Then we'd need to have $\widetilde{\mathbb{P}}(\{1\}) = 0$ since $\widetilde{\mathbb{P}}(\{1, 2, 3, 4\}) = \widetilde{\mathbb{P}}(\{1\} \uplus \{2, 3, 4\}) = \widetilde{\mathbb{P}}(\{1\}) + \widetilde{\mathbb{P}}(\{2, 3, 4\})$ but $\widetilde{\mathbb{P}}(\{1, 2, 3, 4\}) = \mathbb{P}(\{1, 2, 3, 4\}) = 1$ and $\widetilde{\mathbb{P}}(\{2, 3, 4\}) = \mathbb{P}(\{2, 3, 4\}) = 1$ as well.

The same reasoning shows we would need $\widetilde{\mathbb{P}}(2) = 0$ and $\widetilde{\mathbb{P}}(3) = 0$. But such a scenario is impossible because $\widetilde{\mathbb{P}}(\{1, 2, 3\}) = \mathbb{P}(\{1, 2, 3\}) = 1 \neq 0$.

b. Give an example of a collection $\mathcal{A} \subset \mathcal{F}$ and two maps $\mathbb{P}_1, \mathbb{P}_2 : \mathcal{F} \to [0, 1]$ such that:

1. $\sigma(\mathcal{A}) = \mathcal{F}$

2. $\mathbb{P}_1$ and $\mathbb{P}_2$ are valid probability measures

3. $\mathbb{P}_1(A) = \mathbb{P}_2(A)$ for all $A \in \mathcal{A}$

and yet $\mathbb{P}_1 \neq \mathbb{P}_2$. This shows that the uniqueness part of Caratheodory's Extension Theorem fails without the assumption that $\mathcal{A}$ is an algebra.

Take the set $\mathcal{A} = \{\{1, 2\}, \{2, 3\}\}$ and consider the probability measures $\mathbb{P}_1, \mathbb{P}_2 : \mathcal{F} \to [0, 1]$ given by $\mathbb{P}_1(A) = \frac{1}{2}(\mathbb{1}_{1 \in A} + \mathbb{1}_{3 \in A})$ and $\mathbb{P}_2(A) = \frac{1}{2}(\mathbb{1}_{2 \in A} + \mathbb{1}_{4 \in A})$.

First we show the sigma algebra generated by $\mathcal{A}$ is $\mathcal{F} = 2^{\Omega}$. We generate the singletons as follows: the intersection of $\{1, 2\} \in \mathcal{A}$ and $\{2, 3\} \in \mathcal{A}$ is $\{2\}$. The complement of $\{1, 2\} \in \mathcal{A}$ is $\{3, 4\}$, whose intersection with $\{2, 3\} \in \mathcal{A}$ is $\{3\}$. The complement of $\{2, 3\} \in \mathcal{A}$ is $\{1, 4\}$, whose intersection with $\{1, 2\} \in \mathcal{A}$ is $\{1\}$. The complement of the union of the three singletons generated above gives the fourth singleton, after which we can generate the whole power set.

Next we show the probability measures are valid on $\mathcal{F}$. The full sample space has probability one since 1 and 3 are in Ω for $\mathbb{P}_1$ and since 2 and 4 are in Ω for $\mathbb{P}_2$. The probability of disjoint unions is equivalent to the sum of the probabilities of the sets making up the disjoint unions since $\mathbb{P}_1$ gives a uniform probability on $\{1\}$ and $\{3\}$ and $\mathbb{P}_2$ gives a uniform probability on $\{2\}$ and $\{4\}$.

Finally, we show the probability measures agree on $\mathcal{A}$. There are only two cases to check: $\mathbb{P}_1(\{1, 2\}) = \frac{1}{2}(\mathbb{1}_{1 \in A} + \mathbb{1}_{3 \in A}) = \frac{1}{2}(1 + 0) = \frac{1}{2} = \frac{1}{2}(1 + 0) = \frac{1}{2}(\mathbb{1}_{2 \in A} + \mathbb{1}_{4 \in A}) = \mathbb{P}_2(\{1, 2\})$ and $\mathbb{P}_1(\{2, 3\}) = \frac{1}{2}(\mathbb{1}_{1 \in A} + \mathbb{1}_{3 \in A}) = \frac{1}{2}(0 + 1) = \frac{1}{2} = \frac{1}{2}(1 + 0) = \frac{1}{2}(\mathbb{1}_{2 \in A} + \mathbb{1}_{4 \in A}) = \mathbb{P}_2(\{2, 3\})$. So $\mathbb{P}_1$ and $\mathbb{P}_2$ agree on $\mathcal{A}$, but not on all of $\mathcal{F}$ (see for example that $\left[\mathbb{P}_1(\{1\}) = \frac{1}{2}\right] \neq [0 = \mathbb{P}_2(\{1\})]$).

2 Measurable Functions And Random Variables

2.1 Definitions

Definition 2.1. Measurable Function: A function $X : \Omega \to S$ between measure spaces $(\Omega, \mathcal{F})$ and $(S, \mathcal{S})$ is measurable if whenever $B \in \mathcal{S}$, $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$ (the inverse image of every measurable set is measurable). To emphasize that dependency on the respective sigma-algebras and to be precise, we might say "X is $(\mathcal{F}, \mathcal{S})$ measurable" (or just "X is $\mathcal{F}$-measurable" when $\mathcal{S}$ is understood) and write $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$.

Definition 2.2. Random Variable: A measurable function $X : \Omega \to \mathbb{R}$ between measure spaces $(\Omega, \mathcal{F})$ and $(\mathbb{R}, \mathbb{B}(\mathbb{R}))$ (it is just a specific case of a measurable function where the codomain is fixed). Note that the "randomness" from a random variable comes from the random experiment of choosing the $\omega \in \Omega$. Note further that to emphasize the fact that a random variable is a function, we may often write $X(\omega)$ (though X may be used for brevity).

Example 2.1: Every random variable X is $(\sigma(X), \mathbb{B}(\mathbb{R}))$ measurable (see Definition 2.4).

Example 2.2: Consider a measure space $(\Omega_3, \mathcal{F}_3 = 2^{\Omega_3})$ where Ω_3 denotes the outcomes of three coin flips and 2^{Ω_3} is the power set. Label each of the eight elements $\omega \in \Omega_3$ in accordance to the coin flip outcomes, (e.g. ω_{HHT}), and label each set in the event space likewise (e.g. $A_{HT} = \{\omega_{HTH}, \omega_{HTT}\}$). Then $Y(\omega) = (\# \text{ number of heads in first two flips})$ is a random variable. Informally, this is because the "information" contained in $\mathcal{F}_3$ is sufficient to completely determine the output of the Y; Y is $\mathcal{F}_3$-measurable.

Non-example 2.1: Consider the same sample space and function Y as Example 2.2. If we replace $\mathcal{F}_3$ with $\mathcal{F} = \{\emptyset, \Omega_3, A_{\bullet\bullet H}, A_{\bullet\bullet T}\}$ (where $A_{\bullet\bullet H}$ is the event that the third flip is "heads"), then Y is not a $\mathcal{F}$-measurable function. Informally, this is because the "information" in $\mathcal{F}$ is insufficient to determine the output of Y; knowing which events ω belongs to in $\mathcal{F}$ does not allow you to determine $Y(\omega)$. Concretely, $\omega_1 = \omega_{HHT}$ and $\omega_2 = \omega_{HTT}$ are both in $A_{\bullet\bullet T}$ and Ω_3 (and not in the other two sets) and yet $Y(\omega_2) = 1 \neq 2 = Y(\omega_1)$. Another way to look at it is to take $B = \{2\} \in \mathbb{B}(\mathbb{R})$. Then $Y^{-1}(\{2\}) = \{A_{HHH}, A_{HHT}\} \notin \mathcal{F}$.

Example 2.3: Where $(\Omega, \mathcal{F})$ is a measure space and $A \in \mathcal{F}$, the **indicator function** $\mathbb{1}_A : \Omega \to \mathbb{R}$ where $\mathbb{1}_A(\omega) = \begin{cases} 1, \omega \in A \\ 0, \omega \notin A \end{cases}$ is a random variable.

We need to check that the inverse image of every set in $\mathbb{B}(\mathbb{R})$ is measurable. That is, we need to check that the inverse of all subsets B from $\mathbb{B}(\mathbb{R})$ are subsets of $\mathcal{F}$.

So let B be given. The pre-image of B is dependent on whether B contains either 0 or 1. If B contains 1 but not 0, then $\mathbb{1}_A^{-1}(B)$ is A. If B contains 0 but not 1, then $\mathbb{1}_A^{-1}(B)$ is A^c. If B contains both 0 and 1, then $\mathbb{1}_A^{-1}(B)$ contains both A^c and A, i.e. is Ω. Finally, if B contains neither 0 and 1, then since any element in Ω maps to either 1 or 0, $\mathbb{1}_A^{-1}(B)$ is the empty set. Since we start with $A \in \mathcal{F}$ and sigma algebras are closed under compliment, $A^c \in \mathcal{F}$. Sigma algebras also necessarily contain the full and empty set, so this proves $\mathbb{1}_A$ is measurable.

Definition 2.3. Random Vector: A measurable function $(X_1, X_2, \ldots, X_n) : (\Omega^n, \mathcal{F}^n) \to (\mathbb{R}^n, \mathbb{B}(\mathbb{R}^n))$. This is essentially just n random variables placed next to each other.

Definition 2.4. σ-algebra (Generated By A Random Variable X, $\sigma(X)$): Where X is a random variable, the sigma-algebra generated by X is $\sigma(X) = \{X^{-1}(B) : B \in \mathbb{B}(\mathbb{R})\} = \{\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F} : B \in \mathbb{B}(\mathbb{R})\}$. Informally, it is the minimally small sigma algebra that completely captures the information revealed by the values of the random variable.

Example 2.4: Consider the random variable Y from Example 2.2, where the realization of Y is the number of heads in the first two flips of a coin flipped three times.

There are a few Borel sets to check to help us build the sigma-algebra $\sigma(Y)$. Note that the exact borel sets below are not unique (e.g. B_7 could just as well be $\{3\}$).

- $B_1 = \{2\} \implies Y^{-1}(B_1) = A_{HH}$
- $B_2 = \{1\} \implies Y^{-1}(B_2) = A_{HT} \cup A_{TH}$
- $B_3 = \{0\} \implies Y^{-1}(B_3) = A_{TT}$
- $B_4 = \{[1,2]\} \implies Y^{-1}(B_4) = A_H \cup A_{TH}$

- $B_5 = \{[0,1]\} \implies Y^{-1}(B_5) = A_T \cup A_{HT}$
- $B_6 = \{[0,2]\} \implies Y^{-1}(B_6) = \Omega_3$
- $B_7 = \{[0.25, 0.75]\} \implies Y^{-1}(B_7) = \emptyset$

Do these seven elements actually form a sigma-algebra? Since every element's complement must be in the set, and since there are currently an odd number of elements, we know the answer is "no". Using the systematic approach from Example 1.8, we see that the only element whose complement is missing is $A_{HT} \cup A_{TH}$. So we add in $A_{HH} \cup A_{TT}$, and after checking each of the pairwise unions, see that these eight elements are indeed a sigma-algebra, $\sigma(Y) = \{\emptyset, \Omega_3, A_{HH}, A_{TT}, A_{HT} \cup A_{TH}, A_{TT} \cup A_{HH}, A_H \cup A_{TH}, A_T \cup A_{HT}\}$.

Notice that this sigma-algebra is but a small subset of $\mathcal{F}_3 = 2^{\Omega_3}$ which has $2^{2^3} = 256$ elements. We repeat that Y is $\mathcal{F}_3$-measurable; the "information" in $\mathcal{F}_3$ is more than sufficient to determine the value of Y. But what about all this "extra" information? For instance, A_H is an element of $\mathcal{F}_3$ (and even $\mathcal{F}_1$!), but A_H only appears in $\sigma(Y)$ as a union with other elements. This is because knowing the value of Y is not enough to know if the first flip was "heads". For example, if $Y = 1$, the first flip might have been heads, or the first flip might have been tails– both ω_{HTT} and ω_{THT} map to 1, after all.

Definition 2.5. Distribution (Push-forward, Law) Of A Random Variable, μ_X: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathbb{B}(\mathbb{R}))$ be a random variable. Then the law of X (distributional measure, push-forward) is the function $\mu_X : \mathbb{B}(\mathbb{R}) \to [0,1]$ given by $\mu_X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\})$.

Example 2.5: If the measure space in the domain of a random variable is equipped with a probability measure, then the random variable will have a distribution. But distributions and random variables are different concepts– different random variables can have the same distribution, and a single random variable can have two different distributions (by changing the probability measure).

Consider the probability measures $\mathbb{P}([a,b]) = b - a$ and $\widetilde{\mathbb{P}}([a,b]) = b^2 - a^2$ acting on the borel set $\mathbb{B}([0,1])$. Consider further the random variables $X(\omega) = \omega$ and $Y(\omega) = 1 - \omega$ for all $\omega \in ([0,1] = \Omega)$. Even though $X \neq Y$, $\mu_X = \mu_Y$ under $\mathbb{P}$. See that $\mu_X([a,b]) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in [a,b]\}) = b - a = (1-a)-(1-b) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in [1-b, 1-a]\}) = \mathbb{P}(\{\omega \in \Omega : Y(\omega) \in [a,b]\}) = \mu_Y([a,b])$ where the second to last equality follows from the observation that $a \leq 1 - \omega \leq b \implies -a \geq \omega - 1 \geq -b \implies 1 - a \geq \omega \geq 1 - b$.

On the other hand, $\widetilde{\mu}_X \neq \widetilde{\mu}_Y$. See that $\widetilde{\mu}_X([a,b]) = \widetilde{\mathbb{P}}(\{\omega \in \Omega : X(\omega) \in [a,b]\}) = b^2 - a^2$ but that $\widetilde{\mu}_Y([a,b]) = \widetilde{\mathbb{P}}(\{\omega \in \Omega : Y(\omega) \in [a,b]\}) = \widetilde{\mathbb{P}}(\{\omega \in \Omega : X(\omega) \in [1-a, 1-b]\}) = (1-b)^2 - (1-a)^2 = a^2 - b^2 - 2a + 2b$.

Example 2.6: Occasionally, a random variable X may have a **density function** (PDF), $f_X(x)$. This happens if $\mu_X([a,b]) = \mathbb{P}(a \leq X \leq b) = \int_a^b f_X(x)\,dx$ for all $a, b \in \mathbb{R}$ (where f_X is necessarily non-negative). Random variables with PDFs are called **continuous random variables**.

Example 2.7: Occasionally, random variables may have a **probability mass function** (PMF). This happens when there is a countable sequence of numbers $x_1, x_2, \ldots$ which the random variable takes on with probability one, and we have a p_i such that $\mu_X(B) = \mathbb{P}(X \in B) = \sum_{i, x_i \in B} p_i$. Random variables with PMFs are often called **discrete random variables**.

Non-example 2.2: Random variables need not have either a density or a probability mass function. Consider the random variable $Y = \sum_{n=1}^{\infty} \frac{2Y_n}{3^n}$ where $Y_n \overset{i.i.d}{\sim} \text{Bern}(0.5)$. See that through the first n summands, Y takes on values from the Cantor Set given in Example 1.9, Page 6 (if $Y_1 = 0$, which happens with probability $\frac{1}{2}$, then $Y \in \left[0, \frac{1}{3}\right]$; if $Y_1 = 0$ and $Y_2 = 1$ which happens with probability $\frac{1}{4}$, then $Y \in \left[\frac{2}{9}, \frac{1}{3}\right]$, etc.).

If there was a density function for Y, call it f_Y, then we would need to see $\int_{-\infty}^{\infty} f_Y(y)\,dy = \int_0^1 f_Y(y)\,dy = 1$. However, we have shown in the details to the Cantor Set explanation that C has Lebesgue Measure 0 and thus is almost everywhere zero; $\int_0^1 f_Y(y)\,dy = 0$.

If there was a probability mass function for Y, then we would need to have $\mathbb{P}(Y = x) > 0$ for some $x \in C$. See that x can be expressed as a base-three expansion $x = \sum_{n=1}^{\infty} \frac{1}{3^n} x_n$ where $x_n \in \mathbb{Z}_3$ (if $x \in C$, then $x \in C_n$ for all n, which requires x_1 to be either 0 or 2, and so on). There is some subtlety in that infinite expansions can be represented in two different ways. For example, $C \ni \frac{1}{9} = 0 \cdot \frac{1}{3} + 0 \cdot \frac{1}{9} + 2 \cdot \frac{1}{27} + 2 \cdot \frac{1}{81} + 2 \cdot \frac{1}{243} + \cdots$ (this is an example of a geometric series $\sum_{n=0}^{\infty} ar^n$ with $r = \frac{1}{3}$ and $\frac{1}{9} = \frac{a}{1-r} \implies a = \frac{2}{27}$). From this perspective, it is clear that there is at most two choices of $\omega \in \Omega$ which yield any given $x \in C$. Since C is uncountable, this means $\mathbb{P}(\{\omega \in \Omega : X(\omega) = x\}) = 0$; there can not be a mass function.

Definition 2.6. Cumulative Density Function (CDF) Of A Random Variable: Where μ_X is the law (Definition 2.5, Page 13) of a random variable X, the CDF of X is the function $F_X : \mathbb{R} \to [0, 1]$ given by:

$$F_X(x) = \mu_X\big((-\infty, x]\big) = \mathbb{P}(X^{-1}(-\infty, x]) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}) = \mathbb{P}(X \leq x)$$

Here x (lowercase) denotes a generic element of the domain $\mathbb{R}$, and X (uppercase) denotes the random variable. So, we might have, e.g. $X(\omega) = x$. Compare this definition to Theorem 1.3, which doesn't require a random variable. This is really the exact same idea, it just maps the image of the random variable back to the sample space.

Definition 2.7. Quantile Function: Where F_X is a valid CDF for a random variable X, the quantile function for F_X is the function $F_X^{-1} : [0, 1] \to \mathbb{R}$ given by $F_X^{-1}(u) = \inf\{t \in \mathbb{R} : F_X(t) \geq u\}$. We capture the intuition behind the quantile function at the cost of precision (since F_X may not have an inverse) when we use the notation F_X^{-1}.

2.2 Theorems And Examples

Theorem 2.1. Composition of Measurable Maps Is Measurable. Let $(\Omega, \mathcal{F})$, $(S, \mathcal{S})$, and $(T, \mathcal{T})$ be measure spaces. Further let $X : \Omega \to S$ and $Y : S \to T$ be $(\mathcal{F}, \mathcal{S})$ and $(\mathcal{S}, \mathcal{T})$ measurable respectively. Then $Z = Y \circ X : \Omega \to T$ is $(\mathcal{F}, \mathcal{T})$ measurable.

Proof. Let $B \in \mathcal{T}$ be given. Then $Z^{-1}(B) = X^{-1}(Y^{-1}(B))$. Since Y is measurable, $Y^{-1}(B) \in \mathcal{S}$. And since X is measurable and $Y^{-1}(B) \in \mathcal{S}$, $X^{-1}(Y^{-1}(B)) \in \mathcal{F}$ as desired. ∎

Theorem 2.2. Check generating set. A trick to ensuring measurability is to check a generating set. If $(\Omega, \mathcal{F})$ and $(S, \mathcal{S})$ are measure spaces and $B \subseteq \mathcal{S}$ such that $\sigma(B) = \mathcal{S}$, then if $X : \Omega \to S$ satisfies $X^{-1}(a) \in \mathcal{F}$ for all $a \in \mathcal{S}$; X is measurable.

Proof. Consider the set $\mathcal{S}' = \{B \subseteq S : X^{-1}(B) \in \mathcal{F}\}$. If we can show $\mathcal{S}'$ is a sigma-algbera, then we will have arrived at our conclusion; since $\sigma(B)$ is the smallest sigma-algebra containing B, we would see $\mathcal{S} \subseteq \mathcal{S}'$ and from how $\mathcal{S}'$ is defined, the inverse image of any set in $\mathcal{S}'$ is in $\mathcal{F}$, which is the definition of measurable. To that end, let $B_1, B_2, \ldots$ be given.

Since $B_1 \in \mathcal{S}'$, $X^{-1}(B_1) \in \mathcal{F}$, and then since $\mathcal{F}$ is a sigma-algebra, $X^{-1}(B_1^c) \in \mathcal{F}$. But $\mathcal{S}'$ is the set of all elements in S whose inverse mapping is in $\mathcal{F}$, so B_1^c must be in $\mathcal{S}'$. This proves $\mathcal{S}'$ is closed under compliment.

Since $B_1, B_2, \cdots \in \mathcal{S}'$, each of $X^{-1}(B_i) \in \mathcal{F}$. Since $\mathcal{F}$ is a sigma-algebra, $\bigcup_{i=1}^{\infty} X^{-1}(B_i) \in \mathcal{F}$. But then $X^{-1}(\bigcup_{i=1}^{\infty} B_i) \in \mathcal{F}$ and so $\bigcup_{i=1}^{\infty} B_i \in \mathcal{S}'$. This proves $\mathcal{S}'$ is closed countable union and thus is a sigma-algebra and we've reached our result. ∎

Corollary 2.2.1. Sup And Inf Are Random Variables.

Recall that we say L is the least upper bound (**supremum**) of a set A provided $L \geq a$ for all a in A (L is an upper bound) and provided whenever $M \geq a$ for all $a \in A$, $L \leq M$ (L is the least upper bound). Similarly define the greatest lower bound (**infimum**). Relatedly, we can define the **limit superior** of a sequence $\{s_n\}_{n \in \mathbb{N}}$ as the value L such that $L = \lim_{m \to \infty} \sup \{s_n : n > m\}$ (it is an infimum of supremums) and similarly define the **limit inferior**. As an illustrative example take: $S_n = \left\{1 + \frac{1}{2}, 0 - \frac{1}{2}, 1 + \frac{1}{4}, 0 - \frac{1}{4}, 1 + \frac{1}{8}, 0 - \frac{1}{8}, \ldots\right\}$. Then $\sup \{S_n\} = \frac{3}{2}$, $\inf \{S_n\} = \frac{-1}{2}$, $\limsup \{S_n\} = 1$, and $\liminf \{S_n\} = 0$. A sequence $\{S_n\}$ only has a limit if $\liminf \{S_n\} = \limsup \{S_n\}$.

Now to our statement. Where $X_1, X_2, \ldots, X_n$ are random variables on $(\Omega, \mathcal{F})$, define a random variable $X : \Omega \to \mathbb{R}$ by $X(\omega) = \inf \{X_i(\omega)\}_{i \in [1,n]}$. We claim X is a random variable. By Theorem 2.2, we just need to verify measurability on a generating set $B = (-\infty, a)$. Since the function is the infimum, we see $\inf X_n \leq a \implies \cup \{\omega \in \Omega : X_n(\omega) < a\} \in \mathcal{F}$. Similar reasoning proves the supremum, and then by Theorem 2.1, we also see the limit inf/superior is a random variable.

2.3 Problems

Problem 2.1) Let $(\Omega, \mathcal{F})$ and $(S, \mathcal{S})$ be measurable spaces, and $X: \Omega \to S$ a measurable map. We denote the pre-image of any $B \subseteq S$ by $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\}$. Define the collection of all pre-images of measurable sets $\omega(X) = \{X^{-1}(B) : B \in \mathcal{S}\}$. Since we assume X is measurable, we know $\sigma(X) \subset \mathcal{F}$. Show that $\sigma(X)$ is a σ-algebra.

We go directly for the definition. Assume $A, A_1, A_2, \dots$ are sets in $\sigma(X)$.

Since $A \in \sigma(X)$, there exists a $B \in \mathcal{S}$ such that $X^{-1}(B) = A$. Further, since A is the set of all elements in Ω that map into B under X, A^c is the set of all elements in Ω that map into B^c under X. Then $A^c \in \omega(X)$ if and only if $B^c \in \mathcal{S}$. But $\mathcal{S}$ is a sigma-algebra, so is closed under compliment. As $B \in \mathcal{S}$, this means $B^c \in \mathcal{S}$ and we have shown $\sigma(X)$ is closed under compliment.

Since the A_i are all in $\sigma(X)$, there exists corresponding $B_i \in \mathcal{S}$ such that for every i, $X^{-1}(B_i) = A_i$. Thus if the union of the B_i's are in $\mathcal{S}$, it must be the case that the union of the A_i's are in $\sigma(X)$. As each B_i resides in $\mathcal{S}$ and $\mathcal{S}$ is a sigma-algebra (and so is closed under countable union), the union of the B_i's is also in $\mathcal{S}$ and we have shown $\sigma(X)$ is closed under countable union.

Problem 2.2) Let U be a uniform random variable on the open unit interval. Define $f : (0, 1) \to \mathbb{R}_+$ by $f(u) = -\ln(1 - u)$ Compute the distribution function F_X of $X = f(U)$.

Given $U \sim (0, 1)$, we know the CDF is $F_U(u) = \frac{u-0}{1-0} = u$ for values of u in the open unit interval. We are asked to compute the CDF of the transformation. The general CDF method is shown below:

$$F_X(x) = \mathbb{P}(X \leq x) = \mathbb{P}(f(U) \leq x) = \mathbb{P}(U \leq f^{-1}(x)) = F_U\big(f^{-1}(x)\big)$$

Since $f(u) = -\ln(1 - u)$, we compute the inverse as:

$$u = -\ln\big(1 - f^{-1}(u)\big)$$
$$e^{-u} = 1 - f^{-1}(u)$$
$$f^{-1}(u) = 1 - e^{-u}$$

So substituting from above (with a restriction on $x \geq 0$), we have:

$$F_X(x) = F_U\big(f^{-1}(x)\big) = F_U\big(1 - e^{-x}\big) = 1 - e^{-x}$$

Problem 2.3) Let U be a uniform random variable on the open unit interval. Let F be any distribution function (i.e. is non-decreasing, right continuous, mapping to the closed unit interval, and has left and right limits at $\pm\infty$ of 0 and 1 respectively), Find a function $f : (0,1) \to \mathbb{R}$ such that the random variable $X = f(U)$ has F as its distribution function.

From the reasoning in Problem 2.2, we want f to be the inverse of F (this is the quantile function). However the restrictions on $F : \mathbb{R} \to [0,1]$ don't imply injectivity; F may not have an inverse. We can get around this by defining $f(u) = \inf \{x \in \mathbb{R} : F(x) \geq u\}$ (in non-precise words, the smallest member of the domain of F whose image under F is at least u).

First check f is well-defined. For any u, $f(u) < \infty$ since $F(x)$ goes to 1 as x goes to infinity, and since u is maximally 1. Similarly, $f(u) > -\infty$ since $F(x)$ goes to 0 as x goes to minus infinity, and since u in minimally 0.

Ultimately, we want to show $X = f(U)$ satisfies $F(x) = \mathbb{P}(X \leq x) = \mathbb{P}(f(U) \leq x) = \mathbb{P}(U \leq F(x)) = F(U)$. The third equality is what remains to be shown (i.e. $f(U) \leq x \iff U \leq F(x)$). The second direction follows immediately from f being an infimum of the x's. The first direction follows from the monotonicity and right-continuity of F. This proves the distribution function of f is F as desired.

Problem 2.4) Let $\{X_i\}_{i=1}^{\infty}$ be independent exponential random variables of rate 1, i.e $\mathbb{P}(X_i \geq x) = e^{-x}$ for $x \geq 0$. Let $M_n = \max_{1 \leq i \leq n} X_i$. Show that for any $t \in \mathbb{R}$, $\lim_{n\to\infty} \mathbb{P}(M_n - \ln(n) \leq t) = e^{-e^{-t}}$. The double exponential on the right-hand side is called the Gumbel distribution. Roughly speaking, this results tells us that for large n, $M_n \cong \ln(n) + Z$, where Z is a random variable with the Gumbel distribution.

See that $\mathbb{P}(M_n \leq t) = \mathbb{P}(X_1, X_2, \ldots, X_n \leq t) = \mathbb{P}(X_1 \leq t)\mathbb{P}(X_2 \leq t) \cdots \mathbb{P}(X_n \leq t)$ by independence, and further that $\mathbb{P}(X_1 \leq t)\mathbb{P}(X_2 \leq t) \cdots \mathbb{P}(X_n \leq t) = [\mathbb{P}(X_1 \leq t)]^n$ by the identical distribution. So the distribution of M_n is $\mathbb{P}(M_n \leq t) = [1 - e^{-t}]^n$ (each of the X_i's is $\sim \text{Exp}(1)$).

Then $\lim_{n\to\infty} \mathbb{P}(M_n - \ln(n) \leq t) = \lim_{n\to\infty} \mathbb{P}(M_n \leq t + \ln(n)) = \lim_{n\to\infty} \left[1 - e^{-(t+\ln(n))}\right]^n$. Expanding, we have $\lim_{n\to\infty} \left[1 - e^{-(t+\ln(n))}\right]^n = \lim_{n\to\infty} \left[1 - e^{-t}e^{-\ln(n)}\right]^n = \lim_{n\to\infty} \left[1 + \frac{-e^{-t}}{n}\right]^n$. Using the limit definition of e, this is precisely $e^{-e^{-t}}$ as desired.

Problem 2.5) Check that the "bell curve" is in fact a density function. That is, show that $\int_{-\infty}^{\infty} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}\, dt = 1$.

First note for all positive a, $a = 1$ if and only if $a^2 = 1$. To the problem at hand, we have:

$$\left(\int_{-\infty}^{\infty} \frac{e^{-\frac{t^2}{2}}}{\sqrt{2\pi}}\, dt\right)^2 = \int_{-\infty}^{\infty} \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}}\, dx \cdot \int_{-\infty}^{\infty} \frac{e^{-\frac{y^2}{2}}}{\sqrt{2\pi}}\, dy \qquad \text{Variable change}$$

$$= \int_{\mathbb{R}}\int_{\mathbb{R}} \frac{e^{-(x^2+y^2)/2}}{2\pi}\, dx\, dy \qquad \text{Integration is bilinear}$$

$$= \int_0^{2\pi}\int_0^{\infty} \frac{e^{-r^2/2}}{2\pi} r\, dr\, d\theta \qquad \text{Polar coordinates}$$

$$= \int_0^{2\pi} \frac{1}{2\pi}\left(\int_0^{\infty} r\cdot e^{-r^2/2}\, dr\right) d\theta \qquad \text{Pull out constants}$$

$$= \int_0^{\infty} r\cdot e^{-r^2/2}\, dr \qquad \text{Integrate over angle}$$

$$= \int_0^{\infty} e^{-u}\, du = -e^{-u}\,\big|_0^{\infty} = 1 \qquad \text{Integrate over radius}$$

The step transforming to polar coordinates follows from the substitution $x = r\cos\theta$ and $y = r\sin\theta$. For the integrand, we then have $x^2 + y^2 = r^2(\cos^2\theta + \sin^2\theta) = r^2$. For the variables of integration, we have $J = \begin{bmatrix} \frac{dx}{dr} & \frac{dx}{d\theta} \\ \frac{dy}{dr} & \frac{dy}{d\theta} \end{bmatrix} = \begin{bmatrix} \cos\theta & -r\sin\theta \\ \sin\theta & r\cos\theta \end{bmatrix}$ and so $dx\, dy = |\det(J)| dr\, d\theta = |r\cos^2\theta + r\sin^2\theta|\, dr\, d\theta = r\, dr\, d\theta$. For the limits of integration, we transform the xy-plane to polar coordinates; the radius must stretch $(0, \infty)$, and the angle must rotate $(0, 2\pi)$.

The step integrating over the radius comes from the substitution $u = \frac{r^2}{2}$. Then $\frac{du}{dr} = r$ and so $du = r\, dr$ and the integrand is changed from $r\cdot e^{-r^2/2}\, dr$ to $e^{-u}\, du$.

Problem 2.6) Let $\{G_n\}_{n=1}^{\infty}$ be such that $\mathbb{P}(G_n) \to 1$ as $n \to \infty$. Show that for any other sequence $\{A_n\}_{n=1}^{\infty}$, we have $\liminf\limits_{n\to\infty} \mathbb{P}(A_n) = \liminf\limits_{n\to\infty} \mathbb{P}(A_n \cap G_n)$ and $\limsup\limits_{n\to\infty} \mathbb{P}(A_n) = \limsup\limits_{n\to\infty} \mathbb{P}(A_n \cap G_n)$. This justifies the practice of "restricting to a good event", provided the good event occurs with probability tending to 1.

By properties of probability, we can write:

$$\mathbb{P}(A_n) = \mathbb{P}(A_n \cap G_n) + \mathbb{P}(A_n \cap G_n^c) \le \mathbb{P}(A_n \cap G_n) + \mathbb{P}(G_n^c)$$
$$\implies \mathbb{P}(A_n) - \mathbb{P}(G_n^c) \le \mathbb{P}(A_n \cap G_n)$$

Then since $\mathbb{P}(A_n \cap G_n) \le \mathbb{P}(A_n)$ for any n, we can squeeze $\mathbb{P}(A_n \cap G_n)$:

$$\mathbb{P}(A_n) - \mathbb{P}(G_n^c) \le \mathbb{P}(A_n \cap G_n) \le \mathbb{P}(A_n)$$

As $G_n \to 1$, we must have $G_n^c \to 0$. Then as $n \to \infty$, we see:

$$\mathbb{P}(A_n) \le \mathbb{P}(A_n \cap G_n) \le \mathbb{P}(A_n)$$

This proves our result.

3 Expectations Of Random Variables

3.1 Definitions

Definition 3.1. Lebesgue Integral: Recall the definition of the Riemann Integral for a differentiable function f. Partition the domain $a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$, let $M_k = \max\limits_{x_{k-1} \leq x \leq x_k} f(x)$, $m_k = \min\limits_{x_{k-1} \leq x \leq x_k} f(x)$, $\Pi = \{x_0, \ldots, x_n\}$, and $\|\Pi\| = \max\limits_{1 \leq k \leq n} (x_k - x_{k-1})$, then see the Upper Riemann Sum $(RS_{\Pi^+}(f) = \sum\limits_{k=1}^{n} M_k \cdot (x_k - x_{k-1})$ and Lower Riemann Sum $(RS_{\Pi^-}(f) = \sum\limits_{k=1}^{n} m_k \cdot (x_k - x_{k-1})$ converge to the same value as $\|\Pi\|$ goes to zero, namely $\int_a^b f(x) \, dx$. Integrating in this way necessitates a natural ordering of the domain, which is a property that Ω, unlike $\mathbb{R}$, may not have. For that reason, instead of partitioning the domain, we partition the range in the Lebesgue Integral.

So assume for now that $0 \leq X(\omega) < \infty$. Partition the range of the random variable X as $0 = y_0 < y_1 < \ldots$ and as before denote $\Pi = \{y_0, \ldots, y_n\}$ and $\|\Pi\| = \max\limits_{1 \leq k \leq n} (y_k - y_{k-1})$. Consider the event $A_k = \{\omega \in \Omega : y_k \leq X(\omega) \leq y_{k+1}\}$. Then the Lebesgue Integral is the limit of the Lower Lebesgue Sum as $\|\Pi\|$ goes to zero; $\lim\limits_{\|\Pi\| \to 0} \sum\limits_{k=1}^{\infty} y_k \mathbb{P}(A_k) = \int_\Omega X(\omega) \, d\mathbb{P}(\omega)$.

Define $X^+(\omega) = \max\{X(\omega), 0\}$ and $X^-(\omega) = \max\{-X(\omega), 0\}$ (in the future we may abbreviate maximum as $X \vee 0$). If $\mathbb{P}(\{\omega \in \Omega : X^+(\omega) = \infty\}) = \mathbb{P}(\{\omega \in \Omega : X^-(\omega) = \infty\}) = 0$, then we say X is **integrable** and have $\int_\Omega X(\omega) \, d\mathbb{P}(\omega) = \int_\Omega X^+(\omega) \, d\mathbb{P}(\omega) - \int_\Omega X^-(\omega) \, d\mathbb{P}(\omega)$. If both $\mathbb{P}(\{\omega \in \Omega : X^+(\omega) = \infty\}) > 0$ and $\mathbb{P}(\{\omega \in \Omega : X^-(\omega) = \infty\}) > 0$, then the Lebesgue Integral is undefined. If only one of the positive or negative parts of X takes values of infinity with non-zero probability, then the Lebesgue Integral is either ∞ (in the case where $0 = \mathbb{P}(\{\omega \in \Omega : X^-(\omega) = \infty\}) < \mathbb{P}(\{\omega \in \Omega : X^+(\omega) = \infty\})$) or $-\infty$ (in the other case).

We may be interested in integrating our random variable over a subset A of Ω. In such cases, we write $\int_A X(\omega) \, d\mathbb{P}(\omega) = \int_\Omega \mathbb{1}_A(\omega) X(\omega) \, d\mathbb{P}(\omega)$ where $\mathbb{1}_A(\omega)$ is the indicator function previously defined. Note that in all cases, we are integrating with respect to the probability measure in question, since the same event may have different probabilities under different measures. We define the expectation of X as it's Lebesgue Integral, and write $\mathbb{E}(X) = \int_\Omega X(\omega) \, d\mathbb{P}(\omega)$. As we'll see below, this is just one of many ways to define expectation.

Example 3.1: Consider the function $f : [0,1] \to \{0,1\}$ taking on values of zero if x is rational and 1 otherwise. A result from analysis is that $\mathbb{Q}$ is dense in $\mathbb{R}$ (that is, in any interval of $\mathbb{R}$ there will be both rational and irrational numbers). For this reason, regardless of the size of $\|\Pi\|$, the Upper Riemann Sum will always be 1 and the Lower Riemann Sum will always be 0; the Riemann Sum is undefined. In contrast, if $X(\omega)$ is a random variable defined in the same manner, then the Lebesgue Integral *is* defined (in fact, it is 1). To see why this is the case, recall that $\mathbb{Q}$ is countable, and thus, by countable additivity and the fact that any individual point has probability zero, $\mathbb{P}(\{\omega \in \Omega : X(\omega) = 0\}) = 0$. So, since $\mathbb{P}([0,1]) = 1$, $\mathbb{P}(\{\omega \in \Omega : X(\omega) = 1\}) = 1$.

Definition 3.2. Riemann-Stieljes Integral: While the Lebesgue Integral allows for maximum generality (for the purposes of these notes), to actually compute expectations, it often suffices to use the integrals more familiar to us. The expectation of a function g of any random variable X with cumulative distribution function F_X is calculated as $\mathbb{E}\big(g(X)\big) = \int_{-\infty}^{\infty} g(x) \, dF_X(x)$. By definition, $F_X(x) = \int_{-\infty}^{x} f_X(t) \, dt$ where f_X is the density function of X. By the fundamental theorem of calculus, this means $dF_X(x) = f_X(x) \, dx$. In particular, $\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) \cdot f_X(x) \, dx$.

Definition 3.3. Random Variable (Simple): A random variable X is simple whenever there are only finitely many values that X can take, that is, if there exists $x_1, x_2, \ldots, x_n \in \mathbb{R}$ such that for all $\omega \in \Omega$, $\mathbb{P}(X(\omega) \in \{x_1, x_2, \ldots, x_n\}) = 1$.

Definition 3.4. Random Variable (Bounded): A random variable X is bounded whenever there exists a $c \in \mathbb{R}$ such that for all $\omega \in \Omega$, $\mathbb{P}(|X(\omega)| < c) = 1$.

Definition 3.5. Random Variable (Non-negative): A random variable X is non-negative if for all $\omega \in \Omega$, $\mathbb{P}(X(\omega) \geq 0) = 1$.

Definition 3.6. Expectation: The expectation of a random variable X, denoted $\mathbb{E}(X)$, obeys

1. Linearity: for all random variables X, Y and constants c, $\mathbb{E}(cX + Y) = c\mathbb{E}(X) + \mathbb{E}(Y)$.

2. Non-negativity: if $\mathbb{P}(X > 0) = 1$ then $\mathbb{E}(X) \geq 0$.

We define the calculation for $\mathbb{E}$ in four stages in the theorem section below: first for simple random variables, then for bounded random variables, then for non-negative random variables, then for general random variables. At each stage, we calculate the expectation differently, and check that it agrees with previous calculations and meet the criteria for expectations above. While this is a useful exercise, actually computing expectations is usually easier done with the previous two pieces of machinery (Lebesgue and Riemann-Stieljes Integration).

Definition 3.7. Variance: The variance of a random variable X, denoted $\mathbb{V}(X)$, is the value $\mathbb{E}\left[(X - \mathbb{E}(X))^2\right] = \mathbb{E}(X^2) - \big(\mathbb{E}(X)\big)^2$. The square root of the variance is called the **standard deviation**; $\sqrt{\mathbb{V}(X)} = \sigma$.

Definition 3.8. Covariance: The covariance of random variables X and Y is $\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}\left[\big(X - \mathbb{E}(X)\big)\big(Y - \mathbb{E}(Y)\big)\right]$. This is a generalization of variance, since $\mathbb{V}(X) = \mathrm{Cov}(X, X)$. When the covariance is zero, we say the random variables are uncorrelated.

Example 3.2: Everyone knows that "correlation doesn't equal causation". The reverse can also be shown to be true. For example if X is a uniform random variable on $(-1, 1)$, and $Y = X^2$ is another random variable, then $\mathbb{E}(XY) = \mathbb{E}(X^3) = \int_{-1}^{1} x^3 \, dF(t) = \frac{1}{2} \int_{-1}^{1} x^3 \, dx = \frac{x^4}{8}\big|_{-1}^{1} = 0$ and by the symmetry of the support of X, $\mathbb{E}(X) = 0$. So even though Y is literally caused by X, $\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathrm{Cov}(X, Y) = 0$.

Definition 3.9. Correlation: The correlation coefficient between random variables X and Y is $\rho(X, Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}}$. Note $\rho \in [-1, 1]$.

Definition 3.10. Coefficient Of Determination: Where $\rho(X, Y)$ is the correlation between random variables X and Y, the coefficient of determination is simply it's square; $r^2 = \rho(X, Y)^2$. Note $r^2 \in [0, 1]$.

Definition 3.11. π-system: A collection of sets $\mathcal{P}$ from a non-empty set Ω is a pi-system provided $\mathcal{P}$ is closed under finite intersection. That is, $\mathcal{P}$ is a pi-system if whenever $A_1, A_2, \ldots A_n \in \mathcal{P}$ we have $\bigcap_{i=1}^{n} A_i \in \mathcal{P}$.

Example 3.3: Where $\Omega = \{1, 2, 3\}$, $\mathcal{P} = \{\emptyset, \{1, 2\}, \{2, 3\}, \{2\}, \{3\}\}$ is a pi-system but not an algebra (Definition 1.9, Page 5). It is not an algebra since, e.g., $\{1, 2\} \cup \{3\} = \{1, 2, 3\} \notin \mathcal{P}$.

Non-example 3.1: Where $\Omega = \{1, 2, 3\}$, $\mathcal{P} = \{\emptyset, \{1, 2\}, \{2, 3\}, \Omega\}$ is not a pi-system since $\{1, 2\} \cap \{2, 3\} = \{2\} \notin \mathcal{P}$.

Definition 3.12. λ-system: A collection of sets $\mathcal{L}$ from a non-empty set Ω is a lambda-system provided $\mathcal{L}$ is closed under compliment and countable disjoint union. That is, $\mathcal{L}$ is a lambda-system if whenever $A_1, A_2, \cdots \in \mathcal{L}$ are disjoint, we have $\biguplus_{i=1}^{n} A_i \in \mathcal{L}$ and $A_1^c \in \mathcal{L}$.

Example 3.4: Where $\Omega = \{1, 2, 3, 4, 5\}$, the set $\mathcal{L} = \{\emptyset, \{1, 5\}, \{4, 5\}, \{2, 3, 4\}, \{1, 2, 3\}, \Omega\}$ is a lambda-system (but not a pi-system (Definition 3.11, Page 22), since e.g., $\{2, 3\} = \{2, 3, 4\} \cap \{1, 2, 3\} \notin \mathcal{L}$, and not an algebra (Definition 1.9, Page 5), since e.g., $\{1, 5\} \cup \{4, 5\} = \{1, 4, 5\} \notin \mathcal{L}$).

Non-example 3.2: Where $\Omega = \mathbb{R}$, $\mathcal{L} = \{(a, b) : a, b \in \mathbb{R}\}$ is a pi-system (Definition 3.11, Page 22) but not a lambda-system. For example, $(1, 2)$ and $(3, 4)$ are disjoint open intervals that are both in $\mathcal{L}$, but their union is not an open interval.

Definition 3.13. Semi-Algebra: A collection of sets $\mathcal{S}$ from a non-empty set Ω is a semi-algebra provided $\mathcal{S}$ is closed under intersection and each compliment is some finite disjoint union from $\mathcal{S}$ (even if the compliment is not in $\mathcal{S}$). That is, $\mathcal{S}$ is a semi-algebra if whenever $A_1, A_2, \ldots, A_n \in \mathcal{S}$, we have $A_i \bigcap A_j \in \mathcal{S}$ and $A_j^c = \biguplus_{i=1}^{n} A_i$.

Example 3.5: Where $\Omega = \{1, 2, 3\}$, $\mathcal{S} = \{\emptyset, \{1\}, \{2\}, \{3\}\}$ is a semi-algebra but not an algebra (Definition 1.9, Page 5).

Non-example 3.3: Where $\Omega = \mathbb{N}$, $\mathcal{S} = \{\emptyset, \{1\}, \{2\}, \ldots\}$ (the set of singletons) is a pi-system (Definition 3.11, Page 22) but not a semi-algebra since every compliment of a singleton is an infinite union.

3.2 Theorems And Examples

Theorem 3.1. Consequences Of Expectations:

1. Respects dominance, if $\mathbb{P}(X \leq Y) = 1$ then $\mathbb{E}(X) \leq \mathbb{E}(Y)$.

2. Respects equality, if $\mathbb{P}(X = Y) = 1$ then $\mathbb{E}(X) = \mathbb{E}(Y)$.

3. Triangle Inequality, $|\mathbb{E}(X)| \leq \mathbb{E}(|X|)$.

Proof. For 1, $\mathbb{P}(X \leq Y)=1 \implies \mathbb{P}(Y-X\geq 0)=1 \implies \mathbb{E}(Y-X) \geq 0 \implies \mathbb{E}(Y)-\mathbb{E}(X) \geq 0$.
For 2, apply the proof of 1 twice.
For 3, $-|X|\leq X\leq |X| \implies \mathbb{E}(-|X|) \leq \mathbb{E}(X) \leq \mathbb{E}(|X|) \implies -\mathbb{E}(|X|) \leq \mathbb{E}(X) \leq \mathbb{E}(|X|)$. ∎

Theorem 3.2. Expectations Of Simple Random Variables: Where X is a simple random variable taking values $x_i, \ldots, x_n$, $\mathbb{E}(X) = \sum_{i=1}^{n} x_i \mathbb{P}(X = x_i)$.

Proof. The proof of this and all the below verify that the specific type of random variable meets the definition for expectations (Definition 3.6, Page 21). So let c be any constant and let $X, Y : \Omega \to \mathbb{R}$ be any simple random variables.

Does $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$? Label the elements in the codomain of X $\{x_1, \ldots, x_n\}$ and the elements in the codomain of Y $\{y_1, \ldots, y_m\}$. Then

$$\begin{aligned}
\mathbb{E}(X + Y) &= \sum_{k=1}^{l} z_k \mathbb{P}(X + Y = z_k) && \text{Definition of simple random variable} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} (x_i + y_j) \mathbb{P}(X = x_i, Y = y_j) && \text{Possibly many ways to get same } z_k \\
&= \sum_{i=1}^{n} \sum_{j=1}^{m} x_i \mathbb{P}(X = x_i, Y = y_j) + \sum_{i=1}^{n} \sum_{j=1}^{m} y_j \mathbb{P}(X = x_i, Y = y_j) \\
&= \sum_{i=1}^{n} x_i \mathbb{P}(X = x_i) + \sum_{j=1}^{m} y_j \mathbb{P}(Y = y_j) = \mathbb{E}(X) + \mathbb{E}(Y)
\end{aligned}$$

Does $\mathbb{E}(cX) = c\mathbb{E}(X)$ for all $c \in \mathbb{R}$? If $c = 0$ this is trivially true, so assume otherwise. Then the finite list of elements in the image of X, call them $x_1, x_2, \ldots, x_n$ remain finite upon being multiplied by c (they are $cx_1, cx_2, \ldots, cx_n$). By how we defined expectation for simple random variables, $\mathbb{E}(cX)=\sum_{i=1}^{n} cx_i \mathbb{P}(cX=cx_i)=c\sum_{i=1}^{n} x_i \mathbb{P}(cX=cx_i)=c\sum_{i=1}^{n} x_i \mathbb{P}(X=x_i)=c\mathbb{E}(X)$.

Does the definition respect non-negativity? If $\mathbb{P}(X > 0) = 1$, then $x_1, \ldots, x_n$ are all greater than zero. Then for every $i \in [1, n]$, $x_i \mathbb{P}(X = x_i) \geq 0$. So $\mathbb{E}(X) = \sum_{i=1}^{n} x_i \mathbb{P}(X = x_i) \geq 0$. So we have verified linearity and negativity, which proves that our definition for simple random variables fits. ∎

Theorem 3.3. Expectations Of Bounded Random Variables: Where Y is bounded, use approximation and define the expected value of Y as $\mathbb{E}(Y) = \sup\limits_{\substack{X \text{simple} \\ X \leq Y}} \mathbb{E}(X) = \inf\limits_{\substack{X \text{simple} \\ X \geq Y}} \mathbb{E}(X).$

Proof. Since simple random variables are bounded, we need to make sure this definition agrees with the definition we gave for simple random variables. If Y is simple, then by the sup part of the equation, the expectation as defined by bounded random variables is at least as big as the expectation as defined by simple random variables. On the other hand, by the inf part of the equation, the expectation as defined by the bounded random variable is at most as small as the expectation as defined by simple random variables. So the definitions agree.

First, we must prove that the definition is valid (i.e. that the supremum and infimum actually agree). Consider any simple random variables X_1, X_2 such that $X_1 \leq Y \leq X_2$. Expectations respect dominance (Theorem 3.1, Page 23), so $\mathbb{E}(X_1) \leq \mathbb{E}(X_2)$. Since this holds regardless of the choice of X_1, X_2, $\sup\limits_{\substack{X \text{simple} \\ X \leq Y}} \mathbb{E}(X) < \inf\limits_{\substack{X \text{simple} \\ X \geq Y}} \mathbb{E}(X)$. For the other direction, let $\varepsilon > 0$ be given and partition Ω into the sets $A_k = \{\omega \in \Omega : k\varepsilon \leq Y(\omega) \leq (k+1)\varepsilon\}$. Consider the random variable $X_1 = \sum\limits_k k\varepsilon \mathbb{1}_{A_k}$. By the way Ω is partitioned, any $\omega \in \Omega$ is in exactly one A_k. So $X_1(\omega) = k\varepsilon$ whenever $k\varepsilon \leq Y(\omega) \leq (k+1)\varepsilon$ (note X_1 is not necessarily a constant because the k is changing). We have $X_1 \leq Y \leq X_1 + \varepsilon$. Further since Y is bounded, there are only finitely many A_k's needed to cover Ω, and thus X_1 is simple. So we see our definition is valid by taking ε to zero and observing:

$$\sup\limits_{\substack{X \text{simple} \\ X \leq Y}} \mathbb{E}(X) \geq \mathbb{E}(X_1) \qquad X_1 \text{ is a simple random variable less than } Y$$

$$= \mathbb{E}(X_1 + \varepsilon) - \varepsilon \qquad \text{Simple expectations are linear}$$

$$\geq \inf\limits_{\substack{X \text{simple} \\ X \geq Y}} \mathbb{E}(X) - \varepsilon \qquad X_1 + \varepsilon \text{ is a simple random variable greater than } Y$$

Now that we've proved the definition is both valid and matches previous definitions, we can move on to proving it meets the criteria for expectations. So let Y_1 and Y_2 be bounded random variables and $c \in \mathbb{R}$ a non-zero constant.

Is $\mathbb{E}(Y_1 + Y_2) = \mathbb{E}(Y_1) + \mathbb{E}(Y_2)$? If X_1, X_2 are simple random variable such that $X_1 \leq Y_1$ and $X_2 \leq Y_2$, then by the supremum definition and the linearity of simple expectations, $\mathbb{E}(Y_1 + Y_2) \geq \mathbb{E}(X_1 + X_2) \geq \mathbb{E}(X_1) + \mathbb{E}(X_2) \geq \mathbb{E}(Y_1) + \mathbb{E}(Y_2)$. On the other hand if X_1, X_2 are simple random variables such that $X_1 \geq Y_1$ and $X_2 \geq Y_2$, then by the infimum definition and the linearity of simple expectations, we get the reverse inequality. Is $\mathbb{E}(cY_1) = c\mathbb{E}(Y_1)$? If $c > 0$, then $\mathbb{E}(cY_1) = \sup\limits_{\substack{cX \text{simple} \\ cX \leq cY_1}} \mathbb{E}(cX) = \sup\limits_{\substack{X \text{simple} \\ X \leq Y_1}} c\mathbb{E}(X) = c \sup\limits_{\substack{X \text{simple} \\ X \leq Y_1}} \mathbb{E}(cX) = c\mathbb{E}(Y_1)$. If $c < 0$, then $\mathbb{E}(cY_1) = \sup\limits_{\substack{cX \text{simple} \\ cX \leq cY_1}} \mathbb{E}(cX) = \sup\limits_{\substack{X \text{simple} \\ X \geq Y_1}} c\mathbb{E}(X) = c \inf\limits_{\substack{X \text{simple} \\ X \geq Y_1}} \mathbb{E}(X) = c\mathbb{E}(Y_1)$.

Does the definition respect non-negativity? If $\mathbb{P}(Y > 0) = 1$ then the constant random variable $X = 0$ is simple with $X \leq Y$, so $\mathbb{E}(Y) \geq \mathbb{E}(X) = 0$ from the supremum definition. ∎

Theorem 3.4. Expectations Of Non-Negative Random Variables: Where Z is a non-negative random variable, use truncation and define $\mathbb{E}(Z) = \sup \{\mathbb{E}(Y) : 0 \leq Y \leq Z, Y \text{ bounded}\}$.

Lemma 3.4.1. If Z is non-negative, then $\mathbb{E}(Z) = \lim_{n \to \infty} \mathbb{E}(Z \wedge n)$ (where $Z \wedge n$ denotes the minimum between $Z(\omega)$ and n). For every n, we know $Z \wedge n$ is bounded and non-negative. For any fixed n, $(Z \wedge n) \leq Z$, and, since expectations respect dominance, $\mathbb{E}(Z \wedge n) \leq \mathbb{E}(Z)$. Since this relationship holds for every n, $\limsup_{n \to \infty} \mathbb{E}(Z \wedge n) \leq \mathbb{E}(Z)$. On the other hand, if Y is a bounded, non-negative random variable such that $Y \leq Z$, then we can find a m large enough such that $\mathbb{P}(Y \leq m) = 1$. In other words, we have $\mathbb{P}(\{Y \leq Z\} \cap \{Y \leq m\}) = \mathbb{P}(Y \leq Z \wedge m) = 1$ and thus $\mathbb{E}(Y) \leq \mathbb{E}(Z \wedge m)$. Since this relationship holds for every $n \geq m$, $\mathbb{E}(Y) \leq \liminf_{n \to \infty} \mathbb{E}(Z \wedge n)$. Taking the supremum of the Y's and in in conjunction with the previous inequality, we have proved the result.

Proof. We first need to verify that this definition agrees with the definition we gave for bounded random variables if those bounded random variables are also non-negative. If Z is bounded and non-negative, then by the new definition, the expectation of Z is at least as big as the old definition (because it is the supremum). On the other hand, if Y is any other bounded non-negative random variable less than Z, then by the dominance of expectations, $\mathbb{E}(Y) \leq \mathbb{E}(Z)$ (in light of the old definition), and then taking the supremum of the left-hand side, we see the expectation with the new definition is at most the expectation with the old definition. So the definitions agree.

Now we aim to prove linearity and non-negativity. Let Z_1 and Z_2 be non-negative random variables and $c \in \mathbb{R}_+$ a real constant (we can ignore the case where $c < 0$ since the random variables are non-negative).

Is $\mathbb{E}(Z_1 + Z_2) = \mathbb{E}(Z_1) + \mathbb{E}(Z_2)$? Observe $\mathbb{E}(Z_1 + Z_2) \leq \mathbb{E}(Z_1) + \mathbb{E}(Z_2)$ since

$$
\begin{aligned}
\mathbb{E}(Z_1 + Z_2) &= \lim_{n \to \infty} \mathbb{E}\big((Z_1 + Z_2) \wedge n\big) && \text{By Lemma 3.4.1} \\
&\leq \lim_{n \to \infty} \mathbb{E}\big((Z_1 \wedge n) + (Z_2 \wedge n)\big) && \text{If } Z_1 + Z_2 > n \text{ but } Z_1 < n \text{ and } Z_2 < n \\
&= \lim_{n \to \infty} \mathbb{E}(Z_1 \wedge n) + \mathbb{E}(Z_2 \wedge n) && \text{Properties of bounded random variables} \\
&= \mathbb{E}(Z_1) + \mathbb{E}(Z_2) && \text{By Lemma 3.4.1}
\end{aligned}
$$

Now let Y_1 and Y_2 be bounded non-negative random variables with $Y_1 \leq Z_1$ and $Y_2 \leq Z_2$. Then $\mathbb{E}(Z_1 + Z_2) \geq \mathbb{E}(Y_1 + Y_2) = \mathbb{E}(Y_1) + \mathbb{E}(Y_2)$ and taking the supremum on the right gives $\mathbb{E}(Z_1 + Z_2) \geq \mathbb{E}(Z_1) + \mathbb{E}(Z_2)$.

Is $\mathbb{E}(cZ) = c\mathbb{E}(Z)$? By definition, $\mathbb{E}(cZ) = \sup \{\mathbb{E}(cY) : cY \leq cZ\}$ where cY is bounded and non-negative. We've proved that bounded random variables preserve linearity, so this is equivalent to $c \sup \{\mathbb{E}(Y) : Y \leq Z, Y \text{ bounded and non-negative}\} = c\mathbb{E}(Z)$ as desired. It is clear $\mathbb{P}(Z \geq 0) = 1 \implies \mathbb{E}(Z) \geq 0$ since each Y in $\mathbb{E}(Z) = \{\mathbb{E}(Y) : Y \leq Z\}$ where Y is bounded and non-negative is itself bounded, and we already showed that bounded random random variables preserve non-negativity. Since $\mathbb{E}(Z)$ is the supremum of these non-negative expectations, $\mathbb{E}(Z)$ is non-negative. ∎

Theorem 3.5. Expectations Of General Random Variable: Where X is a general random variable, use partitioning and define $\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-)$ so long as the right-hand side isn't $\infty - \infty$. Here, the positive part of X is denoted $X^+ = \max\{X, 0\}$ and the negative part of X denoted $X^- = \max\{(-X), 0\}$. Since both of these random variables are non-negative, and functions of random variables are random variables, expectation here is well defined.

Proof. Does this definition agree with previous definitions? If X is non-negative, then $\mathbb{P}(X^- = 0) = 1$, so $\mathbb{E}(X) = \mathbb{E}(X^+) - 0$ as desired.

For generic X and Y, does $\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$? Observe:

$$\left[(X + Y)^+ - (X + Y)^- = (X + Y)\right] = \left[(X) + (Y) = (X^+ - X^-) + (Y^+ - Y^-)\right]$$
$$(X + Y)^+ + X^- + Y^- = X^+ + Y^+ + (X + Y)^-$$
$$\mathbb{E}\left((X + Y)^+\right) + \mathbb{E}\left(X^-\right) + \mathbb{E}\left(Y^-\right) = \mathbb{E}\left(X^+\right) + \mathbb{E}\left(Y^+\right) + \mathbb{E}\left((X + Y)^-\right)$$
$$\mathbb{E}\left((X + Y)^+\right) - \mathbb{E}\left((X + Y)^-\right) = \mathbb{E}\left(X^+\right) - \mathbb{E}\left(X^-\right) + \mathbb{E}\left(Y^+\right) - \mathbb{E}\left(Y^-\right)$$
$$\mathbb{E}(X + Y) = \mathbb{E}(X) + \mathbb{E}(Y)$$

For any c, is $\mathbb{E}(cX) = c\mathbb{E}(X)$? If $c > 0$, then $\mathbb{E}(cX) = \mathbb{E}(cX^+) - \mathbb{E}(cX^-) = c(\mathbb{E}(X^+) - \mathbb{E}(X^-)) = c\mathbb{E}(X)$. Similarly if $c < 0$, then $\mathbb{E}(cX) = \mathbb{E}(cX^+) - \mathbb{E}(cX^-) = -c\mathbb{E}(X^-) + c\mathbb{E}(X^+) = c\left(\mathbb{E}(X^+) - \mathbb{E}(X^-)\right) = c\mathbb{E}(X)$.

Does the definition respect non-negativity? If $\mathbb{P}(X > 0) = 1$, then $\mathbb{P}(X^- = 0) = 1$ and so $\mathbb{E}(X^-) = 0$ and we get $\mathbb{E}(X) = \mathbb{E}(X^+) - 0 \geq 0$ as desired and we've proved our claim. ∎

Lemma 3.5.1. If $\mathcal{L}$ is both a λ-system and a π-system, then $\mathcal{L}$ is a σ-algebra.

Proof. We need to show $\mathcal{L}$ is closed under countable union (as opposed to just countable disjoint union). So let $A_1, A_2, \cdots \in \mathcal{L}$ be given. Consider the events $A'_1 = A_1$, $A'_2 = A2 \cap A^c_1$, $A'_3 = A3 \cap A^c_2 \cap A^c_1$, etc. We know these events are in $\mathcal{L}$ as each of the compliments are in $\mathcal{L}$ (since $\mathcal{L}$ is a lambda-system) and as each of the intersections are in $\mathcal{L}$ (since $\mathcal{L}$ is a pi-system). Since each of these new events are disjoint, by the properties of lambda-systems, $\biguplus_{i=1}^{\infty} A'_i \in \mathcal{L}$.

However by the way we defined the A'_i events, $\biguplus_{i=1}^{\infty} A'_i = \bigcup_{i=1}^{\infty} A_i$ and we have our result. ∎

Lemma 3.5.2. The intersection of lambda-systems is a lambda-system.

Proof. Let $\mathcal{L}_1$ and $\mathcal{L}_2$ be lambda-systems and let $A_1, A_2, \cdots \in \mathcal{L}_1 \cap \mathcal{L}_2$ be given. Since $A_k \in \mathcal{L}_1$, $A^c_k \in \mathcal{L}_1$. Likewise since $A_k \in \mathcal{L}_2$, $A^c_k \in \mathcal{L}_2$. So $A^c_k \in \mathcal{L}_1 \cap \mathcal{L}_2$; it is closed under compliment. Likewise since the disjoint union is in both $\mathcal{L}_1$ and $\mathcal{L}_2$, it is in the intersection ∎

Lemma 3.5.3. Where $\mathcal{L}$ is a lambda-system and $A \in \mathcal{L}$, $\mathcal{L}_A = \{B \in \Omega : A \cap B \in \mathcal{L}\}$ is a lambda-system.

Proof. First see that $\mathcal{L}_A$ is non-empty since $\Omega \cap A = A \in \mathcal{L}$ so $\Omega \in \mathcal{L}_A$.

Next see $\mathcal{L}_A$ is closed under compliment. Consider any $B \in \mathcal{L}_A$. We want to show $B^c \in \mathcal{L}_A$, i.e. that $B^c \cap A \in \mathcal{L}$. Since $A \in \mathcal{L}$ and $\mathcal{L}$ is a lambda-system, $A^c \in \mathcal{L}$. Then since A^c and $A \cap B$ are disjoint elements in $\mathcal{L}$ and $\mathcal{L}$ is a lambda-system, $A^c \cup (A \cap B) \in \mathcal{L}$. Then $\left(A^c \cup (A \cap B) \right)^c \in \mathcal{L}$ and by DeMorgan's Laws, $A \cap (A^c \cup B^c) = A \cap B^c \in \mathcal{L}$. So $\mathcal{L}_A$ is closed under compliment.

Finally see $\mathcal{L}_A$ is closed under countable disjoint union. Consider disjoint $B_1, B_2, \cdots \in \mathcal{L}_A$. We want to show $\biguplus_{i=1}^{\infty} B_i \in \mathcal{L}_A$. This is equivalent to showing $\bigcup_{i=1}^{\infty} (B_i \cap A) \in \mathcal{L}$. Since each $B_i \in \mathcal{L}_A$, by the construction of $\mathcal{L}_A$, each $(B_i \cap A) \in \mathcal{L}$. But then since $\mathcal{L}$ is a lambda system and each $(B_i \cap A)$ is disjoint, $\bigcup_{i=1}^{\infty} (B_i \cap A) \in \mathcal{L}$ as desired and we're done. ∎

Theorem 3.6. Dynkin's π-λ Theorem: If $\mathcal{P}$ is a pi system and $\mathcal{L}$ a lambda system such that $\mathcal{P} \subseteq \mathcal{L}$, then $\sigma(\mathcal{P}) \subseteq \mathcal{L}$. This says that although $\mathcal{L}$ may fail to be a sigma-algebra, we may be able to find a portion of $\mathcal{L}$ that is a sigma-algebra.

Proof. Consider a pi-system $\mathcal{P}$ and a lambda-system $\mathcal{L}$ such that $\mathcal{P} \subseteq \mathcal{L}$. Call $\mathcal{L}'$ the smallest lambda-system containing $\mathcal{P}$. If we can show $\mathcal{L}'$ is a sigma-algebra we will have our results since $\sigma(\mathcal{P}) \subseteq \mathcal{L}'$ by the minimality of $\sigma(\mathcal{P})$ (from the perspective of sigma-algebras) and since $\mathcal{L}' \subseteq \mathcal{L}$ by the minimality of $\mathcal{L}'$ (from the perspective of lambda-systems). By Lemma 3.5.1, it suffices to show $\mathcal{L}'$ is a pi-system.

If $A \in \mathcal{P} \subseteq L'$, then for all $B \in \mathcal{P}$, $A \cap B \in \mathcal{P} \subseteq L'$ by the definition of pi-system. So for any $B \in \mathcal{P}$, $B \in \mathcal{L}_A$, which we showed is a lambda system in Lemma 3.5.3. But L' is the *smallest* Lambda system containing $\mathcal{P}$. So $L' \subseteq \mathcal{L}_A$.

On the other hand, if $C \in \mathcal{L}'$, then from the above argument, $C \in \mathcal{L}_A$ and also $A \subseteq \mathcal{L}_C$ for every $A \in \mathcal{P}$. So $\mathcal{P} \subseteq \mathcal{L}_C$, and again by the minimality of $\mathcal{L}'$, $\mathcal{P} \subseteq \mathcal{L}' \subseteq \mathcal{L}_C$. Since this holds for any C, from how $\mathcal{L}_C$ is defined, for any two events $X, Y \in \mathcal{L}'$, we must have $X \cap Y \in \mathcal{L}'$ as desired. ∎

3.3   Problems

Problem 3.1) Let $\lambda > 0$. Recall that X is said to have a Poisson(λ) distribution if $\mathbb{P}(X = k) = e^{-\lambda}\frac{\lambda^k}{k!}$ for $k \in \{0, 1, 2, \dots\}$.

a. Show that $\mathbb{E}(X) = \lambda$ (For this reason, λ is called the mean parameter).

Recall the Taylor Series expansion $e^x = \sum\limits_{n=0}^{\infty} \frac{x^n}{n!}$. Then using the formula for expectations:

$$\mathbb{E}(X) = \sum_{k=0}^{\infty} k \cdot \frac{\lambda^k e^{-\lambda}}{k!} = \sum_{k=0}^{\infty} k \cdot \frac{\lambda \lambda^{k-1} e^{-\lambda}}{k!} \qquad \text{Goal is to put in the form of Taylor Series above}$$

$$= \lambda e^{-\lambda} \sum_{k=0}^{\infty} \frac{k \lambda^{k-1}}{k!} \qquad\qquad \text{Terms don't depend on summand}$$

$$= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{k \lambda^{k-1}}{k!} \qquad\qquad \text{Limit change doesn't alter sum}$$

$$= \lambda e^{-\lambda} \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda} \sum_{n=0}^{\infty} \frac{\lambda^n}{(n)!} \qquad \text{Simplifying and substituting } n = k - 1$$

$$= \lambda e^{-\lambda} e^{\lambda} = \lambda \qquad\qquad\qquad \text{Taylor Series}$$

b. Compute $\mathbb{V}(X)$.

We can compute the variance as $\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$, so first need to find $\mathbb{E}(X^2)$. Using a similar method as above, see that:

$$\mathbb{E}(X^2) = \lambda e^{-\lambda} \sum_{k=1}^{\infty} k \cdot \frac{\lambda^{k-1}}{(k-1)!} \qquad\qquad \text{Same as part a with one more } k \text{ term}$$

$$= \lambda e^{-\lambda} \left(\sum_{k=1}^{\infty} (k-1) \cdot \frac{\lambda^{k-1}}{(k-1)!} + \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \right) \quad \text{Goal is to put in form of Taylor Series}$$

$$= \lambda e^{-\lambda} \left(\sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-2)!} + \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \right) \qquad \text{Simplifying}$$

$$= \lambda e^{-\lambda} \left(\lambda \sum_{k=2}^{\infty} \frac{\lambda^{k-2}}{(k-2)!} + \sum_{k=1}^{\infty} \frac{\lambda^{k-1}}{(k-1)!} \right) \qquad \text{Limit change doesn't alter sum}$$

$$= \lambda e^{-\lambda} \left(\lambda \sum_{n=0}^{\infty} \frac{\lambda^n}{(n)!} + \sum_{m=0}^{\infty} \frac{\lambda^m}{(m)!} \right) \qquad \text{Substituting } n = k - 2 \text{ and } m = k - 1$$

$$= \lambda e^{-\lambda} (\lambda e^{\lambda} + e^{\lambda}) = \lambda^2 + \lambda$$

Then using part a and the variance formula, $\mathbb{V}(X) = (\lambda^2 + \lambda) - \lambda^2 = \lambda$.

Problem 3.2) Let X and Y be any random variables on the same probability space. Show that $\mathbb{E}(|X+Y|) \leq \mathbb{E}(|X|)+\mathbb{E}(|Y|)$ and $\mathbb{E}(|X-Y|) \leq \mathbb{E}(|X|)+\mathbb{E}(|Y|)$.

For the first part, see:

$|X + Y| \leq |X| + |Y|$ The regular triangle inequality

$\mathbb{E}(|X + Y|) \leq \mathbb{E}(|X| + |Y|)$ Expectations respect dominance

$\mathbb{E}(|X + Y|) \leq \mathbb{E}(|X|) + \mathbb{E}(|Y|)$ Expectations are linear

For the second part, see:

$|X - Y| \leq |X| + |Y|$ The regular triangle inequality

$\mathbb{E}(|X - Y|) \leq \mathbb{E}(|X| + |Y|)$ Expectations respect dominance

$\mathbb{E}(|X - Y|) \leq \mathbb{E}(|X|) + \mathbb{E}(|Y|)$ Expectations are linear

Problem 3.3) Let Z be an integrable random variable. Show that for any $\varepsilon > 0$, there exists a simple random variable X such that $\mathbb{E}(|X - Z|) \leq \varepsilon$.

Let $\varepsilon > 0$ be given.

Since Z is integrable, $\mathbb{E}(Z) = \mathbb{E}(Z^+) - \mathbb{E}(Z^-)$ with each term finite and non-negative. From how we defined expectations for non-negative random variables, $\mathbb{E}(Z^\pm) = \sup\limits_{\substack{0 \leq Y \leq Z \\ Y \text{ bounded}}} \mathbb{E}(Y)$. By the definition of supremum, there exists non-negative bounded random variables Y_1 and Y_2 such that $Y_1 \leq Z^+$ and $Y_2 \leq Z^-$ and such that $\mathbb{E}(Z^+) \leq \mathbb{E}(Y_1)+\frac{\varepsilon}{4}$ and $\mathbb{E}(Z^-) \leq \mathbb{E}(Y_2)+\frac{\varepsilon}{4}$. In particular, $\mathbb{E}(|Z^+ - Y_1|) \leq \frac{\varepsilon}{4}$ and $\mathbb{E}(|Z^- - Y_2|) \leq \frac{\varepsilon}{4}$.

Since Y_1 and Y_2 are both bounded, from how we defined expectations for bounded random variables, $\mathbb{E}(Y_i) = \sup\limits_{\substack{X \leq Y_i \\ X \text{ simple}}} \mathbb{E}(X)$. Then we can likewise find simple functions X_1, X_2 such that $\mathbb{E}(Y_1) \leq \mathbb{E}(X_1) + \frac{\varepsilon}{4}$ and $\mathbb{E}(Y_2) \leq \mathbb{E}(X_2) + \frac{\varepsilon}{4}$. In particular, $\mathbb{E}(|Y_1 - X_1|) \leq \frac{\varepsilon}{4}$ and $\mathbb{E}(|Y_2 - X_2|) \leq \frac{\varepsilon}{4}$

Since X_1 and X_2 are random variables, so too is $X = X_1 - X_2$. Now observe that

$$
\begin{aligned}
\mathbb{E}(|Z - X|) &= \mathbb{E}(|Z^+ - Z^- - X_1 + X_2|) \\
&= \mathbb{E}(|Z^+ - Z^- - X_1 + X_2 + Y_1 - Y_1 + Y_2 - Y_2|) && \text{Creatively add zero} \\
&= \mathbb{E}(|Z^+ - Y_1 + Y_2 - Z^- + Y_1 - X_1 + X_2 - Y_2|) && \text{Reorder terms} \\
&\leq \mathbb{E}(|Z^+ - Y_1|) + \mathbb{E}(|Y_2 - Z^-|) + \mathbb{E}(|Y_1 - X_1|) + \mathbb{E}(|X_2 - Y_2|) && \text{From Problem 3.2} \\
&\leq \frac{\varepsilon}{4} + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} + \frac{\varepsilon}{4} = \varepsilon && \text{Desired result}
\end{aligned}
$$

Problem 3.4) Let X be a non-negative random variable with $\mathbb{E}(X)=0$. Show $\mathbb{P}(X=0) = 1$.

By the non-negative assumption of the proof, $\mathbb{P}(X < 0) = 0$. Due to this assumption, $X \geq \frac{1}{n}\mathbb{1}_{\left\{X(\omega)>\frac{1}{n}\right\}}$ for any $n \in \mathbb{N}$. The expectation of an indicator is the probability of the event. Then by the expectation assumption of the proof and since expectations respect dominance and linearity, we have:

$$0 = \mathbb{E}(X) \geq \mathbb{E}\left(\frac{1}{n}\mathbb{1}_{\left\{X(\omega)>\frac{1}{n}\right\}}\right) = \frac{1}{n}\mathbb{P}\left(X > \frac{1}{n}\right)$$

This inequality implies that $\mathbb{P}(X > \frac{1}{n}) = 0$. Since this holds for any $n \in \mathbb{N}$, choosing a strictly monotone increasing sequence of n and employing continuity from below gives us that $\mathbb{P}(X > 0) = 0$. To be explicit, label A_1 the event that $X > 1$, A_2 the event that $X > \frac{1}{2}$, A_3 the event that $X > \frac{1}{3}$, etc. Then $\mathbb{P}(X > 0) = \lim_{n \to \infty} \mathbb{P}(X > \frac{1}{n}) = \mathbb{P}\left(\bigcup_{i=1}^{\infty} A_i\right) = 0$ where the last equality follows from the union bound and minimality of $\mathbb{P}$. Then by the law of total probability, $\mathbb{P}(X = 0) = 1$ since we have shown $\mathbb{P}(X < 0) = \mathbb{P}(X > 0) = 0$.

Problem 3.5) Prove that a probability measure is uniquely determined by what it does on a generating π-system. Concretely, if $\mathbb{P}_1$ and $\mathbb{P}_2$ are two probability measures on $(\Omega, \mathcal{F})$ such that $\mathbb{P}_1(A) = \mathbb{P}_2(A)$ for all $A \in \mathcal{P}$, then $\mathbb{P}_1(B) = \mathbb{P}_2(B)$ for all $B \in \sigma(\mathcal{P})$.

Consider the set $L = \{B \in \sigma(\mathcal{P}) : \mathbb{P}_1(B) = \mathbb{P}_2(B)\}$. By assumption of the proof, $\mathcal{P} \subseteq L$ since $\mathbb{P}_1(A) = \mathbb{P}_2(A)$ for all $A \in \mathcal{P} \subseteq \sigma(\mathcal{P})$. So if we can show L is a λ-system, we will have $\sigma(\mathcal{P}) \subseteq L$ by the $\pi - \lambda$ Theorem (Theorem 3.6, Page 27). By the construction of L, this will prove our conclusion (if the measures agree on all of L, they also agree on a subset of L).

We go directly for the definition, and thus prove that L is closed under both compliment and countable disjoint union. So let A be any set in L and $A_1, A_2, \ldots$ be arbitrary disjoint sets in L. Since $\mathbb{P}_1(A) = \mathbb{P}_2(A)$, by the law of total probability and compliment rules, $\mathbb{P}_1(A^c) = 1 - \mathbb{P}_1(A) = 1 - \mathbb{P}_2(A) = \mathbb{P}_2(A^c)$; L is closed under compliment. Also, by countable additivity, $\mathbb{P}_1\left(\biguplus_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathbb{P}_1(A_n) = \sum_{n=1}^{\infty} \mathbb{P}_2(A_n) = \mathbb{P}_2\left(\biguplus_{n=1}^{\infty} A_n\right)$ (the middle equality follows since $\mathbb{P}_1(A_i) = \mathbb{P}_2(A_i)$ for all i); L is closed under countable disjoint union. So L is indeed a λ-system and we have our result.

Problem 3.6) Let X and Y be integrable random variables with $\mathbb{P}(X \leq Y) = 1$ and $\mathbb{E}(X) = \mathbb{E}(Y)$. Show that $\mathbb{P}(X = Y) = 1$.

Rearranging the probability, we have $\mathbb{P}\big((Y - X) > 0)\big) = 1$; $Y - X$ is a non-negative random variable. Using properties of expectations and the assumption of the proof, we have $\mathbb{E}(X)=\mathbb{E}(Y) \implies \mathbb{E}(Y) - \mathbb{E}(X)=0 \implies \mathbb{E}(Y - X)=0$. Then using Problem 3.4, $\mathbb{P}(Y - X = 0) = 1$ as desired.

Problem 3.7) Let X be uniformly distributed on $[0,1]$. Compute the expected value of the following random variables:

a. e^{5X}

The Riemann-Stieljes formula is $\mathbb{E}\big(f(X)\big) = \int_{-\infty}^{\infty} f(t)\, dF_x(t)$. The support of X is $[0,1]$ so we can reduce our problem to $\mathbb{E}\big(f(X)\big) = \int_0^1 f(t)\, dF_x(t)$. The CDF of a uniform random variable X is $F_X(t) = t \implies dF_X(t) = 1$. Here $f(t) = e^{5x}$. So our problem becomes $\mathbb{E}\big(e^{5X}\big) = \int_0^1 e^{5t} \cdot 1\, dt = \frac{1}{5}e^{5t}\big|_{t=0}^{t=1} = \frac{e^5-1}{5}$.

b. $1/X$

We again use the Riemann-Stieljes formula. $\mathbb{E}(\frac{1}{X}) = \int_0^1 \frac{1}{x}\, dt = \lim_{y \to 0^+} \ln|t|\big|_y^1 =0-(-\infty)=\infty$.

c. $\cos(\pi X)$

We again use the Riemann-Stieljes formula. $\mathbb{E}(\cos(\pi t)) = \int_0^1 \cos(\pi t)\, dt = \frac{1}{\pi}\sin(\pi t)\big|_0^1 = 0$.

d. $\lfloor 3.5X \rfloor$

The floor function indicates the largest integer less than or equal to the value. So here the random variable takes values 0 when $0 \leq \frac{7}{2}X(\omega) < 1$ or equivalently when $0 \leq X < \frac{2}{7}$. Likewise it takes values 1 when $1 \leq \frac{7}{2}X(\omega) < 2$ or equivalently when $\frac{2}{7} \leq X < \frac{4}{7}$, values 2 when $2 \leq \frac{7}{2}X(\omega) < 3$ or equivalently when $\frac{4}{7} \leq X < \frac{6}{7}$, and values 3 when $2 \leq \frac{7}{2}X(\omega) < 3$ or equivalently when $\frac{6}{7} \leq X \leq 1$.

Since $\frac{7}{2}X$ is a simple random variable, we can just use the formula for expectations of simple random variables, $\mathbb{E}(\frac{7}{2}X) = \sum_{i=0}^{3} i \cdot \mathbb{P}(\frac{7}{2}X = i) = \sum_{i=0}^{2}\big(i \cdot \frac{2}{7}\big) + \big(3 \cdot \frac{1}{7}\big) = \frac{9}{7}$.

e. $\max(X, 2/3)$

We can break up the Riemann-Stieljes formula and write $\mathbb{E}(\max(X, \frac{2}{3})) = \int_0^{\frac{2}{3}} \frac{2}{3}\, dt + \int_{\frac{2}{3}}^1 t\, dt = \left[\frac{2t}{3}\big|_0^{\frac{2}{3}}\right] + \left[\frac{1}{2}t^2\big|_{\frac{2}{3}}^1\right] = \frac{4}{9} + (\frac{1}{2} - \frac{4}{18}) = \frac{13}{18}$.

4 Norms And Important Inequalities

4.1 Definitions

Definition 4.1. Convex: A function whose second derivative is everywhere positive. Equivalently, a function $f : \mathbb{R} \to \mathbb{R}$ such that for all $t \in [0,1]$ and for all $x, y \in \mathbb{R}$, we have $f\big(tx + (1-t)y\big) \leq tf(x) + (1-t)f(y)$.

Definition 4.2. Raw Moment: The nth raw moment of a random variable X is the value $\mathbb{E}(X^n)$.

Definition 4.3. Central Moment: The nth central moment of a random variable X is the value $\mathbb{E}\left[(X - \mathbb{E}(X))^n\right]$.

Example 4.1: The 2nd central moment is the variance (Definition 3.7, Page 21).

Definition 4.4. Standard Moment: The nth central moment of a random variable X is the value $\mathbb{E}\left[\left(\frac{X - \mathbb{E}(X)}{\sigma}\right)^n\right]$ (where $\sigma = \sqrt{\mathbb{V}(X)}$, the standard deviation).

Example 4.2: The third standard moment is the **skewness**, which measures the symmetry of a distribution of a random variable. A random variable that is skewed to the right (the tail of the distribution is longer to it's right) will have a positive skew, and a random variable that is skewed to the left will have a negative skew.

Example 4.3: The fourth standard moment is the **kurtosis**, which measures how heavy-tailed a distribution is (how likely rare events are to occur). In Problem 4.2, Page 37, we show that the kurtosis of the distribution of a normal random variable is always 3. The **excess kurtosis** measures the kurtosis of a distribution of a random variable in relation to the normal distribution– a random variable with kurtosis greater than 3 (excess kurtosis greater than zero) indicates that the distribution of the random variable is **leptokurtic** (fatter-tailed than a normal distribution). A random variable with kurtosis less than 3 (excess kurtosis less than zero) indicates that the distribution of the random variable is **platykurtic** (thinner-tailed than a normal distribution).

Definition 4.5. Moment Generating Function: The moment generating function (MGF) for a random variable X is $M_X(t) = \mathbb{E}(e^{tX})$. The name of the function comes from the fact that the nth derivative of the MGF with respect to t, evaluated at 0, is the nth raw moment.

Definition 4.6. P-norm: The p norm of a random variable X is $\|X\|_p = \mathbb{E}\big(|X|^p\big)^{1/p}$. By Jensen's Inequality (Theorem 4.3, Page 33), if $p \leq q$, then $\|X\|_p \leq \|X\|_q$.

Definition 4.7. L^p Space: Fix a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. The space of random variables with finite p-norm is denoted $L^p(\mathbb{P}) = \{X : \Omega \to \mathbb{R} : \|X\|_p < \infty\}$. Since $p \leq q \implies \|X\|_p \leq \|X\|_q$, $L^p(\mathbb{P}) \supseteq L^q(\mathbb{P})$ (the spaces get more exclusive as p grows). In that sense, the most exclusive space is L^∞. In the conditions for which X belong in L^∞, define $\|X\|_\infty = \inf\{L \geq 0 : \mathbb{P}(|X| \leq L) = 1\}$.

Example 4.4: Where X is a random variable, $\mathbb{V}(X) < \infty$ if and only if $X \in L^2(\mathbb{P})$. If $X \in L^2(\mathbb{P})$, then $\mathbb{E}(|X|^2)^{1/2} < \infty \implies \mathbb{E}(|X|^2) < \infty$ by the definition of L^2. So since $\mathbb{E}(|X|^2) = \mathbb{E}(X^2)$, we know that $\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$ must be finite. The same reasoning works going the other direction.

4.2 Theorems And Examples

Theorem 4.1. Markov's Inequality: For any integrable random variable X and for any $t, p > 0$, $\mathbb{P}(|X|^p \geq t) \leq \frac{\mathbb{E}(|X|^p)}{t}$.

Proof.

$$
\begin{aligned}
\mathbb{P}(|X|^p \geq t) &= \mathbb{E}(\mathbb{1}_{\{|X|^p \geq t\}}) && \text{Expectation of indicator is probability of event} \\
&\leq \mathbb{E}\left(\mathbb{1}_{\{|X|^p \geq t\}}\frac{|X|^p}{t}\right) && \text{Indicator is only one when } \frac{|X|^p}{t} \geq 1 \\
&\leq \mathbb{E}\left(\frac{|X|^p}{t}\right) && \text{Removing cases where could be zero} \\
&= \frac{\mathbb{E}(|X|^p)}{t} && \text{Linearity of expectations}
\end{aligned}
$$

∎

Theorem 4.2. Cheyshev's Inequality: For any random variable $X \in L^2(\mathbb{P})$ and for any $t > 0$, $\mathbb{P}(|X - \mathbb{E}(X)| \geq t) \leq \frac{\mathbb{V}(X)}{t^2}$.

Proof. Squaring the inside of the probability, we have $\mathbb{P}(|X - \mathbb{E}(X)|^2 \geq t^2)$. Then from Marvov's Inequality (Theorem 4.1, Page 33), $\mathbb{P}(|X - \mathbb{E}(X)|^2 \geq t^2) \leq \frac{\mathbb{E}(X - \mathbb{E}(X))^2}{t^2} = \frac{\mathbb{V}(X)}{t^2}$. ∎

Theorem 4.3. Jensen's Inequality: When f is convex, $f\big(\mathbb{E}(X)\big) \leq \mathbb{E}\big(f(X)\big)$ for any integrate random variable X.

Proof. Since f is convex, the left derivative at any point is no greater than the right derivative at the point. In particular, the left derivative at $\mathbb{E}(X)$, call it $L_1 = \lim_{h \to 0+} \frac{f(\mathbb{E}(X)) - f(\mathbb{E}(X) - h)}{h}$, is less than or equal to the right derivative at $\mathbb{E}(X)$, call it $L_2 = \lim_{h \to 0+} \frac{f(\mathbb{E}(X) + h) - f(\mathbb{E}(X))}{h}$.

Let $a = \left(\frac{L_1 + L_2}{2}\right)$ and consider the real-valued function $l(x) = a(x - \mathbb{E}(X)) + f(\mathbb{E}(X))$. See that $l(x) \leq f(x)$ with equality holding at $x = \mathbb{E}(X)$. This can be shown in cases.

Since f is convex and a is the midpoint between L_1 and L_2, for all $h > 0$, we have $\frac{f(\mathbb{E}(X)) - f(\mathbb{E}(X) - h)}{h} \leq a \leq \frac{f(\mathbb{E}(X) + h) - f(\mathbb{E}(X))}{h}$. In the first case, $x > \mathbb{E}(X)$, choose $h = x - \mathbb{E}(X)$ and see $a(x - \mathbb{E}(X)) \leq f(\mathbb{E}(X) + (x - \mathbb{E}(X))) - f(\mathbb{E}(X)) = f(x) - f(\mathbb{E}(X))$ and so $l(x) = a(x - \mathbb{E}(X)) + f(\mathbb{E}(X)) \leq f(x)$. In the second case, $x < \mathbb{E}(X)$, choose $h = \mathbb{E}(X) - x$ and the same arithmetic follows.

Since $l(x) \leq f(x)$ and expectations respect dominance (Theorem 3.1, Page 23), we reach our conclusion: $[\mathbb{E}(l(X)) = \mathbb{E}(aX - a\mathbb{E}(X) + f(\mathbb{E}(X))) = f(\mathbb{E}(X))] \leq \mathbb{E}(f(X))$ ∎

Corollary 4.3.1. Variance is always non-negative. Take $f(x) = x^2$ which is clearly convex. Then $\mathbb{E}(X)^2 = f(\mathbb{E}(X)) \leq \mathbb{E}(f(X)) = \mathbb{E}(X^2)$ and so $\mathbb{V}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2 \geq 0$.

Theorem 4.4. Holder's Inequality: Given $p \in [1, \infty]$, let q be such that $\frac{1}{p} + \frac{1}{q} = 1$. Then $\|XY\|_1 \leq \|X\|_p \|Y\|_q$.

Proof. If either $\|X\|_p$ or $\|Y\|_q$ is zero, then so too is $\mathbb{E}(|XY|)$ and the result holds. So assume $\|X\|_p$ and $\|Y\|_q$ are both strictly greater than zero.

Where $y \geq 0$ is some fixed constant, consider a function $f(x) = \frac{x^p}{p} + \frac{y^q}{q} - xy$ defined for $x \geq 0$. Since $f'(x) = x^{p-1} - y$, f has a local extrema at $x = y^{\frac{1}{p-1}}$, call it x_0. Since $f''(x) = (p-1)x^{p-2} \geq 0$ (since $p \geq 1$ and $x \geq 0$), $f(x_0)$ is a minimum.

Further, $f(x_0) = \frac{x_0^p}{p} + \frac{y^q}{q} - x_0 y = \frac{y^{\frac{p}{p-1}}}{p} + \frac{y^q}{q} - (y^{\frac{1}{p-1}})y = \frac{y^q}{p} + \frac{y^q}{q} - y^q = y^q(\frac{1}{p} + \frac{1}{q}) - y^q = 0$ since $\frac{1}{p} + \frac{1}{q} = 0$ and thus $\frac{1}{q} = 1 - \frac{1}{p} \implies q = \frac{p}{p-1}$. As x_0 is the minimum of f and $f(x_0) = 0$, the term being subtracted, xy, must never be greater than the terms being added, $\frac{x^p}{p} + \frac{y^q}{q}$.

Now let $x = \frac{|X|}{\|X\|_p}$ and $y = \frac{|Y|}{\|Y\|_q}$. From the above, we see:

$$\frac{|X|}{\|X\|_p}\frac{|Y|}{\|Y\|_q} \leq \frac{\left(\frac{|X|}{\|X\|_p}\right)^p}{p} + \frac{\left(\frac{|Y|}{\|Y\|_q}\right)^q}{q}$$

$$\frac{1}{\|X\|_p\|Y\|_q}\mathbb{E}(|XY|) \leq \frac{\mathbb{E}(|X|^p)}{p\|X\|_p^p} + \frac{\mathbb{E}(|Y|^q)}{q\|Y\|_q^q}$$

$$\frac{1}{\|X\|_p\|Y\|_q}\|XY\|_1 \leq \frac{\|X\|_p^p}{p\|X\|_p^p} + \frac{\|Y\|_p^p}{q\|Y\|_q^q}$$

$$\frac{1}{\|X\|_p\|Y\|_q}\|XY\|_1 \leq \frac{1}{p} + \frac{1}{q}$$

$$\|XY\|_1 \leq \|X\|_p\|Y\|_q$$

∎

Corollary 4.4.1. Cauchy-Schwarz: The special case of Holder's Inequality where $p = q = 2$ is the Cauchy-Schwarz Inequality.

Proposition 4.1. $(1 + x) \leq e^x$ and $(1 - x) \leq e^{-x}$ for all $x \in \mathbb{R}_+$.

Proof. Using Taylor Series expansion, we have $(1 + x) \leq (1 + x + \frac{x^2}{2!} + \cdots) = e^x$. ∎

Theorem 4.5. Minkowski's Inequality: $\|X + Y\|_p \le \|X\|_p + \|Y\|_p$.

Proof. Recall that the Holder conjugate of $p \in [1, \infty]$ is the q such that $\frac{1}{p} + \frac{1}{q} = 1$; i.e. $q = \frac{1}{(1-\frac{1}{p})} = \frac{1}{\frac{p-1}{p}} = \frac{p}{p-1}$. Next keep in mind the Holder inequality, that $\mathbb{E}(|XY|) \le \|X\|_p \cdot \|Y\|_q$. Returning to the problem at hand, we see the following:

$$
\begin{aligned}
\mathbb{E}\big(|X + Y|^p\big) &= \mathbb{E}\big(|X + Y| \cdot |X + Y|^{p-1}\big) \\
&\le \mathbb{E}\big(|X| \cdot |X + Y|^{p-1} + |Y| \cdot |X + Y|^{p-1}\big) && \text{Triangle inequality on } |X + Y| \\
&= \mathbb{E}\big(|X| \cdot |X + Y|^{p-1}\big) + \mathbb{E}\big(|Y| \cdot |X + Y|^{p-1}\big) && \text{Linearity of expectations} \\
&\le \|X\|_p \cdot \||X + Y|^{p-1}\|_q + \|Y\|_p \cdot \||X + Y|^{p-1}\|_q && \text{Holder's Inequality on both terms} \\
&= \big(\|X\|_p + \|Y\|_q\big) \cdot \||X + Y|^{p-1}\|_q && \text{Grouping terms} \\
&= \big(\|X\|_p + \|Y\|_q\big) \cdot \mathbb{E}\left[\big((|X + Y|^{p-1})^{\frac{p}{p-1}}\big)\right]^{\frac{p-1}{p}} && \text{The Holder conjugate } q \\
&= \big(\|X\|_p + \|Y\|_q\big) \cdot \mathbb{E}\big(|X + Y|^p\big)^{\frac{p-1}{p}} && \text{Simplifying}
\end{aligned}
$$

Since $\frac{p-1}{p} - 1 = \frac{p-1}{p} - \frac{p}{p} = \frac{-1}{p}$, after dividing terms, we see:

$$
\frac{1}{\mathbb{E}\big(|X + Y|^p\big)^{\frac{-1}{p}}} \le \big(\|X\|_p + \|Y\|_q\big) \implies \mathbb{E}\big(|X + Y|^p\big)^{\frac{1}{p}} \le \big(\|X\|_p + \|Y\|_q\big)
$$

$\blacksquare$

Theorem 4.6. AM-GM Inequality: The arithmetic mean is always at least as large as the geometric mean. More formally, where $p_1, p_2, \ldots$ are real numbers such that $\sum\limits_{i=1}^{\infty} p_i = 1$, then for any non-negative real numbers $x_1, x_2, \ldots$, we must have $\sum\limits_{n=1}^{\infty} x_n p_n \ge \prod\limits_{n=1}^{\infty} x_n^{p_n}$

Proof. Fix a $n \in \mathbb{N}$, a set of positive real numbers $\Omega = \{x_1, x_2, \ldots, x_n\}$, and a simple random variable $X : \Omega \to \mathbb{R}$ given by $X(x_i) = \ln(x_i)$ with $\mathbb{P}(X = \ln(x_i)) = p_i$. Now consider the function $f(x) = e^x$ and with the help of Jensen's Inequality (Theorem 4.3, page 33) observe:

$$
\begin{aligned}
f\big(\mathbb{E}(X)\big) &\le \mathbb{E}\big(f(X)\big) && \text{Jensen's Inequality} \\
\left[e^{\mathbb{E}(X)} = e^{\sum\limits_{i=1}^{n} \ln(x_i)\mathbb{P}(X=\ln(x_i))}\right] &\le \left[\mathbb{E}\big(e^X\big) = \sum_{i=1}^{n} e^{\ln(x_i)} \cdot \mathbb{P}(X = \ln(x_i))\right] && \text{How } f \text{ was defined} \\
\prod_{i=1}^{n} e^{\ln(x_i)p_i} &\le \sum_{i=1}^{n} x_i \cdot p_i && \text{Properties of exponential} \\
\prod_{i=1}^{n} x_i^{p_i} &\le \sum_{i=1}^{n} x_i \cdot p_i && \text{Desired result}
\end{aligned}
$$

$\blacksquare$

Theorem 4.7. Payley-Zygmund Inequality: For any non-negative random variable X and any $\theta \in (0,1)$, $\mathbb{P}(X > \theta\mathbb{E}(X)) \geq (1-\theta)^2 \frac{\mathbb{E}(X)^2}{\mathbb{E}(X^2)}$

Proof. Observe:

$$
\begin{aligned}
\mathbb{E}(X) &= \mathbb{E}(X\mathbb{1}_{\{X \leq \theta\mathbb{E}(X)\}}) + \mathbb{E}(X\mathbb{1}_{\{X > \theta\mathbb{E}(X)\}}) \\
&\leq \theta\mathbb{E}(X) + \mathbb{E}(X\mathbb{1}_{\{X > \theta\mathbb{E}(X)\}}) \qquad &&\text{Properties of indicator function} \\
&\leq \theta\mathbb{E}(X) + \mathbb{E}(X^2)^{\frac{1}{2}} \cdot \mathbb{E}(\mathbb{1}^2_{\{X > \theta\mathbb{E}(X)\}})^{\frac{1}{2}} \qquad &&\text{Holder inequality} \\
&= \theta\mathbb{E}(X) + \mathbb{E}(X^2)^{\frac{1}{2}} \cdot \mathbb{E}(\mathbb{1}_{\{X > \theta\mathbb{E}(X)\}})^{\frac{1}{2}} \qquad &&\text{Square of 0 or 1 is still 0 or 1}
\end{aligned}
$$

Then after recalling the expected value of the indicator is the probability of the event:

$$
\begin{aligned}
\mathbb{E}(X) - \theta\mathbb{E}(X) &\leq \mathbb{E}(X^2)^{\frac{1}{2}}\mathbb{P}(X > \theta\mathbb{E}(X))^{\frac{1}{2}} \qquad &&\text{Subtracting} \\
\mathbb{E}(X)^2 - 2\theta\mathbb{E}(X)^2 + \theta^2\mathbb{E}(X)^2 &\leq \mathbb{E}(X^2)\mathbb{P}(X > \theta\mathbb{E}(X)) \qquad &&\text{Square both sides} \\
(1-\theta)^2\frac{\mathbb{E}(X)^2}{\mathbb{E}(X^2)} &\leq \mathbb{P}(X > \theta\mathbb{E}(X)) \qquad &&\text{Group for desired result}
\end{aligned}
$$

$\blacksquare$

4.3 Problems

Problem 4.1) Assume $\mathbb{E}(|X|) < \infty$. Show that $t \mapsto \mathbb{E}\left[(X - t)^2\right]$ achieves a unique minimum at $t = \mathbb{E}(X)$. That is, the expected value is the best deterministic approximation of X with respect to L^2 error.

Using properties of expectations, we can write $\mathbb{E}\left[(X - t)^2\right] = \mathbb{E}\left[X^2 - 2tX - t^2\right] = \mathbb{E}\left[X^2\right] - \mathbb{E}\left[2tX\right] - \mathbb{E}\left[t^2\right] = \mathbb{E}(X^2) - 2t\mathbb{E}(X) + t^2$. As a function of t, this equation has a derivative of $-2\mathbb{E}(X) + 2t$. When $t = \mathbb{E}(X)$, the derivative is zero, and so is an extrema. When $t > 0$, the derivative is positive. When $t < 0$, the derivative is negative. So the extrema is a minimum and we've shown the result.

Problem 4.2) Compute the kurtosis of a normal random variable X with generic parameters μ and σ^2.

The moment generating function (Definition 4.5, Page 32) of a random variable $Y \sim N(0, t)$ is given by:

$$M_Y(u) = \mathbb{E}(e^{uY}) = \int_{\mathbb{R}} e^{uy} f_Y(y)\, dy = \int_{\mathbb{R}} e^{uy} \frac{1}{\sqrt{2\pi t}} e^{\frac{-1}{2}\left(\frac{y^2}{t}\right)}\, dy$$

$$= \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi t}} \exp\left(uy - \frac{y^2}{2t}\right) dy$$

$$= e^{\frac{u^2 t}{2}} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi t}} \exp\left(uy - \frac{y^2}{2t} + \frac{u^2 t}{2}\right) dy$$

$$= e^{\frac{u^2 t}{2}} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi t}} \exp\left(\frac{uy2t}{2t} - \frac{y^2}{2t} + \frac{u^2 t^2}{2t}\right) dy$$

$$= e^{\frac{u^2 t}{2}} \int_{\mathbb{R}} \frac{1}{\sqrt{2\pi t}} e^{\frac{-1}{2}\left(\frac{(y - ut)^2}{t}\right)}\, dy$$

$$= e^{\frac{u^2 t}{2}}$$

Let $Z = X - \mu$. Using the above derivation and repeated applications of the chain and product rule, we see:

- $\frac{d}{du} M_Z(u) = u\sigma^2 e^{\frac{u^2 \sigma^2}{2}}$

- $\frac{d^2}{du^2} M_Z(u) = (\sigma^2) e^{\frac{u^2 \sigma^2}{2}} + u\sigma^2 \left(u\sigma^2 e^{\frac{u^2 \sigma^2}{2}}\right) = (\sigma^2 + u^2 \sigma^4) e^{\frac{u^2 \sigma^2}{2}}$

- $\frac{d^3}{du^3} M_Z(u) = (2u\sigma^4) e^{\frac{u^2 \sigma^2}{2}} + (\sigma^2 + u^2 \sigma^4)\left(u\sigma^2 e^{\frac{u^2 \sigma^2}{2}}\right) = (3u\sigma^4 + u^3 \sigma^6) e^{\frac{u^2 \sigma^2}{2}}$

- $\frac{d^4}{du^4} M_Z(u) = \left(3\sigma^4 + 3u^2 \sigma^6\right) e^{\frac{u^2 \sigma^2}{2}} + \left(3u\sigma^4 + u^3 \sigma^6\right)\left(u\sigma^2 e^{\frac{u^2 \sigma^2}{2}}\right)$

  $$= \left(3\sigma^4 + 3u^2 \sigma^6 + 3u^2 \sigma^6 + u^4 \sigma^8\right) e^{\frac{u^2 \sigma^2}{2}}$$

Then the kurtosis of X is: $\mathbb{E}\left[\left(\frac{(X - \mathbb{E}(X))}{\sigma}\right)^4\right] = \mathbb{E}\left[\left(\frac{\frac{d^4}{du^4} M_Z(0)}{\sigma^4}\right)\right] = \mathbb{E}\left[\frac{3\sigma^4}{\sigma^4}\right] = 3.$

5   Modes Of Convergence

5.1   Definitions

Definition 5.1. Convergence In Probability: A sequence of random variables X_n converges in probability to a random variable X, if for any $\varepsilon > 0$, $\lim_{n\to\infty} \mathbb{P}(|X_n - X| < \varepsilon) = 1$. We denote this $X_n \xrightarrow{\mathbb{P}} X$. To be precise, $X_n \xrightarrow{\mathbb{P}} X$ if $\lim_{n\to\infty} \mathbb{P}\left(\{\omega \in \Omega : |X_n(\omega) - X(\omega)| < \varepsilon\}\right) = 1$.

Example 5.1: Consider the typewriter sequence, $X_n = \mathbb{1}_{\left[\frac{n-2^k}{2^k}, \frac{n+1-2^k}{2^k}\right]}$, where for every n, k is the unique integer where $2^k \le n < 2^{k+1}$. After $X_1 = 1$ the first random variables are:

$$X_2(\omega) = \begin{cases} 1, \omega \le \frac{1}{2} \\ 0, \text{else} \end{cases}, X_3(\omega) = \begin{cases} 1, \omega \ge \frac{1}{2} \\ 0, \text{else} \end{cases}, X_4(\omega) = \begin{cases} 1, \omega \le \frac{1}{4} \\ 0, \text{else} \end{cases}, X_5(\omega) = \begin{cases} 1, \omega \in \left[\frac{1}{4}, \frac{1}{2}\right] \\ 0, \text{else} \end{cases}$$

Then $X_n \xrightarrow{\mathbb{P}} 0$ since $\mathbb{P}(|X_n| \ge \varepsilon) = \mathbb{P}(X_n = 1) = 2^{-k}$. As n grows to infinity, k does as well (since $n < 2^{k+1}$), and so $\lim_{n\to\infty} \mathbb{P}(X_n = 1) = \lim_{k\to\infty} 2^{-k} = 0$.

Non-example 5.1: The sequence $Y_n = \mathbb{1}_{\left[\frac{n \mod 3}{3}, 1\right]}$ does not converge in probability. Unlike the typewriter sequence, the "strip" under which $Y_n(\omega) = 1$ does not shrink to zero.

Definition 5.2. Almost Sure Convergence: A sequence of random variables X_n converges almost surely to a random variable X, denoted $X_n \xrightarrow{a.s.} X$, if $\mathbb{P}(\lim_{n\to\infty} X_n = X) = 1$. To be precise, this is saying $X_n \xrightarrow{a.s.} X$ if $\mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\right\}\right) = 1$.

Example 5.2: Consider the "escape to vertical infinity" sequence given by $X_n = n\mathbb{1}_{\left\{(0, \frac{1}{n})\right\}}$. Then for every $\omega \in [0, 1]$, $X_n(\omega) = 0$ for all $n > \frac{1}{\omega}$, and so $\lim_{n\to\infty} X_n(\omega) = 0$.

Non-example 5.2: The typewriter sequence (Example 5.1, Page 38) converges in probability but not almost surely since X_n never converges to a point (e.g. for any $\omega \in \Omega = [0, 1]$ $\limsup_{n\to\infty} X_n(\omega) = 1$ but $\liminf_{n\to\infty} X_n(\omega) = 0$). In general, almost sure convergence is a stronger result than convergence in probability (Theorem 5.1, Page 41).

Definition 5.3. Converges in L^p: A sequence of random variables X_n converges in L^p to X, denoted $X_n \xrightarrow{L^p} X$, if $X \in L^p(\mathbb{P})$ and $\lim_{n\to\infty} \|X_n - X\|_p = 0$ (Definition 4.6, Page 32). When dealing with $p = 1$, we may say "X_n converges in mean to X". When dealing with $p = 2$, we may say "X_n converges in mean-square to X".

Example 5.3: Convergence in L^p does not imply almost sure convergence. Take the "typewriter" sequence (Non-example 5.2, Page 38). We know X_n converges in mean to 0 since $\lim_{n\to\infty} \|X_n - 0\|_1 = \lim_{n\to\infty} \mathbb{E}(|X_n|) = \lim_{n\to\infty} 2^{-k} = 0$ (since k is such that $n < 2^{k+1}$).

Non-example 5.3: Convergence in probability (and therefore, convergence almost surely) does not imply converge in L^p. Take the "escape to vertical infinity" (Example 5.1, Page 38). We know that if $X_n \xrightarrow{\mathbb{P}} X$ and if $X_n \xrightarrow{L^p} Y$ then $X \overset{a.s.}{=} Y$. But in this example, $X_n \xrightarrow{a.s.} 0$ and yet for $p = 1$, $\lim_{n\to\infty} \|X_n - 0\|_p = \lim_{n\to\infty} \mathbb{E}(|X_n|) = \lim_{n\to\infty} n \cdot \frac{1}{n} = \lim_{n\to\infty} 1 = 1 \ne 0$.

Definition 5.4. Convergence in Distribution (Weak Convergence): A sequence of random variables X_n converges in distribution to a random variable X, denoted $X_n \xrightarrow{d} X$, if $\lim_{n \to \infty} F_{X_n}(x) = F_X(x)$ for all points x where the CDF (Theorem 1.3, Page 7) F_X is continuous. An equivalent definition is that $X_n \xrightarrow{d} X$ provided $\lim_{n \to \infty} \mathbb{E}(f(X_n)) = \mathbb{E}(f(X))$ for all bounded and continuous $f : \mathbb{R} \to \mathbb{R}$.

Example 5.4: We echo Example 2.5 and say that random variables and the distributions of said random variables are distinct concepts– different random variables can have the same distribution and a single random variable can have two different distributions (by changing the probability measure relevant to the sample space). For example, where $\Omega = \{H, T\}$ is the outcome of a single fair coin-flip, and where X and Y are random variables such that $X(H) = Y(T) = 1$ and $X(T) = Y(H) = 0$, then $X(\omega) \neq Y(\omega)$ for all $\omega \in \Omega$, and yet $X \overset{d}{=} Y$. For this reason, convergence in distribution is substantially weaker than the three convergence results above (convergence almost surely, convergence in probability, and convergence in L^p). Convergence in distribution doesn't even require random variables to be defined on the same probability space!

Take the following example. For every $n \in \mathbb{N}$, let $(\Omega_n = \{1, \ldots, n\}, \mathcal{F}_n = 2^{\Omega_n}, \mathbb{P})$ be a probability space where $\mathbb{P}$ is the uniform measure (i.e. for all $\omega \in \Omega_n$, $\mathbb{P}(\omega) = \frac{1}{n}$), and let $X_n : \Omega_n \to \mathbb{R}$ be the random variable such that $X_n(\omega) = \frac{\omega}{n}$. Further consider a probability space $(\Omega = [0, 1], \mathcal{F} = \mathbb{B}([0, 1]), \widehat{\mathbb{P}})$ where $\widehat{\mathbb{P}}$ is the Lebegue Measure (Example 1.5, Page 5), and let $X : \Omega \to \mathbb{R}$ be the random variable such that $X(\omega) = \omega$. Then $X_n \xrightarrow{d} X$ even though each X_n is a discrete random variable and X is a continuous random variable. See

that $F_{X_n}(x) = \begin{cases} 0, & x < \frac{1}{n} \\ \frac{\lfloor nx \rfloor}{n}, & \frac{1}{n} \leq x < 1 \\ 1, & 1 \leq x \end{cases}$ and $F_X(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \leq x < 1 \\ 1, & 1 \leq x \end{cases}$, which converge by the

squeeze theorem as $n \to \infty$ (since $\left[x - \frac{1}{n} = \frac{nx-1}{n} \right] < \frac{\lfloor nx \rfloor}{n} \leq x$).

Definition 5.5. Vague Convergence: A sequence of random variables converges vaguely if their distribution functions F_n converges to a monotone, right-continuous function $F : \mathbb{R} \to [0, 1]$, at all continuity points t of F. Note that F need not be a valid Cumulative Distribution Function (it's missing the condition that $\lim_{n \to \infty} F(x_n) = 1$, for example).

Example 5.5: "All mass escapes to infinity". Let $(\Omega = [0, 1], \mathbb{B}([0, 1]), \mathbb{P}([a, b]) = b - a)$ be a probability space, and for every $n \in \mathbb{N}$ consider the random variable $X_n(\omega) = n + \omega$. Then the sequence of random variables converges vaguely to $F(x) = 0$ since each random variable is a uniform on a unit interval that slides further and further along the real line.

Example 5.6: "Some mass escapes to infinity". Let $(\Omega = [0, 1], \mathbb{B}([0, 1]), \mathbb{P}([a, b]) = b - a)$ be a probability space, and for every $n \in \mathbb{N}$ consider the random variable $X_n(\omega) = \begin{cases} n, & \omega < \frac{1}{3} \\ 2, & \frac{1}{3} \leq \omega \end{cases}$.

Then $F_{X_n}(x) = \begin{cases} 0, & x < 2 \\ \frac{2}{3}, & 2 \leq x < n \\ 1, & x \geq n \end{cases}$ converges vaguely to $F(x) = \begin{cases} 0, & x < 2 \\ \frac{2}{3}, & 2 \leq x \end{cases}$.

Definition 5.6. Tightness: A sequence of random variables $\{X_n\}_{n\in N}$ are tight if for all $\varepsilon > 0$, there exists $a, b \in \mathbb{R}$ such that $\mathbb{P}(X_n \in [a, b]) \geq 1 - \varepsilon$. Equivalently, the sequence is tight if there exists $a, b \in \mathbb{R}$ such that $F_{X_n}(a) \leq \varepsilon$ and $F_{X_n}(b) \geq 1 - \varepsilon$. Also equivalently, the sequence is tight if there exists a $M > 0$ such that $\sup_n \mathbb{P}(|X_n| > M) < \varepsilon$.

Example 5.7: "No mass escapes to ∞". Let $(\Omega = [0,1], \mathbb{B}([0,1]), \mathbb{P}([a,b]) = b-a)$ be a probability space, and for every $n \in \mathbb{N}$ consider the random variable $X_n(\omega) = \begin{cases} n, & \omega < \frac{1}{n} \\ 2, & \frac{1}{n} \leq \omega \end{cases}$. Then

$$F_{X_n}(x) = \begin{cases} 0, & x < 2 \\ 1 - \frac{1}{n}, & 2 \leq x < n \\ 1, & n \leq x \end{cases}$$ converges weakly (and thus vaguely) to $F_X(x) = \begin{cases} 0, & x < 2 \\ 1, & 2 \leq x \end{cases}$.

The property that this example has which Examples 5.5 and 5.6 don't is the notion of tightness. See a choice of $a = 2$ and $b \geq \frac{1}{\varepsilon}$, yields $\mathbb{P}(X_n \in [a,b]) \geq 1 - \varepsilon$ since if $b \geq n$, $\mathbb{P}(X_n \in [a,b]) = 1$, and if $b < n$, $\mathbb{P}(X_n \in [a,b]) \geq \mathbb{P}(X_n = a) \geq 1 - \frac{1}{n} \geq 1 - \frac{1}{b} \geq 1 - \varepsilon$. In general, tightness upgrades vague convergence to weak convergence (see Theorem 5.12, page 46).

5.2   Theorems And Examples

Theorem 5.1. Almost Sure Convergence Implies Convergence in Probability: If $X_n \xrightarrow{a.s.} X$, then $X_n \xrightarrow{\mathbb{P}} X$. The converse is not true (Non-example 5.2, Page 38).

Proof. Since $X_n \xrightarrow{a.s.} X$, we have $\mathbb{P}\left(\limsup_{n\to\infty} \{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \varepsilon\}\right) = 0$, i.e. the event that for all $k \in \mathbb{N}$ there exists an $n \geq k$ such that $|X_n(\omega) - X(\omega)| > \varepsilon$ has probability zero. Identifying "there exists" with union, and "for all" with intersection, we have:

$$
\begin{aligned}
0 &= \mathbb{P}\left(\bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} \{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \varepsilon\}\right) \\
&= \lim_{k\to\infty} \mathbb{P}\left(\bigcup_{n=k}^{\infty} \{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \varepsilon\}\right) \\
&\geq \lim_{k\to\infty} \mathbb{P}\left(\{\omega \in \Omega : |X_k(\omega) - X(\omega)| > \varepsilon\}\right)
\end{aligned}
$$

The first step follows from Continuity From Above (Theorem 1.1, Page 7), since for all k,

$$
\left(\bigcup_{n=k}^{\infty} \{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \varepsilon\}\right) \supseteq \left(\bigcup_{n=k+1}^{\infty} \{\omega \in \Omega : |X_n(\omega) - X(\omega)| > \varepsilon\}\right)
$$

The second step follows from the fact that $A_k \subseteq \bigcup_{n=k}^{\infty} A_n$. ∎

Theorem 5.2. Convergence In Probability Implies Convergence In Distribution: If $X_n \xrightarrow{\mathbb{P}} X$, then $X_n \xrightarrow{d} X$. The converse is not true (Example 5.4, Page 39).

Proof. Pick any point t where F_X is continuous and let $\varepsilon > 0$ be given. Observe:

$$
\begin{aligned}
F_{X_n}(t) = \mathbb{P}(X_n \leq t) &= \mathbb{P}(X_n \leq t, X \leq t + \varepsilon) + \mathbb{P}(X_n \leq t, X > t + \varepsilon) \\
&\leq \mathbb{P}(X \leq t + \varepsilon) + \mathbb{P}(X_n \leq t, X > t + \varepsilon) \\
&\leq F_X(t + \varepsilon) + \mathbb{P}(X_n - X \leq t - X, t - X < -\varepsilon) \\
&\leq F_X(t + \varepsilon) + \mathbb{P}(X_n - X \leq -\varepsilon) \\
&\leq F_X(t + \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon)
\end{aligned}
$$

Since this holds for all $\varepsilon > 0$, since F_X is right continuous at t, and since our assumption is that $\limsup_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$, $\limsup_{n\to\infty} F_{X_n} \leq F_X$. Using the same setup as above, we know $F_{X_n}(t) \geq F_X(t - \varepsilon) - \mathbb{P}(|X_n - X| > \varepsilon)$ and so by left continuity have $\liminf_{n\to\infty} F_{X_n} \geq F_X$. ∎

Theorem 5.3. Convergence In L^p Implies Convergence In Probability: If $\|X_n - X\|_p \to 0$, then $X_n \xrightarrow{\mathbb{P}} X$. The converse is not true (Non-example 5.3, Page 38)

Proof. Let $\varepsilon > 0$ be given. We want to show $\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0$. Since $p \in [1, \infty)$, $\mathbb{P}(|X_n - X| \geq \varepsilon) = \mathbb{P}(|X_n - X|^p \geq \varepsilon^p)$. So by Markov's Inequality (Theorem 4.1, Page 33), we can write $\mathbb{P}(|X_n - X| \geq \varepsilon) \leq \frac{\mathbb{E}(|X_n - X|^p)}{\varepsilon^p}$. But since X_n converges in L^p, the numerator on the right-side of the inequality becomes 0 in the limit and we've proven our result. ∎

Lemma 5.3.1. There Are Only Countably Many Discontinuity Points In A CDF:
If $g : \mathbb{R} \to \mathbb{R}$ is monotone, then the set of discontinuities is countable.

Proof. Without loss of generality, assume g is monotone increasing (otherwise replace the arguments with $-g$). Denote the left limit of g at t as $g(t^-) = \lim_{\varepsilon \to 0+} g(t - \varepsilon)$. Similarly denote the right limit of g at t as $g(t^+) = \lim_{\varepsilon \to 0+} g(t + \varepsilon)$. The points of discontinuity are precisely the points where the left and right limits disagree; they are the set $D = \{t : g(t^-) < g(t^+)\}$. For any $s, t \in D$ with $s < t$, we have $g(s^+) \leq g(t^-)$, so $\big(g(s^-), g(s^+)\big) \cap \big(g(t^-), g(t^+)\big) = \emptyset$. Then $\big\{ \big(g(q^-), g(q^+)\big) : q \in D \big\}$ is a collection of disjoint intervals, each containing a distinct rational number (since $\mathbb{Q}$ is dense in $\mathbb{R}$). Since there are only countably many rationals, this proves the lemma. ∎

Theorem 5.4. Slutsky's Theorem: If $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{\mathbb{P}} c$ where c is a real constant, then $(X_n + Y_n) \xrightarrow{d} X + c$ and $X_n Y_n \xrightarrow{d} cX$.

Proof. Let t be any continuity point of F_{X+c} and $\varepsilon > 0$ be given. Observe that:

$$
\begin{aligned}
\mathbb{P}(X_n + Y_n \leq t) &= \mathbb{P}(X_n + Y_n \leq t, Y_n - c \geq -\varepsilon) + \mathbb{P}(X_n + Y_n \leq t, Y_n - c < -\varepsilon) \\
&\leq \mathbb{P}(X_n + Y_n \leq t, Y_n - c \geq -\varepsilon) + \mathbb{P}(Y_n - c < -\varepsilon) \\
&\leq \mathbb{P}(X_n + c \leq t + \varepsilon) + \mathbb{P}(Y_n - c < -\varepsilon)
\end{aligned}
$$

Since $Y_n \xrightarrow{\mathbb{P}} c$, the right term vanishes as n grows large. By Lemma 5.2, we know that there are only countably many discontinuity points, and so can send ε to zero in a way that F_{X+c} is continuous at the point $t + \varepsilon$. So since $X_n + c \xrightarrow{d} X + c$, $\limsup_{n \to \infty} F_{X_n + Y_n}(t) \leq F_{X+c}(t)$. The same type of argument shows the limit inferior and we reach our result. ∎

Theorem 5.5. Bounded Convergence Theorem: If X_n is a sequence of random variables converging in probability to X and there is a constant L such that $\mathbb{P}(|X_n| \leq L) = 1$ for all n, then $\lim_{n \to \infty} \mathbb{E}(X_n) = \mathbb{E}(X)$.

Proof. Let $\varepsilon > 0$ be given. Then observe

$$
\begin{aligned}
&|\mathbb{E}(X_n - X)| \\
&= \left| \mathbb{E}\big((X_n - X)\mathbb{1}_{\{|X_n - X| \leq \varepsilon\}}\big) + \mathbb{E}\big((X_n - X)\mathbb{1}_{\{|X_n - X| > \varepsilon\}}\big) \right| && \text{Exactly one indicator is zero} \\
&\leq \mathbb{E}\big(\big|(X_n - X)\mathbb{1}_{\{|X_n - X| \leq \varepsilon\}}\big|\big) + \mathbb{E}\big(\big|(X_n - X)\mathbb{1}_{\{|X_n - X| > \varepsilon\}}\big|\big) && \text{Triangle inequality for real numbers} \\
&\leq \varepsilon + 2L \cdot \mathbb{E}(\mathbb{1}_{\{|X_n - X| > \varepsilon\}}) && |X_n - X| \leq |X_n| + |X| \leq 2L \\
&\leq \varepsilon + 2L \cdot \mathbb{P}(|X_n - X| > \varepsilon) && \text{Expectation of indicator is probability of event} \\
&\leq \varepsilon && \lim_{n \to \infty} \mathbb{P}(|X_n - X| \leq \varepsilon) = 1 \implies \lim_{n \to \infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0
\end{aligned}
$$

∎

Theorem 5.6. Fatou's Lemma: If X_n is a non-negative random variable for all $n \in \mathbb{N}$, then $\liminf\limits_{n \to \infty} \mathbb{E}(X_n) \geq \mathbb{E}(\liminf\limits_{n \to \infty} X_n)$.

Proof. The limit inferior is a supremum of infimums. So $\liminf\limits_{n \to \infty} X_n = \sup\limits_{n \geq 1} \inf\limits_{m \geq n} X_m$. For notational ease, call $Y_n(\omega) = \inf\limits_{m \geq n} X_m(\omega)$. Then $X_n(\omega) \geq Y_n(\omega)$ and so $\mathbb{E}(X_n) \geq \mathbb{E}(Y_n)$ for every n; $\liminf\limits_{n \to \infty} \mathbb{E}(X_n) \geq \liminf\limits_{n \to \infty}(Y_n)$. It thus suffices to prove $\liminf\limits_{n \to \infty}(Y_n) \geq \mathbb{E}(Y)$ where $Y = \liminf\limits_{n \to \infty} X_n$.

We do so by way of truncation (i.e. reducing the non-negative case to the bounded case). For every $L \in \mathbb{R}_+$, $\liminf\limits_{n \to \infty} \mathbb{E}(Y_n) \geq \liminf\limits_{n \to \infty} \mathbb{E}(Y_n \wedge L)$. Since $(Y_n \wedge L) \xrightarrow{\mathbb{P}} (Y \wedge L)$ ($[Y_n = \inf\limits_{m \geq n} X_m] \searrow [Y = \sup\limits_{n \geq 1} \inf\limits_{m \geq n} X_m]$ since the supremum can only grow smaller as points in the sequence are removed) and since $\mathbb{P}(|Y_n \wedge L| \leq L) = 1$, we can apply the Bounded Convergence Theorem (Theorem 5.5, Page 42) to say $\liminf\limits_{n \to \infty} \mathbb{E}(Y \wedge L) = \mathbb{E}(Y \wedge L)$. Taking L to infinity and applying Lemma 3.4.1, page 25, this is precisely $\mathbb{E}(Y)$. ∎

Theorem 5.7. Monotone Convergence Theorem: If $\{X_n\}_{n \in \mathbb{N}}$ are non-negative monotonically increasing random variables such that $X_n \xrightarrow{a.s.} X$, then $\lim\limits_{n \to \infty} \mathbb{E}(X_n) = \mathbb{E}(X)$.

Proof. By Fatou's Lemma (Theorem 5.6, page 43) and the convergence of X_n, $\lim\limits_{n \to \infty} \mathbb{E}(X_n) \geq \mathbb{E}(\lim\limits_{n \to \infty} X_n) = \mathbb{E}(X)$.

On the other hand, since X_n is a monotonically increasing sequence of random variables, $X_n \leq X$ for all n and then since expectations respect dominance, $\mathbb{E}(X_n) \leq \mathbb{E}(X)$ and so $\limsup\limits_{n \to \infty} \mathbb{E}(X_n) \leq \mathbb{E}(X)$. Taken together, this proves our result. ∎

Theorem 5.8. Dominated Convergence Theorem: If X_n converges almost surely to X and if Z is an integrable random variable such that $\mathbb{P}(|X_n| \leq Z) = 1$ for all $n \in \mathbb{N}$, then $\mathbb{E}(X_n)$ converges to $\mathbb{E}(X)$

Proof. Since $X_n + Z$ is non-negative by assumption, $X_n + Z$ is a non-negative random variable that converges almost surely to $X + Z$. By Fatou's Lemma (Theorem 5.6, page 43), $\liminf\limits_{n \to \infty} \mathbb{E}(X_n + Z) \geq \mathbb{E}(\liminf\limits_{n \to \infty} X_n + Z) = \mathbb{E}(X + Z)$ and after using the linearity of expectations and canceling, we see $\liminf\limits_{n \to \infty} \mathbb{E}(X_n) \geq \mathbb{E}(X)$.

On the other hand, $-X_n + Z$ is also non-negative by assumption and so we can repeat virtually the same argument. Again by Fatou, $\liminf\limits_{n \to \infty} \mathbb{E}(-X_n + Z) \geq \mathbb{E}(\liminf\limits_{n \to \infty} -X_n + Z) = \mathbb{E}(-X + Z)$ and after using the linearity of expectations and canceling, we see $\liminf\limits_{n \to \infty} \mathbb{E}(-X_n) \geq \mathbb{E}(-X)$. This is the same as saying $-\limsup\limits_{n \to \infty} \mathbb{E}(X_n) \geq -\mathbb{E}(X)$ and equivalently $\limsup\limits_{n \to \infty} \mathbb{E}(X_n) \leq \mathbb{E}(X)$. So when seen with the above, this proves our result. ∎

Theorem 5.9. Conditions For Convergence In Probability: A sequence of random variables X_n converges in probability to the random variable X if and only if every subsequence has a further subsequence that converges almost surely.

Proof. Assume $X_n \xrightarrow{\mathbb{P}} X$. We want to show the existence of a sequence $\{n_k\}$ such that $X_{n_k} \xrightarrow{a.s.} X$ as $k \to \infty$. By the definition of convergence in $\mathbb{P}$, $\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0$ for all positive ε. So choose n_k such that $n_k > n_{k-1}$ and $\mathbb{P}(|X_{n_k} - X| \geq \frac{1}{k}) \leq \frac{1}{k^2}$. Then we have $\sum_{k=1}^{\infty} \mathbb{P}(|X_{n_k} - X| \geq \frac{1}{k}) < \infty$, and applying the first Borel-Cantelli Theorem (Theorem 7.1, Page 60), $\mathbb{P}(|X_{n_k} - X| \geq \frac{1}{k}$ i.o.$) = 0$, i.e. for all large k, $|X_{n_k} - X| \to 0$ almost surely.

Now assume any generic sequence $\left\{ X_{n_{k_l}} \right\}$ converges almost surely and let $\varepsilon > 0$ be given. We'd like to show that $\lim_{n\to\infty} \mathbb{P}(|X_n - X| \geq \varepsilon) = 0$. Call the inside of this limit p_n for notational ease. Suppose there is not convergence in probability. Then by definition, there exists a $\delta > 0$ and a sequence $\{n_k\}$ such that $p_{n_k} \geq \delta$ for all k, i.e. for all $k \in \mathbb{N}$, $P(|X_{n_k} - X| \geq \varepsilon) \geq \delta$. But this means that no subsequence can even converge in probability, never mind converge almost surely. This gives us our contradiction and we've proved our claim. ∎

Theorem 5.10. Skorohod's Representation Theorem: If $X_n \xrightarrow{d} X$, then there exists random variables Y_n and Y such that $Y_n \overset{d}{=} X_n$, $Y \overset{d}{=} X$, and $Y_n \xrightarrow{a.s.} Y$.

Proof. Let $U \sim \text{Unif}(0,1)$. Define f_{X_n} to be the quantile function (Definition 2.7, Page 15) for F_{X_n} and then the coupling we seek will be given by $Y_n = f_{X_n}(U_n)$ and $Y = f_X(U)$. We know that $F_{Y_n} = F_{X_n}$ and $F_Y = F_X$. So all that remains to be shown is $Y_n \xrightarrow{a.s.} Y$. Since the quantile function is non-decreasing, we know it's set of discontinuities, call it D, is finite (Lemma 5.3.1, Page 42). So as $\mathbb{P}(U \in D) = 0$, it suffices to show $f_{X_n}(u) \to f_X(u)$ whenever $u \notin D$.

To do so, we can check two inequalities: $\liminf_{n\to\infty} f_{X_n}(u) \geq f_X(u)$ and $\limsup_{n\to\infty} f_{X_n}(u) \leq f_X(u)$. For the first inequality, consider any $t < f_X(u)$ where F_X is continuous. Then since $t < f_X(u)$, $F_X(t) < u$, and since F_X is continuous, $f_{X_n}(t) \to F_X(t)$. Then $F_{X_n}(t) < u$ for all large enough n, and thus $t \leq F_{X_n}(u)$ for all large enough n, which proves $\liminf_{n\to\infty} f_{X_n}(u) \geq t$. By the above lemma (Lemma 5.2, Page 42), since the points of discontinuity are countable, we can take t arbitrarily close to $f(u)$ and we have our desired inequality. A near identical argument gives the second inequality. ∎

Lemma 5.10.1. Portmanteau Lemma: There are many equivalent definitions of weak convergence. Here we list some of them:

- $\mathbb{E}\big(f(X_n)\big) \to \mathbb{E}\big(f(X)\big)$ for all bounded Lipschitz f
- $\mathbb{E}\big(f(X_n)\big) \to \mathbb{E}\big(f(X)\big)$ for all bounded f such that $\mathbb{P}(X \in \{\text{discontinuity points of } f\}) = 0$
- $\limsup\limits_{n\to\infty} \mathbb{E}\big(f(X)\big) \leq \mathbb{E}\big(f(x)\big)$ for all upper-semicontinuous f that is bounded from above
- $\liminf\limits_{n\to\infty} \mathbb{E}\big(f(X)\big) \geq \mathbb{E}\big(f(x)\big)$ for all lower-semicontinuous f that is bounded from below
- $\limsup\limits_{n\to\infty} \mathbb{P}(X_n \in K) \leq \mathbb{P}(X \in K)$ for every closed $K \subset \mathbb{R}$

Theorem 5.11. Equivalent Definitions Of Weak Convergence: A sequence of random variables X_n converges weakly to X if and only if $\mathbb{E}\big(f(X_n)\big) \to \mathbb{E}\big(f(X)\big)$ for all bounded and continuous $f : \mathbb{R} \to \mathbb{R}$.

Proof. Assume $X_n \xrightarrow{d} X$, i.e. that $\lim\limits_{n\to\infty} F_{X_n}(t) = F_X(t)$ for all t where F_X is continuous. Take Y_n and Y as from Skorohod's Theorem (Theorem 5.10, Page 44). Then for all bounded and continuous $f : \mathbb{R} \to \mathbb{R}$, $f(Y_n) \xrightarrow{a.s.} f(Y)$ since $Y_n \xrightarrow{a.s.} Y$ and f is continuous. Then by the Bounded Convergence Theorem (Theorem 5.5, Page 42) $\mathbb{E}\big(f(Y_n)\big) \to \mathbb{E}\big(f(Y)\big)$, and by the equality in distribution portion of Skorohod's Theorem, $\mathbb{E}\big(f(X_n)\big) \to \mathbb{E}\big(f(X)\big)$.

Now assume $\mathbb{E}\big(f(X_n)\big) \to \mathbb{E}\big(f(X)\big)$ for all bounded and continuous $f : \mathbb{R} \to \mathbb{R}$. We want to show that if t is a continuity point, then $\big[\mathbb{P}(X_n \leq t) = \mathbb{E}(\mathbb{1}_{\{X_n \leq t\}})\big] \to \big[\mathbb{P}(X \leq t) = \mathbb{E}(\mathbb{1}_{\{X \leq t\}})\big]$ (these are the definition of CDF's). The issue is that the indicator function isn't continuous. So we aim to approximate the indicator with a sequence of continuous functions.

Given $\varepsilon > 0$, define $f_\varepsilon(x) = \begin{cases} 1, & x \leq t \\ \frac{t+\varepsilon-x}{\varepsilon}, & t < x \leq t+\varepsilon \\ 0, & t+\varepsilon \leq x \end{cases}$. As ε goes to zero, the slope of f_ε between t and $t+\varepsilon$ grows vertical; $f_\varepsilon(x)$ goes to $\mathbb{1}_{\{x \leq t\}}$. Then since $\mathbb{1}_{\{X_n \leq t\}} \leq f_\varepsilon(x)$, we have $\limsup\limits_{n\to\infty} \mathbb{E}\big(\mathbb{1}_{\{X_n \leq t\}}\big) \leq \lim\limits_{n\to\infty} \mathbb{E}\big(f_\varepsilon(X_n)\big)$ and since f_ε is bounded and continuous, $\lim\limits_{n\to\infty} \mathbb{E}\big(f_\varepsilon(X_n)\big) = \mathbb{E}\big(f_\varepsilon(X)\big)$. Taking ε to zero, we have $\limsup\limits_{n\to\infty} \mathbb{E}\big(\mathbb{1}_{\{x_n \leq t\}}\big) \leq \lim\limits_{\varepsilon \to 0} \mathbb{E}\big(f_\varepsilon(X)\big) = \mathbb{E}\big(\mathbb{1}_{\{x \leq t\}}\big)$ by the Dominated Converge Theorem (Theorem 5.8, Page 43).

For the other direction, define $g_\varepsilon(x) = \begin{cases} 1, & x \leq t-\varepsilon \\ \frac{t-x}{\varepsilon}, & t-\varepsilon < x < t \\ 0, & t \leq x \end{cases}$. For the same rationale, as ε goes to zero, $g_\varepsilon(x)$ goes to $\mathbb{1}_{\{x < t\}}$. And by the same argument for the first inequality, since $g_\varepsilon(x) \leq \mathbb{1}_{\{x \leq t\}}$, $\liminf\limits_{n\to\infty} \mathbb{E}\big(\mathbb{1}_{\{x_n \leq t\}}\big) \geq \lim\limits_{n\to\infty} \mathbb{E}\big(g_\varepsilon(x_n)\big) = \mathbb{E}\big(g_\varepsilon(x)\big)$. Then taking ε to 0, again by the Dominated Convergence Theorem, $\liminf\limits_{n\to\infty} \mathbb{E}\big(\mathbb{1}_{\{x_n \leq t\}}\big) \geq \mathbb{E}\big(\mathbb{1}_{\{x < t\}}\big)$. Then since t is a continuity point of F_X, $\mathbb{P}(X \leq t) = \mathbb{P}(X < t)$ and $\mathbb{E}\big(\mathbb{1}_{\{x \leq t\}}\big) = \mathbb{E}\big(\mathbb{1}_{\{x < t\}}\big)$. In light of the two inequalities above, we reach our conclusion, that $\mathbb{P}(X_n \leq t) \to \mathbb{P}(X \leq t)$. $\blacksquare$

Theorem 5.12. Helly's Selection Theorem ("No free lunch theorem"): Let $\{F_n\}_{n=1}^{\infty}$ be a sequence of distribution functions. Then there exists a subsequence $\{F_{n_k}\}_{k=1}^{\infty}$ and a right-continuous, non-decreasing $F : \mathbb{R} \to \mathbb{R}$ such that $F_{n_k}(t) \to F(t)$ for all continuity points t of F. If the subsequence is tight, then F is a valid CDF (so the subsequence converges weakly).

Proof. Consider any $t \in \mathbb{R}$. Since $F_n(t) \in [0,1]$ for all n, there exists a convergent subsequence $F_{n_k}(t)$ as $k \to \infty$. The issue is that this subsequence depends on t, but we want to use the same subsequence for all t. So we try the diagonalization trick. Enumerate the rationals. By the above, there exists a $\left\{ n_k^{(1)} \right\}_{k=1}^{\infty}$ such that $F_{n_k^{(1)}}(q_1)$ as $k \to \infty$. Now proceed inductively; given $(n_k^{(l)})_{k=1}^{\infty}$, choose a subsequence $(n_k^{(l+1)})_{k=1}^{\infty}$ such that $F_{n_k}^{(l+1)}(q_l + 1)$ as $k \to \infty$. Finally, set $d_k = n_k^{(k)}$, i.e. $n_1^{(1)}, n_2^{(2)}, n_3^{(3)}$, etc. Then $F_{d_k}(q_l)$ converges as $k \to \infty$ for all l since $(d_k)_{k \geq l}$ is a subsequence $(n_k^{(l)})_{k=1}^{\infty}$. Now set $\hat{F}(q) = \lim_{k \to \infty} F_{d_k}(q)$ for all $q \in \mathbb{Q}$. This defines a non-decreasing $\hat{F} : \mathbb{Q} \to [0,1]$ since F_{d_k} is non-decreasing for all k. To get the function defined on all of $\mathbb{R}$, use $F(t) = \inf\left\{ \hat{F}(q) : q \in \mathbb{Q}, q > t \right\}$. So F remains non-decreasing.

Then when $s \leq t$, $\{q \in \mathbb{Q} : q > s\} \supset \{q \in \mathbb{Q} : q > t\}$ and thus $F(s) \leq F(t)$. F is also right-continuous. Given $\varepsilon > 0$, choose rational $q > t$ such that $\hat{F}(q) \leq F(t) + \varepsilon$. Then for all $t' \in (t, q)$ we have $F(t) \leq F(t') \leq \hat{F}(q) \leq F(t) + \varepsilon$. Then we conclude that $F(t')$ monotonically decreases to $F(t)$ as t' goes to t.

Finally, we need to show that $F_{n_k}(t) \to F(t)$ if t is a continuity point. Let $\varepsilon > 0$ be given and use continuity to find $\delta > 0$ such that $|s - t| \leq \delta$ (and thus $|F(s) - F(t)| \leq \varepsilon$). Then choose rationals q', q'' such that $t - \delta < q' < t < q''$ and $\hat{F}(q'') \leq F(t) + \varepsilon$. Now we have $F(t) - \varepsilon \leq F(t - \delta) \leq \hat{F}(q') \leq \hat{F}(q'') \leq F(t) + \varepsilon$ by the definition of $F(t - \delta)$. Since $F_{d_k}(q) \to \hat{F}(q)$ for all $q \in \mathbb{Q}$, it follows that $\limsup_{k \to \infty} F_{d_k}(t) \leq \limsup_{k \to \infty} F_{d_k}(q'') = \hat{F}(q'') \leq F(t) + \varepsilon$ while $\liminf_{k \to \infty} F_{d_k}(t) \geq \limsup_{k \to \infty} F_{d_k}(q') = \hat{F}(q') \geq F(t) - \varepsilon$. Then as ε goes to 0, we conclude that $F_{d_k}(t) \to F(t)$ as $k \to \infty$.

For part 2, assuming tightness, we can find $a, b \in \mathbb{R}$ such that $F_n(a) \leq \varepsilon$ and $F_n(b) \geq 1 - \varepsilon$ for all $n \in \mathbb{N}$. Then selecting continuity points s, t of F such that $s \leq a \leq b \leq t$, we have $F(s) = \lim_{n \to \infty} F_n(s) \leq \liminf_{k \to \infty} F_{n_k}(a) \leq \varepsilon$ and $F(t) = \lim_{n \to \infty} F_n(t) \geq \limsup_{k \to \infty} F_{n_k}(b) \leq 1 - \varepsilon$. As ε was arbitrary, we conclude that $\lim_{s \to \infty} F(s) = 0$ and $\lim_{t \to \infty} F(t) = 1$. So F is a valid distribution. $\blacksquare$

5.3   Problems

Problem 5.1) Suppose that $\{X_n\}_{n=1}^{\infty}$, X, and $\{Y_n\}_{n=1}^{\infty}$ are random variables defined on the same probability space. Assume $X_n \xrightarrow{d} X$ and $Y_n \xrightarrow{d} c$ where c is a constant. Prove that $X_n + Y_n \xrightarrow{d} X + c$.

Let $\varepsilon > 0$ be given. The distribution function of a constant takes a value of 1 once the constant is reached, and 0 prior to the constant being reached. So by assumption of the proof,

$$\lim_{n\to\infty} F_{Y_n} = F_c = \begin{cases} 1, & x \geq c \\ 0, & x < c \end{cases} \text{ and thus } \lim_{n\to\infty} \mathbb{P}\left[Y_n \leq (c-\varepsilon)\right] = 0 \text{ and } \lim_{n\to\infty} \mathbb{P}\left[Y_n \leq (c+\varepsilon)\right] = 1.$$

Equivalently, $\lim_{n\to\infty} \mathbb{P}\left[Y_n > (c-\varepsilon)\right] = 1$.

Intersection with a probability one event is just the original probability, so for any $t \in \mathbb{R}$, we have $\limsup_{n\to\infty} \mathbb{P}\left[(X_n + Y_n) \leq (t+c)\right] = \limsup_{n\to\infty} \mathbb{P}\left[\{X_n + Y_n \leq t+c\} \cap \{Y_n > (c-\varepsilon)\}\right]$. Note that if $(X_n + Y_n) \leq (t+c)$ and if $(Y_n) \geq (c-\varepsilon)$, then $(X_n + c - \varepsilon) \leq (t+c)$ and so $X_n \leq t + \varepsilon$. Taken together, $\limsup_{n\to\infty} \mathbb{P}\left[(X_n + Y_n) \leq (t+c)\right] \leq \limsup_{n\to\infty} \mathbb{P}\left[X_n \leq (t+\varepsilon)\right]$.

We know $(-\infty, t+\varepsilon]$ is closed, so by the Portmanteau Lemma (Lemma 5.10.1, Page 45), $\limsup_{n\to\infty} \mathbb{P}(X_n \leq t+\varepsilon) \leq \mathbb{P}(X \leq t+\varepsilon)$. When combined with the previous inequality, we see:

$$\limsup_{n\to\infty} \mathbb{P}\left[(X_n + Y_n) \leq (t+c)\right] \leq \limsup_{n\to\infty} \mathbb{P}\left[X_n \leq (t+\varepsilon)\right] \leq \mathbb{P}(X \leq t+\varepsilon)$$

Using the same reasoning as outlined above, we also have:

$$\liminf_{n\to\infty} \mathbb{P}\left[(X_n + Y_n) \leq (t+c)\right] \geq \liminf_{n\to\infty} \mathbb{P}\left[X_n \leq (t-\varepsilon)\right] \geq \mathbb{P}(X \leq t-\varepsilon)$$

When t is a continuity point of F_X and we take ε to 0, we squeeze the lim sup/inf like so:

$$\mathbb{P}(X \leq t) = \mathbb{P}(X \leq t-\varepsilon) \leq \liminf_{n\to\infty} \mathbb{P}\left[(X_n + Y_n) \leq (t+c)\right]$$
$$\leq \limsup_{n\to\infty} \mathbb{P}\left[(X_n + Y_n) \leq (t+c)\right] \leq \mathbb{P}(X \leq t+\varepsilon) = \mathbb{P}(X \leq t)$$

Since the limit superior and limit inferior agree, we can write:

$$\lim_{n\to\infty} \mathbb{P}\left[(X_n + Y_n) \leq (t+c)\right] = \mathbb{P}(X \leq t) = \mathbb{P}(X + c \leq t+c)$$

As $t + c$ is a continuity point of $F_{X+c}(x)$ if and only if t is a continuity point of $F_X(x)$, this proves our result.

Problem 5.2) Assume $\{X_n\}_{n=1}^{\infty}$ is a weakly convergent sequence of random variables. Show that this sequence is tight.

A sequence of random variables are tight if for all $\varepsilon > 0$, there exist $a, b \in \mathbb{R}$ such that $F_i(a) \leq \varepsilon$ and $F_i(b) \geq 1 - \varepsilon$ where F_i are distribution functions.

So let $\varepsilon > 0$ be given and choose continuity points a^* and b^* such that $F(a^*) \leq \frac{\varepsilon}{2}$ and $F(b^*) \geq (1 - \frac{\varepsilon}{2})$. By assumption, $F_n(a^*) \to F(a^*)$ and $F_n(b^*) \to F(b^*)$. Then by definition, there exists an $N \in \mathbb{N}$ such that for all $n > N$, $F_n(a^*) \leq \varepsilon$ and $F_n(b^*) \geq (1 - \varepsilon)$.

For any $n < N$, we can choose values a_n, b_n such that $F_n(a_n) \leq \varepsilon$ and $F_n(b_n) \geq (1 - \varepsilon)$. Call $a = \min\{a_1, \ldots, a_N, a^*\}$ and $b = \min\{b_1, \ldots, b_N, b^*\}$. Then for all $n \geq 1$, $F_n(a) \leq \varepsilon$ and $F_n(b) \geq (1 - \varepsilon)$.

Problem 5.3) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space such that $\mathbb{P}(A) \in \{0, 1\}$ for every $A \in \mathcal{F}$. Prove that every random variable on this space is an almost sure constant. That is, show that for any measurable function $X : \Omega \to \mathbb{R}$, there exists a $a \in \mathbb{R}$ such that $\mathbb{P}(X = a) = 1$.

Consider the set $(-\infty, k)$ in the Borel sigma-algebra of $\mathbb{R}$. Since X is measurable, $X^{-1}\big((-\infty, k)\big) \in \mathcal{F}$. By assumption of proof, we can say

$$\mathbb{P}\Big(X^{-1}\big((-\infty, k)\big)\Big) = \mathbb{P}\big(\{\omega \in \Omega : X(\omega) \leq k\}\big) = \mathbb{P}(X \leq k) = F_X(k) \in \{0, 1\}$$

where F_X is understand to mean the cumulative distribution function of X. Since $\lim_{n \to -\infty} F_X(n) = 0$, $\lim_{n \to \infty} F_X(n) = 1$, and F_x is monotone, there exists values $a, b \in \mathbb{R}$ such that $a = \sup\{x \in \mathbb{R} : F_X(x) = 0\}$ and $b = \inf\{x \in \mathbb{R} : F_X(x) = 1\}$.

If $a < b$, then there would exist a $c \in \mathbb{R}$ such that $a < c < b$. Then by the construction of a and b and by the monotonicity of F_X, $0 = F_x(a) < F_X(c) < F_X(b) = 1$, a contradiction to $F_X \in \{0, 1\}$. So $a = b$.

Now let $\varepsilon > 0$ be given. We have:

$$
\begin{aligned}
&\mathbb{P}\big((a - \varepsilon) \leq X \leq (a + \varepsilon)\big) && \text{Would like this to be 1} \\
=&\mathbb{P}\big(X \leq (a + \varepsilon)\big) - \mathbb{P}\big(X \leq (a - \varepsilon)\big) && \text{Breaking up inequality} \\
=&F_X(b + \varepsilon) - F_X(a - \varepsilon) && \text{Definition of CDF and } a = b \\
=&1 - 0 = 1 && \text{How } a \text{ and } b \text{ were constructed}
\end{aligned}
$$

Since this holds for all $\varepsilon > 0$, we must have $\mathbb{P}(X = a) = 1$ as desired.

Problem 5.4) Show that if $X_1, X_2, \ldots$ are non-negative, $\mathbb{E}\left(\sum\limits_{i=1}^{\infty} X_i\right) = \sum\limits_{i=1}^{\infty} \mathbb{E}(X_i)$.

For all finite n, $\mathbb{E}(X_1 + X_1 + \cdots + X_n) = \mathbb{E}(X_1) + \mathbb{E}(X_2) + \cdots + \mathbb{E}(X_n)$ by linearity of expectations. Define a sequence of random variables by $Y_1 = X_1$, $Y_2 = Y_1 + X_2$, $Y_3 = Y_2 + X_3$, etc. Then $\mathbb{P}\left(\lim\limits_{n \to \infty} Y_n = \sum\limits_{i=1}^{\infty} X_i\right) = 1$. Since each X_i is non-negative, $Y_i \leq Y_{i+1}$ for every i; the sequence is monotone increasing. So we have a non-negative sequence of random variables that converges almost surely and we can use the monotone convergence theorem (Theorem 5.7, Page 43) to say $\lim\limits_{n \to \infty} \mathbb{E}(Y_n) = \mathbb{E}\left(\sum\limits_{i=1}^{\infty} X_i\right)$. This proves the claim.

Problem 5.5) Let X be a nonnegative random variable. Show that there is a sequence of nonnegative simple random variables $\{X_n\}_{n \geq 1}$ such that $X_n \nearrow X$ as $n \to \infty$.

Consider the sequence of random variables given by $X_n(\omega) = \min\left\{\frac{1}{2^n}\lfloor 2^n \cdot X(\omega)\rfloor, n\right\}$.

For every n in the sequence and any $\omega \in \Omega$, $X_n(\omega) = 0$ when $2^n \cdot X(\omega) < 1$. Similarly, whenever $X(\omega) > n$, $X_n(\omega) = n$. Since each X_n is non-negative, the sequence is bounded by these values, $0 \leq X_n \leq n$. Then by inspection, the floor function can take on values $0, 1, \ldots, n \cdot 2^n$ to stay within the bounds. In other words, $X_n(\omega) \in \left\{0, \frac{1}{2^n}, \frac{2}{2^n}, \ldots, n\right\}$; each random variable in the sequence is simple since it can only take on finitely many values.

Consider a random variable in this sequence X and an element $\omega \in \Omega$. Observe that $2\lfloor 2^n \cdot X(\omega)\rfloor \leq \lfloor 2 \cdot 2^n \cdot X(\omega)\rfloor = \lfloor 2^{n+1} \cdot X(\omega)\rfloor$ (the inequality is only an equality when $X(\omega)$ is an integer). Then dividing both sides of the inequality by 2^{n+1}, we see $\frac{\lfloor 2^n \cdot X(\omega)\rfloor}{2^n} \leq \frac{\lfloor 2^{n+1} \cdot X(\omega)\rfloor}{2^{n+1}}$. Since the right-most part of the minimum is $n < n+1$, this proves the sequence is monotone increasing.

We prove the sequence of random variables converges to X with the squeeze theorem. At any n, $\left[X_n(\omega) = \min\left\{\frac{1}{2^n}\lfloor 2^n \cdot X(\omega)\rfloor, n\right\}\right] \leq \min\left\{\lfloor\frac{1}{2^n} \cdot 2^n \cdot X(\omega)\rfloor, n\right\} \leq \min\{X(\omega), n\}$. On the other hand $\left[X_n(\omega) = \min\left\{\frac{1}{2^n}\lfloor 2^n \cdot X(\omega)\rfloor, n\right\}\right] \geq \min\left\{X(\omega) - \frac{1}{2^n}, n\right\}$. Then taking both limits we see $\lim\limits_{n \to \infty} \min\{X(\omega), n\} = X(\omega) = \lim\limits_{n \to \infty} X(\omega) - \frac{1}{2^n} = \lim\limits_{n \to \infty} \min\left\{X(\omega) - \frac{1}{2^n}, n\right\}$.

Problem 5.6) Assume $X_n \xrightarrow{\mathbb{P}} X$ and that there is an integrable random variable Y where $X_n \leq Y$ for each n. Show that $\lim\limits_{n \to \infty} \mathbb{E}(X_n) = \mathbb{E}(X)$.

Let $\{n_k\}_{k=1}^{\infty}$ be a sequence of positive integers. Since $X_n \xrightarrow{\mathbb{P}} X$, we must have $X_{n_k} \xrightarrow{\mathbb{P}} X$ as well. Then by the properties of convergence in probability, there is a further subsequence $X_{n_{k_l}} \xrightarrow{a.s.} X$. Since each $X_{n_{k_l}}$ is bounded by Y, by the Bounded Convergence Theorem (Theorem 5.5, Page 42), $\lim\limits_{n \to \infty} \mathbb{E}(X_{n_{k_l}}) = \mathbb{E}(X)$. Then because every subsequence has a further convergent subsequence, the full sequence converges.

Problem 5.7) Let $\{X_n\}_{n\geq 1}$ be a sequence of random variables such that $X_n \nearrow X$ almost surely and $\mathbb{E}(X_1^-) < \infty$. Prove that $\mathbb{E}(X_n) \nearrow \mathbb{E}(X)$.

Consider the sequence of random variables given by $Y_n = X_n + X_1^-$. By the monotonicity of the X_i's, $\left[Y_n = X_n + X_1^-\right] \leq \left[Y_{n+1} = X_{n+1} + X_1^-\right]$; $\{Y_n\}_{n\in\mathbb{N}}$ is monotone increasing.

Again by the monotonicity of the X_i's, $\left[Y_n = X_n + X_1^-\right] \geq X_1 + X_1^- = X_1^+ \geq 0$; $\{Y_n\}_{n\in\mathbb{N}}$ is non-negative.

Since $X_n \nearrow X$ almost surely, $\left[Y_n = X_n + X_1^-\right] \nearrow \left[X + X_1^-\right]$ almost surely. So $\{Y_n\}_{n\in\mathbb{N}}$ is a monotonically increasing non-negative sequence that converges almost surely. We can then apply the monotone convergence theorem and linearity of expectations to reach our result.

$$\mathbb{P}\left(\lim_{n\to\infty} \mathbb{E}(Y_n) = \mathbb{E}(X + X_1^-)\right) = 1 \qquad \text{Monotone Convergence Theorem}$$

$$\mathbb{P}\left(\lim_{n\to\infty} \mathbb{E}(X_n + X_1^-) = \mathbb{E}(X + X_1^-)\right) = 1 \qquad \text{How } Y_n \text{ was defined}$$

$$\mathbb{P}\left(\lim_{n\to\infty} \mathbb{E}(X_n) + \mathbb{E}(X_1^-) = \mathbb{E}(X) + \mathbb{E}(X_1^-)\right) = 1 \quad \text{Linearity of expectations since } \mathbb{E}(X_1^-) < \infty$$

$$\mathbb{P}\left(\lim_{n\to\infty} \mathbb{E}(X_n) = \mathbb{E}(X)\right) = 1 \qquad \text{Desired result}$$

Problem 5.8) Let X be an integrable random variable on the measurable space $(\Omega, \mathcal{F})$. Show that for any $\varepsilon > 0$, there exists a $\delta > 0$ such that the following implication is true: $A \in \mathcal{F}, P(A) \leq \delta \implies \mathbb{E}(|X|\mathbb{1}_A) \leq \varepsilon$.

Let $\varepsilon > 0$ be given. First note that for any $\omega \in \Omega$ and any $n \in \mathbb{N}$, the random variable

$$\mathbb{1}_{\{|X(\omega)|>n\}}(\omega) = \begin{cases} 1, & |X(\omega)| > n \\ 0, & |X(\omega)| \leq n \end{cases}$$ goes to zero almost surely as n grows large. Further since

$|X|\mathbb{1}_{\{|X|>n\}} \leq |X|$ and $|X|$ is integrable by the assumption of the proof, we can use the Dominated Convergence Theorem (Theorem 5.8, Page 43) to show $\mathbb{E}\left(|X|\mathbb{1}_{\{|X|>n\}}\right)$ goes to zero in the limit (and in particular that we can choose an n such that $\mathbb{E}\left(|X|\mathbb{1}_{\{|X|>n\}}\right) \leq \frac{\varepsilon}{2}$).

With $\varepsilon > 0$ and $n \in \mathbb{N}$ in mind, consider $\delta = \frac{\varepsilon}{2n}$. Then observe

$$
\begin{aligned}
\mathbb{E}\left(|X|\mathbb{1}_A\right) &= \mathbb{E}\left(|X|\mathbb{1}_{\{A\cap|X|\leq n\}} + |X|\mathbb{1}_{\{A\cap|X|>n\}}\right) \\
&= \mathbb{E}\left(|X|\mathbb{1}_{\{A\cap|X|\leq n\}}\right) + \mathbb{E}\left(|X|\mathbb{1}_{\{A\cap|X|>n\}}\right) && \text{Linearity of expectations} \\
&\leq \mathbb{E}(n\mathbb{1}_A) + \mathbb{E}\left(|X|\mathbb{1}_{\{A\cap|X|>n\}}\right) && |X| \leq n \text{ if indicator isn't zero} \\
&= n\mathbb{P}(A) + \mathbb{E}\left(|X|\mathbb{1}_{\{A\cap|X|>n\}}\right) && \text{Expectation of indicator is probability of event} \\
&= n\delta + \frac{\varepsilon}{2} && \text{By assumption and how we chose } n \\
&= n \cdot \frac{\varepsilon}{2n} + \frac{\varepsilon}{2} && \text{Desired result}
\end{aligned}
$$

6   Independence

6.1   Definition

Definition 6.1. Sigma-Algebra Generated By Random Variables: The sigma-algebra generated by a sequence of random variables $\{X_i\}_{i \in I}$ is the smallest sigma-algebra containing $\sigma(X_i)$ for all i; $\sigma(\{X_i\}_{i \in I}) = \sigma\left(\bigcup_{i \in I} \sigma(X_i)\right)$. Here, $\sigma(X) = \{\{\omega \in \Omega : X(\omega) \in B\} : B \in \mathbb{B}(\mathbb{R})\}$.

Example 6.1: Consider a sample space whose sets indicate the outcome of three fair coin flips (e.g. A_{HTT} denotes a head followed by two tails). Further consider the random variable $S_2(\omega)$ which returns the number of heads that come up after the first two flips. It is an easy exercise to find the pre-images that belong to each borel set. For example, $\{S_2 \in \{2\}\} = \{A_{HHH}, A_{HHT}\} = A_{HH}$ and $\{S_2 \in [1, 2]\} = \{A_{HH} \cup A_{HT} \cup A_{TH}\} = A_{TT}^c$.

In total, $\sigma(S_2) = \left\{\begin{array}{c} \emptyset, \Omega, A_{HH}, A_{TT}, A_{HT} \cup A_{TH}, \\ A_{HH}^c, A_{TT}^c, A_{HH} \cup A_{TT} \end{array}\right\}$, which is a substantially different

sigma-algebra than $\mathcal{F}_2 = \left\{\begin{array}{c} \emptyset, \Omega, A_H, A_T, \\ A_{HH}, A_{HT}, A_{TH}, A_{TT}, \\ A_{HH}^C, A_{HT}^C, A_{TH}^C, A_{TT}^C, \\ A_{HH} \cup A_{TT}, A_{HH} \cup A_{TH}, A_{HT} \cup A_{TT}, A_{HT} \cup A_{TH} \end{array}\right\}$.

For example, $A_{HT} \in \mathcal{F}_2$ but $A_{HT} \notin \sigma(S_2)$. This is because only knowing the value of S_2 (e.g. that $S_2(\omega) = 1$) does not allow one to distinguish if the initial flip was a head or tail (just that there was one total head in the first two flips). Since $\mathcal{F}_2$ has enough information to determine the value of S_2, we say that S_2 is $\boldsymbol{\mathcal{F}_2}$**-measurable** (Definition 2.1, Page 12).

Definition 6.2. Tail $\boldsymbol{\sigma}$-Algebra: Where $\{X_i\}_{i \in I}$ is a sequence of random variables, the tail sigma algebra is denoted $\mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(X_{n+1}, X_{n+2}, \dots)$. The idea is that the Tail σ-algebra is the collection of events whose occurrence is unaffected when finitely many of the random variables are changed.

Example 6.2: The set $A = \left\{\lim_{i \to \infty} X_i \text{ Exists}\right\}$ is in the tail sigma-algebra. Intuitively, this is because removing any finite number of elements does not change the limit. More formally, we recall from real analysis that a limit exists if and only if it is Cauchy. The definition of Cauchy convergence is that for any $\varepsilon > 0$, there exists a $N \in \mathbb{N}$ such that whenever $i, j > N$, it must be the case that $|X_i - X_j| < \varepsilon$. Since this holds *for all* $\varepsilon > 0$, we can pick a particular element $k \geq 1$, and get the result that $|X_i - X_j| \leq \frac{1}{k}$. Using the usual method of converting qualifiers to unions (there exists) and intersections (for all), A can be expressed as $\bigcap_{k \geq 1} \bigcup_{N \geq 1} \bigcap_{i \geq N} \bigcap_{j \geq N} \{|X_i - X_j| \leq \frac{1}{k}\}$.

Non-example 6.1: The set $C = \left\{\sup_{i \geq 1} X_i \geq 5\right\}$ is not in the tail sigma-algebra. For example, the first random variable could be doing all the lifting (e.g. is a constant 6), and removing it would change the supremum of the sequence (if all the other random variables, were, e.g. a constant 4).

Definition 6.3. Independence (Of Events): A finite set of events $A_1, A_2, \ldots, A_n$ is **mutually independent** if for all $I \subseteq \{1, \ldots, n\}$ we have $\mathbb{P}\left(\bigcap_{i \in I} A_i\right) = \prod_{i \in I} \mathbb{P}(A_i)$. We say $A_1, A_2, \ldots, A_n$ are **pairwise independent** if for all $i \neq j$, $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$. Note that mutual independence implies pairwise independence, but not vise-versa. Infinite collection of events are independent when any finite subset of the events are independent.

Example 6.3: Take $\Omega = \{1, 2, 3, 4\}$, $\mathcal{F} = 2^\Omega$, and $\mathbb{P}(A) = |A|$. Then consider events $A = \{1, 4\}$, $B = \{2, 4\}$, and $C = \{3, 4\}$. These are pairwise independent but not mutually independent since $\mathbb{P}(A \cap B) = \mathbb{P}(\{4\}) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(A) \cdot \mathbb{P}(B)$ (the same results hold for the other two intersections), but $\mathbb{P}(A \cap B \cap C) = \mathbb{P}(\{4\}) = \frac{1}{4} \neq \frac{1}{8} = \frac{1}{2} \cdot \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(A) \cdot \mathbb{P}(B) \cdot \mathbb{P}(C)$.

Non-example 6.2: Independence isn't transitive. Take $\Omega = \{1, 2, \ldots, 20\}$, $\mathcal{F} = 2^\Omega$, and $\mathbb{P}(A) = |A|$. Then consider events $A = \{1, 2, \ldots, 10\}$, $B = \{1, 3, 12, 13\}$, and $C = \{3, 4, \ldots, 12\}$. We see $A \cap B \cap C = \{3\}$, so $\mathbb{P}(A \cap B \cap C) = \frac{1}{20} = \frac{1}{2} \cdot \frac{1}{5} \cdot \frac{1}{2} = \mathbb{P}(A)\mathbb{P}(B)\mathbb{P}(C)$. We also see $A \cap B = \{1, 3\}$ and $B \cap C = \{3, 12\}$, so $\mathbb{P}(A \cap B) = \frac{1}{10} = \frac{1}{2} \cdot \frac{1}{5} = \mathbb{P}(A)\mathbb{P}(B)$ and likewise $\mathbb{P}(B \cap C) = \frac{1}{10} = \frac{1}{5} \cdot \frac{1}{2} = \mathbb{P}(B)\mathbb{P}(C)$. But $A \cap C = \{3, 4, \ldots, 10\}$ so $\mathbb{P}(A \cap C) = \frac{2}{5} \neq \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(A)\mathbb{P}(C)$.

Definition 6.4. Independence (Of Sigma Algebras): A finite collection of sigma-algebras $\{\mathcal{F}_i\}_{i \in I}$ is independent if for every $A_i \in \mathcal{F}_i$, $\{A_i\}_{i \in I}$ is independent. Note that this specifically is not saying anything about events within any one sigma-algebra (i.e the events within a sigma-algebra may not be independent, see example), but rather is saying that selecting one event from each sigma-algebra results in independence.

Example 6.4: Take $(\Omega, \mathcal{F}, \mathbb{P})$ where $\Omega = \{1, 2, 3, 4\}$, $\mathcal{F} = 2^\Omega$, and $\mathbb{P}(A) = |A|$. Consider the sigma-algebras $\mathcal{F}_1 = \{\emptyset, \Omega, \{1, 2\}, \{3, 4\}\}$ and $\mathcal{F}_2 = \{\emptyset, \Omega, \{1, 3\}, \{2, 4\}\}$. We see that the two sigma-algebras are independent (and may write $\mathcal{F}_1 \perp \mathcal{F}_2$) since the two non-trivial sets in $\mathcal{F}_1$ share precisely one element with the two non-trivial sets in $\mathcal{F}_2$. For example, $\mathbb{P}(\{1, 2\} \cap \{1, 3\}) = \mathbb{P}(\{1\}) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(\{1, 2\}) \cdot \mathbb{P}(\{1, 3\})$ and $\mathbb{P}(\{1, 2\} \cap \{2, 4\}) = \mathbb{P}(\{2\}) = \frac{1}{4} = \frac{1}{2} \cdot \frac{1}{2} = \mathbb{P}(\{1, 2\}) \cdot \mathbb{P}(\{2, 4\})$.

Non-example 6.3: Take $(\Omega, \mathcal{F}, \mathbb{P})$ where $\Omega = \{1, 2, 3, 4\}$, $\mathcal{F} = 2^\Omega$, and $\mathbb{P}(A) = |A|$. Consider the sigma-algebras $\mathcal{F}_1 = \{\emptyset, \Omega, \{1, 2\}, \{3, 4\}\}$ and $\mathcal{F}_2 = \{\emptyset, \Omega, \{1, 2, 3\}, \{4\}\}$. These two sigma-algebras aren't independent. For example, $\mathbb{P}(\{1, 2\} \cap \{1, 2, 3\}) = \mathbb{P}(\{1, 2\}) = \frac{1}{2} \neq \frac{1}{2} \cdot \frac{3}{4} = \mathbb{P}(\{1, 2\}) \cdot \mathbb{P}(\{1, 2, 3\})$.

Definition 6.5. Independence (Of Random Variables): A finite collection of random variables $\{X_i\}_{i \in I}$ is independent if $\{\sigma(X_i)\}_{i \in I}$ (Definition 2.4, Page 13) is independent. For two random variables, this is equivalent to checking that $\mathbb{P}(X \leq t_1, Y \leq t_2) = F_X(t_1)F_Y(t_2)$.

Non-example 6.4: If S_2 denotes the number of heads in the first 2 tosses of a fair coin and S_1 denotes the number of heads in the first flip of a fair coin, then S_2 and S_1 aren't independent. Informally, knowing the value of S_1 influences the knowledge of S_2 (for example, if $S_1 = 0$, $S_2 \neq 2$). More formally take $A_H \in \sigma(S_1)$ and $A_{HH} \in \sigma(S_2)$. Then $\mathbb{P}(A_H \cap A_{HH}) = \mathbb{P}(A_{HH}) = \frac{1}{4} \neq \frac{1}{8} = \frac{1}{2} \cdot \frac{1}{4} = \mathbb{P}(A_H) \cdot \mathbb{P}(A_{HH})$.

6.2   Theorems And Examples

Theorem 6.1. Information Is Lost Through Composition: Where $f : \mathbb{R} \to \mathbb{R}$ is a measurable function, X is a random variable, and $\sigma(X)$ is the sigma-algebra generated by the random variable, $\sigma\big(f(X)\big) \subseteq \sigma(X)$. In particular, if X and Y are independent, then $f(X)$ and $f(Y)$ are independent.

Proof. Take any $S \in \sigma(f(X))$. Then by definition, there exists a $B \in \mathbb{B}(\mathbb{R})$ such that $S = (f \circ X)^{-1}(B) = X^{-1}(f^{-1}(B))$. Since f is measurable by assumption, $f^{-1}(B) \in \mathbb{B}(\mathbb{R})$ (the sigma-algebra associated with the domain of f). So, there is a $D \in \mathbb{B}(\mathbb{R})$ (namely $f^{-1}(B)$) such that $S = X^{-1}(D)$; $S \in \sigma(X)$. ∎

Theorem 6.2. Equivalent Definitions of Independent Random Variables: If X and Y are independent random variables whose expectations are defined, then $\mathbb{E}(XY) = \mathbb{E}(X)\mathbb{E}(Y)$. This result extends to multiple independent random variables, but the converse is not true (see Example 3.2, Page 21). Additionally, we must have the joint distribution, joint MGF, and joint density factor as well.

Proof. First, we assume X and Y are simple, i.e. $X(\omega) \in \{x_1, \ldots, x_n\}$ and $Y(\omega) \in \{y_1, \ldots, y_m\}$. Define the events $A_i = X^{-1}(\{x_i\})$ and $B_j = Y^{-1}(\{y_j\})$ for all $1 \le i \le n$ and $1 \le j \le m$. Then clearly $X(\omega) = \sum_{i=1}^{n} x_i \mathbb{1}_{\{A_i\}}(\omega)$ and $Y(\omega) = \sum_{j=1}^{m} y_j \mathbb{1}_{\{B_i\}}(\omega)$. We then have:

$$
\begin{aligned}
\mathbb{E}(XY) &= \mathbb{E}\left(\left(\sum_{i=1}^{n} x_i \mathbb{1}_{\{A_i\}}\right)\left(\sum_{j=1}^{m} y_j \mathbb{1}_{\{B_i\}}\right)\right) & \\[2mm]
&= \mathbb{E}\left(\sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j \mathbb{1}_{\{A_i\}}\mathbb{1}_{\{B_i\}}\right) & \text{Grouping} \\[2mm]
&= \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j \mathbb{E}\left(\mathbb{1}_{\{A_i\}}\mathbb{1}_{\{B_i\}}\right) & \text{Linearity} \\[2mm]
&= \sum_{i=1}^{n}\sum_{j=1}^{m} x_i y_j \mathbb{E}\left(\mathbb{1}_{\{A_i\}}\right)\mathbb{E}\left(\mathbb{1}_{\{B_i\}}\right) & A_i \perp B_j \\[2mm]
&= \sum_{i=1}^{n} x_i \mathbb{E}\left(\mathbb{1}_{\{A_i\}}\right) \sum_{j=1}^{m} y_j \mathbb{E}\left(\mathbb{1}_{\{B_i\}}\right) & \text{Grouping} \\[2mm]
&= \mathbb{E}(X)\mathbb{E}(Y) &
\end{aligned}
$$

The third step follows since X and Y are independent, and so every A_i and B_j are as well. It is easy to see that indicator functions of independent events are independent, since $\mathbb{E}(\mathbb{1}_{\{A_i\}}\mathbb{1}_{\{B_j\}}) = \mathbb{E}\left(\mathbb{1}_{\{A_i \cap B_j\}}\right) = \mathbb{P}(A_i \cap B_j) = \mathbb{P}(A_i)\mathbb{P}(B_j) = \mathbb{E}(\mathbb{1}_{\{A_i\}})\mathbb{E}(\mathbb{1}_{\{B_j\}})$.

Next, we assume X and Y are non-negative. Consider the function $f_n(t) = \min\left\{\frac{1}{2^n}\lfloor 2^n t\rfloor, n\right\}$ and call $X_n = f_n(X)$ and $Y_n = f_n(Y)$. Since f_n is measurable, $\sigma(X_n) = \sigma(f_n(X)) \subseteq \sigma(X)$ and $\sigma(Y_n) = \sigma(f_n(Y)) \subseteq \sigma(Y)$ (Theorem 6.1, Page 53). Due to the independence of X and Y, this means that X_n and Y_n are independent for every n.

Since X_n and Y_n are simple random variables (they can only take on $n2^n$ different values due to the bound of n and the discretization of size $\frac{1}{2^n}$) that are independent, we can use the above argument to see $\mathbb{E}(X_n Y_n) = \mathbb{E}(X_n)\mathbb{E}(Y_n)$. Further, since X_n and Y_n are increasingly better approximations of X and Y (the floor function ensures that for every ω, $X(\omega)$ is no more than $\frac{1}{2^n}$ larger than $X_n(\omega)$), $0 \le X_n \nearrow X$ and $0 \le Y_n \nearrow Y$ (as n grows large, the difference $X(\omega) - X_n(\omega) \le \frac{1}{2^n}$ becomes arbitrarily small). So $0 \le X_n Y_n \nearrow XY$, and applying the Monotone Convergence Theorem (Theorem 5.7, Page 43), we have $\lim_{n\to\infty} \mathbb{E}(X_n Y_n) = E(XY)$. Applying the same theorem to X_n and Y_n individually, we have
$$\lim_{n\to\infty} \mathbb{E}(X_n Y_n) = \lim_{n\to\infty} \mathbb{E}(X_n)\mathbb{E}(Y_n) = \lim_{n\to\infty} \mathbb{E}(X_n) \lim_{n\to\infty} \mathbb{E}(Y_n) = \mathbb{E}(X)\mathbb{E}(Y).$$

Finally, we assume X and Y are integrable. Each of $X^+ = \min\{X, 0\}$, $X^- = \min\{-X, 0\}$, and $Y^\pm$ is non-negative and a function of X and Y; we can again apply Theorem 6.1 to say every combination of $X^\pm$ is independent of every combination of $Y^\pm$. Then observe:

$$
\begin{aligned}
\mathbb{E}(XY) &= \mathbb{E}\left(\left(X^+ - X^-\right)\left(Y^+ - Y^-\right)\right) &&\text{Definition} \\
&= \mathbb{E}\left(X^+ Y^+ - X^+ Y^- - X^- Y^+ + X^- Y^-\right) &&\text{Grouping} \\
&= \mathbb{E}\left(X^+ Y^+\right) - \mathbb{E}\left(X^+ Y^-\right) - \mathbb{E}\left(X^- Y^+\right) + \mathbb{E}\left(X^- Y^-\right) &&\text{Linearity} \\
&= \mathbb{E}\left(X^+\right)\mathbb{E}\left(Y^+\right) - \mathbb{E}\left(X^+\right)\mathbb{E}\left(Y^-\right) - \mathbb{E}\left(X^-\right)\mathbb{E}\left(Y^+\right) + \mathbb{E}\left(X^-\right)\mathbb{E}\left(Y^-\right) &&\text{Independence} \\
&= \mathbb{E}(X^+)\left[\mathbb{E}(Y^+) - E(Y^-)\right] - \mathbb{E}(X^-)\left[\mathbb{E}(Y^+) - E(Y^-)\right] &&\text{Grouping} \\
&= \left[\mathbb{E}(X^+) - \mathbb{E}(X^-)\right]\left[\mathbb{E}(Y^+) - E(Y^-)\right] = \mathbb{E}(X)\mathbb{E}(Y) &&\text{Grouping}
\end{aligned}
$$

■

Lemma 6.2.1. Let $\{X_i\}_{i\in I}$ be a collection of independent random variables and define $\mathcal{P} = \left\{\bigcap_{i\in I} A_i : A_i \in \sigma(X_i)\right\}$. Then $\mathcal{P}$ is a pi-system.

Proof. If $A, B \in \mathcal{P}$, then $A \cap B = \left(\bigcap_{i\in I_1} A_i\right) \cap \left(\bigcap_{i\in I_2} B_i\right) = \bigcap_{i\in I}(A_i \cap B_i)$. ■

Lemma 6.2.2. Let $\{X_i\}_{i\in I}$ be a collection of independent random variables and define $\mathcal{P} = \left\{\bigcap_{i\in I} A_i : A_i \in \sigma(X_i)\right\}$. Then $\sigma(\mathcal{P}) = \sigma\left(\{X_i\}_{i\in I}\right)$, call it $\mathcal{G}$.

Proof. Consider any $A \in \mathcal{P}$. By definition, $A = \bigcap_{i\in I} A_i$ where each $A_i \in \sigma(X_i)$ (and thus also an element of $\mathcal{G}$). Since sigma-algebras are closed under intersection, we have $A \in \mathcal{G}$ and thus $\mathcal{P} \subseteq \mathcal{G}$. Since $\sigma(\mathcal{P})$ is the smallest sigma-algebra containing $\mathcal{P}$, and since $\mathcal{G}$ is a sigma-algebra, we therefore have $\sigma(\mathcal{P}) \subseteq \mathcal{G}$. Now consider any fixed $i \in I$. If $B \in \sigma(X_i)$, then $B \in \mathcal{P} \subseteq \sigma(\mathcal{P})$ (to see this, just take $A_j = \Omega$ for every $j \ne i$ in the intersection; $\bigcap_{i\in I} A_i = B \cap \bigcap_{j\ne i, j\in I} A_j$). Since this holds for every $i \in I$ and since sigma-algebras are closed under unions, we have $\bigcup_{i\in I} \sigma(X_i) \subseteq \sigma(\mathcal{P})$. Since $\mathcal{G} = \sigma\left(\bigcup_{i\in I} \sigma(X_i)\right)$ is the smallest sigma-algebra containing $\bigcup_{i\in I} \sigma(X_i)$, and since $\sigma(\mathcal{P})$ is a sigma-algebra, we have $\mathcal{G} \subseteq \sigma(\mathcal{P})$ and have proved $\sigma(\mathcal{P}) = \mathcal{G}$. ■

Lemma 6.2.3. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Suppose $\mathcal{P}_1, \ldots, \mathcal{P}_n \subseteq \mathcal{F}$ are π-systems such that for any $A_1 \in \mathcal{P}_1, \ldots, A_n \in \mathcal{P}_n$, the collection $\{A_i\}_{i=1}^{\infty}$ is independent. Prove that $\sigma(\mathcal{P}_1), \ldots, \sigma(\mathcal{P}_n)$ are independent.

Proof. Fix any $A_2 \in \mathcal{P}_2, \ldots, A_n \in \mathcal{P}_n$. Define $\mathcal{L} = \{B \in \mathcal{F} : \{B, A_2, \ldots, A_n\}$ is independent$\}$. If we can show $\mathcal{L}$ is a lambda-system, then since $\mathcal{P}_1$ is a pi-system contained in $\mathcal{L}$ by assumption, $\sigma(\mathcal{P}_1) \subset \mathcal{L}$ by the Pi-Lambda Theorem (Theorem 3.6, Page 27). Then we would have $A_i, A_2, \ldots, A_n$ are independent for any $A_i \in \sigma(\mathcal{P}_1)$. Since $\sigma(\mathcal{P}_1)$ is a pi-system, we can repeat this argument replacing $\mathcal{P}_2$ with $\sigma(\mathcal{P}_2)$, $\mathcal{P}_3$ with $\sigma(\mathcal{P}_3)$, etc. So all we need to do is show that $\mathcal{L}$ is a lambda-system.

Let $B \in \mathcal{L}$ and $I \subseteq \{2, \ldots, n\}$ be given. To show $\mathcal{L}$ is closed under compliment, we need to show that $P\left(B^c \cap \bigcap_{i \in I} A_i\right) = P(B^c) \prod_{i \in I} P(A_i)$. Note that we check I for *any* subset of $\{2, \ldots, n\}$ instead of just the entirety of $\{2, \ldots, n\}$ because we need to show it is mutually independent. Then observe:

$$\mathbb{P}\left(B^c \cap \bigcap_{i \in I} A_i\right) = \mathbb{P}\left(\bigcap_{i \in I} A_i\right) - \mathbb{P}\left(B \cap \bigcap_{i \in I} A_i\right) \qquad \text{Everything in } \cap_{i \in A} \text{ not in } B$$

$$= \prod_{i \in I} \mathbb{P}(A_i) - \mathbb{P}(B) \prod_{i \in I} \mathbb{P}(A_i) \qquad \text{By assumption}$$

$$= \left(1 - \mathbb{P}(B)\right) \prod_{i \in I} \mathbb{P}(A_i) \qquad \text{Grouping terms}$$

$$= \mathbb{P}(B^c) \prod_{i \in I} \mathbb{P}(A_i) \qquad \text{Desired result}$$

To show $\mathcal{L}$ is closed under countable disjoint union, let $B_1, B_2, \ldots$ be disjoint events in $\mathcal{L}$ and call $B = \biguplus_{i=1}^{\infty} B_i$. Then observe that:

$$\mathbb{P}\left(B \cap \bigcap_{i \in I} A_i\right) = \mathbb{P}\left(\biguplus_{i=1}^{\infty} B_i \cap \bigcap_{i \in I} A_i\right) \qquad \text{How } B \text{ is defined}$$

$$= \mathbb{P}\left[\biguplus_{i=1}^{\infty} \left(B_i \cap \bigcap_{i \in I} A_i\right)\right] \qquad \text{If } B_1, B_2 \text{ disjoint, then } A \cap B_1 \text{ and } A \cap B_2 \text{ disjoint}$$

$$= \sum_{i=1}^{\infty} \mathbb{P}\left(B_i \cap \bigcap_{i \in I} A_i\right) \qquad \text{Properties of disjoint union}$$

$$= \sum_{i=1}^{\infty} \mathbb{P}(B_i) \prod_{i \in I} \mathbb{P}(A_i) \qquad \text{Independence assumption}$$

$$= \mathbb{P}\left(\biguplus_{i=1}^{\infty} B_i\right) \prod_{i \in I} \mathbb{P}(A_i) \qquad \text{Properties of disjoint union}$$

$$= \mathbb{P}(B) \prod_{i \in I} \mathbb{P}(A_i)$$

■

Lemma 6.2.4. Let $\{X_i\}_{i \in I}$ be a collection of independent random variables. Let I_1 and I_2 be two disjoint subsets of I. Then $\sigma\left(\{X_i\}_{i \in I_1}\right) = \mathcal{G}_1$ is independent of $\sigma\left(\{X_i\}_{i \in I_2}\right) = \mathcal{G}_2$.

Proof. In the spirit of Lemma 6.2.1 and Lemma 6.2.2, we can define $\mathcal{P}_1 = \left\{ \bigcap_{i \in I_1} A_i : A_i \in \sigma(X_i) \right\}$ and $\mathcal{P}_2 = \left\{ \bigcap_{i \in I_2} A_i : A_i \in \sigma(X_i) \right\}$. By Lemma 6.2.1, $\mathcal{P}_1$ and $\mathcal{P}_2$ are pi-systems. By Lemma 6.2.2, $\sigma(P_1) = \mathcal{G}_1$ and $\sigma(P_2) = \mathcal{G}_2$. Then applying Lemma 6.2.3, $\sigma(P_1) = \mathcal{G}_1$ is independent of $\sigma(P_2) = \mathcal{G}_2$. ∎

Theorem 6.3. Kolmogorov's 0-1 Law: Let $\{X_i\}_{i=1}^{\infty}$ be independent random variables and $\mathcal{T}$ the associated tail algebra. Then for all $A \in \mathcal{T}$, either $\mathbb{P}(A) = 0$ or $\mathbb{P}(A) = 1$.

Proof. Since the X_i's are independent, by Lemma 6.2.4 we have $\sigma\left(X_1, \ldots, X_n\right)$ independent of $\sigma\left(X_{n+1}, \ldots\right)$ for any $n \in \mathbb{N}$. By the definition of tail-algebra, $\mathcal{T} \in \sigma\left(X_{n+1}, \ldots\right)$ and so $\sigma\left(X_1, \ldots, X_n\right)$ is independent of $\mathcal{T}$. Since this holds for all n, we have $\bigcup_{n=1}^{\infty} \sigma\left(X_1, \ldots, X_n\right)$ independent of $\mathcal{T}$. But again by the definition of tail-algebra, $\mathcal{T} \in \bigcup_{n=1}^{\infty} \sigma\left(X_1, \ldots, X_n\right)$. So $\mathcal{T}$ is independent of itself. In particular, any event $A \in \mathcal{T}$ is independent of itself. That is, $\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A)$ and so $\mathbb{P}(A) \in \{0, 1\}$. ∎

6.3   Problems

Problem 6.1) Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and consider events $A, B \in \mathcal{F}$. Show that A and B are independent if and only if $\mathbb{1}_A$ and $\mathbb{1}_B$ are independent random variables.

Recall that random variables are independent if and only if the sigma-algebra generated by the random variables are independent. Here $\sigma(\mathbb{1}_A) = \{\mathbb{1}_A^{-1}(B) : B \in \mathbb{B}(\mathbb{R})\} = \{\emptyset, A, A^c, \Omega\}$ (this was shown on another homework). Similarly, $\sigma(\mathbb{1}_B) = \{\emptyset, B, B^c, \Omega\}$. For the sigma-algebras to be independent, we need each event in one sigma-algebra, call the event A_i, to be independent from each event in the other sigma-algebra, call the event B_j; $\mathbb{P}(A_i \cap B_j) = \mathbb{P}(A_i)\mathbb{P}(B_j)$.

Assume A and B are independent, i.e. $\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B)$. Then proceed one-by-one through the cases.

- If $A_i = \emptyset$, then $\mathbb{P}(\emptyset \cap B_j) = \mathbb{P}(\emptyset) = 0 = 0 \cdot \mathbb{P}(B_j) = \mathbb{P}(\emptyset)\mathbb{P}(B_j)$ for all $B_j \in B$

- If $A_i = \Omega$, then $\mathbb{P}(\Omega \cap B_j) = \mathbb{P}(B_j) = 1 \cdot \mathbb{P}(B_j) = \mathbb{P}(\Omega)\mathbb{P}(B_j)$

- If $A_i = A$, then there we can assume $B_j = B^c$ (the reverse argument for the points above give cases where $B_j = \emptyset$ and where $B_j = \Omega$, and we have assumed the case where $B_j = B$ in the proof). Then $\mathbb{P}(A \cap B^c) = \mathbb{P}(B) - \mathbb{P}(A \cap B) = \mathbb{P}(B) - \mathbb{P}(A)\mathbb{P}(B) = \mathbb{P}(B)\big(1 - \mathbb{P}(A)\big) = \mathbb{P}(B)\mathbb{P}(A^c)$

- if $A_i = A^c$, then we can assume $B_j = B^c$ and use the same argument as above.

On the other hand, if the sigma-algebras are independent, then a choice of $A \in \sigma(\mathbb{1}_A)$ and a choice of $B \in \sigma(\mathbb{1}_B)$ shows us that A and B are independent. This proves our result.

Problem 6.2) Show that covariance is bilinear, i.e. for any random variables $X_1, \ldots, X_n$ and $Y_1, \ldots, Y_m$ and any constants $\alpha_1, \ldots, \alpha_n$ and $\beta_1, \ldots, \beta_m \in \mathbb{R}$, that $\mathrm{Cov}\big(\sum\limits_{i=1}^{n} \alpha_i X_i, \sum\limits_{j=1}^{m} \beta_j Y_j\big) = \sum\limits_{i=1}^{n}\sum\limits_{j=1}^{m} \alpha_i\beta_j\mathrm{Cov}(X_i, Y_j)$, provided X_i, Y_j, and X_iY_j are integrable for each i, j.

The covariance of random variables X and Y is $\mathrm{Cov}(X, Y) = \mathbb{E}\big[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))\big] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$. So by linearity of expectations, we see:

$$
\begin{aligned}
\mathrm{Cov}\Big(\sum_{i=1}^{n} \alpha_i X_i, \sum_{j=1}^{m} \beta_j Y_j\Big) &= \mathbb{E}\left[\Big(\sum_{i=1}^{n}\alpha_i X_i - \mathbb{E}\Big(\sum_{i=1}^{n}\alpha_i X_i\Big)\Big)\Big(\sum_{j=1}^{m}\beta_j Y_j - \mathbb{E}\Big(\sum_{j=1}^{m}\beta_j Y_j\Big)\Big)\right] \\
&= \mathbb{E}\left[\Big(\sum_{i=1}^{n}\alpha_i X_i - \sum_{i=1}^{n}\alpha_i\mathbb{E}(X_i)\Big)\Big(\sum_{j=1}^{m}\beta_j Y_j - \sum_{j=1}^{m}\beta_j\mathbb{E}(Y_j)\Big)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n}\alpha_i\big(X_i - \mathbb{E}(X_i)\big)\sum_{j=1}^{m}\beta_j\big(Y_j - \mathbb{E}(Y_j)\big)\right] \\
&= \mathbb{E}\left[\sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_i\beta_j\big(X_i - \mathbb{E}(X_i)\big)\big(Y_j - \mathbb{E}(Y_j)\big)\right] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_i\beta_j\mathbb{E}\big[\big(X_i - \mathbb{E}(X_i)\big)\big(Y_j - \mathbb{E}(Y_j)\big)\big] \\
&= \sum_{i=1}^{n}\sum_{j=1}^{m}\alpha_i\beta_j\mathrm{Cov}(X_i, Y_j)
\end{aligned}
$$

Problem 6.3) Show that if $X_1, \ldots, X_n$ are uncorrelated, then $\mathbb{V}\big(\sum\limits_{i=1}^{n} X_i\big) = \sum\limits_{i=1}^{n}\mathbb{V}(X_i)$ provided $\mathbb{E}(X_i^2) < \infty$ for each i.

$$
\begin{aligned}
\mathbb{V}\Big(\sum_{i=1}^{n} X_i\Big) &= \mathrm{Cov}\Big(\sum_{i=1}^{n} X_i, \sum_{j=1}^{n} X_j\Big) && \text{Definition of Variance} \\
&= \sum_{i=1}^{n}\sum_{j=1}^{n}\mathrm{Cov}(X_i, X_j) && \text{From Problem 6.2} \\
&= \sum_{i=1}^{n}\mathrm{Cov}(X_i, X_i) && \text{Ignore cases where } i \neq j \text{ since } \mathrm{Cov}(X_i, X_j) = 0 \\
&= \sum_{i=1}^{n}\mathbb{V}(X_i) && \text{Desired result}
\end{aligned}
$$

7   Law Of Large Numbers

7.1   Definitions

Definition 7.1. Identically Distributed: Two random variables X and Y are identically distributed if $\mathbb{P}(X \in B) = \mathbb{P}(Y \in B)$ for all $B \in \mathbb{B}(\mathbb{R})$. Equivalently, we can check if $\mathbb{P}(X_i \leq t) = \mathbb{P}(X_1 \leq t)$ for all $t \in \mathbb{R}$ (equivalent by taking $B = (-\infty, t]$). Another equivalence is $\mathbb{E}\big(f(X_i)\big) = \mathbb{E}\big(f(X_1)\big)$ for all measurable f and all $i \geq 1$ provided the expectations exist (equivalent by using the push-forward formula in the above).

Example 7.1: Consider random variables $X \sim \text{Bern}(0.5)$ and $Y = 1 - X$. Then X and Y are identically distributed (they are both fair coin flips, with "heads" counted as 1 and "tails" as 0), but not independent (X literally causes Y; XY is always 0, so $\mathbb{E}(XY) = 0 \neq \mathbb{E}(X)\mathbb{E}(Y) = \frac{1}{2}\frac{1}{2} = \frac{1}{4}$).

Non-example 7.1: Consider random variables X and Y where X models a coin flip and Y models a dice role. Then X and Y are independent (XY has 12 possible outcomes, 6 of which are 0, so $\mathbb{E}(XY) = \frac{1}{2}\frac{1}{6}\sum_{n=1}^{6} n = \frac{21}{12} = \mathbb{E}(X)\mathbb{E}(Y) = \frac{1}{2}\frac{21}{6}$), but not identically distributed (for example, X does not even take the value 2).

Example 7.2: Consider random variables X and Y which are two independent copies of a normal random variable. Then X and Y are both independent and identically distributed. When random variables are both independent and identically distributed, we may abbreviate them as **iid**.

Definition 7.2. Infinitely Often: If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and if $\{A_i\}_{i=1}^{\infty}$ is a sequence of events in $\mathcal{F}$, then A_i occurs infinitely often if $\mathbb{P}\left(\bigcap_{n=1}^{\infty} \bigcup_{i \geq n} A_i\right) = 1$. Identifying intersection with "for all" and union with "there exists", this is saying "for all $n \in \mathbb{N}$, there exists an $i \geq n$ such that A_i occurs with probability 1", which is precisely the definition of $\limsup$;
$$\mathbb{P}\left(\bigcap_{n=1}^{\infty} \bigcup_{i \geq n} A_i\right) = 1 \iff \mathbb{P}\left(\limsup_{i \to \infty} A_i\right) = \mathbb{P}\left(\{\omega \in \Omega : \omega \in A_i \text{ for infinitely many } i\}\right) = 1.$$
We will often abbreviate this to $\mathbb{P}(A_i \text{ i.o.}) = 1$.

Definition 7.3. Eventually Always: If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and if $\{A_i\}_{i=1}^{\infty}$ is a sequence of events in $\mathcal{F}$, then eventually, A_i will always occur if $\mathbb{P}\left(\bigcup_{n=1}^{\infty} \bigcap_{i \geq n} A_i\right) = 1$. Identifying union with "there exists" and intersection with "for all", this is saying "there exists an $n \in \mathbb{N}$ such that for all $i \geq n$, A_i occurs with probability 1", which is the definition of $\liminf$;
$$\mathbb{P}\left(\bigcup_{n=1}^{\infty} \bigcap_{i \geq n} A_i\right) = 1 \iff \mathbb{P}\left(\liminf_{i \to \infty}\right) = \mathbb{P}\left(\{\omega \in \Omega : \omega \in A_i \text{ for all large enough } i\}\right) = 1.$$
While not universal notation, we may abbreviate this to $\mathbb{P}(A_i \text{ e.a.}) = 1$.

7.2   Theorems And Examples

Theorem 7.1. Borel-Cantelli: If $\{A_i\}_{i=1}^{\infty}$ is a sequence of events that satisfy $\sum_{i=1}^{\infty} \mathbb{P}(A_i)<\infty$, then $\mathbb{P}(A_i$ i.o.$) = 0$; i.e. only finitely many of the A_i's occur.

As a partial converse, if $\{B_i\}_{i=1}^{\infty}$ is a sequence of independent events (independence is not needed in the first part of the theorem), such that $\sum_{i=1}^{\infty} \mathbb{P}(B_i) = \infty$, then $\mathbb{P}(B_i$ i.o.$) = 1$.

Proof. For the first part of Borel-Cantelli, label $A_n' = \bigcup_{i \geq n} A_i$ for every n. As n increases, events are removed from the union; $A_n' \supseteq A_{n+1}' \supseteq A_{n+2}' \supseteq \cdots$. Then by continuity from above (Theorem 1.1, Page 7), $\mathbb{P}\left(\bigcap_{n=1}^{\infty} \bigcup_{i \geq n} A_i\right) = \mathbb{P}\left(\bigcap_{n=1}^{\infty} A_n'\right) = \lim_{n \to \infty} \mathbb{P}(A_n') = \lim_{n \to \infty} \mathbb{P}\left(\bigcup_{i \geq n} A_i\right)$. For any fixed n in the limit, we can use the union bound (Theorem 1.1, Page 7) to show that the probability is no more than $\sum_{i \geq n} \mathbb{P}(A_i)$; $\mathbb{P}\left(\bigcap_{n=1}^{\infty} \bigcup_{i \geq n} A_i\right) = \lim_{n \to \infty} \mathbb{P}\left(\bigcup_{i \geq n} A_i\right) = \lim_{n \to \infty} \sum_{i \geq n} \mathbb{P}(A_i)$. Since the infinite sum is less than infinity by assumption, so too is the tail, and we've proved our result.

For the second part of Borel-Cantelli, we want to show $\mathbb{P}(B_i$ i.o.$) = 1$ or equivalently $\mathbb{P}(B_i$ i.o.$^c) = 0$. See that:

$$\mathbb{P}\left(\left[\bigcap_{n=1}^{\infty} \bigcup_{i \geq n} B_i\right]^c\right) = \mathbb{P}\left(\left[\bigcup_{n=1}^{\infty} \bigcap_{i \geq n} B_i^c\right]\right) \quad \text{DeMorgan's Laws}$$

$$= \mathbb{P}\left(\bigcup_{n=1}^{\infty} B_n'\right) \quad \text{After calling } B_n' = \bigcap_{i \geq n} B_i^c, \text{ so } B_n' \subseteq B_{n+1}' \subseteq \cdots$$

$$\leq \sum_{n=1}^{\infty} \mathbb{P}(B_n') \quad \text{Union bound (Theorem 1.1, Page 7)}$$

It suffices to show $\mathbb{P}(B_n') = 0$ for all n. See that:

$$\mathbb{P}(B_n') = \lim_{n \to \infty} \mathbb{P}\left(\bigcap_{i \geq n} B_i^c\right) \quad \text{Continuity from above (Theorem 1.1, Page 7)}$$

$$= \prod_{i \geq n} \left(1 - \mathbb{P}(B_i)\right) \quad \text{By independence and compliment rules}$$

$$\leq \prod_{i \geq n} e^{-\mathbb{P}(B_i)} \quad \text{Using the inequality in Lemma 4.1, Page 34}$$

$$= e^{-\sum_{i=n}^{\infty} \mathbb{P}(B_i)} \quad \text{Product of exponentials is exponential of sums}$$

$$= e^{-\infty} = 0 \quad \text{By assumption of the proof}$$

∎

Example 7.3: If you flip a one-sided die followed by flipping a two-sided die, followed by flipping a three-sided die, etc. then the probability that a one is rolled infinitely often is one. This follows from Borel-Cantelli's second theorem since on a uniform probability measure, $\mathbb{P}(X_n = 1) = \frac{1}{n}$, and $\sum\limits_{n=1}^{\infty} \frac{1}{n} = \infty$

Non-example 7.2: If you roll a one-sided die, followed by flipping a four-sided die, followed by flipping a nine-sided die, etc. then the probability that a one is rolled infinitely often is zero (i.e. there is a *last* time a 1 is rolled). This follows from Borel-Cantelli's first theorem since on a uniform probability measure, $\mathbb{P}(X_n = 1) = \frac{1}{n^2}$ and $\sum\limits_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} < \infty$.

Proposition 7.1. L^4 Strong Law Of Large Numbers: The sample mean of independent $X_i \in L^4(\mathbb{P})$ converges almost surely to the true mean. More precisely, if $\{X_i\}_{i=1}^{\infty}$ is a sequence of i.i.d. random variables such that $\mathbb{E}(X_i^4) \leq c < \infty$, then $\left(\overline{X} = \frac{1}{n} \sum\limits_{i=1}^{n} X_n \right) \xrightarrow{a.s.} \mathbb{E}(X)$.

Proof. Without loss of generality, assume $\mathbb{E}(X) = 0$ (if not, replace X_i with $X_i - \mathbb{E}(X)$). Label $S_n = \sum\limits_{i=1}^{n} X_i$ so that $S_n^4 = \left(\sum\limits_{i=1}^{n} X_i \right)^4 = \sum\limits_{i=1}^{n} \sum\limits_{j=1}^{n} \sum\limits_{k=1}^{n} \sum\limits_{l=1}^{n} X_i X_j X_k X_l = \sum\limits_{i,j,k,l}^{n} X_i X_j X_k X_l$. By linearity of expectations, $\mathbb{E}(S_n^4) = \sum\limits_{i,j,k,l}^{n} \mathbb{E}(X_i X_j X_k X_l)$, a sum with n^4 terms.

Since the random variables are mutually independent, whenever i, j, k, l are distinct, $\mathbb{E}(X_i X_j X_k X_l) = \mathbb{E}(X_i)\mathbb{E}(X_j)\mathbb{E}(X_k)\mathbb{E}(X_l) = 0$. Similarly, whenever $i \neq j \neq k$, terms in the form $\mathbb{E}(X_i^3 X_j)$ factor as $\mathbb{E}(X_i^3)\mathbb{E}(X_j) = \mathbb{E}(X_i^3) \cdot 0 = 0$ and terms in the form $\mathbb{E}(X_i^2 X_j X_k)$ factor as $\mathbb{E}(X_i^2)\mathbb{E}(X_j)\mathbb{E}(X_k) = 0$. So only terms in the form $\mathbb{E}(X_i^4)$ and $\mathbb{E}(X_i^2 X_j^2)$ contribute to the sum.

There are $\binom{n}{1} = n$ terms in the form $\mathbb{E}(X_i^4)$ and $\binom{n}{2}\binom{4}{2} = \frac{4!}{2!2!} \frac{n!}{2!(n-2)!} = 3n(n-1)$ terms in the form $\mathbb{E}(X_i^2 X_j^2)$. This follows since there are $\binom{n}{2}$ ways to choose pairs (i,j) as our index in $\mathbb{E}(X_i^2 X_j^2)$, and, once the (i,j) pair is determined, there are $\binom{4}{2}$ distinct permutations (e.g., $\mathbb{E}(X_i X_i X_j X_j) = \mathbb{E}(X_i X_j X_i X_j)$) since fixing the location of the i terms determines the location of the j terms.

So our sum is $\mathbb{E}(S_n^4) = n\mathbb{E}(X_i^4) + 3n(n-1)\mathbb{E}(X_i^2 X_j^2)$ which, by assumption of the proof, is no more than $nc + 3n(n-1)\mathbb{E}(X_i^2 X_j^2)$. Since $\mathbb{E}(X_i^2 X_j^2) = \mathbb{E}(|X_i^2 X_j^2|)$, by Cauchy-Schwartz (Theorem 4.4.1, Page 34), $\mathbb{E}(X_i^2 X_j^2) \leq \mathbb{E}(|X_i^2|^2)^{\frac{1}{2}}\mathbb{E}(|X_j^2|^2)^{\frac{1}{2}} = \sqrt{\mathbb{E}(X_i^4)\mathbb{E}(X_j^4)} = \sqrt{c^2} = c$, and our bound becomes $\mathbb{E}(S_n^4) \leq nc + 3n(n-1)c = (3n^2 - 2n)c \leq 3n^2 c$.

Now let $\varepsilon > 0$ be given. Then $\mathbb{P}(\frac{S_n}{n} \geq \varepsilon) = \mathbb{P}(S_n^4 \geq n^4 \varepsilon^4) \leq \frac{\mathbb{E}(S_n^4)}{n^4 \varepsilon^4} = \frac{3n^2 c}{n^4 \varepsilon^4} = \frac{3c}{\varepsilon^4} \frac{1}{n^2}$ by Markov's Inequality (Theorem 4.1, Page 33) and the previously established bound. So we have $\sum\limits_{n=1}^{\infty} \mathbb{P}\left(\frac{S_n}{n} \geq \varepsilon \right) = \frac{3c}{\varepsilon^4} \sum\limits_{n=1}^{\infty} \frac{1}{n^2} = \frac{3c}{\varepsilon^4} \frac{\pi^2}{6} < \infty$. By the second Borel-Cantelli Lemma (Theorem 7.1, Page 60), $\mathbb{P}\left(\frac{S_n}{n} \geq \varepsilon \text{ i.o.} \right) = 0$. Since this holds for all $\varepsilon > 0$, $\frac{S_n}{n} \xrightarrow{a.s.} 0$ and we have proven our result. ∎

Lemma 7.1.1. Conditions For Almost Sure Convergence: A sequence of random variables X_n converges almost surely to X if $\mathbb{P}(|X_n - X| \leq \varepsilon$ for all large $n) = 1$ for all $\varepsilon > 0$.

Proof. By definition, $X_n \xrightarrow{a.s.} X$ if for all $\varepsilon > 0$, there exists an $N \in \mathbb{N}$ such that for all $n \geq N$, $\mathbb{P}\left(|X_n - X| \leq \varepsilon\right) = \mathbb{P}\left(\bigcup\limits_{N=1}^{\infty} \bigcap\limits_{n=N}^{\infty} |X_n - X| \leq \varepsilon\right) = 1$. Since this holds for all $\varepsilon > 0$, this must also hold for all $\frac{1}{k} > 0$. So $X_n \xrightarrow{a.s.} X$ if $\mathbb{P}\left(\bigcap\limits_{k=1}^{\infty} \bigcup\limits_{N=1}^{\infty} \bigcap\limits_{n=N}^{\infty} |X_n - X| \leq \frac{1}{k}\right) = 1$. Label the event $\bigcup\limits_{N=1}^{\infty} \bigcap\limits_{n=N}^{\infty} |X_n - X| \leq \frac{1}{k}$ as A_k. If $\mathbb{P}(A_k) = 1$ for all k, then $\mathbb{P}\left(\bigcap\limits_{k=1}^{\infty} A_k\right) = 1$. So to prove almost sure convergence, it suffices to show $\mathbb{P}\left(|X_n - X| \leq \varepsilon$ for all large $n\right) = 1$ ∎

Proposition 7.2. Law Of Large Numbers For Infinite Expectation: If $\{X_i\}_{i=1}^{\infty}$ is a sequence of i.i.d. random variables such that $\mathbb{E}(X_i) = \infty$, then the probability that the sample mean exists and is finite is 0.

Proof. We know that $\mathbb{E}(|X_1|) = \int_0^{\infty} \mathbb{P}(|X_1| \geq t)\, dt \leq \int_0^{\infty} \mathbb{P}(|X_1| \geq \lfloor t \rfloor)\, dt = \sum\limits_{n=0}^{\infty} \mathbb{P}(|X_1| \geq n) = \sum\limits_{n=0}^{\infty} \mathbb{P}(|X_n| \geq n)$. Since $\mathbb{E}(|X_n|) = \infty$, $\sum\limits_{n=0}^{\infty} \mathbb{P}(|X_n| \geq n) = \infty$ as well. Then by the second Borel-Cantelli Lemma (Theorem 7.1, Page 60), $\mathbb{P}(|X_n| \geq n$ i.o.$) = \mathbb{P}(\frac{|X_n|}{n} \geq 1$ i.o.$) = 1$.

Label the event that $\lim\limits_{n\to\infty} \sum\limits_{i=1}^{n} \frac{X_i}{n}$ exists and is finite A. We claim that $A \cap \left\{\frac{|X_n|}{n} \geq 1$ i.o.$\right\} = \emptyset$. If this claim is true, then since $\mathbb{P}(\left\{\frac{|X_n|}{n} \geq 1$ i.o.$\right\}) = 1$, we must have $\mathbb{P}(A) = 0$. So we focus our aim on proving the claim.

Call $S_n = \sum\limits_{i=1}^{n} X_n$. Observe $\left|\frac{S_n}{n} - \frac{S_{n-1}}{n-1}\right| = \left|\frac{S_{n-1}+X_n}{n} - \frac{S_{n-1}}{n-1}\right| = \left|\frac{X_n}{n} - \frac{S_{n-1}}{n(n-1)}\right| \geq \frac{|X_n|}{n} - \frac{|S_{n-1}|}{n(n-1)}$ by the reverse triangle inequality. If A occurs, then $\left|\frac{S_n}{n} - \frac{S_{n-1}}{n-1}\right| \to 0$ as n grows large (by Cauchy criteria for limits). On the other hand, if both A and $\left\{\frac{|X_n|}{n} \geq 1$ i.o.$\right\}$ occur, then $\limsup\limits_{n\to\infty} \frac{|X_n|}{n} - \frac{S_{n-1}}{n(n-1)} = \limsup\limits_{n\to\infty} \frac{|X_n|}{n} \geq 1$. But this violates the above inequality.   ∎

Theorem 7.2. Strong Law Of Large Numbers: Let $\{X_j\}_{j=1}^{\infty}$ be a sequence of iid random variables such that $-\infty < \mathbb{E}(X_1) = \mu < \infty$. Then $\sum_{j=1}^{n} \frac{X_j}{n} \xrightarrow{\text{a.s.}} \mu$.

Proof. First, we reduce to the non-negative case. Since $X_j = X_j^+ - X_j^-$ and $X_j^+ = \max(X_j, 0)$ are iid, it suffices to show that $\frac{X_1^+ + X_2^+ + \cdots + X_n^+}{n} \xrightarrow{a.s.} \mathbb{E}(X_1^+)$ and $\frac{X_1^- + X_2^- + \cdots + X_n^-}{n} \xrightarrow{a.s.} \mathbb{E}(X_1^-)$ as doing so would imply $\frac{X_1 + X_2 + \cdots + X_n}{n} \xrightarrow{a.s.} \mathbb{E}(X_1^+) - \mathbb{E}(X_1^-) = \mathbb{E}(X_1)$. So henceforth we can assume that $X_j \geq 0$.

Second, we can truncate the random variables. Call $Y_j = X_j \mathbb{1}_{\{X_j \leq j\}}$. Then see that $\sum_{j=1}^{\infty} \mathbb{P}(X_j > j) = \sum_{j=1}^{\infty} \mathbb{P}(X_1 > j) \leq \int_0^{\infty} \mathbb{P}(X_1 \geq t)\, dt = \mathbb{E}(X_1) < \infty$ by assumption of the proof. Then using Borel-Cantelli (Theorem 7.1, Page 60), we have $\mathbb{P}(X_j > j \;\; \text{i.o.}) = 0$. In other words, $\mathbb{P}(X_j \leq j) = 1$ for all large j and thus $Y_j = X_j$ for all large j. So if we define $S_n = X_1 + X_2 + \cdots + X_n$ and $T_n = Y_1 + Y_2 + \cdots + Y_n$, then $S_n - T_n$ is finite with probability 1, and further $\limsup_{n \to \infty} \left| \frac{S_n - T_n}{n} \right| = 0$ almost surely.

Third, we can apply Chebyshev's Inequality (Theorem 4.2, Page 33). For any $\varepsilon > 0$, we have $\mathbb{P}(|\frac{T_n - \mathbb{E}(T_n)}{n}| \geq \varepsilon) = \mathbb{P}(|T_n - \mathbb{E}(T_n)| \geq n\varepsilon) \leq \frac{\mathbb{V}(T_n)}{\varepsilon^2 n^2}$. We can further bound the probability as $\mathbb{P}(|\frac{T_n - \mathbb{E}(T_n)}{n}| \geq \varepsilon) \leq \frac{\mathbb{V}(T_n)}{\varepsilon^2 n^2} \leq \frac{\mathbb{E}(X_1^2 \mathbb{1}_{\{X_1 \leq n\}})}{n\varepsilon^2}$ after observing:

$$
\begin{aligned}
\mathbb{V}(T_n) &= \sum_{j=1}^{n} \mathbb{V}(Y_j) \leq \sum_{j=1}^{n} \mathbb{E}(Y_j^2) && \text{By independence and variance formula} \\
&= \sum_{j=1}^{n} \mathbb{E}(X_j^2 \mathbb{1}_{\{X_j \leq j\}}) \leq \sum_{j=1}^{n} \mathbb{E}(X_j^2 \mathbb{1}_{\{X_j \leq n\}}) && \text{How } Y_j \text{ was defined and } j \leq n \\
&= n\mathbb{E}(X_1^2 \mathbb{1}_{\{X_1 \leq n\}}) && \text{Since the } Y_j\text{'s are identically distributed}
\end{aligned}
$$

Fourth, we can examine subsequences in the form $n_k = \lfloor \alpha^k \rfloor$ for some $\alpha > 1$. By the bound established above and the knowledge that $\frac{1}{2}\alpha^k \leq n_k \leq \alpha^k$, we have:

$$
\sum_{k=1}^{\infty} \mathbb{P}\left(\left| \frac{T_{n_k} - \mathbb{E}(T_{n_k})}{n_k} \right| \geq \varepsilon \right) \leq \sum_{k=1}^{\infty} \frac{\mathbb{E}(X_1^2 \mathbb{1}_{\{X_1 \leq n_k\}})}{\varepsilon^2 n_k} \leq \sum_{k=1}^{\infty} \frac{\mathbb{E}(X_1^2 \mathbb{1}_{\{X_1 \leq \alpha^k\}})}{\frac{1}{2}\varepsilon^2 \alpha^k}
$$

Note that for all $x \in \mathbb{R}$, $\sum_{k=1}^{\infty} \frac{\mathbb{1}_{\{x \leq \alpha^k\}}}{\alpha^k} = \sum_{k : \alpha^k \geq x} \frac{1}{\alpha^k} \leq \frac{x^{-1}}{1 - \alpha^{-1}}$ since the sum is a geometric series with a common ratio of α^{-1} and first term less than x^{-1}. So using linearity of expectations and the assumption of the proof, we can continue to examine the sum as follows:

$$
\sum_{k=1}^{\infty} \frac{\mathbb{E}(X_1^2 \mathbb{1}_{\{X_1 \leq \alpha^k\}})}{\frac{1}{2}\varepsilon^2 \alpha^k} = \frac{2}{\varepsilon^2} \mathbb{E}\left(\sum_{k=1}^{\infty} \frac{X_1^2 \mathbb{1}_{\{X_1 \leq \alpha^k\}}}{\alpha^k} \right) \leq \frac{2}{\varepsilon^2} \mathbb{E}\left(\frac{X_1}{1 - \alpha^{-1}} \right) = \frac{2\mu}{\varepsilon^2 (1 - \alpha)^{-1}}
$$

In particular, this bound is finite, and we can thus use the first Borel-Cantelli Theorem (Theorem 7.1, Page 60) to say that $\mathbb{P}\left(\left| \frac{T_{n_k} - \mathbb{E}(T_{n_k})}{n_k} \right| \geq \varepsilon \;\; \text{i.o.} \right) = 0$. Since ε was arbitrary,

this means that $\frac{T_{n_k}-\mathbb{E}(T_{n_k})}{n_k} \xrightarrow{a.s.} 0$. Further, since we have assumed each X_j is positive and since we know $\lim_{j\to\infty} Y_j = \lim_{j\to\infty} X_j \mathbb{1}_{\{X_j \leq j\}} = X_1$, we can apply the Monotone Convergence Theorem (Theorem 5.7, Page 43) to conclude that $\lim_{j\to\infty} \mathbb{E}(Y_j) = \mathbb{E}(X_1) = \mu$. From how we defined T_n, we have $\mathbb{E}(T_n) = \mathbb{E}(Y_1) + \mathbb{E}(Y_2) + \cdots + \mathbb{E}(Y_n) \implies \frac{\mathbb{E}(T_n)}{n} = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(Y_i) \xrightarrow{a.s.} \mu$. Combined with the argument that $\lim_{k\to\infty} \frac{T_{n_k}-\mathbb{E}(T_{n_k})}{n_k} \overset{a.s.}{=} 0$, this means $\lim_{k\to\infty} \frac{T_{n_k}}{n_k} \overset{a.s.}{=} \mu$ and so, from our second step, $\lim_{n\to\infty} |\frac{S_n-T_n}{n}| \overset{a.s.}{=} 0 \implies \frac{S_{n_k}}{n_k} \xrightarrow{a.s.} \mu$.

Finally, we can use interpolation. Since we have assumed $X_i \geq 0$, we know that $\{S_n\}_{n=1}^{\infty}$ is non-decreasing. As such, given some n, there is a k such that $n_k \leq n \leq n_{k+1}$ and consequently $\frac{S_{n_k}}{n_{k+1}} \leq \frac{S_n}{n} \leq \frac{S_{n_{k+1}}}{n_k} \implies \frac{n_k}{n_{k+1}}\frac{S_{n_k}}{n_k} \leq \frac{S_n}{n} \leq \frac{S_{n_{k+1}}}{n_{k+1}}\frac{n_{k+1}}{n_k}$. We have already shown $\frac{S_{n_k}}{n_k} \xrightarrow{a.s.} \mu$ and know $\frac{n_{k+1}}{n_k} \xrightarrow{a.s.} \alpha$ since $n_k = \lfloor \alpha^k \rfloor$. So in the limit, the aforementioned inequality becomes $\frac{1}{\alpha}\mu \leq \liminf_{n\to\infty}\frac{S_n}{n} \leq \limsup_{n\to\infty}\frac{S_n}{n} \leq \alpha\mu$. Since this holds for all $\alpha \geq 1$, we can take a monotonically decreasing sequence $\alpha_k \searrow 1$ and then with probability 1 we have $\frac{\mu}{\alpha_k} \leq \left[\liminf_{n\to\infty}\frac{S_n}{n} \leq \limsup_{n\to\infty}\frac{S_n}{n}\right] \leq \alpha_k\mu$. Since this holds for all k in our sequence, we have $\mu = \lim_{k\to\infty}\frac{\mu}{\alpha_k} \leq \left[\liminf_{n\to\infty}\frac{S_n}{n} \leq \limsup_{n\to\infty}\frac{S_n}{n}\right] \leq \lim_{k\to\infty}\alpha_k\mu = \mu$. This sandwich proves that $\lim_{n\to\infty}\frac{S_n}{n} = \mu$ almost surely. ∎

7.3   Problems

Problem 7.1) Let $\{X_i\}_{i=1}^{\infty}$ be iid Exponential(1) random variables, i.e. $\mathbb{P}(X_i \geq x) = e^{-x}$ for $x \geq 0$. Let $M_n = \max_{1 \leq i \leq n} X_i$.

a. Show that $\mathbb{P}\left(\limsup\limits_{n \to \infty} X_n / \ln(n) = 1\right) = 1.$

For all $\varepsilon > 0$, we have:

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\frac{X_n}{\ln(n)} \geq (1 + \varepsilon)\right) = \sum_{n=1}^{\infty} \mathbb{P}(X_n \geq \ln(n)(1 + \varepsilon)) = \sum_{n=1}^{\infty} e^{-\ln(n)(1+\varepsilon)} = \sum_{n=1}^{\infty} \frac{1}{n^{1+\varepsilon}} < \infty$$

Then by the first Borel-Cantelli Theorem, $\mathbb{P}\left(\frac{X_n}{\ln(n)} \geq (1 + \varepsilon) \quad \text{i.o.}\right) = 0$, and in particular $\limsup\limits_{n \to \infty} \frac{X_n}{\ln(n)} \leq (1 + \varepsilon)$ almost surely. Since this is true for every $\varepsilon > 0$, it must be the case that $\limsup\limits_{n \to \infty} \frac{X_n}{\ln(n)} \leq 1$ almost surely.

On the other hand, we have:

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\frac{X_n}{\ln(n)} \geq 1\right) = \sum_{n=1}^{\infty} \mathbb{P}(X_n \geq \ln(n)) = \sum_{n=1}^{\infty} e^{-\ln(n)} = \sum_{n=1}^{\infty} \frac{1}{n} = \infty$$

Then since the random variables are independent, by the second Borel-Cantelli Theorem, $\mathbb{P}\left(\frac{X_n}{\ln(n)} \geq 1 \quad \text{i.o.}\right) = 1$, and in particular $\limsup\limits_{n \to \infty} \frac{X_n}{\ln(n)} \geq 1$ almost surely. This proves both directions.

b.   Show that $\mathbb{P}\left(\liminf\limits_{n \to \infty} M_n / \ln(n) \geq 1\right) = 1.$ **As a hint, use the fact that** $\sum\limits_{n=1}^{\infty} e^{-n^{1-c}} < \infty$ **for any** $c < 1.$

For all $\varepsilon > 0$, we have:

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\frac{M_n}{\ln(n)} < (1 - \varepsilon)\right) = \sum_{n=1}^{\infty} \mathbb{P}(M_n < \ln(n)(1 - \varepsilon))$$

$$= \sum_{n=1}^{\infty} \mathbb{P}(X_1 < \ln(n)(1 - \varepsilon))^n \qquad \text{Independence and definition of maximum}$$

$$= \sum_{n=1}^{\infty} \left(1 - e^{-(\ln(n)(1-\varepsilon))}\right)^n = \sum_{n=1}^{\infty} \left(1 - \frac{1}{n^{1-\varepsilon}}\right)^n \qquad \text{Identical distributions and simplifying}$$

$$\leq \sum_{n=1}^{\infty} e^{-n^{1-\varepsilon}} < \infty$$

So by the first Borel-Cantelli Theorem, $\mathbb{P}\left(\frac{M_n}{\ln(n)} \leq (1 - \varepsilon) \quad \text{i.o.}\right) = 0$ and in particular $\liminf\limits_{n \to \infty} \frac{M_n}{\ln(n)} \geq 1.$

Problem 7.2) Let $\{X_n\}_{i=1}^{\infty}$ be any sequence of random variables. Show that there exists a sequence of constants $\{c_n\}_{n=1}^{\infty}$ such that $\frac{1}{c_n}X_n \xrightarrow{a.s.} 0$.

Let $\varepsilon > 0$ be given. Then we see

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\frac{X_n}{c_n} \geq \varepsilon\right) = \sum_{n=1}^{\infty} \mathbb{P}\left(X_n \geq c_n\varepsilon\right) = \sum_{n=1}^{\infty} 1 - \mathbb{P}\left(X_n < c_n\varepsilon\right) = \sum_{n=1}^{\infty} 1 - F_{X_n}(c_n\varepsilon)$$

where F_{X_n} is understood to be the cumulative distribution function of X_n. Then a choice of $c_n = n$ (eventually) forces $F_{X_n}(c_n\varepsilon) \geq (1 - \frac{1}{n^2})$ since $\lim_{t\to\infty} F_{X_n}(t) = 1$ and since ε is positive. Then using this sequence, we have (for some k):

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\frac{X_n}{c_n} \geq \varepsilon\right) = \sum_{n=1}^{\infty} 1 - F_{X_n}(c_n\varepsilon) \leq \sum_{n=1}^{k} 1 - F_{X_n}(c_n\varepsilon) + \sum_{n=k+1}^{\infty} \frac{1}{n^2} \leq k + \sum_{n=k+1}^{\infty} \frac{1}{n^2} < \infty$$

Then since $\sum_{n=1}^{\infty} \mathbb{P}\left(\frac{X_n}{c_n} \geq \varepsilon\right) < \infty$, we use the first Borel-Cantelli Theorem to say that $\mathbb{P}\left(\frac{X_n}{c_n} \geq \varepsilon \quad \text{i.o.}\right) = 0$ and in particular that $\mathbb{P}\left(\lim_{n\to\infty} \frac{X_n}{c_n} < \varepsilon\right) = 1$. Since this holds for all $\varepsilon > 0$, we must have $\mathbb{P}\left(\lim_{n\to\infty} \frac{X_n}{c_n} = 0\right) = 1$ as desired.

Problem 7.3) Let $\{A_n\}_{n=1}^{\infty}$ be a sequence of independent events such that $\mathbb{P}(A_n) < 1$ for all n and such that $\mathbb{P}(\bigcup_{i=1}^{\infty} A_i) = 1$. Show that $\mathbb{P}(A_n \text{ i.o.}) = 1$.

By the second Borel-Cantelli Theorem, since $\{A_n\}_{n=1}^{\infty}$ are independent, if we can show that $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$, then we will have $\mathbb{P}(A_n \text{ i.o.}) = 1$. If $\limsup_{n\to\infty} \mathbb{P}(A_n) > 0$, then clearly we have this result. So we can assume that $\limsup_{n\to\infty} \mathbb{P}(A_n) = 0$.

First notice that $\mathbb{P}(\bigcup_{n=1}^{\infty} A_n) = 1 \implies \mathbb{P}\left(\left(\bigcup_{n=1}^{\infty} A_n\right)^c\right) = 0 \implies \mathbb{P}(\bigcap_{n=1}^{\infty} A_n^c) = \prod_{n=1}^{\infty} \mathbb{P}(A_n^c) = 0$ by DeMorgan's Laws and the independence assumption. Then applying the negative natural log to both sides, we have $-\ln(0) = -\ln\left(\prod_{n=1}^{\infty} \mathbb{P}(A_n^c)\right) \implies \infty = \sum_{n=1}^{\infty} -\ln\left(\mathbb{P}(A_n^c)\right) = \sum_{n=1}^{\infty} -\ln\left(1 - \mathbb{P}(A_n)\right) = \sum_{n=1}^{\infty} \ln\left(\frac{1}{1-\mathbb{P}(A_n)}\right).$

We have proved earlier that for any real number x, $1 + x \leq e^x$. In particular, for any $y < 1$ we have $\left[1 + \left(\frac{y}{1-y}\right) = \frac{1}{1-y}\right] \leq e^{\frac{y}{1-y}}$. Applying this to the equality obtained above, we have $\infty = \sum_{n=1}^{\infty} \ln\left(\frac{1}{1-\mathbb{P}(A_n)}\right) \leq \sum_{n=1}^{\infty} \ln\left(e^{\frac{\mathbb{P}(A_n)}{1-\mathbb{P}(A_i)}}\right) = \sum_{n=1}^{\infty} \frac{\mathbb{P}(A_n)}{1-\mathbb{P}(A_n)} = \sum_{n=1}^{\infty} \frac{1}{1-\mathbb{P}(A_n)}\mathbb{P}(A_n)$. Then since $\limsup_{n\to\infty} \mathbb{P}(A_n) = 0$, $\lim_{n\to\infty} \frac{1}{1-\mathbb{P}(A_n)} = 1$ and we see $\sum_{n=1}^{\infty} \mathbb{P}(A)_n = \infty$ as desired.

Problem 7.4) Without appealing to the Strong Law Of Large Numbers, prove the weak L^2 law of large numbers ("Weak" refers to convergence in probability as opposed to almost sure convergence, and L^2 refers to the assumption of bounded variance.). Assume $X_1, \ldots, X_n$ are uncorrelated random variables with $\mathbb{E}(X_i) = \mu$ for every i, and $\mathbb{V}(X_i) \leq C$ for every i, where C is some finite constant. Define $S_n = X_1 + \cdots + X_n$. Show that S_n/n converges to μ in probability as $n \to \infty$.

Let $\varepsilon > 0$ be given. It is easy to show that since the random variables are uncorrelated; the sum of their variances is the variance of their sums. Upon multiplying by $\frac{1}{n}$, we see that $\mathbb{V}(\sum_{i=1}^{n} \frac{1}{n} X_i) = \mathbb{V}(\frac{1}{n} \sum_{i=1}^{n} X_i) = \frac{1}{n^2} \mathbb{V}(\sum_{i=1}^{n} X_i) \leq \frac{1}{n^2} nC \leq \frac{C}{n}$. Then by Chebyshev's inequality, $\mathbb{P}\left(|\frac{S_n}{n} - \mu| \geq \varepsilon\right) \leq \frac{\mathbb{V}(S_n/n)}{\varepsilon^2} \leq \frac{C/n}{\varepsilon^2} = \frac{C}{n\varepsilon^2} \xrightarrow{n \to \infty} 0$.

8 Central Limit Theorem

8.1 Definitions

Definition 8.1. Characteristic Function: The characteristic function of a random variable X is $\varphi_X(t) = \mathbb{E}(e^{itX})$. This is essentially the Fourier Transform for random variables.

Example 8.1: Let X be a binary random variable that takes the values ± 1 (these are called Rademacher Random Variables), each with probability $\frac{1}{2}$. Label the support of X as x_0, x_1. Then the characteristic function of X is calculated as:

$$\mathbb{E}(e^{itX}) = \sum_{k=0}^{\infty} e^{itx_k} \mathbb{P}(X = x_k) = \sum_{k=0}^{1} e^{itx_k} \mathbb{P}(X = x_k) = e^{it(-1)}\frac{1}{2} + e^{it(1)}\frac{1}{2}$$
$$= \frac{1}{2}(\cos(-t) + i\sin(-t)) + \frac{1}{2}(\cos(t) + i\sin(t))$$
$$= \frac{1}{2}(\cos(-t) + \cos(t)) + \frac{1}{2}(\sin(-t) + \sin(t)) = \cos(t) \quad \text{sine is odd, cosine is even}$$

Example 8.2: After recalling the Taylor Expansion for e^x, we calculate the characteristic function for $X \sim \text{Poi}(\lambda)$ as:

$$\mathbb{E}(e^{itX}) = \sum_{k=0}^{\infty} e^{itx_k} \mathbb{P}(X = x_k) = \sum_{k=0}^{\infty} e^{itx_k} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{\infty} \frac{(\lambda e^{it})^k}{k!} = e^{-\lambda}e^{\lambda e^{it}} = e^{\lambda(e^{it}-1)}$$

Definition 8.2. Laplace Transform: The laplace transform of a random variable X is $L_X(t) = \mathbb{E}(e^{-tX})$.

Definition 8.3. Complex Modulus: Where z and w are complex numbers, a and b are real numbers, and i is the complex unit, recall the following facts about complex numbers (the proofs of which follow from Taylor Series expansions and elementary results from calculus).

1. $e^{itx} = \cos(tx) + i\sin(tx)$

2. $e^{a+bi} = e^a e^{bi}$. The distance from the origin is $e^a = r$ and the angle from the origin is $e^{bi} = \theta$.

3. where $|\cdot|$ denotes the complex modulus, $|e^{itx}| = |a + bi| = \sqrt{a^2 + b^2}$

4. $|wz| = |w| \cdot |z|$

5. $|e^z| = e^{|z|}$

8.2 Theorems And Examples

Lemma 8.0.1. Properties of Characteristic Functions: For independent random variables X and Y and constants c, we have

1. $\varphi_{X+Y}(t) = \mathbb{E}(e^{it(X+Y)}) = \mathbb{E}(e^{itx})\mathbb{E}(e^{itY}) = \varphi_X(t)\varphi_Y(t)$

2. $\varphi_{X+c}(t) = \mathbb{E}(e^{it(x+c)}) = \mathbb{E}(e^{itX})\mathbb{E}(e^{itc}) = e^{itc}\varphi_X(t)$

3. $\varphi_{cX}(t) = \mathbb{E}(e^{itcX}) = \mathbb{E}(e^{i(ct)X}) = \varphi_X(ct)$

4. $\varphi_X(0) = \mathbb{E}(e^{i\cdot 0 \cdot X}) = 1$

5. $\varphi_X^{(n)}(0) = i^n\mathbb{E}(X^n)$ (can recover moments, see theorem below (Theorem 8.1, Page 69))

Lemma 8.0.2. Exchanging derivatives and exponents: If $f : (\mathbb{R} \times \mathbb{R}) \to \mathbb{C}$ is continuously differentiable in t and there exists a g such that $|\frac{\partial}{\partial t}f(t,X)| \leq g(X)$ for all $t \in \mathbb{R}$ with $\mathbb{E}(g(X)) < \infty$, then $t \mapsto \mathbb{E}(f(t,X))$ is differentiable and $\frac{d}{dt}\mathbb{E}(f(t,X)) = \mathbb{E}(\frac{\partial}{\partial t}f(t,X))$.

Proof.

$$\frac{d}{dt}\mathbb{E}(f(t,X))$$

$$= \lim_{h\to 0} \frac{\mathbb{E}(f(t+h,X)) - \mathbb{E}(f(t,X))}{h} \qquad \text{Limit definition of derivative}$$

$$= \lim_{h\to 0} \mathbb{E}\left(\frac{f(t+h,X) - f(t,X)}{h}\right) \qquad \text{Linearity of expectations}$$

$$= \lim_{h\to 0} \mathbb{E}\left(\frac{\partial}{\partial t}f(c,X)\right) \qquad \text{Existence of } t < c < t+h \text{ is by Mean-Value Theorem}$$

$$= \mathbb{E}\left(\frac{\partial}{\partial t}f(t,X)\right) \qquad \text{Dominated Convergence Theorem (Theorem 5.8, Page 43)}$$

Where the last inequality follows by assumption of the proof (there is an integrable $g(X)$ such that $|\frac{\partial}{\partial t}f(t,X)| \leq g(X)$ for all $t \in \mathbb{R}$ with probability one) and the definition of derivative ($\frac{\partial}{\partial t}f(c,X) \overset{a.s.}{\to} \frac{\partial}{\partial t}f(t,X)$ as $h \to 0$). ∎

Theorem 8.1. Derivatives Of Characteristic Functions: If $\mathbb{E}(|X|^n) < \infty$, then φ has n continuous derivatives and $\varphi^{(n)}(t) = \mathbb{E}((ix)^n e^{itX})$. In particular $\varphi^{(n)}(0) = \mathbb{E}(i^n X^n e^0) = i^n\mathbb{E}(X^n)$, so we can recover the nth raw moment of a random variable from finding the nth derivative of it's characteristic function at zero.

Proof. We have $\frac{d^n}{dt^n}e^{itX} = (iX)^n e^{itX}$. So if $\mathbb{E}(|X|^n) < \infty$, applying Lemma 8.0.2 n times gives the result. ∎

Lemma 8.1.1. If $z \in \mathbb{C}$ with $|z| \leq 1$, then $|e^z - (1 + z)| \leq |z|^2$.

Proof. By Taylor Expansion, $f(z) = e^z = \sum\limits_{n=0}^{\infty} \frac{\partial^n}{\partial z^n} f(0) \frac{z^n}{n!} = 1 + z + \frac{z^2}{2!} + \cdots$. Then observe:

$$
\begin{aligned}
|e^z - (1 + z)| &= \sum_{n=2}^{\infty} \frac{z^n}{n!} && \text{Taylor Expansion about 0 for } e^z \\
&\leq |z|^2 \sum_{n=2}^{\infty} \frac{1}{n!} && \text{Since } |z| < 1 \text{ by assumption, } |z|^2 > |z|^n \text{ for } n > 2 \\
&\leq |z|^2 \sum_{n=2}^{\infty} \frac{1}{2^{n-1}} && \text{For } n \geq 1,\ n! \geq 2^{n-1} \\
&= |z|^2 \sum_{n=1}^{\infty} \frac{1}{2^n} = |z|^2
\end{aligned}
$$

∎

Lemma 8.1.2. For all $z, w \in \mathbb{C}$ and $n \in \mathbb{N}$, $|z^n - w^n| \leq n(\max\{|z|, |w|\})^{n-1}$.

Proof. Without loss of generality, assume $|w| \leq |z|$. Then:

$$
\begin{aligned}
z^n - w^n &= (z - w)(z^{n-1} + wz^{n-2} + \cdots + w^{n-2}z + w^{n-1}) \\
|z^n - w^n| &\leq |z - w|(|z|^{n-1} + |w||z|^{n-2} + \cdots + |w|^{n-2}|z| + |w|^{n-1}) && \text{Triangle Inequality} \\
&\leq |z - w|(|z|^{n-1} + |z||z|^{n-2} + \cdots + |z|^{n-2}|z| + |z|^{n-1}) && \text{Assumed } |w| \leq |z| \\
&\leq |z - w| \cdot n \cdot |z|^{n-1}
\end{aligned}
$$

∎

Lemma 8.1.3. If $\{c_n\} \in \mathbb{C}$ is a sequence of numbers converging to c, then $\lim\limits_{n \to \infty} (1 + \frac{c_n}{n})^n = e^c$.

Proof. By the triangle inequality, $\left|(1 + \frac{c_n}{n})^n - e^c\right| \leq \left|(1 + \frac{c_n}{n})^n - e^{c_n}\right| + |e^{c_n} - e^c|$. In the limit, the second term in the sum vanishes and, after multiplying an exponent by one and flipping the order of the difference, we have $\left|(1 + \frac{c_n}{n})^n - e^c\right| \leq \left|(e^{\frac{c_n}{n}})^n - (1 + \frac{c_n}{n})^n\right|$. Notice that $|1 + \frac{c_n}{n}| \leq \left|e^{\frac{c_n}{n}}\right|$ from the Taylor Expansion in (Lemma 4.1, Page 34). So identifying $\frac{c_n}{n}$ with w and $e^{\frac{|c_n|}{n}}$ with z in (Lemma 8.1.2, Page 70), we can further bound our expression with $\left|(e^{\frac{c_n}{n}})^n - \left(1 + \frac{c_n}{n}\right)^n\right| \leq n \left|e^{\frac{c_n}{n}}\right|^{n-1} |e^{\frac{c_n}{n}} - (1 + \frac{c_n}{n})|$. Since c_n converges to c in the limit, for all large n, $|\frac{c_n}{n}| < 1$. With this in mind we can apply (Lemma 8.1.1, Page 70) to improve our bound to $n \left|e^{\frac{c_n}{n}}\right|^{n-1} |e^{\frac{c_n}{n}} - (1 + \frac{c_n}{n})| \leq n \left|e^{\frac{c_n}{n}}\right|^{n-1} |\frac{c_n}{n}|^2$. Some final arithmetic gives us $n \left|e^{\frac{c_n}{n}}\right|^{n-1} |\frac{c_n}{n}|^2 = n \left|e^{\frac{c_n}{n}}\right|^{n-1} \frac{|c_n|}{n^2} \leq \frac{e^{|c_n| \frac{n-1}{n}} |c_n|}{n} \leq \frac{e^{|c_n||c_n|}}{n}$, which goes to zero as n goes to infinity. ∎

Lemma 8.1.4. $\lim\limits_{T\to\infty}\int_0^T \frac{\sin(\theta t)}{t}\,dt = \begin{cases} \frac{\pi}{2}, & \theta > 0 \\ 0, & \theta = 0 \\ \frac{-\pi}{2}, & \theta < 0 \end{cases}$

Proof. If $\theta = 0$, then the integrand is 0. So assume $\theta \neq 0$. Using the substitution $u = \theta t$ we have $\frac{d}{dt}u = \theta$ and so $du = \theta\,dt$. Then $\int_0^T \frac{\sin(\theta t)}{t}\,dt = \int_0^{\theta t} \frac{\sin([u])}{[\frac{u}{\theta}]}[\frac{du}{\theta}] = \int_0^{\theta t} \frac{\sin(u)}{u}\,du$ since as t varies from 0 to T, $u = \theta t$ varies from 0 to θT. This is the sine integral, which evaluates to $\text{sgn}(\theta) \cdot \int_0^{|\theta T|} \frac{\sin(\theta t)}{t}\,dt$. Since we are interested in the limit of this quantity as $T \to \infty$, it suffices to show $\lim\limits_{T\to\infty}\int_0^T \frac{\sin(t)}{t}\,dt = \frac{\pi}{2}$.

Our strategy is to write the integral as a double integral that will help us get a usable limit. Since $\int_0^\infty e^{-ty}\,dy = \frac{1}{t}$, $t\int_0^\infty e^{-ty}\,dy = 1$ and so:

$$\int_0^T \frac{\sin(t)}{t}\,dt = \int_0^T \frac{\sin(t)}{t}\cdot\left(t\int_0^\infty e^{-ty}\,dy\right)dt = \int_0^T \left(\int_0^\infty \sin(t)e^{-ty}\,dy\right)dt \qquad (8.1)$$

See that the integrand is integrable:

$$\int_0^T \left(\int_0^\infty \left|\sin(t)e^{-ty}\right|\,dy\right)dt$$

$$= \int_0^\pi \left(\int_0^\infty \sin(t)e^{-ty}\,dy\right)dt + \int_\pi^T \left(\int_0^\infty |\sin(t)|\,e^{-ty}\,dy\right)dt$$

$$\leq \int_0^\pi \left(\int_0^\infty \sin(t)e^{-ty}\,dy\right)dt + \int_\pi^T \left(\int_0^\infty e^{-ty}\,dy\right)dt$$

$$= \int_0^\pi \sin(t)\frac{1}{t}\,dt + \int_\pi^T \frac{1}{t}\,dt$$

$$\leq \int_0^\pi 1\,dt + \int_\pi^T 1\,dt < \infty$$

So we can apply Fubini's Theorem (Theorem **??**, Page **??**) to flip the integrals and write:

$$\int_0^T \frac{\sin(t)}{t}\,dt = \int_0^\infty \left(\int_0^T \sin(t)e^{-ty}\,dt\right)dy \qquad (8.2)$$

We now try to evaluate the inner integral, call it $I(t) = \int_0^T \sin(t)e^{-ty}\,dt$, using integration by parts. Choose $u = e^{-ty}$ and $dv = \sin(t)\,dy$ so that $\frac{d}{dt}u = -ye^{-ty}$ and $v = -\cos(t)$. Then we can write:

$$I(t) = \left(-e^{-ty}\cos(t)\right)\Big|_0^T - \int_0^T y\cos(t)e^{-ty}\,dt \qquad (8.3)$$

We again apply integration by parts to the remaining integral. Choose $u = ye^{-ty}$ and $dv = \cos(t)\,dt$ so that $\frac{d}{dt}u = -y^2e^{-ty}$ and $v = \sin(t)$. Then:

$$\int_0^T y\cos(t)e^{-ty}\,dy = \left(ye^{-ty}\sin(t)\right)\big|_0^T - \int_0^T -y^2 e^{-ty}\sin(t)\,dt = \left(ye^{-ty}\sin(t)\right)\big|_0^T + y^2 I(t) \quad (8.4)$$

Equations 8.3 and 8.4 combine to say:

$$I(t) = \frac{1}{1+y^2}\left(\left(-e^{-ty}\cos(t)\right)\big|_0^T - \left(ye^{-ty}\sin(t)\right)\big|_0^T\right)$$

$$= \frac{1}{1+y^2}\left(1 - e^{-Ty}\cos(T) - ye^{-Ty}\sin(T)\right)$$

When $y > 0$, the second factor is bounded and tends to 1 as T grows large. Thus there exists a c such that $\frac{1}{1+y^2}\left(1 - e^{-Ty}\cos(T) - ye^{-Ty}\sin(T)\right) < \frac{c}{1+y^2}$ for all $T, y > 0$ and we can use the Dominated Convergence Theorem (Theorem 5.8, Page 43) to say:

$$\lim_{T\to\infty}\int_0^T \frac{\sin(t)}{t}\,dt = \lim_{T\to\infty}\int_0^\infty \left(\int_0^T \sin(t)e^{-ty}\,dt\right)dy \quad \text{Equation 8.2}$$

$$= \int_0^\infty \lim_{T\to\infty}\left(\int_0^T \sin(t)e^{-ty}\,dt\right)dy \quad \text{Dominated Convergence Theorem}$$

$$= \int_0^\infty \frac{1}{1+y^2}\,dy$$

$$= \tan^{-1}(y)\big|_0^\infty = \frac{\pi}{2}$$

∎

Lemma 8.1.5. There is a $c \in \mathbb{R}$ so that $|\int_0^T \frac{\sin(\theta t)}{t}\,dt| \le c$ for all $\theta \in \mathbb{R}$ and for all $T \ge 0$.

Proof. From Lemma 8.1.4, there exists a t_o such that for all $T \ge t_0$, $\left|\int_0^T \frac{\sin(\theta t)}{t}\,dt\right| \le \pi$ (since in the limit, the difference between $\left|\int_0^T \frac{\sin(\theta t)}{t}\,dt\right|$ and $\frac{\pi}{2}$ becomes arbitrarily small).

For $T < t_0$, notice that $\left|\int_0^T \frac{\sin(\theta t)}{t}\,dt\right| \le |\int_0^T \left|\frac{\sin(\theta t)}{t}\right|\,dt \le \int_0^T 1\,dt = T$. So a choice of $c = \max\{\pi, t_0\}$ yields our result. ∎

Theorem 8.2. Inversion Formula: The characteristic function uniquely determines the law of a random variable. Specifically, for all continuity points a, b of F_X, $\mathbb{P}(a < X < b) = \lim\limits_{T \to \infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi_X(t) \, dt$.

Proof. For notational ease, define $I(t) = \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi_X(t) \, dt$. By the Riemann-Stieltjes (Definition 3.2, Page 21) view of expectation, this is $I(t) = \int_{-T}^{T} \left(\int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it} e^{itx} \mu_X(dx) \right) dt$. We want to flip the integrals, so aim to apply Fubini's Theorem (Theorem **??**, Page **??**).

To do so, we must show that the integrand is integrable. Since $\left| e^{i\theta} \right| = 1$ for all θ, observe:

$$\left| \frac{e^{-ita} - e^{-itb}}{it} e^{itx} \right| = \left| \frac{e^{-ita} - e^{-itb}}{it} \right| \cdot \left| e^{itx} \right| = \left| \int_{a}^{b} e^{-ity} \, dy \right| \cdot \left| e^{itx} \right| \leq \int_{a}^{b} \left| e^{-ity} \right| \, dy = (b - a)$$

So $\int_{-T}^{T} \left(\int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it} e^{itx} \mu_X(dx) \right) dt \leq \int_{-T}^{T} \left(\int_{-\infty}^{\infty} (b - a) \, \mu_X(dx) \right) dt = 2T(b - a) < \infty$, which proves we can apply Fubini. Thus $I(T) = \int_{-\infty}^{\infty} \left(\int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} e^{itx} \, dt \right) \mu_X(dx)$. We now closely examine the integrand.

Using Euler's Formula, write the integrand in the form $A + B$:

$$\frac{e^{-ita} - e^{-itb}}{it} e^{itx} = \frac{e^{it(x-a)} - e^{it(x-b)}}{it}$$

$$= \frac{\cos(t(x - a)) + i \sin(t(x - a))}{it} - \frac{\cos(t(x - b)) + i \sin(t(x - b))}{it}$$

$$= \frac{\cos(t(x - a)) - \cos(t(x - b))}{it} + \frac{\sin(t(x - a)) - \sin(t(x - b))}{t}$$

$$= A + B$$

Using the linearity of the intergral, $\int_{-T}^{T} (A + B) \, dt = \int_{-T}^{T} A \, dt + \int_{-T}^{T} B \, dt$. Since cosine is an even function in t, the numerator in A is an even function. Since it is an odd function in t, the denominator in A is an odd function. So A is an odd function being integrated over a symmetric interval, and thus evaluates to zero. After abbreviating $S(\theta, T) = \int_{0}^{T} \frac{\sin(\theta t)}{t} \, dt$, our integral becomes:

$$I(T) = \int_{-\infty}^{\infty} \left(\int_{-T}^{T} \frac{\sin(t(x - a)) - \sin(t(x - b))}{t} \, dt \right) \mu_X(dx)$$

$$= 2 \int_{-\infty}^{\infty} \left(\int_{0}^{T} \frac{\sin(t(x - a)) - \sin(t(x - b))}{t} \, dt \right) \mu_X(dx)$$

$$= 2 \int_{-\infty}^{\infty} [S(x - a, T) - S(x - b, T)] \, \mu_X(dx)$$

Taking the limit and applying Lemma 8.1.5, we see:

$$\lim_{T \to \infty} I(T) = 2 \int_{-\infty}^{\infty} \lim_{T \to \infty} [S(x - a, T) - S(x - b, T)] \, \mu_X(dx)$$

We can compute the inside limit directly from Lemma 8.1.4. Since $a < b$, we have:

$$\lim_{T \to \infty} [S(x-a,T) - S(x-b,T)]$$

$$= \lim_{T \to \infty} S(x-a,T) - \lim_{T \to \infty} S(x-b,T)$$

$$= \begin{cases} \frac{\pi}{2}, & (x-a) > 0 \\ 0, & (x-a) = 0 \\ \frac{-\pi}{2}, & (x-a) < 0 \end{cases} - \begin{cases} \frac{\pi}{2}, & (x-b) > 0 \\ 0, & (x-b) = 0 \\ \frac{-\pi}{2}, & (x-b) < 0 \end{cases}$$

$$= \begin{cases} \pi, & a < x < b \\ \frac{\pi}{2}, & a < (x=b) \text{ or } (x=a) < b \\ 0, & x < a \text{ or } x > b \end{cases}$$

Then breaking up the integral, we reach our desired conclusion:

$$\lim_{T \to \infty} I(T) = 2 \left(\int_{x \in (a,b)} \pi \, \mu_X(dx) + \int_{x\{a,b\}} \frac{\pi}{2} \mu_X(dx) + \int_{x \notin [a,b]} 0 \, \mu_X(dx) \right)$$

$$= 2 \left(\pi \mathbb{P}(a < X < b) + \frac{\pi}{2} \mathbb{P}(X \in \{a,b\}) \right)$$

$$= 2\pi \mathbb{P}(a < X < b)$$

∎

Theorem 8.3. Continuity Theorem: $X_n \xrightarrow{d} X$ if and only if $\varphi_{X_n}(t) \to \varphi_X(t)$ for all $t \in \mathbb{R}$.

Proof. Assume $X_n \xrightarrow{d} X$. By definition, X_n converges in distribution to X if for all bounded and continuous f, $\lim_{n\to\infty} \mathbb{E}(f(X_n)) = \mathbb{E}(f(X))$. Since $f(x) = e^{itx} = \cos(tx) + i\sin(tx)$ is bounded and continuous, we have $\lim_{n\to\infty} \mathbb{E}\left(e^{itX_n}\right) = \lim_{n\to\infty} \varphi_{X_n}(t) = \mathbb{E}\left(e^{itX}\right) = \varphi_X(t)$.

Now assume $\varphi_{X_n}(t) \to \varphi_X(t)$ for all t. We focus our efforts on proving that $\{X_n\}_{n=1}^\infty$ is tight, i.e. we want to show that there is a $M > 0$ so that $\mathbb{P}(|X_n| > M)$ is arbitrarily small for any n.

We start with the definition. The random variable may or may not have a density function, so to keep to the most general terms, let μ_{X_n} denote the law of X_n. By the Riemann–Stieltjes (Definition 3.2, Page 21) view of expectation, we have:

$$\varphi_{X_n}(t) = \mathbb{E}(e^{itX_n}) = \int_{\mathbb{R}} e^{itx}\mu_{X_n}(dx) \tag{8.5}$$

Now by Lemma 8.0.1, $\varphi_{X_n}(0) = 1$. If X_n is tightly concentrated around 0, then small perturbations t about 0 should yield values of $\varphi_{X_n}(t)$ close to one. On the symmetric interval $(-\delta, \delta)$, the total deviation of $\varphi_{X_n}(t)$ from 1 is $\int_{-\delta}^{\delta}(1 - \varphi_{X_n}(t))\,dt$, so the average deviation is $\frac{1}{2\delta}\int_{-\delta}^{\delta}(1 - \varphi_{X_n}(t))\,dt$. Plugging Equation 8.5 into this, we see the average deviation is:

$$\frac{1}{2\delta}\int_{-\delta}^{\delta}\left(1 - \left(\int_{\mathbb{R}} e^{itx}\mu_{X_n}(dx)\right)\right)dt \tag{8.6}$$

Since μ_{X_n} is a probability mass, $1 - \left(\int_{\mathbb{R}} e^{itx}\mu_{X_n}(dx)\right) = \int_{\mathbb{R}} \mu_{X_n}(dx) - \left(\int_{\mathbb{R}} e^{itx}\mu_{X_n}(dx)\right) = \int_{\mathbb{R}}(1 - e^{itx})\mu_{X_n}(dx)$ and we can write the average deviation as:

$$\frac{1}{2\delta}\int_{-\delta}^{\delta}\left(\int_{\mathbb{R}}\left(1 - e^{itx}\right)\mu_{X_n}(dx)\right)dt = \frac{1}{2\delta}\int_{\mathbb{R}}\left(\int_{-\delta}^{\delta}\left(1 - e^{itx}\right)dt\right)\mu_{X_n}(dx) \tag{8.7}$$

After applying Fubini's Theorem (Theorem **??**, Page **??**) to exchange the order of the integrals. Since $\frac{d}{dt}e^{itx} = ixe^{itx}$, we can evaluate the inner integral in Equation 8.7 as:

$$\frac{1}{2\delta}\int_{\mathbb{R}}\left(\left(t - \frac{e^{itx}}{ix}\right)\Big|_{-\delta}^{\delta}\right)\mu_{X_n}(dx) = \frac{1}{2\delta}\int_{\mathbb{R}}\left(2\delta + \frac{e^{-i\delta x} - e^{i\delta x}}{ix}\right)\mu_{X_n}(dx) \tag{8.8}$$

By Euler's formula, $e^{-i\delta x} - e^{i\delta x} = (\cos(-\delta x) + i\sin(-\delta x)) - (\cos(\delta x) + i\sin(\delta x))$. Cosine is an even function, and sine an odd function, so the numerator can be written in the form $(\cos(\delta x) - i\sin(\delta x)) - (\cos(\delta x) + i\sin(\delta x)) = -2i\sin(\delta x)$. So Equation 8.8 can be written:

$$\frac{1}{2\delta}\int_{\mathbb{R}}\left(2\delta + \frac{-2i\sin(\delta x)}{ix}\right)\mu_{X_n}(dx) = \frac{1}{2\delta}\int_{\mathbb{R}}\left(2\delta - \frac{2\sin(\delta x)}{x}\right)\mu_{X_n}(dx) \tag{8.9}$$

It is natural to factor out 2δ. So Equation 8.9 becomes:

$$\frac{1}{2\delta} \int_{\mathbb{R}} 2\delta \left(1 - \frac{\sin(\delta x)}{\delta x}\right) \mu_{X_n}(dx) = \int_{\mathbb{R}} \left(1 - \frac{\sin(\delta x)}{\delta x}\right) \mu_{X_n}(dx) \tag{8.10}$$

Since $\frac{\sin(\delta x)}{\delta x} \leq 1$ (this follows since for $f(y) = y - \sin(y)$, $\frac{d}{dy} f(y) = 1 - \cos(y) \geq 0$, and $f'(0) = 0$, so $y - \sin(y) \geq 0$), the integrand $1 - \frac{\sin(\delta x)}{\delta x}$ is always positive. And since $\int_{\mathbb{R}} \left(1 - \frac{\sin(\delta x)}{\delta x}\right) \mu_{X_n}(dx) = \int_{\{|X_n| \geq \frac{2}{\delta}\}} \left(1 - \frac{\sin(\delta x)}{\delta x}\right) \mu_{X_n}(dx) + \int_{\{|X_n| < \frac{2}{\delta}\}} \left(1 - \frac{\sin(\delta x)}{\delta x}\right) \mu_{X_n}(dx)$, the average deviation is:

$$\frac{1}{2\delta} \int_{-\delta}^{\delta} (1 - \varphi_{X_n}(t)) \, dt = \int_{\mathbb{R}} \left(1 - \frac{\sin(\delta x)}{\delta x}\right) \mu_{X_n}(dx) \geq \int_{\{|X_n| \geq \frac{2}{\delta}\}} \left(1 - \frac{\sin(\delta x)}{\delta x}\right) \mu_{X_n}(dx) \tag{8.11}$$

When $|X_n| \geq \frac{2}{\delta}$, $|\delta x| \geq 2$ and so $\left|\frac{\sin(\delta x)}{\delta x}\right| \leq \frac{1}{2}$ and $1 - \frac{\sin(\delta x)}{\delta x} \geq \frac{1}{2}$. Then we can write:

$$\frac{1}{2\delta} \int_{-\delta}^{\delta} (1 - \varphi_{X_n}(t)) \, dt \geq \int_{\{|X_n| \geq \frac{2}{\delta}\}} \left(1 - \frac{\sin(\delta x)}{\delta x}\right) \mu_{X_n}(dx) = \int_{\{|X_n| \geq \frac{2}{\delta}\}} \frac{1}{2} \mu_{X_n}(dx) \tag{8.12}$$

Integrating Equation 8.12, we have:

$$\frac{1}{2\delta} \int_{-\delta}^{\delta} (1 - \varphi_{X_n}(t)) \, dt \geq \int_{\{|X_n| \geq \frac{2}{\delta}\}} \frac{1}{2} \mu_{X_n}(dx) = \frac{1}{2} \mathbb{P}\left(|X_n| \geq \frac{2}{\delta}\right) \tag{8.13}$$

Now let $\varepsilon > 0$ be given. Since $\lim_{t \to 0} \varphi_X(t) = \varphi_X(0) = 1$, the definition of limit guarantees the existence of a δ_1 such that whenever $|t - 0| = |t| < \delta_1$, $|\varphi_X(t) - 1| < \varepsilon$. Further, since $\varphi_X(t) < 1$, $1 - \varphi_X(t) \leq |\varphi_X(t) - 1| < \varepsilon$. So choosing $\delta < \delta_1$, we see:

$$\frac{1}{2\delta} \int_{-\delta}^{\delta} (1 - \varphi_X(t)) \, dt \leq \frac{1}{2\delta} \int_{-\delta}^{\delta} \varepsilon \, dt = \frac{1}{2\delta} (\delta \varepsilon + \delta \varepsilon) = \varepsilon \tag{8.14}$$

Since $\lim_{n \to \infty} (1 - \varphi_{X_n}(t)) = (1 - \varphi_X(t))$ and since $(1 - \varphi_{X_n}(t)) \leq 2$, we can apply the Dominated Convergence Theorem (Theorem 5.8, Page 43) to say that $\lim_{n \to \infty} \frac{1}{2\delta} \int_{-\delta}^{\delta} (1 - \varphi_{X_n}(t)) \, dt = \frac{1}{2\delta} \int_{-\delta}^{\delta} \lim_{n \to \infty} (1 - \varphi_{X_n}(t)) \, dt = \frac{1}{2\delta} \int_{-\delta}^{\delta} (1 - \varphi_X(t)) \, dt$. So Equation 8.13 and Equation 8.14 become:

$$\frac{1}{2} \mathbb{P}\left(|X_n| \geq \frac{2}{\delta}\right) \leq \lim_{n \to \infty} \frac{1}{2\delta} \int_{-\delta}^{\delta} (1 - \varphi_{X_n}(t)) \, dt = \frac{1}{2\delta} \int_{-\delta}^{\delta} (1 - \varphi_X(t)) \, dt \leq \varepsilon \tag{8.15}$$

Since $\varepsilon > 0$ was arbitrary, this proves that $\{X_n\}_{n=1}^{\infty}$ is tight. Then by Helly's Selection Criteria (Theorem 5.12, Page 46), every subsequence of $\{X_n\}_{n=1}^{\infty}$ admits a further subsequence that convergences weakly. Since $\varphi_{X_n}(t) \to \varphi_X(t)$, and since characteristic functions uniquely determine the law (Theorem 8.2, Page 73), that weak limit must be X. Since this limit holds for every subsequence, we must have $X_n \xrightarrow{d} X$ as desired.   ∎

Theorem 8.4. Central Limit Theorem: Let $\{X_n\}_{n=1}^{\infty}$ be an iid sequence of random variables. If $\mathbb{V}(X_i) = \sigma^2 < \infty$, $\mathbb{E}(X_i) = \mu$, and $S_n = \sum\limits_{i=1}^{n} X_i$, then $\frac{S_n - n\mu}{\sqrt{n\sigma^2}} \xrightarrow{d} N(0,1)$.

Proof. By the Continuity Theorem (Theorem 8.3, Page 75), it suffices to show the characteristic functions converge in distribution. In Problem 4.2 Page 37, we derived the moment generating function of a standard normal random variable Y as $M_Y(t) = e^{\frac{t^2}{2}}$. Since $M_Y(t) = \mathbb{E}(e^{tY})$ and since $\varphi_Y(t) = \mathbb{E}(e^{itY})$, substituting t for it in the computation in Problem 4.2 yields $\varphi_Y(t) = e^{\frac{-t^2}{2}}$. So, we aim to show the characteristic functions on the left side converge to this value.

Assume $\mathbb{E}(X) = \mu = 0$ (if not, just replace each X_i with $X_i - \mu$). With this assumption, $\mathbb{V}(X) = \mathbb{E}(X^2) = \sigma^2$ and so for a single random variable X, we have:

$$\varphi_X(t) = \varphi_X(0) + \varphi_X'(0) + \varphi_X''(0)\frac{t^2}{2} + o(t^2) \qquad \text{Taylor Expansion about } c = 0$$

$$= 1 + \sigma^2\frac{t^2}{2} + o(t^2) \qquad\qquad \text{Part 5 of Lemma 8.0.1}$$

Since each X_i is independent, by part 1 of Lemma 8.0.1:

$$\varphi_{S_n}(t) = \left(1 + \sigma^2\frac{t^2}{2} + o(t^2)\right)^n$$

For any fixed n, $\frac{1}{\sqrt{n\sigma^2}}$ is a constant, so by part 3 of Lemma 8.0.1:

$$\varphi_{\frac{S_n}{\sqrt{n\sigma^2}}}(t) = S_n\left(\frac{t}{\sqrt{n\sigma^2}}\right) = \left(1 + \sigma^2\frac{\left[\frac{t}{\sqrt{n\sigma^2}}\right]^2}{2} + o\left(\left[\frac{t}{\sqrt{n\sigma^2}}\right]^2\right)\right)^n = \left(1 + \frac{t^2}{2n} + o\left(\frac{t^2}{n\sigma^2}\right)\right)^n$$

Then call $c_n = \frac{-t^2}{2} + n \cdot o\left(\frac{t^2}{n\sigma^2}\right)$. See that $\lim\limits_{n\to\infty} c_n = -\frac{t^2}{2}$ since, by the definition of $o\left(\frac{t^2}{n\sigma^2}\right)$, there exists a sequence $a_n \xrightarrow{n\to\infty} 0$ such that $o\left(\frac{t^2}{n\sigma^2}\right) = \frac{t^2}{n\sigma^2}a_n$ and then $n \cdot o\left(\frac{t^2}{n\sigma^2}\right) = \frac{t^2}{\sigma^2}a_n \xrightarrow{n\to\infty} 0$. So by Lemma 8.1.3:

$$\lim_{n\to\infty} \varphi_{\frac{S_n}{\sqrt{n\sigma^2}}}(t) = \lim_{n\to\infty}\left(1 + \frac{t^2}{2n} + o\left(\frac{t^2}{n\sigma^2}\right)\right)^n = \lim_{n\to\infty}\left(1 + \frac{c_n}{n}\right)^n = e^{\frac{-t^2}{2}}$$

∎

8.3   Problems

Problem 8.1) Roll 166 six-sided dice and add the outcomes. Using the Central Limit Theorem, estimate the probability that the total is at least 537.

The mean of one roll is $\mu = \sum_{n=1}^{6} \frac{n}{6} = \frac{7}{2}$.

The variance of one roll is $\mathbb{E}(X^2) - \mathbb{E}(X)^2 = \left(\sum_{n=1}^{6} \frac{n^2}{6} \right) - \left(\frac{7}{2} \right)^2 = \frac{\frac{6(6+1)(2 \cdot 6+1)}{6}}{6} - \frac{49}{4} = \frac{35}{12}$

Then using the central limit theorem, we get an approximation of $\mathbb{P}\left(Z \geq \frac{537 - \frac{166 \cdot 7}{2}}{\sqrt{166 \cdot \frac{35}{12}}} \right)$, or about $\mathbb{P}(Z \geq -2) \approx 0.975$.

Problem 8.2) The Mega Millions lottery is played as follows. You pick five distinct integers between 1 and 70 (order doesn't matter), as well as one integer between 1 and 25 (which could be a repeat of one of the five other numbers). The lottery does the same, uniformly at random from all possible choices. If all six numbers match, you win (a share of) the jackpot.

a. Compute the probability of winning the jackpot.

There are $\binom{70}{5}$ ways to pick the five numbers without replacement or regard to order. So the probability of matching all five is $\frac{1}{\binom{70}{5}}$. We consider the sixth selection independent of the first five picks, so the total probability of winning the jackpot is $\frac{1}{25 \cdot \binom{70}{5}} = \frac{1}{302,575,350}$.

b. Now suppose 1 million people play (where each person selects their six numbers independently at random, uniformly from all possible choices.) Estimate the probability that no one wins the jackpot.

We have $n = 1,000,000$ and $p = \frac{1}{302,575,350}$, so parameter $\lambda = \frac{n}{p} \approx \frac{1}{300}$. Then using the random variable $X \sim \text{Poi}\left(\frac{1}{300}\right)$, we see $\mathbb{P}(X = 0) = \frac{\lambda^{[0]} e^{-\frac{1}{300}}}{[0]!} = e^{-\frac{1}{300}} \approx 0.9967$.

Problem 8.3) Suppose $\varphi_X : \mathbb{R} \to \mathbb{C}$ is a characteristic function. That is, there is a random variable X such that $\varphi_X(t) = \mathbb{E}(e^{itX})$ for all $t \in \mathbb{R}$.

a. Show that $\varphi_X(-t)$ is the complex conjugate of $\varphi_X(t)$.

Observe $\varphi_X(-t) = \mathbb{E}(e^{-itX}) = \mathbb{E}\big(\cos(-tX) + i\sin(-tX)\big) = \mathbb{E}\big(\cos(tX) - i\sin(tX)\big)$ since cosine is an even function and sine an odd function. This is exactly the complex conjugate of $\varphi_X(t) = \mathbb{E}\big(e^{itX}\big) = \mathbb{E}\big(\cos(tX) + i\sin(tX)\big)$.

b. Show that $|\varphi_X(t)|^2$ is also a characteristic function.

First note that for any complex number z, say $z = a + bi$, $|z|^2 = \big(\sqrt{a^2 + b^2}\big)^2 = a^2 + b^2$. This is precisely $z \cdot \bar{z} = (a + bi)(a - bi) = a^2 - b^2 i^2 = a^2 + b^2$.

So from part a, $|\varphi_X(t)|^2 = \varphi_X(t) \cdot \varphi_X(-t) = \mathbb{E}(e^{itX})\mathbb{E}(e^{-itX})$. We want to show this is a characteristic function for some random variable Z. Consider $Z = X - Y$ where Y is an independent copy of X. The characteristic function of Z is then $\varphi_X(t) = \mathbb{E}(e^{it(X-Y)}) = \mathbb{E}(e^{itX}e^{-itY}) = \mathbb{E}(e^{itX})\mathbb{E}(e^{-itY})$ where the last equality follows from the independence of X and Y (Lemma 8.0.1, Page 69). Since characteristic functions uniquely characterize distributions (Theorem 8.2, Page 73), and since Y was chosen to have the same distribution as X, we can write the final equality as $\mathbb{E}(e^{itX})\mathbb{E}(e^{-itX})$ as desired.

c. Show that $\mathrm{Re}\big(\varphi_X(t)\big)$, the real part of $\varphi_X(t)$, is also a characteristic function.

For any $z = a + bi \in \mathbb{C}$, $\mathrm{Re}(z) = \frac{z + \bar{z}}{2}$ since $\frac{(a+bi)+(a-bi)}{2} = \frac{2a}{2} = a$.

From part a, see:

$$\mathrm{Re}\,[\varphi_X(t)] = \frac{1}{2}\varphi(t) + \frac{1}{2}\varphi(-t) = \frac{1}{2}\Big(\mathbb{E}\big(\cos(tX) + i\sin(tX)\big) + \mathbb{E}\big(\cos(-tX) + i\sin(-tX)\big)\Big)$$

Using linearity of expectations, we have:

$$\mathrm{Re}\,[\varphi_X(t)] = \frac{1}{2}\Big(\mathbb{E}\big(\cos(tX) + \cos(-tX)\big) + i\mathbb{E}\big(\sin(tX) + \sin(-tX)\big)\Big)$$

Using the parity of sine and cosine, we further have:

$$\mathrm{Re}\,[\varphi_X(t)] = \frac{1}{2}\Big(\mathbb{E}\big(2\cos(tX)\big) + i\mathbb{E}(0)\Big) = \frac{1}{2}\Big(2\mathbb{E}(\cos(tX))\Big) = \mathbb{E}\big(\cos(tX)\big)$$

We want to show this is a characteristic function for some random variable Z, and our steps have already revealed the solution. Consider Y independent of X with $\mathbb{P}(Y = 1) = \mathbb{P}(Y = -1) = \frac{1}{2}$, and the random variable $Z = XY$. Then we have:

$$\varphi_Z(t) = \mathbb{E}(e^{itZ}) = \frac{1}{2}\mathbb{E}\big(e^{itX}\big) + \frac{1}{2}\mathbb{E}\big(e^{-itX}\big) = \mathbb{E}(\cos(tX))$$

Problem 8.4) Show that if X and Y are independent and $X + Y \overset{d}{=} X$, then $Y = 0$ almost surely.

Since characteristic functions uniquely determine distributions of random variables (Theorem 8.2, Page 73), and $X + Y$ is equal in distribution to X by assumption, $\varphi_X(t) = \varphi_{X+Y}(t)$ for all $t \in \mathbb{R}$. Since X and Y are independent, by properties of characteristic functions (Lemma 8.0.1, Page 69), $\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t)$. Then as $\varphi_X(t) = \varphi_X(t)\varphi_Y(t)$ for all t, we must have $\varphi_Y(t) = 1$ whenever $\varphi_X(t) \neq 0$.

Further, since $\varphi_X(0) = 1$ and since characteristic functions are continuous, there exists a $\delta > 0$ such that whenever $t \in [-\delta, \delta]$, $\varphi_X(t) \neq 0$ (and consequently $\varphi_Y(t) = 1$). In Equation 8.11 from the proof of the Continuity Theorem (Theorem 8.3, Page 75), we showed that:

$$\frac{1}{2\delta} \int_{-\delta}^{\delta} (1 - \varphi_Y(t)) \, dt = \mathbb{E}\left(1 - \frac{\sin(\delta Y)}{\delta Y}\right)$$

Since $\varphi_Y(t) = 1$ for every t in $(-\delta, \delta)$, the integrand becomes a constant 0. Since sine is an odd function, $\frac{\sin(\delta Y)}{\delta Y} = \left|\frac{\sin(\delta Y)}{\delta Y}\right| \leq 1$ and thus $1 - \frac{\sin(\delta Y)}{\delta Y} \geq 0$.

A non-negative random variable whose expectation is 0 is almost surely 0. So $\frac{\sin(\delta Y)}{\delta Y} \overset{a.s.}{=} 1$, which implies $Y \overset{a.s.}{=} 0$ from the result $\lim_{x \to 0} \frac{\sin(x)}{x} = 1$.

9   Glossary

Algebra: A collection of sets $\mathcal{A}$ from a non-empty set Ω is an algebra provided $\mathcal{A}$ is closed under finite unions and complements. That is, $\mathcal{A}$ is an algebra if whenever $A_1, A_2, \ldots, A_n \in \mathcal{A}$ we have $\bigcup_{i=1}^{n} A_i \in \mathcal{A}$ and $A_1^C \in \mathcal{A}$. (Definition 1.9, Page 5).

Characteristic Function: The characteristic function of a random variable X is given by $\varphi_X(t) = \mathbb{E}(e^{itX})$. (Definition 8.1, Page 68).

Coefficient Of Determination: Where $\rho(X, Y)$ is the correlation between random variables X and Y, the coefficient of determination is simply it's square; $r^2 = \rho(X, Y)^2$. Note $r^2 \in [0, 1]$. (Definition 3.10, Page 22).

Convergence (Almost Surely): A sequence of random variables X_n converges almost surely to a random variable X, denoted $X_n \xrightarrow{a.s.} X$, if $\mathbb{P}(\lim_{n \to \infty} X_n = X) = 1$. To be precise, this is saying $X_n \xrightarrow{a.s.} X$ if $\mathbb{P}\left(\left\{\omega \in \Omega : \lim_{n \to \infty} X_n(\omega) = X(\omega)\right\}\right) = 1$. (Definition 5.2, Page 38).

Convergence (in Distribution; Weak Convergence): A sequence of random variables X_n converges in distribution to a random variable X, denoted $X_n \xrightarrow{d} X$, if $\lim_{n \to \infty} F_{X_n}(x) = F_X(x)$ for all points x where the CDF (Theorem 1.3, Page 7) F_X is continuous. An equivalent definition is that $X_n \xrightarrow{d} X$ provided $\lim_{n \to \infty} \mathbb{E}(f(X_n)) = \mathbb{E}(f(X))$ for all bounded and continuous $f : \mathbb{R} \to \mathbb{R}$. (Definition 5.4, Page 39).

Convergence (in L^p): A sequence of random variables X_n converges in L^p to X, denoted $X_n \xrightarrow{L^p} X$, if $X \in L^p(\mathbb{P})$ and $\lim_{n \to \infty} \|X_n - X\|_p = 0$ (Definition 4.6, Page 32). When dealing with $p = 1$, we may say "X_n converges in mean to X". When dealing with $p = 2$, we may say "X_n converges in mean-square to X". (Definition 5.3, Page 38).

Convergence (In Probability): A sequence of random variables X_n converges in probability to a random variable X, if for any $\varepsilon > 0$, $\lim_{n \to \infty} \mathbb{P}(|X_n - X| < \varepsilon) = 1$. We denote this $X_n \xrightarrow{\mathbb{P}} X$. To be precise, $X_n \xrightarrow{\mathbb{P}} X$ if $\lim_{n \to \infty} \mathbb{P}\left(\{\omega \in \Omega : |X_n(\omega) - X(\omega)| < \varepsilon\}\right) = 1$. (Definition 5.1, Page 38).

Convergence (Vaguely): A sequence of random variables converges vaguely if their distribution functions F_n converges to a monotone, right-continuous function $F : \mathbb{R} \to [0, 1]$, at all continuity points t of F. Note that F need not be a valid Cumulative Distribution Function (it's missing the condition that $\lim_{n \to \infty} F(x_n) = 1$, for example). (Definition 5.5, Page 39).

Convex: A function whose second derivative is everywhere positive. Equivalently, a function

$f : \mathbb{R} \to \mathbb{R}$ such that for all $t \in [0, 1]$ and for all $x, y \in \mathbb{R}$, we have $f\big(tx + (1 - t)y\big) \leq tf(x) + (1 - t)f(y)$.(Definition 4.1, Page 32).

Covariance: The covariance of random variables X and Y is $\mathrm{Cov}(X, Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}\big[\big(X - \mathbb{E}(X)\big)\big(Y - \mathbb{E}(Y)\big)\big]$. This is a generalization of variance, since $\mathbb{V}(X) = \mathrm{Cov}(X, X)$. When the covariance is zero, we say the random variables are uncorrelated. (Definition 3.8, Page 21).

Correlation: The correlation coefficient between random variables X and Y is $\rho(X, Y) = \frac{\mathrm{Cov}(X,Y)}{\sqrt{\mathbb{V}(X)\mathbb{V}(Y)}}$. Note $\rho \in [-1, 1]$. (Definition 3.9, Page 21).

Cumulative Density Function (CDF) Of A Random Variable: Where μ_X is the law (Definition 2.5, Page 13) of a random variable X, the CDF of X is the function $F_X : \mathbb{R} \to [0, 1]$ given by:

$$F_X(x) = \mu_X\big((-\infty, x]\big) = \mathbb{P}(X^{-1}(-\infty, x]) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}) = \mathbb{P}(X \leq x)$$

Here x (lowercase) denotes a generic element of the domain $\mathbb{R}$, and X (uppercase) denotes the random variable. So, we might have, e.g. $X(\omega) = x$. Compare this definition to Theorem 1.3, which doesn't require a random variable. This is really the exact same idea, it just maps the image of the random variable back to the sample space. (Definition 2.6, Page 15).

Distribution (Push-forward, Law) Of A Random Variable, μ_X: Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X : (\Omega, \mathcal{F}) \to (\mathbb{R}, \mathbb{B}(\mathbb{R}))$ be a random variable. Then the law of X (distributional measure, push-forward) is the function $\mu_X : \mathbb{B}(\mathbb{R}) \to [0, 1]$ given by $\mu_X(B) = \mathbb{P}(X^{-1}(B)) = \mathbb{P}\left(\{\omega \in \Omega : X(\omega) \in B\}\right)$. (Definition 2.5, Page 13).

Event Space $\mathcal{F}$: a σ-algebra consisting of unions, intersections, and complements from elements in the sample space. (Definition 1.3, Page 4).

Eventually Always: If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and if $\{A_i\}_{i=1}^{\infty}$ is a sequence of events in $\mathcal{F}$, then eventually, A_i will always occur if $\mathbb{P}\left(\bigcup_{n=1}^{\infty} \bigcap_{i \geq n} A_i\right) = 1$. Identifying union with "there exists" and intersection with "for all", this is saying "there exists an $n \in \mathbb{N}$ such that for all $i \geq n$, A_i occurs with probability 1", which is the definition of $\liminf$; $\mathbb{P}\left(\bigcup_{n=1}^{\infty} \bigcap_{i \geq n} A_i\right) = 1 \iff \mathbb{P}\left(\liminf_{i \to \infty}\right) = \mathbb{P}\left(\{\omega \in \Omega : \omega \in A_i \text{ for all large enough } i\}\right) = 1$. While not universal notation, we may abbreviate this to $\mathbb{P}(A_i \text{ e.a.}) = 1$. (Definition 7.3, Page 59)

Expectation: The expectation of a random variable X, denoted $\mathbb{E}(X)$, obeys

1. Linearity: for all random variables X, Y and constants c, $\mathbb{E}(cX + Y) = c\mathbb{E}(X) + \mathbb{E}(Y)$.

2. Non-negativity: if $\mathbb{P}(X > 0) = 1$ then $\mathbb{E}(X) \geq 0$.

We define the calculation for $\mathbb{E}$ in four stages in the theorem section below: first for simple random variables, then for bounded random variables, then for non-negative random variables, then for general random variables. At each stage, we calculate the expectation differently, and check that it agrees with previous calculations and meet the criteria for expectations above. While this is a useful exercise, actually computing expectations is usually easier done with the previous two pieces of machinery (Lebesgue and Riemann-Stieljes Integration). (Definition 3.6, Page 21).

Filtration: Where Ω is a sample space, where T is some fixed positive number, and where $\mathcal{F}_t$ is a sigma-algebra for all $t \in [0, T]$, then if $\mathcal{F}_s \subseteq \mathcal{F}_t$ whenever $s \leq t$, we say the collection $\mathcal{F}_t$ for $t \in [0, T]$ is a filtration. (Definition 1.12, Page 6).

Identically Distributed: Two random variables X and Y are identically distributed if $\mathbb{P}(X \in B) = \mathbb{P}(Y \in B)$ for all $B \in \mathbb{B}(\mathbb{R})$. Equivalently, we can check if $\mathbb{P}(X_i \leq t) = \mathbb{P}(X_1 \leq t)$ for all $t \in \mathbb{R}$ (equivalent by taking $B = (-\infty, t]$). Another equivalence is $\mathbb{E}\big(f(X_i)\big) = \mathbb{E}\big(f(X_1)\big)$ for all measurable f and all $i \geq 1$ provided the expectations exist (equivalent by using the push-forward formula in the above). (Definition 7.1, Page 59)

Independence (Of Events): A finite set of events $A_1, A_2, \ldots, A_n$ is **mutually independent** if for all $I \subseteq \{1, \ldots, n\}$ we have $\mathbb{P}\big(\bigcap_{i \in I} A_i\big) = \prod_{i \in I} \mathbb{P}(A_i)$. We say $A_1, A_2, \ldots, A_n$ are **pairwise independent** if for all $i \neq j$, $\mathbb{P}(A_i \cap A_j) = \mathbb{P}(A_i)\mathbb{P}(A_j)$. Note that mutual independence implies pairwise independence, but not vise-versa. Infinite collection of events are independent when any finite subset of the events are independent. (Definition 6.3, Page 52).

Independence (Of Random Variables): A finite collection of random variables $\{X_i\}_{i \in I}$ is independent if $\{\sigma(X_i)\}_{i \in I}$ (Definition 2.4, Page 13) is independent. For two random variables, this is equivalent to checking that $\mathbb{P}(X \leq t_1, Y \leq t_2) = F_X(t_1)F_Y(t_2)$.
(Definition 6.5, Page 52).

Independence (Of Sigma Algebras): A finite collection of sigma-algebras $\{\mathcal{F}_i\}_{i \in I}$ is independent if for every $A_i \in \mathcal{F}_i$, $\{A_i\}_{i \in I}$ is independent. Note that this specifically is not saying anything about events within any one sigma-algebra (i.e the events within a sigma-algebra may not be independent, see example), but rather is saying that selecting one event from each sigma-algebra results in independence. (Definition 6.4, Page 52).

Infinitely Often: If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space and if $\{A_i\}_{i=1}^{\infty}$ is a sequence of events in $\mathcal{F}$, then A_i occurs infinitely often if $\mathbb{P}\left(\bigcap_{n=1}^{\infty} \bigcup_{i \geq n} A_i\right) = 1$. Identifying intersection with "for all" and union with "there exists", this is saying "for all $n \in \mathbb{N}$, there exists an $i \geq n$ such that A_i occurs with probability 1", which is precisely the definition of $\limsup$; $\mathbb{P}\left(\bigcap_{n=1}^{\infty} \bigcup_{i \geq n} A_i\right) = 1 \iff \mathbb{P}\left(\limsup_{i \to \infty} A_i\right) = \mathbb{P}\left(\{\omega \in \Omega : \omega \in A_i \text{ for infinitely many } i\}\right) = 1$. We will often abbreviate this to $\mathbb{P}(A_i \text{ i.o.}) = 1$. (Definition 7.2, Page 59)

λ-system: A collection of sets $\mathcal{L}$ from a non-empty set Ω is a lambda-system provided $\mathcal{L}$ is closed under compliment and countable disjoint union. That is, $\mathcal{L}$ is a lambda-system if whenever $A_1, A_2, \cdots \in \mathcal{L}$ are disjoint, we have $\biguplus_{i=1}^{n} A_i \in \mathcal{L}$ and $A_1^c \in \mathcal{L}$.
(Definition 3.12, Page 22).

Lebesgue Integral: Recall the definition of the Riemann Integral for a differentiable function f. Partition the domain $a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$, let $M_k = \max_{x_{k-1} \leq x \leq x_k} f(x)$, $m_k = \min_{x_{k-1} \leq x \leq x_k} f(x)$, $\Pi = \{x_0, \ldots, x_n\}$, and $\|\Pi\| = \max_{1 \leq k \leq n}(x_k - x_{k-1})$, then see the Upper Riemann Sum $(RS_{\Pi^+}(f) = \sum_{k=1}^{n} M_k \cdot (x_k - x_{k-1})$ and Lower Riemann Sum $(RS_{\Pi^-}(f) = \sum_{k=1}^{n} m_k \cdot (x_k - x_{k-1})$ converge to the same value as $\|\Pi\|$ goes to zero, namely $\int_a^b f(x)\,dx$. Integrating in this way necessitates a natural ordering of the domain, which is a property that Ω, unlike $\mathbb{R}$, may not have. For that reason, instead of partitioning the domain, we partition the range in the Lebesgue Integral.

So assume for now that $0 \leq X(\omega) < \infty$. Partition the range of the random variable X as $0 = y_0 < y_1 < \ldots$ and as before denote $\Pi = \{y_0, \ldots, y_n\}$ and $\|\Pi\| = \max_{1 \leq k \leq n}(y_k - y_{k-1})$. Consider the event $A_k = \{\omega \in \Omega : y_k \leq X(\omega) \leq y_{k+1}\}$. Then the Lebesgue Integral is the limit of the Lower Lebesgue Sum as $\|\Pi\|$ goes to zero; $\lim_{\|\Pi\| \to 0} \sum_{k=1}^{\infty} y_k \mathbb{P}(A_k) = \int_{\Omega} X(\omega)\,d\mathbb{P}(\omega)$.

Define $X^+(\omega) = \max\{X(\omega), 0\}$ and $X^-(\omega) = \max\{-X(\omega), 0\}$ (in the future we may abbreviate maximum as $X \vee 0$). If $\mathbb{P}(\{\omega \in \Omega : X^+(\omega) = \infty\}) = \mathbb{P}(\{\omega \in \Omega : X^-(\omega) = \infty\}) = 0$, then we say X is **integrable** and have $\int_{\Omega} X(\omega)\,d\mathbb{P}(\omega) = \int_{\Omega} X^+(\omega)\,d\mathbb{P}(\omega) - \int_{\Omega} X^-(\omega)\,d\mathbb{P}(\omega)$. If both $\mathbb{P}(\{\omega \in \Omega : X^+(\omega) = \infty\}) > 0$ and $\mathbb{P}(\{\omega \in \Omega : X^-(\omega) = \infty\}) > 0$, then the Lebesgue Integral is undefined. If only one of the positive or negative parts of X takes values of infinity with non-zero probability, then the Lebesgue Integral is either ∞ (in the case where $0 = \mathbb{P}(\{\omega \in \Omega : X^-(\omega) = \infty\}) < \mathbb{P}(\{\omega \in \Omega : X^+(\omega) = \infty\})$) or $-\infty$ (in the other case).

We may be interested in integrating our random variable over a subset A of Ω. In such cases, we write $\int_A X(\omega)\,d\mathbb{P}(\omega) = \int_{\Omega} \mathbb{1}_A(\omega) X(\omega)\,d\mathbb{P}(\omega)$ where $\mathbb{1}_A(\omega)$ is the indicator function previously defined. Note that in all cases, we are integrating with respect to the probability measure in question, since the same event may have different probabilities under different measures. We define the expectation of X as it's Lebesgue Integral, and write $\mathbb{E}(X) = \int_{\Omega} X(\omega)\,d\mathbb{P}(\omega)$. As we'll see below, this is just one of many ways to define expectation. (Definition 3.1, Page 20).

L^p Space: Fix a probability triple $(\Omega, \mathcal{F}, \mathbb{P})$. The space of random variables with finite p-norm is denoted $L^p(\mathbb{P}) = \{X : \Omega \to \mathbb{R} : \|X\|_p < \infty\}$. Since $p \leq q \implies \|X\|_p \leq \|X\|_q$, $L^p(\mathbb{P}) \supseteq L^q(\mathbb{P})$ (the spaces get more exclusive as p grows). In that sense, the most exclusive space is L^∞. In the conditions for which X belong in L^∞, define $\|X\|_\infty = \inf\{L \geq 0 : \mathbb{P}(|X| \leq L) = 1\}$. (Definition 4.7, Page 32).

Measurable Function: A function $X : \Omega \to S$ between measure spaces $(\Omega, \mathcal{F})$ and $(S, \mathcal{S})$ is measurable if whenever $B \in \mathcal{S}$, $X^{-1}(B) = \{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F}$ (the inverse image of every measurable set is measurable). To emphasize that dependency on the respective sigma-algebras and to be precise, we might say "X is $(\mathcal{F}, \mathcal{S})$ measurable" (or just "X is $\mathcal{F}$-measurable" when $\mathcal{S}$ is understood) and write $X : (\Omega, \mathcal{F}) \to (S, \mathcal{S})$. (Definition 2.1, Page 12).

Measurable Space (X, Σ): A set X (for example a sample space) along with a sigma-algebra Σ on the set. mydefdef.measurespace.

Measure μ: In the context of a measure space (X, Σ), a measure $\mu : \Sigma \to \mathbb{R}$ is a function from the sigma-algebra to the real line such that $\mu(\emptyset) = 0$ and μ is countably additive, i.e. for all disjoint $A_1, A_2, \cdots \in \Sigma$, $\mu\left(\biguplus_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} \mu(A_i) \geq 0$. (Definition 1.5, Page 4)

Measure Space (X, Σ, μ): A measurable space along with a measure acting on the space. (Definition 1.6, Page 4).

(Central) Moment: The nth central moment of X is the value $\mathbb{E}\left[(X - \mathbb{E}(X))^n\right]$. (Definition 4.3, Page 32).

(Raw) Moment: The nth raw moment of a random variable X is the value $\mathbb{E}(X^n)$. (Definition 4.2, Page 32).

(Standard) Moment: The nth central moment of a random variable X is the value $\mathbb{E}\left[\left(\frac{X - \mathbb{E}(X)}{\sigma}\right)^n\right]$ (where $\sigma = \sqrt{\mathbb{V}(X)}$, the standard deviation). (Definition 4.4, Page 32).

Moment Generating Function: The moment generating function (MGF) for a random variable X is $M_X(t) = \mathbb{E}(e^{tX})$. The name of the function comes from the fact that the nth derivative of the MGF with respect to t, evaluated at 0, is the nth raw moment. (Definition 4.5, Page 32).

π-system: A collection of sets $\mathcal{P}$ from a non-empty set Ω is a pi-system provided $\mathcal{P}$ is closed under finite intersection. That is, $\mathcal{P}$ is a pi-system if whenever $A_1, A_2, \ldots A_n \in \mathcal{P}$ we have $\bigcap_{i=1}^{n} A_i \in \mathcal{P}$. (Definition 3.11, Page 22).

P-norm: The p norm of a random variable X is $\|X\|_p = \mathbb{E}\left(|X|^p\right)^{1/p}$. By Jensen's Inequality (Theorem 4.3, Page 33), if $p \leq q$, then $\|X\|_p \leq \|X\|_q$. (Definition 4.6, Page 32).

Probability Measure $\mathbb{P}$: A probability measure $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a function on a sigma-algebra $\mathcal{F}$ of the sample space Ω such that $\mathbb{P}(\Omega) = 1$ and $\mathbb{P}$ is countably additive, i.e. for disjoint A_i's, $\mathbb{P}\left(\biguplus_{i=1}^{\infty} A_i \right) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$. This is a specific case of a general measure. Note that

$\mathbb{P}(\emptyset) = 0$ as a consequence of the other two conditions. (Definition 1.7, Page 4).

Probability Space $(\Omega, \mathcal{F}, \mathbb{P})$: A triple consisting of a sample space Ω, an event space $\mathcal{F}$, and a probability measure $\mathbb{P}$ acting on the measurable space $(\Omega, \mathcal{F})$.
(Definition 1.8, Page 5).

Quantile Function: Where F_X is a valid CDF for a random variable X, the quantile function for F_X is the function $F_X^{-1} : [0, 1] \to \mathbb{R}$ given by $F_X^{-1}(u) = \inf\{t \in \mathbb{R} : F_X(t) \geq u\}$. We capture the intuition behind the quantile function at the cost of precision (since F_X may not have an inverse) when we use the notation F_X^{-1}. (Definition 2.7, Page 15).

Random Variable: A measurable function $X : \Omega \to \mathbb{R}$ between measure spaces $(\Omega, \mathcal{F})$ and $(\mathbb{R}, \mathbb{B}(\mathbb{R}))$ (it is just a specific case of a measurable function where the codomain is fixed). Note that the "randomness" from a random variable comes from the random experiment of choosing the $\omega \in \Omega$. Note further that to emphasize the fact that a random variable is a function, we may often write $X(\omega)$ (though X may be used for brevity).
(Definition 2.2, Page 12).

Random Variable (Bounded): A random variable X is bounded whenever there exists a $c \in \mathbb{R}$ such that for all $\omega \in \Omega$, $\mathbb{P}(|X(\omega)| < c) = 1$. (Definition 3.4, Page 21).

Random Variable (Non-negative): A random variable X is non-negative if for all $\omega \in \Omega$, $\mathbb{P}(X(\omega) \geq 0) = 1$. (Definition 3.5, Page 21).

Random Variable (Simple): A random variable X is simple whenever there are only finitely many values that X can take, that is, if there exists $x_1, x_2, \ldots, x_n \in \mathbb{R}$ such that for all $\omega \in \Omega$, $\mathbb{P}(X(\omega) \in \{x_1, x_2, \ldots, x_n\}) = 1$. (Definition 3.3, Page 21).

Random Vector: A measurable function $(X_1, X_2, \ldots, X_n) : (\Omega^n, \mathcal{F}^n) \to (\mathbb{R}^n, \mathbb{B}(\mathbb{R}^n))$. This is essentially just n random variables placed next to each other. (Definition 2.3, Page 13).

Resolved Sets: Suppose we are given a measure space $(\Omega, \mathcal{F})$ and an outcome $\omega \in \omega$. The sets in the event space $\mathcal{F}$ which are resolved by some level of information are those sets $A \in \mathcal{F}$ that either definitely contain or definitely don't contain ω. For this reason, it may be helpful to informally think of sigma-algebras as "information". (Definition 1.11, Page 6).

Riemann-Stieljes Integral: While the Lebesgue Integral allows for maximum generality (for the purposes of these notes), to actually compute expectations, it often suffices to use the integrals more familiar to us. The expectation of a function g of any random variable X with cumulative distribution function F_X is calculated as $\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x)\, dF_X(x)$. By definition, $F_X(x) = \int_{-\infty}^{x} f_X(t)\, dt$ where f_X is the density function of X. By the fundamental theorem of calculus, this means $dF_X(x) = f_X(x)\, dx$. In particular, $\mathbb{E}(g(X)) = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)\, dx$. (Definition 3.2, Page 21).

Sample Space Ω: any set containing outcomes (e.g. heads/tails, 1:6, etc.). (Definition 1.1, Page 4).

Semi-Algebra: A collection of sets $\mathcal{S}$ from a non-empty set Ω is a semi-algebra provided $\mathcal{S}$ is closed under intersection and each compliment is some finite disjoint union from $\mathcal{S}$ (even if the compliment is not in $\mathcal{S}$). That is, $\mathcal{S}$ is a semi-algebra if whenever $A_1, A_2, \ldots, A_n \in \mathcal{S}$, we have $A_i \bigcap A_j \in \mathcal{S}$ and $A_j^c = \biguplus_{i=1}^{n} A_i$.(Definition 3.13, Page 22).

σ-Algebra: A collection of sets $\mathcal{F}$ from a non-empty set Ω is a sigma-algebra provided $\mathcal{F}$ is closed under countable union and complements. That is, $\mathcal{F}$ is a sigma-algebra if whenever $A_1, A_2, \cdots \in \mathcal{F}$ we have $\bigcup_{i=1}^{\infty} A_i \in \mathcal{A}$ and $A_1^C \in \mathcal{F}$. (Definition 1.2, Page 4).

σ-Algebra (Generated By An Event A, $\sigma(A)$): It is trivial to see that the intersection of sigma algebras is itself a sigma-algebra. So we can define $\sigma(A)$ to be the intersection of all sigma-algebras containing A (in this sense, it is the smallest such set). Constructively, this means we start with the sets in A, and allow for countably many unions, intersections, and complements until we run out of ability to add more. (Definition 1.10, Page 5)

σ-algebra (Generated By A Random Variable X, $\sigma(X)$): Where X is a random variable, the sigma-algebra generated by X is $\sigma(X) = \{X^{-1}(B) : B \in \mathbb{B}(\mathbb{R})\}$. Unwinding the definition, this is $\{\{\omega \in \Omega : X(\omega) \in B\} \in \mathcal{F} : B \in \mathbb{B}(\mathbb{R})\}$. Informally, it is the minimally small sigma algebra that completely captures the information revealed by the values of the random variable. (Definition 2.4, Page 13).

Sigma-Algebra (Generated By Random Variables): The sigma-algebra generated by a sequence of random variables $\{X_i\}_{i \in I}$ is the smallest sigma-algebra containing $\sigma(X_i)$ for all i; $\sigma(\{X_i\}_{i \in I}) = \sigma\left(\bigcup_{i \in I} \sigma(X_i)\right)$. Here, $\sigma(X) = \{\{\omega \in \Omega : X(\omega) \in B\} : B \in \mathbb{B}(\mathbb{R})\}$. (Definition 6.1, Page 51).

Tail σ-Algebra: Where $\{X_i\}_{i \in I}$ is a sequence of random variables, the tail sigma algebra is denoted $\mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(X_{n+1}, X_{n+2}, \ldots)$. The idea is that the Tail σ-algebra is the collection of events whose occurrence is unaffected when finitely many of the random variables are changed. (Definition 6.2, Page 51).

Tightness: A sequence of random variables $\{X_n\}_{n \in N}$ are tight if for all $\varepsilon > 0$, there exists $a, b \in \mathbb{R}$ such that $\mathbb{P}(X_n \in [a, b]) \geq 1 - \varepsilon$. Equivalently, the sequence is tight if there exists $a, b \in \mathbb{R}$ such that $F_{X_n}(a) \leq \varepsilon$ and $F_{X_n}(b) \geq 1 - \varepsilon$. (Definition 5.6, Page 40).

Variance: The variance of a random variable X, denoted $\mathbb{V}(X)$, is the value $\mathbb{E}\left[(X - \mathbb{E}(X))^2\right] = \mathbb{E}(X^2) - \left(\mathbb{E}(X)\right)^2$. The square root of the variance is called the **standard deviation**; $\sqrt{\mathbb{V}(X)} = \sigma$. (Definition 3.7, Page 21).

10   Acknowledgments

These notes are the byproduct of my thinking on lectures given by Professor Erik Bates to students in the graduate probability sequence at North Carolina State University from Fall 2023 - Spring 2024. Some problems, proof ideas, and examples come from Rick Durrett's *Probability* textbook, Sourav Chatterjee's Probability class notes, and Steven Shreve's *Stochastic Calculus For Finance* textbook.