Advanced probability: a measure theoretic introduction

Gourab Ray

Last updated: June 7, 2025

These lecture notes were written to serve as the course material for Math 451/ Math 555 at University of Victoria. If you find any typos/errors I would be grateful if you let me know.

Contents

1 Measure theory

1.1 The probability space

In the first level probability course, we learnt about an elementary version of the probability space. This consisted of a *sample space* Ω, which is the set of all possible outcome of an experiment. Then we defined an event to be a subset of the sample space. We also defined probability $\mathbb{P} : \Omega \mapsto [0, 1]$ as a function satisfying certain rules. The issue with this setup is that it is too naive to work in general, in particular for continuous sample space (e.g. when we measure lifetime of a bulb). If you remember, certain technicalities were swept under the rug. For example, how do we know that a probability measure satisfying the properties we outline exist? Where did the probability density functions (pdf) come from? Probably there were many more such questions, which are left unanswered.

The goal now is to construct a rigorous version of the probability space. It consists of three objects.

- Sample space
- sigma-algebra (or σ-algebra)
- Probability measure.

The *sample space* Ω is simply a set, not necessarily a subset of $\mathbb{R}$. (We think of it as an abstract set with no additional structure, and with no relation to any experiment for the moment). The "space of events" is going to be replaced by an object called a σ-algebra which we now define.

Definition 1.1 (σ-algebra)**.** *A* collection of subsets $\mathcal{F}$ *of* Ω *is called a* σ*-algebra if it satisfies the following conditions,*

- $\Omega \in \mathcal{F}$
- *If* $A \in \mathcal{F}$ *then* $A^c \in \mathcal{F}$
- *If* $\{A_n\}_{n\geq 1}$ *is a* countable *collection in* $\mathcal{F}$*, then* $\cup_{n=1}^{\infty} A_n \in \mathcal{F}$*.*

The sets in $\mathcal{F}$ *are called* ***measurable sets****.*

Exercise 1.1. Let $\mathcal{F}$ be a σ-algebra. Suppose $\{A_i\}_{i\geq 1}$ be a countable collection taken from $\mathcal{F}$. Which of the following is in $\mathcal{F}$ as well?

a. $A_1 \cap A_2$.

b. $\cap_{i=1}^{\infty} A_i$.

c. $A_1 \setminus A_2$.

d. A_1^c (complement of A_1).

e. $\emptyset$ (the emptyset).

f. $\cap_{i=1}^{\infty} A_i^c$, $\cup_{i=1}^{\infty} A_i^c$

The collection $\mathcal{F} := \{A : A \subset \Omega\}$ is called the *power set* of Ω. This is clearly a σ-algebra (Check!). In fact, this is the largest [1] σ-algebra possible. For a concrete example, think of $\Omega = \mathbb{N}$ and $\mathcal{F}$ to be all the subsets of $\mathbb{N}$.

On the other hand, what is the smallest σ-algebra? It is easy to see that the candidate for this is $\{\emptyset, \Omega\}$. Let's venture a bit more: take $A \subset \Omega$. What is the smallest σ-algebra containing a set A? Notice that this σ-algebra must contain $A, A^c, \Omega, \emptyset$. In fact a bit of inspection shows that $\mathcal{F} = \{A, A^c, \Omega, \emptyset\}$ is the smallest σ-algebra containing A (Check!).

Having gotten used to σ-algebras a bit, let us consider what happens if we take unions and intersections of σ-algebras themselves. It is very easy to see that taking unions of two σ-algebra does not necessarily produce a σ-algebra. Take for example $\Omega = \{1, 2, \ldots, 6\}$. $A = \{1\}$, $B = \{2\}$,

$$\mathcal{F}_1 = \{\emptyset, A, A^c, \Omega\}, \qquad \mathcal{F}_2 = \{\emptyset, B, B^c, \Omega\}, \qquad \mathcal{F}_1 \cup \mathcal{F}_2 = \{\emptyset, A, A^c, B, B^c, \Omega\}$$

Since $A \cup B = \{1, 2\} \notin \mathcal{F}_1 \cup \mathcal{F}_2$, it is not a σ-algebra.

Lemma 1.2. *Verify that if $\{\mathcal{F}_i\}_{i \in \mathcal{I}}$ is a collection of σ-algebras (not necessarily countable!) then*

$$\mathcal{F} := \cap_{i \in \mathcal{I}} \mathcal{F}_i := \{A : A \in \mathcal{F}_i \quad \forall i \in \mathcal{I}\}$$

is also a σ-algebra.

Proof. We verify the three criterions which make a σ-algebra. Note Ω is in all of $\mathcal{F}_i$ hence $\Omega \in \mathcal{F}$. Suppose $A \in \mathcal{F}$. Then $A \in \mathcal{F}_i$ for all i. Therefore, $A^c \in \mathcal{F}_i$ for all i since $\mathcal{F}_i$ is a σ-algebra. Thus $A^c \in \mathcal{F}$. Similarly, if $A_n \in \mathcal{F}_i$ for all n then $A := \cup_{n \geq 1} A_n \in \mathcal{F}_i$ since $\mathcal{F}_i$ is a σ-algebra. Therefore, $A \in \mathcal{F}$. □

Proposition 1.1 (Smallest σ-algebra exists)**.** *Consider a collection*

$$\mathcal{A} := \{A_i : A_i \subset \Omega, i \in \mathcal{I}\}.$$

Then there exists a smallest σ-algebra containing all the sets in $\mathcal{A}$.

Proof. Let Σ denote the collection of *all* σ-algebras containing $\mathcal{A}$. The set Σ is nonempty, since clearly the power set 2^{Ω} contains all the sets in $\mathcal{A}$. Then $\mathcal{G} := \cap_{\mathcal{F} \in \Sigma} \mathcal{F}$ is our required σ-algebra (we employed Lemma 1.2 to ensure that this is a σ-algebra). □

[1] largest in the sense of the 'number' of sets in them. To be more precise, $\mathcal{F}_1$ is smaller than $\mathcal{F}_2$ if $\mathcal{F}_1 \subseteq \mathcal{F}_2$

Notation: The smallest σ-algebra containing a collection of sets $\mathcal{A}$ is denoted by $\sigma(\mathcal{A})$.

Example 1.3. Suppose the countable collection $\mathcal{A} := \{A_i\}_{i=1}^{\infty}$ partition Ω, i.e., $\cup A_i = \Omega$ and $A_i \cap A_j = \emptyset$ if $i \neq j$. Then $\sigma(\mathcal{A})$ is described by all possible unions of the elements of $\mathcal{A}$. That is,

$$\sigma(\mathcal{A}) = \{B : B = \cup_{i \in S \subseteq \mathbb{N}} A_i\}$$

Prove this by

- first checking that $\{B : B = \cup_{i \in S \subseteq \mathbb{N}} A_i\}$ satisfies all the conditions in Definition 1.1,
- and then arguing that any sigma algebra containing $\mathcal{A}$ must contain $\{B : B = \cup_{i \in S \subseteq \mathbb{N}} A_i\}$.

Exercise 1.4. Find the smallest σ-algebra containing two sets A, B, $A \neq B$.

Exercise 1.5. Let $\Omega = \mathbb{N}$. Let $\mathcal{A} = \{S \subset \mathbb{N} : |S| < \infty\}$. Argue that this is NOT a σ-algebra. Find $\sigma(\mathcal{A})$. [2]

Example 1.6. Let $S = \{H, T\}$ be a set of two elements. Let

$$\Omega = \bigotimes_{i=1}^{\infty} S = \{(x_1, x_2, \dots,) : x_i \in \{H, T\}\}.$$

We can think of Ω as the set of all possible outcomes when a coin is flipped (countably) infinitely many times. Let $\mathcal{G}_n = \{A \times \Omega : A \subset S^n\}$. In words: $\mathcal{G}_n$ is all possible outcomes where the first n outcomes are constrained to be in the some set $A \subset S^n$. Prove that $\mathcal{G}_n$ is a σ-algebra by arguing that it satisfies all the conditions of Definition 1.1. Also convince yourself that

$$\mathcal{G}_1 \subset \mathcal{G}_2 \subset \mathcal{G}_3 \dots$$

Here one can define $\mathcal{G}_\infty$ to be the smallest σ-algebra containing all of $\mathcal{G}_n$. Can you find an event which is in $\mathcal{G}_\infty$ but not in $\mathcal{G}_n$ for any n?

Exercise 1.7 (Thought experiment)**.** Let $\mathcal{B} := \{(a, b), -\infty < a < b < \infty \in \mathbb{R}\}$ denote the set of all finite intervals. Argue this is not a σ-algebra (easy). Think about how $\sigma(\mathcal{B})$ might look like? You should convince yourself that there is no reasonably simple way to 'describe' any such set in English.

We now define a very important σ-algebra.

Definition 1.2 (Borel σ-algebra on $\mathbb{R}$)**.** *Borel σ-algebra on the real line (i.e. when $\Omega = \mathbb{R}$, the real line), is defined as $\sigma(\tau)$ where*

$$\tau = \{(a, b) - \infty < a < b < \infty\}$$

*That is the Borel σ-algebra on the real line is the smallest σ-algebra containing all the open intervals. It is denoted by $\mathcal{B}(\mathbb{R})$ and it's elements are called **Borel sets**.*

[2] Ans: Power set of $\mathbb{N}$.

Borel sets are named after Emile Borel. See https://en.wikipedia.org/wiki/%C3%89mile_Borel#/media/File:Emile_Borel-1932.jpg to learn more about this illustrious gentleman.

Figure 1: Prof. Emile Borel (taken from Wiki).

Let us think a bit on what kind of sets are Borel sets.

Exercise 1.8. Show that $\mathcal{B}(\mathbb{R})$ contains

- All closed intervals of the form $[a, b]$.
- All singletons $\{a\}$. All countable union of singletons.
- The cantor set. See https://en.wikipedia.org/wiki/Cantor_set.

In fact, it is quite hard to *think* of a set which is not Borel! But there exists such sets and $\mathcal{B}(\mathbb{R})$ is not equal to the power set 2^{Ω}. Check Wikipedia https://en.wikipedia.org/wiki/Borel_set or your favourite website in the internet to find such horrible (or beautiful, depending on the taste) beasts.

What kind of sets are not Borel? Is the set of Borel sets the power set? Remarkably, there are such sets and it leads to weird paradoxes. We will come back to it briefly in the next section.

Exercise 1.9. Let $\mathcal{I}$ be an interval. Let $\mathcal{B}(\mathcal{I}) := \{A \cap I : A \in \mathcal{B}(\mathbb{R})\}$. Prove that $\mathcal{B}(\mathcal{I})$ is a σ-algebra. This σ algebra is called the Borel σ-algebra on $\mathcal{I}$.

Generalization The definition of Borel sets can be generalized to arbitrary topological space. If (Ω, τ) is a topological space with τ being the open sets, then $\mathcal{B}(\Omega)$ denotes the smallest σ-algebra containing all the open sets in τ. For example, for any **metric space** (M, d_M), we can take the topology to be one defined by the metric (i.e. open sets τ are generated by $B(x, r) := \{y \in M : d_M(x, y) < r\}$ for $x \in M, r > 0$) and then the Borel sigma algebra can be taken to be $\sigma(\tau)$. You can imagine how this expands the scope of defining probability measures.

Exercise 1.10 (Borel sets in $\mathbb{R}^n$)**.** Use the above idea to define Borel σ-algebra in higher dimensions $\{\mathbb{R}^n\}_{n\geq 2}$. This σ-algebra is denoted $\mathcal{B}(\mathbb{R}^n)$.

1.2 Measures and random variables

Definition 1.3 (Measure (space)/ probability (space))**.** *Suppose* $(\Omega, \mathcal{F})$ *is a measure space. A* measure μ *is a function* $\mu : \mathcal{F} \mapsto [0, \infty]$ *which satisfies:*

- $\mu(\emptyset) = 0$,
- *(Countable Additivity) For any collection* $\{A_n\}_{n\geq 1}$ *of disjoint sets in* $\mathcal{F}$,

$$\mu(\cup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n).$$

The triplet $(\Omega, \mathcal{F}, \mu)$ *is sometimes called a* ***measure space*** *or* ***measurable space****. If* μ *additionally satisfies* $\mu(\Omega) = 1$*,* μ *is called a* ***probability measure*** *and the triplet* $(\Omega, \mathcal{F}, \mu)$ *is called a* ***probability space****.*

Here are a few examples. Sometimes μ is called a **distribution.** We will come back to it when we talk about random variables.

i. Let $\Omega = \mathbb{N}$ and $\mathcal{F} = 2^{\Omega}$. Let $\mu(\{i\}) = p_i$ where $p_i > 0$ for all $i \in \mathbb{N}$ and $\sum_{i\in\mathbb{N}} p_i = 1$. Finally, define $\mu(A) = \sum_{i\in A} p_i$. It is a simple exercise to check that μ is a probability measure. Another name for the function p_i is a **probability mass function**. Some special cases: $\mu(i) = (1-p)p^{i-1}$; $p \in [0,1]$ is the probability mass function of a geometric random variable . If $p_i = \frac{e^{-\lambda}\lambda^i}{i!}$ for some $\lambda > 0$ is the pmf of Poisson(λ).

ii. Take $\Omega = \mathbb{R}$, $\mathcal{F} = \mathcal{B}(\mathbb{R})$ (Borel sigma algebra) and a function $f : \mathbb{R} \to [0, \infty)$ satisfying $\int_{-\infty}^{\infty} f(x)dx = 1$. Define $\mu(A) = \int_A f(x)dx$ for any $A \in \mathcal{B}(\mathbb{R})$. This integral may not make sense if you are not familiar with Lebesgue integration. We will get to it later. But at this point it is enough to note that this defines a probability measure as soon as Lebesgue integral satisfies certain obvious properties of integration. This measure corresponds to the so called continuous distributions with f being the probability density function. We will come back to it later.

iii. Let $\Omega = \mathbb{N}$ and $\mathcal{F} = 2^{\Omega}$ and $\mu(\{i\}) = 1$. This corresponds to the so-called **counting measure**. Indeed $\mu(A) = |A|$ where $|\cdot|$ is the cardinality.

iv. Take $\Omega = (a, b)$, $\mathcal{F} = \mathcal{B}((a, b))$ (Borel sigma algebra on the interval (a, b), recall Exercise 1.9). For any $A \in \mathcal{B}((a, b))$ define $\lambda(A) = \int_A dx$. This measure is called the **Lebesgue measure** on (a, b) and corresponds to the 'length' of an interval, our usual notion of a 'measure'.

A simple example when μ is not a measure is the following: take $\Omega = \{0, 1\}$ and $\mathcal{F}$ to its power set. Let $\mu(\{0\}) = \mu(\{1\}) = \frac{1}{2}$ and $\mu(\{0, 1\}) = 2$. Clearly this is not a measure as $\frac{1}{2} + \frac{1}{2} \neq 2$.

Exercise 1.11. Show that if $A \subset B$, and $\mu(B) < \infty$ then $\mu(B \setminus A) = \mu(B) - \mu(A)$. Conclude that $\mu(A) \leq \mu(B)$.

We will mostly deal with probability spaces in what is to come. Please note in the definition that measures only take non-negative values. Usually a probability measure will be denoted by $\mathbb{P}$.

Example 1.12. Here are two ways to explicitly construct a probability space modelling a fair coin toss. Take Ω to be any set (e.g. [0,1]). Take $A \subset \Omega$ such that $A \neq \Omega$. Take $\mathcal{F} = \{\emptyset, A, A^c, \Omega\}$. Define the probability measure $\mathbb{P}$ such that $\mathbb{P}(A) = 1/2$, $\mathbb{P}(\emptyset) = 0 = 1 - \mathbb{P}(\Omega)$, and $\mathbb{P}(A^c) = 1/2$. Check that this definition satisfies all the conditions of the definition above. On the other hand, we can define $\tilde{\Omega} = \{H, T\}$, $\mathcal{F} = 2^{\tilde{\Omega}}$ and $\mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = \frac{1}{2}$, $\mathbb{P}(\{H, T\}) = 1$ and $\mathbb{P}(\emptyset) = 0$.

Remark 1.13. Example 1.12 illustrates an important concept. Whenever we talk of an experiment with a random outcome (e.g. lifetime of a bulb, rolling a die, counting the number of days it rained in July), there is always a probability space in the background. It is usually kept implicit. It is also important to note that one could work with different probability spaces for the same random experiment.

Example 1.14 (Non measurable set.)**.** Let us argue that there is a set which is not Borel. To do that imagine there is a measure μ which measures 'length' of subsets of $[0, 1]$. Roughly, we want to define a measure μ on subsets of $[0, 1]$ such that $\mu((a, b)) = b - a$ for all $0 \leq a < b \leq 1$. We will see later that Kolmogorov ensured (via the so-called Kolmogorov extension theorem) that μ extends to all Borel sets, which recall is the smallest σ-algebra containing the intervals.

Now let us define a set which is not Borel. Let $x \sim y$ if $x - y \in \mathbb{Q}$. It is not too hard to check that this is an equivalence relation, which means that we can divide $[0, 1]$ into equivalence classes. Let $\mathcal{N}$ be the set obtained by picking exactly one element from each equivalence class (this is a non-trivial step, and employs the Axiom of choice). We claim $\mathcal{N}$ cannot be Borel. To prove this, note that for different rationals q, q', $(\mathcal{N} + q) \mod 1$ is disjoint from $\mathcal{N} + q' \mod 1$ (Exercise: check this.) Also $\cup_{q \in \mathbb{Q}, 0 < q < 1} (\mathcal{N} + q) \mod 1 = [0, 1]$ (exercise: check this as well). Furthermore, it is not hard to convince yourself that $\mu((\mathcal{N} + q) \mod 1) = \mu(\mathcal{N})$ for any q since length does not change by translating (this needs a serious proof, which we will learn later). But this is a contradiction as this would mean

$$1 = \mu([0, 1]) = \mu(\cup_{q \in \mathbb{Q}, 0 < q < 1} (\mathcal{N} + q) \mod 1) = \sum_{q \in \mathbb{Q}, 0 < q < 1} \mu((\mathcal{N} + q) \mod 1) = \sum_{q \in \mathbb{Q}, 0 < q < 1} \mu(\mathcal{N}),$$

which is impossible as either $\mu(\mathcal{N}) = 0$, in which case we get $1 = 0$, or $\mu(\mathcal{N}) > 0$ in which case we get $1 = \infty$.

Consequently, $\mu(\mathcal{N})$ is undefinable. On the other hand $\mu(A)$ is a real number for every Borel set A. Consequently $\mathcal{N}$ cannot be Borel.

We now state and prove a useful lemma about the 'continuity' properties of a measure.

Lemma 1.15. • *If $A_1 \subseteq A_2 \subseteq \ldots$ be a sequence of measurable sets (sometimes called an **increasing sequence**), then $\mu(\cup_{n=1}^{\infty} A_n) = \lim_{n\to\infty} \mu(A_n)$ (in particular the limit exists). This is sometimes called continuity of the measure from below.* [3]

- *If $A_1 \supseteq A_2 \supseteq \ldots$ be a sequence of measurable sets, with $\mu(A_1) < \infty$. Then $\mu(\cap_{n\geq 1} A_n) = \lim_{n\to\infty} \mu(A_n)$ (in particular the limit exists). This is sometimes called continuity of the measure from above.* [4].

- *(Subadditivity/ Union bound) Show that for any sequence $\{A_n\}_{n\geq 1}$ of measurable sets (not necessarily disjoint)*

$$\mu(\cup_{n\geq 1} A_n) \leq \sum_{n=1}^{\infty} \mu(A_n) \tag{1.1}$$

Proof. For the first part, we use a standard trick called **disjointification**. To that end, define $B_n = A_n \setminus (A_1 \cup A_2 \cup \ldots \cup A_{n-1})$. Notice $B_n \subset A_n$ and $B_n \cap A_i = \emptyset$ for all $1 \leq i \leq n-1$. Therefore, the collection $\{B_n\}_{n\geq 1}$ are *mutually disjoint* that is $B_i \cap B_j = \emptyset$ for all $i \neq j$. Also note

$$\cup_{k=1}^{n} B_k = \cup_{k=1}^{n} A_k$$

Since A_k is increasing, that is, $A_k \subseteq A_{k+1}$, we have

$$\cup_{k=1}^{n} A_k = A_n$$

In particular,

$$A_n \uparrow \cup_{k=1}^{\infty} A_k = \cup_{k=1}^{\infty} B_k$$

Thus using the additivity property of measures

$$\mu(A_n) \overset{A_k \text{ increasing}}{=} \mu(\cup_{k=1}^{n} A_k) = \mu(\cup_{k=1}^{n} B_k) \overset{\text{additivity}}{=} \sum_{k=1}^{n} \mu(B_k)$$

Since μ of any set is non-negative, $\sum_{k=1}^{n} \mu(B_k)$ is nondecreasing in n and in particular, the limit as $n \to \infty$ exists. Therefore, the limit of $\mu(A_n)$ also exists and

$$\lim_{n\to\infty} \mu(A_n) = \lim_{n\to\infty} \sum_{k=1}^{n} \mu(B_k) = \sum_{k=1}^{\infty} \mu(B_k) \overset{\text{countable additivity}}{=} \mu(\cup_{k=1}^{\infty} B_k) = \mu(\cup_{k=1}^{\infty} A_k)$$

Note that both sides could be ∞ in this equation. Indeed, we will see later that this can indeed be the case if $\mu(\Omega) = \infty$. [5]

[3] Can both sides of the equation be ∞?

[4] We need the assumption $\mu(A_1) < \infty$, look at the example below

[5] For example, special and very important measure called 'Lebesgue measure' λ(Section 1.4) which basically measures length. For now it will be enough to know that for any interval, $\lambda((a,b)) = b - a$ for any interval $(a,b) \subset \mathbb{R}$ (including the case $a = -\infty$ or $b = \infty$). Note that

$$\infty = \lambda((0,\infty)) = \lambda(\cup_{n=1}^{\infty}(0,n)) \overset{\text{first part of lemma}}{=} \lim_{n\to\infty} \lambda((0,n)) = \lim_{n\to\infty} n = \infty$$

For the second part, define $C_i = A_1 \setminus A_i$. Since $A_1 \supseteq A_2 \ldots$, we have $C_1 \subseteq C_2 \subseteq \ldots$. Thus by previous part,

$$\lim_{n\to\infty} \mu(A_1 \setminus A_n) = \lim_{n\to\infty} \mu(C_n) = \mu(\cup_{i\geq 1} C_i) = \mu(A_1 \setminus (\cap_{i\geq 1} A_i))$$

By Exercise 1.11, we have

$$\lim_{n\to\infty} (\mu(A_1) - \mu(A_n)) = \mu(A_1) - \mu(\cap_{i\geq 1} A_i).$$

(Note here we crucially used $\mu(A_1) < \infty$). Cancelling $\mu(A_1)$, we are done.
[6]

The third item is left as an exercise. □

1.3 Random variables

Definition 1.4 (Random variable)**.** *Let $(\Omega, \mathcal{F})$ be a measurable space. A random variable is a function $X : \Omega \mapsto \mathbb{R}$ such that for any open interval I,*

$$X^{-1}(I) := \{\omega \in \Omega : X(\omega) \in I\} \in \mathcal{F}.$$

Lemma 1.16. *If X is a random variable, then $X^{-1}(B) \in \mathcal{F}$ for any Borel set B.*

Proof. One can show this fact using the following standard (and very useful!) technique in measure theory. Let $\tilde{\mathcal{B}} = \{A \subset \mathbb{R} : X^{-1}(A) \in \mathcal{F}\}$. Clearly $\tilde{\mathcal{B}}$ contains all the open intervals. Now we wish to argue that $\tilde{\mathcal{B}}$ is a σ-algebra. If we can do that, then this finishes the argument since $\mathcal{B}(\mathbb{R})$ is the smallest σ-algebra containing the open intervals which in turn implicates $\tilde{\mathcal{B}} \supset \mathcal{B}(\mathbb{R})$, as desired.

Now we finish this argument, i.e., show $\tilde{\mathcal{B}}$ is a σ-algebra. To that end, we need to verify the following.

- Clearly $\mathbb{R} \in \tilde{\mathcal{B}}$ as $\Omega = X^{-1}(\mathbb{R})$.
- If $A \in \tilde{\mathcal{B}}$ then $X^{-1}(A) \in \mathcal{F}$ which implies $X^{-1}(A^c) = X^{-1}(\mathbb{R}) \setminus X^{-1}(A) \in \tilde{\mathcal{F}}$ as well. Thus $A^c \in \tilde{B}$.
- If $\{A_n\}_{n\geq 1}$ is a *countable* collection in $\tilde{\mathcal{B}}$, then $X^{-1}(\cup_{n\geq 1} A_n) = \cup_{n\geq 1} X^{-1}(A_n)$[7] $\in \mathcal{F}$ as well. Thus $\cup_{n\geq 1} A_n \in \tilde{B}$.

This completes the proof □

[6] The assumption $\mu(A_1) < \infty$ is important. Indeed, if we take $A_n = (n, \infty)$ and λ to be Lebesgue measure (yet to be defined), then $\lambda(\cap_{n=1}^{\infty} A_n) = \lambda(\emptyset) = 0$ by the first property of measures. On the other hand $\lim_{n\to\infty} \lambda((n, \infty)) = \lim_{n\to\infty} \infty = \infty$. Thus in this case, the second part of the lemma is false.

[7] verify!

Definition 1.5. *We denote by $\sigma(X)$ the smallest σ-algebra which makes X measurable. In other words,*

$$\sigma(X) = \cap\{\mathcal{F} : \mathcal{F} \text{ is a } \sigma\text{-algebra}, X^{-1}(E) \in \mathcal{F} \text{ for all } E \in \mathcal{B}(\mathbb{R})\}.$$

In other words,

$$\sigma(X) = \sigma(\{X^{-1}((-\infty, x]), x \in \mathbb{R}\}).$$

Exercise 1.17. Prove that

$$\sigma(X) = \{X^{-1}(A) : A \in \mathcal{B}(\mathbb{R})\}.$$

Hint: Use the same technique as in Lemma 1.16.

Example 1.18 (Modelling a coin toss)**.** Assume $\mathcal{F} = (\emptyset, \Omega)$. Consider $X : \Omega \mapsto \mathbb{R}$ so that $X(\omega) = 1$ if $\omega \in A$ and $X(\omega) = 0$ if $\omega \in A^c$. Then X is NOT a ($\mathcal{F}$- $\mathcal{B}(\mathbb{R})$) random variable, since $X^{-1}(\{1\}) = A \notin \mathcal{F}$.

However if we make $\mathcal{F} = (\emptyset, A, A^c, \Omega)$ then X becomes a random variable. Indeed, for any $U \in \mathcal{B}$

$$X^{-1}(U) = \begin{cases} \emptyset \text{ if } 1 \notin U \text{ and } 0 \notin U \\ A \text{ if } 1 \in U \text{ and } 0 \notin U \\ A^c \text{ if } 1 \notin U \text{ and } 0 \in U \\ \Omega \text{ if } 1 \in U \text{ and } 0 \in U. \end{cases} \tag{1.2}$$

all of which are in $\mathcal{B}$. In fact $\sigma(X) = \mathcal{F}$ in this case.

Example 1.19 (Modelling a discrete random variable)**.** Let $\mathcal{S} = \{s_1, s_2, \ldots\}$ be a countable subset of $\mathbb{R}$. Let $p : \mathcal{S} \to \mathbb{R}$ is a function such that $\sum_{x \in \mathcal{S}} p_x = 1$. Let $F_i = \sum_{j \le i} p_{s_j}$ with $F_0 = 0$. Define $A_i = [F_i, F_i + p_{s_{i+1}})$. Note that A_i forms a partition of $[0, 1]$.

Now let us construct a random variable with pmf p. Define $X : [0, 1] \to \mathbb{R}$ to be the following piecewise constant function: $X_t = s_{i+1}$ in $[F_i, F_i + p_{s_{i+1}})$ for $i \ge 0$. With this definition it is not too hard to see that

$$\sigma(X) = \{B : B := \cup_{j \in S \subseteq \mathbb{N}} A_j\}.$$

Indeed $X^{-1}(\{s_{i+1}\}) = A_i$ for all $i \ge 0$. Thus all the sets of the form $\cup_{j \in S \subseteq \mathbb{N}} A_j$ must be in $\sigma(X)$. Furthermore, $\sigma(X)$ is a σ-algebra as we already saw in Example 1.3.

Now let us show that the space of random variables is closed under arithmetic operations. To that end, let X, Y be random variables defined on $(\Omega, \mathcal{F})$. Then

- $X + Y$ is a random variable. To see this note that it is enough to show that for all $x \in \mathbb{R}$

$$\{X + Y \le x\} \in \mathcal{F}$$

which is equivalent to showing

$$\{X+Y \leq x\}^c = \{X+Y > x\} \in \mathcal{F}$$

which is equivalent to showing

$$\cup_{r\in\mathbb{Q}}\{X > r\} \cap \{Y > x-r\}$$

where $\mathbb{Q}$ is the set of rationals. The last statement is clear from definition.

- $X - Y$ is a random variable. (Similar proof, exercise)
- X^2 is a random variable. To see this, note that

$$\{\omega : X^2(\omega) > x\} = \begin{cases} \Omega \text{ if } x < 0 \\ X \in \{\sqrt{x}, \infty\} \cup (-\infty, -\sqrt{x}) \text{ if } x > 0. \end{cases}$$

  both of which is in $\mathcal{F}$.

- cX is a random variable where $c \in \mathbb{R}$. This is easy, left as exercise.
- XY is a random variable. To see this write

$$4XY = (X+Y)^2 - (X-Y)^2$$

  and appeal to the previous items.

- $|X|$ is a random variable. To see this we introduce a very important random variable called **indicator random variable**. For any $A \in \mathcal{F}$, define

$$1_A(\omega) = \begin{cases} 1 \text{ if } \omega \in A \\ 0 \text{ if } \omega \notin A. \end{cases} \tag{1.3}$$

  Now note

$$|X| = X1_{\{\omega:X(\omega)>0\}} - X1_{\{\omega:X(\omega)<0\}}$$

  and appeal to the previous items.

- $\max\{X, Y\}$ is a random variable. To see this, note

$$\max\{X, Y\} = \frac{|X+Y| + |X-Y|}{2}$$

  and appeal to the previous items.

- $\min\{X, Y\}$ is a random variable. (Exercise).
- $\max\{X_1, \ldots, X_n\}$ is a random variable and $\min\{X_1, \ldots, X_n\}$ is a random variable.

Terminology: Whenever we write a relation between random variables without specifying anything else, we mean it is valid "pointwise" which is an euphemism for saying that it is valid for "all ω". For example, $X \leq Y$ means that $X(\omega) \leq Y(\omega)$ for all $\omega \in \Omega$. It is usually the case that probabilists do not care for sets of measure 0 (although there are very good reasons in other branches of mathematics to care about them!), and are happy for a relation to be valid "up to measure 0 sets". Usually this is denoted by $X \leq Y$ **a.e.** or **a.s.** (**almost everywhere** or **almost surely**). For example "$X \leq Y$ a.s." means that if $\Omega_0 = \{\omega \in \Omega : X(\omega) > Y(\omega)\}$ then $\mu(\Omega_0) = 0$.

Let us try to generalize the above. Sometimes, a random variable is also used to denote arbitrary maps $X : \Omega_1 \mapsto \Omega_2$. In that case, we need to specify two measurable spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ and it must be the case that for all $A \in \mathcal{F}_2$, $X^{-1}(A) \in \mathcal{F}_1$. In that case, sometimes we specify X is $\mathcal{F}_1$-$\mathcal{F}_2$ **measurable** to make sure there is no confusion with the σ-algebras involved. Usually in this course, $(\Omega_2, \mathcal{F}_2)$ will be $(\mathbb{R}, \mathcal{B})$ (or at the most $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$).

We assume $(\Omega_2, \mathcal{F}_2)$ is $(\mathbb{R}, \mathcal{B})$ if nothing else is specified.

Fact: (Obvious) If $\mathcal{F}_1 \subset \mathcal{F}_2$, then if X is $\mathcal{F}_1$ - $\mathcal{B}$ measurable then X is $\mathcal{F}_2$ - $\mathcal{B}$ measurable. If $\mathcal{F}_1, \mathcal{F}_2$ are both $\mathbb{R}$, then we shall usually denote the random variables using small letters like f, g just like we did in Calculus. We need a slightly general definition of real valued measurable functions.

Definition 1.6. *Suppose $I, J \subset \mathbb{R}$ are intervals (open or closed). A function $g : I \mapsto J$ is $\mathcal{B}(I)$-$\mathcal{B}(J)$ measurable if for all $A \in \mathcal{B}(J)$, $\{x : g(x) \in A\} \in \mathcal{B}(I)$. Again, this is equivalent to saying that*

$$\{x : g(x) < t\} \in \mathcal{B}(I), \qquad \forall t \in J.$$

Lemma 1.20. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $f : \mathbb{R} \mapsto \mathbb{R}$ be $\mathcal{B}(\mathbb{R})$-$\mathcal{B}(\mathbb{R})$ measurable. Let X be a random variable. Then $f(X)$ is a random variable.*

Proof. This simply follows from definition. Let $Y = f(X)$. For any $A \in \mathcal{B}(\mathbb{R})$, $Y^{-1}(A) = X^{-1}(f^{-1}(A))$. Since A is Borel, $f^{-1}(A)$ is Borel and $X^{-1}(f^{-1}(A))$ is also Borel as X is a random variable. □

Recall that a set U in $\mathbb{R}$ is open if for any $x \in U$, there exists an interval $x \in (a, b) \subseteq U$.

Lemma 1.21. *Any open set in $\mathbb{R}$ is measurable.*

Proof. This follows from the fact that the topology in $\mathbb{R}$ has a countable base. In fact, one can look at

$$\mathcal{G} := \{(r - q, r + q) : q \in \mathbb{Q}, q > 0, r \in \mathbb{Q}\}.$$

Any open set can be written as union of elements in $\mathcal{G}$[8]. Since $\mathcal{G}$ is countable, any open set is measurable. □

[8]If this is unknown to you, you can assume this, or look it up in the internet if you are curious, we won't need to go much in this direction.

Lemma 1.22. *Let $f : \mathbb{R} \mapsto \mathbb{R}$ be a continuous function. Then f is $\mathcal{B}(\mathbb{R})$-$\mathcal{B}(\mathbb{R})$ measurable.*

Proof. By the definition of continuity $f^{-1}((a, b))$ is an open set. Since an open set in $\mathbb{R}$ is measurable by Lemma 1.21, we are done. □

Combining Lemmas 1.20 and 1.22, we conclude that if X is a random variable, $\sin(X), e^X, \log(|X|)$ etc. In particular,

Lemma 1.23. *If X is measurable and $g : \mathbb{R} \mapsto \mathbb{R}$ is Borel measurable, then any measurable function g, $g(X)$ is a random variable.*

1.4 Existence of measures

Now that we have defined the Borel sets, we want to define measures defined on Borel sets, which satisfies the conditions of Definition 1.3.

Let us start with a prototypical example, $\Omega = \mathbb{R}$ and μ is a measure which measures "length". Such a measure should satisfy $\mu((a, b)) = b - a$ for all $-\infty < a < b < \infty$. But how do we know that any such measure on $\mathbb{R}$ exists which satisfies $\mu((a, b)) = b - a$ for all $-\infty < a < b < \infty$?

Exercise 1.24. Thought experiment: try to construct such a measure directly. Firstly try to define this for "unions of finite intervals" (easy), then to countable unions and intersections (moderate), and then extend this to *all Borel sets* (*hard*, we will not try to do this in this course).

Indeed, the whole structure of probability theory is built upon the existence of such measures. We will now state two theorems, which as you can imagine are fundamental to everything that follows. Due to lack of time, we will not prove the theorems, but take the theorems as a *black box* and move on. To introduce the two theorems, we need two additional definitions.

Definition 1.7. *Let $\mathcal{A}$ be a collection of subsets of Ω.*

- *(ring) We say $\mathcal{A}$ is a ring if (i) $A, B \in \mathcal{A} \implies A \cup B \in \mathcal{A}$, (ii) $A, B \in \mathcal{A} \implies B \setminus A \in \mathcal{A}$.*
- *(π-system) We say $\mathcal{A}$ is a π-system if $A, B \in \mathcal{A} \implies A \cap B \in \mathcal{A}$. (e.g. The collection of all open intervals of $\mathbb{R}$.)*

Note that a ring must necessarily contain the emptyset. For example, let $\mathcal{A}$ is the collection of all subsets of $\mathbb{R}$ which are finite unions of intervals of the form $(a, b]$ for $a, b \in \mathbb{R}$. That is,

$$\mathcal{A} := \{\cup_{j=1}^{n}(a_j, b_j] : -\infty < a_1 \leq b_1 \leq a_2 \leq b_2 \leq \ldots \leq b_n < \infty\} \cup \{\emptyset\}. \tag{1.4}$$

Exercise 1.25. Verify that the example $\mathcal{A}$ above satisfy the conditions of a ring and a π-system respectively.

Exercise 1.26. Let $\mathcal{P} = \{(a, b) : -\infty < a < b < \infty\}$. Prove that $\mathcal{P}$ is a π-system.

The point is that both a ring and a π-system are subcollections of much less complexity than a σ-algebra. In particular, starting with a desirable definition of measure on such structures are potentially much easier. This motivates the following definition

Definition 1.8 (Pre-measure)**.** *Let $\mathcal{A}$ be a* ring*. We say $\mu : \mathcal{A} \mapsto [0, \infty)$ is a* ***pre-measure*** *if $\mu(\emptyset) = 0$ and for any countable collection of* disjoint *elements $\{A_n\}_{n\geq 1}$ of $\mathcal{A}$ such that $\cup_{n\geq 1} A_n$ is in $\mathcal{A}$,*

$$\mu(\bigcup_{n=1}^{\infty} A_n) = \sum_{n=1}^{\infty} \mu(A_n).$$

Theorem 1.1 (Carathéodory extension theorem)**.** *Let $\mathcal{A}$ be a ring over a sample space Ω and let μ be a* pre-measure *on $\mathcal{A}$. Then there exists a σ-algebra $\mathcal{F}$ containing $\mathcal{A}$ such that μ* ***extends*** *to a measure on $\mathcal{F}$.*

Figure 2: Constantin Carathéodory (1873-1950)

Here by "extends", we mean that there exists a measure on $\tilde{\mu}$ on $\mathcal{F}$ such that for any $A \in \mathcal{A}$, $\tilde{\mu}(A) = \mu(A)$.

Example 1.27. Take $\mathcal{A}$ to be the ring of the collection of finite union of intervals of the form $(a, b]$. In other words,

$$\mathcal{A} := \{\cup_{j=1}^{n}(a_j, b_j] : -\infty < a_1 \leq b_1 \leq a_2 \leq b_2 \leq \ldots \leq b_n < \infty\}. \tag{1.5}$$

Define

$$\mu(\cup_{j=1}^{n}(a_j, b_j]) = \sum_{j=1}^{n}(b_j - a_j)$$

Lemma 1.28. *The function μ on $\mathcal{A}$ defined above is a pre-measure.*

Proof. Take a countable collection $(A_n)_{n\geq 1}$ in $\mathcal{A}$. Since the union of A_n is in $\mathcal{A}$ we can write $A := \cup_{n\geq 1} A_n = \cup_{j=1}^{k} I_j$ where I_j are disjoint intervals of the form $(a, b]$. The crucial point here is that even though A is a countably infinite union of A_n, it is assumed to be in $\mathcal{A}$ so we can actually write it as a *finite* union of intervals.

Since A_ns are disjoint and each A_n is in $\mathcal{A}$, we can also write $\cup_{n\geq 1} A_n$ as a disjoint union of countably many intervals $\mathcal{J} := \{J_n : n \in \mathbb{N}\}$. (Finitely many coming from each A_n and they are all disjoint.) Note that

$$\sum_{n\geq 1} \mu(A_n) = \sum_{n\geq 1} \mu(J_n)$$

since we can rearrange the terms in the infinite series as everything is non-negative. Therefore we need to show

$$\sum_{n\geq 1} \mu(J_n) = \mu(A) = \sum_{j=1}^{k} \mu(I_j) \tag{1.6}$$

To prove (1.6), we make the following claim:

Claim 1.29. *Let U_j be the set of indices n such that J_n intersects I_j for $1 \leq j \leq k$.*

- *$\{U_j\}_{1\leq j\leq k}$ are disjoint and $\cup_{j=1}^{k} U_j = \mathbb{N}$. In other words, $\{U_j\}_{1\leq j\leq k}$ forms a partition of $\mathbb{N}$.*
- *$\cup_{n\in U_j} J_j = I_j$*

Let us prove the lemma assuming Claim 1.29. Since the series $\sum_{n\geq 1} \mu(J_n)$ consists of nonnegative terms, we can rearrange the terms without changing the value of the series. Thus writing

$$\sum_{n\geq 1} \mu(J_n) = \sum_{j=1}^{k} \sum_{n\in U_j} \mu(J_n)$$

Arrange the intervals with index in U_j in order, call them $(a_1, b_1], (a_2, b_2], \ldots$. Note that $a_2 = b_1$ as otherwise one can pick an element in (a_2, b_1) which is not in $\cup_{n\geq 1} A_n$ but is in $\cup_{j=1}^{k} I_j$. Hence

$$\sum_{n\in U_j} \mu(J_n) = \mu(I_j).$$

which completes the proof. □

Proof of Claim 1.29. Let U_1 be indices n such that J_n which intersect I_1. Since A is equal to $\cup_{j=1}^{k} I_j$, it must be the case that $\cup_{n\in U_1} J_n = I_1$. Indeed, suppose that an interval L with index in U_1 contains an element outside I_1. Since L is an interval, it must contain a point arbitrarily close to I_1 but outside I_1. Then it must contain an element outside $\cup_{j=1}^{k} I_j$ since the intervals I_j are disjoint. But this is impossible. Thus all the intervals with index in U_1 must be contained in I_1. Their union must be I_1 as well as otherwise $\cup_{n\geq 1} A_n = \cup_{j=1}^{k} I_j$ is false.

Applying the same reasoning for each I_ℓ for $1 \le \ell \le k$, we see that we can similarly define U_ℓ for $1 < \ell \le k$. Also $\cup_{1\le\ell\le k} U_\ell = \mathbb{N}$ since each interval J_n must intersect some I_j. Thus

$$\sum_{n\ge 1} \mu(J_n) = \sum_{1\le\ell\le k} \mu(I_1) = \mu(A).$$

Thus (1.6) is established, and hence μ is a pre-measure. □

By Theorem 1.1, we immediately get that μ extends to a measure on $\mathcal{B}(\mathbb{R})$. However, the next thing we need to make sure is that there can be **only one** such extentsion. This is provided by the following theorem. Recall the definition of a π-system from Definition 1.7.

Theorem 1.2. *Let $(\Omega, \mathcal{F})$ be a measure space and let $\mathcal{P} \subset \mathcal{F}$ be a π-system with $\Omega \in \mathcal{P}$ and $\sigma(\mathcal{P}) = \mathcal{F}$. Suppose there are two measures on $\mathcal{F}$ such that*

- *$\mu_1(A) = \mu_2(A)$ for all $A \in \mathcal{P}$*
- *There exists a sequence $\{\Omega_n\}_{n\ge1}$ with $\Omega_n \in \mathcal{P}$ for all n and $\Omega_n \uparrow \Omega$ as $n \to \infty$ with $\mu_1(\Omega_n) < \infty$ for all n. Then μ_1 and μ_2 are equal as measures (i.e. $\mu_1(E) = \mu_2(E)$ for all $E \in \mathcal{F}$.)*

Let us come back to Example 1.27. Take $\Omega = \mathbb{R}$ and $\mathcal{F} = \mathcal{B}(\mathbb{R})$ (Borel Sets). Take $\mathcal{P}$ to be the set of all open intervals of $\mathbb{R}$. You have already verified in Exercise 1.26 that $\mathcal{P}$ is a π-system. Now we verify that μ defined in Example 1.27 has a unique extension to a measure in $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ by applying Theorem 1.2. That is, suppose that there are two measures μ_1, μ_2 with $\mu_1((a,b)) = \mu_2((a,b)) = b - a$ for all open interval $(a,b) \subset \mathbb{R}$. All we need to find is a sequence $\Omega_n \in \mathcal{P}$ with $\mu_1(\Omega_n) < \infty$ for all n with $\Omega_n \uparrow \Omega$. Simply take $\Omega_n = (-n, n)$ which clearly satisfies the two conditions. We conclude that μ_1 is equal to μ_2. This unique extension is called a **Lebesgue measure** on $\mathbb{R}$.

Example 1.30. Take your favourite function $f : \mathbb{R} \mapsto [0, \infty)$ such that $\int_{\mathbb{R}} f(x)dx = 1$. Let $\mathcal{A}$ be as in (1.5) and let

$$\tilde{\mu}_f(\cup_{j=1}^n (a_j, b_j]) = \sum_{j=1}^n \int_{a_j}^{b_j} f(u)du.$$

Using simple properties of integrals, we can check that μ_f is a pre-measure. Thus $\tilde{\mu}_f$ extends to a unique measure μ_f on $\mathcal{B}(\mathbb{R})$ using Theorem 1.1 and Theorem 1.2 as before. This measure corresponds to a probability measure with density f.

Distribution of random variables One can take Example 1.27 quite a bit further. Recall the definition of a cumulative distribution function (cdf) from Math 352/ Stat 350: for any random variable X defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$[9], the cdf F_X can be taken to be

$$F_X(t) = \mathbb{P}(X \le t). \tag{1.7}$$

[9]It does not matter what we take this probability space to be. For example we can take $\Omega = [0,1]$, $\mathcal{F}$ to be the Borel sets and $\mathbb{P}$ to be the Lebesgue measure restricted to subsets of $[0,1]$ if we like.

Take $\mathcal{A}$ to be the ring in Equation (1.5) and for any $A = \cup_{i=1}^{n}(a_i, b_i) \in \mathcal{A}$

$$\mu_X(A) = \sum_{j=1}^{n}(F_X(b_j) - F_X(a_j))$$

Exercise 1.31. Show that the μ_X defined above is a pre-measure.

Using Theorem 1.1, we get that μ_X extends to a measure on all Borel sets.

Exercise 1.32. Using Theorem 1.2 show that μ_X extends to a unique measure on Borel sets.

Once we have verified this extension is unique, this is called the measure corresponding to the random variable X. The following exercise is the reason behind such nomenclature

Exercise 1.33. Show that for any Borel set $A \subset \mathbb{R}$, $\mu_X(A) = \mathbb{P}(X \in A)$ (μ_X is as defined above). Hint: Start by showing this for $\mathcal{A}$ and then extend to every element in $\mathcal{B}(\mathbb{R})$ using the trick in the proof of Lemma 1.16, or Theorem 1.2.

Definition 1.9. *Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space. Distribution of a random variable X is the probability measure μ_X defined on $(\mathbb{R}, \mathcal{B})$ defined by*

$$\mu_X(A) = \mathbb{P}(X \in A) = \mathbb{P}(\omega : X(\omega) \in A) = \mathbb{P}(X^{-1}(A)); \qquad \forall A \in \mathcal{B}.$$

Maybe it is not too hard to convince oneself that in general doing explicit computations with μ_X might be quite cumbersome. Also the description of μ_X is, unless we are in the trivial finite valued random variable case, somewhat explicit: take the unique extension of a set function satisfying certain properties. Perhaps a more practical representation of a probability measure on $\mathbb{R}$ is through a function called cumulative distribution function, or cdf.

Definition 1.10. *The **cumulative distribution function** is a function $F : \mathbb{R} \to \mathbb{R}$ corresponding to a probability measure $\mu : \mathbb{R} \mapsto \mathbb{R}$ is given by,*

$$F(t) = \mu((-\infty, t]), \qquad \forall t \in \mathbb{R}.$$

Exercise 1.34. Let F be a cumulative distribution function.

- $F \in [0, 1]$
- $F(t)$ is non-decreasing in t
- F is a right-continuous function. Also the left limit exist at all t, i.e.,

  $$\lim_{u \to t_-} F(u)$$

  exists.

- $\mu_X(\{a\}) = F(a) - F(a-)$ for all $a \in \mathbb{R}$.

Note that Definition 1.10 did not involve any random variable. But actually, we can recover the cumulative distribution function, which you learned in an introductory prob/ Stat course corresponding to a random variable via the following lemma. The proof is obvious from definition.

Lemma 1.35. *The **cumulative distribution function** F_X corresponding to a random variable defined in (1.7) satisfies*

$$F_X(t) = \mu_X((-\infty, t]), \qquad \forall t \in \mathbb{R}.$$

where μ_X is defined as above.

Let us finish this section which tells us that random variables, cdfs and probability measures on $\mathbb{R}$ are in one to one correspondence to one another.

Firstly, given a probability measure μ on $\mathbb{R}$, we can define a cdf $F(t) = \mu((-\infty, t]), t \in \mathbb{R}$ and given a cdf F we can define for any $A = \cup_{i=1}^n (a_i, b_i) \in \mathcal{A}$

$$\mu(A) = \sum_{j=1}^{n} (F(b_j) - F(a_j))$$

The same argument as in Exercise 1.31 tells us that μ can be extended to a probability measure on $\mathbb{R}$ which clearly corresponds to the cdf F in the sense that $\mu((-\infty, t]) = F(t)$. Thus we have established a one to one correspondence between cdfs and probability measures on $\mathbb{R}$.

Now let us prove a similar correspondence between probability measures on $\mathbb{R}$ and random variables. Clearly given any random variable, we have a unique measure μ_X such that $\mathbb{P}(X \in A) = \mu_X(A)$ for all Borel A. We now prove the converse.

Proposition 1.36. *Suppose $\mu : \mathcal{B}(\mathbb{R}) \mapsto [0, 1]$ is a probability measure. There exists a random variable X defined on $([0, 1], \mathcal{B}([0, 1]), \lambda|_{[0,1]})$ where $\lambda\ |_{[0,1]}$ is the restriction of the Lebesgue measure on $[0, 1]$ such that*

$$\mu_X(A) = P(X \in A) = \mu(A), \qquad \forall A \in \mathcal{B}(\mathbb{R}).$$

Proof. We straightaway construct the random variables by hand as follows. Take the probability space to be $([0, 1], \mathcal{B}([0, 1]), \lambda|_{[0,1]})$ where $\lambda\ |_{[0,1]}$ is the restriction of the Lebesgue measure on $[0, 1]$ as usual. Let us write this λ to shorten notation. Let us first construct a random variable U such that $\mu_U = \lambda$. This is easy, define $U(x) = x$ for all $x \in [0, 1]$. Then

$$\mu_U(a, b) = \lambda(\omega : U(\omega) \in (a, b)) = \lambda((a, b)) = b - a.$$

Clearly this means (by the uniqueness theorem)

$$\mu_U = \lambda.$$

Now let us move on to the general case. Take the cdf F corresponding to the probability measure μ given by

$$F(t) = \mu((-\infty, t]), \qquad t \in \mathbb{R}.$$

The idea now is to define '$G = F^{-1}$'. If F is not strictly increasing, this definition does not make sense. However let us assume F is strictly monotonic for the moment. Then define $X = G(U)$. Note that

$$\mathbb{P}(X \leq t) = \mathbb{P}(G(U) \leq t) = \mathbb{P}(U \leq G^{-1}(t)) = \mathbb{P}(U \leq F(t)) = F(t)$$

For general F, the idea is the same but we need to define G more carefully, which is as follows: Define $G : [0,1] \mapsto \mathbb{R}$ defined by

$$G(p) = \inf\{x : F(x) \geq p\}.$$

(This is a kind of inverse, sometimes called "right inverse" and matches with the inverse function if F is actually injective).

Claim. For every $q \in [0,1]$, $\{r \in \mathbb{R} : F(r) \geq q\} = \{r : G(q) \leq r\}$.

Given the claim, the proof is complete as we simply take

$$X = G(U).$$

Then

$$\mu_X((-\infty, r]) = \lambda(G(U) \leq r) = \lambda(U \leq F(r)) = \lambda((0, F(r)]) = F(r).$$

Now we can prove the claim by hand. Take r in the left hand side so that $F(r) \geq q$. Then$\{x : F(x) \geq q\} \subset [r, \infty)$. By definition $G(q)$ is the infimum of $\{x : F(x) \geq q\}$, and thus $G(q) \geq r$ and we conclude r is on the right hand side.

On the other hand, take r from the right hand side so that $G(q) \leq r$. Now if by contradiction $F(r) < q$ then by right continuity of F there is an $s > r$ such that $F(s) < q$. So $\{x : F(x) \geq q\} \subset (s, \infty)$ so $G(q) \geq s > r$ a contradiction, so $F(r) \geq q$. This completes the proof. □

Definition 1.11. *From now on, we shall refer to* $([0,1], \mathcal{B}([0,1]), \lambda|_{[0,1]})$ *where* $\lambda\,|_{[0,1]}$ *is the restriction of the Lebesgue measure to* $[0,1]$ *as the* ***standard probability space.***

Remark 1.37. One can have two random variables X, Y, defined on different probability spaces, but still $\mu_X = \mu_Y$. In this case, we say X **and** Y **have the same distribution.**

Example 1.38. Let $\Omega_1 = \{H, T\}$, $\mathcal{F}_1 = \{\emptyset, \{H\}, \{T\}, \Omega_1\}$ and $\mathbb{P}(\{H\}) = \mathbb{P}(\{T\}) = \frac{1}{2}$. Let $X : \Omega_1 \mapsto \mathbb{R}$ be defined as $X(H) = 1$, $X(T) = 0$. Now let $\Omega_2 = [0,1]$, $\mathcal{F}_2 = \mathcal{B}([0,1])$ and let λ be the Lebesgue measure on $[0,1]$. Now define $X(\omega) = 1$ if $\omega \in [0, 1/2]$ and $X(\omega) = 0$ if $\omega \in [1/2, 1]$. Then $\mu_X = \mu_Y$ which satisfies $\mu_X(\{1\}) = \mu_X(\{0\}) = \frac{1}{2}$. So X and Y have the same distribution, but they are defined on different probability spaces.

A random variable can be of different types depending on it's 'support'. If there is a countable set $\mathcal{S}$ such that $\mathbb{P}(X \in \mathcal{S}) = 1$ then the random variable is discrete and its probability mass function or pmf is defined as $p(i) = \mu_X(\{i\})$ for all $i \in \mathcal{S}$. Note here that $\mathcal{S}$ can be much more complicated than the integers, e.g. $\mathcal{S}$ could be the set of rationals. In the world of measures, we say μ_X has *atoms* at $\mathcal{S}$.

On the other hand, if X might have a probability density function pdf f_X, so that the cdf

$$F_X(a) = \int f_X(t)dt.$$

Using the arguments above, we know that F_X corresponds to a probability measure μ_X. Such a random variable is called a **continuous random variable.** [10]

Random variables can be much more complicated. For example, it can be a mixture of a discrete and a continuous random variables. Take the standard probability space. Let $U(\omega) = \omega$ be the identity function on $[0, 1]$. Define $Z = 10 \cdot 1_{[0,1/2]}$. Define

$$X = Z1_{[0,1/3]} + U1_{[1/2,1]}.$$

Spelling it out,

$$X(\omega) = \begin{cases} Z(\omega) = 10 \text{ if } 0 \leq \omega \leq 1/3 \\ 0 \text{ if } 1/3 < \omega < 1/2 \\ U(\omega) \text{ if } 1/2 \leq \omega \leq 1. \end{cases} \tag{1.8}$$

Note that X cannot be We now claim that for any Borel set A,

$$\mu_X(A) = \lambda(A \cap [1/2, 1]) + \frac{1}{3}\delta_{\{10 \in A\}} + \frac{1}{6}\delta_{\{0 \in A\}}.$$

where we interpret $\delta_B = 1$ if and only if B is true. Let us compute the cdf. Convince yourself that the cdf is

$$F_X(t) = \begin{cases} 0 \text{ if } t < 0 \\ \frac{1}{6} \text{ if } 0 \leq t < 1/2 \\ \frac{1}{6} + (t - \frac{1}{2}) \text{ if } 1/2 \leq t \leq 1 \\ \frac{1}{6} + \frac{1}{2} \text{ if } 1 \leq t \leq 10 \\ 1 \text{ if } t \geq 10. \end{cases}$$

Thus F_X is continuous at except at 0 and 10 where it has 'jumps' of size 1/6 and 1/3 respectively. By the last item of Exercise 1.34, we conclude that $\mu_X(\{0\}) = 1/6$ and $\mu_X(\{10\}) = \frac{1}{3}$, so μ_X has atoms at 0 and 10. Furthermore,

$$\begin{aligned} \mu_X((-\infty, t]) = F_X(t) &= (t - 1/2)1_{1/2<t\leq 1} + \frac{1}{6}\delta_{0\leq t} + \frac{1}{3}\delta_{10\leq t} \\ &= (t - 1/2)1_{1/2<t\leq 1} + \frac{1}{6}\delta_{0\leq t} + \frac{1}{3}\delta_{10\leq t} \\ &= \lambda((-\infty, t] \cap [1/2, 1]) + \frac{1}{3}\delta_{\{10\in(-\infty,t]\}} + \frac{1}{6}\delta_{\{0\in(-\infty,t]\}} \end{aligned}$$

[10] Later, we will call the measures corresponding to such random variables absolutely continuous measures.

Thus the claimed equality is satisfied for sets of the form $A = (-\infty, t]$. Now to extend this equality to all Borel sets, we employ the same trick as in Lemma 1.16. Define

$$\mathcal{B} = \{A : \mu_X(A) = \lambda(A \cap [1/2, 1]) + \frac{1}{3}\delta_{\{10 \in A\}} + \frac{1}{6}\delta_{\{0 \in A\}}\}.$$

By the argument just before, we conclude all sets of the form $(-\infty, t]$ is in $\mathcal{B}$. Of course $\emptyset$ is in $\mathcal{B}$ as both sides are 0 in this case. Finally we need to show that $\mathcal{B}$ is a σ-algebra. We leave it as a simple exercise to check this. Once we do this, we conclude that $\mathcal{B} \supset \mathcal{B}(\mathbb{R})$ and therefore our claim is proved.

Note μ_X cannot have a density function as μ_X has atoms. Also, μ_X cannot be discrete as $\mu_X(\{a\}) \neq 0$ if and only if $a \in \{0, 10\}$ and $\mu_X(\{0, 10\}) = \frac{1}{3} + \frac{1}{6} = \frac{1}{2} < 1$, so in a sense 'half' of its mass is carried by the Uniform variable U which is continuous.

1.5 Some pathological distributions

If a random variable is discrete, is it's cdf something nice that we can practically plot by hand? Unfortunately even discrete random variables can have very complicated cdf's. Take for example an enumeration of rationals $\{r_i : i \in \mathbb{N}\}$ and take a probability mass function $p := (p_i)_{i \in \mathbb{N}}$ and consider the random variable X with pmf p. The cdf of X will have a positive jump p_i at a dense (but countable) many points in $\mathbb{R}$.

Since the integral of a function is always continuous, the cdf of a continuous random variable is always continuous. Is the converse true? Unfortunately no. There is a cdf known as the **devil's staircase** whose corresponding random variable has no probability density function. See this Wiki article. The key insight is that the measure μ corresponding to this cdf satisfies $\mu(C) = 1$ where C is the cantor set. But you saw in an exercise that the Lebesgue measure of a Cantor set is 0. We will see in the next section that when we integrate any function over a Lebesgue measure 0 set, the integral is always 0. Thus there cannot exist such a density function. We will come back to the devil's staircase at an opportune time.

2 Integration

The goal of this subsection is to give a brief outline of construction of Expectation of X, where $X : \Omega \mapsto \mathbb{R}$ is a measurable random variable on the measure space $(\Omega, \mathcal{F}, \mu)$. There are various notations for this object that we construct:

$$\mathbb{E}(X) \equiv \int X d\mu \equiv \int X(\omega) d\mu(\omega).$$

They **all have the same meaning** and is a real number [11]. In this sense Expectation of a random variable and integration of a measurable function also has the same meaning. Perhaps a probabilist will view the above as expectation more often and integration is how an analyst will view the above quantity, maybe for a general measure. The above integral is constructed step by step for functions of increasing complexity, and is very similar to the construction of a Riemann integral. The difference is that the function (random variable) which we want to integrate is defined on an abstract space Ω rather than $\mathbb{R}$.

A review of Riemann integral: Let us get back to Riemann integration in $\mathbb{R}$ for a moment, and let us consider a concrete function, say $f(x) = x^2$ in the interval $[0, 1)$. The way Riemann integral works is of course to break up $[0, 1]$ into intervals $[i/n, (i+1)/n)$, $0 \leq i \leq n-1$. Then consider the function

$$f_{*,n} = \sum_{i=0}^{n-1} (i/n)^2 1_{[i/n,(i+1)/n)}$$

which approximates f from below. (Here $1_{[i/n,(i+1)/n)}$ is the indicator function as defined in (1.3).) Similarly we can have a function $f_n^* = \sum_{i=0}^{n-1} ((i+1)/n)^2 1_{[i/n,(i+1)/n)}$ which approximates f from above. Then we define

$$\int_0^1 f_{*,n} = \sum_{i=0}^{n-1} (i/n)^2 \frac{1}{n}.$$

A similar expression can be written for the integral f_n^*. Then we took a limit as $n \to \infty$ and called the limit (if it exists, which in most cases does) to be the Riemann integral. Here, maybe you remember that instead of the word 'define', we said 'clearly this is the case as we are finding the area under a bunch of rectangles'.

Here, instead of discretizing the x-axis, one can also discretize the y-axis as follows:

$$f_n(x) = \sum_{k=0}^{n2^n-1} \frac{k}{2^n} 1_{\frac{k}{2^n} \leq f(x) < \frac{k+1}{2^n}}(x).$$

[11] Sometimes people talk about expectation of a vector X in which case the expectation is a vector obtained by taking the expectation of each co-ordinate

and define

$$\int f_n(x)dx = \sum_{k=0}^{n2^n-1} \frac{k}{2^n}\lambda(\{x : f(x) \in [\frac{k}{2^n}, \frac{k+1}{2^n})\})$$

Clearly, this is a better approach if the domain is not $\mathbb{R}$ (recall that the range is always $\mathbb{R}$ for random variables). But we pay the price of changing the axis by having to deal with more complicated sets of the form $\{x : f(x) \in [\frac{k}{2^n}, \frac{k+1}{2^n})\} = f^{-1}([\frac{k}{2^n}, \frac{k+1}{2^n}))$.

In what follows, we make the setup more abstract:

- We replace $[0, 1]$ by Ω.
- We replace intervals by Borel sets.
- There is no nice way to find an approximation from above or below, so we need to do something more abstract.

2.1 Construction of Lebesgue integral:

We assume that we have a probability space $(\Omega, \mathcal{F}, \mu)$ given to us. All our random variables are defined on the same space.

- **Step 1.** Let X be a *simple function*, i.e., a function of the form $X = \sum_{i=1}^k a_i 1_{X\in A_i}$.

  Exercise 2.1. If X is a simple function, we can write $X = \sum_{i=1}^k b_i 1_{B_i}$ where B_i s are disjoint.

  Now take a representation of X of the form $X = \sum_{i=1}^k b_i 1_{B_i}$ where B_i s are disjoint and *define*

  $$\mathbb{E}(X) = \sum_{i=1}^k b_i \mu(B_i).$$

  We say a simple function X is integrable if $\mu(B_i) < \infty$ for all i. Note that if μ is a probability measure, every simple function is integrable.

  This definition has a slight problem: the representation of $X = \sum_{i=1}^k b_i 1_{B_i}$ might not be unique. For example:

  $$2 \cdot 1_{(0,1/2)} + 2 \cdot 1_{[1/2,3/4)} + 3 \cdot 1_{[3/4,1)} = 2 \cdot 1_{(0,3/4)} + 3 \cdot 1_{[3/4,1]}.$$

  We need to show that the above definition is well defined, that is,

  Lemma 2.2. *If $X = \sum_{i=1}^k b_i 1_{B_i} = \sum_{j=1}^\ell c_j 1_{C_j}$ are two different representations of X (but some B_i might intersect C_j) then*

  $$\sum_{i=1}^k b_i \mu(B_i) = \sum_{j=1}^\ell c_j \mu(C_j).$$

Proof. Note that if $B_i \cap C_j \neq \emptyset$ for some $i \neq j$ then $b_i = c_j$. Therefore in this case we can write

$$\sum_{i=1}^{k}\sum_{j=1}^{\ell} b_i 1_{B_i \cap C_j} = \sum_{i=1}^{k}\sum_{j=1}^{\ell} c_j 1_{B_i \cap C_j}.$$

Thus, taking expectation of both sides,

$$\sum_{i=1}^{k}\sum_{j=1}^{\ell} b_i \mu(B_i \cap C_j) = \sum_{i=1}^{k} b_i \mu(B_i)$$

and exchanging summation

$$\sum_{i=1}^{k}\sum_{j=1}^{\ell} c_j \mu(B_i \cap C_j) = \sum_{j=1}^{\ell}\sum_{i=1}^{k} c_j \mu(B_i \cap C_j) = \sum_{j=1}^{\ell} c_j \mu(C_j).$$

□

Lemma 2.3. *Suppose X, Y be simple integrable functions.*

a. *If $X \geq 0$ a.s. then $\mathbb{E}(X) \geq 0$.*
b. *If $a \in \mathbb{R}$, then $\mathbb{E}(aX + Y) = c\mathbb{E}(X) + \mathbb{E}(Y)$.*
c. *If $X \leq Y$ a.s. then $\mathbb{E}(X) \leq \mathbb{E}(Y)$.*

Proof. Suppose $X = \sum_{i=1}^{k} c_i 1_{E_i}$. Since $X \geq 0$ a.s., if $c_i < 0$ for some i then $\mathbb{P}(E_i) = 0$. Therefore $\mathbb{E}(X) = \sum_{i=1}^{k} c_i \mathbb{P}(E_i) \geq 0$.
Next, if $X = \sum_{i=1}^{k} c_i 1_{E_i}$ and $Y = \sum_{i=1}^{\ell} d_i 1_{F_i}$, then

$$aX + Y = \sum_{i=1}^{k} ac_i 1_{E_i} + \sum_{i=1}^{\ell} d_i 1_{F_i} = \sum_{i=1}^{k}\sum_{j=1}^{l} (ac_i + d_j) 1_{E_i \cap F_j}$$

. Thus

$$\begin{aligned}
\mathbb{E}(cX + Y) &= \sum_{i=1}^{k}\sum_{j=1}^{l} (ac_i + d_j)\mu(E_i \cap F_j) \\
&= \sum_{i=1}^{k}\sum_{j=1}^{l} ac_i \mu(E_i \cap F_j) + \sum_{i=1}^{k}\sum_{j=1}^{l} d_j \mu(E_i \cap F_j) \\
&= \sum_{i=1}^{k} ac_i \mu(E_i) + \sum_{j=1}^{l} d_j \mu(F_j) \\
&= a\mathbb{E}(X) + \mathbb{E}(Y)
\end{aligned}$$

For the final item, note $X - Y \geq 0$ a.s. and by the first item, $\mathbb{E}(X - Y) \geq 0$ and by the second item, $\mathbb{E}(X - Y) = \mathbb{E}(X) - \mathbb{E}(Y)$. Combining the two information, we have $\mathbb{E}(X) \geq \mathbb{E}(Y)$. □

- **Step 2.** Now assume $X \geq 0$ but not necessarily simple. Define

$$\mathbb{E}(X) = \sup\{\mathbb{E}(Z) : Z \leq X, Z \text{ is simple and integrable}\}. \tag{2.1}$$

The above quantity is allowed to be infinite. To clarify, $Z \leq X$ means that $Z(\omega) \leq X(\omega)$ for all ω. The detail missing here is to show that for simple functions this definition matches with step 1. To that end it is enough to show that if $Z \leq Z'$ and Z, Z' are both simple integrable, then $\mathbb{E}(Z) \leq \mathbb{E}(Z')$ for the previous definition. This is the last item of Lemma 2.3.

Digression. Before moving into step 3, we prove a quick Lemma which should make the second step above slightly less abstract.

Lemma 2.4. *Given any $X \geq 0$, there exists a (rather explicit) sequence of simple random variables X_n such that for every $\omega \in \Omega$, $X_n(\omega) \uparrow X(\omega)$.*

Proof. Let

$$B_{n,k} = \left\{\omega \in \Omega : \frac{k}{2^n} \leq X(\omega) < \frac{k+1}{2^n}\right\}; \qquad 0 \leq k < n \cdot 2^n.$$

Note changing from n to $n+1$ breaks up the intervals as

$$[\frac{k}{2^n}, \frac{k+1}{2^n}) = [\frac{2k}{2^{n+1}}, \frac{2k+1}{2^{n+1}}) \cup [\frac{2k+1}{2^{n+1}}, \frac{2k+2}{2^{n+1}}).$$

for each k. Thus $B_{n,k} = B_{n+1,2k} \cup B_{n+1,2k+1}$. Define

$$X_n(\omega) = \sum_{k=0}^{n2^n-1} \frac{k}{2^n} 1_{B_{n,k}}(\omega).$$

Thus on $B_{n,k}$, $X_n(\omega) = \frac{k}{2^n} \leq X_{n+1}(\omega)$. Thus $X_n(\omega)$ increases in n for every ω. Thus $\sup_n X_n(\omega) = \lim_{n\to\infty} X_n$.

Also, for every ω ,

$$X_n(\omega) \leq X(\omega) \text{ (easy to check!)} \implies \sup_n X_n(\omega) \leq X(\omega).$$

Also

$$X_n(\omega) \geq X1_{X\leq n} - \frac{1}{2^n} \implies \liminf X_n(\omega) \geq \liminf_n (X1_{X\leq n} - \frac{1}{2^n}) = X(\omega).$$

Consequently,

$$\lim_{n\to\infty} X_n(\omega) = X(\omega).$$

Thought experiment. Does this construction imply $\mathbb{E}(X_n) \uparrow \mathbb{E}(X)$. (It is ok if you find it hard to justify it. It is not OK if you think this is "obvious" (unless you are confident in analysis).) □

- **Step 3**. Back to construction of Expectation. For general random variables, decompose $X(\omega) = X^+(\omega) - X^-(\omega)$ for $\omega \in \Omega$ where $X^+ = \max(X, 0)$ and $X^- = -\min(X, 0)$. If either $\mathbb{E}(X^+) < \infty$ or $\mathbb{E}(X^-) < \infty$, define

$$\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-). \tag{2.2}$$

We say expectation of X is undefined if both $\mathbb{E}(X^+) = \mathbb{E}(X^-) = \infty$. Note here that $\mathbb{E}(X)$ could by ∞ or $-\infty$ and we do not say expectation is undefined if either of this occurs. It is easy to check that this definition matches the definition in step 2. That is, if $X \geq 0$ then $\mathbb{E}(X^+) = \mathbb{E}(X)$ and similarly for X^-.

This completes the construction of Expectation.

Notation: When μ is the Lebesgue measure, sometimes $d\mu(x)$ is shortened to dx.[12]

Example 2.5. Let X be a Cauchy random variable, i.e., for any $A \subset \mathbb{R}$,

$$\mathbb{P}(X \in A) = \int_A \frac{2}{\pi(1+x^2)} dx$$

We will show that $\mathbb{E}(X^+) = \mathbb{E}(X^-) = \infty$ in Section 2.3. In other words, Expectation of X is not well-defined.

To summarize, we say Expectation of X exists if and only if one of $\mathbb{E}(X^+)$ or $\mathbb{E}(X^-)$ is finite. On the other hand, we say X is **integrable** or **in** L^1 if $\mathbb{E}(|X|) = \mathbb{E}(X^+) + \mathbb{E}(X^-) < \infty$.

Next, we state a proposition which says that the nice properties which we expect in an integral is well maintained.

Proposition 2.6. *Suppose $X, Y, (X_n)_{n\geq 1}$ are measurable defined on a measure space with a σ-finite measure μ and expectation $\mathbb{E}(Z) = \int Z d\mu$. Then*

a. If $X \leq Y$ and $\mathbb{E}(X)$, $\mathbb{E}(Y)$ are well defined, then $\mathbb{E}(X) \leq \mathbb{E}(Y)$.

*b. **(Weak monotone convergence)** If $X_n \geq 0$ a.s. and $X_n \uparrow X$, then $\mathbb{E}(X_n) \uparrow \mathbb{E}(X)$.*

*c. **(Linearity of expectation)**If $X, Y \geq 0$ a.s. and $a > 0$ then*

$$\mathbb{E}(aX + Y) = a\mathbb{E}(X) + \mathbb{E}(Y).$$

Proof of Part a. If $X \geq 0$ then $\mathbb{E}(X) \leq \mathbb{E}(Y)$ is obvious as the supremum in the definition (2.1) is taken over a larger set. Similarly for general measurable $X \leq Y$, $X^+ \geq Y^+$ and $X^- \leq Y^-$, and therefore using the definition (2.2), the result follows. □

[12] which matches the notation of elementary integration learnt in Calculus I

Proof of Part b. Note that since X_n is non-decreasing a.s., $\mathbb{E}(X_n)$ is non-decreasing and upper bounded by $\mathbb{E}(X)$ by item a. Therefore

$$\lim_{n\to\infty} \mathbb{E}(X_n) = \sup_n \mathbb{E}(X_n) \leq \mathbb{E}(X).$$

We now concentrate on proving the reverse inequality. First assume $\mathbb{E}(X) < \infty$. Take a simple function $Z = \sum_{i=1}^k c_i 1_{E_i}$ with $Z \leq X$. Fix $\delta > 0$ and let $\varepsilon = \frac{\delta}{\mathbb{E}(Z)}$. Let

$$E_{i,n} = E_i \cap \{X_n \geq c_i(1-\varepsilon)\}.$$

Notice that since $Z \leq X$, $E_{i,n} \uparrow E_i$. Also note

$$X_n \geq \sum_{i=1}^k (c_i(1-\varepsilon))1_{E_{i,n}}$$

so by part a.

$$\begin{aligned}\mathbb{E}(X_n) &\geq \mathbb{E}(\sum_{i=1}^k (c_i(1-\varepsilon))1_{E_{i,n}}) \\ &= \sum_{i=1}^k c_i \mathbb{P}(E_{i,n}) - \varepsilon \sum_{i=1}^k c_i \mu(E_{i,n}).\end{aligned}$$

Notice by continuity of measures $\lim_n \mu(E_{i,n}) = \mu(E_i)$. Thus

$$\begin{aligned}\lim_n \mathbb{E}(X_n) &\geq \sum_{i=1}^k c_i \mu(E_i) - \varepsilon \sum_{i=1}^k c_i \mu(E_i) \\ &= \mathbb{E}(Z) - \delta \text{ by choice of } \varepsilon.\end{aligned}$$

Notice now that since $Z \leq X$ was arbitrary and simple,

$$\lim_n \mathbb{E}(X_n) \geq \sup_{Z \text{ simple }, Z \leq X} \mathbb{E}(Z) - \delta = \mathbb{E}(X) - \delta.$$

We are done since δ was arbitrary.

If $\mathbb{E}(X) = \infty$, the proof follows similar lines and we define $Z = \sum_{i=1}^k c_i 1_{E_i}$ be a simple function such that $Z \leq X$ as before. Let

$$E_{i,n} = E_i \cap \{X_n \geq c_i/2\}.$$

Note $E_{i,n} \uparrow E_i$, again as before. Thus

$$\begin{aligned}\lim_n \mathbb{E}(X_n) = \liminf_n X_n &\geq \liminf_n \mathbb{E}(\sum_{i=1}^k (c_i/2)1_{E_{i,n}}) \\ &= \sum_{i=1}^k c_i/2 \mu(E_i) \\ &= \frac{1}{2}\mathbb{E}(Z).\end{aligned}$$

Notice now that since $Z \leq X$ was arbitrary and simple,

$$\lim_n \mathbb{E}(X_n) \geq \frac{1}{2} \sup_{Z \text{ simple }, Z \leq X} \mathbb{E}(Z) = \infty.$$

as desired. □

Proof of part c. Using Lemma 2.4, find $X_n \uparrow X$ and $Y_n \uparrow Y$ such that $X_n \geq 0$ and $Y_n \geq 0$ are simple functions. Using linearity of expectation of simple functions (Lemma 2.3, item b.) and weak monotone convergence,

$$\mathbb{E}(aX+Y) \overset{\text{weak mon.}}{=} \lim_{n\to\infty} \mathbb{E}(aX_n+Y_n) \overset{Lemma\ 2.3b.}{=} \lim_{n\to\infty} (a\mathbb{E}(X_n)+\mathbb{E}(Y_n)) \overset{\text{weak mon.}}{=} a\mathbb{E}(X)+\mathbb{E}(Y)$$

□

2.2 Further properties of Expectation

Definition 2.1 (Almost sure convergence)**.** *We say X_n almost surely converges to a random variable X if we set*

$$\Omega_0 = \{\omega \in \Omega : X_n(\omega) \text{ converges to } X(\omega)\}$$

then

$$\mathbb{P}(\Omega_0) = 1.$$

We specify that we will liberally use the term almost surely or a.s. to describe events which have probability measure 1. Here is a few illustrations of the use of this terminology.

- $X \leq Y$ almost surely means $\mathbb{P}(\Omega_0) = 1$ where $\Omega_0 = \{\omega \in \Omega : X(\omega) \leq Y(\omega)\}$.
- $X_n \to 5$ almost surely means that $\mathbb{P}(\Omega_0) = 1$ where $\Omega_0 = \{\omega \in \Omega : X_n(\omega) \to 5\}$.

Lemma 2.7. *If X_n converges to X almost surely and $g : \mathbb{R} \mapsto \mathbb{R}$ is a continuous function then $g(X_n)$ converges to $g(X)$ almost surely.*

Proof. Let $\mathcal{G} = \{\omega : \lim_{n\to\infty} X_n(\omega) = X(\omega)\}$. By definition of a.s. convergence, $\mathbb{P}(\mathcal{G}) = 1$. But by continuity of g, for every ω in $\mathcal{G}$, $g(X_n(\omega)) \to g(X(\omega))$ (just basic definition of continuity from real analysis is invoked here). Thus $\{\omega : g(X_n(\omega)) \to g(X(\omega))\} \supset \mathcal{G}$. Therefore

$$\mathbb{P}(\omega : g(X_n(\omega)) \to g(X(\omega))) = 1,$$

as desired. □

As a corollary, we can say for example that if $X_n \to X$ almost surely, then $X_n^2 \to X^2$ almost surely, or $e^{X_n}\cos(X_n) \to e^X \cos(X)$ almost surely.

Exercise 2.8. If $X_n \to 2$ almost surely and Y_n converges to U almost surely, show that $X_n + Y_n$ converges to $5 + U$ almost surely.

We now address the following question:

Question 2.9. *If $X_n \to X$, can we always say $\mathbb{E}(X_n) \to \mathbb{E}(X)$ (recall we assumed the expectations exist at the beginning of the chapter, so this is a nontrivial question)? If not, when can we assure the convergence of expectations?*

The answer unfortunately is no. Here is a working example to keep in mind when going through the next set of propositions.

Example 2.10. Let Z_n be a sequence of random variables defined on the standard probability space (Definition 1.11) defined as $Z_n = n1_{[0,1/n]}$. Note Z_n is a simple function and $\mathbb{E}(Z_n) = 1$. However, for all $\omega \in (0,1)$, $Z_n(\omega) \to 0$ and hence $Z_n \to 0$ almost surely as $\lambda((0,1)) = 1$.

In this subsection, X, Y and $\{X_n\}_{n\geq 1}$ are random variables defined on a σ-finite measure space $(\Omega, \mathcal{F}, \mu)$. We also assume that their expectations are all well-defined.

Lemma 2.11. *Assume $X \geq 0$. Then $\mathbb{E}(X) = 0$ if and only if $X = 0$ almost surely.*

Proof. First assume $X = 0$ almost surely. Let $A = \{\omega : X(\omega) = 0\}$. Since $X = 0$ almost surely, $\mathbb{P}(A) = 1$. By Lemma 2.4, we can find a sequence of simple functions $0 \leq X_n \leq X$ such that $X_n(\omega) \uparrow X(\omega)$ for all ω. Therefore $X_n 1_{A^c} \uparrow X 1_{A^c}$. Since X_n is a simple function, we must have a number K_n such that $X_n(\omega) \leq K_n$ for all ω. By weak monotone convergence theorem (proved above),

$$0 \leq \mathbb{E}(X_n 1_{A^c}) \leq K_n \mathbb{P}(A^c) = 0 \overset{\text{(taking } n \to \infty \text{)}}{\Longrightarrow} \mathbb{E}(X 1_{A^c}) = 0.$$

Also by definition $X 1_A = 0$ for all ω. Thus $\mathbb{E}(X 1_A) = 0$. Hence

$$\mathbb{E}(X) = \mathbb{E}(X 1_A) + \mathbb{E}(X 1_{A^c}) = 0.$$

For the converse, assume $\mathbb{E}(X) = 0$. Let $A_n = \{\omega : X(\omega) \in [0, 1/n)\}$. Then $A_n \downarrow A = \{\omega : X(\omega) = 0\}$ and $A_n^c \uparrow A^c$. Therefore by continuity of measures, $\mathbb{P}(A_n) \to \mathbb{P}(A)$ and $\mathbb{P}(A_n^c) \to \mathbb{P}(A^c)$ as $n \to \infty$. Thus

$$0 = \mathbb{E}(X) \overset{\text{since } X \geq 0}{\geq} \mathbb{E}(X 1_{A_n^c}) \geq \frac{1}{n} \mathbb{P}(A_n^c) \implies \mathbb{P}(A_n^c) = 0$$

which means $\mathbb{P}(A^c) = 0$ hence $\mathbb{P}(A) = 1$ which implies $X = 0$ almost surely. □

Lemma 2.12. *If $X \geq 0$ almost surely then $\mathbb{E}(X) \geq 0$ and $\mathbb{E}(X) = 0$ if and only if $X = 0$ almost surely.*

Proof. Notice that $X^+(\omega) \geq 0$ for all ω. Thus $\mathbb{E}(X^+) \geq 0$. Also $X^-(\omega) \geq 0$ for all ω and $X^- = 0$ almost surely since $X \geq 0$ a.s. Thus $\mathbb{E}(X^-) = 0$ by previous part. Hence $\mathbb{E}(X) = \mathbb{E}(X^+) - \mathbb{E}(X^-) \geq 0$.

Suppose $\mathbb{E}(X) = 0$. This means $\mathbb{E}(X^+) = 0$ which implies $X^+ = 0$ a.s. Let $\Omega_+ = \{\omega : X^+ \neq 0\}$ and $\Omega^- := \{X^- \neq 0\}$. Also $\Omega_0 := \{\omega : X(\omega) \neq 0\} \subseteq \Omega_+ \cup \Omega_-$. We conclude that $\mathbb{P}(\Omega_0) \leq \mathbb{P}(\Omega_+) + \mathbb{P}(\Omega_-) = 0$. Thus $X = 0$ a.s.

Suppose $X = 0$ a.s. Note $X^+ \geq 0$ and $\{X^+ \neq 0\} \subseteq \{X \neq 0\}$. Thus $X^+ = 0$ a.s. as well. Thus $\mathbb{E}(X^+) = 0$. Similarly $\mathbb{E}(X^-) = 0$ as well and we have $\mathbb{E}(X) = 0$. □

Lemma 2.13. *If $X \geq Y$ almost surely, then $\mathbb{E}(X) \geq \mathbb{E}(Y)$. If $X = Y$ almost surely, then $\mathbb{E}(X) = \mathbb{E}(Y)$.*

Proof. Consider $Z = X - Y$ and apply the previous item. We leave details as an exercise. □

Proposition 2.1 (Monotone convergence theorem or MCT)**.** *If $0 \leq X_n$ for all n and $X_n \uparrow X$ almost surely, then $\mathbb{E}(X_n) \uparrow \mathbb{E}(X)$.*

Proof. Let $E = \{\omega : X_n(\omega) \uparrow X(\omega)\}$. By definition $\mathbb{P}(E) = 1$. Thus $X_n 1_E = X_n$ almost surely and $X 1_E = X$ almost surely.

$$\mathbb{E}(X_n 1_E) \uparrow \mathbb{E}(X 1_E) \qquad \text{by weak monotone convergence theorem.}$$

and

$$\mathbb{E}(X_n 1_E) = \mathbb{E}(X_n); \qquad \mathbb{E}(X 1_E) = \mathbb{E}(X)$$

Combining the above two displays, the proof is complete. □

Exercise 2.14. What condition in the monotone convergence theorem does Example 2.10 violate?

Proposition 2.2 (Fatou's lemma)**.** *If $X_n \geq 0$ almost surely for all n then*

$$\mathbb{E}(\liminf_{n\to\infty} X_n) \leq \liminf_{n\to\infty} \mathbb{E}(X_n) \tag{2.3}$$

Proof. Notice that $\inf_{k\geq n} X_k$ is increasing in n and its limit is $\liminf_n X_n$. Thus by monotone convergence theorem,

$$\mathbb{E}(\liminf_{n} X_n) = \lim_{n\to\infty} \mathbb{E}(\inf_{k\geq n} X_k)$$

However

$$\inf_{k\geq n} X_k \leq X_n \implies \mathbb{E}(\inf_{k\geq n} X_k) \leq \mathbb{E}(X_n).$$

Consequently,

$$\liminf_{n} \mathbb{E}(\inf_{k\geq n} X_k) = \lim_{n\to\infty} \mathbb{E}(\inf_{k\geq n} X_k) \leq \liminf_{n} \mathbb{E}(X_n) \tag{2.4}$$

But the left hand side is $\mathbb{E}(\liminf_{n\to\infty} X_n)$ which yields (2.3). □

Exercise 2.15. Does Fatou's lemma hold for Example 2.10?

Proposition 2.3 (Dominated convergence theorem or DCT)**.** *Suppose $X_n \to X$ almost surely and there exists a random variable Y with $\mathbb{E}(|Y|) < \infty$ such that $|X_n| \leq Y$ almost surely, then,*

$$\mathbb{E}(X_n) \to \mathbb{E}(X).$$

Proof. We learnt in Lemma 2.12 that changing a random variable on a set of measure zero does not change its expectation. Therefore, we assume $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega$. Furthermore, $|X_n| \leq Y$ for all n, so $|X| \leq Y$ pointwise. Now apply Fatou's lemma to $X_n + Y$ and $Y - X_n$ (we can do this since $-Y \leq X_n \leq Y$ hence both these random variables are non-negative) .

$$\liminf_{n\to\infty} \mathbb{E}(X_n + Y) \geq \mathbb{E}(\liminf_{n\to\infty}(X_n + Y)) = \mathbb{E}(X + Y) \implies \liminf_{n\to\infty} \mathbb{E}(X_n) \geq \mathbb{E}(X) \quad (2.5)$$

and

$$\liminf_{n\to\infty} \mathbb{E}(Y - X_n) \geq \mathbb{E}(\liminf_{n\to\infty}(Y - X_n)) = \mathbb{E}(Y - X) \implies \liminf_{n\to\infty} \mathbb{E}(-X_n) \geq \mathbb{E}(-X). \quad (2.6)$$

This means $\limsup_{n\to\infty} \mathbb{E}(X_n) \leq \mathbb{E}(X)$. Combining with $\liminf_{n\to\infty} \mathbb{E}(X_n) \geq \mathbb{E}(X)$, the proof is complete. □

Exercise 2.16. What condition in the dominated convergence theorem does Example 2.10 violate?

2.3 Special cases and computing Expectations

In this subsection, we show how this theory encompasses the things we learned about random variables in Math 352/ Stat 350.

Riemann integrable implies Lebesgue integrable: First let us think about Riemann integration and how it is encompassed by Lebesgue integration. The setup is a measurable function $f : \mathbb{R} \to \mathbb{R}$. The measure space on which f is defined is $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \lambda)$ where λ is the Lebesgue measure. (If the random variable is defined on $\mathbb{R}$, then we usually replace capital letters like X, Y by usual small letters like f, g). If we want to integrate f over an interval $[0, 1]$, Let $I_i = [i/n, (i + 1)/n]$ and $f_{*,i} = \inf\{f(x) : x \in [i/n, (i + 1)/n\}$. then note that the approximation

$$f_{*,n} = \sum_{i=0}^{n-1} f_{*,i} 1_{[i/n,(i+1)/n)}$$

is an approximation of f from below by simple function, whose limit is the Riemann integral. By MCT, we can conclude that this limit is also the same as the Lebesgue integral of f over $[0, 1]$. It is a good exercise to think about this on your own and convince yourself:

Exercise 2.17. Suppose f is Riemann integrable on an interval $[a, b]$. Prove that the Riemann integral of f over $[a, b]$ is the same as it's Lebesgue integral.

Now we can use the Riemann integral calculations from calculus we are used to. For example, if we want to compute $\int e^{-x} dx$, we simply note that

$$\int_0^n e^{-x} dx = \int e^{-x} 1_{x\in[0,n]} dx \uparrow \int e^{-x} dx$$

by MCT. On the other hand, since Lebesgue and Riemann integrals coincide, $\int_0^n e^{-x}dx = 1 - e^{-n}$. Thus $\int e^{-x}dx = 1$.

Let us introduce some further notations:

- If E is a Borel set then $\int_E f(x)d\mu(x) := \int f1_E d\mu(x)$.
- If λ is the lebesgue measure then $\int f(x)d\lambda(x) = \int_{\mathbb{R}} f(x)dx = \int_{-\infty}^{\infty} f(x)dx$.
- If λ is the lebesgue measure and $E = (a,b)$ or $(a,b]$ or $[a,b)$ or $[a,b]$, we write $\int_E f(x)d\lambda(x) = \int_a^b f(x)dx$. (The integral is the same for all such E, and is equal to the Riemann integral).

Let us now consider a general probability measure. Suppose X is defined on an abstract probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The main idea is to translate the expectation in such a space to the real line where computation and calculus is possible. This is given by the following change of variable formula.

Theorem 2.18. *Suppose $g : \mathbb{R} \mapsto \mathbb{R}$ is measurable and either $g(X) \geq 0$ or $\mathbb{E}(|g(X)|) < \infty$. Then*

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(y)d\mu_X(y). \tag{2.7}$$

Furthermore, if μ_X comes from a density (see Example 1.30), that is, $\int_A d\mu_X = \int_A f_x(y)dy$ for a pdf f_X, then

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(y)d\mu_X(y) = \int_{-\infty}^{\infty} g(y)f_X(y)dy$$

Proof. The proof follows essentially the construction of integral in Section 2, by proving (2.7) for functions of increasing level of complexity. We will write $\int f d\mu$ in place of expectation with respect to the measure μ_X on $\mathbb{R}$ to differentiate from expectation with respect to the measure $\mathbb{P}$ on $(\Omega, \mathcal{F}, \mathbb{P})$.

Step 1. Step function . Let $g(x) = 1_A(x)$ for some Borel $A \subset \mathbb{R}$. Then

$$\mathbb{E}(g(X)) = \mathbb{E}(1_A(X)) = \mathbb{P}(X \in A) = \mu_X(A) = \int_A d\mu_X = \int 1_A(y)d\mu_X(y) = \int g(y)d\mu_X(y).$$

If μ_X comes from a density f_X then using our notations defined before the proposition,

$$\mu_X(A) = \int_A f_X(y)dy = \int 1_A(y)f_X(y)dy = \int g(y)f_X(y)dy.$$

Step 2. Simple function. Let $g(x) = \sum_{m=1}^{n} c_m 1_{B_m}$, $c_m \in \mathbb{R}$ and B_ms are disjoint and Borel for all m. Then by linearity of expectation

$$\begin{aligned}\mathbb{E}(g(X)) &= \mathbb{E}(\sum_{m=1}^{n} c_m 1_{B_m}(X)) = \sum_{m=1}^{n} c_m \mathbb{E}(1_{B_m}(X)) \\ &= \sum_{m=1}^{n} c_m \mu_X(B_m) \\ &= \int g(x) d\mu_X(x).\end{aligned}$$

Step 3. Non-negative function. Suppose $g \geq 0$. Then find a sequence of simple functions $0 \leq g_n \uparrow g$ as in Lemma 2.4. Then by Monotone convergence theorem, $\mathbb{E}(g_n(X)) \uparrow \mathbb{E}(g(X))$ and $\int g_n(y) d\mu_X(y) \uparrow \int g(y) d\mu_X(y)$. By step 2, $\mathbb{E}(g_n(X)) = \int g_n(y) d\mu_X(y)$, which completes the proof.

Step 4. Integrable functions. If $g(X)$ is integrable, that is, $\mathbb{E}(|g(X)|) < \infty$, we can write $g(X) = (g(X))^+ - (g(X))^-$ and continue as in the construction of Lebesgue integral. This step is left as an exercise. □

Now it is easy to see that if $\mu_X(A) = 0$ then $\int_A g(y) d\mu_X(y) = 0$. This is because the integral is, by Theorem 2.18,

$$\int g(y) 1_{y \in A} d\mu_X(y) = \mathbb{E}(g(X) 1_{\{\omega : X(\omega) \in A\}}).$$

But $g(X) 1_{\{\omega : X(\omega) \in A\}}$ is 0 outside $X^{-1}(A)$ and is non-zero on $X^{-1}(A)$ which has probability 0. Thus $g(X) 1_{\omega : X(\omega) \in A}$ is 0 almost surely, and consequently, it's expectation is 0. As a corrolary, we get that for Lebesgue measure λ, if $\lambda(A) = 0$ then $\int_A g dx = 0$.

Discrete Random variables Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A random variable is discrete if there exists a countable set $\mathcal{S} \subset \mathbb{R}$ such that $\mathbb{P}(X \in \mathcal{S}) = 1$. The set $\mathcal{S}$ is called the support of X.

How will the measure induced by X, μ_X behave? For any $y \in \mathcal{S}$,

$$\mu_X(\{y\}) = \mathbb{P}(X = y).$$

So μ_X is actually the **probability mass function** (denoted p_X there) that we learn in a first level probability course.

Proposition 2.19. *If X is discrete with support $\mathcal{S}$. Let $g : \mathbb{R} \mapsto \mathbb{R}$ is measurable. Then* $\mathbb{E}(g(X)) = \sum_{x \in \mathcal{S}} g(x) \mu_X(\{x\})$.

Proof. By Theorem 2.18, we have

$$\mathbb{E}(g(X)) = \int g(y)d\mu_X(y).$$

Therefore to prove the formula, we need to show that $\int g(y)d\mu_X(y) = \sum_{x\in\mathcal{S}} g(x)\mu_X(\{x\})$. The proof of this follows similar steps as in the proof of Theorem 2.18. In the first step, assume $g = 1_A$ for some Borel A. Then note:

$$\int g(y)d\mu_X(y) = \int 1_A(y)d\mu_X(y) = \mu_X(A) = \sum_{x\in A}\mu_X(\{x\}) = \sum_{x\in\mathcal{S}} 1_A(x)\mu_X(\{x\}) = \sum_{x\in\mathcal{S}} g(x)\mu_X(\{x\}).$$

The next steps involve taking g to be a simple function which is a simple application of linearity of expectation. Then using monotone convergence theorem we need to prove the statement for $g \geq 0$ and then finally for integrable g.

Exercise 2.20. Finish the proof by mimicking the proof of Theorem 2.18.

□

2.4 Applications of MCT and DCT

We finish this section with some applications of DCT. Before we start we need to extend our viewpoint of random variables slightly, and assume they can take values ∞ or $-\infty$. In this case, we need to be wary of three cases.

- If $X^{-1}(\{\infty\})$ has positive probability but $X^{-1}(\{-\infty\})$ has zero probability then $\mathbb{E}(X) = \infty$.
- If $X^{-1}(\{\infty\})$ has zero probability but $X^{-1}(\{-\infty\})$ has positive probability then $\mathbb{E}(X) = -\infty$
- If both $X^{-1}(\{\infty\})$ and $X^{-1}(\{-\infty\})$ have positive probability then $\mathbb{E}(X)$ is undefined.

Lemma 2.21 (Bounded convergence theorem)**.** *Suppose M is a constant such that $|X_n| \leq M$ for all $n \geq 1$. Then $\mathbb{E}(X_n) \to \mathbb{E}(X)$.*

Proof. This is a simple application of DCT where we take Y in Proposition 2.3 to be the constant random variable. □

Example 2.22. Suppose $X_n \sim$Binomial $(n, 1/n)$ and $\{X_n\}_{n\geq 1}$ is defined on the same probability space. Suppose we know $X_n \to X$ almost surely where $X \sim$Poisson(1). Calculate

$$\lim_{n\to\infty} \mathbb{E}(e^X \cos(X_n)).$$

Using Exercise 2.8 we know that $\cos(X_n)$ converges almost surely to $\cos(X)$. Also note$|e^X \cos(X_n)| \leq e^X$ and $\mathbb{E}(e^X) < \infty$ (check). Thus by DCT,

$$\lim_{n\to\infty} \mathbb{E}(e^X \cos(X_n)) = \mathbb{E}(e^X \cos(X)) = \sum_{k=0}^{\infty} \frac{e^{1-k}\cos(k)}{k!}.$$

Lemma 2.23 (Reverse Fatou). *If $X_n \leq Y$ for all $n \geq 1$ with $\mathbb{E}(|Y|) < \infty$, then*

$$\limsup_{n\to\infty} \mathbb{E}(X_n) \leq \mathbb{E}(\limsup_{n\to\infty} X_n).$$

Proof. Apply Fatou to $Y - X_n$ (Exercise: Fill in the details.) □

We now take a quick detour into two very useful inequalities.

Definition 2.24. *A function $\varphi : \mathbb{R} \mapsto \mathbb{R}$ is convex if for all $x \in \mathbb{R}$, for all $p \in [0,1]$, $\varphi(px + (1-p)y) \leq p\varphi(x) + (1-p)\varphi(y)$.*

Usual examples include $\varphi(x) = x^2$ or $|x|^p$ for $p \geq 1$. A convex function is always continuous and hence Borel measurable (exercise), but may not be differentiable (e.g.$|x|$).

Proposition 2.4 (Jensen's inequality). *Suppose X is a random variable with $\mathbb{E}(|X|) < \infty$ and φ is convex. Then $\mathbb{E}(\varphi(X)) \geq \varphi(\mathbb{E}(X))$.*

Proof. It is a fact that for any convex function one can find a linear function $ax + b$ such that $ax + b \leq \varphi(x)$ for all $x \in \mathbb{R}$ and such that $ax_0 + b = \varphi(x_0)$ [13]. Choose $x_0 = \mathbb{E}(X)$. Then

$$\mathbb{E}(\varphi(X)) \geq \mathbb{E}(aX + b) = a\mathbb{E}(X) + b = ax_0 + b = \varphi(x_0) = \varphi(\mathbb{E}(X)).$$

which is the required inequality. □

Corollary 2.25. *We have $\mathbb{E}(X^2) \geq (\mathbb{E}(X))^2$ and hence $\mathrm{Var}(X) \geq 0$ where $\mathrm{Var}(X) := \mathbb{E}(X^2) - (\mathbb{E}(X))^2$. In general $\mathbb{E}(|X|^p) \geq |\mathbb{E}(X)|^p$ if $p \geq 1$.*

Proof. Apply Jensen with $\varphi(x) = |x|^p$ which is convex for $p \geq 1$. □

A particularly useful application is taking the absolute value inside the expectation, specially for sums. For example, considering the random variable X uniform over the finite set $\{x_1, \ldots, x_n\}$, we get

$$|\mathbb{E}(X)| \leq \mathbb{E}(|X|) \implies |\frac{x_1 + \ldots + x_n}{n}| \leq \frac{|x_1| + \ldots + |x_n|}{n} \implies |x_1 + \ldots + x_n| \leq |x_1| + \ldots + |x_n|.$$

See Proposition 2.19 for a discussion on the above formula. Also,

$$|\mathbb{E}(X_1 + \ldots + X_n)| \leq \mathbb{E}(|X_1 + \ldots + X_n|) \leq \mathbb{E}(|X_1| + \ldots + |X_n|) = \sum_{k=1}^{n} \mathbb{E}(|X_k|).$$

In other words, we took the absolute value inside at the expense of an inequality.

[13] See https://en.wikipedia.org/wiki/Subderivative

Example 2.26. Take a sequence $(X_n)_{n\geq 1}$ defined on the same probability space, but $X_n \sim$Binomial $(n, \frac{1}{n^3})$. Does the series $\sum_{n\geq 1} X_n$ converge with probability 1? Note that $\mathbb{E}(X_n) = \frac{1}{n^2}$ (Recall if $X \sim$Binomial (n, p) then $\mathbb{E}(X) = np$). Then

$$\sum_{n\geq 1} \mathbb{E}(X_n) = \sum_{n\geq 1} \frac{1}{n^2} < \infty.$$

If *we could push the expectation inside the sum* without changing the value, then we would get

$$\mathbb{E}(\sum_{n\geq 1} X_n) = \sum_{n\geq 1} \mathbb{E}(X_n) = \sum_{n\geq 1} \frac{1}{n^2} < \infty.$$

which would imply that $\sum_{n\geq 1} X_n < \infty$ almost surely.

Lemma 2.27. *Suppose $\sum_{n=1}^{\infty} \mathbb{E}(|X_n|) < \infty$. Show that $\sum_{n=1}^{\infty} X_n$ is almost surely absolutely convergent series, and furthermore,*

$$\sum_{n=1}^{\infty} \mathbb{E}(X_n) = \mathbb{E}(\sum_{n=1}^{\infty} X_n).$$

Proof. The idea is to combine MCT and DCT. We first employ MCT to ensure DCT is applicable, and then apply DCT. Let $Y_n = \sum_{k=1}^{n} |X_k|$. Hence $\lim_n Y_n = \sum_{k=1}^{\infty} |X_k|$. Notice that by MCT,

$$\mathbb{E}(\lim_{n\to\infty} Y_n) = \lim_{n\to\infty} \mathbb{E}(Y_n) = \sum_{k=1}^{\infty} \mathbb{E}(|X_k|) < \infty$$

Thus $\lim_{n\to\infty} Y_n < \infty$ almost surely (since it's expectation is finite), and consequently, $\sum_{n=1}^{\infty} X_n$ is absolutely convergent almost surely. Let

$$Y = \lim_n Y_n = \sum_{k=1}^{\infty} |X_k|.$$

Now let

$$Z_n = \sum_{k=1}^{n} X_k$$

So $|Z_n| \leq \sum_{k=1}^{n} |X_k| \leq \sum_{k=1}^{\infty} |X_k| = Y$ and $\mathbb{E}(Y) < \infty$ as proved above. Thus by DCT,

$$\lim_n \mathbb{E}(Z_n) = \mathbb{E}(\lim_n Z_n) \implies \lim_n \mathbb{E}(\sum_{k=1}^{n} X_k) = \mathbb{E}(\sum_{k=1}^{\infty} X_k)$$

This gives

$$\sum_{k=1}^{\infty} \mathbb{E}(X_k) = \mathbb{E}(\sum_{k=1}^{\infty} X_k)$$

□

Exercise 2.28. Let $X_n \sim N(0, \frac{1}{n^{2.1}})$. Is $\sum_{n\geq 1} X_n$ finite almost surely?

Example 2.29 (Integrable but not Riemann integrable)**.** A classic example of a function which is not Riemann integrable but Lebesgue integrable is the following. Take our favourite probability space $([0,1], \mathcal{B}([0,1]), \tilde{\lambda})$ where $\tilde{\lambda}$ is the Lebesgue measure restricted to $[0,1]$. Consider function 1_I where I is the set of irrationals in $[0,1]$. This is a function which is Lebesgue integrable, and in fact since this is 1 a.s. $\mathbb{E}(1_I) = 1$. We leave it as an exercise to verify that this function is not Riemann integrable. The punchline is that the infimum of the function in any interval is 0 and supremum on any interval is 1 as rationals are dense.

Example 2.30. Suppose X is integrable. Then

$$\mathbb{E}(X) = \lim_{n\to\infty} \mathbb{E}(X 1_{X\in[-n,n]}).$$

Indeed, $|X|1_{X\in[-n,n]} = |X|1_{|X|\in[0,n]} \leq |X|$ and $\mathbb{E}(|X|) < \infty$ since X is integrable. Also $X1_{X\in[-n,n]} \to X$ almost surely. Thus by DCT, we are done.

Example 2.31. Let $X \sim$Unif$[0,1]$. Let us calculate

$$\lim_{n\to\infty} \mathbb{E}\left((1 - 10e^{-\frac{X^2}{n}})\frac{1}{\sqrt{X}}\right)$$

We could try to apply MCT, but since the integrand is not always non-negative, there is an issue with applying it. Nevertheless, we can apply DCT. We see

$$|(1 - 10e^{-\frac{X^2}{n}})\frac{1}{\sqrt{X}}| \leq \frac{1}{\sqrt{X}} \text{ and } \mathbb{E}(\frac{1}{\sqrt{X}}) = \int_0^1 \frac{1}{\sqrt{x}}dx = 2.$$

Now simply observe,

$$\lim_{n\to\infty} (1 - 10e^{-\frac{X^2}{n}})\frac{1}{\sqrt{X}} = \frac{1}{\sqrt{X}}.$$

Thus by DCT,

$$\lim_{n\to\infty} \mathbb{E}((1 - 10e^{-\frac{X^2}{n}})\frac{1}{\sqrt{X}}) = \mathbb{E}(\frac{1}{\sqrt{X}}) = 2.$$

Example 2.32. Using DCT, we can show that

$$\int_0^\infty e^{-x}\cos(\pi t x)dx$$

is continuous in t. We are going to assume $e^{-x}\cos(\pi t x)$ is continuous in t for a fixed x, which is something we learnt in basic analysis. Now note that $|e^{-x}\cos(\pi t x)| \leq e^{-x}$ for all x, t. Also $\int_0^\infty e^{-x} < \infty$. Thus for any $t_n \to t$, by DCT,

$$\lim_{n\to\infty} \int_0^\infty e^{-x}\cos(\pi t_n x)dx = \int_0^\infty \lim_{n\to\infty} e^{-x}\cos(\pi t_n x)dx = \int_0^\infty e^{-x}\cos(\pi t x)dx$$

by continuity of the function $e^{-x}\cos(\pi t x)$.

Another, perhaps more probabilistic approach, is to note:

$$\int_0^\infty e^{-x}\cos(\pi t x)dx = \mathbb{E}(\cos(\pi t X))$$

and then use bounded convergence theorem.

3 Independence

The goal of this section is to define the notion of **independence** and how to use it. We will define it for objects of increasing complexity.

Definition 3.1 (Independence of events)**.** *We say the events $\{E_i\}_{i\in I}$ are independent (sometimes called **mutually** independent) if for any finite $J \subset I$*

$$\mathbb{P}(\cap_{j\in J} E_j) = \prod_{j\in J} \mathbb{P}(E_j).$$

Definition 3.2. *We say the events $\{E_i\}_{i\in I}$ are **pairwise independent** if*

$$\mathbb{P}(E_i \cap E_j) = \mathbb{P}(E_i)\mathbb{P}(E_j) \text{ for all } i \neq j, \qquad i, j \in I.$$

Exercise 3.1. Suppose $\{A_k\}_{k\geq 1}$ are independent. Then show that any sequence of the form $\{B_k\}_{k\geq 1}$ where each B_k is either A_k or A_k^c is independent. Use induction.

Independence of σ-algebras The definition is similar. In fact we will define independence for an arbitrary collection of sets rather than sigma algebras.

Definition 3.3 (independence of collections of events)**.** *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Define a collection $\{\mathcal{F}_i\}_{i\in I}$ such that $\mathcal{F}_i \subset \mathcal{F}$ for all i. We say the collection $\{\mathcal{F}_i\}_{i\in I}$ is independent if for all $J \subset I$, and $\{E_j\}_{j\in J}$, where $E_j \in \mathcal{F}_j$ for all $j \in J$, the collection $\{E_j\}_{j\in J}$ is independent.*

Note that the definition makes sense even if $\mathcal{F}_i$ are not σ-algebras. Let us emphasize that the independence criterion must be valid for *all* events in the σ-algebra. Fortunately, there is a simpler way to test the independence of σ-algebras.

Proposition 3.2. *Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $\mathcal{P}, \mathcal{Q}$ be π systems in $\mathcal{F}$. Suppose*

$$\mathbb{P}(A \cap B) = \mathbb{P}(A)\mathbb{P}(B) \text{ for all } A \in \mathcal{P}, B \in \mathcal{Q}.$$

Then $\sigma(\mathcal{P}), \sigma(\mathcal{Q})$ are independent.

Proof. We are going to apply the uniqueness theorem Theorem 1.2 twice. First fix $A \in \mathcal{P}$. Consider the measure

$$\mu_A(B) := \mathbb{P}(A)\mathbb{P}(B), \qquad B \in \sigma(\mathcal{P}). \qquad \mathbb{P}_A(B) := \mathbb{P}(A \cap B), \qquad B \in \sigma(\mathcal{P})$$

Since μ_A and $\mathbb{P}_A$ both agree on $\mathcal{Q}$, by the uniqueness Theorem 1.2, $\mu_A(B) = \mathbb{P}_A(B)$ for all $B \in \sigma(\mathcal{Q})$ (we don't need to check the second condition in Theorem 1.2 as we are dealing with probability measures.)

Since A is an arbitrary element in $\sigma(\mathcal{Q})$ we are done. □

Exercise 3.3. Extend the above to a general collection of σ-fields. That is, if $\{\mathcal{A}_j\}_{j\in J}$ are collections of sets such that $\mathcal{A}_j \subset \mathcal{F}$ for all $j \in J$, and $\mathcal{A}_j$ are π-systems then $\{\sigma(\mathcal{A}_j)\}_{j\in J}$ are independent.

Independence of random variables Let us start by recalling the definition of the σ-algebra which is generated by a random variable.

Definition 3.4. *Suppose Ω is a sample space, and suppose $X : \Omega \mapsto \mathbb{R}$ is a random variable defined on it. We define $\sigma(X)$ to be the smallest sigma algebra which makes X measurable. In other words,*

$$\sigma(X) = \cap\{\mathcal{F} : \mathcal{F} \text{ is a } \sigma\text{-algebra}, X^{-1}(E) \in \mathcal{F} \text{ for all } E \in \mathcal{B}(\mathbb{R})\}.$$

In other words,

$$\sigma(X) = \sigma(\{X^{-1}(A), A \in \mathcal{B}(\mathbb{R})\}) = \sigma(\{X^{-1}((-\infty, x]), x \in \mathbb{R}\}).$$

Definition 3.5. *σ-algebra generated by a collection of random variables $\{X_i\}_{i\in I}$, denoted $\sigma(\{X_i\}_{i\in I})$, is the smallest σ-algebra which makes all of the X_is measurable. In other words, we have*

$$\sigma(\{X_i\}_{i\in I}) = \sigma(X_i^{-1}(A), i \in I, A \in \mathcal{B}(\mathbb{R})).$$

Note that

$$\sigma(\{X_i\}_{i\in I}) = \sigma(\{\sigma(X_i) : i \in I\}).$$

Definition 3.6 (Independence)**.** *We say a collection of random variables $\{X_i\}_{i\in I}$ are independent if $\{\sigma(X_i) : i \in I\}$ are independent.*

Definition 3.7 (Independence)**.** *If we have two collections $\{X_i\}_{i\in I}$ and $\{Y_j\}_{j\in J}$, we say $\{X_i\}_{i\in I}$ is independent of $\{Y_j\}_{j\in J}$ if $\sigma(\{\sigma(X_i) : i \in I\})$ is independent of $\{Y_j\}_{j\in J}$*

Exercise 3.4. Suppose $\{X_i\}_{i\in I}$ are independent. Suppose $I_1 \subset I$ and $I_2 \subset I$ such that $I_1 \cap I_2 = \emptyset$. Then $\{X_i\}_{i\in I_1}$ and $\{X_i\}_{i\in I_2}$ are independent. Hint: Use the same idea as in the proof of Proposition 3.2.

Similarly if we have a countable partition $I = \cup_{i=1}^{\infty} I_j$ and $I_j \cap I_{j'} = \emptyset$ if $j \neq j'$ then $\{\sigma(\{X_m\}_{m\in I_j}) : j \geq 1\}$ are independent.

Proposition 3.5 (Test for independence)**.**

$$\mathbb{P}(\cap_j X_j \leq b_j \forall j \in J) = \prod_{j\in J} \mathbb{P}(X_j \leq b_j).$$

Proof. This follows from exercise 3.3 as $\{(-\infty, b] : b \in \mathbb{R}\}$ is a π-system. □

3.1 Construction of independent random variables

In general it is not at all easy to **construct** a sequence of mutually independent events directly. However, here is a construction.

Example 3.6 (A hands on construction of independent events(due to James Norris))**.** Take the probability space $([0,1], \mathcal{B}([0,1]), \lambda|_{[0,1]})$ where λ is as usual the Lebesgue measure restricted to $[0,1]$. Take $A_1 = (0, 1/2]$, $A_2 = (0, 1/4] \cup (1/2, 3/4]$ and in general

$$A_k = \cup_{0 \leq i < 2^{k-1}} (\frac{2i}{2^k}, \frac{2i+1}{2^k}). \text{ Note that } \lambda(A_k) = \frac{1}{2}.$$

Exercise 3.7. Show that $\{A_k\}_{k \geq 1}$ are mutually independent. Start by showing $\{A_1, A_2\}$ are independent. Note that 'half' of A_j is a subset of A_k if $j > k$.

Proposition 3.8. *For any sequence of distributions $(\mu_n)_{n \geq 1}$ (or equivalently distribution functions F_n (cf. Definition 1.10)), we can construct a probability space and a sequence of* ***independent*** *random variables $(X_n)_{n \geq 1}$ on it such that X_n has distribution μ_n.*

Proof sketch. Take the probability space $([0,1], \mathcal{B}([0,1], \lambda)$ where λ is the Lebesgue measure restricted to $[0,1]$. Let A_k be as in Example 3.6. Let $\xi_k = 1_{A_k}$. Note that $\sigma(\xi_k) = \{\emptyset, A_k, A_k^c, \Omega\}$. Thus by exercise 3.1, we see that $(\xi_k)_{k \geq 1}$ are independent. Thus we have constructed an i.i.d. sequence of Bernoulli $(1/2)$ random variables since $\lambda(\xi_k = 1) = \lambda(A_k) = \frac{1}{2}$ (see Example 3.6).

Now let us consider

$$U := \sum_{k \geq 1} \frac{\xi_k}{2^k}.$$

We claim that $U \sim$Unif$[0,1]$. The idea is that this is roughly like a binary expansion where each entry in the expansion is ξ_k which has equal probability to be 0 or 1, hence this leads to a uniform distribution. Here is a rigorous argument. Take any $n \geq 1$ and let $0 \leq m < 2^n$, and binary expand

$$\frac{m}{2^n} = .b_1 \dots b_n.$$

Recall how binary expansion works: once we know m, we divide the interval $[0,1]$ into two equal halves and then let b_1 to be 0 or 1 depending on whether m falls in the right or the left half. Once we know b_1, we divide the half we chose into two equal halves again, and choose b_2 to equal 0 or 1 depending on whether m falls on the left or the right half, and so on. In otherwords,

$$\sum_{i=1}^{n} \frac{b_i}{2^i} = \frac{m}{2^n}.$$

Once we have this, we note that

$$\mathbb{P}(U \in [\frac{m}{2^n}, \frac{m+1}{2^n}) = \mathbb{P}(\xi_1 = b_1, \dots, \xi_n = b_n) = \frac{1}{2^n}$$

To see this, simply note

$$\sum_{k\geq n+1}\frac{\xi_k}{2^k}\leq\sum_{k\geq n+1}\frac{1}{2^k}=\frac{1}{2^n}.$$

Thus we conclude that

$$\mathbb{P}(U\leq r)=r$$

for every number $r\in[0,1]$ of the form $r=\frac{m}{2^n}$, $m\in\mathbb{N},n\in\mathbb{N}$ (such an r is called a *dyadic rational*). Now for any $x\in[0,1]$, choose a sequence of dyadic rational $r_n\downarrow x$. By continuity from above (Lemma 1.15),

$$\mathbb{P}(U\leq x)=\lim_{n\to\infty}\mathbb{P}(U\leq r_n)=\lim_{n\to\infty}r_n=x.$$

Thus $U\sim$Uniform $[0,1]$.

Now the point is that we can sum over *any countable collection of indices* and get a uniform random variable in this way. Now if we can now partition $\mathbb{N}=\cup_{j=1}^{\infty}I_j$ such that I_j s are disjoint and define $U_j=\sum_{k\in I_j}\frac{\xi_k}{2^k}$, we have created i.i.d. uniform random variables. By the proof of proposition 1.36, we are done as it is proved there that X_m has distribution μ_m. □

The extreme opposite of X,Y being independent is that one is completely determined by the other.

We now state a theorem without proof, which is very much believable. For the interested reader, the proof involves measure theory and monotone class theorem.

Theorem 3.1. *Suppose $X_1,\ldots,X_n$ are independent and defined over the probability space $(\Omega,\mathcal{F},\mathbb{P})$. Let $f_1,\ldots,f_n$ be measurable functions. Then*

$$\mathbb{E}(\prod_{i=1}^{n}f_i(X_i))=\prod_{i=1}^{n}\mathbb{E}(f_i(X_i)).$$

Proof. (For the interested reader, not part of the course) See the proof of Theorem 4.9 in [14] □

Applications. The applications of Theorem 3.1 are rather widespread. For example, if X,Y are independent, then

$$\mathbb{E}(\sin(X)e^{X+Y})=\mathbb{E}(\sin(X)e^{X})\mathbb{E}(e^{Y}).$$

[14] Addario–Berry's notes.

3.2 Borel Cantelli Lemmas.

The Borel Cantelli Lemmas are perhaps the most used statements in probability theory. We need two definitions

Definition 3.8.

$$\limsup_{n\to\infty} E_n = \bigcap_{n\geq 1}\bigcup_{m\geq n} E_m = \{\omega \in \Omega : \omega \in E_n \textit{ for infinitely many } n\}. \tag{3.1}$$

$$\liminf_{n\to\infty} E_n = \bigcup_{n\geq 1}\bigcap_{m\geq n} E_m = \{\omega \in \Omega : \omega \in E_n \textit{ for all but finitely many } n\}. \tag{3.2}$$

$$\tag{3.3}$$

In probabilistic language, or in "English",

$$\omega \in \limsup_{n\to\infty} E_n \implies \omega \in E_n \text{ for infinitely many } n.$$

Thus we also say

$$\limsup_n E_n = \{E_n \text{ occurs infinitely often}\} \text{ or simply } \{E_n \text{ occurs i.o.}\}$$

Similarly,

$$\liminf_n E_n = \{E_n \text{ occurs for all but finitely many } n\} \text{ or simply } \{E_n \text{ occurs } \textit{eventually}\}$$

Example 3.9. Suppose there are infinitely many clocks, each running an amount of time given by X_n where $X_n \sim$Exp(1). Assume all the X_ns are defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let

$$A_n = \{X_n \geq n\}$$

and

$$B_n = \{X_n \geq \log n\}$$

and

$$C_n = \{X_n \geq 2\log n\}.$$

(notice how the events are different for each n). Then

$$\limsup A_n = \{A_n \text{ occurs i.o.}\} = \{X_n \geq n \text{ for infinitely many } n\}.$$

Similarly we can write down

$$\limsup_n B_n = \{X_n \geq \log n \text{ for infinitely many } n\}$$

and

$$\limsup_n C_n = \{X_n \geq 2\log n \text{ for infinitely many } n\}.$$

Also

$$\liminf_n A_n = \{A_n \text{ occurs evenually for all large enough } n\} = \{X_n \geq n \text{ evenually for all large enough } n\}.$$

Another lengthy way of saying the same thing is that

$$\liminf A_n = \{\omega : \exists N(\omega) \text{ such that } X_n(\omega) \geq n \text{ for all } n \geq N(\omega)\}.$$

Exercise 3.10. Write down similarly $\liminf_n B_n$ and $\liminf_n C_n$ in "English" or "probabilistic language".

Exercise 3.11. Show that

$$(\limsup_n E_n)^c = \liminf_n (E_n^c)$$

by simply writing down the definitions and applying De-Morgan.

Lemma 3.12 (First Borel–Cantelli lemma)**.** *Let $\{E_n\}_{n\geq 1}$ be a collection of events defined on the same probability space $(\Omega, \mathcal{F}, P)$. Then if*

$$\sum_{n\geq 1} \mathbb{P}(E_n) < \infty$$

then

$$\mathbb{P}(E_n \text{ occurs infinitely often}) = \mathbb{P}(\limsup_n E_n) = 0.$$

Proof. Fix $\varepsilon > 0$. Since $\sum_{n\geq 1} \mathbb{P}(E_n) < \infty$, there exists an n_0 such that for all $n \geq n_0$,

$$\sum_{n\geq n_0} \mathbb{P}(E_n) < \varepsilon.$$

Now notice that

$$\mathbb{P}(\limsup_n E_n) = \mathbb{P}(\bigcap_{n\geq 1} \bigcup_{m\geq n} E_m) \leq \mathbb{P}(\bigcup_{m\geq n_0} E_m) \leq \sum_{n\geq n_0} \mathbb{P}(E_m) < \varepsilon.$$

whihc completes the proof. □

Lemma 3.13 (Second Borel–Cantelli lemma)**.** *Suppose $\{E_n\}_{n\geq 1}$ is a collection of* ***independent*** *events defined on the same probability space $(\Omega, \mathcal{F}, P)$. Then if*

$$\sum_{n=1}^{\infty} \mathbb{P}(E_n) = \infty,$$

then

$$\mathbb{P}(E_n \text{ occurs infinitely often}) = \mathbb{P}(\limsup_n E_n) = 1.$$

The assumption of independence in the second Borel–Cantelli lemma is crucial as we shall see in applications later.

Proof. By Exercise 3.11, $(\limsup_n E_n)^c = \liminf E_n^c$. Thus

$$\mathbb{P}((\limsup_n E_n)^c) = \mathbb{P}(\liminf E_n^c) = \mathbb{P}(\bigcup_{n\geq 1}\bigcap_{m\geq n} E_m^c) \leq \sum_{n=1}^{\infty} \mathbb{P}(\bigcap_{m\geq n} E_m^c)$$

We will show that each summand above is 0. For any $N \geq n$, simply by inclusion of events,

$$\mathbb{P}(\bigcap_{m\geq n} E_m^c) \leq \mathbb{P}(\bigcap_{m=n}^{N} E_m^c)$$

Let $p_m = \mathbb{P}(E_m)$ to simplify notation. Notice

$$\prod_{m=n}^{N} \mathbb{P}(E_m^c) = \prod_{m=n}^{N} (1 - \mathbb{P}(E_m)) = \prod_{m=n}^{N} (1 - p_m) \leq e^{-\sum_{m=n}^{N} p_m}$$

since $\sum_{m\geq n} p_m = \infty$ by assumption, for any $\varepsilon > 0$, we can choose N large enough so that $e^{-\sum_{m=n}^{N} p_m} < \varepsilon$. The proof is complete since the choice of ε is arbitrary. □

Remark 3.14 (Infinite monkey theorem)**.** Borel-Cantelli lemmas lie at the heart of various 'paradoxes' like the infinite monkey theorem, see this wiki link. Namely, a monkey randomly hitting the typewriter, will almost surely type the full works of Shakespeare almost surely. In fact, they will produce the full works infinitely many times with prob. 1! In other words, if some event has positive probability, no matter how small, is bound to happen.

Here is a quick argument of why this is the case. Let M be the number of characters in the full works of Shakespeare. The probability that the monkey randomly produces exactly this sequence of characters is 27^{-M} [15] (an astronomically small number, still positive!). If we break up the characters produced by the monkey into segments $(I_1, I_2, \ldots)$ of length M, and let $\xi_i \sim$Bernoulli (27^{-M}) counts if segment i produced the desired full works of Shakespeare, then by Borel–Cantelli, assuming independence of ξ_i, $\sum_{i\geq 1} \xi_i = \infty$ almost surely.

Let us get back to the example we were dealing with.

Example 3.15. Suppose there are infinitely many clocks, each running an amount of time given by X_n where $X_n \sim$Exp(1). Assume all the X_ns are defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let

$$A_n = \{X_n \geq n\}$$

Notice

$$\mathbb{P}(A_n) = \mathbb{P}(X_n \geq n) = \int_n^{\infty} e^{-t} dt = e^{-n}$$

[15] including 'space'

so

$$\sum_n \mathbb{P}(A_n) = \sum_{n=1}^{\infty} e^{-n} < \infty$$

So

$$\mathbb{P}(A_n \text{ occurs i.o. }) = 0.$$

In words,

$$\text{almost surely, } X_n \leq n \text{ for all large enough } n.$$

Let us turn to

$$B_n = \{X_n \geq \log n\}$$

By the same logic,

$$\sum_n \mathbb{P}(B_n) = \sum_{n=1}^{\infty} e^{-\log n} = \sum_{n=1}^{\infty} \frac{1}{n} = \infty$$

So we cannot say anything more. If we had the further information that B_n are independent, then

$$\mathbb{P}(B_n \text{ occurs i.o.}) = 1$$

or in other words, if X_ns are independent then

$$\text{almost surely, } X_n \geq \log n \text{ for infinitely many } n.$$

Exercise 3.16. Show that almost surely $\{X_n \leq 2 \log n\}$ for all large enough n.

Example 3.17. Suppose X_n is a sequence of random variables. Does there exist a sequence of numbers $c_n > 0$ such that $X_n/c_n \to 0$ a.s.? Indeed, yes. Since $\mathbb{P}(X_n > t) \to 0$ as $t \to \infty$, we can find c_n such that $\mathbb{P}(|X_n|/c_n > 2^{-n}) < 2^{-n}$. Since $\sum_{n\geq 1} 2^{-n} < \infty$, we must have $|X_n|/c_n \leq 2^{-n}$ for all large enough n almost surely. Check from the definitions of convergence of sequences that $X_n/c_n \to 0$ a.s.

3.3 Basics of moments

For a random variable X with $\mathbb{E}(X) = \mu \in \mathbb{R}$, we define its **variance** as

$$\text{Var}(X) = \mathbb{E}[(X-\mu)^2] = \mathbb{E}(X^2) - \mu^2 \text{ where } \mu = \mathbb{E}(X).$$

and **Covariance** as

$$\text{Cov}(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y).$$

Exercise 3.18. Prove the second equality in the definition of Variance above.

Definition 3.19. $Cov(X,Y) = \mathbb{E}[(X - \mathbb{E}(X))(Y - \mathbb{E}(Y))]$

Exercise 3.20. Show $\mathbb{E}[(X-\mathbb{E}(X))(Y-\mathbb{E}(Y))] = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)$.

Note that $Cov(X,X) = Var(X)$. Furthermore, from Theorem 3.1, we see that if X, Y are independent, then

$$Cov(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \mathbb{E}(X)\mathbb{E}(Y) - \mathbb{E}(X)\mathbb{E}(Y) = 0.$$

Example. Suppose (X,Y) are indicator random variables, that is, $X = 1_A$ and $Y = 1_B$. Then

$$Cov(X,Y) = \mathbb{P}(X=1, Y=1) - \mathbb{P}(X=1)\mathbb{P}(Y=1)$$

So

$$Cov(X,Y) > 0 \implies \mathbb{P}(X=1|Y=1) > \mathbb{P}(X=1).$$

Tells us that $\mathrm{Cov}(X,Y) > 0$ means that Y increasing means X is increasing (in other words, " B attracts A "). On the other hand, the same logic shows that $Cov(X,Y) < 0$ implies that B repels A.

Definition 3.21. *Also we define Correlation between two random variables X, Y as*

$$Correlation(X,Y) = Cor(X,Y) = \frac{Cov(X,Y)}{\sqrt{Var(X)\cdot Var(Y)}}.$$

Example The joint density of (X,Y) is given by

$$f(x,y) = \begin{cases} 3x, & 0 < y \le x \le 1. \\ 0 \text{ otherwise.} \end{cases}$$

Calculate $Cov(X,Y)$. $Var(X)$, $Var(Y)$, $Corr(X,Y)$.

Solution. Marginal of X: $\int_0^x 3x dy = 3x^2$, if $0 < x < 1$.
Marginal of Y: $\int_y^1 3x dx = \frac{3x^2}{2}|_y^1 = \frac{3}{2}(1-y^2)$, $0 < y < 1$
$\mathbb{E}(X) = \int_0^1 3x^3 dx = \frac{3}{4}$. $\mathbb{E}(X^2) = \int_0^1 3x^4 dx = \frac{3}{5}$.
$Var(X) = \frac{3}{5} - \frac{9}{16} = \frac{3}{80}$
$\mathbb{E}(Y) = \int_0^1 y\frac{3}{2}(1-y^2)dy = \frac{3}{2}[\frac{1}{2} - \frac{1}{4}] = \frac{3}{8}$.
$\mathbb{E}(Y^2) = \int_0^1 y^2\frac{3}{2}(1-y^2)dy = \frac{3}{2}[\frac{1}{3} - \frac{1}{5}] = \frac{1}{5}$.
$Var(Y) = \frac{1}{5} - (3/8)^2 = \frac{19}{320}$.
$\mathbb{E}(XY) = \int_0^1 \int_0^x xy3xdydx = \int_0^1 3x^2(\int_0^x ydy)dx = \int_0^1 \frac{3}{2}x^4 dx = \frac{3}{10}$. So $\mathrm{Cov}(X,Y) = \mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y) = \frac{3}{10} - \frac{9}{32}$.

Properties of Covariance

a. $Cov(X, X) = Var(X)$.

b. $Cov(X, Y) = Cov(Y, X)$ (Verify)

c. $Cov(cX, Y) = cCov(X, Y)$.

d. $Cov(X, Y + Z) = Cov(X, Y) + Cov(Y, Z)$

Exercise 3.22. Verify the above properties, they follow from the definition of Covariance and some algebra.

Example 3.23 (Uncorrelated does not imply independence)**.** Let $X = +1$ with probability $1/2$ and -1 with probability $1/2$, (i.e. $X = \xi - 1$ where $\xi \sim Ber(1/2)$.) If $X = 1$ then $Y = 1000$ with prob $1/2$ and -1000 with prob. $1/2$. If $X = -1$ then $Y = 0$. Then $\mathbb{E}(Y) = 0$, $\mathbb{E}(X) = 0$ and $\mathbb{E}(XY) = 1000 \times 1/4 + (-1000) \times 1/4 = 0$. Therefore $Cov(X, Y) = 0$. But $\mathbb{P}(Y = 0|X = 1) = 0 \neq \mathbb{P}(Y = 0) = \mathbb{P}(X = -1) = 1/2$. So X and Y are not independent.

We can generalize property d. to

$$Cov(\sum_{i=1}^{n} X_i, \sum_{j=1}^{m} Y_j) = \sum_{i=1}^{n} \sum_{j=1}^{m} Cov(X_i, Y_j)$$

Therefore, we obtain a special case:

$$\begin{aligned} \mathrm{Var}(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} \sum_{j=1}^{n} Cov(X_i, X_j) &= \sum_{i=1}^{n} Cov(X_i, X_i) + \sum_{i=1}^{n} \sum_{j \neq i}^{n} Cov(X_i, X_j) \\ &= \sum_{i=1}^{n} Var(X_i) + 2 \sum_{i=1}^{n} \sum_{j<i}^{n} Cov(X_i, X_j) \end{aligned}$$

Corollary 3.24. *If $X_i : 1 \leq i \leq n$ are pairwise uncorrelated ,i.e., $Cov(X_i, X_j) = 0$ for all $i \neq j$ then*

$$Var(\sum_{i=1}^{n} X_i) = \sum_{i=1}^{n} Var(X_i).$$

Example. Here is a neat way to compute the variance of Bin (n, p). We know if $X \sim Bin(n, p)$ then $X = \sum_{i=1}^{n} \xi_i$ where $\xi_i \sim Ber(p)$. Note that $Var(\xi_1)$ is easy to compute

$$Var(\xi_1) = \mathbb{E}(\xi_1^2) - (\mathbb{E}(\xi_1))^2 = p - p^2 = p(1 - p).$$

Therefore

$$Var(X) = nVar(\xi_1) = np(1 - p).$$

Definition 3.25. *For random variables $X_1, \ldots, X_n$,*

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n}$$

is called the ***sample mean.***

Warning: This is not the "mean" or the "expectation" of a random variable which is a number. Sample mean is itself a random variable!

Proposition 3.26. *If $X_1, \ldots, X_n$ are i.i.d. with mean μ and Variance σ^2. Then*

1. $\mathbb{E}(\bar{X}) = \mu$.
2. $Var(\bar{X}) = \sigma^2/n$.
3. $Cov(\bar{X}, X_i - \bar{X}) = 0$

Proof. $\mathbb{E}(\bar{X}) = n\mu/n = \mu$ and $Var(\bar{X}) = \frac{1}{n^2} Var(\sum_{i=1}^n X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$. Finally note

$$Cov(\bar{X}, X_i) = \frac{1}{n} \sum_{j=1}^{n} Cov(X_j, X_i) = \frac{\sigma^2}{n}.$$

Therefore

$$Cov(\bar{X}, X_i - \bar{X}) = Cov(\bar{X}, X_i) - Cov(\bar{X}, \bar{X}) = \frac{\sigma^2}{n} - \sigma^2/n = 0.$$

□

4 Modes of convergence of random variables.

Recall the definition of almost sure convergence, that we already introduced.

Definition 4.1 (Almost sure convergence)**.** *We say X_n almost surely converges to a random variable X if we set*

$$\Omega_0 = \{\omega \in \Omega : X_n(\omega) \text{ converges to } X(\omega)\}$$

then

$$\mathbb{P}(\Omega_0) = 1.$$

We now introduce a new notion of convergence, which will turn out to be weaker than a.s. convergence.

Definition 4.2 (Convergence in probability)**.** *Suppose $\{X_n\}_{n\geq 1}$, X be random variables defined on the same probability space. We say X_n converges to X in probability if for all $\varepsilon > 0$,*

$$\mathbb{P}(|X_n - X| > \varepsilon) \xrightarrow[n\to\infty]{} 0$$

The notation is

$$X_n \xrightarrow[n\to\infty]{P} X.$$

To prove convergence in probability, the following inequalities are extremely useful.

Proposition 4.1. *Suppose X is a.s. non-negative. For any $a > 0$,*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}X}{a}$$

Proof. Simply note

$$\begin{aligned}\mathbb{E}(X) &= \mathbb{E}(X1_{X\geq a}) + \mathbb{E}(X1_{X\leq a}) \\ &\geq a\mathbb{P}(X \geq a).\end{aligned}$$

which after rearranging yields the desired inequality. □

Proposition 4.2 (Chebyshev's inequality)**.** *Let X be a random variable with mean μ and Variance σ^2. Then for any $a > 0$,*

$$\mathbb{P}(|X - \mu| \geq a) \leq \frac{\sigma^2}{a^2}$$

Proof. Since $(X - \mu)^2 \geq 0$, applying Markov's inequality,

$$\mathbb{P}((X - \mu)^2 \geq a^2) \leq \frac{\mathbb{E}((X - \mu)^2)}{a^2} = \frac{\sigma^2}{a^2}.$$

□

We can now present a weak version of a law of large numbers. Also it is useful to recall the following useful properties of variance.

Theorem 4.3 (Weak law of large numbers)**.** *Let $X_1, X_2, \ldots$ be i.i.d. with $\mathbb{E}(X_1) = \mu$ and $Var(X_1) = \sigma^2$. Let $S_n = \sum_{i=1}^n X_i$. Then*

$$\frac{S_n}{n} \xrightarrow[n\to\infty]{P} \mu.$$

Proof. Note that for all $\varepsilon > 0$, by Chebyshev

$$\mathbb{P}(|S_n/n - \mu| > \varepsilon) \leq \frac{\mathrm{Var}(S_n/n)}{\varepsilon^2} = \frac{n\sigma^2}{n^2\varepsilon^2} = \frac{\sigma^2}{n\varepsilon^2} \to 0$$

as desired. □

In fact we will see that a much stronger version of the weak law of large numbers is true, which is called the strong law of large numbers. We will prove it soon. But let us prove some relationships between these various notions of convergence.

Lemma 4.4. *If X_n converges almost surely to X, then X_n converges to X in probability.*

Proof. Fix $\varepsilon > 0$. Notice that

$$\Omega_0 := \{\omega : \lim_{n\to\infty} |X_n(\omega) - X(\omega)| = 0\} \subseteq \{\omega : \limsup_n |X_n(\omega) - X(\omega)| \leq \varepsilon\}.$$

Thus letting $Z_n = 1_{|X_n(\omega)-X(\omega)|\leq\varepsilon}$, we see that $\{\omega : Z_n(\omega) \to 1\} \supseteq \Omega_0$ and consequently $Z_n \to 1$ almost surely. Also $|Z_n| \leq 1$ for all n. Thus by DCT

$$\lim \mathbb{E}(Z_n) = \mathbb{E}(1) = 1 \implies \lim \mathbb{P}(|X_n - X| \leq \varepsilon) = 1$$

Consequently

$$\lim_{n\to\infty} \mathbb{P}(|X_n - X| > \varepsilon) = 0$$

which completes the proof. □

The converse is not true. Here is an example.

Example 4.5. Suppose ξ_n be i.i.d. Bernoulli $(1/n)$. Then for any $\varepsilon > 0$, $\mathbb{P}(|\xi_n| > \varepsilon) = \mathbb{P}(\xi_n = 1) = 1/n \to 0$. But

$$\sum_{n=1}^{\infty} \mathbb{P}(\xi_n = 1) = \sum_{n=1}^{\infty} \frac{1}{n} = \infty$$

Thus by Borel–Cantelli,

$$\mathbb{P}(\xi_n = 1 \text{ i.o.}) = 1$$

which means

$$\mathbb{P}(\{\omega : \lim_{n\to\infty} \xi_n(\omega) = 0\} = 0.$$

So the probability that X_n converges to X has to decay "fast enough" so that almost sure convergence also holds.

Lemma 4.6. *If X_n converges to X in probability then there exists a subsequence $\{n_k\}_{k\geq 1}$ such that $X_{n_k} \to X$ almost surely as $k \to \infty$.*

Proof. We will leverage the fact that $\mathbb{P}(|X_n - X| > \varepsilon) \to 0$ *for all* $\varepsilon > 0$ and also employ Borel Cantelli to choose a fast growing subsequence.

For any $k \in \mathbb{N}$, we know that

$$\mathbb{P}(|X_n - X| \geq \frac{1}{k}) \to 0.$$

as $n \to \infty$. Therefore, we can choose an n_k large enough such that

$$\mathbb{P}(|X_{n_k} - X| \geq \frac{1}{k}) \leq 2^{-k}.$$

The key is that the right hand side above is chosen to decay fast enough so that it is summable. Thus for a fixed $m \in \mathbb{N}$,

$$\sum_{k=1}^{\infty} \mathbb{P}(|X_{n_k} - X| \geq \frac{1}{m}) \leq m + \sum_{k\geq m} 2^{-k} < \infty.$$

Since for $k \geq m$, $\mathbb{P}(|X_{n_k} - X| \geq \frac{1}{m}) \leq \mathbb{P}(|X_{n_k} - X| \geq \frac{1}{k}) \leq 2^{-k}$ and we bound the probability trivially by 1 for $k < m$. Overall, applying the first Borel–Cantelli lemma, for every $m \in \mathbb{N}$, $|X_{n_k} - X| < \frac{1}{m}$ for all large enough k. In other words, for every $m \in \mathbb{N}$,

$$\mathbb{P}(\limsup_k |X_{n_k} - X| \geq \frac{1}{m}) = 0.$$

But if $\limsup_k |X_{n_k} - X| > 0$, there must be some $m \in \mathbb{N}$, such that $\limsup_k |X_{n_k} - X| \geq \frac{1}{m}$. Thus

$$\begin{aligned}\mathbb{P}(\limsup_k |X_{n_k} - X| > 0) &\leq \mathbb{P}(\cup_{m=1}^{\infty}\{\limsup_k |X_{n_k} - X| \geq \frac{1}{m}\}) \\ &\leq \sum_{m=1}^{\infty} \mathbb{P}(\limsup_k |X_{n_k} - X| \geq \frac{1}{m}) = \sum_{m=1}^{\infty} 0 = 0.\end{aligned}$$

□

We record a partial converse.

Proposition 4.7. *$X_n \xrightarrow[n\to\infty]{P} X$ if and only if for every subsequence $\{n_k\}_{k\geq 1}$ there exists a further subsequence $\{n_{k_\ell}\}_{\ell\geq 1}$ such that $X_{n_{k_\ell}} \xrightarrow[\ell\to\infty]{a.s.} X$*

Proof. $X_n \xrightarrow[n\to\infty]{P} X$ implies $X_{n_k} \xrightarrow[k\to\infty]{P} X$, so one can extract a subsequence using Lemma 4.6.

For the converse, fix an $\varepsilon > 0$ consider the sequence $\{a_n\}_{n\geq 1} = \{\mathbb{P}(|X_n - X| > \varepsilon)\}_{n\geq 1}$. For this sequence, for every subsequence there is a further subsequence that converges to 0 (since almost sure convergence of the subsequence implies convergence in probability). This means $\{a_n\}$ converges to 0 as the set of all subsequential limits of $(a_n)_{n\in\mathbb{N}}$ has to be 0. □

Remark 4.8. Proposition 4.7 shows that almost sure convergence cannot describe a topology on the space of random variables. If it did, then if for a sequence, we could extract a subsequence of every subsequence which converges almost surely, then the sequence would have to converge almost surely. But we know that there are random variables which converge in probability but not almost surely, yet Proposition 4.7 ensures that for every subsequence we can extract a further subsequence which converges almost surely.

Convergence in distribution The third notion of convergence is convergence in distribution. This is a statement about convergence of the probability measures μ_{X_n} so in general the X_n need not be defined on arbitrary probability spaces. One would have liked the definition to be '$\mu_n \to \mu$ if $\mu_n(A) \to \mu(A)$ for all Borel A', but in reality that does not work, we need something more sophisticated. The overarching theorem in this context is known as the *Portmanteau theorem*, but we will be content with a weaker version of that theorem for real values random variables.

Recall the definition of cumulative distribution function $F_X = \mathbb{P}(X \leq x)$. This notion of convergence only deals with convergence of the probability measures without referring at all to the probability spaces on which the random variables are defined.

Definition 4.3 (Convergence in distribution)**.** *We say X_n converges in distribution to X, with notation*

$$X_n \xrightarrow[(d)]{n\to\infty} X$$

if for all continuity points x *of* F_X*,*

$$F_{X_n}(x) \xrightarrow[n\to\infty]{} F_X(x).$$

The assumption of convergence at all continuity points is necessary as otherwise many natural examples will fail to converge in distribution. For example, define

$$\mathbb{P}(\xi_n = 1 + \frac{1}{n}) = \frac{1}{2} = \mathbb{P}(\xi_n = 0) \tag{4.1}$$

We would like ξ_n to converge to a Bernoulli $(1/2)$ random variable ξ. But at $x = 1$, (a non-continuity point of ξ),

$$F_\xi(1) = 1 \text{ but } F_{\xi_n}(1) = \frac{1}{2} \text{ for all } n$$

Thus $F_{\xi_n}(1) \not\to F_\xi(1)$. However, the convergence holds at all other points. Hence ξ_n does converge in distribution to ξ according to the definition.

Proposition 4.9. *If $X_n \xrightarrow[n\to\infty]{P} X$ in probability, then $X_n \xrightarrow[n\to\infty]{(d)} X$.*

Proof. Notice that since F_X is monotone, there can be at most countably many points where $F_X(x-) < F_X(x)$ (Exercise: check!). Thus we can ignore those points and still uncountably points remaining which are continuity points of X. Take a continuity point x of F_X and fix an ε. By continuity of F_X at x, we can ensure that

$$F_X(x+\delta) \in (F_X(x), F_X(x) + \varepsilon) \text{ and } F_X(x-\delta) \in (F_X(x), F_X(x) - \varepsilon).$$

Now note

$$\{X_n \le x\} \subseteq \{X \le x+\delta\} \cup \{|X_n - X| > \delta\} \text{ and } \{X \le x-\delta\} \subseteq \{X_n \le x\} \cup \{|X_n - X| > \delta\}.$$

Thus taking limsup of the probability of the left hand side,

$$\limsup_n \mathbb{P}(X_n \le x) \le \mathbb{P}(X \le x+\delta) + \limsup_n \mathbb{P}(|X_n - X| > \delta) < \varepsilon$$

for the choice of δ by the convergence in probability assumption. Similarly, taking the liminf of the probability of the right hand side,

$$\begin{aligned} \mathbb{P}(X \le x-\delta) &\le \mathbb{P}(X_n \le x) + \mathbb{P}(|X_n - X| > \delta) \\ &\implies \liminf_n \mathbb{P}(X_n \le x) \ge \mathbb{P}(X \le x-\delta) - \liminf_n \mathbb{P}(|X_n - X| > \delta) < \varepsilon. \end{aligned}$$

which completes the proof. □

The converse of Proposition 4.9 is not true. If $X, X_1, X_2, \ldots$ are i.i.d. Bernoulli(1/2) defined on the same probability space then clearly $X_n \to X$ in distribution, but $\mathbb{P}(|X_n - X| > 1/4) \geq \mathbb{P}(X_n = 0, X = 1) \geq \frac{1}{4} \not\to 0$.

However in the special case when the limiting random variable is a constant, the following is true.

Proposition 4.10. *If $X_n \to c$ in distribution where $c \in \mathbb{R}$ is a constant, then $X_n \to c$ in probability.*

Proof. Suppose $X_n \to c$ in distribution. Then for all $\varepsilon > 0$, $F_{X_n}(c + \varepsilon) \to 1 \implies \mathbb{P}(X_n > c + \varepsilon) \to 0$ and $F_X(c - \varepsilon) \to 0$ (c is the only point of discontinuity of the distribution function of the degenerate random variable which takes the value c with probability 1). Then

$$\mathbb{P}(|X_n - c| > \varepsilon) = \mathbb{P}(\{X_n > c + \varepsilon\} \cup \{X_n < c - \varepsilon\}) \leq \mathbb{P}(X_n - c > \varepsilon) + \mathbb{P}(X_n - c < -\varepsilon) \to 0 + 0 = 0$$

as desired. □

To summarize:

A.s. convergence $\implies$ Convergence in probability $\implies$ convergence in distribution.

And finally, convergence in distribution to a constant implies convergence in probability.

4.1 Laws of large numbers

Theorem 4.1 ((Weak version of) Strong law of large numbers)**.** *Suppose $X_1, \ldots,$ are i.i.d. with $\mathbb{E}(X_1^4) < \infty$ and $\mathbb{E}(X_1) = \mu$. Then*

$$\frac{X_1 + \ldots + X_n}{n} \xrightarrow[n \to \infty]{a.s.} \mu.$$

Proof. We can assume without loss of generality that $\mu = 0$, since otherwise we can take $\tilde{X}_i = X_i - \mu$ and prove the result for $\tilde{X}_i$. Let $S_n = X_1 + \ldots + X_n$. By Markov's inequality Proposition 4.1

$$\mathbb{P}\left(\frac{|S_n|}{n} \geq \varepsilon\right) \leq \frac{\mathbb{E}(S_n^4)}{\varepsilon n^4} = \frac{\sum_{1 \leq i_1, i_2, i_3, i_4} \mathbb{E}(X_{i_1} X_{i_2} X_{i_3} X_{i_4})}{n^4 \varepsilon^4}$$

Now we claim

$$\frac{\mathbb{E}(X_{i_1} X_{i_2} X_{i_3} X_{i_4})}{n^4 \varepsilon^4} \leq \frac{C}{n^2}$$

for some constant $C > 0$, which is summable. Using the claim we can conclude by first Borel–Cantelli and and argument similar to the proof of Lemma 4.6.

To prove the claim, note that by independence and since the mean is assumed to be 0, any term where $i_1 \notin \{i_2, i_3, i_4\}$ vanishes and similarly for i_2, i_3, i_4. Thus the only terms which remain are those with $i_1 = i_2$ and $i_3 = i_4$. The number of such terms is at most

$$n + \binom{n}{2}\binom{4}{2} \leq Cn^2$$

as desired. □

Proposition 4.11 (Durrett Theorem 2.3.9.)**.** *Suppose* $(A_n)_{n\geq 1}$ *are pairwise independent events. Let* $S_n = \sum_{k=1}^n 1_{A_k}$ *and* $\mathbb{E}(S_n) \to \infty$ *as* $n \to \infty$*. Then*

$$\frac{S_n}{\mathbb{E}(S_n)} \to 1 \qquad a.s.$$

Proof. We first prove something weaker: convergence in probability. Note

$$\mathbb{P}(\left|\frac{S_n}{\mathbb{E}(S_n)} - 1\right| > \varepsilon) \leq \frac{\mathrm{Var}(S_n)}{\varepsilon^2(\mathbb{E}(S_n))^2}$$

Since 1_{A_k} are pairwise independent, $\mathrm{Var}(S_n) = \sum_{k=1}^n \mathrm{Var}(1_{A_k}) = \sum_{k=1}^n \mathbb{P}(A_k)(1-\mathbb{P}(A_k)) \leq \sum_{k=1}^n \mathbb{P}(A_k) = \mathbb{E}(S_n)$. Thus plugging it back

$$\mathbb{P}(\left|\frac{S_n}{\mathbb{E}(S_n)} - 1\right| > \varepsilon) \leq \frac{\mathbb{E}(S_n)}{\varepsilon^2(\mathbb{E}(S_n))^2} = \frac{1}{\varepsilon^2\mathbb{E}(S_n)} \to 0. \tag{4.2}$$

since $\mathbb{E}(S_n) \to \infty$.

Now we could directly use Lemma 4.6 to get almost sure convergence. However, to turn the convergence along subsequence to a full convergence, we need more control on the 'sparseness' of the subsequence. To that end, our hope we want go for the slowest rate of convergence possible so that we can still have almost sure convergence. It turns out that a polynomial decay is enough.

Let us implement the above idea. We can choose a subsequence n_k such that n_k is the smallest m such that $\mathbb{E}(S_m) \geq k^2$. Note

$$\mathbb{P}(\left|\frac{S_{n_k}}{\mathbb{E}(S_{n_k})} - 1\right| > \varepsilon) \leq \frac{1}{\varepsilon\mathbb{E}(S_{n_k})} \leq \frac{1}{\varepsilon k^2}.$$

which is summable. Hence by the first Borel-Cantelli, $|S_{n_k} - \mathbb{E}(S_{n_k})| \leq \varepsilon\mathbb{E}(S_{n_k})$ for all large enough k almost surely. Let A_m be the set on which this inequality holds for all large enough k with $\varepsilon = \frac{1}{m}$. On $\cap_{m\geq 1}A_m$, $S_{n_k}/\mathbb{E}(S_{n_k}) \to 1$ and $\cap_{m\geq 1}A_m$ has probability 1. Thus overall, $S_{n_k}/\mathbb{E}(S_{n_k}) \to 1$ a.s.

Since S_t increases by at most 1 as t increases to $t+1$ (we are only adding indicators), we have $k^2 \leq \mathbb{E}(S_{n_k}) \leq k^2 + 1$ for all k. Furthermore, for every m between n_k and n_{k+1}, $S_{n_k} \leq S_m \leq S_{n_{k+1}}$ a.s. since S_n is non-decreasing (consequently $\mathbb{E}(S_{n_k}) \leq \mathbb{E}(S_m) \leq \mathbb{E}(S_{n_{k+1}})$ as well). Thus for every n such that $n_k \leq n \leq n_{k+1}$,

$$\frac{S_{n_k}}{\mathbb{E}(S_{n_{k+1}})} \leq \frac{S_n}{\mathbb{E}(S_n)} \leq \frac{S_{n_{k+1}}}{\mathbb{E}(S_{n_k})} \qquad \text{a.s.}$$

Note

$$\frac{k^2}{(k+1)^2+1} \leq \frac{\mathbb{E}(S_{n_k})}{\mathbb{E}(S_{n_{k+1}})} \leq \frac{k^2+1}{(k+1)^2}$$

In fact this is the reason for choosing a polynomial sequence k^2 rather than e^{-k}.

and consequently, $\frac{\mathbb{E}(S_{n_k})}{\mathbb{E}(S_{n_{k+1}})} \to 1$ as $k \to \infty$. Thus

$$\frac{S_{n_{k+1}}}{\mathbb{E}(S_{n_k})} = \frac{S_{n_{k+1}}}{\mathbb{E}(S_{n_{k+1}})} \frac{\mathbb{E}(S_{n_{k+1}})}{\mathbb{E}(S_{n_k})} \to 1 \text{ a.s.}$$

and similarly $\frac{S_{n_k}}{\mathbb{E}(S_{n_{k+1}})} \to 1$ a.s. as well. Thus $\frac{S_n}{\mathbb{E}(S_n)}$ is sandwiched between two random variables for all n, both of which converge to 1 a.s. and hence it converges to 1 a.s. as well. □

4.2 Skorokhod representation theorem

Note that the properties in Section 2.2 mostly require almost sure convergence. Can we replace a.s. convergence by convergence in probability? The following theorem allows us to bypass this issue.

Theorem 4.2 (Skorokhod representation theorem)**.** *Suppose*

$$X_n \xrightarrow[n\to\infty]{(d)} X.$$

Then there exists a probability space and a collection of random variables $\{Y_n\}_{n\geq 1}$ and Y defined on that space such that Y_n and X_n have the same distribution, X and Y have the same distribution, and furthermore

$$Y_n \xrightarrow[n\to\infty]{a.s.} Y$$

Proof sketch. Take our favourite probability space $([0,1], \mathcal{B}([0,1]), \lambda)$ where λ is the Lebesgue measure on $[0,1]$. Let $F_n = F_{X_n}$ denote the cdf of X_n. Define

$$Y_n(p) = \inf\{x : F_n(x) \geq p\}.$$

We now complete the proof on the special case that $\{F_n\}_{n\geq 1}, F$ are all one-one and onto with range $(0,1)$ so that the inverse exists. This means that the definition simplifies to

$$Y_n(p) = F_n^{-1}(p); \qquad Y(p) = F^{-1}(p) \text{ for } p \in (0,1).$$

Note

$$\lambda(Y_n \leq t) = \lambda(\{p : F_n^{-1}(p) \leq t\}) = \lambda(\{p : p \leq F_n(t)\}) = F_n(t) = \mathbb{P}(X_n \leq t)$$

So Y_n and X_n have the same distribution. Similarly Y and X also have the same distribution. All that is left to show is that Y_n converges almost surely to Y. By assumption, any $t \in \mathbb{R}$ is a continuity point of F (since they are continuous by assumption), we have $F_n(t) \to F(t)$ by definition. So take $t_n = F_n^{-1}(p) = Y_n(p)$ and $t = F^{-1}(p) = Y(p)$. So $F_n(t_n) = p = F(t)$ and $F_n(t) \to F(t) = p$. We need to show $t_n \to t$.

Fix $\varepsilon > 0$ and pick $t - 2\varepsilon < z < t - \varepsilon$. If $t_n < t - 2\varepsilon$ infinitely often, then $F_n(t_n) \leq F_n(z)$ infinitely often by monotonicity. But $F_n(z) \to F(z)$ by continuity and $F(z) < F(t) - \delta = p - \delta$ for some $\delta > 0$ by strict monotonicity. Thus $F_n(t_n) < p - \delta$ infinitely often, which is a contradiction since $F_n(t_n) = p$ for all n. One can similarly show that $t_n > t + 2\varepsilon$ infinitely often leads to a contradiction. Since ε is arbitrary, $t_n \to t$.

Exercise 4.12. Read the proof of the general case from Addario–Berry's notes http://problab.ca/louigi/courses/2019/math587/587notes.pdf Theorem 3.10.

□

Remark 4.13. The joint law of Y_ns are no longer necessarily equal to the joint law of X_n, if specified. The main application of Skorokhod representation is to recover some theorems about distributions of X_n which only assumed almost sure convergence before (like DCT, Fatou etc.) in situations where almost sure convergence is replaced by distributional convergence or convergence in probability.

We now state a couple of applications of Skorokhod representation theorem.

Lemma 4.14 (Fatou's lemma)**.** *Suppose $X_n \geq 0$ for all n and $X_n \to X$ in probability. Then*

$$\liminf_n \mathbb{E}(X_n) \geq \mathbb{E}(\liminf_n X_n)$$

Proof. Apply Skorokhod, use the previous Fatou (2.3), and conclude. □

Proposition 4.15. *$X_n \xrightarrow[n\to\infty]{(d)} X$ if and only if for every bounded continuous function $f : \mathbb{R} \mapsto \mathbb{R}$, $\mathbb{E}(f(X_n)) \to \mathbb{E}(f(X))$ as $n \to \infty$ (where the expectations are taken with respect to the respective probability spaces if the variables are defined on different probability spaces).*

Proof sketch. $X_n \xrightarrow[n\to\infty]{(d)} X$, lift to a space with random variables Y_n, Y defined on them having the same distribution as X_n, X respectively so that Y_n converges to Y almost surely. By Dominated convergence theorem, $\mathbb{E}(f(Y_n)) \to \mathbb{E}(f(Y))$. We conclude since $\mathbb{E}(f(X_n)) = \mathbb{E}(f(Y_n))$ and $\mathbb{E}(f(Y)) = \mathbb{E}(f(X))$ as they have the same distribution.

For the converse, note that we need to show that for every continuity point of F_X, $\mathbb{E}(1_{X_n \leq x}) \to \mathbb{E}(1_{X \leq x})$. Thus we need to use the bounded function $f(y) = 1_{y \leq x}$ for $y \in \mathbb{R}$. Unfortunately this function is not continuous. So there is some work needed to approximate it by a bounded continuous function and take limit, which we skip (one way is to convolve it with a bump function). □

4.3 Kolmogorov 0-1 law.

We start with the picture of Andrey Nikolaevich Kolmogorov, who can be safely considered to be one of the founding fathers of probability theory. He had a pretty wild life, check out the wiki page for a quick idea.

Suppose we have a countable collection of random variables $\{X_n : n \geq 1\}$ defined on a probability space $\{\Omega, \mathcal{F}, \mathbb{P}\}$. First, we need to introduce a special σ-algebra $\mathcal{T} \subset \mathcal{F}$ called the **tail σ-algebra**. This brings us to the following definition.

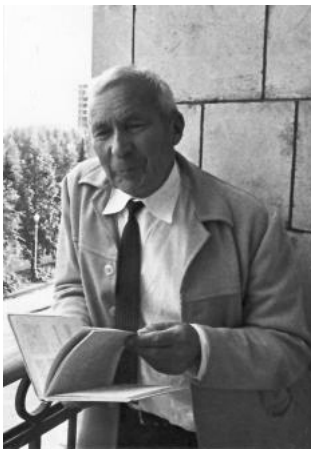

Figure 3: Andrey Nikolaevich Kolmogorov 1903-87.

Definition 4.16 (Tail σ-algebra)**.** *Given $\{X_n : n \geq 1\}$, the tail σ-algebra is given by*

$$\bigcap_{n\geq 1} \sigma(\{X_m : m \geq n\}) = \bigcap_{n\geq 1} \sigma(X_n, X_{n+1}, \ldots).$$

What kind of events are there in this σ-algebra.

- The event
$$\mathcal{L} := \{\omega : \lim X_n \text{ exists } \}$$
is in $\mathcal{T}$. The "hand wavy" argument is that *pick any any $\omega \in \mathcal{L}$ and change the values of of the sequence $\{X_n(\omega)\}$ (to anything) for **finitely many** indices n. Then ω is still in $\mathcal{L}$. Conversely, pick any any $\omega \notin \mathcal{L}$ and change the values of of the sequence $\{X_n(\omega)\}$ (to anything) for **finitely many** indices n. Then ω is still $\notin \mathcal{L}$* . In other words, the values of X_n for finitely many n does not determine whether the limit exists or not.

- The same logic gives
$$\{\omega : X_n(\omega) = 0 \text{ infinitely often}\}$$
is in $\mathcal{T}$.

- In fact for any sequence of Borel sets $B_n \in \mathcal{B}$, we have
$$\{\omega : X_n(\omega) \in B_n \text{ infinitely often}\}$$
is in $\mathcal{T}$.

Exercise 4.17. Prove that the two examples above form examples of tail events.

Theorem 4.3. *Suppose $\{X_n\}_{n\geq 1}$ is a collection of* ***independent*** *random variables. Let $A \in \mathcal{T}$ where*

$$\mathcal{T} = \bigcap_{n\geq 1} \sigma(X_n, X_{n+1}, \ldots)$$

is the tail σ-algebra. Then $\mathbb{P}(A)$ is either 0 or 1.

Proof. Notice for any $n \geq 1$, $A \in \sigma(X_{n+1}, X_{n+2}, \ldots)$ and hence is independent of all events in $\sigma(X_1, \ldots, X_n)$. Since this is true for any n, A is independent of every event in $\mathcal{G} := \bigcup_{n\geq 1} \sigma(X_1, \ldots, X_n)$. In particular, by Proposition 3.2, we see that A is independent of every event in $\sigma(\bigcup_{n\geq 1} \sigma(X_1, \ldots, X_n))$.

But $\mathcal{T} \subset \sigma(\mathcal{G})$. Thus $A \in \sigma(\mathcal{G})$ as well. So A is independent of itself, in particular,

$$\mathbb{P}(A \cap A) = \mathbb{P}(A)\mathbb{P}(A) = \mathbb{P}(A) \implies \mathbb{P}(A)(1 - \mathbb{P}(A)) = 0 \implies \mathbb{P}(A) \in \{0, 1\}.$$

which completes the proof. □

Example 4.18. If $\{X_n\}_{n\geq 1}$ are not independent, then there is no 0-1 law. For example let $X \sim$ Bernoulli $(1/2)$. Then if with prob $1/2$, $X_n = X$ for all n and with prob. $1/2$ $X_n = 1 - X$, then the event $\{X_n = 1$ for infinitely many $n\}$ has prob $1/2$ (X_n is either all 1 with prob. $1/2$ and all 0 with prob. $1/2$.)

Example 4.19. Suppose $\{X_n\}$ is a collection of independent random variables. Let $S_n = \sum_{k=1}^n X_k$. Then $\{\limsup S_n/n \geq x\}$ for any $x \in \mathbb{R}$ is a tail event hence by Kolmogorov's 0-1 law

$$\mathbb{P}(\limsup S_n/n \geq x) \in \{0, 1\}.$$

Let $x_+ = \sup\{y : \mathbb{P}(\limsup S_n/n \geq y) = 1\}$. By definition, for $y > x_+$, $\mathbb{P}(\limsup S_n/n \geq y) = 0$. Thus $\mathbb{P}(\limsup S_n/n = x_+) = 1$, or in other words, $\limsup S_n/n$ is a constant almost surely.

5 L^p spaces

We cover a bit of the theory of L^p-spaces here. For $p \geq 1$, define the L^p-norm of X as

$$\|X\|_p = (\mathbb{E}(|X|^p))^{\frac{1}{p}}. \tag{5.1}$$

As we will see later, this is indeed a *norm* in the usual sense for $p \geq 1$. We denote the set of all random variables X defined on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ with finite L^p-norm as the L^p**-space**. The notation is $X \in L^p(\Omega, \mathcal{F}, \mathbb{P})$. For $p = \infty$, we set (this is the L^∞-norm)

$$\|X\|_\infty = \inf\{\lambda : |X| < \lambda, \text{ almost surely}\}.$$

Lemma 5.1. *For $1 \leq p \leq q$, if $X \in L^q$ then $X \in L^p$. In this sense, the L^p spaces are nested.*

Proof. This is an application of Jensen for the convex function $x \mapsto x^{\frac{q}{p}}$. □

We state a basic theorem of functional analysis which we will not prove here.

Proposition 5.2. *For $p \geq 1$, and a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, L^p-space is a Banach space. That is,*

- *$\|X\|_p = 0$ implies $X = 0$ almost surely.*
- *$\|aX\|_p = |a|\|aX\|_p$ for any $a \in \mathbb{R}$.*
- *$\|X + Y\| \leq \|X\| + \|Y\|$.*
- *The space is complete. This means that if $\{X_n\}$ is a Cauchy sequence (i.e. for all $\varepsilon > 0$, $\|X_n - X_m\| < \varepsilon$ for all large enough m, n) then there exists an $X \in L^p$ so that $\|X_n - X\|_p$ converges to 0 as $n \to \infty$.*

Proof. Look at Section 9 of http://problab.ca/louigi/courses/2019/math587/587notes.pdf. □

Definition 5.1 (L^p-convergence)**.** *For any $p > 0$, we say that a sequence X_n converges in L^p to X if $\mathbb{E}(|X_n - X|^p) \to 0$ as $n \to \infty$.*

For $p \geq 1$, X_n converges to X in L^p is equivalent to saying that $\|X_n - X\|_p \to 0$. The notation is

$$X_n \xrightarrow[L^p]{n\to\infty} X.$$

Proposition 5.3. *If $X_n \to X$ in L^p for some $p > 0$ then $X_n \to X$ in probability.*

Proof. Fix $\varepsilon > 0$. Then

$$\mathbb{P}(|X_n - X| > \varepsilon) = \mathbb{P}(|X_n - X| > \varepsilon) \leq \frac{\mathbb{E}(|X_n - X|^p)}{\varepsilon^p} \to 0$$

by Markov's inequality. □

Convergence in L^p does not guarantee almost sure convergence. For example, let $\xi_n \sim$ i.i.d. Bernoulli$(1/n)$. We showed in (4.1) that ξ_n does not converge to 0 almost surely but $\xi_n \to 0$ in L^p (easy exercise). On the other hand, convergence in probability does not guarantee L^p convergence.

Exercise 5.4. Let $\xi_n/n \sim$ *Bernoulli*$(1/n)$. Then show that $\mathbb{E}(\xi_n^p) \not\to 0$ for any $p \geq 1$ but $\xi_n \to 0$ in probability.

We finish with two very useful inequalities. The proof involves some tricky version of Jensen. See Theorem 9.3 of Addario–Berry's notes for a proof. [16].

[16] http://problab.ca/louigi/courses/2019/math587/587notes.pdf

Proposition 5.5. *Suppose $1 \le p, q \le \infty$ (note the equality to ∞) with $\frac{1}{p} + \frac{1}{q} = 1$. Then for any two random variables X, Y defined on a common probability space,*

$$\|XY\|_1 \le \|X\|_p \|Y\|_q \tag{5.2}$$

This is called ***Hölder's inequality*** *For $p = q = 2$, this inequality is called* ***Cauchy–Schwarz inequality.***

$$\|XY\|_1 \le \|X\|_2 \|Y\|_2 \text{ or equivalently } \mathbb{E}(|XY|) \le \sqrt{\mathbb{E}(X^2)\mathbb{E}(Y^2)} \tag{5.3}$$

The right hand side is used more commonly.

5.1 Geometric structure of L^2

We now focus on the special case of L^2 spaces. As it turns out, this is a Hilbert space, so we can talk about *angles*. To that end we define the *inner product*

$$\langle X, Y \rangle = \mathbb{E}(XY), \qquad X, Y \in L^2$$

The right hand side is finite by Cauchy–Schwarz inequality. We can define the angle in the sense that the angle between X and Y is measured as $\theta \in [0, \pi)$

$$\cos(\theta) = \frac{\langle X, Y \rangle}{\|X\|_2 \|Y\|_2}$$

So for example

$$\|X + Y\|_2^2 = \mathbb{E}(X + Y)^2 = \mathbb{E}(X^2) + \mathbb{E}(Y^2) + 2\mathbb{E}(XY) = \|X\|_2^2 + \|Y\|_2^2 + 2\langle X, Y \rangle.$$

This gives the **parallelogram law**

$$\|U + V\|_2^2 + \|U - V\|_2^2 = 2(\|U\|_2^2 + \|V\|_2^2). \tag{5.4}$$

Theorem 5.1 (Projection theorem)**.** *Let $\mathcal{G} \subset \mathcal{F}$ be a sub σ-field of $\mathcal{F}$. For every $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$, there is an (almost surely) unique random variable $Y \in L^2(\Omega, \mathcal{G}, \mathbb{P})$ such that*

$$\|X - Y\|_2 = \inf_{Z \in L^2(\Omega, \mathcal{G}, \mathbb{P})} \|X - Z\| =: \Delta.$$

Furthermore, Y is the minimizer (i.e. $\|X - Y\| = \Delta$) if and only if $\langle X - Y, Z \rangle = 0$ for all $Z \in L^2(\Omega, \mathcal{G}, \mathbb{P})$.

Proof. Let

$$\Delta = \inf_{Z \in L^2(\Omega, \mathcal{G}, \mathbb{P})} \|X - Z\|.$$

Take a sequence of random variables $Y_n \in L^2(\Omega, \mathcal{G}, \mathbb{P})$ such that $\|Y_n - X\| = \Delta + \frac{1}{n}$. Now apply the parallelogram law (5.4) with

$$U + V = X - Y_m \qquad U - V = X - Y_n.$$

which gives

$$\|X - Y_m\|_2^2 + \|X - Y_n\|_2^2 = 2\|X - \frac{Y_n + Y_m}{2}\|_2^2 + \|Y_m - Y_n\|_2^2.$$

Note $\|X - Y_m\|_2^2 \leq (\Delta + \frac{1}{m})^2$, and $\|X - Y_m\|_2^2 \leq (\Delta + \frac{1}{m})^2$, also $2\|X - \frac{Y_n - Y_m}{2}\|_2^2 \geq 2\Delta^2$. Thus

$$\|Y_m - Y_n\|_2^2 \leq (\Delta + \frac{1}{m})^2 + (\Delta + \frac{1}{n})^2 - 2\Delta^2 = 2\Delta(\frac{1}{n} + \frac{1}{m}) + 2(\frac{1}{m^2} + \frac{1}{n^2}).$$

The right hand side converges to zero for $n > m \to \infty$. Thus, Y_m is a Cauchy sequence and it converges to a Y in $L^2(\Omega, \mathcal{G}, \mathbb{P})$ by completeness. This Y is the required Y. Indeed, by the triangle inequality,

$$\Delta \leq \|X - Y\| \leq \|X - Y_n\| + \|Y_n - Y\| \leq \Delta + \frac{1}{n} + \|Y_n - Y\| \to \Delta.$$

where the leftmost inequality is by definition of Δ.

Now suppose Z is another random variable with $\Delta = \|X - Z\|$. Then by the parallelogram law,

$$2\Delta^2 = \|X - Y\|_2^2 + \|X - Z\|_2^2 = \|2X - Y - Z\|_2^2 + \|Z - Y\|_2^2 \geq 2\Delta^2 + \|Z - Y\|_2^2$$

which means $\|Z - Y\|_2^2 = 0$ which means $Z = Y$ almost surely.

Now suppose $Y \in L^2(\Omega, \mathcal{G}, \mathbb{P})$ is such that for any $Z \in L^2(\Omega, \mathcal{G}, \mathbb{P})$, $\langle X - Y, Z \rangle = 0$. Then for any $Z \in L^2(\Omega, \mathcal{G}, \mathbb{P})$

$$\|X - Z\|_2^2 = \|X - Y\|_2^2 + \|Z - Y\|^2 + 2\langle X - Y, Z - Y \rangle \geq \|X - Y\|^2$$

Note that $\langle X - Y, Z - Y \rangle = 0$ as $Z - Y \in L^2(\mathcal{G})$ and the inner product of $X - Y$ with any element in $L^2(\mathcal{G})$ is assumed to be 0. So taking infimum over Z on the left hand side, we are done.

Now for the converse, take Y to be the minimizer. Then for any $Z \in L^2(\Omega, \mathcal{G}, \mathbb{P})$,

$$\Delta^2 \leq \mathbb{E}(X - Y - tZ)^2 = \mathbb{E}((X - Y)^2) + t^2\mathbb{E}(Z^2) - 2t\mathbb{E}((X - Y)Z) = \Delta^2 + t^2\mathbb{E}(Z^2) - 2t\mathbb{E}((X - Y)Z)$$

Thus

$$t^2\mathbb{E}(Z^2) - 2t\mathbb{E}((X - Y)Z) \geq 0$$

This cannot hold for small t if $\mathbb{E}((X - Y)Z) \neq 0$, which concludes the proof. □

6 Conditional Expectation

The goal of this section is to make sense of the notion of conditioning by a σ-algebra. In particular, we want to define

$$\mathbb{E}(X|\mathcal{G})$$

where X is a random variable, and $\mathcal{G}$ is a $\mathcal{G}$-algebra. Usually, $\mathcal{G}$ will be taken to be $\sigma(Y)$ for another random variable Y, and then $\mathbb{E}(X|\sigma(Y))$ is the quantity we need to look at if we want to understand *Expectation of X conditioned on Y*. First of all, observe that once conditional expectation is defined, conditional distribution will fall out as the corollary of that defintion as we can write

$$\mathbb{P}(X \in A|\sigma(Y)) = \mathbb{E}(1_{X\in A}|\sigma(Y)).$$

Let us start with something simple. We assume that we are on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let us first recall what it means to condition on an event E where $\mathbb{P}(E) > 0$. For any measurable event E, with $\mathbb{P}(E) > 0$ we can define the conditional expectation of a random variable as X conditioned on E as

$$\mathbb{E}(X|E) = \frac{\mathbb{E}(X1_E)}{\mathbb{P}(E)}. \tag{6.1}$$

Note that $\mathbb{E}(X|A)$ is just a real number.

Now let us see how far this definition takes us through a simple example. Let $a \in [0, 1/2]$. Suppose the joint distribution of Y_1, Y_2 is given by the following table.

$Y_1 \downarrow Y_2 \rightarrow$	0	1	Marginal(Y_1)
0	a	$1/2 - a$	$1/2$
1	$1/2 - a$	a	$1/2$
Marginal (Y_2)	$1/2$	$1/2$	1

Here

$$\mathbb{E}(1_{Y_1=0}|Y_2 = 1) = \mathbb{P}(Y_1 = 0|Y_2 = 1) = \frac{1/2 - a}{1/2} = 1 - 2a,$$
$$\mathbb{E}(1_{Y_1=0}|Y_2 = 0) = \mathbb{P}(Y_1 = 0|Y_2 = 0) = \frac{a}{1/2} = 2a.$$

and

$$\mathbb{P}(Y_1 = 1|Y_2 = 1) = \frac{a}{1/2} = 2a, \qquad \mathbb{P}(Y_1 = 1|Y_2 = 0) = \frac{1/2 - a}{1/2} = 1 - 2a.$$

It is reasonable to think of the 'conditional distribution of Y_1 conditioned on the random variable Y_2 (as opposed to conditioning on an event $\{Y_2 = 0\}$ or $\{Y_2 = 1\}$) as a *random variable*

$$\mathbb{P}(Y_1 = 0|Y_2) = (1 - 2a)1_{Y_2=1} + 2a1_{Y_2=0}; \qquad \mathbb{P}(Y_1 = 1|Y_2) = 2a1_{Y_2=1} + (1 - 2a)1_{Y_2=0}.$$

Here, the σ-algebra generated by Y_2 is $\{\emptyset, \{Y_2 = 1\}, \{Y_2 = 0\}, \Omega\}$. Thus both the random variables defined in the above display is measurable with respect to Y_2.

Furthermore, in both cases, observe that

$$\mathbb{E}(\mathbb{E}(1_{Y_1=0}|Y_2)) = \mathbb{E}(\mathbb{P}(Y_1 = 0|Y_2)) = (1-2a)\frac{1}{2} + 2a\frac{1}{2} = \frac{1}{2} = \mathbb{P}(Y_1 = 0) = \mathbb{E}(1_{Y_1=0});$$
$$\mathbb{E}(\mathbb{E}(1_{Y_1=1}|Y_2)) = \mathbb{E}(\mathbb{P}(Y_1 = 1|Y_2)) = 2a\frac{1}{2} + (1-2a)\frac{1}{2} = \mathbb{P}(Y_1 = 1) = \mathbb{E}(1_{Y_1=1}).$$

Not only this, more is true:

$$\mathbb{E}(\mathbb{E}(1_{Y_1=0}|Y_2)1_{Y_2=1}) = \mathbb{E}(1_{Y_1=0}1_{Y_2=1}) = \frac{(1-2a)}{2}$$
$$\mathbb{E}(\mathbb{E}(1_{Y_1=0}|Y_2)1_{Y_2=0}) = \mathbb{E}(1_{Y_1=0}1_{Y_2=0}) = \frac{2a}{2} = a.$$

Let us try to generalize this idea to random variables taking more values than just $\{0, 1\}$. If you look closely at the above examples, it is clear that conditioning on a random variable X supported on $\{0, 1\}$ is the same as conditioning on the events $\{X = 0\}, \{X = 1\}$, which are the only non-trivial events present in $\sigma(X)$ and are complements of each other. Thus for a σ-algebra $\mathcal{F} := \{\emptyset, A, A^c, \Omega\}$, we can define

$$Y := \mathbb{E}(X|\mathcal{F}) := \mathbb{E}(X|A)1_A + \mathbb{E}(X|A^c)1_{A^c}$$

This random variable satisfies the following properties (can be easily checked just like the example above)

- Y is $\{\mathcal{F}\}$-measurable.
- $\mathbb{E}(Y1_A) = \mathbb{E}(X1_A)$ for all $A \in \mathcal{F}$.
- Taking $A = \Omega$ in the above item, we get $\mathbb{E}(Y) = \mathbb{E}(X)$.

We enlarge this idea to condition on a σ-algebra generated by finitely many events $\mathcal{F} = \sigma(B_1, B_2, \ldots, B_n)$ with $\sqcup_{j=1}^n B_i = \Omega$ and B_is disjoint. Namely, introduce

$$\mathbb{E}(X|\mathcal{F}) = Y = \sum_{j=1}^{n} \mathbb{E}(X|B_j)1_{B_j}.$$

Notice that

$$\mathbb{E}(X1_{B_j}) = \mathbb{E}(Y1_{B_j}) \tag{6.2}$$

The following theorem generalizes this idea to a general σ-algebra. In particular we want to define a random variable Y as the conditional expectation of X given a σ-algebra $\mathcal{G}$ to satisfy

$$\mathbb{E}(X1_B) = \mathbb{E}(Y1_B) \text{ for all } B \in \mathcal{G}.$$

It turns out that this is a strong enough property which guarantees existence of Y which is almost surely unique. This is the content of the next theorem, and this Y will be defined to be the conditional expectation of X given a σ-algebra $\mathcal{G}$.

Theorem 6.1. *Let $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{G} \subset \mathcal{F}$ be a sub σ-algebra. Then $\exists$ a random variable Y on $(\Omega, \mathcal{F}, \mathbb{P})$ such that*

- *Y is $\mathcal{G}$-measurable.*
- *$Y \in L^1$ (integrable) with*

$$\mathbb{E}(Y 1_A) = \mathbb{E}(X 1_A) \text{ for all } A \in \mathcal{G}$$

Note above how we require the equality for events in $\mathcal{G}$ only. This random variable is almost surely unique, in the sense that if there is another random variable Z with the above two properties then $Z = Y$ almost surely.

Since the conditional expectation is defined only up to almost sure sets, any random variable satisfying the two criterions in Theorem 6.1 is called a **version of** conditional expectation. We write

$$\mathbb{E}(X|\mathcal{F}) = Y \text{ a.s.}$$

Also, conditioning by a random variable is simply conditioning by the sigma algebra generated by X, denoted $\sigma(X)$. So

$$\mathbb{E}(Y|X) = \mathbb{E}(Y|\sigma(X)).$$

Proof. First let us simplify our lives and assume that $X \in L^2(\Omega, \mathcal{F}, \mathbb{P})$. Then the theorem is an application of Theorem 5.1. Indeed, using it, find the (a.s. unique) Y such that

$$\|X - Y\|_2 = \inf_{Z \in L^2(\Omega, \mathcal{G}, \mathbb{P})} \|X - Z\|.$$

We claim that Y satisfies the two assumptions. Firstly Y is $\mathcal{G}$-measurable and in $L^2(\mathcal{G})$ by definition. Now we write

$$X = X - Y + Y$$

Notice that

$$\mathbb{E}(X 1_A) = \mathbb{E}((X - Y) 1_A) + \mathbb{E}(Y 1_A) = \langle X - Y, 1_A \rangle + \langle Y, 1_A \rangle$$

But again by Theorem 5.1, $\langle X - Y, 1_A \rangle = 0$ if $A \in \mathcal{G}$, thereby completing the proof.

For the general case of $X \in L^1$, we use the idea of truncation [17]. We need the following claim.

Claim 6.1. *If $X \geq 0$ a.s. and $Y = \mathbb{E}(X|\mathcal{G})$ is a version of conditional expectation (which we assume to exist), then $\mathbb{E}(X|\mathcal{G}) \geq 0$ almost surely.*

Proof of Claim 6.1. Take $B = \{\omega : Y < 0\}$. Notice that by definition of conditional expectation, $B \in \mathcal{G}$. So

$$0 \leq \mathbb{E}(X 1_B) = \mathbb{E}(Y 1_B)$$

[17] this is a general idea, which will be used elsewhere as well

where the first inequality comes from Lemma 2.12 of properties of Expectation. But note that $Y1_B \leq 0$ by definition, thus $Y1_B = 0$ almost surely. This means $\mathbb{P}(B) = \mathbb{P}(Y < 0) = 0$, which is exactly the claim. □

Now assume $X \geq 0$ but we assume $X \in L^1$ only. Now we look at $X_n = X \wedge n = \min\{X, n\}$ for some $n \in \mathbb{N}$. Notice that X_n is bounded, and in particular in L^2. Thus we can find a Y_n which is a version of $\mathbb{E}(X_n|\mathcal{G})$. Now we claim Y_n will converge to a version of the conditional expectation of X given $\mathcal{G}$. Notice that for any $A \in \mathcal{G}$, $X_n 1_A \uparrow X1_A$ and by Claim 6.1, Y_n is also non-decreasing almost surely as

$$Y_{n+1} - Y_n = \mathbb{E}(X_{n+1} - X_n|\mathcal{G}) \geq 0 \text{ a.s.}$$

Thus $Y_n \uparrow Y$ almost surely and by monotone convergence theorem,

$$\mathbb{E}(X_n 1_A) \uparrow \mathbb{E}(X1_A) \text{ and } \mathbb{E}(Y_n 1_A) \uparrow \mathbb{E}(Y1_A).$$

Since $\mathbb{E}(X_n 1_A) = \mathbb{E}(Y_n 1_A)$ for all $A \in \mathcal{G}$, $\mathbb{E}(X1_A) = \mathbb{E}(Y1_A)$ for all $A \in \mathcal{G}$. Since X is integrable, we also have Y is integrable simply by plugging in $A = \Omega$ in this equation.

For general X, not necessarily non-negative, apply the same argument separately to X^+ and X^- and then obtain versions of conditional expectation Y^+ and Y^-. By linearity of expectation, we conclude that $Y := Y^+ - Y^-$ is a version of the conditional expectation.

For uniqueness, assume Z is another such random variable. Let $A = \{Z < Y\}$ which is in $\mathcal{G}$ since both Z, Y are $\mathcal{G}$-measurable. Then

$$\mathbb{E}(Y1_A) = \mathbb{E}(X1_A) = \mathbb{E}(Z1_A) \implies \mathbb{E}((Y - Z)1_A) = 0$$

which means that $Y \leq Z = 0$ almost surely by definiton of A. By symmetry, $Z \leq Y$ a.s. as well and thus $Y = Z$ a.s. □

Example 6.2. Suppose $X \sim$ Bernoulli (Θ) where $\Theta = 1/2$ with prob. $1/4$ and $\Theta = 1/3$ with prob. $3/4$. Let us calculate $\mathbb{E}(X|\Theta)$. Note $\sigma(\Theta) = \sigma(\{\Theta = 1/2\}, \{\Theta = 1/3\})$. Thus on $\{\Theta = 1/2\}$, $\mathbb{E}(X|\Theta) = 1/2$ and on $\{\Theta = 1/3\}$, $\mathbb{E}(X|\Theta) = 1/3$. In other words, $\mathbb{E}(X|\Theta) = \Theta$ almost surely.

Proposition 6.3 (Properties of conditional expectation)**.** *Suppose X is defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{G} \subset \mathcal{F}$. Then,*

a. Suppose $X \geq 0$ a.s. and $X \in L^1$ and $Y = \mathbb{E}(X|\mathcal{G})$ is a version of conditional expectation, then $\mathbb{E}(X|\mathcal{G}) \geq 0$ almost surely.

b. If $a, b \in \mathbb{R}$, then $\mathbb{E}(aX + bY|\mathcal{G}) = a\mathbb{E}(X|\mathcal{G}) + b\mathbb{E}(Y|\mathcal{G})$ a.s.

c. $X \geq Y$ a.s. implies $E(X|\mathcal{G}) \geq \mathbb{E}(Y|\mathcal{G})$.

d. (Conditional Jensen) For any convex function φ with $\mathbb{E}(|\varphi(X)|) < \infty$ or if φ is non-negative, $\mathbb{E}(\varphi(X)|\mathcal{G}) \geq \varphi(\mathbb{E}(X|\mathcal{G}))$. In particular, for $\varphi(x) = |x|$, $\mathbb{E}(|X||\mathcal{G}) \geq |\mathbb{E}(X|\mathcal{G})|$ a.s.

e. $\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(X)$.

f. If X is $\mathcal{G}$-measurable, then $\mathbb{E}(X|\mathcal{G}) = X$ a.s.

g. If X is independent of $\mathcal{G}$, then $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}(X)$ a.s.

Proof. a. Since Claim 6.1 was only used for bounded random variables in the previous proof, this item follows.

b. This follows immediately from the linearity of usual expectation. (Exercise)

c. This follows by applying *a.* with $X - Y$ and using linearity of conditional expectation.

d. This follows by repeating the proof of unconditional Jensen (Proposition 2.4). (Exercise: figure out the details)

e. Use the definition of conditional expectation for the set $A = \Omega$.

f. Follows immediately from the definition (Exercise: convince yourself.)

g. Take $A \in \mathcal{G}$. By independence,

$$\mathbb{E}(X1_A) = \mathbb{E}(X)\mathbb{E}(1_A) = \mathbb{E}(\mathbb{E}(X)1_A).$$

Thus $\mathbb{E}(X)$ is a version of conditional expectation of X given $\mathcal{G}$.

□

Proposition 6.4. *Suppose $\{X_n\}_{n\geq 1}$ is defined on $(\Omega, \mathcal{F}, \mathbb{P})$ and $\mathcal{G} \subset \mathcal{F}$. Then,*

a. (Conditional MCT) If $X_n \geq 0$ and $X_n \uparrow X$ a.s. then

$$\mathbb{E}(X_n|\mathcal{G}) \uparrow \mathbb{E}(X|\mathcal{G})$$

b. (Conditional Fatou's lemma) If $X_n \geq 0$ then

$$\liminf_n \mathbb{E}(X_n|\mathcal{G}) \geq \mathbb{E}(\liminf X_n|\mathcal{G}).$$

c. (Conditional DCT) If $|X_n| \leq Y$ for some random variable Y almost surely for all $n \geq 1$, with $\mathbb{E}(Y) < \infty$ and $X_n \to X$ almost surely, then

$$\lim_n \mathbb{E}(X_n|\mathcal{G}) \to \mathbb{E}(X|\mathcal{G}).$$

Proof. These proofs follows more or less in the same way as in the unconditional case, the additional ingredient needed are the basic properties of conditional expectation outlined in Proposition 6.3. We leave this as an exercise and move on for now. □

Proposition 6.5 (Tower property: smaller σ-algebra wins)**.** *Let $\mathcal{H} \subset \mathcal{G}$ and $X \in L^1(\Omega, \mathcal{F}, \mathbb{P})$.*

$$\mathbb{E}(\mathbb{E}(X|\mathcal{G})|\mathcal{H}) = \mathbb{E}(\mathbb{E}(X|\mathcal{H})|\mathcal{G}) = \mathbb{E}(X|\mathcal{H}) \text{ almost surely.}$$

Here is a rewriting of the above which might be easier to the eye. Let $H = \mathbb{E}(X|\mathcal{H})$ and $G = \mathbb{E}(X|\mathcal{G})$ then

$$\mathbb{E}(G \mid \mathcal{H}) = \mathbb{E}(H \mid \mathcal{G}) = H \text{ almost surely}$$

Proof. Notice that H is $\mathcal{H}$-measurable, and hence $\mathcal{G}$-measurable. Thus $\mathbb{E}(H|\mathcal{G}) = H$ almost surely. On the other hand, for any $A \in \mathcal{H} \subset \mathcal{G}$,

$$\mathbb{E}(\mathbb{E}(G \mid \mathcal{H})1_A) = \mathbb{E}(G1_A) = \mathbb{E}(X1_A)$$

By uniqueness, $\mathbb{E}(G \mid \mathcal{H}) = \mathbb{E}(X|\mathcal{H})$ almost surely. □

Proposition 6.6. *Suppose $X \in L^1(\mathcal{F})$ and $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra. Then if Y is $\mathcal{G}$-measurable random variable, then*

$$\mathbb{E}(XY|\mathcal{G}) = Y\mathbb{E}(X|\mathcal{G}) \text{ almost surely}$$

(Conditioned on $\mathcal{G}$, Y becomes "non random").

Proof. The proof goes through a standard technique called measure theoretic induction. Namely, we start with $Y = 1_A$. Since Y is $\mathcal{G}$-measurable, we must have $A \in \mathcal{G}$. Thus for any $B \in \mathcal{G}$ by definition, Thus

$$\mathbb{E}(Y\mathbb{E}(X|\mathcal{G})1_B) = \mathbb{E}(\mathbb{E}(X|\mathcal{G})1_{A\cap B}) = \mathbb{E}(X1_{A\cap B}) = \mathbb{E}((X1_B)1_A) = \mathbb{E}(YX1_B)$$

The second equality holds since $A \cap B \in \mathcal{G}$. Also since the above equality holds for all B, we have $Y\mathbb{E}(X|\mathcal{G}) = YX$ almost surely, by the uniqueness of conditional probability. By linearity of conditional expectation, (item b. of Proposition 6.3), we have that the equality holds for any simple function.

Now for $Y \geq 0$, take a sequence of simple functions $Y_n \uparrow Y$ (using, e.g., Lemma 2.4). Now by conditional monotone convergence theorem and the previous step, (item a, Proposition 6.4), we have $\mathbb{E}(XY|\mathcal{G}) = \mathbb{E}(Y\mathbb{E}(X|\mathcal{G}))$.

Finally for any Y, break up as $Y = Y^+ - Y^-$, and use the previous step.

Remark 6.7. This general technique is called **measure theoretic induction.**

Exercise: Fill in the details above. □

Using Proposition 6.6, and the fact that $\mathbb{E}(\mathbb{E}(Z|\mathcal{G})) = \mathbb{E}(Z)$, we have

Lemma 6.8. *If Y is $\mathcal{G}$-measurable,*

$$\mathbb{E}(XY) = \mathbb{E}[\mathbb{E}(XY|\mathcal{G})] = \mathbb{E}(Y\mathbb{E}(X|\mathcal{G})).$$

Measure theoretic induction is a pretty robust technique. For example, we can get the following equivalent definition of conditional expectation.

Lemma 6.9 (Alternate definition of conditional expectation.). *Y is a version of $\mathbb{E}(X|\mathcal{G})$ if and only if*

$$\mathbb{E}(YZ) = \mathbb{E}(XZ)$$

for all Z which is $\mathcal{G}$-measurable and bounded.

Proof Sketch. Of course if this statement is true then just by plugging in $Z = 1_A$ for $A \in \mathcal{G}$ works. For the other direction, use measure theoretic induction. (Exercise). □

6.1 (Absolutely) continuous random variables

Recall that we say a random variable X is continuous if there is a measurable function $f_X : \mathbb{R} \mapsto \mathbb{R}$ (called its density) such that for all $A \in \mathcal{B}(\mathbb{R})$, we have

$$\mu_X(A) = \mathbb{P}(X \in A) = \int_A f_X(t)dt.$$

Recall from Proposition 2.19 that if g is a measurable random variable and X has density f_X, then

$$\mathbb{E}(g(X)) = \int_{\mathbb{R}} g(t)f_X(t)dt.$$

Now we state a very useful theorem (without proof) called Fubini's theorem. It essentially says that we can exchange the integrals under certain conditions. Recall that if μ, ν are probability measures on $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$, then $\mu \otimes \nu$ is a probability measure on $(\Omega, \mathcal{G})$ where $\Omega = \Omega_1 \times \Omega_2$ and $\mathcal{G} = \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$, where

$$\mu \otimes \nu((A, B)) = \mu(A)\nu(B)$$

For example, if $\mu = \nu =$ Lebesgue measure on $[0, 1]$ then $\mu \otimes \nu((a, b), (c, d)) = (b - a)(d - c)$.

Theorem 6.2 (Fubini's theorem). *Let $(\Omega_1, \mathcal{F}_1, \mu)$ and $(\Omega_2, \mathcal{F}_2, \nu)$ be two probability spaces. Let $(\Omega, \mathcal{G}, \mu \otimes \nu)$ be the product space described as above. Then for any $f : \Omega \to \mathbb{R}$, with either $\int_\Omega |f|d(\mu \otimes \nu) < \infty$ or $f \geq 0$,*

$$\int_\Omega f(x, y)d(\mu \otimes \nu)(x, y) = \int_{\Omega_1} \left(\int_{\Omega_2} f(x, y)d\nu(y) \right) d\mu(x)$$
$$\int_{\Omega_2} \left(\int_{\Omega_1} f(x, y)d\mu(y) \right) d\nu(x)$$

We say (X, Y) are jointly continuous, simply if they have a joint density function $f_{X,Y}$ such that for all $A \in \mathcal{B}(\mathbb{R}^2)$

$$\mathbb{P}((X, Y) \in A) = \int_A f_{X,Y}(x, y)dxdy$$

where "$dxdy$" can be short for Lebesgue measure in $\mathbb{R}^2$. It coincides with the "usual" Riemann integral taught in calculus courses. It is easy to see that the marginal density of Y can be written as $f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x,y)dx$, since

$$\mathbb{P}(Y \in A) = \mathbb{P}((X,Y) \in A \times \mathbb{R}) = \int_A \left(\int_{\mathbb{R}} f_{X,Y} dy \right) dx$$

Now we want to find a concrete expression for $\mathbb{E}(h(X)|Y)$ where (X,Y) are jointly continuous.

Proposition 6.10 (Conditional density)**.** *Suppose (X,Y) are jointly continuous. Let us define for every $y \in \mathbb{R}$*

$$f_{X|y}(x) = \begin{cases} \frac{f_{X,Y}(x,y)}{f_Y(y)} \text{ if } f_Y(y) > 0 \\ 0 \text{ otherwise} \end{cases} \tag{6.3}$$

Then

$$\mathbb{E}(h(X)|Y) = \int_{\mathbb{R}} h(x) f_{X|Y}(x) dx \text{ almost surely.}$$

(Note that y changed to the random variable Y in the subscript of f).

Proof. Let $\mathcal{N} := \{y : f_Y(y) = 0\}$. Since

$$\mathbb{P}(Y \in \mathcal{N}) = \int 0 dy = 0$$

Let $A = \{Y \in B\}$ and assume $A \subset \{Y \in \mathcal{N}\}$. By Fubini,

$$\mathbb{E}(h(X)1_A) = \int_{\mathbb{R}} \int_{\mathbb{R}} h(x) f_{X,Y}(x,y) 1_{y \in B} dxdy = \int_{\mathbb{R}} \left(\int_{\mathbb{R}} h(x) \frac{f_{X,Y}(x,y)}{f_Y(y)} dx \right) 1_{y \in B} f_Y(y) dy = \mathbb{E}(\mathbb{E}(h(X)|Y)1_A)$$

Note $\mathbb{E}(h(X)|Y)$ is $\sigma(Y)$-measurable, hence we can write $\mathbb{E}(h(X)|Y)1_A = \phi(Y)1_{Y \in B}$. And hence using Theorem 2.18,

$$\mathbb{E}(\mathbb{E}(h(X)|Y)1_A) = \int \phi(y) 1_{y \in B} f_Y(y) dy.$$

By uniqueness of conditional expectation

$$\phi(Y) = \int_{\mathbb{R}} h(x) \frac{f_{X,Y}(x,Y)}{f_Y(Y)} dx$$

almost surely. Therefore, $f_{X|Y}(x)$ is justifiably the "density" of the random variable X conditioned on Y, sometimes called the conditional density.

Since $\{Y \in \mathcal{N}\}$ is a null-set, we can define a version of conditional expectation $\mathbb{E}(h(X)|Y)(\omega) = 0$ for $\omega \in \{Y \in \mathcal{N}\}$. Then for any Borel B. and $A = \{Y \in B\}$,

$$\mathbb{E}(h(X)|Y)1_A = \mathbb{E}(h(X)|Y)1_{A \cap \{Y \notin \mathcal{N}\}},$$

and get back the above formula. □

7 Martingales

Let $(X_n)_{n\geq 1}$ be a collection of random variables defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We sometimes call this collection a **stochastic process** where we think of n as a 'time step'. A stochastic process is **integrable** if $\mathbb{E}(|X_n|) < \infty$ for all $n \geq 1$. A **filtration** is an non-decreasing sequence of sigma algebras $\mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots \mathcal{F}$. A stochastic process $(X_n)_{n\geq 1}$ generates a natural filtration $\mathcal{F}_n^X := \sigma(X_1, \ldots, X_n)$. In particular, X_n is $\mathcal{F}_k^X$ measurable for all $k \geq n$. Conversely, given a filtration $(\mathcal{F}_n)_{n\geq 1}$, we say a stochastic process $(X_n)_{n\geq 1}$ is $\mathcal{F}_n$ adapted if X_n is $\mathcal{F}_n$ measurable for all $n \geq 1$. A probability space with a filtration is called a **filtered space.**

Definition 7.1. *A stochastic process $(X_n)_{n\geq 1}$ defined on a filtered probability space $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 1}, \mathbb{P})$ is a martingale (resp. submartingale, resp. supermartingale) if it is integrable, $\mathcal{F}_n$-adapted and $\mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n$ (resp. $\geq X_n$, resp. $\leq X_n$) a.s. for all $n \geq 1$.*

Using the tower property, actually one gets that $\mathbb{E}(X_{n+1}|\mathcal{F}_m) = X_m$ (resp. $\geq X_m$, resp. $\leq X_m$) a.s. for all $m \leq n$ if X_n is a martingale (resp. submartingale, resp. supermartingale). (Exercise: prove it.)

In a typical situation, the filtration is taken to be $\mathcal{F}_n^X = \sigma(X_1, \ldots, X_n)$. In this case, we say the Martingale is adapted to the **filtration generated by itself.**

Example 7.2. Suppose $X_1, X_2, \ldots \sim$i.i.d. with $\mathbb{E}(X_1) = 0$. Then $(S_n)_{n\geq 1}$ is a martingale with respect to the filtration generated by the X_ns. Indeed $\mathbb{E}(S_{n+1}|\mathcal{F}_n) = \mathbb{E}(S_n + X_{n+1}|\mathcal{F}_n) = S_n$ a.s. using the fact that S_n is $\mathcal{F}_n$-measurable and X_{n+1} is independent of $\mathcal{F}_n$. In fact show using the same ideas that

Exercise 7.3. $T_n = \sum_{k=1}^n H_k X_k$ is a martingale where H_k is $\mathcal{F}_{k-1}$-measurable.

The martingale T_n has an interpretation in terms of gambling. Suppose Bob the gambler enters the casino which plays the fair game(!) in the sense that in each game, if Bob bets a dollars then he gets back $2a$ dollars with prob. $1/2$ and loses the a dollars with prob. $1/2$, independently of what has happened in the past. Suppose in the nth step, Bob bets an amount which depends on what has happened in the first $n-1$ steps (in other words, Bob tries to come up with a strategy). This setup can be modelled by the martingale T_n. Indeed, in step n, Bob strategises and places a bet of H_n dollars and we assume $H_n = \sigma(X_1, \ldots, X_{n-1})$. Thus his winning in nth step is $H_n X_n$ where $X_n = 1$ or -1, each with prob. $1/2$ and $X_1, \ldots, X_n$ are independent. Thus Bob's total winning after n steps is $\sum_{k=1}^n H_k X_k$.

Exercise 7.4. Suppose $\mathbb{E}(|Y|) < \infty$ and $\mathcal{F}_n$ is a filtration. Then show that $X_n = \mathbb{E}(Y|\mathcal{F}_n)$ is a Martingale. (Use the tower property)

Going back to Exercise 7.3, one can ask that if Bob starts with x dollars, and stops when he either wins a Million dollars or goes broke, what is his expected winning when he stops. Note that he stops at a random time which depends on potentially infinitely many random variables. An important concept we need to introduce in order to formalize this idea is the notion of a **stopping time**.

Definition 7.5. *Let $(\Omega, \mathcal{F}, (\mathcal{F}_n)_{n\geq 1}, \mathbb{P})$ be a filtered probability space. A random variable*

$$T : \Omega \mapsto \{1, 2, \dots, \} \cup \{\infty\}$$

is a stopping time if for all $n \geq 1$, $\{T \leq n\} \in \mathcal{F}_n$.

For a stochastic process $(X_n)_{n\geq 1}$, we say T is a stopping time if it is a stopping time for the filtration generate by the process.

Note that we allow $T = \infty$, so this random variable is defined on the so-called 'extended real line'. This causes some minor technicalities with the definitions of measurability, but for the sake of brevity, we ignore them for now.

Example 7.6. Suppose A is a Borel set and let $T = \inf\{k \geq 1, X_k \in A\}$. Then T is a stopping time with respect to the filtration generated by $(X_n)_{n\geq 1}$. Indeed, $\{T \leq n\} = \cup_{i=1}^{n}\{X_i \in A\} \in \mathcal{F}_n$ since $\{X_i \in A\} \in \mathcal{F}_i \subseteq \mathcal{F}_n$.

On the other hand, convince yourself that $T = \sup\{k : X_k \in A\}$ is not a stopping time.

Exercise 7.7. Let T be a stopping time. Then $T \wedge n := \min\{T, n\}$ is also a stopping time.

Now we define the notion of a stopped sigma algebra.

Definition 7.8. *Let T be a stopping time. A stopped sigma algebra is defined as*

$$\mathcal{F}_T := \{A \in \mathcal{F} : A \cap \{T \leq n\} \in \mathcal{F}_n \text{ for all } n \geq 1\}.$$

For all ω such that $T(\omega) < \infty$, define $X_T(\omega) = X_{T(\omega)}(\omega)$.

Definition 7.9. *Let T be a stopping time. Then a stopped process is defined as $(X_n^T)_{n\geq 1} = (X_{T\wedge n})_{n\geq 1}$.*

Now we note a few properties of stopping times.

Lemma 7.10. *If $S \leq T$ a.s. be stopping times, $\mathcal{F}_S \subseteq \mathcal{F}_T$.*

Proof. Since $S \leq T$ a.s., $\{T \leq n\} \subseteq \{S \leq n\}$ and hence $\{T \leq n\} = \{T \leq n\} \cap \{S \leq n\}$. Thus for any $A \in \mathcal{F}_S$,

$$A \cap \{T \leq n\} \subseteq A \cap \{S \leq n\} \cap \{T \leq n\}$$

since $A \cap \{S \leq n\} \in \mathcal{F}_n$ and $\{T \leq n\} \in \mathcal{F}_n$, $A \cap \{T \leq n\} \in \mathcal{F}_n$. Thus $A \in \mathcal{F}_T$ and we are done. □

Lemma 7.11. *$\{T = k\} \in \mathcal{F}_k$ if T is a stopping time.*

Proof. Note $\{T = k\} = \{T \leq k\} \cap (\{T \leq k-1\})^c$. Since the first event on the right hand side is in $\mathcal{F}_k$ and the second is in $\mathcal{F}_{k-1} \subseteq \mathcal{F}_k$, we are done. □

Lemma 7.12. *Suppose $(X_n)_{n\geq 1}$ be an $\mathcal{F}_n$-adapted process and T is a stopping time. Then $X_T 1_{T<\infty}$ is $\mathcal{F}_T$ measurable.*

Proof. We need to show that for any A Borel,

$$\{X_T 1_{T<\infty} \in A\} \in \mathcal{F}_T$$

or equivalently, $\{X_T 1_{T<\infty} \in A, \{T \leq n\}\} \in \mathcal{F}_n$. Note that this event is equal to

$$\cup_{i=1}^n \{X_T 1_{T<\infty} \in A, \{T = i\}\} = \cup_{i=1}^n \{X_i \in A, \{T = i\}\}$$

since the event inside the union is in $\mathcal{F}_i \subseteq \mathcal{F}_n$ for all $i \leq n$, we are done. □

Lemma 7.13. *Suppose $(X_n)_{n\geq 1}$ is $(\mathcal{F}_n)_{n\geq 1}$-adapted and T is a stopping time. Then $(X_{n\wedge T})_{n\geq 1}$ is also $(\mathcal{F}_n)_{n\geq 1}$-adapted. Furthermore, If $\mathbb{E}(|X_n|) < \infty$ then $\mathbb{E}(|X_{n\wedge T}|) < \infty$.*

Proof. Using Exercise 7.7, $T \wedge n$ is a stopping time which is a.s. finite. Thus by Lemma 7.12 $X_{T\wedge n}$ is $\mathcal{F}_{T\wedge n}$-measurable. Thus for $k \leq n$,

$$\{T \wedge n \leq k\} = \{T \leq k\}$$

and if $k > n$, $\mathbb{P}(\{T \wedge n \leq k\}) = 1$. Thus for any Borel A,

$$\{X_{T\wedge n} \in A\} = \cup_{k=1}^{n-1} \{X_k \in A, T = k\} \cup \{X_n \in A, T \geq n\}$$

Note $X_k \in A \in \mathcal{F}_k$, $T = k \in \mathcal{F}_k$, and $T \geq n \in \mathcal{F}_n$. Using these facts we can conclude that $X_{T\wedge n} \in A \in \mathcal{F}_n$ as desired.

For integrability, note that $|X_{T\wedge n}| \leq |X_1| + \ldots + |X_n|$. From this integrability follows. □

7.1 Optional stopping theorem

Note that if X_n is a martingale, then $\mathbb{E}(X_n) = \mathbb{E}(X_{n-1})$, taking expectation on both sides of $\mathbb{E}(X_n|\mathcal{F}_{n-1}) = X_{n-1}$. This immediately yields that $\mathbb{E}(X_n) = \mathbb{E}(X_0)$. Optional stopping theorem will tell us the conditions under which $\mathbb{E}(X_T) = \mathbb{E}(X_0)$ if T is a stopping time.

Example 7.14. We give a simple example to illustrate that $\mathbb{E}(X_T) = \mathbb{E}(X_0)$ is not true in general. Take $X_0 = 0$ and $X_1, X_2, \ldots$ i.i.d. with $X_1 = \pm 1$ with prob. $1/2$ each. Let $T = \min\{k \geq 1, S_k = 10\}$ where $S_k = X_1 + X_2 + \ldots + X_k$. Then it can be shown that $T < \infty$ a.s. (this follows from the fact that a random walk on $\mathbb{Z}$ hits every point a.s.) and $\mathbb{E}(S_T) = 10 \neq 0$. The issue here is that $\mathbb{E}(T) = \infty$.

In the next theorem, we list several conditions under which we can conclude $\mathbb{E}(X_T) = \mathbb{E}(X_0)$ for a stopping time T.

Theorem 7.1. *Let $(X_n)_{n\geq 0}$ be a martingale defined on a filtered probability space $(\Omega, (\mathcal{F}_n)_{n\geq 1}, \mathcal{F}, \mathbb{P})$. Let T be a $\mathcal{F}_n$-stopping time.*

(1) The process $(X_n^T)_{n\geq 1} = (X_{n\wedge T})_{n\geq 1}$ is an $(\mathcal{F}_n)$-adapted martingale. (In particular, $\mathbb{E}(X_{n\wedge T}) = \mathbb{E}(X_0)$ for all $n \geq 1$.)

(2) If $S \leq T$ a.s. and T is bounded a.s (i.e. $\exists M > 0$ such that $T < M$ a.s.) then $\mathbb{E}(X_T|\mathcal{F}_S) = X_S$ a.s. In particular, $\mathbb{E}(X_T) = \mathbb{E}(X_S)$ (take $S = 0$ to get $\mathbb{E}(X_T) = \mathbb{E}(X_0)$).

(3) If $\mathbb{E}(T) < \infty$ and $\exists C > 0$ such that

$$\mathbb{E}(|X_{n+1} - X_n||\mathcal{F}_n)1_{T>n} \leq C \text{ a.s.}$$

Then $\mathbb{E}(X_T) = \mathbb{E}(X_0)$.

(4) If $T < \infty$ a.s. and $\exists C < \infty$ such that $|X_{n\wedge T}| < C$ a.s. for all $n \geq 1$, then $\mathbb{E}(X_T) = \mathbb{E}(X_0)$.

If X_n is a sub(resp. super) martingale, then (1) is true by replacing martingale by sub (resp. super) martingale, (2), (3), (4) is true by replacing $=$ by $\geq$ (resp. $\leq$)

Proof. We prove each item separately.

Proof of (1) Note that

$$\mathbb{E}(X_{n\wedge T}|\mathcal{F}_{n-1}) = \mathbb{E}(\sum_{k\geq 1} X_{n\wedge T}1_{T=k}|\mathcal{F}_{n-1}) = \mathbb{E}(\sum_{k=1}^{n-1} X_k 1_{T=k}|\mathcal{F}_{n-1}) + \mathbb{E}(X_n 1_{T\geq n}|\mathcal{F}_{n-1})$$

Now observe that $\{T \geq n\} = \{T \leq n-1\}^c \in \mathcal{F}_{n-1}$ and $X_k 1_{T=k}$ is $\mathcal{F}_k$ measurable. Thus

$$\begin{aligned}&\mathbb{E}(\sum_{k=1}^{n-1} X_k 1_{T=k}|\mathcal{F}_{n-1}) + \mathbb{E}(X_n 1_{T\geq n}|\mathcal{F}_{n-1}) \\ &\qquad = \sum_{k=1}^{n-1} X_k 1_{T=k} + 1_{T\geq n}\mathbb{E}(X_n|\mathcal{F}_{n-1}) = \sum_{k=1}^{n-1} X_k 1_{T=k} + 1_{T\geq n}X_{n-1} = X_{n-1\wedge T},\end{aligned}$$

as desired.

Proof of (2) Assume $S \leq T < M$ a.s. First we breakup X_T as follows

$$\begin{aligned} X_T &= X_S + (X_T - X_S) \\ &= X_S + \sum_{k=0}^{M-1} (X_T - X_S) 1_{S=k<T} \\ &= X_S + \sum_{k=0}^{M-1} (X_{k+1} - X_k) 1_{S \leq k < T} \end{aligned}$$

Now take $A \in \mathcal{F}_S$. Note that it is enough to show that $\{S \leq k < T\} \cap A \in \mathcal{F}_k$ for all k, as then by the definition of conditional expectation

$$\mathbb{E}(X_S)1_A + \mathbb{E}(\sum_{k=0}^{M-1} (X_{k+1} - X_k) 1_{S \leq k < T} 1_A) = \mathbb{E}(X_S)1_A + \sum_{k=0}^{M-1} \mathbb{E}((X_{k+1} - X_k)|\mathcal{F}_k) 1_{S \leq k < T} 1_A) = \mathbb{E}(X_S)1_A. \tag{7.1}$$

and consequently $\mathbb{E}(X_T 1_A) = \mathbb{E}(X_S 1_A)$ for all $A \in \mathcal{F}_S$ which implies $\mathbb{E}(X_T|\mathcal{F}_S) = X_S$ a.s. Now let us prove that $\{S \leq k < T\} \cap A \in \mathcal{F}_k$. Note $\{S \leq k\} \cap A \in \mathcal{F}_k$ as $A \in \mathcal{F}_S$ and $\{T > k\} \in \mathcal{F}_{k-1} \subseteq \mathcal{F}_k$. This completes the proof.

Proof of (3). Note that by (1), $\mathbb{E}(X_{n \wedge T}) = \mathbb{E}(X_0)$ for every $n \geq 1$, and $X_{n \wedge T} \to X_T$ a.s. as $T < \infty$ a.s. The strategy is to employ Dominated convergence theorem. Note that

$$X_{n \wedge T} = X_0 + \sum_{k=0}^{n-1} (X_{k+1} - X_k) 1_{T > k}$$

Thus

$$|X_{n \wedge T}| \leq |X_0| + \sum_{k=0}^{\infty} |(X_{k+1} - X_k)| 1_{T>k} \text{ a.s.}$$

Now note $\{T > k\} \in \mathcal{F}_{k-1}$. Thus

$$\mathbb{E}(|(X_{k+1} - X_k)| 1_{T>k} | \mathcal{F}_k) = 1_{T>k} \mathbb{E}(|(X_{k+1} - X_k)||\mathcal{F}_k) \leq C 1_{T>k}.$$

by the hypothesis. Plugging this back in, and using MCT (see, for example, Lemma 2.27),

$$|X_0| + \sum_{k=0}^{\infty} |(X_{k+1} - X_k)| 1_{T>k} \leq \mathbb{E}(X_0) + \sum_{k \geq 1} C \mathbb{E}(1_{T>k}) \leq \mathbb{E}(|X_0|) + \mathbb{E}(T) < \infty.$$

by hypothesis. Thus by DCT, $\mathbb{E}(X_0) = \mathbb{E}(X_{n \wedge T}) \to \mathbb{E}(X_T)$, and we are done.

Proof of (4)**.** Thus immediately follows from DCT. Left as an exercise. □

We now present some applications.

Proposition 7.15 (Wald's identity)**.** *Suppose $X_1, X_2, \ldots$ are i.i.d. with $\mathbb{E}(X_1) = \mu < \infty$ and T is a stopping time with respect to the filtration generated by X_is. Suppose $S_n = \sum_{i=1}^n X_i$ and $\mathbb{E}(T) < \infty$. Then*

$$\mathbb{E}(S_T) = \mu \mathbb{E}(T).$$

Proof. Note that $Z_n := S_n - n\mu$ is a martingale. It is an exercise to check that condition (3) of Theorem 7.1 is satisfied by Z_n. Hence using the Optional stopping theorem, $\mathbb{E}(Z_T) = \mathbb{E}(Z_0)$ which means $\mathbb{E}(S_T) = \mu \mathbb{E}(T)$ □

We now get back to the gambler's example from Exercise 7.3. Assume that the Gambler starts with x dollars and let us assume that the gambler bets only 1 dollar per bet. Let $T = \inf\{k \geq 1, S_k \in \{0, M\}\}$ where $0 < x < M$. By Example 7.6, we already know that T is a stopping time. Note that $|S_{n \wedge T}| \leq M$. To apply item (4), we need the additional fact that $T < \infty$ a.s. There are many ways to show this, essentially it follows from the fact that simple random walk in $\mathbb{Z}$ is recurrent, i.e., it visits every vertex with a.s. However here is a way to prove this. Divide $\mathbb{N} = \cup I_i$ where $I_i = \{Mi + 1, \ldots, Mi + M\}$. Note that the probability that all the outcomes of times in I_i is $+1$ is $(\frac{1}{2})^M$. Also if any such event occurs, the gambler hits M and hence $T < \infty$. By Borel Cantelli, one of these events occur a.s. (Exercise).

Applying item (4) of the Optional stopping theorem, we conclude that $\mathbb{E}(S_T) = \mathbb{E}(S_0) = x$. Note that we can find the distribution of S_T from this fact:

$$\mathbb{E}(S_T) = 0 \times \mathbb{P}(S_T = 0) + M \times \mathbb{P}(S_T = M) = x \implies \mathbb{P}(S_T = M) = \frac{x}{M}.$$

7.2 Martingale convergence theorem

We now state the Martingale convergence theorem which illustrates the power of Martingale theory.

Theorem 7.2. *Let $(X_n)_{n \geq 0}$ be a supermartingale that is uniformly bounded in L^1, i.e., $\exists M > 0$ such that $\mathbb{E}(|X_n|) \leq M$ for all $n \geq 1$. Then $X_n \to X_\infty$ a.s. for a random variable X_∞. Furthermore $X_\infty \in L^1$.*

The proof of Theorem 7.2 uses a fact from real analysis which gives a (technical) necessary an sufficient condition for a sequence to converge. Essentially the condition is the following: take any $a < b$. If a real sequence $(x_k)_{k \geq 1}$ converges, it must be the case that the sequence x_k does not oscillate to be above a or below b infinitely many times. This motivates the following definition. Let $N_0 = -1$ and for $k \geq 1$

$$N_{2k-1} = \inf\{m \geq N_{2k-2}, x_m \leq a\} \tag{7.2}$$

$$N_{2k} = \inf\{m \geq N_{2k-1} : x_m \geq b\}. \tag{7.3}$$

The number of **upcrossings** completed by the sequence up to time n is defined as

$$U_n[a, b] = U_n((x_k)_{k \geq 1}, [a, b]) = \sup\{k \geq 0 : N_{2k} \leq n\}$$

Note that $U_n[a,b]$ is non-decreasing in n. We now state a lemma about deterministic sequences.

Lemma 7.16. *Suppose $(x_k)_{k\geq 1}$ is a real sequence. Then $(x_k)_{k\geq 1}$ converges in the extended real line $\mathbb{R}\cup\{\pm\infty\}$ if and only if for all rational $a<b$*

$$\sup_n U_n[a,b] = \lim_n U_n[a,b] < \infty.$$

Proof sketch. If $\limsup x_n > \liminf x_n$, then by choosing rationals a, b such that

$$\limsup x_n > b > a > \liminf x_n$$

we can easily see that $U[a,b]=\infty$ (exercise: fill in details), a contradiction. □

Lemma 7.17. *Suppose X_n is an $\mathcal{F}_n$-adapted submartingale and φ is a non-decreasing convex function with $\mathbb{E}(|\varphi(X_n)|)<\infty$. Then $\varphi(X_n)$ is a submartingale.*

Proof. By conditional Jensen's inequality

$$\mathbb{E}(\varphi(X_{n+1})|\mathcal{F}_n) \geq \varphi(\mathbb{E}(X_{n+1}|\mathcal{F}_n)) \geq \varphi(X_n)$$

since φ is non-decreasing. □

As an immediate corollary we get that

Corollary 7.18. *If X_n is a submartingale, then $(X_n-a)^+$ is a submartingale. If X_n is a supermartingale, $X_n\wedge a$ is a supermartingale.*

Proof. Exercise in application of Lemma 7.17. □

Lemma 7.19. *Suppose X_n is an $\mathcal{F}_n$-adapted sub (resp. super) martingale and H_n is $\mathcal{F}_{n-1}$ measurable. Then $\sum_{k=1}^n H_k(X_k-X_{k-1})$ is a sub (resp. super) martingale.*

Proof. The proof of this is exactly the same as Exercise 7.3. □

Lemma 7.20 (Doob's upcrossing lemma)**.** *Suppose X_n is a submartingale and $a<b$. Then*

$$(b-a)\mathbb{E}(U_n) \leq \mathbb{E}(X_n-a)^+ - \mathbb{E}(X_0-a)^+.$$

Proof. Let $Y_n = a+(X_n-a)^+$. It is straightforward to see that the number of upcrossings of $[a,b]$ is the same as that of X_n. Using Lemma 7.17, Y_n is a submartingale. Recall the definitions of N_{2k-1}, N_{2k} from (7.2). Let $H_m=1$ if $N_{2k-1}<m\leq N_{2k}$ for some $k\geq 1$ and $H_m=0$ otherwise. (That is, drawing the analogue from Exercise 7.3, we only bet when there is an upward trend.) Let $K_m = 1-H_m$. Note that H_m and K_m are $\mathcal{F}_{m-1}$-measurable. Also note that $Z_n=\sum_{k=1}^n(H_k(Y_k-Y_{k-1}))$ and $W_n := \sum_{k=1}^n(K_k(Y_k-Y_{k-1}))$ are submartingales by Lemma 7.19. Furthermore,

$$Z_n \geq (b-a)U_n$$

since we gain at least $(b-a)$ for each upcrossing, and there is a final incomplete crossing which contributes something non-negative. However,

$$Y_n - Y_0 = Z_n + W_n$$

Since W_n is a submartingale $\mathbb{E}(W_n) \geq \mathbb{E}(W_0) = 0$. Thus

$$(b-a)\mathbb{E}(U_n) \leq \mathbb{E}(Z_n) \leq \mathbb{E}(Y_n - Y_0) = \mathbb{E}(X_n - a)^+ - \mathbb{E}(X_0 - a)^+$$

as desired. □

Proof of Theorem 7.2. Fix $a < b$. Note $(X_n - a)^+ \leq X_n^+ + a$. By Doob's upcrossing lemma, $\mathbb{E}(U_n[a,b]) \leq \frac{1}{b-a}(\mathbb{E}(X_n^+) + |a|)$. Taking supremum over n on both sides, and since $\sup_n \mathbb{E}(X_n^+) \leq \sup_n \mathbb{E}(|X_n|) < \infty$, we conclude that

$$\mathbb{E}(U[a,b]) < \infty$$

which means $U[a,b] < \infty$ almost surely. Taking intersection over all rationals a, b, $U[a,b] < \infty$ for all $a, b \in \mathbb{Q}$ almost surely. Thus by Lemma 7.16, we conclude X_n converges almost surely, to some random variable X_∞.

Now let us conclude X_∞ is in L^1. Note by Fatou,

$$\infty > \liminf_n \mathbb{E}(X_n^+) \geq \mathbb{E}(\liminf X_n^+) = \mathbb{E}(X_\infty^+)$$

For X_∞^-, note that

$$\mathbb{E}(X_n^-) = \mathbb{E}(X_n^+) - \mathbb{E}(X_n) \leq \mathbb{E}(X_n^+) - \mathbb{E}(X_0)$$

since X_n is a martingale tells us that $\mathbb{E}(X_n) \geq \mathbb{E}(X_0)$. Using Fatou again, we conclude $\mathbb{E}(X^-) < \infty$ as well. □

7.3 Applications of Martingales

7.3.1 Levy's 0-1 law

We know that $\mathbb{E}(X_n|\mathcal{F}_m) = X_m$ for $m \leq n$. The following corollary shows that this corollary persists even in the limit.

Corollary 7.21. *If X_n is a martingale and $X_n \to X$ in L^1 then $X_n = \mathbb{E}(X|\mathcal{F}_n)$.*

Proof. Take $A \in \mathcal{F}_m$. We know for all $n \geq m$, $\mathbb{E}(X_n|\mathcal{F}_m) = X_m$ a.s. Thus by properties of conditional expectation

$$\mathbb{E}(X_n 1_A) = \mathbb{E}(X_m 1_A)$$

Also,

$$|\mathbb{E}(X_n 1_A) - \mathbb{E}(X 1_A)| \leq \mathbb{E}(|X_n - X| 1_A) \leq \mathbb{E}(|X_n - X|) \to 0,$$

since $X_n \to X$ in L^1. Thus $\mathbb{E}(X_m 1_A) = \mathbb{E}(X 1_A)$. Since A is an arbitrary element of $\mathcal{F}_m$, we conclude using the definition of conditional expectation. □

Now we prove a nice property of the martingale from Exercise 7.4.

Theorem 7.3. *Let $\mathbb{E}(|X|) < \infty$. Suppose $(\mathcal{F}_n)_{n\geq 1}$ be an increasing sequence of σ-algebras and $\mathcal{F}_\infty = \sigma(\cup_{n\geq 1}\mathcal{F}_n)$. Then*

$$\mathbb{E}(X|\mathcal{F}_n) \to \mathbb{E}(X|\mathcal{F}_\infty) \text{ a.s. and in } L^1.$$

Proof. We know from Exercise 7.4 that $Z_n := \mathbb{E}(X|\mathcal{F}_n)$ is a martingale. Also

$$\mathbb{E}(|Z_n|) = \mathbb{E}(|\mathbb{E}(X|\mathcal{F}_n)|) \leq \mathbb{E}(\mathbb{E}(|X||\mathcal{F}_n)) = \mathbb{E}(|X|).$$

Consequently Z_n is uniformly bounded in L^1. Applying the martingale convergence theorem Theorem 7.2, we obtain that $Z_n \to Z$ almost surely and in L^1. By Corollary 7.21, $\mathbb{E}(Z|\mathcal{F}_n) = \mathbb{E}(X|\mathcal{F}_n)$ for all $n \geq 1$. Thus for any $A \in \mathcal{F}_n$, $\mathbb{E}(Z1_A) = \mathbb{E}(X1_A)$ a.s. for all $A \in \cup\mathcal{F}_n$. Thus using the same logic as in Lemma 1.16, we conclude that $\mathbb{E}(Z1_A) = \mathbb{E}(X1_A)$ for all $A \in \mathcal{F}_\infty$ (Exercise: prove it). We conclude using the definition of conditional expectation. □

One powerful consequence is the following theorem

Theorem 7.4. *If $\mathcal{F}_n \uparrow \mathcal{F}_\infty$ and $A \in \mathcal{F}_\infty$ then $\mathbb{E}(1_A|\mathcal{F}_n) \to 1_A$ almost surely.*

Proof. Immediately follows from Theorem 7.3 as 1_A is measurable with respect to $\mathcal{F}_\infty$. □

Note that we immediately recover Kolmogorov 0-1 law from Theorem 7.4. Indeed, if A is tail σ-algebra measurable, the it is independent of $\mathcal{F}_n$ for any n. So $\mathbb{E}(1_A|\mathcal{F}_n) = \mathbb{P}(A)$ for all n. On the other hand, using Levy's 0-1 law, $\mathbb{E}(1_A|\mathcal{F}_n) \to 1_A \in \{0, 1\}$ a.s. Thus $\mathbb{P}(A) \in \{0, 1\}$.

7.3.2 Branching process

Branching processes are used to model the growth of a population. Suppose we start with a single individual of a certain specie who gives rise to Z many offsprings where Z has some pmf given by

$$\mathbb{P}(Z = i) = p_i \text{ for } i \geq 0.$$

Call this offsprings members of generation 1. Next, each offspring of generation 1 gives rise to a certain number of offsprings distributed as Z and these are independent of each other. Let X_n be the number of offsprings in the nth generation for $n \geq 0$ with $X_0 = 1$.

We are interested in the question: does the specie die out? If so, can we compute/estimate its probability?

What happens to $\mathbb{P}(X_n = 0)$ as n increases? Clearly if the population has died out in step n, $X_{n+1} = 0$ is trivially true. Thus $\{X_n = 0\} \subseteq \{X_{n+1} = 0\}$. Thus

$$\{\text{Population eventually dies }\} = \{X_n = 0 \text{ for some } n \geq 1\} = \cup_{n\geq 1}\{X_n = 0\} = \lim_{n\to\infty}\{X_n = 0\}.$$

Said otherwise, $\mathbb{P}(X_n = 0)$ is non-decreasing, therefore must have a limit. Let $d_n = \mathbb{P}(X_n = 0)$ and $d = \lim_{n\to\infty} d_n$. Clearly $d_n \in [0, 1]$ and hence so does d.

Let $\mu \in \mathbb{E}(Z)$. The following observation illustrates the utility of martingales in this setup.

Lemma 7.22. *$(\frac{X_n}{\mu^n})_{n\geq 1}$ is a martingale which is uniformly bounded in L^1.*

Proof. Observe that for every $n \geq 1$,

$$X_n = \sum_{i=1}^{X_{n-1}} Z_{n-1,i}$$

where conditioned on X_n, $(Z_{n-1,i})_{1\leq i\leq X_n}$ are i.i.d. and distributed as Z. Now observe that

$$\mathbb{E}(X_n|\mathcal{F}_{n-1}) = X_{n-1}\mu$$

and hence

$$\mathbb{E}(X_n/\mu^n \mid \mathcal{F}_{n-1}) = X_{n-1}/\mu^{n-1}.$$

□

As a consequence of Lemma 7.22 and theorem 7.2, we conclude

$$\frac{X_n}{\mu^n} \xrightarrow[a.s.]{n\to\infty} X_\infty.$$

Martingale theory unfortunately does not tell us anything about X_∞. However

Proposition 7.23. *If $\mu < 1$, $X_\infty = 0$ almost surely.*

Proof. Since $\mathbb{E}(X_n/\mu^n) = \mathbb{E}(X_0) = 1$, we conclude using Markov's inequality:

$$\mathbb{P}(X_n \geq 1) \leq \mu^n \xrightarrow[n\to\infty]{} 0 \text{ if } \mu < 1.$$

Thus $X_n \to 0$ in probability. Since

$$\{X_n = 0\} \uparrow \cup_n \{X_n = 0\} \subset \{X_\infty = 0\}$$

and since $\mathbb{P}(X_n = 0) \to 1$, we must have $\mathbb{P}(X_\infty = 0) = 1$ as well. □

If $\mu \geq 1$, it is not so straightforward to figure out what X_∞ is. We do not pursue this further in this section.

7.3.3 Discrete harmonic function

Let $G = (V, E)$ be a finite graph and call $\partial \subset V$ the boundary of the graph. Let $g : \partial \to \mathbb{R}$ be a function which is called the *boundary condition*. A function $h : V \to \mathbb{R}$ is called harmonic on (G, ∂) with boundary condition g if $h(v) = g(v)$ for all $v \in \partial$ and

$$h(v) = \frac{1}{\deg(v)} \sum_{u\sim v} h(u)$$

where $u \sim v$ means u is adjacent to v and $\deg(v)$ is the degree of v. Recall degree of a vertex is simply the number of its neighbours.

Let $(X_n)_{n\geq 1}$ be a simple random walk. That is, conditioned $X_i = v$, X_{i+1} is distributed uniformly among the neighbours of v. Let

$$\tau = \inf\{k \geq 0 : X_k \in \partial\}.$$

Observe that τ is a stopping time with respect to the natural filtration generated by the simple random walk.

Lemma 7.24. *$(h(X_{n\wedge\tau}))_{n\geq 1}$ is a martingale.*

Proof. Note $|h(X_{n\wedge\tau})| \leq \max_{v\in V} |h(v)|$, and hence $|h(X_n)|$ is uniformly bounded, and thus is in L^1. Note

$$\mathbb{E}(h(X_{(n+1)\wedge\tau})|\mathcal{F}_n)1_{\tau>n} = \frac{1}{\deg(X_n)} \sum_{v\sim X_n} h(v)1_{\tau>n} = h(X_n)1_{\tau>n} = h(X_{n\wedge\tau})1_{\tau>n}.$$

Indeed, if $\tau > n$, then $X_n \notin \partial$, and hence h is harmonic on X_n. Also,

$$\mathbb{E}(h(X_{(n+1)\wedge\tau})|\mathcal{F}_n)1_{\tau\leq n} = h(X_\tau)1_{\tau\leq n} = h(X_{\tau\wedge n})1_{\tau\leq n}.$$

Overall,

$$\mathbb{E}(h(X_{(n+1)\wedge\tau})|\mathcal{F}_n) = \mathbb{E}(h(X_{(n+1)\wedge\tau})|\mathcal{F}_n)(1_{\tau>n} + 1_{\tau\leq n}) = h(X_{\tau\wedge n})((1_{\tau>n} + 1_{\tau\leq n})) = h(X_{\tau\wedge n}).$$

as desired. □

A harmonic function on a graph with a boundary condition g is called the harmonic extension on g. Finding such a harmonic extension is a linear algebra problem (convince yourself!) Martingale theory allows us to prove, fairly easily, that such a harmonic extension exists and is unique.

Lemma 7.25. *Given a graph G with boundary ∂ and a boundary condition g, there exists a unique harmonic extension given by*

$$h(v) = \mathbb{E}(h(X_\tau))$$

where $(X_n)_{n\geq 0}$ is a simple random walk started at $X_0 = v$.

Proof. Note that by Lemma 7.24, $h(X_{n\wedge\tau})$ is a martingale. It is easy to check that condition (3) of the optional stopping theorem holds, and hence $h(v) = \mathbb{E}(h(X_0)) = \mathbb{E}(X_\tau)$. Now suppose there are two function h and $\tilde{h}$ which are harmonic extensions of g. Then $h - \tilde{h}$ is a harmonic extension of a function which is identically 0 on ∂. It is easy to see that such a harmonic extension must be identically 0. □

Another interesting consequence of martingale theory is the nonexistence of bounded harmoinic functions on the square lattice. The proof of this requires us to assume that simple random walk on the square lattice is recurrent, i.e., it visits every point on the graph.

Proposition 7.26. *There cannot exist a harmonic function h on the square lattice $\mathbb{Z}^2$ which is uniformly bounded and is not identically a constant function. (A function h is uniformly bounded if there exists a $C > 0$ such that $|h(v)| \leq C$ for all $v \in V$.)*

Proof. The same argument as in Lemma 7.24 entails that $(h(X_n))_{n\geq 0}$ is a martingale. Since h is bounded, $(h(X_n))_{n\geq 0}$ is a bounded martingale. Therefore $h(X_n)$ converges almost surely by the martingale convergence theorem. Now suppose there are two vertices u and v such that $h(u) \neq h(v)$. Since $h(X_n)$ visits both u and v infinitely often by recurrence of simple random walk on $\mathbb{Z}^2$, $h(X_n)$ takes two different values infinitely often, which contradicts the conclusion that it almost surely converges. □