Notes on Probability Theory and Statistics*

Joël Terschuur

Contents

*All material presented here can be traced down to the course Probability Foundation for Electrical Engineers by Dr. Krishna Jagannathan,Department of Electrical Engineering,IIT Madras in the Youtube Channel nptelhrd, Econometrics I lectures by Juan Carlos Escanciano at UC3M and other sources in the bibliography.

1 Introduction

In real life we encounter many events which we cannot predict perfectly or for which we have imperfect knowledge. Examples range from tossing of a coin, to the state of the weather in the next three days or the extent to which GDP will decrease after a pandemic. However, while it might seem we are clueless about many of these events, there are patterns we can study. We know that if we toss a coin a thousand times, we should be very close to half heads and half tails. We are able to forecast weather and GDP growth to some extent. In essence, probability theory is a science of randomness which allows us to make some reasonable predictions. It provides a foundation to the patterns we observe with more regularity in real life.

We have been playing games of chance and computing probabilities for centuries. In the 17^{th} century a gambler called Chevallier de Mere asked Laplace and Fermat for help. He was playing two betting games he considered equivalent but was consistently losing with one of them and not with the other. Pascal and Fermat showed why these two games were not the same and by doing this paved the way for computation of probabilities. However, modern probability theory is roughly 100 years old. The main founder of the axiomatic approach to probability which we encounter today is the Russian mathematician Andréi Kolmogórov (1903-1987). He noticed that probability theory is just a special case of measure theory which was developed by French mathematicians Émile Borel (1871-1956) and Henri Léon Lebesgue (1875-1941). While people already knew how to compute many probabilities before Kolmogórov, it was usual to run into puzzling results and paradoxes. With a strong axiomatic foundation, the answer to these apparent puzzles and paradoxes was resolved by unifying all that was known into one single logical framework.

We will first do a short review of set theory and cardinality of sets and then we will start diving into probability theory.

2 Mathematical Review

This section follows closely Kolmogorov and Fomin (1975) and De la Fuente (2000). A set is a collection of objects we call elements. For instance the set of all integers or the set of all even numbers. It started to be developed at the end of the 19th century with the paper Cantor (1874). It is a subject within mathematics in its own right and is crucial in many fields in modern mathematics. Specifically, a basic knowledge of set theory is essential to understand probability theory.

We will denote sets with capital letters like $A, B, ...$ and their elements with lower case letters such as $a, b, ...$ A set with elements $a, b, c, ...$ is often denoted by $\{a, b, c, ...\}$. For instance, the set of all positive integers is {1,2,3,...}. The singleton set containing only one element, for instance the number one, is $\{1\}$. If a is an element of a set A, "a belongs in A", we write $a \in A$. If a does not belong to a set A we write $a \notin A$. For instance, $2 \in \{1, 2, 3\}$ but $4 \notin \{1, 2, 3\}$. If every element of a set A belongs to a set B, we say that A is a subset of

B, denoted $A \subseteq B$ or $B \supseteq A$, the latter meaning "B contains A". Formally,

$$A \subseteq B \iff (x \in A \implies x \in B)$$

So from now on, if we want to show that a set A is a subset of a set B we show that all elements belonging to A also belong to B. For instance $\{1,2\} \subseteq \{1,2,3\}$. We say that two sets A and B are equal, $A = B$, if all elements belonging in A also belong in B and vice versa. That is, every time we want to show two sets are equal we need to show that $A \subseteq B$ and $B \subseteq A$. Also, if $A \subseteq B$ and $A \neq B$, then A is a proper subset of B, $A \subset B$. Sometimes we might not know whether a set contains any elements at all. For instance, the set of roots of a given equation. A set containing no elements is called the empty set and is denoted by $\emptyset$.

A collection or family or class of sets is a set whose elements are sets themselves. For instance if we have sets A, B, and C. A collection of sets containing these is $\mathcal{D} = \{A, B, C\}$. Collections of sets are usually denoted by calligraphic capital letters. It is important to see that now the elements of $\mathcal{D}$ are sets, that is, we write $A \in \mathcal{D}$. Hence, we do not say A is a subset of $\mathcal{D}$, it is an element belonging to it in the same way that an element a belongs in A, $a \in A$. Hence it is important to always have in mind what is the typical element of a set. Given a set A, an important collection of subsets is the collection of all subsets of A. This is called the power set of A, denoted by 2^A. For instance, if $A = \{1,2,3\}$ we have that

$$2^A = \Big\{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}, \{2,3\}, \{1,2,3\}\Big\}$$

Again note that 2^A is a set of sets[1]. From now own suppose there is some universal set X and that there is nothing outside of it. We will work only with subsets of X.There are two operations on sets which will be used often. The union and the intersection. If we have two sets A and B, $A, B \subseteq X$, we define their union, $A \cup B$, as the set

$$A \cup B = \{x \in X : x \in A \text{ or } x \in B\}$$

The right hand side (RHS) above reads as: "all x belonging in X such that x belongs in A or x belongs in B". It means all those x which belong *at least* in either A or B. Hence if x belongs to A, to B, or to both, x belongs to the set $A \cup B$. The intersection $A \cap B$, is the set

$$A \cap B = \{x \in X : x \in A \text{ and } x \in B\}$$

That is, x belongs to the set $A \cap B$ if it belongs to *both* A and B. This operations can be extended to more than two sets. Suppose we have a collection of sets $\{A_i, i \in I\}$ where I is

[1] $\emptyset$ is a subset of all sets. This is because $\emptyset \subseteq A$ means that $x \in \emptyset \implies x \in A$. However, there is nothing in $\emptyset$, so the antecedent in the previous implication ($x \in \emptyset$) is false. Then, the consequence ($x \in A$) is true. If antecedent is false any consequence is true. See `https://math.stackexchange.com/questions/439987/assumed-true-until-proven-false-the-curious-case-of-the-vacuous-truthifinterested`.

some index set, for example the natural numbers. Then

$$\bigcup_{i\in I} A_i = \{x \in X : \exists i \in I \text{ such that } x \in A_i\}$$
$$\bigcap_{i\in I} A_i = \{x \in X : x \in A_i \ \forall i \in I\}$$

Hence, the union above consists in the set of all $x \in X$ which belong to *at least* one of the A_i's while the intersection consists in the set of all $x \in X$ which belong to *all* A_i's.

Let us examine some properties of unions and intersections

Proposition 2.1. *Let A, B and C be subsets of X. Then the following hold*

(i) Commutative law: $A \cup B = B \cup A$ and $A \cap B = B \cap A$.

(ii) Associative law: $(A \cup B) \cup C = A \cup (B \cup C) = A \cup B \cup C$ and $(A \cap B) \cap C = A \cap (B \cap C) = A \cap B \cap C$.

(iii) Distributive law: $(A \cup B) \cap C = (A \cap C) \cup (B \cap C)$ and $(A \cap B) \cup C = (A \cup C) \cap (B \cup C)$.

The proof of Proposition 2.1 follows form the definitions of union and intersection. I encourage you to do them.

Two sets A and B are disjoint if they have no elements in common, that is, if $A \cap B = \emptyset$. Generally, given a family of sets $\mathcal{A} = \{A_i, i \in I\}$, we say that the elements of $\mathcal{A}$ are pairwise disjoint if

$$A_i \cap A_j = \emptyset \ \forall \, i \neq j$$

A partition of X is a class of pairwise disjoint sets in X such that their union is X. Formally, $\mathcal{A} = \{A_i, i \in I\}$ is a partition of X if for all $i \neq j$

$$A_i \cap A_j = \emptyset \text{ and } \bigcup_{i\in I} A_i = X$$

Given two sets A and B, both subsets of X, $A \setminus B$ denotes the set of elements belonging to A and not to B

$$A \setminus B = \{x \in X : x \in A \text{ and } x \notin B\}$$

The complement of a set $A \subset X$, denoted by A^c is the set containing all elements in X which are not in A

$$A^c = \{x \in X : x \notin A\}$$

Note that $A \setminus B = A \cap B^c$. Another important property is the following

Proposition 2.2 (De Morgan's Laws)**.** *Let $\mathcal{A} = \{A_i, i \in I\}$, then*

(i) $\left(\bigcup_{i\in I} A_i\right)^c = \bigcap_{i\in I} A_i^c$, and

(ii) $\left(\bigcap_{i\in I} A_i\right)^c = \bigcup_{i\in I} A_i^c$.

Proof. Note that if we want to prove that two sets are equal we need to show that they are subsets of each other. For $i)$ suppose that $x \in (\cup_{i\in I} A_i)^c$, then $x \notin \cup_{i\in I} A_i$, then there exists no $i \in I$ such that $x \in A_i$, this implies that $x \in A_i^c$ for all $i \in I$, hence $x \in \cap_{i\in I} A_i^c$. We conclude that $(\cup_{i\in I} A_i)^c \subseteq \cap_{i\in I} A_i^c$. For the other direction suppose that $x \in \cap_{i\in I} A_i^c$, then $x \in A_i^c$ for all $i \in I$, then $x \notin \cup_{i\in I} A_i$ and $x \in (\cup_{i\in I} A_i)^c$. Hence, $\cap_{i\in I} A_i^c \subseteq (\cup_{i\in I} A_i)^c$. We conclude that $\cap_{i\in I} A_i^c = (\cup_{i\in I} A_i)^c$. $ii)$ is left as an exercise. □

Exercise 2.1. *Let $A_1, A_2, ...$ be subsets of X. Define $B_1 = A_1$, $B_2 = A_2 \setminus A_1$,..., $B_k = A_k \setminus \cup_{i=1}^{k-1} A_i$,... Show that $\{B_i\}_{i=1}^{\infty}$ are disjoint and that*

$$\bigcup_{i=1}^{\infty} A_i = \bigcup_{i=1}^{\infty} B_i \text{ and } \bigcup_{i=1}^{m} A_i = \bigcup_{i=1}^{m} B_i$$

Mappings between sets can be particularly useful. Because of this we introduce the following concepts

Definition 2.1 (Function)**.** *A function $f : A \to B$ is a rule which associates every element in A with a unique element in B.*

Note that a function needs to map every element in A, that is, we cannot leave out any element in A. However we do not need to map to all elements of B. B is called the co-domain and the set of elements of B to which the function maps to is called the range, denoted R, formally

$$R = \{y \in B : \exists x \in A \text{ such that } f(x) = y\}$$

Definition 2.2 (Injective function)**.** *Every element in R has a unique pre-image in A.*

This means that there are not two distinct elements in the domain which map to the same element in B.

Definition 2.3 (Surjective function)**.** $R = B$.

Definition 2.4 (Bijective function)**.** *A function is called bijective if it is both injective and surjective.*

Definition 2.5. *(i) Sets A and B are said to be equicardinal if there exists a bijection $f : A \to B$, $|A| = |B|$,*

(ii) B has cardinality greater than or equal to A if there exists an injective function $f : A \to B$, $|B| \geq |A|$,

(iii) $|B| > |A|$, if there exists an injective function $f : A \to B$, but A and B are not equicardinal.

Definition 2.6. *(i) A set A is said to be countably infinite if it is equicardinal with $\mathbb{N}$.*

(ii) A set A is countable if it is either finite or countably infinite.

Example 2.1. $\mathbb{Q} \cap [0,1]$ *(Rationals in* $[0,1]$*) is a countable set. This means you can form a bijection with* $\mathbb{N}$*, that is, a list enumerating all elements*

$$\{0, 1, 1/2, 1/3, 2/3, 1/4, 3/4, 1/5, 2/5, 3/5, 4/5, ...\}$$

One can also show that a countable union of sets is countable. Hence,

$$\mathbb{Q} = \bigcup_{i \in \mathbb{Z}} \mathbb{Q} \cap [i, i+1]$$

is countable.

Definition 2.7. *A set* A *is uncountable if it has cardinality strictly larger than that of* $\mathbb{N}$*.*

Examples of uncountable sets are $\mathbb{R}, \mathbb{R} \setminus \mathbb{Q}, 2^{\mathbb{N}}, [0,1]$ or $\{0,1\}^{\infty}$ which is the set of all infinite binary strings. Showing that the latter is uncountable is the first step to show the rest are uncountable. The proof is Cantor's diagonal argument which you can find in many textbooks or just in the Wikipedia. The key from this discussion for us is to have a clear notion of countable (finite and infinite) and uncountable sets.

3 Probability spaces

We start from two *undefined* entities. A *Random Experiment* and an *Outcome*. These are to be understood intuitively from their semantic meaning. However, we will give some examples to clarify these notions. Starting from these two entities we will define all the rest.

Definition 3.1 (Sample space)**.** *the sample space* Ω *is the set of all possible outcomes of a random experiment.*

For example, suppose that your random experiment is to toss a coin once. What is the sample space? It depends on what is of interest to you. You determine what the sample space is. If you are interested in which face shows up, then the set of possible outcomes is $\Omega = \{H, T\}$, where H = Heads and T = Tails. However, you might not be interested in which face shows up but in the number of times the coin flips in the air. In this case, the set of all possible outcomes which are of interest to you is $\Omega = \mathbb{N}$. Another possibility is that what is of interest to you is the velocity at which the coin hits the ground. In this case, the sample sample space is $\mathbb{R}^+$. Note that we have given an example of a finite, countably infinite and uncountable sample space.

An (elementary) outcome is denoted by $\omega \in \Omega$. *This* is the source of randomness. You have no control over what ω realizes. You can think about it as $\omega \in \Omega$ being chosen by some Goddess of Chance. Another way of imagining it is as $\omega \in \Omega$ being one of many alternate realities[2]. Suppose that whenever you throw a dice, six alternative realities are created and you do not get to choose in which one you end up. Every time you run the random experiment, an outcome $\omega \in \Omega$ realizes. Another example of a random experiment is to toss

[2]If you like the show Rick and Morty you can think that parallel universe C-137 is the ω which was chosen for the main characters in the show.

a coin n times. If you are interested in the number of faces that show we have $\Omega = \{H, T\}^n$ (finite sample space). Note that we are thinking of this as *one* random experiment and not as n random experiments. We could also think about a random experiment which consists in tossing a coin infinitely many times. Then, $\Omega = \{H, T\}^\infty$ (uncountable sample space). An example of an elementary outcome in this setting is $\omega = \{H, H, T, H, T, T, T, ...\}$. Another random experiment is to throw a dart to the $[0, 1]$ line. Then, $\Omega = [0, 1]$ (uncountable sample space) and an elementary outcome could be $\omega = 0.333$.

Often we are not interested in whether a particular elementary outcome has occurred. We might be interested in whether a *subset* of the sample space has occurred or not. For instance, if your random experiment is tossing a coin once and your sample space is the number of flips in the air ($\Omega = \mathbb{N}$), you might not be interested in the exact number of flips but on whether it flipped more than five times ($\{6, 7, 8, ...\} \subseteq \mathbb{N}$). Or think about the Spanish economy during the last quarter of 2020 as your random experiment and GDP growth as your sample space ($\Omega = \mathbb{R}$). You might not be interested in the exact growth rate but just in whether GDP growth is positive or negative ($\mathbb{R}^+ \subseteq \mathbb{R}$ or $\mathbb{R}^- \subseteq \mathbb{R}$). These kind of subsets of Ω which are of interest to us will be called events (we will give more rigorous definitions in a bit). An event $A \subseteq \Omega$ is said to occur if $\omega \in A$, that is, if the Goddess of Chance has picked an $\omega \in \Omega$ which belongs to A. Importantly, all events are subsets of Ω but not all subsets of Ω are events. For instance, we might not care about GDP growth being a rational number despite the fact that $\mathbb{Q} \subseteq \mathbb{R}$. Ultimately we will want to assign probabilities to events. But first we need to put everything we just said in a more rigorous mathematical scheme. Our goal now is to build a structure of the subsets of Ω to determine what is of interest and what is not of interest, i.e. what is an event and what is not an event. To do this we are going to impose some rules. Intuitively, since Ω always occurs, we should be interested in Ω. Also, if there is some subset A which interests us, we should be also interested in A^c, i.e. in A not occurring. Also, if there are two events A and B which are of interest, we should be interested in at least one of them taking place or in both of them occurring, i.e. $A \cup B$ and $A \cap B$ should be of interest. These rules motivate some mathematical structures of subsets of Ω and some quantities assigned to these subsets with which we will work with. From now on concepts followed by a *-sign are not essential in a first reading and can be postponed to a second reading if needed (and desired)

Definition 3.2 (Algebra*)**.** *Let Ω be the sample space and $\mathcal{F}_0$ be a collection of subsets of Ω. $\mathcal{F}_0$ is called an algebra if*

(i) $\emptyset \in \mathcal{F}_0$,

(ii) if $A \in \mathcal{F}_0$, *then* $A^c \in \mathcal{F}_0$,

(iii) if $A \in \mathcal{F}_0$ *and* $B \in \mathcal{F}_0$, *then* $A \cup B \in \mathcal{F}_0$.

Definition 3.3 (Pre-measure*)**.** *Suppose $\mathcal{F}_0$ is an algebra of Ω. A pre-measure μ_0 on $\mathcal{F}_0$ is a set function $\mu_0 : \mathcal{F}_0 \to [0, \infty]$ such that*

(i) $\mu_0(\emptyset) = 0$,

(ii) if $A_1, A_2, ...$ is a countable collection of disjoint sets in $\mathcal{F}_0$ such that

$$\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}_0,$$

then

$$\mu_0\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mu_0(A_i).$$

One limiting aspect of the algebra is that it is only closed under[3] finite unions (show that it is closed under finite unions). This limitation and others carries through to pre-measures. This is a problem since we might be interested in a countably infinite union of events. This motivates the definition of a broader structure which will be one of our main tools.

Definition 3.4 (σ-algebra)**.** *A collection $\mathcal{F}$ of subsets of Ω is called a σ-algebra if*

(i) $\emptyset \in \mathcal{F}$,

(ii) if $A \in \mathcal{F}_0$, then $A^c \in \mathcal{F}_0$,

(iii) if $A_i \in \mathcal{F}$, $i = 1, 2, ...$, then $\cup A_i \in \mathcal{F}$.

σ-algebras are closed[4] under complementation and under countable unions. You will encounter different countable union notations, $\cup A_i, \cup_{i=1}^{\infty}, \cup_{i\in\mathbb{N}}$ all are defined as in the mathematical review, that is as the set of elements which belong to at least one of the $A_1, A_2, ...$

Exercise 3.1. *Show that a $\sigma-$ algebra is also closed under countable intersections. Hint: you need (ii) and a property in the mathematical review.*

Now, before we informally called events as those subsets of Ω which are of interest. Formally, an event is an element of the σ-algebra. We will also call the events as $\mathcal{F}$-measurable sets. Note, that $\mathcal{F}$ is a *collection* of sets, hence, its elements are sets. A trivial σ-algebra would be $\mathcal{F} = \{\emptyset, \Omega\}$. Another trivial σ-algebra is $\mathcal{F} = 2^{\Omega}$, that is, all subsets of Ω. If we use the power set as a σ-algebra, we do not lose any information, however, as we will see this will not always be possible[5]. Another example of a σ-algebra is the smallest σ-algebra containing $A \subseteq \Omega$, denoted by $\sigma(A)$, $\sigma(A) = \{\Omega, \emptyset, A, A^c\}$.

Exercise 3.2. *Show that all these examples of σ-algebras are indeed σ-algebras.*

So now we already have some more structure. Specifically, we have a sample space Ω and a σ-algebra defined on it. This tuple $(\Omega, \mathcal{F})$ we call a *measurable space*. It is a space to which a measure can be assigned, if not it would not be measurable, hence the logical step now is to define what a measure is.

Definition 3.5 (Measure)**.** *A measure is a function $\mu : \mathcal{F} \to [0, \infty)$ such that*

[3] We say that some set is closed under some operation, if you apply this operation to different elements of the set and you get an element of the set. For instance, (ii) can be read as the algebra being closed under complementation.

[4] see the footnote above ↑

[5] If Ω is uncountable it will not be feasible to take $\mathcal{F} = 2^{\Omega}$.

(i) $\mu(\emptyset) = 0$,

(ii) if $A_1, A_2, ...$ *is a countable collection of* $\mathcal{F}$*-measurable disjoint sets, then*

$$\mu\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mu(A_i).$$

Unsurprisingly, $(\Omega, \mathcal{F}, \mu)$ is called a *measure space*. If $\mu(\Omega) < \infty$, μ is called a finite measure. If there exists a sequence $A_1, A_2, ...$ of subsets of Ω such that $\cup A_i = \Omega$ and $\mu(A_i) < \infty$ for all i, then μ is said to be a σ-finite measure.

Example 3.1 (Counting measure)**.** *Suppose* $\Omega = \{a_1, ..., a_n\}$ *and that* $\mathcal{F}$ *is some* σ*-algebra defined on* Ω*. Let the counting measure be defined as*

$$\nu(A) = \sum_{i=1}^{n} \delta_{a_i}(A) \quad A \in \mathcal{F},$$

where $\delta_{a_i}(A)$ *is a set function*[6] *which is equal to one if* $a_i \in A$ *and zero otherwise. Hence it counts how many elementary outcomes are in the* $\mathcal{F}$*-measurable set* A*.* $(\Omega, \mathcal{F}, \nu)$ *is an example of a measure space.*

Example 3.2 (Lebesgue measure)**.** *Suppose* $\Omega = \mathbb{R}$ *and that we have some* σ*-algebra* $\mathcal{F}$ *defined on it which contains closed intervals (i.e.* $[a, b] \subseteq \mathbb{R}$*)*[7]*. Define the Lebesgue measure as*

$$\lambda([a, b]) = b - a.$$

$(\Omega, \mathcal{F}, \lambda)$ *is another example of a measure space.*

If $\mu(\Omega) = 1$, μ is called a *probability measure*. Since this is really important let us state a proper definition even though it is almost the same as the definition of a measure.

Definition 3.6 (Probability measure)**.** *A probability measure* $\mathbb{P}$ *on* $(\Omega, \mathcal{F})$ *is a function* $\mathbb{P} : \mathcal{F} \to [0, 1]$ *such that*

(i) $\mathbb{P}(\emptyset) = 0$.

(ii) (Countable additivity) If $A_1, A_2, ...$ *is a countable collection of* $\mathcal{F}$*-measurable disjoint sets, then*

$$\mathbb{P}\Big(\bigcup_{i=1}^{\infty} A_i\Big) = \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

$(\Omega, \mathcal{F}, \mathbb{P})$ is called a *probability space*. In summary, we have introduced the concept of a random experiment, we have defined all its possible outcomes as the sample space Ω, we have defined a collection of subsets of Ω, $\mathcal{F}$, which contains the sets which are of interest, these sets are what we have called events or $\mathcal{F}$-measurable sets, then we have defined $(\Omega, \mathcal{F})$ to be a measure space. Finally, we have assigned a probability measure $\mathbb{P}$ on the measure space which gives the probability of all $\mathcal{F}$-measurable sets and we called $(\Omega, \mathcal{F}, \mathbb{P})$ a probability space.

[6] Takes a set as an input and gives a number as an output.

[7] We will talk in detail about how such a σ-algebra can be created, for now take it as given.

Now we introduce some useful properties of probability measures. Note these properties are properties of measures in general for the special case that we deal with a probability measure.

Proposition 3.1 (Properties of Probability Measures)**.**

(i) (Finite additivity) If $A_1, ..., A_n \in \mathcal{F}$ are disjoint then

$$\mathbb{P}(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mathbb{P}(A_i)$$

Proof. Follows from countable additivity, take a countable sequence of sets $B_1, B_2, ...$ such that $B_i = A_i$ if $i \leq 0$ and $B_i = \emptyset$ if $i > 0$. Then $\cup_{i=1}^\infty B_i = \cup_{i=1}^n A_i$. Hence

$$\mathbb{P}(\cup_{i=1}^n A_i) = \mathbb{P}(\cup_{i=1}^\infty B_i) = \sum_{i=1}^\infty \mathbb{P}(B_i) = \sum_{i=1}^n \mathbb{P}(A_i) + \sum_{i+1}^\infty \mathbb{P}(B_i) = \sum_{i=1}^n \mathbb{P}(A_i).$$

□

(ii) $\mathbb{P}(A^c) = 1 - \mathbb{P}(A)$.

(iii) (Monotonicity) If $A \subseteq B$, $A, B \in \mathcal{F}$, then

$$\mathbb{P}(A) \leq \mathbb{P}(B).$$

Proof. Note that $B = A \cup B \setminus A$. A and $B \setminus A$ are disjoint, so

$$\mathbb{P}(B) = \mathbb{P}(A) + \underbrace{\mathbb{P}(B \setminus A)}_{\geq 0} \implies \mathbb{P}(B) \geq \mathbb{P}(A).$$

□

(iv) If $A_1, ..., A_n \in \mathcal{F}$, then

$$\mathbb{P}(\cup_{i=1}^n A_i) = \sum_{i=1}^n \mathbb{P}(A_i) - \sum_{i<j} \mathbb{P}(A_i \cap A_j) + \sum_{i<j<k} \mathbb{P}(A_i \cap A_j \cap A_k) - ... + (-1)^{n-1} \mathbb{P}(\cap_{i=1}^n A_i).$$

Proof. I do it for two sets A and B, the general proof uses induction. We can divide $A \cup B$ into the union of three disjoint sets

$$A \cup B = (A \cap B^c) \cup (B \cap A^c) \cup (A \cap B).$$

We can write

$$\begin{aligned} A = (A \cap B^c) \cup (A \cap B) &\implies \mathbb{P}(A) = \mathbb{P}(A \cap B^c) + \mathbb{P}(A \cap B) \\ B = (B \cap A^c) \cup (A \cap B) &\implies \mathbb{P}(B) = \mathbb{P}(B \cap A^c) + \mathbb{P}(A \cap B). \end{aligned}$$

Hence,

$$\mathbb{P}(A \cup B) = \mathbb{P}(A) - \mathbb{P}(A \cap B) + \mathbb{P}(B) - \mathbb{P}(A \cap B) + \mathbb{P}(A \cap B) = \mathbb{P}(A) + \mathbb{P}(B) - \mathbb{P}(A \cap B).$$

□

(v) (Continuity) If $A_1, A_2, ... \in \mathcal{F}$*, then*

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \lim_{m\to\infty} \mathbb{P}(\cup_{i=1}^{m} A_i)$$

Proof. What this property means is far from obvious. To remember it you can think of it as "taking the limit inside" as you would do with a continuous function. However, that is not really what is going on. It is a really useful property for proofs. Here I give a sketch of the proof. We define $B_1 = A_1$, $B_2 = A_2 \setminus A_1$,..., $B_n = A_n \setminus \cup_{i=1}^{n-1} A_i$,... By Exercise 2.1 we know that the B_i's are disjoint and that $\cup_{i=1}^{\infty} A_i = \cup_{i=1}^{\infty} B_i$. Hence

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \mathbb{P}(\cup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} \mathbb{P}(B_i) = \lim_{m\to\infty} \sum_{i=1}^{m} \mathbb{P}(B_i) = \lim_{m\to\infty} \mathbb{P}(\cup_{i=1}^{m} B_i) = \mathbb{P}(\cup_{i=1}^{m} A_i).$$

We have used Exercise 2.1, countable additivity and Exercise 2.1, definition of infinite sum, finite additivity, Exercise 2.1 respectively in each equality above. □

(vi) (Corollaries of continuity, some textbooks define this as continuity) If $A_1 \subset A_2 \subset A_3$*... (nested increasing) (or* $A_1 \supset A_2 \supset A_3$*... (nested decreasing, like Russian dolls)) and* $A_1, A_2, ... \in \mathcal{F}$*, then*

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \lim_{m\to\infty} \mathbb{P}(A_m) \quad \Big(or \ \mathbb{P}(\cap_{i=1}^{\infty} A_i) = \lim_{m\to\infty} \mathbb{P}(A_m) \Big),$$

which can also be written as $\mathbb{P}(\lim_{m\to\infty} A_m) = \lim_{m\to\infty} \mathbb{P}(A_m)$.

(vii) (Subadditivity) If $A_1, A_2, ... \in \mathcal{F}$*, then*

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i).$$

Proof. Again define $B_i = A_i \setminus \cup_{j=1}^{i-1} A_j$. We know that $\cup_{i=1}^{\infty} A_i = \cup_{i=1}^{\infty} B_i$ and that the B_i's are disjoint by Exercise 2.1. Then

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i) = \mathbb{P}(\cup_{i=1}^{\infty} B_i) = \sum_{i=1}^{\infty} \mathbb{P}(B_i),$$

since $\mathrm{B}_i \subseteq A_i$, $\mathbb{P}(B_i) \leq \mathbb{P}(A_i)$ for all i, so

$$\sum_{i=1}^{n} \mathbb{P}(B_i) \leq \sum_{i=1}^{n} \mathbb{P}(A_i) \text{ for all } n \geq 1 \implies \sum_{i=1}^{\infty} \mathbb{P}(B_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$
$$\implies \mathbb{P}(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mathbb{P}(A_i)$$

□

4 Discrete Probability Spaces

If Ω is countable (finite or countably infinite) we can always take the σ-algebra to be $\mathcal{F} = 2^{\Omega}$, the collection of all subsets of Ω. Now, we have a definition of a probability measure and we have a σ-algebra, our task is to assign probabilities. This means that we have to come up with a probability measure which satisfies the definition and assigns probabilities to all events ($\mathcal{F}$-measurable sets) contained in 2^{Ω}.

The probability of each $A \in \mathcal{F}$ is going to be defined in terms of the probabilities of the singleton subsets[8], $\mathbb{P}(\{\omega\})$. For any $A \in \mathcal{F}$ we are gonna assign a probability measure such that

$$\mathbb{P}(A) = \sum_{\omega \in A} \mathbb{P}(\{\omega\}) \text{ and } \sum_{\omega \in \Omega} \mathbb{P}(\{\omega\}) = 1.$$

Example 4.1.

(i) $\Omega = \{H, T\}$ *and* $\mathcal{F} = 2^{\Omega}$. *We need to assign a probability to each singleton. We can let* $\mathbb{P}(\{H\}) = p$ *and* $\mathbb{P}(\{T\}) = 1 - p$ *for* $0 \leq p \leq 1$.

(ii) $\Omega = \mathbb{N}$ *and* $\mathcal{F} = 2^{\mathbb{N}}$. *We need to assign* $\mathbb{P}(\{k\})$ *for each* $k = 1, 2, ...$ *such that* $\sum_{k=1}^{\infty} \mathbb{P}(\{k\}) = 1$. *There are many ways of doing this. One way would be*

$$\mathbb{P}(\{k\}) = \frac{1}{2^k} \text{ for } k \in \mathbb{N},$$

another way

$$\mathbb{P}(\{k\}) = (1-p)^{k-1} p \text{ for } k \in \mathbb{N} \text{ and } 0 \leq p \leq 1.$$

(iii) $\Omega = \{0\} \cup \mathbb{N}$ *and* $\mathcal{F} = 2^{\Omega}$, *one valid probability assignment would be*

$$\mathbb{P}(\{k\}) = \frac{e^{-\lambda} \lambda^k}{k!} \text{ for } k = 0, 1, 2, ... \text{ and } \lambda > 0.$$

These are examples of different ways of assigning probabilities in discrete probability spaces. Intuitively, you can think about this procedure as having sum mass which adds up to one which you have to distribute among a countable number of points.

5 Uncountable Ω

Here we are going to follow a specific motivating example. Suppose that your random experiment consists in throwing a dart to the $(0, 1]$ line. Then your sample space is $\Omega = (0, 1]$. Suppose you want to assign probabilities in such a way that makes it equally likely for you to hit any part of the line.

Now, assigning probabilities to singletons is not going to work. If you assign a probability to some $\{\omega\}$ it cannot be strictly positive. This is because we want to distribute probability uniformly, implying that we would want all singletons to have the same probability. But

[8] A singleton subset is $\{\omega\}$, which is different from an elementary outcome ω. One is a set (even if it contains only one element) and the other is an element.

then if we assign a strictly positive probability everything blows up. This implies that we can only put zero probability to all singletons.

One of the defining properties of a probability measure is that it is countably additive. If we have assigned $\mathbb{P}(\{\omega\}) = 0$ for all $\omega \in \Omega$ and we want to measure an interval, for instance $[1/2, 1]$, we cannot. $[1/2, 1]$ is an *uncountable* union of singletons. The solution to this problem is to forget about singletons altogether and focus on subsets of $\Omega = (0, 1]$ which are of interest. Remember this is a specific motivating example in which we are trying to assign a *uniform* probability measure to $(0, 1]$, that is, a probability measure which tells us that the dart is equally likely to land anywhere. If we want this probability measure to be uniform, there are two conditions it must satisfy (asides from the definition of a probability measure)

(i) (We do not care about singletons condition) For $a, b \in (0, 1]$, $a \leq b$

$$\mathbb{P}((a, b)) = \mathbb{P}((a, b]) = \mathbb{P}([a, b)) = \mathbb{P}([a, b]).$$

(ii) (Translation invariance): we want intervals of the same length to have the same probability (i.e. probability does not change if you move the set around).

Now we state a general (for any measure not only probability measures) impossibility theorem which we will not prove

Theorem 5.1 (Impossibility Theorem)**.** *There exists no measure $\mu(A)$ defined on 2^Ω (i.e. all subsets of $[0, 1]$), satisfying (i) and (ii).*

The takeaway is that when Ω is uncountable we cannot pick $\mathcal{F} = 2^\Omega$. We cannot keep all subsets, we need to pick less subsets which means we need to specify a smaller σ-algebra. Hence, we cannot dismiss the question of what is interesting to us just by picking all subsets as we do with discrete probability spaces. What subsets should we pick? Borel and Lebesgue when faced with this problem while developing measure theory found that focusing on intervals is a good solution. Of course the collection of all intervals is not a σ-algebra since the complement of an interval is not necessarily an interval. Hence, we have to somehow generate a σ-algebra which contains all intervals. In the following suppose that our collection of subsets of interest is $\mathcal{C}$ (e.g. collection of intervals). We are going to see how to generate a σ-algebra from $\mathcal{C}$.

5.1 Generated σ-algebras

Let $\mathcal{C}$ be an arbitrary collection of subsets of Ω.

Theorem 5.2. *There exists a unique σ-algebra, say $\sigma(\mathcal{C})$, which is the smallest σ-algebra containing $\mathcal{C}$. That is, if $\mathcal{H}$ is any σ-algebra that contains $\mathcal{C}$, then $\sigma(\mathcal{C}) \subseteq \mathcal{H}$. $\sigma(\mathcal{C})$ is called the σ-algebra generated by $\mathcal{C}$.*

Proof. Let $\{\mathcal{F}_i, i \in I\}$ be the collection of *all* σ-algebras which contain $\mathcal{C}$ (a collection of collections of sets!). Note that this collection will not be empty since 2^Ω will always be there. We can prove that

$$\sigma(\mathcal{C}) = \cap_{i \in I} \mathcal{F}_i.$$

To prove it we need to show three things

(i) $\sigma(\mathcal{C})$ is a σ-algebra (Exercise)

(ii) $\mathcal{C} \subseteq \sigma(\mathcal{C})$. This is true since all σ-algebras in the intersection contain $\mathcal{C}$.

(iii) It is the smallest σ-algebra. To see this, let $\mathcal{H}$ a σ-algebra such that $\mathcal{C} \subseteq \mathcal{H}$, then $\mathcal{H} \in \{\mathcal{F}_i, i \in I\}$ since $\{\mathcal{F}_i, i \in I\}$ is the collection of all σ-algebras which contain $\mathcal{C}$. Hence, there exists an $i \in I$ such that $\mathcal{H} = \mathcal{F}_i$ which implies that $\sigma(\mathcal{C}) \subseteq \mathcal{H}$.

□

Exercise 5.1. *Show that the intersection of σ-algebras are σ-algebras (note this is not true for unions).*

5.2 Borel σ-algebra

Let $\Omega = (0, 1]$ and $\mathcal{C}_0$ be the collection of all open intervals of Ω.

Definition 5.1. *$\sigma(\mathcal{C}_0)$ is called the Borel σ-algebra on $(0, 1]$, denoted by $\mathcal{B}((0, 1])$.*

Definition 5.2. *Elements of $\mathcal{B}((0, 1])$ are called Borel-measurable sets, or simply Borel sets.*

Note that $\mathcal{B}((0, 1])$ is well-defined by Theorem 5.2. However, nothing we have done tells us that it is not 2^Ω. However, it turns out (it can be shown) that it is much smaller than 2^Ω. It actually has the same cardinality as $\mathbb{R}$. Also, it is quite hard to find sets in 2^Ω which are not in $\mathcal{B}((0, 1])$. In sum, the Borel σ-algebra buys us a lot with very little sacrifice. Now we show some useful propositions.

Proposition 5.1. *Let $b \in (0, 1]$. Then the singleton $\{b\}$ is a Borel set.*

Proof. It is true because it can be written as a countable intersection of Borel sets

$$\{b\} = \bigcap_{n=1}^{\infty} \left[\left(b - \frac{1}{n}, b + \frac{1}{n} \right) \cap \Omega \right].$$

Since all sets in the intersection contain b, $\{b\}$ is a subset of the intersection. Now, to show that the intersection is a subset of $\{b\}$, take any $c \in (0, 1]$ which is different than b. I can find an n_0 such that for all $n \geq n_0$,

$$c \notin \left(b - \frac{1}{n}, b + \frac{1}{n} \right).$$

□

Proposition 5.2. *$(a, b], [a, b], [a, b)$ are Borel sets.*

Proof. $(a, b] = (a, b) \cup \{b\}$ which is a countable union of Borel sets. $[a, b] = \{a\} \cup (a, b) \cup \{b\}$ which is also a countable union of Borel sets. $[a, b) = \{a\} \cup (a, b)$ which is also a countable union of Borel sets. □

Curiosities: there are really weird sets which are Borel sets. For instance, Cantor sets, which have an interesting fractal like behaviour, are Borel sets. An example of a non Borel-measurable set is the Vitali set. These two sets have extensive Wikipedia articles you can check if interested. An interesting paradox which comes out when dealing with non-measurable sets is the Banach-Tarski paradox[9] which states that you can decompose a ball into a finite number of disjoint subsets and then put these subsets back together in a way in which you get two balls identical to the original one. As you might suspect, these disjoint subsets are not Borel sets.

So remember our motivating example was to put a uniform measure on $(0, 1]$. Now we are going to work with the measurable space $((0, 1], \mathcal{B}((0, 1]))$. You can consider the next part to be optional (if you are studying this for the first part in an Econometrics course for instance). What we are going to do is the same as we did with discrete probability spaces. Which is to assign a measure to all sets belonging to the σ-algebra. In our case we are gonna assign a measure to all Borel sets. However, to do this with Borel sets we need a more advanced mathematical machinery. If you skip the next section, all you need to know is that we are going to assign a measure to half-closed intervals $(a, b]$ proportional to their length (i.e. $\mu((a, b]) = b - a$. Then there will be a theorem (which we will not prove) which says that this measure we have put on half-closed intervals extends uniquely to all Borel sets.

5.3 Caratheodory's Extension*

Define $\mathcal{F}_0$ as the collection of subsets of Ω which are finite unions of disjoint intervals of the form $(a, b]$ plus the empty set. A typical element of $\mathcal{F}_0$ is $(a_1, b_1] \cup (a_2, b_2] \cup ... \cup (a_n, b_n]$, where $a_1 < b_1 \leq a_2 < b_2 ... \leq a_n < b_n$. We want a measure of each of these intervals which is proportional to their length (i.e. $b_1 - a_1$). We want to extend this measure to all (including really weird sets) Borel sets. For this we will use Caratheodory's[10] Extension theorem. But first we need a lemma about the collection $\mathcal{F}_0$.

Lemma 5.1.

(i) $\mathcal{F}_0$ *is an algebra.*

Proof. $\Omega \in \mathcal{F}_0$ clearly and finite unions are also elements of $\mathcal{F}_0$, hence it is an algebra. □

(ii) $\mathcal{F}_0$ *is* not *a* σ*-algebra.*

Proof. Consider

$$A_n = \left(0, \frac{n}{n+1}\right] \; n = 1, 2, ...$$

Note that $A_n \in \mathcal{F}_0$ for all n. However,

$$\cup_{n=1}^{\infty} A_n = \cup_{n=1}^{\infty} \left(0, \frac{n}{n+1}\right] = (0, 1) \notin \mathcal{F}_0$$

[9] A really cool video with amazing visualizations of the paradox: https://www.youtube.com/watch?v=s86-Z-CbaHA.

[10] Greek mathematician who lived from 1873 to 1950

□

(iii) $\sigma(\mathcal{F}_0) = \mathcal{B}((0,1])$.

Proof. To show that $\sigma(\mathcal{F}_0) \subseteq \mathcal{B}((0,1])$, it is enough to show that $\mathcal{F}_0 \subseteq \mathcal{B}((0,1])$, since this implies that $\sigma(\mathcal{F}_0)$ is a sub-σ-algebra of $\mathcal{B}((0,1])$ because it is the smallest σ-algebra containing $\mathcal{F}_0$. We have shown that $(a,b] \in \mathcal{B}((0,1])$ (in Proposition 5.2), so $\mathcal{F}_0 \subseteq \mathcal{B}((0,1])$ which implies that $\sigma(\mathcal{F}_0) \subseteq \mathcal{B}((0,1]) = \sigma(\mathcal{C})$.

To show that $\mathcal{B}((0,1]) \subseteq \sigma(\mathcal{F}_0)$, we proceed as follows

$$(a,b) = \cup_{n=1}^{\infty}\left(a, b - \frac{1}{n}\right] \implies \mathcal{C} \subseteq \sigma(\mathcal{F}_0) \implies \sigma(\mathcal{C}) \subseteq \sigma(\mathcal{F}_0).$$

□

So we want to define a measure. We will call it $\mathbb{P}$ since it will be a probability measure given that we are working with the zero-one interval. This measure has to be proportional to length of the intervals. Hence, one property that we desire is that for $F = (a_1, b_1] \cup (a_2, b_2] \cup ... \cup (a_n, b_n] \in \mathcal{F}_0$ we have that

$$\mathbb{P}(F) = \sum_{i=1}^{n}(b_i - a_i).$$

However, measures are defined on σ-algebras not just algebras. So we cannot define a measure $\mathbb{P}$ on $\mathcal{F}_0$ and call it a measure. The step from algebra to σ-algebra is given by Caratheodory's Extension theorem which we state but do not prove.

Theorem 5.3 (Caratheodory's Extension Theorem)**.** *Let $\mathcal{F}_0$ be an algebra on Ω and let $\mathcal{F} = \sigma(\mathcal{F}_0)$. Suppose $\mathbb{P}_0 : \mathcal{F}_0 \to [0,1]$ such that $\mathbb{P}_0(\Omega) = 1$ and $\mathbb{P}_0$ is countably additive in $\mathcal{F}_0$. Then $\mathbb{P}_0$ can be* uniquely *extended to a probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$. That is, there exists a unique probability measure $\mathbb{P}$ on $(\Omega, \mathcal{F})$ such that*

$$\mathbb{P}(A) = \mathbb{P}_0(A) \text{ for all } A \in \mathcal{F}_0.$$

Note that $\mathbb{P}_0$ is *not* a probability measure. Also, we have stated the theorem for the special case of probability theory, note that the general theorem holds for general measures. In fact, $\mathbb{P}_0$ is a pre-measure with $\mathbb{P}_0(\Omega) = 1$, the general theorem shows that a pre-measure on an algebra can be extended to a measure and uniquely so if the pre-measure is σ-finite. When we say that $\mathbb{P}_0$ is countably additive in $\mathcal{F}_0$, we mean that it is countably additive for those countable union in $\mathcal{F}_0$ (since $\mathcal{F}_0$ is an algebra and not a σ-algebra, a countable unions of elements in $\mathcal{F}_0$ need not belong in $\mathcal{F}_0$).

So, back to our motivating example. We have that $\mathcal{F}_0$ defined as collection of finite unions of half-closed intervals plus the empty set is an algebra. We showed that $\sigma(\mathcal{F}_0) = \mathcal{B}((0,1])$. Let us define $\mathbb{P}_0 : \mathcal{F}_0 \to [0,1]$ such that $\mathbb{P}_0(\emptyset) = 0$ and $\mathbb{P}_0(F) = \sum_{i=1}^{n}(b_i - a_i)$. Next, we need to verify countable additivity of $\mathbb{P}_0$ in $\mathcal{F}_0$.

Exercise 5.2. *For any $F_1, F_2, ... \in \mathcal{F}_0$ such that $\cup_{i=1}^{\infty} F_i \in \mathcal{F}_0$ we have that $\mathbb{P}_0(\cup_{i=1}^{\infty}) = \sum_{i=1}^{\infty}\mathbb{P}(F_i)$.*

Now, all conditions of Caratheodory's Theorem hold. Then, by the theorem we have that there exists a unique probability measure $\mathbb{P}$ on $(\Omega, \mathcal{B}((0,1]))$ which agrees with $\mathbb{P}_0$ on $\mathcal{F}_0$. Which this basically means is that for all Borel sets in $(0,1]$ I can define a unique probability measure which corresponds to the notion of length. This unique measure corresponds to the notion of length because $\mathbb{P}_0$ *is* length and $\mathbb{P}$ is the unique measure defined on all Borel sets which agrees with $\mathbb{P}_0$. Again, since it is a measure, it is defined for all Borel sets. So even if you give it a weird Borel set (such as a Cantor) it will assign to it a measure which corresponds with length. Basically, if you recall the impossibility theorem, $\mathbb{P}$ is a measure defined on $\mathcal{B}((0,1])$ (not on 2^Ω) which satisfies the two conditions we desired[11]. This measure $\mathbb{P}$ is called the *Lebesgue measure* on $(0,1]$.

So we know this Lebesgue measure $\mathbb{P}$ exists and that it is unique, however we only know its explicit form for elements of $\mathcal{F}_0$, not for all Borel sets. Now we are going to see how to construct the Lebesgue measure for some commonly encountered Borel sets which are not in $\mathcal{F}_0$.

5.4 Lebesgue measure on $(0,1]$*

Example 5.1. *Singleton* $\{b\}$*:*

$$\mathbb{P}(\{b\}) = \mathbb{P}\Big(\bigcap_{n=1}^{\infty} \underbrace{\Big(b - \frac{1}{n}, b + \frac{1}{n}\Big] \cap \Omega}_{\equiv B_n}\Big)$$

Note that $B_1, B_2, ...$ *are nested decreasing sets! Hence, we can use continuity of the probability measure*

$$\mathbb{P}(\{b\}) = \lim_{n\to\infty} \mathbb{P}(B_n) \leq \lim_{n\to\infty} \Big(b + \frac{1}{n} - b + \frac{1}{n}\Big) = \lim_{n\to\infty} \frac{2}{n} = 0.$$

The $\leq$ *follows from the fact that* B_n *is intersected with* Ω*. One can show that* $\mathbb{P}((a,b)) = \mathbb{P}([a,b)) = \mathbb{P}((a,b]) = \mathbb{P}([a,b]) = b - a$*.*

Example 5.2. $\mathbb{P}($*all rational numbers*$) = \mathbb{P}(\mathbb{Q} \cap \Omega) = 0$*. Note that rationals are Borel sets since* $\mathbb{Q}$ *is countable and singletons are Borel sets with Lebesgue measure zero. In fact, any countable subset of* $(0,1]$ *will have zero probability.*

Typical confusion: the probability of an event being zero *does not* mean it cannot occur. It is an elementary outcome in your sample space so it can occur. Since the sample space is the set of all possible outcomes. As long as the event is not empty it can occur.

Example 5.3. $\mathbb{P}($*irrationals*$) = 1 - \mathbb{P}(\mathbb{Q} \cap \Omega) = 1$*. However this does not mean that irrationals happen for sure! It is not the case that the set of possible outcomes is equal to the set of all irrationals,* $\Omega \neq \{$*irrationals*$\}$*. What we say is that an irrational happens almost surely (a.s) or with probability one.*

[11] The impossibility theorem stated that there existed not measure in 2^Ω which satisfied those two properties. We have basically settled down for a much smaller σ-algebra, the Borel σ-algebra and found a unique measure on this σ-algebra which does satisfy those two conditions.

6 Lebesgue measure on $\mathbb{R}$*

The Lebesgue measure can be defined also for $\mathbb{R}$ and not only for $(0,1]$. However note that the Lebesgue measure in $\mathbb{R}$ is not a probability measure anymore.

Definition 6.1. *Let $\mathcal{C}_0$ be the collection of all open intervals in $\mathbb{R}$, then $\mathcal{B}(\mathbb{R}) \equiv \sigma(\mathcal{C}_0)$.*

Definition 6.2. *Let $\mathcal{D}$ be the collection of semi-infinite intervals*

$$\mathcal{D} = \{(-\infty, x] : x \in \mathbb{R}\}$$

Then, $\mathcal{B}(\mathbb{R}) \equiv \sigma(\mathcal{D})$.

We have defined the Borel σ-algebra in $\mathbb{R}$ in two different ways. We should prove that both definitions are equivalent, i.e. $\sigma(\mathcal{C}_0) = \sigma(\mathcal{D})$. The first definition is the one that is usually given. The second one is a more operational definition. But they are indeed equivalent. To construct the Lebesgue measure on $\mathbb{R}$ we have to repeat the exact same story as we did for $(0,1]$. Define an algebra $\mathcal{F}_0$, take a pseudo-measure λ_0 which can be interpreted as length. Check that Caratheodory's theorem holds. Then by the theorem there exists a unique measure λ on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ which is the Lebesgue measure on $\mathbb{R}$.

7 Conditional Probability

Let us work now with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let B be an event such that $\mathbb{P}(B) > 0$.

Definition 7.1. *The conditional probability of $A \in \mathcal{F}$ given B is defined as*

$$\mathbb{P}(A \mid B) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)}.$$

We cannot condition of zero probability events. For instance, if we throw a dart to the $[0,1]$ line, the question of what is the probability of hitting within the interval $[0, 1/2]$ given that the dart has landed on a rational number is not well-defined.

Theorem 7.1. *Let $B \in \mathcal{F}$ with $\mathbb{P}(B) > 0$. Then $\mathbb{P}(. \mid B) : \mathcal{F} \to [0,1]$ is a probability measure on $(\Omega, \mathcal{F})$.*

Proof. We need to show that $\mathbb{P}(. \mid B) \in [0,1]$ and that $\mathbb{P}(\cup_{i=1}\infty A_i) = \sum_{i=1}^{\infty} \mathbb{P}(A_i)$ for disjoint $A_1, A_2, ...$ Note the following

$$\begin{aligned}
\mathbb{P}(\Omega \mid B) &= \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(B)}{\mathbb{P}(B)} = 1, \\
\mathbb{P}(\emptyset \mid B) &= \frac{\mathbb{P}(\emptyset \cap B)}{\mathbb{P}(B)} = \frac{\mathbb{P}(\emptyset)}{\mathbb{P}(B)} = 0, \\
0 \leq \mathbb{P}(A \mid B) &= \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(B)} \leq \frac{\mathbb{P}(\Omega \cap B)}{\mathbb{P}(B)} = 1 \text{ for all } A \in \mathcal{F}.
\end{aligned}$$

So $\mathbb{P}(. \mid B) \in [0,1]$. Now take $A_1, A_2, ...$ disjoint, then

$$\mathbb{P}(\cup_{i=1}^{\infty} A_i \mid B) = \frac{\mathbb{P}((\cup_{i=1}^{\infty} A_i) \cap B}{\mathbb{P}(B)} = \frac{\mathbb{P}((\cup_{i=1}^{\infty}(A_i \cap B)))}{\mathbb{P}(B)} = \frac{\sum_{i=1}^{\infty} \mathbb{P}(A_i \cap B)}{\mathbb{P}(B)} = \sum_{i=1}^{\infty} \mathbb{P}(A_i \mid B).$$

Where we have used the definition of conditional probability, then the distributive law for unions and intersections of sets, the fact that $A_i \cap B$ are disjoint sets and countable additivity, and lastly the definition of conditional probability again. □

Note that we are working with the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ all the time. The random experiment does not change. For instance if the random experiment is to throw a dice, the conditioning event could be $B =$ even number and $A =$ getting a two. Everything is done on the basis of the same random experiment. Conditional probabilities are probabilities so they satisfy the properties of probability measures. Now we define further properties which are specific to conditional probabilities.

Proposition 7.1 (Properties of Conditional Probabilities)**.** *(i) (Law of Total Probability) Let $B_i \in \mathcal{F}$, $i = 1, 2, ...$ be a partition of Ω (i.e. $\cup_{i=1}^{\infty} B_i = \Omega$ and $B_i \cap B_j = \emptyset$ for $i \neq j$). Then*

$$\mathbb{P}(A) = \sum_{i=1}^{\infty} \mathbb{P}(A \mid B_i)\, \mathbb{P}(B_i).$$

Proof.

$$\sum_{i=1}^{\infty} \mathbb{P}(A \mid B_i)\, \mathbb{P}(B_i) = \sum_{i=1}^{\infty} \mathbb{P}(A \cup B_i) = \mathbb{P}(\cup_{i=1}^{\infty}(A \cap B_i)) = \mathbb{P}(A \cap (\cup_{i=1}^{\infty} B_i)) = \mathbb{P}(A \cap \Omega) = \mathbb{P}(A).$$

Where we used the definition of conditional probabilities, the fact that $A \cap B_1$, $A \cap B_2$,... are disjoint, countable additivity and the distributive law. One useful example of this property is: $\mathbb{P}(A) = \mathbb{P}(A \mid B)\, \mathbb{P}(B) + \mathbb{P}(A \mid B^c)\, \mathbb{P}(B^c)$. One common example is the following. Suppose you have 1,2,... urns with red and blue balls. The probability of picking a red ball is the sum of the probabilities of picking a red ball in urn 1, times probability of urn 1, plus the probability of picking a red ball in urn 2 times the probability of urn 2 and so on. □

(ii) (Bayes Rule) Let $A \in \mathcal{F}$ with $\mathbb{P}(A) > 0$ and B_i, $i = 1, 2, ...$ as in (i). Then

$$\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(A \mid B_i)\, \mathbb{P}(B_i)}{\sum_{j=1}^{\infty} \mathbb{P}(A \mid B_j)\, \mathbb{P}(B_j)}.$$

Proof.

$$\mathbb{P}(B_i \mid A) = \frac{\mathbb{P}(A \mid B_i)\, \mathbb{P}(B_i)}{\sum_{j=1}^{\infty} \mathbb{P}(A \mid B_j)\, \mathbb{P}(B_j)} = \frac{\mathbb{P}(A \cap B_i)}{\mathbb{P}(A)} = \mathbb{P}(B_i \mid A),$$

where we used property (i). Example: given that you took a red ball what is the probability it came from urn i (posterior). □

(iii) Let $A_i \in \mathcal{F}$, $i = 1, 2, ...$ Then

$$\mathbb{P}(\cap_{i=1}^{\infty} A_i) = \mathbb{P}(A_i) \prod_{i=2}^{\infty} \mathbb{P}(A_i \mid \cap_{j=1}^{i-1} A_j),$$

with $\mathbb{P}(\cap_{j=1}^{i-1} A_j) > 0$ for all i.

Proof.

$$\mathbb{P}(\cap_{i=1}^{\infty} A_i) = \lim_{n\to\infty} \mathbb{P}(\cap_{i=1}^{n} A_i) = \lim_{n\to\infty} \mathbb{P}(A_1) \prod_{i=2}^{n} \mathbb{P}(A_i \mid \cap_{j=1}^{i-1} A_j)$$
$$= \mathbb{P}(A_1) \prod_{i=2}^{\infty} \mathbb{P}(A_i \mid \cap_{j=1}^{i-1} A_j).$$

□

8 Independence

8.1 Independence of events

So we are working with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, we now give a definition of independence of events ($\mathcal{F}$-measurable sets).

Definition 8.1. *Events A and B are said to be independent under $\mathbb{P}$ (or simply independent when measure $\mathbb{P}$ is unambiguous) if $\mathbb{P}(A \cap B) = \mathbb{P}(A)\,\mathbb{P}(B)$.*

We say independent *under* $\mathbb{P}$, since events might be independent under some probability measure but dependent under another probability measure. This is the definition. It is advisable to forget or not pay much attention to previous intuitions you might have. Things like "they are independent if they have nothing to do with each other" can be misleading. Independence of events is what is stated in the definition, not more, not less. You might have encountered another definition, mainly that A and B are independent if $\mathbb{P}(A \mid B) = \mathbb{P}(A)$. This is a consequence of the definition not a definition itself. It adds the requirement that $\mathbb{P}(B) > 0$, which is not required for independence.

Example 8.1. *Statement "A and B are independent if knowing B tells me nothing about A" can be misleading:*

Take $([0,1], \mathcal{B}([0,1]), \mathbb{P})$. Then, $\mathbb{P}(\text{rational} \cap \text{irrational}) = \mathbb{P}(\emptyset) = 0$. And also, $\mathbb{P}(\text{rational}) = 0$ and $\mathbb{P}(\text{irrational}) = 0$. Hence $\mathbb{P}(\text{rational} \cap \text{irrational}) = \mathbb{P}(\text{rational})\,\mathbb{P}(\text{irrational})$. So the event that the elementary outcome is a rational number and the event that it is an irrational number are independent. However, if I know the elementary outcome is an irrational number, I know for sure it is not a rational number. The occurrence of one event rules out the other one, still, they are independent. This happens because they have zero probability, if both events had strictly positive probability the intuition would be fine.

Example 8.2. *Can an event be independent from itself? Yes!*

$$\mathbb{P}(A \cap A) = \mathbb{P}(A)\,\mathbb{P}(A) \iff \mathbb{P}(A) = \mathbb{P}(A)^2 \iff \mathbb{P}(A) = 0 \text{ or } \mathbb{P}(A) = 1.$$

Definition 8.2. *$A_1, A_2, ..., A_n \in \mathcal{F}$ are independent if for all non-empty $I_0 \subseteq \{1, 2, ..., n\}$, we have*

$$\mathbb{P}(\cap_{i\in I_0} A_i) = \prod_{i\in I_0} \mathbb{P}(A_i)$$

The definition above basically tells you that for any finite collection of events, to say they are independent you need to check all combinations. The following extends this to arbitrary collections of events.

Definition 8.3. *Let $\{A_i, i \in I\}$ be an arbitrary collection of events (index does not need to be countable). These events are said to be independent if for all non-empty and finite $I_0 \subseteq I$, we have*

$$\mathbb{P}(\cap_{i\in I_0} A_i) = \prod_{i\in I_0} \mathbb{P}(A_i)$$

Note that in the definition above there might be infinite non-empty finite subsets of I.

Example 8.3. *Suppose we toss a coin infinite times. Then $\Omega = \{0,1\}^\infty$. Let $\mathcal{F}_k$ be the collection of subsets of Ω whose occurrence can be decided by looking at the first k tosses. For instance, the event that there are at least three heads in the first 15 tosses belongs to $\mathcal{F}_{15}$. You can show $\mathcal{F}_n \subseteq \mathcal{F}_m$ for all $n \leq m$ and that $\mathcal{F}_k$ is a σ-algebra (left as an exercise). Then we can show that A_i and A_j, $i \neq j$ are independent events. Without loss of generality suppose that $j > i$, then $A_i, A_j \in \mathcal{F}_j$. Define the following probability measure on $\mathcal{F}_j$, $\mathbb{P}(A) = |A|/2^j$ (check it is a probability measure). Then*

$$\mathbb{P}(A_i \cap A_j) = \frac{2^{j-2}}{2^j} = \frac{1}{4}, \quad \mathbb{P}(A_i) = \frac{2^{i-1}}{2^i}$$

FINISH

Exercise 8.1. *Consider a sequence of events $A_1, A_2, \ldots$. Show that if A_i's are independent, then A_i^c's are independent as well.*

8.2 Independence of σ-algebras

Remember we are working with a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. since $\mathbb{P}$ is defined on $\mathcal{F}$, we are going to look at independence between sub-σ-algebras of $\mathcal{F}$. This is because we need $\mathbb{P}$ to be defined to have a notion of independence. So, for instance, if $\mathcal{F} = \{\emptyset, A, A^c, \Omega\}$ and $\mathcal{F}_1 = \{\emptyset, \Omega\}$, then $\mathcal{F}_1 \subseteq \mathcal{F}$. Or if $\Omega = [0,1]$, $\mathcal{F} = 2^\Omega$ and $\mathcal{F}_1 = \mathcal{B}([0,1])$, then $\mathcal{F}_1 \subseteq \mathcal{F}$.

Definition 8.4. *Two sub-σ-algebras, $\mathcal{F}_1$ and $\mathcal{F}_2$ of $\mathcal{F}$, are said to be independent if for any $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$, A_1 and A_2 are independent.*

Definition 8.5. *Let $\{\mathcal{F}_i, i \in I\}$ be an arbitrary collection of sub-σ-algebras of $\mathcal{F}$ (I might be uncountable). These $\mathcal{F}_i$'s are said to be independent if for any choice of $A_i \in \mathcal{F}_i$, $i \in I$, we have that $\{A_i, i \in I\}$ are independent events*[12].

Comparing all possible events can be extremely difficult. Hence, later we will see that there exists a simpler way of checking independence of sub-σ-algebras. Namely, that if you prove independence in collections of events which are closed under finite intersections (these collections are called π-systems) you are actually done.

[12] As a curiosity: this definition depends on the axiom of choice, which states that for an indexed collection of non-empty sets you can always pick one element of each member of the collection. Even if the index is uncountable. Although this axiom was controversial in its origin, now it is commonly used by most mathematicians and is included in the standard form of axiomatic set theory. Check the wikipedia for more.

9 Borel-Cantelli Lemmas

Now we introduce two important results which will be useful later on.

Lemma 9.1 (First Borel-Cantelli Lemma)**.** *If $A_1, A_2, ...$ is a sequence of events such that $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, then a.s only finitely many A_n's will occur.*

Lemma 9.2 (Second Borel-Cantelli Lemma)**.** *If $A_1, A_2, ...$ are independent events such that $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$, then a.s infinitely many A_n's will occur.*

Before we prove these Borel-Cantelli (BC) lemmas let us make some comments. First let us formally consider the event that infinitely many A_n's occur. This is

$$\{A_n \ i.o\} \equiv \cap_{n=1}^{\infty} \cup_{i=n}^{\infty} A_i$$

To understand this object, let $B_n = \cup_{i=n}^{\infty} A_i$, then $\{A_n \ i.o\} = \cap_{n=1}^{\infty} B_n$. B_n is the event that at least one of $A_n, A_{n+1}, ...$ occurs, we can call it the n-th tail event. Then, $\{A_n \ i.o\}$ is the intersection of these n-th tail events. The event that all B_n's occur, that is, for all n, B_n occurs. Insisting, this means that for every n, at least one of the A_n's occurs. No matter how big your n is, no matter how far you go, you will have at least one of the A_n's after that n occurring. The first BC lemma says that $\mathbb{P}(\{A_n \ i.o\}) = 0$ if $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$. The second BC lemma syas that $\mathbb{P}(\{A_n \ i.o\}) = 1$ if $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ and A_n's are independent. Let us do the proofs now.

Proof of first BC lemma. So we need to show that $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty \implies \mathbb{P}(\{A_n \ i.o\}) = 0$. We can show that $\mathbb{P}(\{A_n \ i.o\}^c) = 1$

$$\{A_n \ i.o\}^c = \left(\cap_{n=1}^{\infty} \cup_{i=n}^{\infty} A_i\right)^c = \cup_{n=1}^{\infty} \cap_{i=n}^{\infty} A_i^c,$$

this event is the event that there exists n_0 such that for all $n \geq n_0$, each of the A_n failed to occur. Note that B_n are nested decreasing ($B_1 \supseteq B_2 \supseteq ...$), hence

$$\mathbb{P}(\cap_{n=1} B_n) = \lim_{n\to\infty} \mathbb{P}(B_n) = \lim_{n\to\infty} \mathbb{P}(\cup_{i=n}^{\infty} A_i) \leq \lim_{n\to\infty} \sum_{i=n}^{\infty} \mathbb{P}(A_i) = 0,$$

where we have used continuity of probability measures, subadditivity of probability measures and the fact that if $\sum_{k=1} b_k < \infty$, then the sequence of tail sums $\sum_{n=k}^{\infty} b_n$ converges to zero. It is in this last step where we used that $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$. □

Proof of second BC lemma. First we prove another lemma which we will need

Lemma. Suppose $0 \leq p_i \leq 1$ is such that $\sum_{i=1}^{\infty} p_i = \infty$. Then $\prod_{i=1}^{\infty}(1 - p_i) = 0$.

Proof. We know that $\ln(1 - x) \leq -x$ for all $x \in [0, 1)$, then

$$\ln \prod_{i=1}^{\infty}(1 - p_i) = \ln\left(\lim_{n\to\infty} \prod_{i=1}^{n}(1 - p_i)\right) = \left(\lim_{n\to\infty} \ln \prod_{i=1}^{n}(1 - p_i)\right) \leq \ln \prod_{i=1}^{k}(1 - p_i).$$

The second equality follows since ln is a continuous function so it can be interchanged with the limit, and the inequality follows since $0 \leq p_i \leq 1$, so if we stop at some $k < \infty$, we will

get a larger number. Now

$$\ln \prod_{i=1}^{k}(1-p_i) = \sum_{i=1}^{k} \ln(1-p_i) \leq \sum_{i=1}^{k}(-p_i) \text{ for all } k \geq 1.$$

Hence, taking $k \to \infty$, $\ln \prod_{i=1}^{\infty}(1-p_i) \to -\infty$ which implies that $\prod_{i=1}^{\infty}(1-p_i) \to 0$. So the lemma is proved.

Now, to show that $\mathbb{P}(\{A_i \text{ } i.o\}) = 1$, note that

$$1 - \mathbb{P}(\{A_i \text{ } i.o\}) = \mathbb{P}(\cup_{n=1}^{\infty} B_n^c) \leq \sum_{n=1}^{\infty} \mathbb{P}(B_n^c).$$

We need to show that the above is zero. This amounts to showing that $\mathbb{P}(B_n^c) = 0$ for all $n \geq 1$. Fix n, and $m \geq n$. Then (if A_i's are independent, A_i^c's are independent too, see exercise 8.1)

$$\mathbb{P}(\cap_{i=n}^{m} A_i^c) = \prod_{i=n}^{m}(1 - \mathbb{P}(A_i)),$$

where we have used independence of events. Hence,

$$\mathbb{P}(B_n^c) = \lim_{m \to \infty} \mathbb{P}(\cap_{i=n} m A_i^c) = \prod_{i=n}^{\infty}(1 - \mathbb{P}(A_i)) = 0 \text{ for all } n \geq 1,$$

where we use continuity of probabilities and the lemma we just proved. Therefore, $\mathbb{P}(\{A_i \text{ } i.o\}) = 1$. □

Example 9.1. *Consider $\Omega = \{0,1\}^{\infty}$, for instance the random experiment of tossing a coin infinite times. Suppose we have on it a σ-algebra which (among others) contains events of the form A_i, denoting the event that the i-th toss is heads. Let $\mathbb{P}$ be a probability measure on $(\Omega, \mathcal{F})$ such that $\mathbb{P}(A_n) = 1/n^2$ for all $n \geq 1$ (there might be many probability measures satisfying this). Since $\sum_{n=1}^{\infty} \mathbb{P}(A_n) < \infty$, the first BC lemma implies that a.s. only finitely many heads will occur. The idea is that if heads are becoming more and more unlikely* fast*, there exists n_0 after which you do not get anymore heads with probability 1.*

Now suppose $\mathbb{P}(A_n) = 1/n$ for all $n \geq 1$ and that the A_n's are independent. Heads are also becoming more and more unlikely but slower. Since now $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$, the second BC lemma implies that a.s. infinitely many heads will occur. That is, there exists no n_0 such that after it you do not get heads anymore with probability one. No matter how big n is, with probability 1 there will be a head in the following tosses. Even though the probability of the head is decreasing at rate n^{-1}! Even if n is one billion, you know that a head will still occur for sure.

Independence in the second BC lemma is sufficient but not necessary. For this reason, there are many more BC lemma's which relax independence in different ways. An example in which there is not "enough independence" for the second BC lemma to hold is the following.

Example 9.2. *Suppose $A_n = E$ for all $n \geq 1$ and that $\mathbb{P}(E) \in (0,1)$. Then $\sum_{n=1}^{\infty} \mathbb{P}(A_n) = \infty$ however, it will not be the case that A_n will occur infinitely often with probability one since $\mathbb{P}(\{A_n \text{ } i.o.\}) = \mathbb{P}(E)$. This happens because of the strong dependence of the A_n's.*

10 Measurable functions and Integration

In this section we will talk about functions between measure spaces and their integration. For now we are going to forget about probabilities and keep a more general measure theoretic approach. However, I will hint at what is coming, which is to define random variables and their expectations. This will just be a special case of what we cover here. Suppose you have two measurable spaces $(\Omega, \mathcal{F})$ and $(\Lambda, \mathcal{G})$. We are going to work with functions $f : \Omega \to \Lambda$. One key aspect we require from these functions is that they are measurable.

Definition 10.1 (Measurable function)**.** *A function $f : \Omega \to \Lambda$ is said to be $\mathcal{F}$-measurable if for every $G \in \mathcal{G}$, the pre-image is $\mathcal{F}$-measurable, $f^{-1}(G) \in \mathcal{F}$, where $f^{-1}(G) = \{\omega \in \Omega : f(\omega) \in G\}$.*

So take any $\mathcal{G}$-measurable set, $G \in \mathcal{G}$. The set of all $\omega \in \Omega$ which map to this set G under the function $f(.)$, is the pre-image of G under f, $f^{-1}(G)$. This subset of Ω need not be an element of $\mathcal{F}$. If it is, and this happens for all possible $G \in \mathcal{G}$, then $f(.)$ is a measurable function. Note that if Ω is countable and we take $\mathcal{F} = 2^{\Omega}$, then all functions defined on Ω are measurable since you cannot have pre-images which are not in the power set.

Example 10.1 (Dirichlet function)**.** *Consider the following function*

$$f_D(x) = \begin{cases} 1 & \text{if } x \in \mathbb{Q} \\ 0 & \text{if } x \in \mathbb{Q}^c \end{cases},$$

where $\mathbb{Q}$ are the rational numbers. The above function is $\mathcal{B}$ measurable since

$$f_D^{-1}(A) = \begin{cases} \mathbb{Q} & \text{if } A = \{1\} \\ \mathbb{Q}^c & \text{if } A = \{0\} \\ \emptyset & \text{if } A = \{0, 1\}. \end{cases}$$

Rational numbers are countable and hence a collection of countable singletons and therefore a Borel set. The complement of a Borel set is a Borel set and the empty set is a Borel set. Hence, the Dirichlet function is measurable.

The concept of measurable functions allows us to define another useful concept about measures

Definition 10.2 (Induced Measure)**.** *Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $(\Lambda, \mathcal{G})$ be a measurable space. Let f be a measurable function from Ω to Λ. The measure induced by f is a measure on $\mathcal{G}$ defined as*

$$\mu \circ f^{-1}(B) = \mu(f^{-1}(B)), \quad B \in \mathcal{G}.$$

The induced measure measures the pre-images of measurable functions. Sometimes it is easier to work with induced measures as we will see when we introduce random variables.

10.1 Quick primer: Riemann Integral

Let $f : \mathbb{R} \to \mathbb{R}$ and suppose we are interested in the area between $f(x)$ and the x-axis for $x \in [a, b] \subset \mathbb{R}$. We could approximate this area by rectangles. There are two ways, to approximate the area from above:

INSERT FIGURE

or approximate it from below:

INSERT FIGURE

Partition the real line into n intervals with cut-offs $(x_1, ..., x_n)$ and let $\Delta x_i \equiv x_i - x_{i-1}$. Define the upper Riemann sum to be:

$$U_n(f) \equiv \sum_{i=1}^{n} \left\{ \sup_{(x_{i-1}, x_i)} f(x) \right\} \Delta x_i,$$

and the lower Riemann sum to be:

$$L_n(f) \equiv \sum_{i=1}^{n} \left\{ \inf_{(x_{i-1}, x_i)} f(x) \right\} \Delta x_i.$$

$U_n(f)$ and $L_n(f)$ are the approximations by rectangles from above and below respectively. The larger n is, the more accurate these approximations are. It can be shown that $U_n(f)$ is a monotonically decreasing sequence and that $L_n(f)$ is a monotonically increasing sequence and both sequences are bounded. Hence, $\lim_{n\to\infty} U_n(f)$ and $\lim_{n\to\infty} L_n(f)$ exist. Also, $U_n(f) \geq L_n(f)$ for all n. f is said to be Riemann integrable over (a, b) if $\lim_{n\to\infty} U_n(f) = \lim_{n\to\infty} L_n(f)$. This common values is denoted by

$$\int_a^b f(s)\, ds.$$

Lebesgue proved a theorem characterising Riemann integrable functions

Theorem 10.1 (Lebesgue's criterion)**.** *Let $f : [a, b] \to \mathbb{R}$, then f is Riemann integrable if and only if f is bounded and the set of discontinuities has Lebesgue measure* 0*.*

The theorem requires boundedness and that f does not have "too many discontinuities".

Example 10.2 (Dirichlet function is not Riemann integrable)**.** *Consider the Dirichlet function on* $[0, 1]$ *(call it f_D) taking value one if $x \in [0, 1]$ is a rational number and zero otherwise. Since the rational numbers are dense in the real numbers,* $\sup_{x\in[a,b]} f_D(x) = 1$ *and* $\inf_{x\in[a,b]} f_D(x) = 0$ *for any interval* $[a, b]$ *in* $[0, 1]$*. Hence, the lower and upper Riemann sums will never be equal and hence the Dirichlet function is not Riemann integrable.*

To see the example in the light of theorem note that the Dirichlet function on $[0, 1]$ is everywhere discontinuous. To see this note that the rationals are dense in the reals which means that for any $x \in \mathbb{R} \setminus \mathbb{Q}$ you can construct a sequence x_n such that $x_n \in \mathbb{Q}$ for all n and $\lim_{n\to\infty} = x$. Now, $f(x_n) = 1$ for all n so $\lim_{n\to\infty} f(x_n) = 1$ but $f(x) = 0$ for all $x \in \mathbb{R} \setminus \mathbb{Q}$. Since you can also approximate any real number with a sequence of irrational numbers hence a similar argument applies at rational numbers. Hence, the set of points at

which the Dirichlet function on $[0,1]$ is discontinuous is $[0,1]$ and the Lebesgue measure of this set is $\lambda([0,1]) = 1 \neq 0$ and hence the Dirichlet function is not Riemann integrable.

10.2 Abstract (or Lebesgue) integration

Let $(\Omega, \mathcal{F}, \mu)$ be a measure space and $f : \Omega \to [0, \infty]$ be a measurable function. We want to define the following object

$$\int_A f \, d\mu, \quad A \in \mathcal{F}.$$

From now on we are going to stop thinking about "integral of a function with respect to a variable over an interval" and start thinking about "integral of a measurable function with respect to a measure over an $\mathcal{F}$-measurable set". This will naturally generalize the concept of the Riemann integral to include more general measures. This will become clearer as we move along. For example if we let $(\Omega, \mathcal{F}, \mu) = (\mathbb{R}, \mathcal{B}, \lambda)$ be the measure space, where $\mathcal{B}$ and λ are the Borel σ-algebra and the Lebesgue measure respectively

$$\int_A f \, d\lambda, \quad A \in \mathcal{B},$$

will be read as the integral of f with respect to the Lebesgue measure over a Borel set A.

Before we go into the definition of this object let us talk about some naming issues and the need for abstract integration even though we have Riemann integration. Unfortunately, Lebesgue integration is used for both abstract integration with respect to any general measure and to integration with respect to the Lebesgue measure. The key difference in the construction of the abstract integral as opposed to the Riemann integral is that while the latter partitions the domain into intervals, abstract integration partitions the range, i.e. the values of the function itself. This will in turn allow for a much more flexible partition of the domain. An important problem mathematicians found with the Riemann integral is that many functions are not Riemann integrable, for example the Dirichlet function. The problem being that Riemann integration does not allow for "too many" discontinuities. Abstract integration works on a much broader set of functions. Also, abstract integration is defined for general measure spaces while the Riemann integral is specifically defined for functions defined on the real line and can be extended (with modifications) to higher-dimensional Euclidean spaces. However, it makes no sense on general measure spaces. Finally, key properties which can be shown for abstract integration do not generally hold for Riemann integration.

We will illustrate all these issues in the next pages. Specifically we will continue with the Dirichlet example to show that it is Lebesgue integrable and we will show that a key property of abstract integration, monotone convergence, fails for Riemann integration. Now it is time to define the abstract integral, we will follow several steps:

1. Define $\int f \, d\mu$ for simple functions (defined in the next section),
2. Define $\int f \, d\mu$ for non-negative f,
3. Define $\int f \, d\mu$ for arbitrary f.
4. Define $\int_A f \, d\mu$ for arbitrary f.

In general these steps will be key whenever we want to compute an abstract integral or prove some result which involves abstract integrals.

10.2.1 Integrating simple functions

Definition 10.3 (Simple function)**.** *A function $f : \Omega \to \mathbb{R} \cup \{\pm\infty\}$ is said to be simple if it can be written as*

$$f(\omega) = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}(\omega) \text{ for all } \omega \in \Omega,$$

where $a_i \geq 0$ for $i = 1, ..., n$ and $A_i \in \mathcal{F}$.

Example 10.3. *$(\Omega, \mathcal{F}) = (\mathbb{R}, \mathcal{B})$, one example of a simple function is $f(x) = \mathbb{1}_{[0,1)}(x) + \sqrt{2}\mathbb{1}_{\mathbb{Q}}(x)$. Another example is $g(x) = \mathbb{1}_{[0,1]}(x) + (3/2)\mathbb{1}_{[1,3]}(x)$ which looks like*
INSERT FIGURE
g can be rewritten in many different ways, for example, $g(x) = \mathbb{1}_{[0,3]}(x) + (1/2)\mathbb{1}_{[1,3]}(x)$ or $g(x) = \mathbb{1}_{[0,1]}(x) + (3/2)\mathbb{1}_{[1,2]}(x) + (3/2)\mathbb{1}_{[2,3]}(x)$.

From the example we see that there are many different representations of a simple function. Now we introduce what it called the canonical representation.

Definition 10.4 (Canonical representation)**.** *A canonical representation of a simple function is one in which $a_i \neq a_j$ and $A_i \cap A_j = \emptyset$ for all $i, j = 1, ..., n$.*

That is in the canonical representation all a_i's are distinct and all A_i's are disjoint. From now on we assume that whenever we write a simple function it is already in canonical form. This is without loss of generality since it can be shown that all simple functions have a canonical representation. Now we are ready to define the integral of simple functions.

Definition 10.5 (Integral of simple function)**.** *Let f be a simple function, then*

$$\int f \, d\mu \equiv \sum_{i=1}^{n} a_i \mu(A_i)$$

Example 10.4. *(i) Note that the indicator function $\mathbb{1}_B$ for some $B \in \mathcal{F}$ is a simple function with $n = 1$, $a_1 = 1$ and $A_1 = B$, hence*

$$\int \mathbb{1}_B \, d\mu = \mu(B),$$

that is, the integral of $\mathbb{1}_A(x)$ with respect to a measure μ is defined to be $\mu(A)$, i.e. the measure of the set that the function indicates. For example, if $\mu = \lambda$ (the Lebesgue measure) and $B = [1, 3]$, then $\int \mathbb{1}_{[1,3]} \, d\lambda = \lambda([1, 3]) = 2$, i.e. the area of the rectangle with base $[1, 3]$ and height 1. Hence, in this particular case with the Lebesgue measure, the abstract integral coincides with the Riemann integral.
(ii) Consider $f(\omega) = \mathbb{1}_{[0,1]}(\omega) + \sqrt{2}\mathbb{1}_{\mathbb{Q}}(\omega)$ and take the Lebesgue measure, then

$$\int f \, d\lambda = \lambda([0, 1]) + \sqrt{2}\overbrace{\lambda(\mathbb{Q})}^{=0} = 1.$$

(iii) Consider $f(\omega) = \mathbb{1}_{[0,1]}(\omega) + (3/2)\mathbb{1}_{[1,3]}(\omega)$ *and take the Lebesgue measure, then*

$$\int f \, d\lambda = \lambda([0,1]) + (3/2)\lambda([1,3]) = 4.$$

Again, in these examples we get the area below f *so it coincides with the Riemann integral, however if* $\mu \neq \lambda$ *then we would not get the area and it would not coincide with the Riemann integral.*

Example 10.5 (Dirichlet function)**.** *Let us come back to the example of the Dirichlet function on* $[0,1]$ *which we have already shown not to be Riemann integrable. The measure space is* $(\Omega, \mathcal{F}, \mu) = ([0,1], \mathcal{B}([0,1]), \lambda)$*. The Dirichlet function can be written as a simple function:* $f(\omega) = \mathbb{1}_{\mathbb{Q}\cap[0,1]}(\omega)$*, hence*

$$\int f \, d\lambda = \int \mathbb{1}_{\mathbb{Q}\cap[0,1]} \, d\lambda = \lambda(\mathbb{Q} \cap [0,1]) = 0.$$

So we see that in this case the Lebesgue integral exists while the Riemann integral does not.

10.2.2 Integrating non-negative measurable functions

Let $g : \Omega \to \mathbb{R} \cup \{\pm\infty\}$ be a non-negative measurable function and let $S(g)$ be the set of all simple functions $q : \Omega \to \mathbb{R} \cup \{\pm\infty\}$ such that $g(\omega) \geq q(\omega)$ for all $\omega \in \Omega$.

Definition 10.6 (Integral of non-negative measurable function)**.**

$$\int g \, d\mu \equiv \sup_{q \in S(g)} \int q \, d\mu.$$

This definition is not very useful for practical purposes (i.e. computing integrals) but is very useful for proving properties about integrals. We will see a more practical definition which can be used to compute integrals and gives some intuition about the above definition in a bit.

10.2.3 Integral for arbitrary measurable functions

Let $f : \Omega \to \mathbb{R} \cup \{\pm\infty\}$ be an arbitrary measurable function. Write

$$f = f_+ - f_-, \text{ where } f_+(\omega) \equiv \max\{f(\omega), 0\} \text{ and } f_-(\omega) \equiv max\{-f(\omega), 0\}.$$

This is a decomposition of the function into its positive and negative parts (convince yourself that the equality holds). Note that f_+ and f_- are both general non-negative functions which we know how to integrate.

Definition 10.7 (Integral of arbitrary measurable function)**.**

$$\int f \, d\mu \equiv \int f_+ \, d\mu - \int f_- \, d\mu.$$

This is well-defined (we say it exists) as long as at least one of the integrals is finite and undefined (does not exist) otherwise. If both are finite we say f is integrable.

10.2.4 Integral over a measurable set

Let $A \in \mathcal{F}$, and f be an arbitrary measurable function, then

$$\int_A f \, d\mu \equiv \int f \mathbb{1}_A \, d\mu.$$

10.3 Properties

Now we introduce useful properties of abstract integrals.

Proposition 10.1 (Properties of abstract integrals)**.**

(i) If $A \in \mathcal{F}$, then $\int \mathbb{1}_A \, d\mu = \mu(A)$. This property is a corollary of the definition of the integral of simple functions and we have already shown it as an example.

(ii) If $g \geq 0$ and measurable, then $\int g \, d\mu \geq 0$.

Proof. By the definition of the integral of non-negative measurable functions

$$\int g \, d\mu = \sup_{q \in S(g)} \int q \, d\mu,$$

note that the integral of a simple function $q = \sum_{i=1}^n a_i \mathbb{1}_{A_i}$ is $\sum_{i=1}^n a_i \mu(A_i)$. Since $a_i \geq 0$ for $i = 1, ..., n$ and measures are positive we know that the integral of simple functions is always positive. Now the set $\{\int q \, d\mu : q \in S(g)\}$ is a set of positive numbers and hence its supremum is positive. Note that there might not be a q in $S(g)$ which attains the supremum, but still the supremum of a set of positive numbers is positive. □

(iii) If $g = 0$ μ a.e., then $\int g \, d\mu = 0$.

Proof. Remember that $S(g)$ is the set of simple functions q such that $q \leq g$. Hence, in this case $S(g)$ is formed by simple functions of the form $\sum_{i=1}^n a_i \mathbb{1}_{A_i} \leq g$ where $\mu(A_i) = 0$ for all $i = 1, ..., n$, i.e. by simple functions which are 0 a.e. and less than or equal to g in sets of measure 0. Hence, for all $q \in S(g)$, we have that $\int q \, d\mu = 0$ and hence $\int g \, d\mu = 0$. □

(iv) Let $h \geq g \geq 0$, then $\int h \, d\mu \geq \int g \, d\mu$.

Proof. Let us prove it first assuming h and g are non-negative. Since $S(h)$ and $S(g)$ are the sets of simple functions bounding h and g, respectively, from below and $h \geq g$, it follows that $S(g) \subseteq S(h)$. Hence

$$\int h \, d\mu = \sup_{q \in S(h)} \int q \, d\mu \geq \sup_{q \in S(g)} \int q \, d\mu = \int g \, d\mu.$$

For arbitrary measurable functions notice that $h_+ \geq g_+$ and $h_- \leq g_-$, hence $S(g_+) \subseteq S(h_+)$ and $S(h_-) \subseteq S(g_-)$ and therefore

$$\begin{aligned}\int h\,d\mu &= \int h_+\,d\mu - \int h_-\,d\mu \\ &= \sup_{q\in S(h_+)} \int q\,d\mu - \sup_{q\in S(h_-)} \int q\,d\mu \\ &\geq \sup_{q\in S(g_+)} \int q\,d\mu - \sup_{q\in S(g_-)} \int q\,d\mu \\ &= \int g\,d\mu.\end{aligned}$$

□

(v) If $g = h$ μ a.e., then $\int h\,d\mu = \int g\,d\mu$.

Proof. **Exercise** □

(vi) If $g \geq 0$ and $\int g\,d\mu = 0$, then $g = 0$ μ a.e.

Proof. Suppose not and define $B \equiv \{\omega \in \omega : g(\omega) > 0\}$ and assume $\mu(B) > 0$ (if not the result follows directly). Let $B_n \equiv \{\omega \in \Omega : g(\omega) > 1/n\}$. It follows that $B_n \subseteq B_{n+1}$ and that $\bigcup_{i=1}^{\infty} B_i = B$ (try to prove this to refresh how to prove set equalities). By continuity of measures $\mu(B) = \lim_{n\to\infty} \mu(B_n)$, this equality states that the limit of a sequence of measures is $\mu(B) > 0$, hence there exists $n_0 \in \mathbb{N}$ such that for all $k \geq n_0$ $\mu(B_k) > 0$, take one such k and note that $(1/k)\mathbb{1}_{B_k}$ is a simple function which is lower or equal than g and hence it belongs to $S(g)$. Hence,

$$\int g\,d\mu = \sup_{q\in S(g)} \int q\,d\mu \geq \int \frac{1}{k}\mathbb{1}_{B_k} = \frac{1}{k}\mu(B_k) > 0,$$

which contradicts $\int g\,d\mu = 0$. □

(vii) (Linearity) $\int (g + h)\,d\mu = \int g\,d\mu + \int h\,d\mu$.
This can be proven by first proving it for simple functions, then for non-negative measurable functions and then for arbitrary measurable functions. However, the way we have defined the integral for non-negative measurable functions makes this very hard. We will prove this property later using the result in the next section.

(viii) (Scaling) Let $a \geq 0$, then $\int ag\,d\mu = a \int g\,d\mu$.

Proof. **Exercise.** □

10.4 The Monotone Convergence Theorem

The monotone convergence theorem (MCT) is a cornerstone result of integration and one of the main reasons to develop abstract integration. In fact, it was with the start of Fourier series that a result such as the MCT was frequently needed, however finding such a result for the Riemann integral turned out to be very hard. We will show this result for abstract integrals and show it does not hold for the Riemann integral. The MCT gives us conditions under which we can exchange limits and integrals, i.e. conditions under which the limit of the integral is the same as the integral of the limit. Let us first introduce pointwise convergence and motivate the need for this result. Let $f_n : \Omega \to \mathbb{R} \cup \{\pm\infty\}$ be a sequence of measurable functions.

Definition 10.8.

(i) We say f_n converges to f pointwise if for all $\omega \in \Omega$,

$$\lim_{n\to\infty} f_n(\omega) = f(\omega),$$

(ii) we say f_n converges to f μ a.e. if $f_n(\omega) \to f$ for all $\omega \in \Omega$ except, perhaps, on a set of μ measure zero.

Regarding (i) it can be shown that if f_n is measurable for all n, then f is measurable. Note that (ii) is weaker than (i) since it allows f_n to not converge to f in measure zero sets.

The question we ask ourselves is whether we can generally interchange limits and integrals. That is, does $\int f_n \, d\mu \to \int f \, d\mu$, i.e. $\lim_{n\to\infty} \int f_n \, d\mu = \int f \, d\mu$. The answer is that not generally as illustrated by the following counterexample.

Example 10.6. *Consider $(\Omega, \mathcal{F}, \mu) = ([0,1], \mathcal{B}, \lambda)$ and let*

$$f_n(\omega) = \begin{cases} n & \text{if } 0 < \omega \leq \frac{1}{n} \\ 0 & \text{otherwise.} \end{cases}$$

noticing that f_n is a simple function we can see that

$$\int f_n \, d\lambda = n\lambda((0, 1/n]) + 0 = n\frac{1}{n} = 1.$$

however, $f(\omega) = \lim_{n\to\infty} f_n(\omega) = 0$ for all $\omega \in \Omega$ and hence $\int f \, d\lambda = 0$.

Try to draw the function in the example for different n to get an idea about why this function is problematic. The conclusion is that it is not true in general that we can exchange limits and integrals. Now we introduce the MCT to give conditions under which this is possible.

Theorem 10.2 (Monotone Convergence Theorem (MCT))**.** *Let $g_n \geq 0$ be a sequence of measurable functions such that $g_n \nearrow g$ μ a.e., i.e. $g_n(\omega) \leq g_{n+1}(\omega)$ for all $n = 1, 2, \ldots$ μ a.e. and $\lim_{n\to\infty} g_n(\omega) = g(\omega)$ μ a.e. Then*

$$\lim_{n\to\infty} \int g_n \, d\mu = \int g \, d\mu.$$

So we do not only require that g_n converges to g but that it converges monotonically.

Proof. The proof is quite long. You can find it step by step in the MCT Wikipedia in the section Beppo Levi's Lemma. It can be made shorter by using Fatou's lemma which we will introduce shortly. □

Exercise 10.1. *Copy the proof of the MCT given in Wikipedia (without invoking Fatou's lemma) and understand all steps.*

Exercise 10.2. *Check why the last example does not satisfy the condition of the MCT.*

The reason why this result does not hold for the Riemann integral is that the limit of a sequence of Riemann integrable functions need not be Riemann integrable as the following example illustrates.

Example 10.7. *(Failure of MCT for Riemann integral) Let $r_0, r_1, r_2, ...$ be the enumeration of all the rationals in $[0,1]$ (they are countable). Then, consider the following function on $[0,1]$*

$$g_n(\omega) = \begin{cases} 1 & \text{if } \omega \in \{r_0, r_1, ..., r_{n-1}\} \\ 0 & \text{otherwise} \end{cases}$$

Since g_n is discontinuous only at finitely many points, the set of discontinuities has Lebesgue measure zero, so by the Lebesgue criterion, g_n is Riemann integrable ($g_n is also bounded$). However, the limit g is the Dirichlet function which we have already shown to not be Riemann integrable.

10.4.1 Proof of linearity property using MCT

Remember that we did not prove the linearity property of Lebesgue integrals, this was because the proof is made much easier if the MCT is invoked.

Proof. First assume that g and h are simple (canonical) functions, then

$$g(\omega) = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}(\omega), \quad h(\omega) = \sum_{i=1}^{m} b_i \mathbb{1}_{B_i}(\omega), \quad g(\omega) + h(\omega) = \sum_{i=1}^{n}\sum_{j=1}^{m} (a_i + b_j)\mathbb{1}_{A_i \cap B_j}(\omega).$$

The sum above follows since $A_1, ..., A_n$ and $B_1, ..., B_n$ are partitions, hence any ω will lie in $A_i \cap B_j$ for some i and j and $g(\omega) + h(\omega)$ will just be the sum of a_i and b_j. Since A_i's and B_j's are disjoint, $A_i \cap B_j$ are disjoint across (i, j). By definition of the Lebesgue integral

$$\begin{aligned} \int (g + h)\, d\mu &= \sum_{i=1}^{n}\sum_{j=1}^{m} (a_i + b_j)\mu(A_i \cap B_j) \\ &= \sum_{i=1}^{n} a_i \sum_{j=1}^{m} \mu(A_i \cap B_j) + \sum_{j=1}^{m} b_j \sum_{i=1}^{n} \mu(A_i \cap B_j) \\ &= \sum_{i=1}^{n} a_i \mu(A_i) + \sum_{j=1}^{m} b_j \mu(B_j) \\ &= \int g\, d\mu + \int h\, d\mu. \end{aligned}$$

In the third equality we use finite additivity, i.e. $\sum_{j=1}^{m} \mu(A_i \cap B_j) = \mu(\cup_{j=1}^{m}(A_i \cap B_j)) = \mu(A_i \cap (\cup_{j=1}^{m} B_j)) = \mu(A_i)$.

Now, for non-negative measurable functions define let g_n and h_n, $n \geq 1$, be simple functions such that $g_n \nearrow g$ and $h_n \nearrow h$ μ a.e. (we will show in the next section that we can always find such simple functions). Then, $g_n + h_n \nearrow g + h$ μ a.e. and

$$\begin{aligned}\int (g+h)\, d\mu &= \lim_{n\to\infty} \int (g_n + h_n)\, d\mu \\ &= \lim_{n\to\infty} \int g_n\, d\mu + \lim_{n\to\infty} \int h_n\, d\mu \\ &= \int g\, d\mu + \int h\, d\mu,\end{aligned}$$

where the first and third equalities follow from the MCT and the second one from the proof we just did for simple functions. The result for arbitrary measurable functions follows since we can write them as sums of non-negative measurable functions (some of them scaled by -1). □

10.5 Approximating a non-negative measurable function from below using simple functions (practical definition)

Let g be a non-negative measurable function. The goal of this section is to show that we can always find a sequence of simple function such that $g_n \nearrow g$ μ a.e. We will do this by giving a specific construction of g_n which works for any non-negative function. There are plenty of ways of constructing these sequences of simple functions but providing one which works in all cases is enough. This sequence is

$$g_n(\omega) = \begin{cases} \frac{i}{2^n} & \text{if } \frac{i}{2^n} \leq g(\omega) < \frac{i+1}{2^n},\ i = 0, 1, ..., n2^n - 1, \\ n & \text{if } g(\omega) \geq n, \end{cases}$$

note that is not a two-piece function, but that there are much more pieces and the above is just a short way to write it. It is a simple function which can be written as

$$g_n(\omega) = \sum_{i=0}^{n2^n-1} \frac{i}{2^n} \mathbb{1}_{\{\omega: i/2^n \leq g(\omega) < (i+1)/2^n\}}(\omega) + n\mathbb{1}_{\{\omega: g(\omega) \geq n\}}(\omega).$$

What g_n is doing is to partition the vertical axis more and more as n increases. For instance, if $n = 1$ we have $i = 0, 1/2, 1$, so $g_n(\omega)$ will take value 0 whenever $0 \leq g(\omega) < 1/2$, value $1/2$ whenever $1/2 \leq g(\omega) < 1$ and value 1 whenever $g(\omega) > 1$. If $n = 2$ we have that $i = 0, 1/4, 2/4, 3/4, 1, 5/4, ..., 2$ (think about which values will g_n take).

Exercise 10.3. *Draw an arbitrary non-negative function and in the same graph draw g_n for $n = 1$ and $n = 2$. Then do the same for a general n.*

Exercise 10.4. *Show the following*

1. *For all n, g_n is a simple function,*

2. $g_n(\omega) \leq g_{n+1}(\omega)$ *for all* n *and all* ω,

3. $\lim_{n\to\infty} g_n(\omega) = g(\omega)$.

The above exercise implies that we can invoke the MCT, that is, $\int g\, d\mu = \lim_{n\to\infty} \int g_n\, d\mu$ for any non-negative measurable function g. Since we have the simple functions g_n and we know how to integrate simple functions, we have a very practical way to evaluate Lebesgue integrals which does not involve searching across all possible simple function and finding the sup. We can construct the abstract integral of a non-negative measurable function g as

$$\int g\, d\mu = \lim_{n\to\infty} \sum_{i=1}^{n2^n-1} \frac{i}{2n} \mu\left(\left\{\omega : \frac{i}{2^n} \leq g(\omega) < \frac{i+1}{2^n}\right\}\right) + n\mu(\{\omega : g(\omega) \geq n\}).$$

Using this approximation via simple functions we can already compute several integrals as in the following examples.

Example 10.8 (Integral with respect to the Dirac measure)**.** *If* f *is a simple function, by definition of the Lebesgue integral for simple functions, its integral with respect to the Dirac measure at some point* x *(*δ_x*) is*

$$\int f\, d\delta_x = \sum_{i=1}^{n} a_i \int \mathbb{1}_{A_i}\, d\delta_x = \sum_{i=1}^{n} a_i \delta_x(A_i) = \sum_{i=1}^{n} a_i \mathbb{1}_{A_i}(x) = f(x).$$

Hence, the integral of any simple function with respect to the Dirac measure at a point is the function itself evaluated at that point. If f *is a general non-negative function, we can find a sequence of simple functions* f_n *which approximates* f *from below and then*

$$\int f\, d\delta_x = \int \lim_{n\to\infty} f_n\, d\delta_x = \lim_{n\to\infty} \int f_n\, d\delta_x = \lim_{n\to\infty} f_n(x) = f(x),$$

where the second equality uses the MCT, the third the fact that the integral of a simple function with respect to the Dirac measure is the simple function itself and the first and last equalities use that f_n *converges to* f.

(Excercise) Show that for an arbitrary measurable function f

$$\int f\, d\delta_x = f(x).$$

Hence, in general, the integral of any measurable function f *with respect to* δ_x *is* $f(x)$.

Example 10.9 (Integral with respect to the counting measure)**.** *Consider some measurable function* f *on* $\Omega = \{a_1, ..., a_n\}$ *and the counting measure* $\nu(A) = \sum_{i=1}^{n} \delta_{a_i}$*, then*

$$\int f\, d\nu = \int f d \sum_{i=1}^{n} \delta_{a_i} = \sum_{i=1}^{n} \int f\, d\delta_{a_i} = \sum_{i=1}^{n} f(a_i),$$

the second equality follows because linear combinations of measures are measures (Exercise), because for any two measures μ_1 *and* μ_2*,* $ntf\, d(\mu_1 + \mu_2) = \int f\, d\mu_1 + \int f\, d\mu_2$ *(Exercise) and by linearity of integrals. The last equality follows from the previous example.*

10.6 Fatou's Lemma

Let g and h be measurable functions. Then it holds that

$$\int \min(g,h)\,d\mu \leq \min\Big(\int g\,d\mu, \int h\,d\mu\Big),$$

since $\min(g(\omega),h(\omega)) \leq g(\omega)$ and $\min(g(\omega),h(\omega)) \leq h(\omega)$ and hence by property (iv) of abstract integrals the result above holds. This extends to n measurable functions. Fatou's lemma extends it to a sequence of measurable functions, if you have infinite measurable functions the minimum might not be attained and hence we speak about the infimum, this is what Fatou's lemma does.

Lemma 10.1. *(Fatou's Lemma)*

(i) Let g_n, $n \geq 1$, be a sequence of measurable functions such that $g_n \geq h$ for all n and $\int |h|\,d\mu < \infty$. Then,

$$\int \liminf_{n\to\infty} g_n\,d\mu \leq \liminf_{n\to\infty} \int g_n\,d\mu,$$

(ii) Let g_n, $n \geq 1$, be a sequence of measurable functions such that $g_n \leq h$ for all n and $\int |h|\,d\mu < \infty$. Then,

$$\int \limsup_{n\to\infty} g_n\,d\mu \geq \limsup_{n\to\infty} \int g_n\,d\mu.$$

Note that (ii) is the same as (i) if you replace g_n with $-g_n$ and h with $-h$. Also, lets recall briefly what lim inf is. For a sequence a_n we have

$$\liminf_{n\to\infty} a_n \equiv \lim_{n\to\infty} \inf_{m\geq n} a_m.$$

That is, first fix some n, then look at the infimum in the set $\{a_n, a_{n+1}, ...\}$, for each n we have one such infimum. In sum, we have a sequence of infimums $(\inf_{n\geq 1} a_n, \inf_{n\geq 2} a_n, ...)$ of which we are taking the limit. Note that the sequence of infimums is a non-decreasing sequence of infimums and hence its limit always exists. You can interpret this as the smallest limit point. Finally, one can show that $\liminf_{n\to\infty} g_n$, $\limsup_{n\to\infty} g_n$ and $\lim_{n\to\infty} g_n$ are measurable functions. Now we can prove Fatou's lemma

Proof of Fatou's lemma. It is enough to show (i). Fix n, then we have that

$$\inf_{k\geq n} g_k - h \leq g_m - h \text{ for all } m \geq n,$$

using property (iv)

$$\int (\inf_{k\geq n} g_k - h)\,d\mu \leq \int (g_m - h)\,d\mu \text{ for all } m \geq n,$$

take the infimum in both sides (nothing happens in LHS)

$$\int (\inf_{k\geq n} g_k - h)\,d\mu \leq \inf_{m\geq n} \int (g_m - h)\,d\mu,$$

now take limits in both sides (RHS becomes lim inf by definition)

$$\lim_{n\to\infty}\int(\inf_{k\geq n} g_k - h)\,d\mu \leq \liminf_{n\to\infty}\int(g_n - h)\,d\mu.$$

Now we want to put the limit inside in the LHS. Let $z_n = \inf_{k\geq n} g_k - h$. z_n is non-decreasing since we are taking the infimum over smaller and smaller sets. Also, $z_n \geq 0$ since $g_n \geq h$. Also, $z = \lim_{n\to\infty} z_n = \liminf_{n\to\infty} g_n - h$ by definition of lim inf. So by the MCT $\lim_{n\to\infty}\int z_n\,d\mu = \int z\,d\mu$. Hence,

$$\begin{aligned}\lim_{n\to\infty}\int(\inf_{k\geq n} g_k - h)\,d\mu &= \int(\liminf_{n\to\infty} g_n - h)\,d\mu \\ &\leq \liminf_{n\to\infty}\int(g_n - h)\,d\mu.\end{aligned}$$

By linearity of integrals we have

$$\int \liminf_{n\to\infty} g_n\,d\mu - \int h\,d\mu \leq \liminf_{n\to\infty}\int g_n\,d\mu - \int h\,d\mu.$$

So,

$$\int \liminf_{n\to\infty} g_n\,d\mu \leq \liminf_{n\to\infty}\int g_n\,d\mu.$$

□

10.7 Dominated Convergence Theorem

The Dominated Convergence Theorem (DCT) provides other conditions under which we can exchange limit and integral.

Proposition 10.2 (DCT)**.** *Consider a sequence of measurable functions g_n $n \geq 1$, such that $\lim_{n\to\infty} g_n(\omega) = g(\omega)$ for all ω. Suppose there exists a measurable function h such that $|g_n| \leq h$ μ a.e. and $\int |h|\,d\mu < \infty$. Then*

$$\lim_{n\to\infty}\int g_n\,d\mu = \int g\,d\mu.$$

Proof. Since $-h \leq g_n \leq h$ for all n, we can invoke both sides of Fatou's lemma. Since also $g_n \to g$, $\liminf_{n\to\infty} g_n = \limsup_{n\to\infty} g_n = \lim_{n\to\infty} g_n$ (i.e. limit is one point not a set), we have that

$$\begin{aligned}\int g\,d\mu &= \int \liminf_{n\to\infty} g_n\,d\mu \\ &\leq \liminf_{n\to\infty}\int g_n\,d\mu \\ &\leq \limsup_{n\to\infty}\int g_n\,d\mu \\ &\leq \int \limsup_{n\to\infty} g_n\,d\mu \\ &= \int g\,d\mu.\end{aligned}$$

It might seem we have proven something trivial, namely that $\int g\,d\mu = \int g\,d\mu$, but what we have actually proven is that all the weak inequalities above must hold with equality. Hence,

$$\liminf_{n\to\infty} \int g_n\,d\mu = \limsup_{n\to\infty} \int g_n\,d\mu = \int g\,d\mu,$$

and hence

$$\lim_{n\to\infty} \int g_n\,d\mu = \int g\,d\mu.$$

□

Exercise 10.5. *Check why example 10.6 does not satisfy the conditions of the DCT.*

10.7.1 Exchanging derivative and integral

The DCT gives us conditions under which we can exchange derivatives and integrals as well. This is because derivatives are limits.

Proposition 10.3. *Take $(\Omega, \mathcal{F}, \mu) = (\mathbb{R}, \mathcal{B}, \lambda)$. Let f be measurable and suppose $\frac{d}{dx}f(x)$ exists a.e. and that $\left|\frac{d}{dx}f(x)\right| \leq g(x)$ a.e. where g is integrable. Then,*

$$\frac{d}{dx}\int f(x)\,d\lambda = \int \frac{d}{dx}f(x)\,d\lambda.$$

Proof. First note that by the definition of a derivative and by linearity of integrals

$$\frac{d}{dx}\int f\,d\lambda = \lim_{t\to 0}\frac{\int f(x+t)\,d\lambda - \int f(x)\,d\lambda}{t} = \lim_{t\to 0}\frac{\int [f(x+t)-f(x)]\,d\lambda}{t}.$$

Now define a sequence $t(n)$ such that $t(n)\to 0$ as $n\to\infty$. Define also

$$f_n = \frac{f(x+t(n))-f(x)}{t(n)}.$$

Then, we want to show that

$$\lim_{n\to\infty}\int f_n\,d\lambda = \int \lim_{n\to\infty} f_n\,d\lambda. \tag{10.1}$$

To do this we invoke the DCT. We need (i) $f_n \to \frac{d}{dx}f(x)$ a.e. and that f_n is bounded by some integrable function for all n. (i) follows easily since by the definition of a derivative $\lim_{n\to\infty} f_n = \frac{d}{dx}f(x)$. (ii) follows by the Mean Value Theorem:

$$f_n(x) = \frac{f(x+t(n))-f(x)}{x+t(n)-x} = \frac{d}{dx}f(c) \text{ for some } c\in[x, x+t(n)],$$

since $\left|\frac{d}{dx}\Big|_{x=c} f(x)\right| \leq g(c)$ with g integrable we have that f_n is bounded by an integrable function. Hence, the condition of the DCT apply and (10.1) holds, meaning that

$$\frac{d}{dx}\int f(x)\,d\lambda = \int \frac{d}{dx}f(x)\,d\lambda.$$

□

10.8 Product measure and Fubini Theorem

Before we introduce Fubini's theorem we need to remind ourselves what a σ-finite measure is and introduce and important result. μ (defined on some space $(\Omega, \mathcal{F})$ is a σ-finite measure if there exists a sequence $A_1, A_2, ...$ of subsets of Ω such that $\cup A_i = \Omega$ and $\mu(A_i) < \infty$ for all i, then μ is said to be a σ-finite measure.

10.8.1 Product measures

In this subsection we motivate the product measure theorem we are going to present in a bit. For a more formal treatment you can see PUT REFERENCE. Suppose that we have several measurable spaces $(\Omega_i, \mathcal{F}_i, \mu_i)$, $i = 1, ..., k$ with k some integer. Suppose we have measurable (with respect their corresponding σ-algebra) sets $A_1 \in \mathcal{F}_1$, $A_2 \in \mathcal{F}_2$, ..., $A_k \in \mathcal{F}_k$. The Cartesian product $A_1 \times ... \times A_k$ is called a measurable rectangle. Our ultimate goal is to construct a measure on what is called the product σ-algebra.

Definition 10.9 (Product σ-algebra)**.** *Suppose that $(\Omega_i, \mathcal{F}_i, \mu_i)$, $i = 1, ..., k$ with k some integer, are measurable spaces. The product σ-algebra $\bigotimes_{i=1}^n \mathcal{F}_i$ is the σ-algebra generated by the collection of all measurable rectangles,*

$$\bigotimes_{i=1}^{n} \mathcal{F}_i = \sigma(\{A_1 \times ... \times A_n : A_1 \in \mathcal{F}_1, ..., A_k \in \mathcal{F}_k\}.$$

The product of $(\Omega_1, \mathcal{F}_1)$, $(\Omega_2, \mathcal{F}_2)$, ... is the measurable space $(\Omega_1 \times ... \times \Omega_k, \bigotimes_{i=1}^n \mathcal{F}_i)$.

Note that the Cartesian product of σ-algebras is not necessarily a σ-algebra, hence the need of the product σ-algebra.

Exercise 10.6. **Show that the collection of finite unions of measurable rectangles forms an algebra (Hint: show that the intersection of measurable rectangles is a measurable rectangle and that the complement of a measurable rectangle is a finite union of measurable rectangles).*

*The exercise above shows that the collection of finite unions of measurable rectangles forms an algebra which we denote as $\mathcal{R}_0$. Now we define a product pre-measure on a given measurable rectangle and then on $\mathcal{R}_0$.

Definition 10.10 (Product pre-measure on measurable rectangles*)**.** *If $(\Omega_i, \mathcal{F}_i, \mu_i)$, $i = 1, ..., k$ with k some integer, are measure spaces, then the product pre-measure $\nu_0(A_1 \times ... \times A_k)$ of a measurable rectangle $A_1 \times ... \times A_k \in \mathcal{F}_1 \times ... \times \mathcal{F}_k$ is*

$$\nu_0(A_1 \times ... \times A_k) = \mu_1(A_1)\mu_2(A_2)...\mu_k(A_k),$$

where $0 \cdot \infty = 0$.

Exercise 10.7. **Show that the pre-measure μ_0 is countably additive in rectangles.*

*Note the above concept is defined only on measurable rectangles but not on the algebra generated my measurable rectangles.

Definition 10.11 (Product pre-measure on algebra*)**.** *If $(\Omega_i, \mathcal{F}_i, \mu_i)$, $i = 1, ..., k$ with k some integer, are measure spaces and $\mathcal{R}_0$ is the algebra generated by measurable rectangles. Then the product pre-measure $\nu_0 : \mathcal{R}_0 \to [0, \infty]$ is given by*

$$\nu_0(R) = \sum_{i=1}^{n} \mu_1(A_{1,i})\mu_2(A_{2,i})...\mu_k(A_{k,i}), \quad R = \bigcup_{i=1}^{n} A_{1,i} \times ... \times A_{k,i},$$

where $R = \cup_{i=1}^{n} A_{1,i} \times ... \times A_{k,i}$ is any representation of $R \in \mathcal{R}_0$ as a disjoint union of measurable rectangles.

Exercise 10.8. **Is ν_0 countably additive on $\mathcal{R}_0$?*

*So we have a pre-measure on an algebra, this should remind us to the Caratheodory extension theorem, in fact what the next theorem does is to make use of this theorem to extend the product pre-measure to a product measure. We will not prove it here.

If you have not followed the concepts indicated with * you can take the next results as granted.

Theorem 10.3 (Product measure theorem)**.** *If $(\Omega_i, \mathcal{F}_i, \mu_i)$, $i = 1, ..., k$ with k some integer, are measure spaces, then*

$$\bigotimes_{i=1}^{n} \mu_i(A_1 \times ... \times A_k) : \bigotimes_{i=1}^{n} \mathcal{F}_i \to [0, \infty],$$

is a measure on $\Omega_1 \times ... \times \Omega_k$ such that

$$\bigotimes_{i=1}^{n} \mu_i(A_1 \times ... \times A_k) = \mu_1(A_1)\mu_2(A_2)...\mu_k(A_k), \text{ for all } A_j \in \mathcal{F}_j \quad j = 1, ..., k.$$

If $\mu_1, ..., \mu_k$ are σ-finite measures, $\bigotimes_{i=1}^{n} \mu_i$ is the unique measure on $\bigotimes_{i=1}^{n} \mathcal{F}_i$ with this property.

The Fubini theorem we introduce now tells us when is it possible to solve an integral with respect to the product measure with iterated integrals, i.e. solving the integrals with respect to each measure in an iterative manner. We state it for two measure spaces but it can be naturally extended to more measure spaces. We omit the proof.

Theorem 10.4 (Fubini theorem)**.** *Let μ_i be a σ-finite measure on $(\Omega_i, \mathcal{F}_i)$ for $i = 1, 2$, and f be a measurable function on $\Omega_1 \times \Omega_2$ such that either $f \geq 0$ or f is integrable with respect to $\mu_1 \bigotimes \mu_2$, then*

$$g(\omega_2) = \int_{\Omega_1} f(\omega_1, \omega_2)\, d\mu_1,$$

exists μ_2 a.e. and defines a measurable function on Ω_2 whose integral w.r.t. μ_2 exists, and

$$\int_{\Omega_1 \times \Omega_2} f(\omega_1, \omega_2)\ d(\mu_1 \otimes \mu_2) = \int_{\Omega_2} \left[\int_{\Omega_1} f(\omega_1, \omega_2)\, d\mu_1 \right] d\mu_2.$$

Example 10.10. *Let $\Omega_1 = \Omega_2 = \{0, 1, 2, ...\}$, $\mu_1 = \mu_2 = \nu$ where $\nu(A) = \sum_{i=1}^{\infty} \delta_{a_i}(A)$ is the counting measure. Let f be a function on $\Omega_1 \times \Omega_2$, i.e. a double sequence $f(i, j)$,*

$i, j = 0, 1, 2, ...$ *If* $f \geq 0$ *or* $\int |f| \, d(\mu_1 \otimes \mu_2) < \infty$*, then*

$$\int f \, d(\mu_1 \otimes \mu_2) = \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} f(i,j) = \sum_{j=0}^{\infty} \sum_{i=0}^{\infty} f(i,j),$$

by Fubini's theorem and Example 10.9.

11 Differentiation*

This section presents the main results in Chapter 9 of Kolmogorov and Fomin (1975) without proof. It amounts to a first superficial reading of that chapter, the interested reader can go deeper by looking in the book. In the previous section we have introduced measurable function and generalized the concept of integration. As we know from the Fundamental Theorem of Calculus, there is a connection between derivatives and integrals in the real line (with the Lebesgue measure). Hence, a question to ask at this point is whether there exists such a connection on a more general level when we deal with abstract integrals. To do this our main working object is going to be the set function

$$\int_A f(x) \, d\mu, \tag{11.1}$$

based on a measurable function f on some space Ω. This set function exists for any measurable set A and hence defines a set function on a σ-algebra of Ω. Before analyzing the problem for general measure spaces we are going to spend some time in the special case of the real line equipped with the Lebesgue measure.

11.1 Real line with the Lebesgue measure

Consider $(\Omega, \mathcal{F}) = (\mathbb{R}, \lambda)$, where λ is the Lebesgue measure. Letting $A = [a, b]$ be some closed interval, (11.1) becomes

$$\int_a^b f(t) \, dt.$$

For our purpose we will want to fix the lower limit and consider the integral above as a function of an upper limit denoted by x

$$\int_a^x f(t) \, dt.$$

Now, the Fundamental Theorem of Calculus (FTC) tells us that if f is continuous, then

$$\frac{d}{dx} \int_a^x f(t) \, dt = f(x), \tag{11.2}$$

i.e. $\int_a^x f(t) \, dt$ is an antiderivative of f in the sense that when you derivate it you get f (the derivative of the antiderivative is the function). A corollary of the FTC is that if a function

$F(x)$ is an antiderivative of f, then

$$\int_a^x f(t)\,dt = F(x) - F(a).$$

We can state this equality just in terms of the antiderivative since $F'(x) \equiv d/dxF(x) = f(x)$, so that if we have some function F with a continuous derivative it holds that

$$\int_a^x F'(t)\,dt = F(x) - F(a). \tag{11.3}$$

So from the FTC we know that (11.2) holds when f is continuous and that (11.3) holds when F has a continuous derivative. Our first goal in this section is to show that (11.2) holds for arbitrary Lebesgue integrable functions. The second goal is to show that (11.3 holds for a larger class of functions than that of function with continuous derivatives. For the first goal the following result is key

Theorem 11.1 (Lebesgue)**.** *A monotonic*[13] *function f defined on an interval $[a, b]$ has a finite derivative almost everywhere on $[a, b]$.*

The proof of this theorem can be found in Kolmogorov and Fomin (1975) and uses several lemmas about monotonic functions. This theorem is important to us because, as we have done in the integration chapter, we can always decompose an integrable function f into its positive and negative components

$$f(t) = f_+(t) - f_-(t),$$

where f_+ and f_- are nonnegative functions. Now, for any nonnegative function g,

$$G(x) = \int_a^x g(t)\,dt,$$

is a nondecreasing function (and hence monotonic) since the larger x is, the larger the region over which we integrate a positive function. Using this observation an invoking Theorem 11.1 the following theorem follows

Theorem 11.2. *Let f be an integrable function on $[a, b]$. Then*

$$\frac{d}{dx}\int_a^x f(t)\,dt,$$

exists and is finite.

Hence we only miss to show it is equal to $f(x)$, the proof of this result is more involved but can be found in Kolmogorov and Fomin (1975), here we state it without proof.

Theorem 11.3. *Let f be any integrable function on $[a, b]$. Then*

$$\frac{d}{dx}\int_a^x f(t)\,dt = f(x) \ \ a.e.$$

[13] A function f is nondecreasing if $x_1 \leq x_2$ implies $f(x_1) \leq f(x_2)$ and nonincreasing if $x_1 \leq x_2$ implies $f(x_1) \geq f(x_2)$. A function is monotonic if it is either nondecreasing or nonincreasing.

So we we have attained the first goal of this section (in fact we have pretty much just presented the result and stated the the proofs behind have a lot to do with the properties of monotonic functions, again, the proofs are in Kolmogorov and Fomin (1975)).

For our second goal we will need the concept of functions of bounded variation and a connection with monotonic functions. Let us first define bounded variation

Definition 11.1 (Bounded variation)**.** *A function f on an interval $[a, b]$ is said to be of bounded variation if there is a constant $C > 0$ such that*

$$\sum_{k=1}^{n} |f(x_k) - f(x_{k-1})| \leq C,$$

for every partition

$$a = x_0 < x_1 < ... < x_n = b,$$

of $[a, b]$ by points of subdivision $x_0, x_1, ..., x_n$.

Note that monotonic functions are functions of bounded variation since the sum in the definition above will always be equal to $|f(b) - f(a)|$. Also, linear combinations of functions of bounded variation are also of bounded variation. The next theorem relates functions of bounded variation with nondecreasing functions.

Theorem 11.4. *If f is of bounded variation on $[a, b]$, then f can be represented as the difference between two nondecreasing functions on $[a, b]$.*

This theorem has two important corollaries, one is that every function of bounded variation has a finite derivative almost everywhere (by Theorem 11.1) and if f is integrable on $[a, b]$, then the indefinite integral

$$\phi(x) = \int_a^x f(t)\, dt,$$

is a function of bounded variation on $[a, b]$. This follows since ϕ can be written as the difference of two nondecreasing functions and nondecreasing functions on $[a, b]$ are of bounded variation. So, let us restate our current goal, we want to find for which class of functions the following equality holds

$$\int_a^x F'(x) = F(x) - F(a).$$

Of course, we need F to be differentiable a.e. for the above to make sense. We know that any function of bounded variation satisfies this (first corollary of previous theorem). Also, we can write equivalently

$$F(x) = F(a) + \int_a^x F'(x),$$

where the RHS is a function of bounded variation by the second corollary of the previous theorem. Hence, the class of functions for which the equality is true must be a subset of the class of functions of bounded variation. To characterize this subset we need to introduce the concept of absolute continuity

Definition 11.2 (Absolute continuity)**.** *A function f defined on an interval $[a, b]$ is said to*

be absolutely continuous on $[a, b]$ *if, given any* $\varepsilon > 0$*, there is a* $\delta > 0$ *such that*

$$\sum_{k=1}^{n} |f(b_k) - f(a_k)| < \varepsilon$$

for every finite system of pairwise disjoint subintervals

$$(a_k, b_k) \subset [a, b] \quad (k = 1, ..., n),$$

of total length

$$\sum_{k=1}^{n} (b_k - a_k),$$

less than δ*.*

In fact we can change "finite" for "finite or countable" in the definition. Absolutely continuous functions are a subset of uniformly continuous functions, hence they are all uniformly continuous but there exist uniformly continuous functions which are not absolutely continuous. Now we introduce some useful theorem about absolutely continuous functions

Theorem 11.5. *If* f *is absolutely continuous in* $[a, b]$*, then* f *is of bounded variation on* $[a, b]$

Theorem 11.6. *If* f *is absolutely continuous in* $[a, b]$*, then* f *can be represented as the difference between two absolutely continuous nondecreasing functions on* $[a, b]$*.*

And now the two key theorems to characterize the class of functions we are after

Theorem 11.7. *The indefinite integral*

$$F(x) = \int_a^x f(t)\, dt.$$

of an integrable function f *is absolutely continuous.*

Theorem 11.8 (Lebesgue)**.** *If* F *is absolutely continuous on* $[a, b]$*, then the derivative* F' *is integrable on* $[a, b]$ *and*

$$F(x) = F(a) + \int_a^x F'(t)\, dt.$$

Finally, combining the two last theorems we get that the formula

$$\int_a^x F'(t)\, dt = F(x) - F(a),$$

or equivalently,

$$F(x) = F(a) + \int_a^x F'(t),$$

holds for all $x \in [a, b]$ if and only if F is absolutely continuous on $[a, b]$. This is because by Theorem 11.8 if F is absolutely continuous the equality holds and because if the equality holds then F has to be absolutely continuous by Theorem 11.7.

11.2 Lebesgue decomposition

Let f be a function of bounded variation on $[a, b]$, it can be shown that f can be represented as

$$f(x) = \varphi(x) + \psi(x),$$

where φ is a continuous function and ψ is a jump function, i.e. a function with only constant parts and jumps between these constant parts. Now define

$$\varphi_1(x) = \int_a^x \varphi'(t)\, dt, \quad \varphi_2(x) = \varphi(x) - \varphi_1(x).$$

Since φ' is integrable, φ_1 is absolutely continuous (and hence continuous) by Theorem 11.7. Since φ is also continuous, φ_2 is a continuous function of bounded variation. Differentiating φ_2 we get

$$\varphi_2'(x) = \varphi'(x) - \frac{d}{dx}\int_a^x \varphi'(t)\, dt = 0 \text{ a.e.}$$

We call a continuous function of bounded variation singular if its derivative is zero a.e. Since $\varphi(x) = \varphi_1(x) + \varphi_2(x)$, we have that

$$f(x) = \varphi_1(x) + \varphi_2(x) + \psi(x).$$

That is, we can decompose f into an absolutely continuous function φ_1, a singular function φ_2 and a jump function ψ. This is called the Lebesgue decomposition. Note that

$$f'(x) = \varphi_1'(x) \text{ a.e.}$$

hence, the integration of the derivative only restores the absolutely continuous component of the function and does not leave a trace of the singular and jump components.

11.3 Abstract differentiation

Now we are ready to deal with general measure spaces $(\Omega, \mathcal{F}, \mu)$. Let f now be a measurable and integrable function on Ω. We are interested in the set function

$$\phi(A) = \int_A f\, d\mu, \quad A \in \mathcal{F}.$$

$\phi(A)$ is countably additive, i.e. if $A = \cup_{i=1}^{\infty} A_i$ with $A_1, A_2, ...$ pairwise disjoint measurable sets, then

$$\phi(A) = \sum_{i=1}^{\infty} \phi(A_i).$$

ϕ has all the defining properties of a measure except that it might be negative. This motivates the next definition

Definition 11.3 (Signed measure)**.** *A countably additive set function ϕ defined on a σ-algebra if subsets of a space Ω and in general taking values of both signs is called a signed*

measure or charge[14] *(on* Ω*).*

Hence, a measure is a non-negative signed measure so what follows applies to measures too. Let us classify signed measured in different types.

Definition 11.4. *Let* μ *be a measure on a* σ*-algebra* $\mathcal{F}$ *of subsets of* Ω*, let* ϕ *be a signed measure defined on* $\mathcal{F}$*. Then* ϕ *is said to be concentrated on a set* $A \in \mathcal{F}$ *if* $\phi(E) = 0$ *for every measurable set* $E \subset A^c$*.*

Definition 11.5. *Let* μ*,* $\mathcal{F}$ *and* ϕ *be defined as in the previous definition. Then* ϕ *is said to be*

1. *Continuous if* $\phi(E) = 0$ *for every singleton set* $E \subset X$ *of measure zero,*
2. *Singular if* ϕ *is concentrated on a set of measure zero,*
3. *Absolutely continuous (with respect to* μ*) if* $\phi(E) = 0$ *for every measurable set* E *such that* $\mu(E) = 0$*. This is sometimes denoted as* $\phi << \mu$*.*

At this point it is natural to wonder whether there is a relationship between absolute continuity of a function on the real line and absolute continuity of signed measures (or measures). As usual, they are related through the particular choice of the Lebesgue measure λ. In fact, any finite measure μ on Borel sets of the real line is absolutely continuous with respect to the Lebesgue measure if and only if the function

$$F(x) = \mu((-\infty, x]),$$

is an absolutely continuous real function.

Note that in general, for an integrable function f, the abstract integral

$$\phi(E) = \int_E f \, d\mu,$$

is absolutely continuous with respect to the measure μ. The next fundamental result tells us that any signed measure (or measure) which is absolutely continuous with respect to a measure μ can be written in the form above.

Theorem 11.9 (Radon-Nikodym)**.** *Let* μ *be a* σ*-finite measure defined on a* σ*-algebra* $\mathcal{F}$ *of subsets of* Ω*, let* ϕ *be a* σ*-finite signed measure (or measure) defined on* $\mathcal{F}$*. Suppose* $\phi << \mu$*. Then, there exists an integrable function* φ *on* Ω *such that*

$$\phi(E) = \int_E \varphi \, d\mu,$$

for every $E \in \mathcal{F}$*. The function* φ *is unique up to values on a set of* μ*-measure zero.*

The function φ is called the Radon-Nikodym derivative (or just density) of the signed measure (or measure) ϕ with respect to μ, it is usually denoted by $d\phi/d\mu$.

[14] This name is analogous to how a surface carrying electrical charge can be divided into a region with positive charge and one with negative charge.

Example 11.1. *Let* $(\Omega, \mathcal{F}) = ([0,1], \mathcal{B}([0,1]))$ *and* $\mu = 2\lambda$ *where* λ *is the Lebesgue measure. Clearly,* $\mu << \lambda$ *and*

$$\mu(E) = 2\lambda(E) = \int_E 2\, d\lambda,$$

so $d\mu/d\lambda = 2$.

Example 11.2. *Let* $(\Omega, \mathcal{F}) = ([0,1], \mathcal{B}([0,1]))$ *and* $\mu = \lambda + \delta_0$ *where* λ *is the Lebesgue measure and* δ_0 *is the Dirac measure giving point mass at* $\{0\}$. *Clearly,* $\lambda << \mu$ *and*

$$\lambda(E) = \lambda(E \cap \Omega \setminus \{0\}) + \delta_0(E \cap \Omega \setminus \{0\}) = \int_E \mathbb{1}_{\Omega\setminus\{0\}}\, d\mu,$$

so $d\lambda/d\mu = \mathbb{1}_{\Omega\setminus\{0\}}$.

Notice that the equality in the Theorem can be written as

$$\phi(E) = \int_E \frac{d\phi}{d\mu}\, d\mu,$$

pointing out that the Radon-Nikodym derivative is a generalization of Theorem 11.8 which states that an absolutely continuous function is the integral of its own derivative (up to a constant). The Radon-Nikodym theorem however does not give you a way of computing the derivatives, it just gives you existence and a.e. uniqueness of the derivative. Usually, the Radon-Nikodym derivative is computed by solving the functional equation in the theorem where φ is unknown. Unfortunately there is no cookbook to solve this functional equation since it varies a lot from case to case. However, the Radon-Nikodym derivative at a point x_0 could be explicitly calculated by solving

$$\lim_{\varepsilon\to 0} \frac{\phi(A_\varepsilon}{\mu(A_\varepsilon)},$$

where A_ε is a system of measurable subsets converging to $\{x_0\}$. This construction gives us the usual interpretation of the derivative being the instantaneous change in the set function ϕ from an instantaneous change in the measure μ. The theorem gives us the interpretation that a set function can be recovered from integrating all its instantaneous changes.

Finally, it can be shown that for any σ-finite signed measure ϕ and σ-finite measure μ, ϕ can be decomposed as

$$\phi = \phi_A + \phi_S + \phi_D,$$

where $\phi_A << \mu$, ϕ_S is singular and ϕ_D is a discrete. This decomposition is analogous to the Lebesgue decomposition for functions and also carries the consequence that by applying the Radon-Nikodym Theorem to only the absolutely continuous part we retrieve the absolutely continuous part of the signed measure.

12 Random variables

We are finally ready to go back to probability theory and study what random variables (RVs) are. As we will see everything boils down to a special case of all the theory we have seen

about measurable functions, abstract integration and abstract differentiation. Therefore, it is advised to go back to the general concepts whenever a concept is said to be a particular case of a more general concept. If you have not read the sections and concepts marked with * (for example the whole section on differentiation) do not worry. While helpful to get a deeper understanding of what comes next is not essential in order to follow it.

The motivation behind RVs comes from the fact that you might not be interested in every possible outcome $\omega \in \Omega$. For instance, if your random experiment consists on tossing a coin 10 times, you might not be interested in the exact sequence of heads and tails but just in how many heads turn up. Note that every outcome ω gives you a specific sequence of heads and tail and hence, the number of heads will be a function of ω. There are cases where the random experiment is much more complicated and you want to define RVs to focus on something simpler. For instance, if your random experiment is the economy in the years 2021-2022 you might just want to look at inflation and not at every possible outcome of the economy. If your random experiment is the weather, you might just be interested in temperature.

As we will see, RVs are functions defined on Ω. Hence, the name "random variable" is somewhat unsatisfactory since random variables are a deterministic function of ω and hence not random and they are functions not variables. The only randomness comes in the choice of ω, but the functions are deterministic.

12.1 Definition and c.d.f.

Definition 12.1 (Random variable)**.** *A random variable on the probability space* $(\Omega, \mathcal{F}, \mathbb{P})$ *is an* $\mathcal{F}$*-measurable function* $X : \Omega \to \mathbb{R}$*.*

We use capital letters to denote RVs, e.g. X, but we use lower case letters to denote realizations, i.e. $x = X(\omega)$. So RVs are just a special case of measurable functions, they are measurable functions defined on a probability space. Ultimately we want to assign probabilities to events $A \in \mathcal{F}$, this is why we require RVs to be measurable. This motivation can be seen in the next definition.

Definition 12.2 (Probability law)**.** *The probability law of a RV* X*,* $\mathbb{P}_X : \mathcal{B} \to [0, 1]$ *is defined for each Borel set as*

$$\mathbb{P}_X(B) \equiv \mathbb{P}(X^{-1}(B)) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\}).$$

The probability law of a RV X takes a subset of $\mathbb{R}$ (some Borel set) and tells you what is the probability that the RV X takes a value in B. To do this it exploits that RVs are measurable so that the pre-image of B (under X) is an event, i.e. a set belonging in the σ-algebra on which we have a probability measure $\mathbb{P}$ defined. The probability law is a particular case of an induced measure, it can be written as $\mathbb{P}_X = \mathbb{P} \circ X^{-1}$.

Theorem 12.1. $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$ *is a probability space.*

Proof. (Exercise, hint: X is a measurable function and $\mathbb{P}$ is a measure.) □

Once you have $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \mathbb{P}_X)$ and if you are interested only on the RV X, then you do not need the original probability space $(\Omega, \mathcal{F}, \mathbb{P})$ anymore since you already have a complete probabilistic description on $\mathbb{R}$. However, this probabilistic description is still quite involved since it requires to know $\mathbb{P}_X$ for all possible Borel sets, for instance it requires to know $\mathbb{P}_X$ for weird Borel sets such as the Cantor set. Fortunately, we can simplify further the probabilistic description. To do this we introduce the concept of a Cumulative Distribution function (c.d.f.).

Definition 12.3 (C.D.F)**.** *Given a RV X, its c.d.f. at $x \in \mathbb{R}$ is given by*

$$F_X(x) = \mathbb{P}_X((-\infty, x]) = \mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\}).$$

Note that $(-\infty, x])$ is a Borel set for all $x \in \mathbb{R}$ $F_X(x)$ is well-defined. With quite an abuse of notation, from now on when we write $\mathbb{P}(X \leq x)$ we mean $\mathbb{P}(\{\omega \in \Omega : X(\omega) \leq x\})$ and when we write $\mathbb{P}(X \in B)$ we mean $\mathbb{P}(\{\omega \in \Omega : X(\omega) \in B\})$.

It turns out (quite amazingly) that the c.d.f. of X determines uniquely $\mathbb{P}_X$ for all Borel sets. So even though $\mathbb{P}_X$ seems much more general, we can work with F_X and still have a full probabilistic description. Intuitively this is because the c.d.f. gives the probability measure of a generating class of $\mathcal{B}(\mathbb{R})$, i.e. it can be proven that the σ-algebra generated by $\{(-\infty, x] : x \in \mathbb{R}\}$ is the Borel σ-algebra.

*The mathematical result behind this very powerful fact is the π-λ theorem, which if applied to probability states that if two probability laws agree on a π-system[15], then they agree on the σ-algebra generated by the π-system. It turns out that the set $\{(-\infty, x] : x \in \mathbb{R}\}$ is a π-system so that if $F_X = F_Y$ then $\mathbb{P}_X = \mathbb{P}_Y$. For those who have read the Caratheodory's extension theorem: algebras are π-systems so the π-λ theorem can be used in its proof.

Hence, stating this in a theorem

Theorem 12.2. *The c.d.f. F_X of a RV X uniquely specifies the probability law $\mathbb{P}_X$.*

from now on we can just work with the c.d.f.

[15] A π-system $\mathcal{T}$ on a set Ω is a non-empty collection of subset of Ω such that if $A \in \mathcal{T}$ and $B \in \mathcal{T}$, then $A \cap B \in \mathcal{T}$.

16 Concentration Inequalities

Concentration inequalities give you probability bounds on r.v.s taking values in some range. This why they are called concentration inequalities, since it is about probability concentrating in some range. I follow Wainwright (2019) very closely.

Proposition 16.1 (Markov's Inequality)**.** *If X is a non-negative r.v. with $\mathbb{E}[X] < \infty$, then for any $\alpha > 0$,*

$$\mathbb{P}(X > \alpha) \leq \frac{\mathbb{E}[X]}{\alpha}.$$

Proof.

$$\mathbb{E}[X] = \underbrace{\mathbb{E}[X\mathbb{1}(X \leq \alpha)]}_{\geq 0} + \mathbb{E}[X\mathbb{1}(X > \alpha)] \geq \mathbb{E}[X\mathbb{1}(X > \alpha)] \geq \alpha\,\mathbb{E}[\mathbb{1}(X > \alpha)] = \alpha\,\mathbb{P}(X > \alpha).$$

□

Note that this is only useful for $\alpha > \mathbb{E}[X]$, otherwise the RHS above is 1. This is a very loose inequality. It just says that the probability of being above α decays at rate $1/\alpha$. However, this probability often decays much faster. If we add further assumptions we can get faster rates. For instance, assuming finite variance (or finite second moment) we get the following

Proposition 16.2 (Chebyshev's Inequality)**.** *If X is a r.v. with mean μ and variance $\sigma^2 < \infty$, then for any $\alpha > 0$*

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

Alternatively, we can write the result as $\mathbb{P}(|X - \mu| > \alpha\sigma) \leq 1/\alpha^2$, which tells us that the probability that $|X - \mu|$ is α times the standard deviation, decays at rate $1/\alpha^2$. The proof follows from applying the Markov Inequality to the r.v. $|X-\mu|^2$. Still, Chebyshev's Inequality is not very sharp. Assuming the existence of the expectation of increasing transformations of X increases our decay rates.

Proposition 16.3 (Extended Markov Inequality)**.** *If $\varphi(\cdot)$ is a monotone increasing non-negative function on the positive reals, X is a r.v. with $\mathbb{E}[\varphi(X)] < \infty$, then*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}[\varphi(X)]}{\varphi(a)}.$$

The proof follows from the fact that $\mathbb{P}(X \geq a) = \mathbb{P}(\varphi(X) \geq \varphi(a))$ and from applying Markov's Inequality to r.v. $\varphi(X)$. This result allows us to find much sharper bounds. For instance, if X has a m.g.f., meaning that for some $b > 0$, $\varphi_X(\lambda) = \mathbb{E}[e^{\lambda X}]$ exists for $\lambda \leq |b|$, we can get exponential decays, for instance

$$\mathbb{P}[(X - \mu) \geq t] = \mathbb{P}(e^{\lambda(X-\mu)} \geq e^{\lambda t}) \leq \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda t}},$$

and we can choose λ as to make the bound as sharp as possible to get the Chernoff bound.

Proposition 16.4 (Chernoff bound)**.** *Let X be a r.v. such that $\mathbb{E}[e^{\lambda(X-\mu)}] < \infty$ for $\lambda \leq |b|$, then*

$$\log \mathbb{P}[(X - \mu) \geq t] \leq \inf_{\lambda \in [0,b]} \Big(\log \mathbb{E}[e^{\lambda(X-\mu)}] - \lambda t \Big). \tag{16.1}$$

Of course, the λ which gives the sharpest bound depends on b, an example in which the m.g.f exists for any $\lambda \in \mathbb{R}$ is the Gaussian case.

16.1 Sub-Gaussian variables and Hoeffding bounds

Let $X \sim \mathcal{N}(\mu, \sigma^2)$, the m.g.f. is

$$\mathbb{E}[e^{\lambda X}] = e^{\mu\lambda + \frac{\sigma^2\lambda^2}{2}}, \text{ for all } \lambda \in \mathbb{R}.$$

The Chernoff bound tells us that it is natural to classify random variables according to the growth rate of their m.g.f. since the tail bounds depend on this growth rate. The simplest classification is that of sub-Gaussian random variables. For our Gaussian X, taking the infimum in (16.1) over the set $\lambda \geq 0$, yields

$$\inf_{\lambda \geq 0} \Big(\log \mathbb{E}[e^{\lambda(X-\mu)}] - \lambda t \Big) = \inf_{\lambda \geq 0} \Big(\frac{\lambda^2\sigma^2}{2} - \lambda t \Big) = -\frac{t^2}{2\sigma^2}.$$

Where we just plugged in the λ which makes the first derivative with respect to λ zero. This means that if $X \sim \mathcal{N}(\mu, \sigma^2)$, then we have the following upper deviation inequality for all $t \geq 0$

$$\mathbb{P}(X \geq \mu + t) \leq e^{-\frac{t^2}{2\sigma^2}}. \tag{16.2}$$

This motivates the definition of sub-Gaussian random variables

Definition 16.1 (Sub-Gaussian random variables)**.** *A random variable X with mean $\mu = \mathbb{E}[X]$ is sub-Gaussian if there is a positive number σ such that for all $\lambda \in \mathbb{R}$*

$$\mathbb{E}[e^{\lambda(X-\mu)}] \leq e^{\frac{\sigma^2\lambda^2}{2}}.$$

Here, σ is called the sub-Gaussian parameter. By the arguments before, it follows that any sub-Gaussian r.v. X with sub-Gaussian parameter σ (and mean μ) satisfies (16.2), since for such a r.v. and for all $\lambda \in \mathbb{R}$ (repeating the same arguments as before)

$$\begin{aligned}\mathbb{P}(X - \mu \geq t) &= \mathbb{P}(e^{\lambda(X-\mu)} \geq e^{\lambda t}) \\ &\leq \frac{\mathbb{E}[e^{\lambda(X-\mu)}]}{e^{\lambda t}} \\ &\leq \frac{e^{\frac{\sigma^2\lambda^2}{2}}}{e^{\lambda t}},\end{aligned}$$

where in the first inequality we use Markov and in the second the fact that X is sub-Gaussian with sub-Gaussian parameter σ. Then, following the arguments of the Chernoff bound

$$\begin{aligned}\log \mathbb{P}(X \geq \mu + t) &\leq \inf_{\lambda \geq 0} \frac{\sigma^2\lambda^2}{2} - \lambda t \\ &= -\frac{t^2}{2\sigma^2},\end{aligned}$$

so

$$\mathbb{P}(X \leq \mu + t) \leq e^{\frac{-t^2}{2\sigma^2}}.$$

It also follows that the r.v. $-X$ is sub-Gaussian if and only if X is sub-Gaussian since, letting $Z = -X$ for all $\tilde{\lambda} \in \mathbb{R}$

$$\begin{aligned}\mathbb{E}[e^{\lambda(Z-\mathbb{E}[Z])}] &= \mathbb{E}[e^{\lambda(-X+\mu)}] \\ &= \mathbb{E}[e^{\tilde{\lambda}(X-\mu)}] \\ &\leq e^{\frac{\sigma^2\tilde{\lambda}}{2}},\end{aligned}$$

where in the first equality we let $\tilde{\lambda} = -\lambda$ (i.e. the statement holds for all $\lambda \in \mathbb{R}$ and hence for all $\tilde{\lambda} \in \mathbb{R}$. This shows that if X is sub-Gaussian then $-X$ is sub-Gaussian, the converse follows using the same logic. The fact that $-X$ is also sub-Gaussian with parameter σ allows us to show a lower deviation inequality for X

$$\begin{aligned}\mathbb{P}(X \leq \mu - t) &= \mathbb{P}(-X \geq -\mu + t) \\ &\leq e^{\frac{-t^2}{2\sigma^2}},\end{aligned}$$

where to get the inequality we again apply sub-Gaussianity together with the Chernoff bound. This discussion allows us to prove the following concentration inequality

$$\mathbb{P}(|X - \mu| \geq t) \leq 2e^{\frac{-t^2}{2\sigma^2}} \text{ for all } t \in \mathbb{R}.$$

This concentration inequality follows since

$$\begin{aligned}\mathbb{P}(|X-\mu|\geq t) &= \mathbb{P}(\{X-\mu\leq -t\}\cup\{X-\mu\geq t\})\\ &= \mathbb{P}(X-\mu\leq -t)+\mathbb{P}(X-\mu\geq t)\\ &\leq 2e^{\frac{-t^2}{2\sigma^2}},\end{aligned}$$

where in the second equality we use that the intersection of the two sets is empty. Now we are going to cover examples of important random variables (we will see later why they are important) which are sub-Gaussian but not Gaussian variables themselves.

Example 16.1 (Rademacher r.v.s)**.** *A Rademacher r.v. ε is a r.v. taking values* 1 *and* -1 *with* $\mathbb{P}(\varepsilon=1)=\mathbb{P}(\varepsilon=-1)=1/2$*. Note that* $\mathbb{E}[\varepsilon]=0$*, the first thing we do is to compute its m.g.f. (all equalities are explained below)*

$$\begin{aligned}\mathbb{E}[e^{\lambda\varepsilon}] &= \frac{1}{2}(e^{\lambda}+e^{-\lambda})\\ &= \frac{1}{2}\Big(\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}+\sum_{k=0}^{\infty}\frac{(-\lambda)^k}{k!}\Big)\\ &= \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{(2k)!}\\ &\leq \sum_{k=0}^{\infty}\frac{\lambda^{2k}}{2^k k!}\\ &= e^{\frac{\lambda^2}{2}}.\end{aligned}$$

The first equality follows from the law of total expectation. The second equality follows form the Taylor series of e^{λ} *and* $e^{-\lambda}$ [16]*. The third equality follows from the fact that when the exponent is odd the expression is* 0*, so only even terms survive. The first inequality follows since* $(2K)!/K!\geq 2^K$ *for all* $K\in\mathbb{N}$ *(you can prove this by induction) so that* $[(2K)!/K!]K!=(2K)!\geq 2^K K!$ *and the last equality follows because of the Taylor series*[17]*. Hence,* $\mathbb{E}[e^{\lambda\varepsilon}]\leq e^{\frac{\sigma^2\lambda^2}{2}}$ *with sub-Gaussian parameter* $\sigma=1$*.*

Example 16.2 (Bounded random variables)**.** *Let* X *be a r.v. with zero mean (otherwise consider a centered version) which takes values in a bounded interval* $[a,b]$*. Also let* X' *be an independent copy of* X *(i.e.* X *and* X' *are two independent r.v.s which are identically distributed). Let* $\mathbb{E}_X$ *and* $\mathbb{E}_{X'}$ *be the expectation operators with respect to* X *and* X' *respectively. Then, since* $\mathbb{E}_{X'}[X']=0$ *we have that*

$$\begin{aligned}\mathbb{E}_X[e^{\lambda X}] &= \mathbb{E}_X[e^{\lambda(X-\mathbb{E}_{X'}[X'])}]\\ &= \mathbb{E}_X[e^{\mathbb{E}_{X'}[\lambda(X-X')]}]\\ &\leq \mathbb{E}_{X,X'}[e^{\lambda(X-X')}].\end{aligned}$$

[16] The Taylor series of e^{λ} at $\lambda=0$ (Maclaurin series) is $e^0+e^{\lambda}|_{\lambda=0}\lambda+e^{\lambda}/2!|_{\lambda=0}\lambda^2+...=1+\lambda+\lambda^2/2!+\lambda^3/3!+...=\sum_{k=0}^{\infty}\lambda^k/k!$.

[17] Define $z=\lambda^2$, then do the Taylor series of of $e^{z/2}$ at $z=0$ and you will get $\sum_{k=0}z^k/(2^k k!)$, replace z with λ^2 and you have it.

In the second equality we use independence (which implies $\mathbb{E}_{X'}[X] = X$) and the inequality follows from Jensen's inequality (the exponential is a convex function). Now we note that $X - X'$ has the same distribution as $\varepsilon(X - X')$ (and hence the same expectation) where ε is an independent Rademacher r.v.[18]*, then*

$$\begin{aligned}\mathbb{E}_{X,X'}[e^{\lambda(X-X')}] &= \mathbb{E}_{X,X'}[\mathbb{E}_{\varepsilon}[e^{\lambda\varepsilon(X-X')}]] \\ &\leq \mathbb{E}_{X,X'}[e^{\frac{\lambda^2(X-X')^2}{2}}],\end{aligned}$$

where the inequality follows from Example 16.1 where what was there λ now is $\lambda(X - X')$. Now, since X can only take values in $[a, b]$, $X - X' \leq b - a$ and hence

$$\mathbb{E}_{X,X'}[e^{\frac{\lambda^2(X-X')^2}{2}}] \leq e^{\frac{\lambda^2(b-a)^2}{2}}.$$

Hence we have shown that X is a sub-Gaussian r.v. with sub-Gaussian parameter of at most $\sigma = b - a$. We say at most since in fact we can get a sharper inequality and show that the m.g.f is also bounded by $(b - a)/2$ (see Exercise 2.4 in Wainwright (2019)).

The trick of using an independent copy X' and an independent Rademacher r.v. ε is a very useful trick called symmetrization which is used in many contexts. Most times we are interested in bounding sums of independent variables, to do this a useful result is that sub-Gaussianity is preserved by linear operations, i.e. if $X1$ and X_2 are independent sub-Gaussian r.v.s. with parameters σ_1 and σ_2, then $X_1 + X_2$ is sub-Gaussian with parameter $\sqrt{\sigma_1^2 + \sigma_2^2}$. This leads to the following important upper deviation inequality called the Hoeffding bound

Proposition 16.5 (Hoeffding bound)**.** *Let $X_1, X_2, ..., X_n$ be independent sub-Gaussian r.v.s where X_i, $i = 1, ..., n$, with mean μ_i and sub-Gaussian parameter σ_i, then, for all $t \geq 0$*

$$\mathbb{P}\Big(\sum_{i=1}^{n}(X_i - \mu_i) \geq t\Big) \leq exp\Big(-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\Big).$$

By the same arguments as before we can also get the following concentration inequality. Bounded r.v.s. are a particular case, if $X_i \in [a, b]$ for all $i = 1, ..., n$, then the Hoeffding bound yield (taking the sub-Gaussian parameter $\sigma = (b - a)/2$ from Exercise 2.4 in Wainwright (2019))

$$\mathbb{P}\Big(\sum_{i=1}^{n}(X_i - \mu_i) \geq t\Big) \leq exp\Big(-\frac{2t^2}{n(b-a)^2}\Big).$$

In many sources (e.g. Wikipedia) the Hoeffding bound is stated as the one above, i.e. for bounded r.v.s. However, as we have seen, this bound applies more generally to sub-Gaussian r.v.s. Finally, by the same arguments as before we can also get the following concentration inequality

$$\mathbb{P}\Big(\Big|\sum_{i=1}^{n}(X_i - \mu_i)\Big| \geq t\Big) \leq 2exp\Big(-\frac{t^2}{2\sum_{i=1}^{n}\sigma_i^2}\Big).$$

[18] $F_{\varepsilon(X-X')}(x) = \mathbb{P}(\varepsilon(X - X') \leq x) = \mathbb{P}((X - X') \leq x \mid \varepsilon = 1)(1/2) + \mathbb{P}((X' - X) \leq x \mid \varepsilon = -1)(1/2) = (1/2)(\mathbb{P}((X - X') \leq x) + \mathbb{P}((X' - X) \leq x)) = \mathbb{P}(X - X' \leq x)$. Where we use independence of ε and the symmetry together with identical distribution of X and X'.

17 Convergence of Random Variables

Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space and $X_1, X_2, ...$ be a sequence of r.v.s.

Definition 17.1 (Pointwise or sure convergence)**.** *The sequence* $\{X_n, n = 1, 2, ...\}$ *is said to converge* pointwise *or* surely *to* X*,* $X_n \to_{pw} X$*, if*

$$X_n(\omega) \to X(\omega) \text{ for all } \omega \in \Omega \text{ as } n \to \infty.$$

Note that for a fixed $\omega \in \Omega$, $\{Xn(\omega), n = 1, 2, ...\}$ is just a sequence of real numbers. Remember that that a sequence of real numbers $\{a_n, n = 1, 2, ...\}$ is said to converge to a, written as $a_n \to a$ or $\lim_{n\to\infty} a_n = a$, iff for any $\varepsilon > 0$, there exists $N \geq 1$ such that for all $n > N$, $|a_n - a| < \varepsilon$. In fact, $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega$ is *exactly* pointwise convergence of functions[19]. Most of the times this is a very strong notion of convergence. This is because it also requires convergence in sets with probability 0. Hence, it is rarely used. Note that a technicality which we do not show here is that for the definition to make sense one would have to show that the limit of a sequence of measurable functions is measurable, i.e. if $X_1, X_2, ...$ are r.v.s so is X. Now we define a convergence notion which allows convergence not to happen in probability zero events.

Definition 17.2 (Almost sure or wp1 convergence)**.** X_n *converges to* X *almost surely or wp1,* $X_n \to_{a.s.} X$*, if* $X_n(\omega) \to X(\omega)$ *on a set of probability 1, that is*

$$\mathbb{P}\Big(\{\omega : X_n(\omega) \to X(\omega)\}\Big) = 1.$$

Sometimes this is also called strong convergence. Here there is also a technicality, for the above to make sense we need $\{\omega : X_n(\omega) \to X(\omega)\}$ to be an event (i.e. to be measurable, to be an element of $\mathcal{F}$, etc...). To do this one can show that the event is a countable union/intersection of events, we do not do it here but the reader is encouraged to do it. We now define a weaker concept of convergence which is widely used.

Definition 17.3 (Convergence in probability)**.** X_n *converges in probability to* X*,* $X_n \to_p X$*, if for all* $\varepsilon > 0$

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

Note that a.s. convergence and convergence in probability are very different. In a.s. convergence, the limit is inside the probability while in convergence in probability the limit is outside of the probability. In fact, it could be said that convergence of X_n in probability is not the most suitable name for this concept since it is not really X_n that is converging but a sequences of probabilities. In fact, there might be sets with probability strictly greater than 0 in which $X_n(\omega)$ and $X(\omega)$ are not "close". What we really have is that the sequence $\mathbb{P}_n(\varepsilon) \to 0$ where $\mathbb{P}_n(\varepsilon) \equiv \mathbb{P}(|X_n - X| > \varepsilon)$. In contrast, a.s. convergence does refer to convergence of the sequence $X_1, X_2, ...$ We will later show that a.s. convergence implies

[19] A sequence of functions $\{f_n, n = 1, 2, ...\}$ with domain D and codomain C is said to converge pointwise to some function $f : D \mapsto C$ iff $\lim_{n\to\infty} f_n(x) = f(x)$ for all $x \in D$.

convergence in probability but not the other way around, meaning a.s. convergence is a stronger concept of convergence.

Definition 17.4 (Convergence in r-th mean)**.** *X_n converges to X in r-th mean, $X_n \to_{r-th} X$, if*

$$\lim_{n\to\infty} \mathbb{E}[|X_n - X|^r] = 0.$$

For $r = 2$, X_n is said to converge in mean-squared sense.

Definition 17.5 (Convergence in distribution)**.** *X_n converges to X in distribution, $X_n \to_d X$ or $X_n \rightsquigarrow X$, if*

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x) \text{ for all } x \text{ where } F_X(x) \text{ is continuous.}$$

Convergence in distribution is also called weak convergence since, as we will show it is the weakest form of convergence. Again, this is really convergence of a sequence of c.d.f.s not really convergence of a sequence of r.v.s, that is X_n and X might be far but have close c.d.fs. Now we state the hierarchy between these convergence notions and in the proofs it will be clear what we mean when we say that for some of them we do not require X_n and X to be "close".

Proposition 17.1 (Hierarchy of Convergence modes)**.** *Do GRAPH.*

1. *$X_n \to_{pw} X$ implies $X_n \to_{a.s.} X$,*
2. *$X_n \to_{a.s.} X$ implies $X_n \to_p X$,*
3. *$X_n \to_p X$ implies $X_n \to_d X$*
4. *$X_n \to_{r-th} X$ for $r \geq 1$ implies $X_n \to_p X$*

Any other relationship does not hold.

To prove the above proposition we need to prove three implications asides from sure convergence implying a.s. convergence which is direct. We also have to provide five counterexamples. So let's start doing this in turn.

Proof: *$X_n \to_{r-th} X$ implies $X_n \to_p X$ for $r \geq 1$.* We use Markov's Inequality:

$$\mathbb{P}\Big(|X_n - X| > \varepsilon\Big) = \mathbb{P}\Big(|X_n - X|^r > \varepsilon^r\Big) \leq \frac{\mathbb{E}[|X_n - X|^r]}{\varepsilon^r},$$

so

$$\lim_{n\to\infty} \mathbb{P}\Big(|X_n - X| > \varepsilon\Big) \leq \lim_{n\to\infty} \frac{\mathbb{E}[|X_n - X|^r]}{\varepsilon^r} = 0$$

□

Proof: *$X_n \to_p X$ implies $X_n \to_d X$.* Note that

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

Where the inequality comes from the fact that $\{X_n \leq x\} \cap \{X \leq x + \varepsilon\} \subseteq \{X \leq x + \varepsilon\}$ and that $\{X_n \leq x\} \cap \{X > x + \varepsilon\} \subseteq \{|X_n - X| > \varepsilon\}$. Similarly, we can show that $F_X(x - \varepsilon) \leq F_{X_n}(x) + \mathbb{P}(|X - X_n| > \varepsilon)$. Putting the two inequalities together we get

$$F_X(x - \varepsilon) - \mathbb{P}(|X - X_n| > \varepsilon) \leq F_{X_n}(x) \leq F_X(x + \varepsilon) + \mathbb{P}(|X_n - X| > \varepsilon).$$

As $n \to \infty$ the above becomes

$$F_X(x - \varepsilon) \leq F_{X_n}(x) \leq F_X(x + \varepsilon).$$

So, by sending $\varepsilon \to 0$ we get that $F_{X_n}(x) \to F_X(x)$ if F_X is continuous at x. □

18 Law of Large Numbers

19 Central Limit Theorem

References

Georg Cantor. On a property of the class of all real algebraic numbers. *Crelle's Journal for Mathematics*, 77:258–262, 1874.

Angel De la Fuente. *Mathematical methods and models for economists*. Cambridge University Press, 2000.

Andrei Nikolaevich Kolmogorov and Sergeui Vasilevich Fomin. *Introductory real analysis*. Courier Corporation, 1975.

Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.