# Probability and Stochastic Processes I, MATH5160
# (Graduate Probability)

Iddo Ben-Ari

Fall 2023

# Contents

# Chapter 1

# Probability Spaces

Idea: the mathematical formalism of chance.

The mathematical theory which formalizes chance is called probability theory. Probability theory provides mathematical structures that can be thought of as abstract versions of "real world experiments" (tossing a coin, gambling) with multiple possible outcomes (Heads/Tails, winning/losing a certain amounts), along with measurements of "chances" of the various outcomes. These structures are called probability spaces, and in this chapter we will develop them.

## 1.1 Measurable spaces

Throughout this chapter $\Omega$ will be assumed to be a nonempty set, which intuitively represents the set of possible outcomes in some experiment. The set $\Omega$ is called the sample space, suggesting that we view the experiment as a way to sample an outcome, an element of $\Omega$. For example, the outcome of each coin toss is a sample from the sample space $\{H, T\}$, and the outcome of a roll of a dice (assuming we observe the number on the top face) is an element from the sample space $\{1, 2, \ldots, 6\}$. We start by developing the "language" we will use to describe possible outcomes. An explanation will follows the definition.

**Definition 1.1.1.** *A collection $\mathcal{A}$ of subsets of $\Omega$ is called an algebra, if*

1. *(not empty) $\Omega \in \mathcal{A}$.*
2. *(closed under complements) If $A \in \mathcal{A}$ then $A^c$ (the complement of $A$) is in $\mathcal{A}$.*
3. *(closed under finite unions) If $A, B \in \mathcal{A}$, then $A \cup B \in \mathcal{A}$.*

By deMorgan's laws an algebra is necessarily closed under finite intersections. Indeed, $A \cap B = (A^c \cap B^c)^c$. If $A, B$ are elements in an algebra $\mathcal{A}$, then so is their difference: $A - B := A \cap B^c$, and of course their symmetric difference, $A \triangle B := (A - B) \cup (B - A)$.

The elements of an algebra (and later $\sigma$-algebra), each being a set of outcomes in the sample space, should be viewed as statements we can make on the outcome of the experiment, and which can be determined by the information available to us at the conclusion of the experiment. Think of rolling a dice and observing the number on the top face. Then the sample space is $\Omega = \{1, 2, \ldots, 6\}$. Let's assume we have complete information. There is exactly one outcome (a single number), but we can make multiple statements on it which will be determined as true or false at the conclusion of the experiment. Here are several examples:

- "landed a 3", the set $A = \{3\}$.
- "Not landed a 3", $A^c$.
- "Landed an even number", $B = \{2, 4, 6\}$.
- "Landed an even number or 3", $A \cup B$.
- "Landed both on an even number and on 3", $A \cap B = \emptyset = \Omega^c$.

Let's now change the experiment a little and assume that we are only told the parity of the number on the top face. Though the sample space has not changed, we have less information on the outcome, and there are only 4 valid statements on the outcome that we can determine as true or false at the conclusion of the experiment: $\emptyset$ ("nothing"), $\{2,4,6\}$ (landed even), $\{1,3,5\}$ (landed odd), $\{1,2,3,4,6\}$ (anything). So each version of the experiment corresponds to a different algebra of sets.

There's a correspondence between an algebra $\mathcal{A}$ and the set of indicator functions of elements in $\mathcal{A}$. Recall that $\mathbf{1}_A$ is the indicator function of the set $A$, if it is a function from $\Omega$ to $\{0,1\}$ which is equal to 1 on $A$ and 0 otherwise. For a collection $\mathcal{A}$ of subsets of $\Omega$, let $I_{\mathcal{A}}$ denote the set of indicator of functions of elements in $\mathcal{A}$. Then $\mathcal{A}$ is an algebra if and only if the constant function $\mathbf{1} \in I_{\mathcal{A}}$, for any $A \in \mathcal{A}$, $\mathbf{1} - \mathbf{1}_A \in I_{\mathcal{A}}$, and for any $A, B \in \mathcal{A}$ the "addition" $\mathbf{1}_A \oplus \mathbf{1}_B$, defined as $\max(\mathbf{1}_A, \mathbf{1}_B)$, is also an element in $I_{\mathcal{A}}$ (you can also replace this by a product of the functions. Why?).

As examples of algebras, consider the collection of subsets of $\Omega$ which are finite or have a finite complement. Another example? Take $\Omega = [0,1)$ and consider all finite unions of intervals of the form $[a,b)$ for $0 \le a < b \le 1$. Why is this closed under complements? Draw such a set. Look at your drawing. Now write your proof.

Here's an important property of algebras.

**Proposition 1.1.1.** *Let $A_1, A_2, \dots$ be elements of an algebra. Then there exists a sequence of disjoint elements in the algebra $B_1, B_2, \dots$ such that for every $n \in \mathbb{N}$, $\cup_{j=1}^n A_j = \cup_{j=1}^n B_j$.*

*Proof.* Exercise. □

Though the first object defined in this book, algebras are not going to be used very much in our class. In order to prove nice theorems (nearly anything involving limits: Law of Large Numbers, Central Limit Theorem. After all probability is a branch of analysis) we need richer structures that allow to take limits. This leads to the definition of a $\sigma$-algebra (AKA $\sigma$-field, I term I will not use).

**Definition 1.1.2.** *A collection $\mathcal{F}$ of subsets of $\Omega$ is a $\sigma$-algebra if*

1. *$\Omega \in \mathcal{F}$.*
2. *If $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$.*
3. *If $A_1, A_2, \dots \in \mathcal{F}$, then $\cup_{i=1}^\infty A_i \in \mathcal{F}$.*

Thus a $\sigma$-algebra is a nonempty collection of subsets of $\Omega$ closed under complements and countable unions. Note that this automatically implies being closed under finite unions, and under countable (and finite) intersections. Of course, every $\sigma$-algebra is an algebra, but the converse is not true (example?).

In probability theory we use $\sigma$-algebras, not algebras, as the collection of statements on the outcome of the experiment which we, the observers, can determine whether true or false at the conclusion of the experiment. The additional requirement allows to form a rich collection of statements. When $\Omega$ is finite every algebra is automatically a $\sigma$-algebra.

Here are some examples of $\sigma$-algebras.

1. The collection $\{\emptyset, \Omega\}$, also known as the trivial $\sigma$-algebra.
2. The power set of $\Omega$, all subset of $\Omega$, denoted by $2^\Omega$.
3. The collection of all subsets of $\Omega$ which are countable (this includes finite) or have a countable complement.

**Exercise 1.1.1.** *Show that the last object in the list above is a $\sigma$-algebra. When is it the power set?*

**Definition 1.1.3.** *A measurable space is a pair $(\Omega, \mathcal{F})$ of a nonempty set $\Omega$, the sample space, and a $\sigma$-algebra $\mathcal{F}$ on $\Omega$. Given a measurable space, the elements of $\mathcal{F}$ are called events.*

A measurable space therefore represents the set of outcomes in the experiment, $\Omega$, and the information we have on the outcome at the conclusion of the experiment, $\mathcal{F}$. By "information" we mean: all statements on the outcome which we can determine as true or false at the conclusion of the experiment.

$\sigma$-algebras have some very nice properties.

**Proposition 1.1.2.** *Suppose that $I$ is a set of indices and that for each $i \in I$, $\mathcal{F}_i$ is a $\sigma$-algebra on $\Omega$. Then $\mathcal{F} = \cap_{i \in I} \mathcal{F}_i$ is a $\sigma$-algebra.*

*Proof.* Exercise. □

**Exercise 1.1.2.** *Is the union of $\sigma$-algebras a $\sigma$-algebra?*

Let's look at some events obtained through countable operations.

**Definition 1.1.4.** *Let $A_1, A_2, \dots$ be events.*

1. $\limsup_{j\to\infty} A_j := \cap_{n=1}^{\infty} \cup_{j\ge n} A_j$. *That is, the* $\limsup$ *is the set* $\{\omega \in \Omega : \omega \in$ *infinitely many* $A_j\}$. *For obvious reasons, this event is also denoted by* $\{A_j$ *ifinitely often*$\}$ *or* $\{A_j$ *i.o.*$\}$. *Similarly, the complement of this event is also denoted by* $\{A_j$ *f.o.*$\}$ *(f.o. for finitely often).*
2. $\liminf_{j\to\infty} A_j := \cup_{n=1}^{\infty} \cap_{j\ge n} A_j$. *That is,* $\liminf$ *is the set* $\{\omega \in \Omega : \omega \in$ *all* $A_j$*'s from some* $j$ *onwards*$\}$.

Note that any collection of subsets of $\Omega$ is a subset of the power set. This leads us to the following definition:

**Definition 1.1.5.** *Let $\mathcal{G}$ be a collection of subsets of $\Omega$. The intersection of all $\sigma$-algebras containing $\mathcal{G}$ is called the $\sigma$-algebra generated by $\mathcal{G}$ and is denoted by $\sigma(\mathcal{G})$.*

Sometimes we refer to $\sigma(\mathcal{G})$ as the minimal $\sigma$-algebra or smallest $\sigma$-algebra containing $\mathcal{G}$. This leads us to some very important $\sigma$-algebras.

**Definition 1.1.6.**

- *Let $\mathcal{T}$ be a topology on $\Omega$ (if you're not familiar with the notion, suppose that we have a metric on $\Omega$ and $\mathcal{T}$ is the collection of all open sets). The Borel $\sigma$-algebra (corresponding to $\mathcal{T}$) is $\sigma(\mathcal{T})$.*
- *Specifically, we write $\mathcal{B}(\mathbb{R}^n)$ for the Borel $\sigma$-algebra on $\mathbb{R}^n$ when the latter is equipped with the standard Euclidean metric.*

**Definition 1.1.7.** *Let $\mathcal{F}_1$ and $\mathcal{F}_2$ be $\sigma$-algebras on the sample spaces $\Omega_1$ and $\Omega_2$, respectively. The product $\sigma$-algebra is the $\sigma$-algebra on thev product space $\Omega_1 \times \Omega_2$ generated by the sets of the form $A \times B = \{(\omega_1, \omega_2) \in \Omega_1 \times \Omega_2 : \omega_1 \in A, \omega_2 \in B\}$.*

**Exercise 1.1.3.** *Extend Definition 1.1.7 to a product of three $\sigma$-algebras.*

**Exercise 1.1.4.** *Show that $\mathcal{B}(\mathbb{R}^2) = \mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathbb{R})$.*

There are also infinite products. In fact, much of our work will be in these measure spaces. Let's introduce one that will serve as faithfully and is perhaps the most important one, as nearly any result in probability theory can be realized on it.

**Example 1.1.1** (Infinite Coin Tossing)**.** *Though traditionally, the outcomes of tosses are labelled as Heads or Tails, we will use 0 and 1. We introduce an important measurable space that will follow us a very long way.*

*Let $\Omega = \{0,1\}^{\mathbb{N}}$, that is elements in $\Omega$ are sequences of the form $\omega = (\omega_1, \omega_2, \dots)$, where $\omega_j \in \{0,1\}$. We will refer to $\omega_j$ as the $j$-th coordinate of the infinite sequence $\omega$, and we refer to the function $\omega \ni \Omega \to \omega_j \in \{0,1\}$ as the $j$-th coordinate mapping.*

*We now introduce an algebra on this space, the algebra generated by the coordinate mappings, the smallest algebra (intersections of algebra is also an algebra - there is less to check than for $\sigma$-algebras!) which contained all sets determined by a single coordinate, namely all events of the form $\{\omega \in \Omega : \omega_j \in A\}$ for some $A \subseteq \{0,1\}$, when ranging over all $j \in \mathbb{N}$. There is an explicit description of this algebra, which we now provide. For every $n$, let $A_1, \dots, A_n$ be each subsets of $\{0,1\}$ and define the corresponding cylinder set $I_{A_1,\dots,A_n}$:*

$$I_{A_1,\dots,A_n} := \{\omega \in \Omega : \omega_1 \in A_1, \dots, \omega_n \in A_n\}.$$

*Now let $\mathcal{A}$ be the set of finite unions of cylinder sets. Clearly $\Omega \in \mathcal{A}$. Also, by construction $\mathcal{A}$ is closed under unions. What about complements? First observe that the intersection of any two cylinder sets is a cylinder set. Next, every cylinder set is a finite intersection of cylinder sets, each determined by a single coordinate. For example $I_{A_1,A_2} = I_{A_1} \cap I_{\{0,1\},A_2}$. Note that the complement of a cylinder set involving a single coordinate is also such a set. Namely $I_{A_1}^c = I_{A_1^c}, I_{\{0,1\},A_2}^c = I_{\{0,1\},A_2^c}$, etc. Therefore by deMorgan's laws, the complement of any cylinder set is the finite union of cylinder sets* *each involving one coordinate. Using this and the distributive properties of intersections of unions, we conclude that the complement of any element in $\mathcal{A}$ is... in $\mathcal{A}$. Done. In words (and you are welcome to make this into a precise mathematical statement. It is an excellent exercise to identify the meaning of "depenends" in the following statement): a subset of $\Omega$ is an element of $\mathcal{A}$ if and only if it depends on a finite number of coordinates.*

*The infinite product $\sigma$-algebra $\mathcal{F}$ is $\sigma(\mathcal{A})$. This is also the $\sigma$-algebra generated the coordinate mappings (all sets determined by a single coordinate). Why?*

*What the difference between $\mathcal{A}$ and $\mathcal{F}$? A world of difference, because the latter allows us to make statements about the entire sequence, not only a finite number of coordinates. Word of warning, $\mathcal{F}$ is not the power set. Again: there are sets not in $\mathcal{F}$, but due to the extensive reach of iterations of unions, intersections and complements, we need the axiom of choice to find such sets. We will show how later.*

*Consider all of the following statements, each being an example of an element in $\mathcal{F}$ but not in $\mathcal{A}$. To do this, let's define a sequence of elements in $\mathcal{A}$: $A_j = \{\omega : \omega_j = 1\}$, $j \in \mathbb{N}$. All are events determined by a single coordinate. Hence $A_j \in \mathcal{A}$. Now, let's consider all of the following statements:*

- *"1 appears", namely $\cup_{j=1}^{\infty} A_j$. This is a countable union of elements in $\mathcal{A}$. Its complement is pretty boring. It consists of one element the constant sequence $\overline{0}$, the sequence $(0, 0, \dots)$. Clearly $\{\overline{0}\}$ is not a finite union of cylinders, because it has one element, and every nonempty cylinder has infinitely many elements.*

- *"1 appears infinitely many times". This is, by definition (Definition 1.1.4), $\limsup_{j\to\infty} A_j$ or $\{\omega_j = 1 \text{ i.o.}\}$. Why? Let's rewrite what we said in words: For every $n \in \mathbb{N}$, there exists some $j \geq n$ such that $\omega_j = 1$. In terms of set operations, this set is*

$$A := \cap_{n=1}^{\infty} \cup_{j=n}^{\infty} A_j = \limsup_{j\to\infty} A_j.$$

- *"The proportion of 1's in the first $n$ coordinates tends to a limit as $n \to \infty$". This one is harder to express through set operations (yet completely possible), and we will develop other methods to show it is an event when we discuss random variables.*

As a closing comment (or exercise for you), a $\sigma$-algebra is either finite or uncountable. It can never be countably infinite (to prove, consider an infinite $\sigma$-algebra. Then you have an countable sequence of disjoint events, from which we can form an uncountable collection of sets through countable unions). Therefore, though we can describe some or many elements in an infinite $\sigma$-algebra, we can rarely describe all elements in the $\sigma$-algebra.

## 1.2 Measures and Probability measures

We are ready to introduce the main subject of this chapter. By now we know what events are. We now present the mathematical notion formalizing chance. It is pretty mundane: a function. Basically a mathematical scale, assigning any event a "weight", the probability of the event, representing how "likely" the event is, in a scale of $[0,1]$. We will define the slightly more general notion of a measure first.

In this section we assume that $(\Omega, \mathcal{F})$ is a measurable space.

**Definition 1.2.1.** *A measure is a function $\mu : \mathcal{F} \to [0,\infty]$ which has the following properties:*

1. $\mu(\emptyset) = 0$
2. *($\sigma$-additivity) If $A_1, A_2, \dots$ are disjoint events, $\mu(\cup_{j=1}^{\infty} A_j) = \sum_{j=1}^{\infty} \mu(A_j)$.*

*A measure is called a finite measure if $\mu(\Omega) < \infty$. A measure is called a probability measure if $\mu(\Omega) = 1$. The triple $(\Omega, \mathcal{F}, \mu)$ is called a measure space, or a probability space if $\mu$ is a probability measure.*

Note that the $\sigma$-additivity requires events to be disjoint, which is not really an obstacle, due to Proposition 1.1.1: any union can be partitioned to a disjoint union through set operations.

Note that if we just wanted to define a probability measure we could replace the first item in the definition by $\mu(\Omega) = 1$ (why?).

In the sequel we will use mostly $P$ and $Q$ for probability measures.

As you see, measures are functions whose domain is a $\sigma$-algebra. Why restrict to $\sigma$-algebra and not work with the entire power set? Sometimes we can use the power set (most often when $\Omega$ is finite), but when $\Omega$ is infinite, and we want additional properties from the measure (analogy: not "just" a continuous function, but a continuously differentiable function), then almost always there will be some serious obstacles which will not allow to define the measure on the power set. We will get to that later. Time for some examples.

**Example 1.2.1.** *Suppose that $\Omega$ is uncountable. Let $\mathcal{F}$ be the $\sigma$-algebra generated by the finite sets. Then an element of $\mathcal{F}$ is a set which is either countable or has a countable complement. On $\mathcal{F}$ define:*

1. $\mu(A) = \begin{cases} 0 & A \text{ is countable;} \\ \infty & \text{otherwise} \end{cases}$. *This is an infinite measure.*
2. $P(A) = \begin{cases} 0 & A \text{ is countable;} \\ 1 & \text{otherwise} \end{cases}$. *This is a probability measure.*

**Exercise 1.2.1.** *Show that the functions defined in Example 1.2.1 are as claimed.*

**Example 1.2.2.** *Uniform measure. Let $\Omega$ be finite, equipped with the power set. Define $P(A) = \frac{|A|}{|\Omega|}$. Then $P$ is a probability measure known as the uniform measure.*

**Example 1.2.3.** *Discrete probability measure. Let $\Omega$ be countable and $\mathcal{F} = 2^{\Omega}$, and let $p : \Omega \to [0,1]$ satisfy $\sum_{\omega} p(\omega) = 1$. Define $P(A) = \sum_{\omega \in A} p(\omega)$. The uniform measure is a special example of a discrete probability measure.*

**Example 1.2.4.** *Dirac's Delta measure. Let $x_0 \in \mathbb{R}^n$. On the measurable space $(\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n))$, define*

$$P(A) = \begin{cases} 1 & x_0 \in A \\ 0 & \text{otherwise} \end{cases}$$

**Exercise 1.2.2.** *Show that the object defined in Example 1.2.4 is a probability measure.*

**Exercise 1.2.3.** *Let $\mu_1$ and $\mu_2$ be measures on a measurable space.*

1. *Show that if $c \geq 0$, then $c\mu_1$ and $\mu_1 + \mu_2$ are also measures (in fact, an infinite sum of measures is a measure, if you want to challenge yourself. We'll be able to show this easily as an application of the Monotone Convergence Theorem later).*
2. *Is $\max(\mu_1, \mu_2)$ necessarily a measure?*

### 1.2.1 Basic Properties of Measures

**Proposition 1.2.1.** *Let $\mu$ be a measure. Then*

1. *(Monotonicity) For events $A \subseteq B$, we have $\mu(A) \leq \mu(B)$.*
2. *(Subadditivity) If $A_1, A_2, \dots$ are events, then $\mu(\cup_{j=1}^{\infty} A_j) \leq \sum_{j=1}^{\infty} \mu(A_j)$.*
3. *(Continuity from below) If $A_1 \subseteq A_2 \subseteq \cdots$ are events, then $\mu(\cup_{j=1}^{\infty} A_j) = \lim_{j\to\infty} \mu(A_j)$.*
4. *(Continuity from above) If $B_1 \subseteq B_2 \subseteq \cdots$ are events and $\mu(B_1) < \infty$, Then $\mu(\cap_{j=1}^{\infty} B_j) = \lim_{j\to\infty} \mu(B_j)$.*

*Proof.* Exercise. □

**Proposition 1.2.2.** *Let $P$ be a probability measure. Then*

1. *$P(A^c) = 1 - P(A)$.*
2. *(Inclusion/Exclusion) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$.*

*Proof.* Exercise. □

**Exercise 1.2.4.** 1. *State and prove the Inclusion/Exclusion Principle: the extension of Proposition 1.2.2-2 for a finite number of events.*

2. *State and prove Bonferroni inequalities (do some research).*

Next we present ways to define new measures from existing measures. There are multiple ways, but at this stage we present only two.

**Definition 1.2.2.** *Let $F \in \mathcal{F}$. The restriction of the measure $\mu$ to $F$, denoted by $\mu|_F$ is the measure defined as follows:*

$$\mu|_F(A) := \mu(A \cap F),\ A \in \mathcal{F}.$$

**Exercise 1.2.5.** *Show that $\mu|_F$ defined above is indeed a measure.*

**Definition 1.2.3.** *Let $P$ be a probability measure and let $F \in \mathcal{F}$ satisfy $P(F) > 0$. The measure $Q$, defined through*

$$Q(A) := \frac{P(A \cap F)}{P(F)},$$

*is a probability measure called $P$ conditioned on $F$. It is denoted by $P(\,\cdot\,|F)$ (that is we write $P(A|F)$ instead of $Q(A)$).*

Conditional probabilities are the mathematical objects representing our model when some concrete information is given. Let's look at a simple example. Consider the uniform probability measure on $\{1, 2, \dots, 6\}$, a roll of a dice. The probability of any number is $\frac{1}{6}$. If I tell you that the the dice was rolled and landed on an even number, then with this new information, the probabilities are different. The probability of $\{1, 3, 5\}$ is now zero, and the probability of each of $2, 4, 6$ is $\frac{1}{3}$. Given this extra information, the set $B = \{2, 4, 6\}$, the probability of each event is now $P(A|B)$.

### 1.2.2 Completion of a measure

To avoid "going out of bounds" when taking limits some types of limits, we need the following technical but important extension result. It is a "set and forget" type of thing.

**Definition 1.2.4.** *A measure space $(\Omega, \bar{\mathcal{F}}, \bar{\mu})$ is complete if for each $N \in \bar{\mathcal{F}}$ satisfying $\bar{\mu}(N) = 0$, all subsets of $N$ are in $\bar{\mathcal{F}}$.*

We will now show how to start from a measure space $(\Omega, \mathcal{F}, \mu)$ and extend it to a complete measure space. A null event is an event $N$ satisfying $\mu(N) = 0$. Let $\mathcal{N}$ be the collection of null events in $\mathcal{F}$.

We start by defining

$$\bar{\mathcal{F}} = \{A \cup L : A \in \mathcal{F}, L \subseteq N \text{ for some } N \in \mathcal{N}\}.$$

**Proposition 1.2.3.** *1. $\bar{\mathcal{F}}$ is a $\sigma$-algebra which contains $\mathcal{F}$.*

*2. Every subset of a null event is in $\bar{\mathcal{F}}$.*

*Proof.* Clearly $\Omega = \Omega \cup \emptyset \in \bar{\mathcal{F}}$. Next, for $A \in \mathcal{F}$, $(A \cup L)^c = A^c \cap L^c$. Now $L \subseteq N$ for some null event $N$. So $N^c \subset L^c$. We can therefore write

$$A^c \cap L^c = (A^c \cap N^c) \cup (A^c \cap (L^c \cap N)).$$

The first insercction is in $\mathcal{F}$ and the second is a subset of $N$, therefore we have an element in $(A \cup L)^c \in \bar{\mathcal{F}}$. The fact that $\bar{\mathcal{F}}$ is closed under countable unions is trivial.

The second assertion follows from the definition of $\bar{\mathcal{F}}$. □

Next define $\bar{\mu}$ a function on $\bar{\mathcal{F}}$ through $\bar{\mu}(A \cup L) := \mu(A)$.

**Exercise 1.2.6.** *Show that $\bar{\mu}$ is a measure.*

We conclude with what seems to be the obvious conclusion:

**Proposition 1.2.4.** *The measure space $(\Omega, \bar{\mathcal{F}}, \bar{\mu})$ constructed above is a complete measure space.*

We call the measure space constructed above the completion of $(\Omega, \mathcal{F}, \mu)$, and refer to $\bar{\mu}$ as the completion of $\mu$. Because we can easily complete every measure space, in the sequel - unless specifically indicated otherwise - we will assume that all measure spaces are complete.

### 1.2.3 The Lebesgue Measure and the Infinite Product Measure

In this section we record the existence and some basic properties of two very important measures. The proofs will not be given in this course.

We begin with the Lebesgue measure.

**Theorem 1.2.1.** *There exists a unique measure $m$ on $\mathcal{B}(\mathbb{R})$ satisfying $m([a, b]) = b - a$ for all $a < b$.*

The completion of this measure is called the Lebesgue measure. We will abuse notation and use $m$ for the Lebesgue measure as well. The domain of the Lebesgue measure is called the Lebesgue $\sigma$-algebra, and is denoted by $\mathcal{L}$.

The Lebesgue measure will be very useful for integration on the real line and beyond. You'll see. This is the single most important measure.

**Exercise 1.2.7.** *1. What is the Lebesgue measure of a countable set?*

*2. Show that for any bounded interval with endpoints $a \leq b$ (open.closed,clopen), the Lebesgue measure of the interval is equal to $b - a$.*

*3. What is the Lebesgue measure of all irrational numbers in the interval $[a, b]$? What about $(a, b)$?*

With the Lebesgue measure introduced it would be easier (and useful) to present another method to generate measures from the ones we discussed before. A function $f : \mathbb{R} \to \mathbb{R}$ is called Borel measurable if for every $B \in \mathcal{B}(\mathbb{R})$, $f^{-1}(B) \in \mathcal{B}(\mathbb{R})$. That is the pre-images of Borel sets are... Borel sets.

For any function $f : \mathbb{R} \to \mathbb{R}$, the collection of sets $\mathcal{F}_f := \{B : f^{-1}(B) \in \mathcal{B}(\mathbb{R})\}$ is a $\sigma$-algebra (why?).

1. Recall that $f$ is continuous then the pre-image of any open interval is open, and therefore $\mathcal{F}_f$ contains all open intervals, and so $\mathcal{F}_f \supseteq \mathcal{B}(\mathbb{R})$ which also implies that $f$ is Borel measurable.

2. If $f$ is piecewise continuous (continuous possibly except for finitely many jumps) then a short argument shows that it is also Borel measurable.

3. If $f$ is monotone (nondecreasing or non-increasing), then the pre-image of any interval is always an interval, and so $\mathcal{F}_f \subseteq \mathcal{B}(\mathbb{R})$, and, yet again: $f$ is Borel measurable.

**Definition 1.2.5** (Pushforward measure)**.** *Let $\mu$ be a Borel measure on $\mathbb{R}$. Let $f : \mathbb{R} \to \mathbb{R}$ Borel-measurable. The pushforward measure $\mu_f$ is a Borel measure defined through $\mu_f(A) := \mu(f^{-1}(A)) = m(\{x \in \mathbb{R} : f(x) \in A\}$.*

**Exercise 1.2.8.** *Show that $\mu_f$ is a measure.*

**Example 1.2.5.** *Consider the restriction of $m$ to the interval $(0,1)$. This is a Borel probability measure.*

1. *Let $f(x) = \mathbf{1}_{[0,p]}$. The pre-image of $0$ under $f$ is the interval $(p,1)$ whose measure is $1-p$, and the pre-image of $1$ is the interval $(0,p)$ whose measure is $p$. Therefore, $m_f$ is the measure on all subsets of $\{0,1\}$ assigning measure $p$ to $1$ and $1-p$ to $0$. Looks familiar? (Hint: Bernoulli random variable).*

2. *Let $f(x) = \ln \frac{1}{1-x}$. Then for any $y \in (0,1)$, the pre-image of the interval $(0,y]$ is $\{x : \ln \frac{1}{1-x} \le y\} = \{x : \frac{1}{1-x} \le e^y\} = \{x : x < 1 - e{-}y\}$, and therefore has measure $1-e^{-y}$. Thus $m_f$ is a measure on the Borel subsets of $(0,\infty)$ satisfying $m_f(\{x : x \le y\}) = 1 - e^{-y}$. Is that familiar? (Hint: Exponential random variable).*

We continue to an infinite product measure.

**Theorem 1.2.2** (Infinite Product Measure)**.** *Consider the infinite product space $(\Omega, \mathcal{F})$ of Example 1.1.1, of infinite coin tosses.*

*Let $p \in (0,1)$. There exists a unique probability measure $P$ on $\mathcal{F}$ satisfying that for every $n \in \mathbb{N}$ and $w_1, \dots, w_n \in \{0,1\}$: $P(\{\omega : \omega_1 = w_1, \dots, \omega_n = w_n\}) p^{\sum_{j=1}^n w_j}(1-p)^{n-\sum_{j=1}^n w_j}$,*

As mentioned before, we can think of each element in $\Omega$ as of an infinite sequence of coin tosses. The probability measure thus defined has the following properties:

**Proposition 1.2.5.**

1. *For every $j$, $P(\{\omega : \omega_j = 1\}) = p = 1 - P(\{\omega : \omega_j = 0\})$. That is the probability of each toss being $1$ is $p$ and being $0$ is $1-p$.*

2. *For every cylinder set $I_{A_1,\dots,A_n}$, $P(I_{A_1,\dots,A_n}) = \prod_{j=1}^n P(I_{A_j})$, that is, probabilities of cylinder events are products of the probabilities of the individual statements on each coordinate.*

Why can't we define the infinite product measure on the power set of $\Omega$? Here's why. This is a sketch, not a complete proof. First, we will make our life a little simpler by replacing the underlying sample space by $\{0,1\}^{\mathbb{Z}}$, all infinite sequences of zeros and ones, indexed by the integers. An element in this space is of the form $(\dots, \omega_{-1}, \omega_0, \omega_1, \dots)$. We can define the product $\sigma$-algebra and the product measure on this space similarly to how we did it for $\{0,1\}^{\mathbb{N}}$, as the $\sigma$-algebra generated by cylinder sets. We omit the simple details. Let's consider the case $p = \frac{1}{2}$ (this does not matter). Define $S : \{0,1\}^{\mathbb{Z}} \to \{0,1\}^{\mathbb{Z}}$ to be the right shift: $(S\omega)_j = \omega_{j-1}$. Observe that by construction, $P$ is shift invariant: $P(SA) = P(A)$ for all $A$ which are cylinder sets. You will be able to show later (Exercise 1.3.1) that this property holds for all $A \in \mathcal{F}$. We will now show that there is a set not in $\mathcal{F}$. We argue by contradiction, assuming $\mathcal{F}$ is the power set. Write $S^{-1}\omega$ for the left shift (inverse of right shift), and more generally $S^j\omega$ for the shift of $\omega$ $|j|$ times to the right or left according to whether $j > 0$ or $j < 0$. Of course $S^0\omega$ is $\omega$.

Clearly $P(\{\omega\}) = 0$ (why?). For every element $\omega \in \{0,1\}^{\mathbb{Z}}$, let $[\omega] := \{S^j\omega : j \in \mathbb{Z}\}$, be the set of all shifts of $\omega$. The set $[\omega]$ is finite if and only if $\omega$ is periodic. As there are only countably many periodic elements in our uncountable sample space, the probability the union of set $[\omega]$ which are finite is zero. Note that for every two elements $\omega, \omega'$, either $[\omega] = [\omega']$ or $[\omega] \cap [\omega'] = \emptyset$. Now pick one exactly one element from each of the distinct infinite sets $[\omega]$ (this requires the axiom of choice) and form a set from all these elements (one from each $[\omega]$). Call this set $A$. Every nonperiodic element in our sample space is in the union $\cup_{j\in\mathbb{Z}} S^jA$. Clearly, for $j \ne j'$, $S^jA \cap S^{j'}A = \emptyset$, and by our assumption, $P(S^jA) = P(A)$. Combining these observations, $1 = P(\cup_{j\in\mathbb{Z}} S^jA) = \sum_{j\in\mathbb{Z}} P(S^jA) = \sum_{j\in\mathbb{Z}} P(A)$. Oh shift! That's a contradiction. To what? the assumption that $\mathcal{F}$ is the power set. The set $A$ is therefore not in $\mathcal{F}$. We can't really describe concretely, but it's out there in the wild.

Bottom lime: we can't construct a probability measure which is both shift invariant and is defined on the power set. Bummer. Have to stick to $\sigma$-algebras.

Let's look at a nice result we will later complement with a partial converse and plays an important role in probability.

**Proposition 1.2.6** (Borel-Cantelli I Lemma)**.** *Let $\mu$ be a measure. Suppose that $A_1, A_2, \dots$ are events satisfying $\sum_{j=1}^\infty \mu(A_j) < \infty$. Then $\mu(A_j \text{ i.o.}) = 0$.*

*Proof.* For $n \in \mathbb{N}$, let $B_n := \cup_{j\ge n} A_j$. Then $B_1 \subseteq B_2 \subseteq \cdots$, and by the subadditivity property,

$$\mu(\cup_{j\ge n} A_j) \le \sum_{j=n}^{\infty} \mu(A_j) < \infty.$$

Since $\mu(B_1) < \infty$, it follows from continuity from above, that

$$\mu(\limsup A_j) = \mu(\cap_{n=1}^{\infty} B_n) = \lim_{n\to\infty} \mu(B_j) \le \lim_{n\to\infty} \sum_{j=n}^{\infty} \mu(A_j) = 0,$$

where the last equality follows from the fact that the tail of a convergent series tends to zero. □

**Example 1.2.6.** *Consider the infinite product measure of Theorem 1.2.2 with $p = \frac{1}{2}$. For each $n$, let $A_n$ be initial segmnet of the sequence of length $n$ is repeated from the $n+1$-th coordinate. This is better explained in symbols:*

$$A_n := \{\omega : \omega_{n+k} = \omega_k : k = 1, \dots, n\}.$$

*We have $P(A_n) = 2^n$, and therefore the Borel-Cantelli lemma $P(A_n \text{ i.o.}) = 0$.*

*Let's remove the restriction of repeating the initial segment from the $n+1$-th coordinate, and define instead $B_n$ as the event that the initial segment of length $n$ is repeated infinitely(!) many times (anywhere along the sequence). We have $P(B_n) = 1$, and therefore $P(\cap_{n=1}^{\infty} B_n) = 1$. Thus, each initial segment is repeated infinitely many times a.s.*

*Patience pays?*

**Exercise 1.2.9.** *Prove the second assertion in Example 1.2.6.*

## 1.3 Monotone Class and $\Pi - \Lambda$ Theorems

Time for our first theorem. As $\sigma$-algebras are often elusive, it may be a good idea to find tools that will allow us to extend results from smaller and easier to manage collections of sets to an entire $\sigma$-algebra. One such tool is the monotone class theorem. To understand the need for such a tool, consider Theorem 1.2.1. It tells us that there exists a measure on $\mathcal{B}(\mathbb{R})$ with a certain property: the measure of an interval is its length. It also says that there's only one such measure. But why? $\mathcal{B}(\mathbb{R})$ is much more than just intervals.

**Definition 1.3.1.** *A collection $\mathcal{M}$ of subsets of $\Omega$ is a monotone class if both of the following hold:*

1. *If $A_1 \subseteq A_2 \subseteq \dots$ are elements in $\mathcal{M}$, then $\cup_j A_j \in \mathcal{M}$.*
2. *If $A_1 \supseteq A_2 \supseteq \dots$ are elements in $\mathcal{M}$, then $\cap_j A_j \in \mathcal{M}$.*

In words, a monotone class is a collection of subsets which is closed under monotone limits. Every $\sigma$-algebra is a monotone class.

**Theorem 1.3.1** (Monotone Class Theorem)**.** *Let $\mathcal{A}$ be an algebra and let $\mathcal{M}$ be a mononone class satisfying $\mathcal{A} \subset \mathcal{M}$. Then $\sigma(\mathcal{A}) \subset \mathcal{M}$.*

Before moving to the proof (I think you'll appreciate its beauty), let's consider an application.

**Proposition 1.3.1.** *Let $\mathcal{A}$ be an algebra on a sample space $\Omega$. Suppose that $P$ and $Q$ are probability measures on $\sigma(\mathcal{A})$, which satisfy $P(A) = Q(A)$ for all $A \in \mathcal{A}$. Then $P = Q$.*

*Proof.* Let $\mathcal{M} = \{B \in \sigma(\mathcal{A}); P(B) = Q(B)\}$. Then $\mathcal{M}$ is a monotone class due to the continuity of probability measures with respect to monotone limits. Also, by assumption $\mathcal{A} \subseteq \mathcal{M}$. The result follows immediately from the Monotone Class Theorem. □

As a practical application, let's consider the restriction of the Lebesgue measure to the interval $[0, 1)$, and define the algebra consisting of all finite unions of sets of the form $[a, b)$ for $0 \le a < b \le 1$. It is easy to see that this is indeed an algebra: it is not empty, it is closed under finite unions, as for the complements, draw a picture, then write a proof: they're also of this form. The restriction of the Lebesgue measure to $[0, 1)$ is a probability measure, and the uniqueness statement holds: there is no other way to form a measure on the Borel subsets of $[0, 1)$ which assigns each interval its length. You can push this argument a little further to prove the uniqueness of the Lebesgue measure on $\mathbb{R}$. Give it a try! The key idea is to exhaust each unbounded set by bounded sets increasing to it.

*Proof.* As with $\sigma$-algebras the intersection of monotone classes all containing some non empty collection of subset of $\Omega$ is by itself a monotone class. To avoid noise, there is no loss of generality assuming that $\mathcal{M}$ is the intersection of all monotone classes containing $\mathcal{A}$, informally, the monotone class generated by $\mathcal{A}$, or the "smallest" monotone class containing $\mathcal{A}$. Remember that.

The proof continues in two steps. The first step is to identify all elements in $\mathcal{M}$ whose intersection with elements of $\mathcal{A}$ remains in $\mathcal{M}$. Formally, define

$$\mathcal{M}_1 := \{B \in \mathcal{M} : A \cap B \in \mathcal{M}, \text{ for all } A \in \mathcal{A}\}.$$

Now guess what? $\mathcal{M}_1$ is a monotone class because $\mathcal{M}$ is. Check! Oh, and $\mathcal{A}$ is an algebra, so clearly $\mathcal{A} \subset \mathcal{M}_1$. Now here's the magic. By construction $\mathcal{M}_1 \subseteq \mathcal{M}$, but by the minimality assumption on $\mathcal{M}$, necessarily $\mathcal{M} \subseteq \mathcal{M}_1$. So equality holds. What did we just prove? The intersection of any element in $\mathcal{M}$ with any element in $\mathcal{A}$ is in $\mathcal{M}$.

The next step is pushing this further. Let

$$\mathcal{M}_2 := \{B \in \mathcal{M} : A \cap B \in \mathcal{M}, \text{ for all } A \in \mathcal{M}\}.$$

Only one thing changed from the definition of $\mathcal{M}_1$: now we look at all element in $\mathcal{M}$ whose intersection with any element in $\mathcal{M}$ (not "just" $\mathcal{A}$) is again in $\mathcal{M}$. Bold move. It is no surprise that $\mathcal{M}_2$ is a monotone class. Also, we showed in the first step that the intersection of every element in $\mathcal{A}$ with any element in $\mathcal{M}$ is in $\mathcal{M}$, so necessarily $\mathcal{A} \subseteq \mathcal{M}_2$. Or, $\mathcal{M}_2$ is a monotone class containing $\mathcal{A}$. As in the first step, we conclude that $\mathcal{M}_2 = \mathcal{M}$.

Let's digest what we have just proven: the intersection of any two elements in $\mathcal{M}$ is in $\mathcal{M}$. That's it. $\mathcal{M}$ is closed under finite intersections. Nice property. What about complements? Well, I guess by now you know what we're going to do:

$$\mathcal{M}_3 := \{B \in \mathcal{M} : B^c \in \mathcal{M}\}.$$

Yes. This is a monotone class (this explains why we require both monotonicity in both directions, right?), and it contains $\mathcal{A}$. The minimality of $\mathcal{M}$ dictates that $\mathcal{M}_3$ is $\mathcal{M}$.

Let's record a few things we know by know about $\mathcal{M}$. It (i) contains $\Omega$ (because $\mathcal{A}$ does); (ii) is closed under complements; (iii) is closed under finite intersections. Of course, with (i) and (ii) (and deMorgan's laws), (iii) is equivalent to "closed under finite unions" Sounds familiar? Yes, $\mathcal{M}$ is an algebra. To show that is is a $\sigma$-algebra, it remains to show it is closed under countable unions. But this is obvious: Let $A_1, A_2, \dots$ be elements in $\mathcal{M}$. For every $n$, the finite union $B_n := \cup_{j=1}^n A_j$ is in $\mathcal{M}$. Of course, $B_1 \subseteq B_2 \subseteq \dots$. Therefore the union $\cup_{j=1}^\infty B_j = \cup_{j=1}^\infty A_j$ is in $\mathcal{M}$.

Time to finish: $\mathcal{M}$ is a $\sigma$-algebra. It contains $\mathcal{A}$. Because every $\sigma$-algebra is a monotone class, the minimality of $\mathcal{M}$ dictates that $\mathcal{M} \subset \sigma(\mathcal{A})$. But hey, $\sigma(\mathcal{A})$ is also minimal (intersection of all $\sigma$-algebras containing $\mathcal{A}$). So $\mathcal{M} = \sigma(\mathcal{A})$. □

**Exercise 1.3.1.** *Consider the probability space presented in the discussion below Proposition 1.2.5. Show that*

1. *$SA \in \mathcal{F}$ for all $A \in \mathcal{F}$.*
2. *$P(SA) = P(A)$ for all $A \in \mathcal{F}$.*

Careful inspection of the monotone class theorem, leads to the proof of another (and perhaps even more useful) theorem. We had an algebra $\mathcal{A}$ contained in a monotone class $\mathcal{M}$. In the first two steps of the proof, we only used the fact that an algebra is closed under finite intersections. This suggests that we may be able to leverage the ideas to prove a different version in which some of the assumptions on $\mathcal{A}$ are passed to over $\mathcal{M}$. Specifically, we define the following:

**Definition 1.3.2.** 1. *A collection $\mathcal{P}$ of subsets of $\Omega$ is called a $\pi$-system if it has at least one element, and for any $A, B \in \mathcal{P}$, $A \cap B \in \mathcal{P}$.*

2. *A collection $\mathcal{L}$ of subsets of $\Omega$ is called a $\lambda$-system if*
   - *(a) $\Omega \in \mathcal{L}$*
   - *(b) (Closure under relative complements) If $B, A \in \mathcal{L}$ and $A \subseteq B$, then $B - A \in \mathcal{L}$*
   - *(c) If $A_1 \subseteq A_2 \subseteq \dots \in \mathcal{L}$, then $\cup_{j=1}^\infty A_j \in \mathcal{L}$.*

So a $\pi$-system requires less than an algebra (and every algebra is a $\pi$-system), and a $\Lambda$ system requires more than a monotone class (every $\lambda$-system is a monotone class, why?). Following arguments we've seen before, if $\mathcal{P}$ is a $\pi$-system, the intersection of all $\lambda$-system containing it is also a $\lambda$-system, the minimal $\lambda$-system containing $\mathcal{P}$. We have the following:

**Theorem 1.3.2** ($\pi - \lambda$ Theorem)**.** *Let $\mathcal{P}$ be a $\pi$-system and let $\mathcal{L}$ be a $\lambda$-system with $\mathcal{P} \subseteq \mathcal{L}$. Then $\sigma(\mathcal{P}) \subseteq \mathcal{L}$.*

*Proof.* This is a sketch. Without loss of generality $\mathcal{L}$ is the minimal $\lambda$-system containing $\mathcal{P}$. Repeating the argument in the first two steps of the proof of the Monotone Class Theorem 1.3.1 (those involving $\mathcal{M}_1, \mathcal{M}_2$), we have that $\mathcal{L}$ is closed under finite intersections. Since by assumption $\mathcal{L}$ contains $\Omega$ and is closed under complements, all that remains to complete the proof (showing that it is a $\sigma$-algebra) is to show that if $(A_j : j \in \mathbb{N})$ is a sequence in $\mathcal{L}$, then $\cup_{j=1}^\infty A_j \in \mathcal{L}$. Thus, if $A, B \in \mathcal{L}$ we have that $A \cup B = (A^c \cap B^c)^c \in \mathcal{L}$. Now look at our sequence $(A_j : j \in \mathbb{N})$. As $A_j \in \mathcal{L}$, so is every finite union $\cup_{j=1}^n A_j$, and by the third assumption we have that $\cup_{j=1}^\infty A_j \in \mathcal{L}$. □

**Exercise 1.3.2.** *Review the proof and find where we use the second assumption (relative complements) in the definition of a $\lambda$-system (it's hidden!)*

## 1.4 Independence

In this section we fix a probability space $(\Omega, \mathcal{F}, P)$.

Independence is one of the most important notions in probability theory. It allows to do (many) computations and prove the classical limit theorems. Some would even go as far to say that it is what distinguishes probability from measure theory. So what is it all about? Actually, products. Let's begin with the classical definition you're probably familiar with. Then delve into a more general definition.

**Definition 1.4.1.** *Let $I$ be an index set and let $(A_i : i \in I)$ be events. We say that these events are independent of for any finite subset $J$ of $I$, $P(\cap_{j\in J} A_j) = \prod_{j\in J} P(A_j)$.*

To see some examples of independent events, let's take our infinite product space from Theorem 1.2.2. Let $I = \mathbb{N}$, and for $i \in \mathbb{N}$, let $A_i$ be an event determined by the $i$-th coordinate of $\omega$, $\omega_i$. Then by construction, $(A_i : i \in \mathbb{N})$ are independent. This can be recast as: the individual coin tosses in this model are independent.

When looking at only two events, say $A$ and $B$, independence is equivalent to the following: either $P(B) = 0$ or $P(A|B) = P(A)$. Note that independence is symmetric in $A$ and $B$, so we can swap the roles. Either way, "knowing" $B$ does not alter the probability of $A$.

Note that in general we can have $A$ and $B$ independent, $B$ and $C$ independent, $A$ and $C$ independent, but not $A, B, C$. So this definition really requires looking at all finite intersections, not just intersections of pairs.

It is a simple exercise to show that if $(A_i : i \in I)$ are independent then the same applies to $(B_i : i \in I)$ where for each $i$, $B_i = A_i$ or $B_i = A_i^c$, with the choices made arbitrarily.

Let's prove a partial converse to the Borel-Cantelli I Lemma 1.2.6.

**Proposition 1.4.1** (Borel-Cantelli II)**.** *Suppose that $A_1, A_2, \dots$ are independent events satisfying $\sum_{j=1}^\infty P(A_j) = \infty$. Then $P(\{A_j \text{ i.o}\}) = 1$.*

*Proof.* We will use the following inequality:

$$1 - x \le e^{-x} \quad x \in \mathbb{R}.$$

We will show that the probability of the complement of the event in question is zero. By definition of $\limsup A_j$, and deMorgan's laws, the complement is $\cup_{n=1}^\infty \cap_{j\ge n} A_j^c$. Because of subadditivity, it is enough to prove that for each $n \in \mathbb{N}$, $P(\cap_{j\ge n} A_j^c) = 0$. By continuity from above independence, the inequality above, and our assumption,

$$P(\cap_{j\ge n} A_j^c) = \lim_{m\to\infty} P(\cap_{j\ge n}^m A_j) = \lim_{m\to\infty} \prod_{j=n}^m P(A_j^c) \le e^{-\sum_{j=n}^\infty P(A_j)} = e^{-\infty} = 0.$$

□

**Example 1.4.1.** *Suppose that $A_1, A_2, \dots$ are independent events with $P(A_j) < 1$ for all $j$ and $P(\cup_{j=1}^\infty A_j) = 1$. We will show that $P(\{A_j \text{ i.o.}\}) = 1$. Clearly, $P(\cap_{j=1}^\infty A_j^c) = \prod_{j=1}^\infty (1 - P(A_j)) = 0$, and therefore $\sum_{j=1}^\infty P(A_j) = \infty$, so Borel Cantelli II holds.*

Here's a nice way to see the difference between the two Borel-Cantelli Lemmas:

**Example 1.4.2.** *Suppose we performing the following sequence of experiments. In the $n$-th experiment we pick an integer from $\{1, \dots, n\}$ uniformly, independently of previous experiments.*

1. *What is the probability that 1 will be selected infinitely often? Let $A_n$ be the event that 1 is selected in the $n$-th experiment. Then $P(A_n) = \frac{1}{n}$. By Borel-Cantelli II, we have $P(A_n \text{ i.o. }) = 1$, and so we we will see 1 infintely often a.s.*

2. *What is the probability that 1 will be repeated successively infinitely many times? Well, the event 1 is sampled at the $n$-th and the $n+1$-th times is $A_n \cap A_{n+1}$ and therefore has probability $\frac{1}{n}\frac{1}{n+1} \le \frac{1}{n^2}$. By Borel-Cantelli I, the probability of 1 being selected successively infinitely often is 0. Or: we will have only finitely many successive pairs of 1 (or any number!) a.s.*

We move to an important generalization of independence of events.

**Definition 1.4.2.** *Let $I$ be an index set and let $(\mathcal{F}_i : i \in I)$ be $\sigma$-algebras with $\mathcal{F}_i \subset \mathcal{F}$ for all $i \in I$. The family $(\mathcal{F}_i : i \in I)$ is independent if every family $(A_i : i \in I, A_i \in \mathcal{F}_i)$ is independent.*

**Example 1.4.3.** *Consider the infinite product measure of Theorem 1.2.2. For $n \in \mathbb{N}$, let $\mathcal{F}_n = \sigma(\{\omega : \omega_n = 1\})$. That is $\mathcal{F}_n$ is the $\sigma$-algebra generated by the $n$-coordinate mapping (it has four elements: all statements that can be made on the $n$-th coordinate). By construction $(\mathcal{F}_n : n \in \mathbb{N})$ are independent $\sigma$-algebras.*

Let's consider two different applications of the $\Pi - \Lambda$ theorem, Theorem 1.3.2, both of which involving independence.

**Theorem 1.4.1.** *Let $(\mathcal{P}_n : n \in \mathbb{N})$ be a sequence of independent $\Pi$-systems. That is for any $n \in \mathbb{N}$ and $A_j \in \mathcal{P}_j$, $j = 1, \dots, n$.*

$$P(\cap_{j=1}^n A_j) = \prod_{j=1}^n P(A_j).$$

*Then $(\sigma(\mathcal{P}_n) : n \in \mathbb{N})$ are independent.*

*Proof.* Since independence is determined by finite intersections, it is sufficient to show that for every $n$, the $\sigma$ algebras in the finite sequence $(\sigma(\mathcal{P}_j : j = 1, \dots, n)$ are independent. To do that, fix any $n$ and define the following:

$$\mathcal{L} = \{A \in \sigma(\mathcal{P}_1) : P(A \cap_{j=2}^n A_j) = P(A) \prod_{j=2}^n P(A_j),\ A_j \in \mathcal{P}_j,\ j = 2, \dots, n\}.$$

By assumption, $\mathcal{P}_1 \subseteq \mathcal{L}$. Also, as you may have already guessed, $\mathcal{L}$ is a $\Lambda$-system. Indeed, if $A \subseteq B$ are events in $\mathcal{L}$, then

$$P((B - A) \cap \cap_{j=2}^n A_j) = P(B \cap \cap_{j=2}^n A_j) - P(B \cap \cap_{j=2}^n A_j).$$

As $A, B \in \mathcal{L}$, and $A \subseteq B$, the righthand side is equal to $P(B - A) \prod_{j=2}^n P(A_j)$, and so $B - A \in \mathcal{L}$. Next, let $L_1 \subseteq L_2 \subseteq \dots$ be a sequence in $\mathcal{L}$. Then the monotonicity of the measure guarantees that

$$P(\cup_{k=1}^\infty L_j \cap \cap_{j=2}^n A_j) = \lim_{N \to \infty} P(L_N \cap \cap_{j=2}^n A_j) = \lim_{N \to \infty} P(L_N) \prod_{j=2}^n P(A_j) = P(\cup_{k=1}^\infty L_j) \prod_{j=2}^n P(A_j).$$

Thus, $\mathcal{L} = \sigma(\mathcal{P}_1)$, and so $\sigma(\mathcal{P}_1), \mathcal{P}_2, \dots, \mathcal{P}_n$ are indepednent. Repeat the argument for this new sequence, now working on $\mathcal{P}_2$, and iterate until the result follows. □

Here's another theorem, whose proof is essentially the same.

**Theorem 1.4.2.** *Let $(\mathcal{G}_{n,i} : n \in \mathbb{N}, i = 1, 2)$ be independent $\sigma$-algebras. Let $\mathcal{H} = \sigma(\cup_{n=1}^\infty \mathcal{G}_{n,2})$. Then $(\mathcal{H}, \mathcal{G}_{n,1} : n \in \mathbb{N})$ are independent.*

Note that without loss of generality we may assume that one or both of the sequences $(\mathcal{G}_{n,j} : n \in \mathbb{N})$ is finite by letting $\mathcal{G}_{k,j} = \{\emptyset, \Omega\}$ for all but finitely many $k$'s.

*Proof.* Let $\mathcal{P}$ be all sets which are finite intersections of elements in $\cup_{n=1}^\infty \mathcal{G}_{n,2}$. This is a $\pi$-system. Now Let $\mathcal{L}$ be the set of all elements in $\mathcal{H}$ which are independent of $(\mathcal{G}_{n,1} : n \in \mathbb{N})$. By assumption $\mathcal{P} \subseteq \mathcal{L}$. We claim $\mathcal{L}$ is a $\lambda$-system. Why? It contains $\Omega$, it is closed under increasing limits (continuity of probability with respect to monotone limits), and it is closed under relative complements. Indeed, take $n \in \mathbb{N}$ and $G_1, \dots, G_n$ satisfy $G_j \in \mathcal{G}_{j,1}$, then letting $G = \cap_{j=1}^n G_j$, and taking elements $A \subseteq B$ in $\mathcal{L}$ we have

$$P(G \cap (B - A)) = P(G \cap B) - P(G \cap A) = P(G)P(B) - P(G)P(A) = P(G)P(B - A).$$

Therefore the $\pi - \lambda$ theorem asserts that $\mathcal{L}$ contains $\sigma(\mathcal{P}) = \mathcal{H}$. □

We're about to reveal an interesting truth. We need a definition:

**Definition 1.4.3.** *Let $(\mathcal{F}_n : n \in \mathbb{N})$ be sub-$\sigma$-algebras. Define $\mathcal{F}_{n,\infty}$ as $\sigma(\cup_{k \geq n} \mathcal{F}_k)$, and let $\mathcal{T} = \cap_{n=1}^\infty \mathcal{F}_{n,\infty}$. Then $\mathcal{T}$ is called the tail $\sigma$-algebra and events in $\mathcal{T}$ are in the Tail $\sigma$-algebra.*

Let's continue Example 1.4.3. The tail $\sigma$-algebra consists of all events which do not depend on any finite set of coordinates. This may sound weird at first, but all events we introduced in Example 1.1.1 are tail events. Why? Each such event is in $\mathcal{F}_{n,\infty}$, and also the events in Example 1.2.6. What events are not in $\mathcal{T}$? Every nontrivial ($\ne \emptyset, \Omega$) cylinder event. Tail $\sigma$-algebras of independent $\sigma$-algebras have an interesting property:

**Theorem 1.4.3** (Kolmogorov's Zero-One Law)**.** *Let $(\mathcal{F}_n : n \in \mathbb{N})$ be independent $\sigma$-algebras and let $\mathcal{T}$ be the corresponding tail $\sigma$-algebra. Then $\mathcal{T}$ is degenerate: for every $A \in \mathcal{T}$, $P(A) \in \{0,1\}$.*

*Proof.* We will show that every $A \in \mathcal{T}$ is independent of itself. That is $P(A) = P(A \cap A) = P(A)^2$, or $P(A) \in \{0,1\}$.

By Theorem 1.4.2, $\mathcal{F}_{n+1,\infty}$ is independent of $\mathcal{F}_1, \dots, \mathcal{F}_n$. Since $\mathcal{T}$ is a subset of $\mathcal{F}_{n+1,\infty}$, it is independent of $\mathcal{F}_1, \dots, \mathcal{F}_n$ for every $n$ which in turn implies it is independent of $(\mathcal{F}_n : n \in \mathbb{N})$. A second application of the theorem then gives $\mathcal{T}$ is independent of $\sigma(\mathcal{F}_n : n \in \mathbb{N})$, which contains $\mathcal{T}$. □

**Example 1.4.4.** *We continue Example 1.4.3. Let $I \subseteq (0,1)$ be an interval. Define $S_n = \sum_{j \le n} \omega_j$, and let $A_I = \{\omega : \lim_{n\to\infty} \frac{S_n}{n}(\omega) \in I\}$. Thus, $A_I$ is the event that $\lim_{n\to\infty} S_n/n$ exists and is in the interval $I$. This is a tail event, and therefore $P(A_I)$ is either $1$ or $0$. No need for further calculations! As you probability know the value is determined according to whether $p \in I$ or not. We will get there soon.*

# Chapter 2

# Random Variables

Not random and not variables. Just an old name that stuck. A more fitting description would be "measurable functions". The idea: assign numerical values to elements in our sample space. This is so we can quantify outcomes (e.g. for assigning a monetary or other value to an outcome in a game of chance).

In this chapter we will introduce the mathematical notion of RVs and will present the Lebesgue integral which is the tool through which define expectations of RVs.

## 2.1 Definitions and Basic Properties

Throughout this section we'll assume $(\Omega, \mathcal{F})$ is a measurable space. In this section, we stress that we do not (and will not) use any specific probability measure to define a RV. The relation between RVs and probability measures will be presented later.

We've already seen random variables. Here. If $A$ is an event, the indicator function $\mathbf{1}_A$ is a random variable. In fact, these are the building blocks of all random variables and Lebesgue integration theory. More later.

**Definition 2.1.1.** *A function $X : \Omega \to [-\infty, \infty]$ is a random variable (RV AKA Borel measurable function) if for any Borel subset $B$ of $[-\infty, \infty]$ the preimage of $B$, $X^{-1}(B) := \{\omega : X(\omega) \in B\}$, is an event.*

The Borel $\sigma$-algebra on $[-\infty, \infty]$, denoted by $\mathcal{B}([-\infty, \infty])$ is the $\sigma$-algebra generated by the all sets of the following forms $(a, b)$, $[-\infty, b)$, $(a, \infty]$, ranging over $-\infty < a < b < \infty$ (equivalently, all intervals of the first form and the singletons $\{\pm\infty\}$). Every interval of any type is in this $\sigma$-algebra because it can be obtained by countable unions and intersections.

Point is: a given function is a RV if statements you can make about its values (Borel sets) are events.

Trivial but important: Pre-images (unlike images!) commute with set operations: $X^{-1}(A^c) = (X^{-1}(A))^c$, $X^{-1}(A \cup B) = X^{-1}(A) \cup X^{-1}(B)$. Same with intersections and countable unions and intersections.

We don't really need to go through all Borel sets to verify that a function is indeed a RV. There's an easy solution.

**Proposition 2.1.1.** *Let $X : \Omega \to [-\infty, \infty]$. Then $X$ is a RV if and only if $X^{-1}([-\infty, b]) \in \mathcal{F}$ for all $b \in \mathbb{R}$.*

*Proof.* If $X$ is an RV, then because $[-\infty, b]$ is a Borel set, it follows that the condition in the theorem holds.

Let's prove that the condition is sufficient. Let $\mathcal{G} = \{A \in \mathcal{B}([-\infty, \infty]) : X^{-1}(A) \in \mathcal{F}\}$. Clearly $\Omega \in \mathcal{G}$ (take $A = [-\infty, \infty]$), clearly $\mathcal{G}$ is closed under complements ($X^{-1}(A^c) = (X^{-1}(A))^c$), and $\mathcal{G}$ is closed under countable unions (why?). So it's a $\sigma$-algebra. It is enough to show it contains all sets of the three forms generating the Borel $\sigma$-algebra. This is done by simple manipulations. Let's go.

1. First $[-\infty, b) = \cup_{n=1}^{\infty} [-\infty, b - \frac{1}{n}]$. Therefore, $X^{-1}([-\infty, b)) = \cup_{n=1}^{\infty} X^{-1}([-\infty, b - \frac{1}{n}])$, a countable union of events.

2. Clearly, $(a,\infty] = [-\infty,a]^c$, so $X^{-1}((a,\infty]) = (X^{-1}([-\infty,a])^c$, an event.
3. Finally, $(a,b) = [-\infty,b) \cap (a,\infty]$, and so $X^{-1}((a,b))$ is the intersection of two events. Done.

□

Of course, we can replace the intervals $[-\infty,b]$ with intervals $[-\infty,b)$ or intervals $(a,\infty]$ or even restrict $a,b$ to any dense set (e.g. rationals). We just gave one choice. The idea should be clear now.

We did allow infinite values because this is the right thing to do, but in the sequel, we will often reduce the discussion to RVs taking only real values or such that may take infinite values but (after we make the connection with probability spaces) this happens with probability zero (think of the first time a coin lands Heads. Theoretically can take - literally - forever).

We will also often use less cumbersome notation. For example instead of writing $X^{-1}((a,b))$ we will write $\{\omega : X(\omega) \in (a,b)\}$ or $\{X \in (a,b)\}$ or $\{a < X < b\}$.

Indicators are so important that we will reintroduce them. If $A$ is an event, the indicator of $A$ (in other areas of analysis it may be called the characteristic function. Not a good choice in probability, as this name is already taken by some other object) is the function:

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & \omega \in A \\ 0 & \text{otherwise.} \end{cases}$$

Why is this a RV? The pre-image of any set is one of the following events: empty, $\Omega$, $A$ or $A^c$ according to whether the set does not include 0 or 1, includes 1 but not 0, includes 0 but not 1, or includes both. Simple. Indicators are also called Bernoulli RVs.

**Definition 2.1.2.** *A RV $X$ is called simple if its image $X(\Omega) = \{X(\omega) : \omega \in \Omega\}$ is a finite subset of $\mathbb{R}$.*

Of course, every indicator is a simple RV. We have the following simple proposition.

**Proposition 2.1.2.** *Let $X$ be a RV with values in $\mathbb{R}$. Then $X$ is simple if and only if it is a linear combination of indicators. Specifically, there exists $n \in \mathbb{N}$, constants $c_1,\dots,c_n \in \mathbb{R}$ and events $A_1,\dots,A_n$ so that $X = \sum_{j=1}^n c_j \mathbf{1}_{A_j}$.*

Here's how to say that using algebraic notions: the vector space of simple functions is spanned by indicators.

*Proof.* If $X$ is simple, let $\{c_1,\dots,c_n\}$ be the distinct elements in its image, and for $j \in 1,\dots,n$, let $A_j = X^{-1}(\{c_j\})$. Then $X = \sum_{j=1}^n c_j \mathbf{1}_{A_j}$.

Conversely, suppose that $X$ is of the given form. We will rearrange it in a way that will lead a a linear combination of indicators of disjoint events. This will be then much simpler. We will not write the details, but describe the algorithm. First we will look at all elements in $\cup_{j=1}^n A_j$ which are exclusive to $A_1$. This is $A_1 \cap (\cup_{j\neq 1} A_j)^c$, an event. Call it $B_1$. Then all events exclusive to $A_2,\dots,A_n$, with respective resulting events $B_2,\dots,B_n$. This step gives all elements in exactly one of the events. We continue to all elements which are in exactly two of the events, starting with all events in $A_1 \cap A_2$ but not in any other pair. More complicated, but still an event (write it down). Call this $B_{1,2}$, and continue for all other pairs. Move to all elements in exactly three, etc. Generally for $k \in 1,\dots,n$, and $1 \le i_1 < i_2 < \dots < i_k$, $B_{i_1,\dots,i_k}$ is the set of all elements which are in $A_{i_1} \cap \dots \cap A_{i_k}$ but in no other intersection of the $A_j$'s. If you're following $B_{1,\dots,n}$ is the elements in all events, namely $\cap_{j=1}^n A_j$. Now by construction all the $B$'s are disjoint and their union is $\cup_{j=1}^n A_j$. On each such event, say $B_{i_1,i_2,\dots,i_k}$, the value of $X$ is $c_{i_1} + \dots + c_{i_k}$. Thus, the pre-image of any set is a union (possibly empty) of the $B$'s, an event. □

We moved from indicators to simple functions by taking linear combinations. We now introduce new operations that allow to generated even more (and eventually all) RVs:

**Proposition 2.1.3.** *Let $X_1, X_2, \dots$ be RVs. Then $\sup_n X_n$ is in RV.*

*Proof.* $\{\sup_n X_n > b\} = \cup_n \{X_n > b\}$. □

**Proposition 2.1.4.** *Let $X$ be a RV with values in $\mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}$ continuous. Then $f(X)$ is a RV.*

**Proposition 2.1.5.**

$$\{\omega : f(X(\omega)) \le b\} = \{\omega : X(\omega) \in f^{-1}((-\infty, b])) = X^{-1}(f^{-1}(-\infty, b])).$$

*Because $f$ is continuous the pre-image $f^{-1}((-\infty, b])$ of the closed set $(-\infty, b]$ is closed, hence a Borel set, and because $X$ is a RV, the pre-image of that set is an event.*

Therefore, if $X$ is a RV, so are all polynomials of $X$, $\sin(X), e^X$, etc. Well, we can't stop here. We need some basic algebraic operations such as sums, products, etc. The idea is pretty much the same, but a little more involved.

**Proposition 2.1.6.** *Let $X, Y$ be real-valued RVs, and let $f : \mathbb{R}^2 \to \mathbb{R}$ be continuous. Then $f(X, Y)$ is a RV.*

*Proof.* What do we need to show? As mentioned earlier we can replace the weak inequality with the strict ineequality, and show that $\{f(X, Y) < b\} = \{\omega : (X, Y)(\omega) \in f^{-1}((-\infty, b))\}$ is an event for every $b \in \mathbb{R}$. Because $f$ is continuous, $f^{-1}((-\infty, b))$ is open in $\mathbb{R}^2$. To complete the proof it is enough to show that if $U$ is open in $\mathbb{R}^2$, then $\{\omega : (X, Y)(\omega) \in U\}$ is an event. Any open set in $\mathbb{R}^2$ is necessarily a countable union of sets of the form $(a, b) \times (c, d)$. Because pre-images commute with set operations, the pre-image of $U$ under the function $\omega \to (X, Y)(\omega)$ a countable union of events of the form $X^{-1}((a, b)) \cap Y^{-1}((c, b))$, therefore an event. □

We gathered enough results to prove a theorem that states exactly what we said in words earlier. It's easy but it's very important. We first present some notation. Let $X$ be any RV. Write $X_+$ for the random variable $\max(X, 0)$ and $X_-$ for the random variable $(-X)_+ = -\min(X, 0)$. Why both are RVs? Because each is a continuous function of the RV $X$. Moreover, $X_+ \ge 0, X_- \ge 0$, $X_+X_- = 0$ and $X = X_+ - X_-$. Or: $X$ is a difference of two nonnegative RVs, with the property that their product is zero. Here's our result:

**Theorem 2.1.1.** *Let $X$ be a nonnegative RV. Then there exists a sequence $(\varphi_n : n \in \mathbb{N})$ of nonegative simple RVs increasing to $X$. Moreover, for every $M > 0$, the convergence is uniform on $\{X < M\}$.*

*Proof.* Define

$$\varphi_n(\omega) := n\mathbf{1}_{\{X \ge n\}} + \sum_{k=0}^{n2^n} k2^{-n}\mathbf{1}_{\{X \in [k2^{-n}, (k+1)2^{-n})\}}.$$

Simply said: $\varphi_n(\omega)$ is the largest dyadic number with denominator expressible as $2^{-n}$ if $\le X(\omega)$ or $n$ according to whether $X \le n$ or $X > n$. Then $\varphi_n$ is simple, because it is a RV and its image is finite. It is non-decreasing (you check), and if $X(\omega) < M$, then for $n > M$ we have $0 \le X - \varphi_n \le 2^{-n}$, hence the claimed uniform convergence holds. □

The pre-images of a RV form a $\sigma$-algebra:

**Definition 2.1.3.** 1. *Let $X$ be a RV. The $\sigma$-algebra generated by $X$, $\sigma(X)$, is defined as*

$$\sigma(X) := \{X^{-1}(B) : B \text{ Borel}\}.$$

2. *Let $I$ be a set of indices and $(X_i : i \in I)$ be RVs. The $\sigma$-algebra generated by $(X_i : i \in I)$, $\sigma(X_i : i \in I)$ is the $\sigma$-algebra generated by all events in $(\sigma(X_i) : i \in I)$.*

In words, $\sigma(X)$ is the smallest $\sigma$-algebra with respect to which $X$ is measurable. For example if $A \in \mathcal{F}$, $\sigma(\mathbf{1}_A) = \{\emptyset, A, A^c, \Omega\}$, even if $\mathcal{F}$ has more elements. While we're at it, let's define

**Definition 2.1.4.** *Let $(X_n : n \in \mathbb{N})$ be a sequence of RVs. The sequence is called independent if the corresponding $\sigma$-algebras $(\sigma(X_n) : n \in \mathbb{N})$ are independent.*

**Example 2.1.1.** *Let $(\Omega, \mathcal{F})$ be the infinite coin tossing space of Example 1.1.1 and here denoted by $\mathcal{F}$. Recall the coordinate mappings: for $n \in \mathbb{N}$, let $X_n : \Omega \to \{0, 1\}$ be defined as $X_n(\omega) = \omega_n$, the value of the $n$-th coin toss.*

- *Note that $\sigma(X_n)$ is a cylinder event, and therefore $X_n$ is a RV. In particular, $\sigma(X_n : n \in \mathbb{N}) \subseteq \mathcal{F}$.*

- *Conversely, every cylinder is in $\sigma(X_n : n \in \mathbb{N})$, and therefore $\mathcal{F} \subseteq \sigma(X_n : n \in \mathbb{N})$.*

*Conclusion: $\mathcal{F}$ is the $\sigma$-algebra generated by the coordinate mappings. Let's look at some examples of new RVs we can form:*

1. *For each $n \in \mathbb{N}$, the partial sum $S_n = X_1 + \cdots + X_n$ is a RV, as a finite sum of RVs, and so is $S_n/n$, as a continuous function of $S_n$ (multiply it by the constant $1/n$).*
2. *$\limsup S_n/n$ is also a random variable. Why? It is $\sup_n \inf_{k\ge n} \frac{S_k}{k}$*
3. *Also, let $T := \inf\{n : X_n = 1\}$, the index of the first appearance of $1$ (by convention, $\inf \emptyset = \infty$). This is a RV. Why? For every $b \in \mathbb{R}$, $\{T > b\} = \cup_{n\le b}\{X_n = 0\}$, a finite intersection of events.*

*Moreover, the RVs $(X_n : n \in \mathbb{N})$ are independent. This is because for each $n \in \mathbb{N}$, $\sigma(X_n)$ is $\mathcal{F}_n$ from Example 1.4.3, and those are independent by construction.*

**Exercise 2.1.1.** *Continue the example. Prove that all of the following are RVs:*

1. *For $k \in N$, the $k$-th appearance of $1$ or $\infty$ if $1$ appears less than $k$ times.*
2. *$\sum_{n=1}^{\infty} 2^{-n} X_n$.*

**Exercise 2.1.2.** *Let $S_n$ be as in Example 2.1.1. Consider the RV $X := \limsup_{n\to\infty} S_n/n$, For $n \in \mathbb{N}$, let $\mathcal{F}_n = \sigma(X_n)$, and let $\mathcal{T}$ be the corresponding Tail $\sigma$-algebra, Definition 1.4.3.*

1. *Show that $\sigma(X) \subseteq \mathcal{T}$.*
2. *Can you find an event in $\mathcal{T}$ which is not in $\sigma(X)$?*

**Proposition 2.1.7.** *Let $(X_n : n \in N)$ be independent and let $(f_n : n \in \mathbb{N})$ be Borel measurable functions from $\mathbb{R}$ to $\mathbb{R}$. Then the sequence of RVs $(f_n(X_n) : n \in \mathbb{N})$ is independent.*

*Proof.* For any $n \in \mathbb{N}$ and Borel set $B$, we have

$$\{f_n(X_n) \in B\} = \{X_n \in f_n^{-1}(B)\} \in \sigma(X_n).$$

Therefore $\sigma(f_n(X_n)) \subset \sigma(X_n)$, and the result follows from the definition of independent $\sigma$-algebras. □

## 2.2 Lebesgue Integral and the Expectation

The expectation is the mathematical notion generalizing arithmetic averages. The Lebesgue integral is the tool through which we define expectation. Nearly all areas of mathematical analysis use Lebesgue integration, probability is just one. Lebesgue integration is more general than Riemann integration (not restricted to Euclidean space), and provides a much more robust structure (limit theorems). In this section we introduce and present the basic properties of the Lebesgue integral with respect to a (general - not restricted to probability) measure.

Throughout this section we fix a measure space $(\Omega, \mathcal{F}, \mu)$.

### 2.2.1 Construction of the Integral

We define the Lebesgue integral of a RV $X$ with respect to $\mu$ in three steps. Recall that if $X$ is a RV, then its positive part $X_+$ and negative part $X_-$ are, respectively, the nonnegative RVs $X_+ := \max(X, 0)$ and $X_- := \max(-X, 0)$. We have $X_+X_- = 0$ and $X = X_+ - X_-$.

**Definition 2.2.1.** 1. *Let $X$ be a simple RV. The integral of $X$, denoted by $\int X d\mu$, is defined as*

$$\sum_{0\ne k\in X(\Omega)} k\mu(X = k),$$

*provided the sum is well-defined (no $\infty - \infty$), even if infinite.*

2. *Let $X$ be a nonnegative RV. The integral of $X$, denoted by $\int X d\mu$, is defined as*

$$\sup_{\varphi} \int \varphi d\mu,$$

*where the supremum is taken over all nonnegative simple RVs $\varphi$ satisfying $\varphi \le X$.*

3. *Let $X$ be a RV. The integral of $X$, denoted by $\int X d\mu$, is defined as $\int X_+ d\mu - \int X_- d\mu$, provided at least one of the integrals is finite. If both integrals are finite, then we say that $X$ is integrable, also denoted by $X \in L^1(\mu)$.*

*If $\mu$ is a probability measure, we often write $E[X]$ for $\int X d\mu$ and refer to the integral as the expectation of $X$ (under $\mu$).*

If $A \in \mathcal{F}$ we write $\int_A X d\mu$ for $\int X \mathbf{1}_A d\mu$, the integral of $X$ on (or over) $A$, and if $\mu$ is a probability measure, we will often write $E[X, A]$ for the same thing.

Note that if $\mu$ is a finite measure ($\mu(\Omega) < \infty$) - which includes all probability measures - then the integral of any simple RV is always well-defined, and that the expectation of a simple RV is weighted average of the values $X$, with the weights being the probabilities. This simple definition which relies on the notion of a measure and is not specific to any particular space (like $\mathbb{R}^n$) leads to a rich and complete theory of integration.

**Proposition 2.2.1.** *Let $X$ be a RV. Then*

1. *$X$ is integrable if and only if $\int |X| d\mu < \infty$.*

2. *(triangle inequality) If $X$ is integrable, then $|\int X d\mu| \le \int |X| d\mu$.*

**Exercise 2.2.1.** *Prove the proposition.*

Let's conisder some examples.

**Example 2.2.1.** *Consider the measure $\delta_0$ on the Borel subsets of $\mathbb{R}$. Let $X$ be any simple function. Then*

$$\int X d\delta_0 = \sum_{0 \ne k \in X(\mathbb{R})} k \delta_0(X = k).$$

*Now $\delta_0(A) = 1$ if $0 \in A$ and zero otherwise and is otherwise $= 0$. There is at most one $k \ne 0$ in the image of $X$ whose preimage $\{X = k\}$ contains $0$, and for such $k$, $k\delta_0(X = k) = X(0)1$. For all other $k$'s $k\delta(X = k) = k * 0 = 0$. Thus, $\int X d\delta_0 = X(0)$. Since this is true for all simple functions, by taking the limits in the construction, it holds for all RVs $X$.*

**Example 2.2.2.** *Let $\Omega$ be finite, $\mathcal{F}$ the power set and $\mu$ any measure. For $\omega \in \Omega$, write $\mu_\omega$ for $\mu(\{\omega\})$. Then*

$$\begin{aligned}\int X d\mu &= \sum_{k \ne 0} X(\omega)\mu(X = k) \\ &= \sum_{\omega \in \Omega} \mu_\omega X(\omega),\end{aligned}$$

*a weighted sum of the $X(\omega)$'s.*

Let's take this a little further:

**Example 2.2.3.** *Let $\Omega = \mathbb{N}$, $\mathcal{F}$ the power set and $\mu$ the counting measure: $\mu(A) := |A|$, the number of elements in $A$. Any function $X : \mathbb{N} \to \mathbb{R}$ is a RV (pre-image of any set is in $\mathcal{F}$, because the latter is the power set).*

*Let $X$ be a nonnegative simple function, and let's assume that the nonzero elements in the image of $X$ are $0 < c_1 < c_2 < \cdots < c_n$. By constrcution, $\mu(X = c_j) = |\{\omega : X(\omega) = c_j\}|$, number of times $c_j$ is repeated in the image of $X$. Then either all $c_j$'s are repeated finitely many times, or at least one of them is repeated infinitely many times. In either case a simple argument shows that $\int X d\mu = \sum_{\omega=1}^{\infty} X(\omega)$. Therefore in this context we can think of $X$ as a nonnegative sequence (any function from $\mathbb{N}$ to $\mathbb{R}$ is a sequence) $X(1), X(2), \ldots$, and its integral as the corresponding series $\sum_{\omega=1}^{\infty} X(\omega)$.*

*When thinking of a general RV $X$ (or sequence in this setting), then the proposition tells us that $X$ is integrable if and only if the corresponding series converges absolutely. To a large extent, the theory of Lebesgue integration generalizes the notion of absolute summability.*

**Exercise 2.2.2.** *Let $(c_n : n \in \mathbb{N})$ be a nonnegative sequence. For any subset $A$ of $\mathbb{N}$, let $\mu(A) := \sum_{n\in A} c_n$ (by convention, a sum over an empty set is zero).*

1. *Prove that $\mu$ is a measure on the power set of $\mathbb{N}$.*
2. *Show that $X : \mathbb{N} \to \mathbb{R}$ is integrable with respect to $\mu$ if and only if $\sum_{n=1}^\infty c_n|X(n)| < \infty$ and in this case $\int X d\mu = \sum_{n=1}^\infty c_n X(n)$.*

## 2.3 Monotone Convergence and Linearity

In this section we present the two most important features of the Lebesgue integral: continuity (in a specific setting, but not too restrictive) and linearity.

The following theorem is probably the key justification for the Lebesgue integral. It is called Lebesgue's Monotone Convergence Theorem (MCT).

**Theorem 2.3.1** (Monotone Convergence Theorem, MCT)**.** *Let $\mu$ be a measure. Let $0 \le X_1 \le X_2 \le \dots$ be a sequence of RVs. Then*

$$\lim_{n\to\infty} \int X_n d\mu = \int \lim_{n\to\infty} X_n d\mu.$$

In words, the integral we have defined (expectation of a RV) commutes with monotone limits (of nonnegative RVs). This is not true for Riemann integrals. This can be viewed as continuity result: integral of the limit (of an increasing nonnegative sequence) is the limit of the integrals. And as you'll see from the proof, the theorem is nothing but an extension of continuity from below of measures, extending the notion from indicators of increasing events to general sequences of nonndecreasing RVs.

Let's highlight one difference from the Riemann integral. If $(q_n : n \in \mathbb{N})$ is a sequence containing all rationals in $[0,1]$ and $f_n(x) = \sum_{k\le n} \mathbf{1}_{\{q_n\}}(x)$, then $f_n \nearrow \mathbf{1}_{\mathbb{Q}\cap[0,1]}$, and $f_n$ is Riemann integrable with a Riemann integral $\int_0^1 f_n(x)dx = 0$. Since $f_n$ is a simple nonnegative function, the Lebesgue integral of $f_n$ with respect to the Lebesgue measure, $\int f_n dm$ exists and is $= 0$. Now let's look at $\lim_{n\to\infty} f_n$, AKA the Dirichlet function. This is the first example we all saw of a nonnegative bounded function which is not Riemann integrable, so we can't take the limit inside the integral... But from the perspective of the Lebesgue integral? No problem. An easy calculation shows that the $\int \lim_{n\to\infty} f_n dm = 0$, as expected (or we just apply the theorem).

*Proof of Theorem 2.3.1.* Let $X = \lim_{n\to\infty} X_n$. Clearly $\int X_n d\mu \le \int X d\mu$. Let $Y$ be a simple nonnegative RV which is $\le X$. Let $c_1 < c_2 < \cdots < c_n$ be the nonzero elements in the image of $Y$. Then $Y = \sum_{j=1}^k c_j \mathbf{1}_{A_j}$, where $A_j = \{Y = c_j\}$, $j = 1, \dots, n$. Fix $\epsilon > 0$. The fact that $X(\omega) = \lim_{n\to\infty} X_n(\omega)$ implies that for every $\omega \in \Omega$, $X_n(\omega) \ge (1-\epsilon)Y$ for sufficiently large $n$. Let $N = N(\epsilon, \omega)$ be the smallest integer that $X_n(\omega) \ge (1-\epsilon)Y$. Then $N$ is a RV (why?). Now

$$\begin{aligned}\int X_n d\mu \ge \int X_n \mathbf{1}_{\{n>N\}} d\mu \ge (1-\epsilon)\int Y \mathbf{1}_{\{n>N\}} d\mu \\ \ge (1-\epsilon)\sum_{j=1}^n c_j \mu(A_j \cap \{n > N\}).\end{aligned}$$

Now $A_j \cap \{n > N\} \underset{n\to\infty}{\uparrow} A_j$, and therefore using the continuity of measures with respect to increasing sequences, the measures on the righthand side increase to the measure of $A_j$. Bottom line,

$$\lim_{n\to\infty} \int X_n d\mu \ge (1-\epsilon)\int Y d\mu.$$

Note that the limit on the lefthand side exists because $n \to \int X_n d\mu$ is a nondecreasing sequence. By definition the integral of a nonnegative RV, we conclude the $\lim_{n\to\infty} \int X_n d\mu \ge (1-\epsilon)\int X d\mu$. Since $\epsilon > 0$ is arbitrary, our work is done. □

Want to see an example? Here's one.

**Example 2.3.1.** *Recall that for every* $|x| < 1$,

$$\ln(1+x) = \sum_{k=1}^{\infty} (-1)^{k+1} \frac{x^k}{k}.$$

*The radius of convergence of the power series on the right is* $1$, *and it clearly diverges at* $x = -1$. *What about* $x = 1$? *Well, the series converges due to the alternating sequence test, but is its value equal to* $\ln(1+1)$? *For that we need to show that as* $x \nearrow 1$, *the values of the series converge to the value of the series at* $X = 1$.

*Let's go. First let's get rid of the alternating signs by combining every two successive terms odd* $k$ *followed by even* $k$ *into one:*

$$\frac{x^{2\omega-1}}{2\omega-1} - \frac{x^{2\omega}}{2\omega} = x^{2\omega-1}\left(\frac{1}{2\omega-1} - \frac{x}{2\omega}\right).$$

*Note that because* $|x| < 1$, *each of these terms is nonnegative. Also note (differentiate the lefthand side) that each of these terms increases as a function of* $x$.

*Now left find a nice measure space. We will take* $\Omega = \mathbb{N}$, *and equip it with the counting measure. As discussed in Example 2.2.3 the integral of nonnegative* $X$ *is* $\int X d\mu = \sum_{\omega=1}^{\infty} X(\omega)$. *Next, let* $(r_n : n \in \mathbb{N})$ *be a sequence in* $(0,1)$ *strictly increasing to* $1$. *For each* $n \in \mathbb{N}$, *let* $X_n(\omega) := \frac{r_n^{2\omega-1}}{2\omega-1} - \frac{r_n^{2\omega}}{2n}$ *(remember that* $\omega$ *is a natural number). From the discussion above* $0 \le X_1 \le X_2 \le \dots$ *and the integral of* $X_n$ *is* $\int X_n d\mu = \sum_{\omega=1}^{\infty} X_n(\omega) = \sum_{\omega=1}^{\infty} \frac{r_n^{2\omega-1}}{2\omega-1} - \frac{r_n^{2\omega}}{2\omega}$. *From the discussion above, this is* $\ln(1+r_n)$. *As* $\lim_{n\to\infty} X_n(\omega) = \frac{1}{2\omega-1} - \frac{1}{2\omega}$, *the MCT gives us*

$$\lim_{n\to\infty} \ln(1+r_n) = \lim_{n\to\infty} \int X_n d\mu = \int \lim_{n\to\infty} X_n d\mu = \sum_{\omega=1}^{\infty} \frac{1}{2\omega-1} - \frac{1}{2\omega} = \sum_{k=1}^{\infty} \frac{(-1)^{k+1}}{k}.$$

*Because the function* $x \to \ln(1+x)$ *is continuous at* $x = 1$, *the lefthand side is equal to* $\ln 2$.

*Nice, right?*

Let's consider another example.

**Example 2.3.2.** *Let* $(\Omega, \mathcal{F}, \mu)$ *be a measure space and let* $X$ *be a nonnegative RV. Let* $I_n = \int n \ln(1 + X/n) d\mu$. *Let's find* $\lim_{n\to\infty} I_n$. *First, recall from calculus (if not, prove), that for every* $c > 0$, *the sequence* $n \to (1+\frac{c}{n})^n$ *increases as* $n \to \infty$ *to* $e^c$. *For* $n \in \mathbb{N}$, *define the RV* $X_n := \ln(1+X/n)^n$. *Then* $0 \le X_1 \le X_2 \dots$, *and* $\lim_{n\to\infty} X_n = \ln e^X = X$. *Therefore, the MCT guarantees that* $\lim_{n\to\infty} I_n = \int X d\mu$.

The next result establishes monotonicity and linearity of the integral, both extremely important in all that will follow.

**Theorem 2.3.2.** 1. *If* $X$ *and* $Y$ *are RVs with* $X \le Y$ *and* $Y$ *integrable, then* $\int X d\mu \le \int Y d\mu$.

2. *If* $X$ *is integrable and* $c \in \mathbb{R}$, *then* $\int cX d\mu = c \int X d\mu$.

3. *If* $X$ *and* $Y$ *are integrable, then* $X + Y$ *is integrable, and*

$$\int X + Y d\mu = \int X d\mu + \int Y d\mu \tag{2.3.1}$$

*Proof of part 1.* Clearly $X \le Y$ and so if $X \ge 0$, $Y \ge 0$. This leads us to $X_+ = X\mathbf{1}_{X\ge 0} \le Y\mathbf{1}_{X \ge 0} \le Y\mathbf{1}_{Y\ge 0} = Y_+$, therefore $\int X_+ d\mu \le \int Y_+ d\mu < \infty$. Similarly, $-Y_- = Y\mathbf{1}_{Y\le 0} \ge X\mathbf{1}_{Y\le 0} \ge X\mathbf{1}_{X\le 0} = -X_-$, that is $Y_- \le X_-$. We have $\int X d\mu = \int X_+ d\mu - \int X_- d\mu \le \int Y_+ d\mu - \int Y_- d\mu = \int Y d\mu$. □

We leave part 2 as an exercise.

Before we prove the last part, we need the following lemma.

**Lemma 2.3.1.** *Let* $X$ *and* $Y$ *be RVs. Then* (2.3.1) *holds in each of the following cases:*

1. $X, Y$ *are both simple and integrable.*

2. *$X, Y$ are both nonnegative.*

*Proof.* Let's first assume that $X = \sum_{j=1}^n c_j \mathbf{1}_{A_j}$ with $A_j$'s disjoint and that $Y$ is of the form $c\mathbf{1}_A$ for some $c \neq 0$. Clearly,

$$X + Y = \sum_{j=1}^n (c_j + c)\mathbf{1}_{A_j \cap A} + \sum_{j=1}^n c_j \mathbf{1}_{A_j - A} + c\mathbf{1}_{A - \cup_{j=1}^n A_j}.$$

Note that all events are disjoint and therefore, the definition of the integral of a simple function gives

$$\int X + Y d\mu = \sum_{j=1}^n (c_j + c)\mu(A_j \cap A) + \sum_{j=1}^n c_j \mu(A_j - A) + c\mu(A - \cup_{j=1}^n A_j).$$

(we may need to consider the case $c = c_j$ for some $j$ separately, but we'll omit the details).

Separating the first sum we obtain

$$\sum_{j=1}^n c_j \mu(A_j) + c\mu(A) = \int X + Y d\mu.$$

Iterating this, we have that the expectation of the sum of two simple integrable RVs is the sum of their expectations. To obtain the conlcusion for nonnegative RVs, apply the MCT. □

*Proof of Theorem 2.3.2 part 3.* First, $|X + Y|$ is a nonnegative RV, and by the triangle inequality, $|X + Y| \leq |X| + |Y|$. The lemma gives $\int |X| + |Y| d\mu = \int |X| d\mu + \int |Y| d\mu < \infty$, and therefore $X + Y$ is integrable with $\int |X + Y| d\mu \leq \int |X| d\mu + \int |Y| d\mu$. Let $Z = X + Y$. We can rewrite this equality as equality on sums of nonnegative RVs: $Z_+ + X_- + Y_- = Z_- + X_+ + Y_+$. Therefore $\int Z_+ + X_- + Y_- d\mu = \int Z_+ + X_+ + Y_+ d\mu$. But since each side is expectation of a sum of nonnegative RVs we have the equality $\int Z_+ d\mu - \int Z_- d\mu = \int X_+ d\mu - \int X_- d\mu + \int Y_+ d\mu - \int Y_- d\mu$, which, by definition, reads $\int Z = d\mu \int X d\mu + \int Y d\mu$, the desired outcome. □

If you followed the proof you'd notice that we also proved the triangle inequality:

**Corollary 2.3.1.** *Let $X$ and $Y$ be RVs. Then*

$$\int |X + Y| d\mu \leq \int |X| d\mu + \int |Y| d\mu.$$

We close this section with a very important corollary of linearity, domain additivity of the integral.

**Corollary 2.3.2.** *Let $X$ be a nonnegative or intgerable RV. Let $A$ be an event. Then $\int X d\mu = \int_A X d\mu + \int_{A^c} X d\mu$.*

**Exercise 2.3.1.** *Use the definition of the integral to show that if $X$ is any RV and $N$ is an event with $\mu(N) = 0$, then the integral of $X$ over $N$ is defined and is equal to zero.*

## 2.4 Fatou's Lemma and Dominated Convergence

While we're at it, let's prove a few other useful results. First, we can relax the monotonity assumption in the MCT at a price of an equality:

**Theorem 2.4.1** (Fatou's Lemma)**.** *Let $(X_n : n \in \mathbb{N})$ be nonnegative RVs. Then $\liminf_{n\to\infty} \int X_n d\mu \geq \int \liminf X_n d\mu$.*

*Proof.* Let $Y_n = \inf_{k \geq n} X_k$. Clearly, $Y_n \leq X_n$, so $\int X_n d\mu \geq \int Y_n d\mu$. Next observe that $0 \leq Y_1 \leq Y_2 \leq \dots$, with $\lim_{n\to\infty} Y_n = \liminf X_n$, and by taking limits we have

$$\liminf_{n\to\infty} \int X_n d\mu \geq \lim_{n\to\infty} \int Y_n d\mu \overset{\text{MCT}}{=} \int \lim_{n\to\infty} Y_n d\mu = \int \liminf X_n d\mu.$$

□

**Exercise 2.4.1.** *We proved Fatou's lemma from the MCT. Show that the reverse implication also holds: Fatou's lemma implies the Monotone Convergence Theorem.*

The last two results in this batch are the dominated and bounded convergence theorems, both consequences of the above, and both very useful.

Before we continue present some language that will be used very often. A statement about RVs in our measure space is said to hold almost everywhere (abbreviated a.e.) if it holds possibly except on a set whose measure under $\mu$ is zero. For example, the Dirichlet function (indicator of rationals in $[0,1]$) is equal to 0 a.e. with respect to the Lebesgue measure. When $\mu$ is a probability measure, we will use "almost surely" (abbreviated a.s.) for "almost everywhere". Due to Corollary 2.3.2 and Exercise 2.3.1, for integration purposes, statements that hold a.e. should be viewed as holding everywhere. We will demonstrate this in the proof of the next theorem.

**Theorem 2.4.2** (Dominated Convergence Theorem, DCT). *Let $(X_n : n \in \mathbb{N}), X$ and $Y$ be RVs, with $Y$ integrable and $\sup_n |X_n| \le Y$ a.e. Suppose that $\lim_{n\to\infty} X_n = X$ a.e. Then $\lim_{n\to\infty} \int |X_n - X| d\mu = 0$, and in particular, $X$ is integrable and $\lim_{n\to\infty} \int X_n d\mu = \int X d\mu$.*

**Example 2.4.1.** *Let $Z$ be an integrable RV. We show that $\lim_{n\to\infty} \int_{|Z|>n} |Z| d\mu = 0$.*

*We apply the DCT with $X_n = |Z|\mathbf{1}_{\{|Z|>n\}}$ and let $Y = |Z|$. Clearly, $X := \lim_{n\to\infty} X_n = |Z|\mathbf{1}_{\{|Z|=\infty\}}$. Now $X$ is integrable, but equal to $\infty\mathbf{1}_{\{|Z|=\infty\}}$, so the construction of the integral of a nonnegative RV garantees that $\int X d\mu = 0$. Therefore the claim follows from the DCT.*

Since in a probability space any constant function is also integrable, we have

**Corollary 2.4.1** (Bounded Convergence Theorem, BCT). *Suppose that $\mu$ is a finite measure. Let $(X_n \in \mathbb{N}), X$ be RVs and $M$ a positive real number satisfying $\sup_n |X_n| \le M$, a.e. and $\lim_{n\to\infty} X_n = X$ a.e. Then the conclusions of the DCT hold.*

The proof of the Corollary from the DCT is immediate, because if $\mu$ is finite, any constant function is integrable.

*Proof of Theorem 2.4.2.* This uses Fatou's lemma and a little trick. Clearly $|X_n - X| \le |X_n| + |X| \le 2Y$ a.e. Therefore, the sequence $2Y - |X_n - X|$ is nonnegative a.e. Let $N$ be the event on which it is negative (which may be empty). Then $\mu(N) = 0$. Fatou's lemma gives

$$\liminf_{n\to\infty} \int_{N^c} 2Y - |X_n - X| d\mu \ge \int_{N^c} 2Y d\mu.$$

and since the integral of any RV on $N$ is zero (exercise 2.3.1), each of the integrals above is equal to the integral of the function (over the entire sample space, not just over $N^c$). In the sequel (including the next paragraph) we skip the intermediate step of integrating over null sets where an a.e. statement fails to hold.

Use the linearity of the expectation to conclude that the lefthand side is $\int 2Y d\mu - \limsup_{n\to\infty} \int |X_n - X| d\mu$, and therefore we obtain $\limsup \int |X_n - X| d\mu \le 0$. Next, note that by Fatou's lemma $\int |X| d\mu \le \liminf_{n\to\infty} \int |X_n| d\mu \le \int Y d\mu < \infty$. Therefore, $X$ is integrable, and we have

$$|\int X_n d\mu - \int X d\mu| = |\int X_n - X d\mu| \le \int |X_n - X| d\mu,$$

and the assertion follows from the first. □

## 2.5 Connection with Riemann Integratrion

Though the new notion of the Lebesgue integral is more abstract than the Riemann integral, both are intricately related. The Lebesgue integral is always with respect to a measure. The Lebesgue measure is a special measure on $\mathbb{R}$.

Let us now consider the measure space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. Fix a bounded closed interval $[a,b]$. Any step function $\phi$ on $[a,b]$ is a simple function, and its Riemann integral coincides with our definition of the Lebesgue integral of the simple function $\int \phi dm$ with respect to the Lebesgue measure. Since Riemann integrability is obtained through approximations of integrals of step functions (lower and upper Darbeaux sums), the following result is not surprising, and its proof is not hard.

**Theorem 2.5.1.** *Let $a < b$ be real numbers and $f : [a, b] \to \mathbb{R}$. Then*

1. *$f$ is Riemann integrable if and only if $f$ is bounded and the set of points where $f$ is not continuous has Lebesgue measure $0$.*
2. *If $f$ is Riemann integrable on $[a, b]$, then it is Lebesgue integrable with respect to the Lebesgue measure and its Lebesgue integral on $[a, b]$, $\int_{[a,b]} f dm$ coincides with its Riemann integral $\int_a^b f(t)dt$.*

**Corollary 2.5.1.** *Suppose that $f \geq 0$ is Riemann integrable on $\mathbb{R}$. Then it is Lebegsue integrable and $\int f(t)dt = \int f dm$.*

*Proof.* Let $f_n(t) := \mathbf{1}_{[-n,n]}(t)f(t)$. Then $(f_n : n \in \mathbb{N})$ is a sequence of nonnegative functions increasing to $f$. Since $f$ is Riemann integrable on any bounded interval, it is Riemann integrable on $[-n, n]$, and therefore $\int_{-n}^n f(t)dt = \int_{[-n,n]} f dm = \int f_n dm$. Now take $n \to \infty$. The lefthand side tends to $\int f(t)dt$ and the MCT gives that the righthand side tends to $\int f dm$. □

**Example 2.5.1.** *Let's find $\lim_{n\to\infty} \int_0^1 \frac{nx^n}{\sin x + x^{3/2}} dx$.*

*We are going to apply the DCT utilizing the theorem. First let's do some manipulation of the Riemann integral:*

$$\int_0^1 \frac{nx^n}{\sin x + x^{3/2}} dx = \int_0^1 \frac{nx^{n-1}}{\frac{\sin x}{x} + x^{1/2}} dx$$

$$\overset{t=x^n}{=} \int_0^1 \frac{1}{\frac{\sin t^{1/n}}{t^{1/n}} + t^{1/(2n)}} dt.$$

*For $n \in \mathbb{N}$, let $f_n(t) := \mathbf{1}_{[0,1]}(t) \frac{1}{\frac{\sin t^{1/n}}{t^{1/n}} + t^{1/(2n)}}$. This is a Riemann integrable on $[0, 1]$. Let $c$ be the infimum of the function $\frac{\sin x}{x}$ on $(0, 1]$. Then $c > 0$, and we clearly have $0 \leq f_n(t) \leq \mathbf{1}_{[0,1]}(t)\frac{1}{c}$. Also $f_n(t) \to \mathbf{1}_{[0,1]}(t)\frac{1}{\sin 1+1}$ $m$-a.e. (the limit exists at all $t \neq 0$), and the DCT imlies $\lim_{n\to\infty} \int f_n d\mu = \int_{[0,1]} \frac{1}{\sin 1+1} dm$. By the theorem, the lefthand side is $\lim_{n\to\infty} \int_0^1 \frac{nx^n}{\sin x+x^{3/2}}$, and the righthand side is $\int_0^1 \frac{1}{\sin 1+1} dt = \frac{1}{\sin 1+1}$.*

The next result allows to calculate Lebesgue integrals (not necessarily with respect to the Lebesgue measure!) through improper Riemann integrals. In what follows we fix a measure space $(\Omega, \mathcal{F}, \mu)$.

**Theorem 2.5.2.** *Let $Z$ be a nonnegative RV. Then*

$$\int Z d\mu = \int_0^\infty \mu(Z \geq t)dt.$$

We note since $t \to \mu(Z \geq t)$ is non-increasing, on the interval it is finite, it is continuous possibly expect for a countable number of jumps. Any countable subset of $\mathbb{R}$ has Lebesgue measure zero, and therefore on any bounded subinterval of this interval, the function is Riemann integrable.

Here's a sketch of the proof. The statement is valid for simple $Z$ by direct computation, and the result follows by approximating $Z$ by simple functions from below and applying the MCT to the righthand side. As an immediate corollary we have

**Corollary 2.5.2.** *The integral of $Z$ (with respect to $\mu$) is defined if and only if $\int_0^\infty \mu(Z \geq t)dt < \infty$ or $\int_0^\infty \mu(Z \leq t)dt < \infty$ and in this case*

$$\int Z d\mu = \int_0^\infty \mu(Z \geq t)dt - \int_0^\infty \mu(Z \leq t)dt.$$

**Example 2.5.2.** *Let's revisit some calculus. Suppose we want to calculate $\int_0^1 x^{1/n} dx$, where $n$ is any integer. Of course, we already know this is equal to $\frac{n}{n+1}$.*

*Let's use the results of this section to obtain this value (we will assume we can calculate the Riemann integral of a polynomial). Our integral is $\int_{[0,1]} x^{1/n} dm = \int Z dm$, where $Z = \mathbf{1}_{[0,1]}(x)x^{1/n}$. Now for $t \in (0, 1]$*

$$\{Z \geq t\} = \{x \in (0, 1) : x^{1/n} \geq t\} = \{x \in (0, 1) : x \geq t^n\}.$$

*The Lebesgue measure of this set is* $1-t^n$. *For* $t>1$, $\{Z\ge t\}=\emptyset$. *Therefore,*

$$\int_0^\infty m(Z\ge t)dt=\int_0^1 m(Z\ge t)=\int_0^1(1-t^n)dt.$$

*The righthand side is* $1-\frac{1}{n+1}=\frac{n}{n+1}$.

## 2.6 Distributions of RVs

In probability theory, a distribution is synonymous with a probability measure, often with some distinguishing specific characteristics. In this section we will review some common probability distributions.

We assume that $(\Omega,\mathcal{F},\mu)$ is a measure space.

**Definition 2.6.1.** *Let* $f$ *be a nonnegative RV. The measure with density* $f$ *(WRT to* $\mu$*), which we denote by* $\mu^f$*, is defined as*

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

**Exercise 2.6.1.** *Prove that* $\mu^f$ *defined above is indeed a measure.*

Note that if $\int f d\mu=1$, then $\mu^f$ is a probability measure.

**Proposition 2.6.1.** *Let* $Z$ *be a RV. Then* $Z$ *is integrable with respect to* $\mu^f$ *if and only if* $\int|Z|fd\mu<\infty$. *In this case*

$$\int Zd\mu^f=\int Zfd\mu.$$

**Exercise 2.6.2.** *Prove the proposition.*

The following two examples represent some common probability distributions as probability measures with density with respect to the Lebesgue measure on $\mathbb{R}$ or the counting measure on $\mathbb{Z}$.

**Example 2.6.1.** *A lot of common probability measures are obtained through this procedure with* $\mu$ *takes as the Lebesgue measure on* $\mathbb{R}$ *and* $f$ *a Riemann integrable function satisfying* $\int_{-\infty}^\infty f(x)dx=1$ *(which - as we saw earlier - implies* $f$ *is Lebesgue integrable with respect to* $m$ *and* $\int fdm=1$*):*

| *Parameter(s)* | $f(x)$ | *Notation* | *Description* |
|---|---|---|---|
| $\lambda>0$ | $\mathbf{1}_{[0,\infty)}(x)\lambda e^{-\lambda x}$ | $Exp(\lambda)$ | *Exponential with parameter* $\lambda$ |
| $\mu\in\mathbb{R},\sigma^2>0$ | $\frac{e^{-(x-\mu)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}}$ | $N(\mu,\sigma^2)$ | *Normal (Gaussian) with mean* $\mu$ *and variance* $\sigma^2$ |
| $-\infty<a<b<\infty$ | $\frac{1_{[a,b]}(x)}{b-a}$ | $U[a,b]$ | *Uniform on* $[a,b]$ |

**Example 2.6.2.** *Let* $\mu$ *be the counting measure on the nonnegative integers,* $\mathbb{Z}_+=\{0,1,2,\dots\}$. *Here is a list of some common probability distributions of the form* $\mu^f$ *for* $f\ge 0$ *with* $\int fd\mu=1$. *Recall that* $\int fd\mu=\sum_{n=0}^\infty f(n)$.

| *Parameter(s)* | $f(x)$ | *Notation* | *Description* |
|---|---|---|---|
| $p\in(0,1)$ | $\begin{cases}1-p & x=0\\ p & x=1\\ 0 & otherwise\end{cases}$ | $Bern(p)$ | *Bernoulli with probabilityp.* |
| $n\in\mathbb{N},p\in(0,1)$ | $\begin{cases}\binom{n}{x}p^x(1-p)^{n-x} & x\in\{0,1,\dots,n\}\\ 0 & otherwise\end{cases}$ | $Bin(n,p)$ | *Binomial with parameters* $n$ *and* $p$ |
| $p\in(0,1)$ | $\begin{cases}(1-p)^{x-1}p & x\in\mathbb{N}\\ 0 & otherwise\end{cases}$ | $Geom(p)$ | *Geometric with parameter* $p$ |
| $\lambda>0$ | $e^{-\lambda}\frac{\lambda^x}{x!}$ | $Pois(\lambda)$ | *Poisson with parameter* $\lambda$ |

## 2.7 RVs and their Distributions

In this section $(\Omega, \mathcal{F}, P)$ is a probability space.

**Definition 2.7.1.** *Let $X$ be a RV taking values in $\mathbb{R}$.*

1. *The distribution of $X$ is a Borel probability measure on $\mathbb{R}$ given by the push-forward measure $P_X$. Specifically:*
$$P_X(B) = P(X \in B),\ B \in \mathcal{B}(\mathbb{R}).$$
2. *The Cumulative Distribution Function (CDF) of $X$, $F_X$ is defined as*
$$F_X(x) := P_X((-\infty, x]) = P(X \le x),\ x \in \mathbb{R}.$$

Why is this so important? Because no matter what the underlying probability measure is the distribution is **always** a Borel probability measure. That is: a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$.

We will say that $X$ has distribution "blah" if its distribution $P_X$ is "blah". Recall, for example, the exponential distribution with parameter $\lambda$ from Example 2.6.1. We will say that $X$ is exponentially distributed with parameter $\lambda$ if $P_X$ is the exponential distribution with parameter $\lambda$, and we will express this statement in symbols by writing $X \sim \text{Exp}(\lambda)$. Another example: $X \sim Bin(n,p)$ means that the distribution of $X$ Binomial with parameters $n$ and $p$, a distribution we defined in Example 2.6.2.

**Example 2.7.1.** *Consider the infinite product measure with $p = \frac{1}{2}$. Let $X$ be the first coordinate mapping. So $X \sim Bern(\frac{1}{2})$.*

*Now consider the Lebesgue measure restricted to $[0,1]$ and let $Y$ be the indicator of the interval $[\frac{1}{4}, \frac{3}{4}]$. Then again $Y \sim Bern(\frac{1}{2})$.*

*Though the two RVs live in different probability spaces, their distributions are the same.*

**Proposition 2.7.1.** *Let $X$ and $F_X$ be as in Definition 2.7.1. Then*

1. *$F_X$ is non-decreasing, right continuous, $\lim_{x\to-\infty} F_X(x) = 0$, and $\lim_{x\to\infty} F_X(x) = 1$.*
2. *$P(X \in (a,b)) = F_X(b-) - F_X(a)$ and $P(X = a) = F_X(a) - F_X(a-)$ $(F_X(x-) := \lim_{y\uparrow x} F_X(x))$.*

**Exercise 2.7.1.** *Prove the proposition.*

As an application of the $\Pi - \Lambda$ theorem we have the following:

**Proposition 2.7.2.** *Let $X$ be a RV taking values in $\mathbb{R}$. Then its distribution $P_X$ is determined by its CDF $F_X$.*

If $P$ is any Borel probability measure on $\mathbb{R}$ and $X$ is the identity mapping, then $P_X = P$. Equivalently, $X$, the identity, has distribution $P$, and in this case $F_X(x) = P((-\infty, x])$. We also have the following:

**Proposition 2.7.3.** *Let $F : \mathbb{R} \to [0,1]$ satisfy all conditions listed in 2.7.1-1. Then there exists a unique Borel probability measure $P$ such that $P(X \le x) = F(x)$ for all $x \in \mathbb{R}$.*

Distributions of RVs are categorized into three groups:

**Definition 2.7.2.** *Let $F_X$ be the CDF of some RV $X$. Then the distribution of $X$ is*

1. *Continuous if $F_X$ is continuous, and absolutely continuous if $P_X$ has a density with respect to the Lebesgue measure.*
2. *Discrete if there exists a countable set $K$ such that $P(X \in K) = 1$, equivalently, $F_X$ increases only through jumps.*
3. *Mixed, otherwise.*

Since $F_X$ is right continuous at all points and has left limits at every point, it is continuous at a point $x$ if and only if $P(X = x) = F_X(x) - F_X(x-) = 0$. Thus the distribution of $X$ is continuous if and only if $P(X = x) = 0$ for all $x \in \mathbb{R}$. As for the discrete case, we have some countable $K$ such that $P(X \in K) = 1$, or $\sum_{k\in K} F_X(k) - F_X(k-) = 1$, which is the same as saying that all the increase of $F_X$ from 0 to 1 is through jumps at the elements in $K$.

All distributions in Example 2.6.1 are continuous and all distributions in Example 2.6.2 are discrete. Here is an example of a mixed distribution.

**Example 2.7.2.** *Suppose that $X \sim U[0,1]$. Let $Y = \max(X, \frac{1}{2})$. Then $P(Y = \frac{1}{2}) = \frac{1}{2}$, so $Y$ is not continuous. But for all $x \neq \frac{1}{2}$, $P(Y = y) = 0$, so $Y$ is not discrete. Therefore $Y$ is mixed. What is the CDF of $Y$?*

$$F_Y(y) = \begin{cases} 0 & y < \frac{1}{2} \\ \min(y,1) & y \geq \frac{1}{2} \end{cases}$$

We continue to some calculations with distributions.

**Example 2.7.3.** *Suppose that $X \sim N(0,1)$. We will show that for $\mu \in \mathbb{R}$ and $\sigma > 0$, the RV $Y := \sigma X + \mu$ satisfies $Y \sim N(\mu, \sigma^2)$.*

*The distribution of $X$ has density $f(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$ with respect to the Lebesgue measure. Therefore its CDF $F_X$ satisfies*

$$F_X(x) = P(X \leq x) = \int_{(-\infty,x]} f_X(t)dm = \int_{-\infty}^{x} f_X(t)dt.$$

*Now let's express the CDF of $Y$, $F_Y$, in terms of the CDF of $X$. Why? Because the CDF determines the distribution, and hope it is of the form we want.*

$$\begin{aligned} F_Y(y) &= P(\sigma X + \mu \leq y) \\ &= P(X \leq \frac{y-\mu}{\sigma}) \\ &= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\frac{y-\mu}{\sigma}} e^{-t^2/2} dt \\ &\overset{x=\sigma t+\mu}{=} \frac{1}{\sqrt{2\pi\sigma^2}} \int_{-\infty}^{y} e^{-(x-\mu)^2/(2\sigma^2)} dx \end{aligned}$$

*Therefore the CDF of $Y$ is the same as the CDF of $N(\mu, \sigma^2)$.*

Suppose $X$ is a RV and $\phi$ is nonnegative Borel measurable function and we want to calculate $E[\phi(X)] = \int \phi(X)dP$. We do not need to work directly with $P$, but instead work with its distribution which - we recall - is a Borel probability measure on $\mathbb{R}$. This makes our life simpler!

**Theorem 2.7.1.** *Let $X$ be a RV and let $\phi : \mathbb{R} \to [0,\infty)$ be Borel measurable. Then $E[\phi(X)] = \int \phi(x)dP_X$*

**Corollary 2.7.1.** *Suppose that $X$ is a RV and $\phi : \mathbb{R} \to \mathbb{R}$ is Borel measurable. Then $\phi(X)$ is integrable if and only if $E[|\phi|(X)] < \infty$ and in this case $E[\phi(X)] = \int \phi(x)dP_X$.*

*Proof of Theorem 2.7.1.* Use Theorem 2.5.2 to obtain:

- The lefthand side is equal to $\int_0^\infty P(\phi(X) \geq t)dt$.
- The righthand side is equal to $\int_0^\infty P_X(\{x \in \mathbb{R} : \phi(x) \geq t\})dt$.

Now

$$P_X(\{x \in \mathbb{R} : \phi(x) \geq t\}) = P_X(\phi^{-1}([t,\infty)] = P(X \in \phi^{-1}([t,\infty)) = P(\phi(X) \geq t).$$

Thus the two Riemann integrals above are of the same function. □

**Example 2.7.4.** *Suppose that $X \sim Exp(\lambda)$. Let's find the expectation of $X$.*

*Since the distribution of $X$ has density $\mathbf{1}_{[0,\infty)}\lambda e^{-\lambda x}$ with respect to the Lebesgue measure, apply the corollary with $\phi(x) = x$, and use Proposition 2.6.1 to obtain*

$$E[X] = \lambda \int_0^\infty x e^{-\lambda x} dx = \frac{1}{\lambda},$$

*where the last equality was obtained through integration byb parts.*

**Example 2.7.5.** *Let's calculate the expectation of $Y \sim N(\mu, \sigma^2)$. Since we already know that this distribution is the same as for $Z := \sigma X + \mu$ where $X \in N(0,1)$, we will apply Proposition 2.6.1 to $Z$. To obtain*

$$E[Z] = \int \sigma X + \mu f_X dm = E[\sigma X + \mu] = \sigma E[X] + \mu.$$

*Now*

$$E[X] = \int x f_X(x) dm = \int_{-\infty}^\infty x f_X(x) dx = 0,$$

*because $f_X$ is symmetric.*

## 2.8 Independent RVs

**Definition 2.8.1.** *The sequence of RVs $(X_n : n \in N)$ is independent if the corresponding sequence of $\sigma$-algebras $(\sigma(X_n) : n \in \mathbb{N})$ is independent.*

**Example 2.8.1.** *Consider the infinite coin tossing measure from Theorem 1.2.2, and let $(X_n : n \in \mathbb{N})$ be the coordinate mappings. Then by definition of the product measure this sequence is independent.*

**Corollary 2.8.1.** *Let $(X_n : n \in \mathbb{N})$ be a sequence of RVs. Then the sequence is independent if and only for any finite subsequence $1 \le i_1 < i_2 < \dots < i_n$ of indices and real numbers $b_1, \dots, b_n$ $P(\cap_{j=1}^n \{X_{i_j} \le b_j\}) = \prod_{j=1}^n P(X_{i_j} \le b_j)$.*

*Proof.* This is a corollary to Theorem 1.4.1. We omit details. □

The next result states that independence is preserved under functions applied to RVs individually.

**Corollary 2.8.2.** *Suppose that $(X_n : n \in \mathbb{N})$ are independent. Let $(f_n : n \in \mathbb{N})$ be sequence of Borel measurable functions. Then the RVs $(f_n(X_n) : n \in \mathbb{N})$ are independent.*

*Proof.* For every borel set $B$ and $n \in \mathbb{N}$ $\{f_n(X_n) \in B\} = \{X_n \in f_n^{-1}(B)\} \in \sigma(X_n)$. Therefore, $\sigma(f_n(X_n)) \subseteq \sigma(X_n)\}$, and the result follows from the definition of independent $\sigma$-algebras. □

**Theorem 2.8.1.** *Let $(X_n : n \in \mathbb{N})$ be independent integrable RVs. Then for every $n \in \mathbb{N}$, $E[\prod_{j=1}^n X_j] = \prod_{j=1}^n E[X_j]$.*

*Proof.* We will prove the statement in two stages. First we show that if $X$ and $Y$ are independent and integrable, then $E[XY] = E[X]E[Y]$, and then we generalize through an inductive argument.

For the first step, we can assume without loss of generality first that $X$ and $Y$ are simple. Why? Theorem 2.1.1 shows that for any nonnegative RV $X$ (you complete the details for a general RV), for each $n \in \mathbb{N}$ there is a Borel measurable function $\phi_n$ such that $\phi_n(X)$ is simple, $|\phi_n(X)| \le |X|$ and $\lim_{n\to\infty} \phi_n(X) = X$ with the analogous statements for $Y$.

Then, assuming both $X$ and $Y$ are simple, then $X = \sum_{k \in X(\Omega)} k \mathbf{1}_{\{X=k\}}$ and $Y = \sum_{k' \in Y(\Omega)} k' \mathbf{1}_{\{Y=k'\}}$ with both sums over finite sets. Then

$$\begin{aligned} E[XY] &= E[\sum_{k,k'} kk' \mathbf{1}_{\{X=k\}\cap\{Y=k'\}}] \\ &= \sum_{k,k'} kk' P(X = k, Y = k') \\ &= \sum_{k,k'} kk' P(X = k) P(Y = k') \\ &= \sum_k k P(X = k) \sum_{k'} k' P(Y = k') \\ &= E[X]E[Y]. \end{aligned}$$

In fact, the same argument shows that $E[|XY|] = E[|X|]E[|Y|]$. Through monotone convergence the equality holds for all nonnegative $X$ and $Y$. If $X$ and $Y$ are both integrable RVs, then this argument gives $E[|XY|] = E[|X|]E[|Y|]$, which in turn implies that $XY$ is integrable. Also, the second corollary states, $X_\pm$ are both independent of each of $Y_\pm$. Write $X = X_+ - X_-$ and $Y = Y_+ - Y_-$. Then $XY = (X_+ - X_-)(Y_+ - Y_-) = X_+Y_+ - X_+Y_- - X_-Y_+ + X_+Y_+$. Each of the four summands is a product of two independent RVs and is an integrable RV. Therefore,

$$\begin{aligned} E[XY] &= E[X_+Y_+] - E[X_+Y_-] - E[X_-Y_+] + E[X_+Y_+] \\ &= E[X_+]E[Y_+] - E[X_+]E[Y_-] - E[X_-]E[Y_+] + E[X_+]E[Y_+] \\ &= E[X]E[Y]. \end{aligned}$$

It remains to expand the statement to finite products. We do it by induction, with the base case $n = 2$ already proved. We want to show that if $X_1, \dots, X_n, X_{n+1}$ are indepenent and integrable, then so is their product and its expectation is the product of the expectations. Let $X = \prod_{j=1}^n X_j$. Then by the induction hypothesis $X$ is integrable and $E[X] = \prod_{j=1}^n E[X_j]$. Now by Theorem 1.4.2, we have that $\sigma(X_1, \dots, X_n)$ and $\sigma(X_{n+1})$ are independent, so $X$ and $X_{n+1}$ are independent integrable RVs so that $E[XX_{n+1}] = E[X]E[X_{n+1}] = \prod_{j=1}^n E[X_j]$. □

# Chapter 3

# Laws of Large Numbers

We know that probability spaces can be viewed as mathematical modeling of chance. To be more precise, the set $\Omega$ consists of all outcomes in an experiment, and a probability measure weighs sets of outcomes, those which we can observe, AKA events.

As you know from coin tossing, we cannot tell anything about the outcome of a single experiment, but if we repeat it, some interesting patterns appear. To make the connection, we need some mathematical notion for "repeating" the experiment. Actually, we have already defined it. This notion is already hidden in concepts we defined before including product spaces and independence, both representing different aspects of it. If we are looking repeat the experiment twice, then our natural sample space is $\Omega \times \Omega$, the set of all ordered pairs of the form $(\omega_1, \omega_2)$, representing the outcome of the first and the second. If we want an infinite sequence, like we did with the coin tossing, then we may want to use an infinite product space $\Omega^{\mathbb{N}}$, equivalently, the set of all functions from $\mathbb{N} \to \Omega$. Since any finite or infinite sequence of experiments can be viewed as one large experiment, we can simply assume that our sample space is large enough to support what we're interested in without providing a concrete description. This may seem somewhat vague, but will be clearer soon. Now that we have this out of our way, let's ponder "repeating an experiment". We basically want each of the experiments to be such that observation of each has no impact on the observation of the others, and that all are identical namely what we see in each is described by the same probability distribution. The mathematical interpretation of the first is independence of the experiments and the second is all experiments having identical distributions. To put this into a concrete setting, let's simply assume that the outcome of each experiment in out sequence is represented by some numerical value. The mathematical interpretation of this is that we represent the outcome of the $n$-th experiment by a random variable, say $X_n$. Going back to independence, we assume that the RVs in the sequence $(X_n : n \in \mathbb{N})$ are independent. Adding to this the identical distribution for each experiment, we obtain an IID sequence. The infinite product measure and the sequence of coordinate mappings provide the simplest example of what we have just described. However, we can also go in the reverse direction: any IID sequence $(X_n : n \in \mathbb{N})$ can and should be viewed as resulting from repeating a certain experiment infinitely many times, independently, regardless of the underlying sample space.

This long introduction was necessary to explain the idea behind laws of large numbers. The key idea is this: repeating an experiment a large number of times (e.g. tossing a coin) allows to reveal the underlying distribution (e.g. the probability that a coin lands Heads). Technically this is done by looking at the "time averages" or "empirical means" of the sequence of the experiments, that is RVs of the form $\frac{1}{N}\sum_{n=1}^{N} X_n$, which as $N \to \infty$ tend (in some appropriate sense) the sample space average, namely the expectation, $E[X_1]$. At first sight, it may not seems as "revealing" the distribution, just its expectation, but as we will see this is enough. A great deal of statistics is based on this type of results: even if we do not know the distribution of something (proportion of left handed people in the population), we can recover it by repeatedly sampling people independently. The empirical proportions will get closer and closer to the answer.

We begin with some basic inequalities, a discussion on convergence of random series, the weak law of large numbers and the strong law. We will close with applications.

## 3.1 Markov's & ChebyChev's Inequalities

Very often expectations are more accessible than probabilities. For example, we can explicitly calculate the expectation of many functions of the normal RVs, but (with one exception) there's no closed form formula for the probability a normal RV is larger than a certain real number... This is a typical situation.

The most effective tools for getting bounds on probabilities are the Makrov inequality and is many derivatives. It may seem too naive, but it can go pretty far.

**Proposition 3.1.1** (Markov's inequality)**.** *Let $Z$ be a nonnegative RV. Then for $z>0$,*

$$zP(Z\ge z)\le E[Z,Z\ge z]\le E[Z].$$

The proof is a one-liner obtained by taking expectations on the following two inequalities:

$$z\mathbf{1}_{\{Z\ge z\}}\le Z\mathbf{1}_{\{Z\ge z\}}\le Z.$$

The bounds Makrov's inequality provides are usually very crude, but often we can improve them significantly.

**Corollary 3.1.1.** *Let $\varphi:\mathbb{R}\to[0,\infty)$ be nondecreasing. Then for every $z>0$*

$$P(Z\ge z)\le\frac{E[\varphi(Z)]}{\varphi(z)}.$$

To prove the corollary, observe that $\{Z\ge z\}\subseteq\{\varphi(Z)\ge\varphi(z)\}$. Now apply Makrov's ineqaulity. To get the best bound from this method we can try minimize the righthand side in the corollary. There's an entire industry of inequalities based on this. Let's present some of the most typical.

**Definition 3.1.1.** *Let $X$ be a RV. The Moment Generating Function of $X$, $\varphi_X$, is defined as*

$$\varphi_X(t):=E[e^{tX}],\ t\in\mathbb{R}.$$

Applying the Corollary, we have that for anty $x>0$,

$$P(X\ge x)\le\inf_{t>0}\varphi_X(t)e^{-tz}.$$

**Example 3.1.1.** *Let $X\sim Pois(\lambda)$. Then*

$$\varphi_X(t)=e^{-\lambda}\sum_{k=0}^{\infty}e^{tk}\frac{\lambda^k}{k!}=\lambda(e^t-1).$$

*Therefore*

$$P(X\ge x)\le\inf_{t>0}\exp(\lambda(e^t-1)-tx).$$

*the exponent we need to minimize is equal to $0$ at $t=0$ and to $\infty$ as $t\to\infty$. Therefore a minimum exists. The derivative is $\lambda e^t-x$, which has one root. Plug this solution into the expression to obtain*

$$P(X\ge x)\le\exp(x-\lambda-x\ln(x/\lambda))=e^{-\lambda}e^{x(1-\ln(x/\lambda))}.$$

*The bound is useless ($\ge 1$) for $x\le\lambda$, but for large $x$ it is pretty good. It can be also used obtain an upper bound on the error term in the Taylor expansion for $x\to e^{\lambda x}$ (why?).*

Another inequality that's very handy is Chebychev's inequality. To introduce it we need to quickly review the notion of the variance. The variance is a very important concept, and will be a key figure on the developments below. The reason is that among all nonlinear functions, the square is the most accessible.

**Definition 3.1.2.** *Let $X$ be a RV. If $X$ is integrable, define the variance of $X$, $Var(X)$ as*

$$Var(X) := E[(X - E[X])^2].$$

**Proposition 3.1.2.** *Let $X$ be a RV. Then*

1. *$Var(X) \in [0, \infty]$ and is finite if and only if $E[X^2] < \infty$ and is zero if and only if the distribution of $X$ is $\delta_x$ for some $x \in \mathbb{R}$.*

   *In the next parts we assume $Var(X) < \infty$.*

2. *$Var(X) = E[X^2] - E[X]^2$.*
3. *For a constant $c \in \mathbb{R}$, $Var(cX) = c^2\, Var(X)$ .*
4. *If $X$ and $Y$ are independent and with finite variance, then*

   $$Var(X + Y) = Var(X) + Var(Y).$$

   *In particular for any constant $c \in \mathbb{R}$, $Var(X + c) = Var(X)$*

We have the following.

**Proposition 3.1.3** (Chebychev's inequality)**.** *Let $X$ be a RV with finite variance. Then for $x > 0$,*

$$P(|X - E[X]| \ge x) \le \frac{Var(X)}{x^2}.$$

*Proof.* Apply Corollary 3.1.1 to $Z = |X - E[X]|$ with $\varphi(z) = \mathbf{1}_{[0,\infty)}(z) z^2$. □

This very simple result is quite important. It is clear that the variance measures deviation from the expectation, but inequality adds a quantitative layer to it. In fact, letting $\sigma := \sqrt{\text{Var}(X)}$, the standard deviation of $X$, and replacing $x$ with $\sigma x$, the inequality becomes

$$P(|X - E[X]| \ge \sigma x) \le \frac{1}{x^2},$$

Thus, measuring the distance from $E[X]$ by units of standard deviation "standardizes" the upper bound in Chebychev's inequality. As with Markov's inequality, Chebychev's inequality is by no means tight, but it is something we can work with.

## 3.2 Convergence in Probability

### 3.2.1 Basic Properties

The proof of the result in the title in its weakest form is immediate from Chebychev's inequality. However, we will first describe a mode of convergence of RVs which we will use to describe the behavior observed in the result.

**Definition 3.2.1.** *Let $(Z_n : n \in \mathbb{N})$ and $Z$ be RVs. We say that the sequence $(Z_n : n \in \mathbb{N})$ convergence to $Z$ in probability if for every $\epsilon > 0$*

$$\lim_{n \to \infty} P(|Z_n - Z| > \epsilon) = 0.$$

Convergence in probability does not imply a.s. convergence.

**Example 3.2.1.** *Let $(A_n : n \in \mathbb{N})$ be a sequence of events with $\lim_{n\to\infty} P(A_n) = 0$.*

1. *Then the sequence of indicators $\mathbf{1}_{A_n}$ converges in probability to the constant RV 0.*

2. *If, in addition, the events are independent and $\sum_{n=1}^\infty P(A_n) = \infty$, then it follows from the Borel-Cantelli II lemma that $\limsup_{n\to\infty} \mathbf{1}_{A_n} = 1$ a.s., therefore the sequence does not converge to $0$ a.s.*

As the next example shows that convergence in probability does not imply convergence of the expectations.

**Example 3.2.2.** *Let $Z$ be any RV taking values in $\mathbb{R}$. Let $Z_n := \mathbf{1}_{\{|Z|\le n\}} Z$.*

1. *Then $\{|Z_n - Z| > \epsilon\} \subseteq \{|Z| > n\}$. Yet, $P(|Z| > n) = 1 - P(|Z| \le n) \to 1 - P(|Z| < \infty) = 1 - 1 = 0$. Therefore $(Z_n : n \in \mathbb{N})$ converges to $Z$ in probability.*

2. *If $Z$ is integrable, then $E[|Z_n - Z|] = E[Z, |Z| > n] \to 0$ due to dominated convergence. However, if $Z$ is nonnegative and not integrable, then $E[|Z_n - Z|] = E[Z, Z > n] \ge \int_n^\infty P(Z \ge t)dt = \infty$.*

3. *Even worse, let $X$ be a nonintegrable RV, and for each $n \in \mathbb{N}$, let $B_n \sim Bern(\frac{1}{n})$. Let $Z_n := B_n X$. Then $(Z_n : n \in \mathbb{N})$ converges in probability to $0$, but $Z_n$ is not integrable.*

There are some connections between the convergence modes we mentioned above:

**Theorem 3.2.1.** *Let $(Z_n : n \in \mathbb{N})$ and $Z$ by RVs.*

1. *If $\lim_{n\to\infty} E[|Z_n - Z|^p] = 0$ for some $p > 0$, then then $(Z_n : n \in \mathbb{N})$ converges to $Z$ in probability.*

2. *If $\lim_{n\to\infty} Z_n = Z$ a.s., then $(Z_n : n \in \mathbb{N})$ converges to $Z$ in probability.*

3. *If $(Z_n : n \in \mathbb{N})$ converges to $Z$ in probability, then there exists a subsequence $(Z_{n_k} : k \in \mathbb{N})$ which converges to $Z$ a.s.*

*Proof.* 1. This follows directly from Markov's inequality.

2. Because $(|Z_n - Z| : n \in \mathbb{N})$ converges to zero a.s. for every $\epsilon > 0$, and for every $\omega \in \Omega$ possibly except on a set of measure zero, there exists $N = N(\epsilon, \omega)$ such that $|Z_n - Z| \le \epsilon$ for $n > N(\epsilon, \omega)$. Let $A_n = \{N(\epsilon, \omega) \le n\}$. $A_1 \subseteq A_2 \subseteq \dots$ and $P(\cup_{n=1}^\infty A_n) = 1$. Therefore $P(A_n) \to 1$. Equivalently, $P(A_n^c) \to 0$. But $\{|Z_n - Z| > \epsilon\} \subseteq A_n^c$, and the result follows.

3. Let $n_1 := \min\{n : P(|Z_n - Z| > \frac{1}{2}) < \frac{1}{2}\}$. Continue inductively, letting

$$n_{k+1} := \min\{n > n_k : P(|Z_n - Z| > 2^{(-k-1)}) < 2^{-k-1}\}.$$

Now let $A_k := \{|Z_{n_k} - Z| > 2^{-k}\}$. By construction $P(A_k) \le 2^{-k}$ and so by Borel-Cantelli, we have that $P(A_k \text{ i.o.}) = 0$, or $|Z_{n_k} - Z| \le 2^{-k}$ eventually, a.s. That is $\lim_{n\to\infty} Z_{n_k} = Z$ a.s. □

### 3.2.2 Convergence Theorems

It important to record the following results.

**Proposition 3.2.1** (Fatou's Lemma (in probability))**.** *Let $(X_n : n \in \mathbb{N})$ and $X$ be nonnegative RVs and suppose that $(X_n : n \in \mathbb{N})$ converges to $X$ in probability. Then $\liminf E[X_n] \ge E[X]$.*

*Proof.* Let $(n_k : k \in \mathbb{N})$ be a subsequence attaining the $\liminf$. Then there exists a further subsequence along which the RVs converge a.s. to $X$. Apply Fatou's lemma to that sequence to obtain the conclusion. □

**Proposition 3.2.2** (Dominated Convergence Theorem (in probability))**.** *Let $(X_n : n \in \mathbb{N})$, $X$ and $Y$ be RVs with $Y$ integrable, $|X_n| \le Y$ a.s. and $(X_n : n \in \mathbb{N})$ converging to $X$ in probability. Then $\lim_{n\to\infty} E[|X_n - X|] \to 0$.*

**Exercise 3.2.1.** *Prove the theorem.*

## 3.3 Uniform Integrability and Vitali's Theorem

In this section we will provide the Vitali's Theorem, the "ultimate" convergence theorem: necessary and sufficient condition to guarantee the conclusion of the dominated convergence theorem. So why do we still used DCT so heavily? The necessary and sufficient condition is uniform integrability. The domination condition is a relatively accessible condition to verify that implies uniform integrability.

We begin with a simple lemma that's quite central in the sense that it gives us some "uniformity" property of integrable RVs.

**Lemma 3.3.1.** *Let $X$ be an integrable RV. Then for every $\epsilon > 0$ there exists $\delta > 0$ such that $E[|X|, A] < \epsilon$ for every event $A$ with $P(A) < \delta$.*

*Proof.* Then for any event $A$, and $M > 0$, we have

$$E[|X|] \le E[|X|, A \cap \{|X| \le M\}] + E[|X|, |X| > M].$$

Noe pick $\epsilon > 0$. The DCT asserts that we can find $M$ large enough so that the second summand on the righthand side is bounded above by $\epsilon/2$. The first summand is bounded above by $MP(A)$, and therefore picking $\delta$ such that $M\delta < \epsilon/2$ then gives $E[|X|, A] \le \frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$, completing the proof. □

We now define the main concept of this section.

**Definition 3.3.1.** *Let $(X_n : n \in \mathbb{N})$ be a sequence of RVs. The sequence is uniformly integrable (UI) if*

$$\lim_{M \to \infty} \sup_n E[|X_n|, |X_n| > M] = 0.$$

**Proposition 3.3.1.** *A sequence of integrable RVs $(X_n : n \in \mathbb{N})$ is UI if and only if*

$$\lim_{M \to \infty} \limsup_{n \to \infty} E[|X_n|, |X_n| > M] = 0.$$

*Proof.* If the sequence is UI, then the stated condition automatically holds because lim sup of a sequence is $\le$ than its sup. Next, suppose that the stated condition holds. Fix some $\epsilon > 0$. Then there exists $M_1 = M_1(\epsilon)$ such that $\limsup_{n\to\infty} E[|X_n|, |X_n| > M] < \epsilon/2$. Therefore, there exists some $N = N(M_1, \epsilon)$ such that for $n \ge N$, $E[|X_n|, |X_n| > M_1] \le \epsilon$. But since each of the RVs $X_1, \dots, X_N$ are integrable, the dominated convergence theorem guarantees the exists of $M_2$ such that $E[|X_j|, |X_j| > M'] \le \epsilon$ for all $j \in \{1, \dots, N\}$. Thus, for every $M \ge \max(M_1, M_2)$, we have that $E[|X_n|, |X_n| > M] < \epsilon$ for all $n \in \mathbb{N}$. As $\epsilon > 0$ is arbitrary, the result follows. □

The domination condition in the DCT implies UI:

**Example 3.3.1.** *Suppose that $(X_n : n \in \mathbb{N})$ is a sequence of RVs and $Z$ is an integrable RV with $|X_n| \le Z$ for all $n \in \mathbb{N}$. Then $(X_n : n \in \mathbb{N})$ is UI. Indeed, for every $n \in \mathbb{N}$, $E[|X_n|, |X_n| > M] \le E[Z, Z > M]$. By dominated convergence, the righthand side tends to zero, and the result follows.*

However, there are simple examples of UI sequences which are not dominated by any integrable RV.

**Example 3.3.2.** *Let $(I_n : n \in \mathbb{N})$ be disjoint intervals in $[0,1]$ with $\cup_{n=1}^{\infty} I_n = [0,1]$, and with length $|I_n| = c/n^2$ for all $n \in \mathbb{N}$ (thus $c = (\sum_{n=1}^{\infty} \frac{1}{n^2})^{-1} = 6/\pi^2$). Let $U \sim U[0,1]$ and define $X_n := n\mathbf{1}_{I_n}(U)$. Then for each $M > 0$, $E[|X_n|, |X_n| > M]$ if $0$ if $M > n$, and is equal to $nP(A_n) = c/n$ otherwise. Therefore the supremum over $n \in \mathbb{N}$ is bounded above by $c/M$, and tends to $0$ as $M \to \infty$. However, on the event $\{X_n > 0\}$, $X_m = 0$ for all $m \ne n$, which implies that if $Z \ge X_n$ for all $m$, then $Z \ge \sum_{n=1}^{\infty} X_n$. The expectation of the righthand side is $c \sum_{n=1}^{\infty} \frac{1}{n} = \infty$. Therefore there does not exist an integrable RV dominating the sequence.*

Here is a nice criterion for uniform integrability.

**Proposition 3.3.2.** *Suppose that $\varphi(x)$ is a function satisfying $\lim_{|x|\to\infty} \varphi(x)/|x| = \infty$. If $\sup_n E[\varphi(X_n)] < \infty$, then $(X_n : n \in \mathbb{N})$ is UI.*

*Proof.* Let $\epsilon > 0$ and let $M$ be such that $\varphi(x)/|x| \geq \frac{1}{\epsilon}$ for $|x| > M$. Then

$$E[\varphi(X_n)] \geq E[|X_n|\varphi(X_n)/|X_n|, |X_n| > M] \geq \frac{1}{\epsilon} E[|X_n|, |X_n| > M].$$

Therefore,

$$\sup_n E[|X_n|, |X_n| > M] \leq \epsilon \sup_n E[\varphi(X_n)].$$

The result now follows by taking $M \to \infty$ as $\epsilon$ is arbitrary. □

**Example 3.3.3.** *Let $(X_n : n \in \mathbb{N})$ be as in Exampled 3.3.2. Take $\varphi(x) = X^2$. Then $E[\varphi(X_n)] = E[X_n^2] = n^2c/n^2 = c$, and the UI follows from the proposition.*

We'd like to give one more result before our main course. It's a nice and simple property of UI which will help down the road when we discuss the optional sampling theorem, for example, and is a generalization of Lemma 3.3.1.

**Proposition 3.3.3.** *Let $(X_n : n \in \mathbb{N})$ be UI. Then for every $\epsilon > 0$ there exists $\delta > \delta(\epsilon)$ such that $\sup_n E[|X_n|, A] < \epsilon$ for all events $A$ with $P(A) < \delta$.*

*Proof.* The UI assumption reduces the proof to an argument identical to the proof of Lemma 3.3.1. For any event $A$ and $M > 0$,

$$E[|X_n|, A] \leq E[|X_n|, A \cap \{|X_n| \leq M\}] + E[|X_n|, |X_n| > M] \leq MP(A) + E[|X_n|, |X_n| > M].$$

Take $M$ large enough so that the second summand on the righhand side is $\leq \epsilon/2$ for all $n \in \mathbb{N}$, which is what we get from UI for free. Next all we need now is to pick $\delta$ such that $M\delta < \epsilon/2$, and our work is done. □

Now for the promised result.

**Theorem 3.3.1** (Vitali's Convergence Theorem)**.** *Let $(X_n : n \in \mathbb{N})$ be RVs converging in probability to an integrable random variable $X$. Then the following are equivalent.*

1. *$(X_n : n \in \mathbb{N})$ is UI.*
2. *$\lim_{n\to\infty} E[|X_n|] = E[|X|]$.*
3. *$\lim_{n\to\infty} E[|X_n - X|] = 0$.*

*Proof.*

$\underline{1 \Rightarrow 2}$. By Fatou's lemma $\liminf E[|X_n|] \geq E[|X|]$. By UI,

$$E[|X_n|] = E[|X_n|, |X_n| \leq M] + E[|X_n|, |X_n| > M],$$

and so

$$\limsup E[|X_n|] \leq \limsup E[|X_n|, |X_n| \leq M] + \sup_n E[|X_n|, |X_n| > M].$$

Clearly $|X_n|\mathbf{1}_{\{|X_n| \leq M\}} \leq |X_n|\mathbf{1}_{\{|X_n| < M+1\}} \leq M + 1$, (we replaced $M$ by $M + 1$ and replaced the weak inequality with a strict inequality (can you see why?). Now apply the DCT for convergence in probability on the sequence in the middle of the inequalities to conclude that $\limsup E[|X_n|, |X_n| \leq M] \leq E[|X|, |X| \leq M + 1]$. Thus,

$$\limsup E[|X_n|] \leq E[|X|, |X| \leq M + 1] + \sup_n E[|X_n|, |X_n| > M].$$

Now take $M \to \infty$ on the righthand side. Monotone convergnece applied to the first summand and the UI applied to the second summand give

$$\limsup E[|X_n|] \leq E[|X|].$$

2 ⇒ 3. By the triangle inequality, $|X_n|+|X|-|X_n-X|\ge 0$. By Fatou's lemma, $\liminf E[|X_n|+|X|-|X_n-X|]\ge E[2|X|]$. By the lefthand side is $E[2|X|]-\limsup E[|X_n-X|]$, and thed result follows from the algebra.

3 ⇒ 1. Again, triangle inequality gives

$$E[|X_n|,|X_n|>M]\le E[|X-X_n|,|X_n|>M]+E[|X|,|X_n|>M]. \tag{3.3.1}$$

We now bound each of the summands on the righthand side. To do that, fix $\epsilon>0$. Then there exists $N=N(\epsilon)$ such that $E[|X_n-X|]<\epsilon/2$. Next, by Markov's and the trjangle inequality,

$$P(|X_n|>M)\le E[|X_n|]/M\le E[|X_n-X|+|X|]/M,$$

but since $\sup_n E[|X_n-X|]<\infty$ and $X$ is integrable, there exists a constant independent of $n$, such that

$$\sup_n P(|X_n|>M)\le C/M.$$

As a result of Lemma 3.3.1, we can therefore find $M=M(\epsilon)$ large enough so that $E[|X|,|X_n|>M]\le\epsilon/2$ for all $n$. Putting both together, we have

$$E[|X_n|,|X_n|>M]\le\epsilon/2+\epsilon/2,\ n\ge N(\epsilon),M>M(\epsilon).$$

This therefore gives $\limsup_{n\to\infty}E[|X_n|,|X_n|>M]\le\epsilon$ if $M>M(\epsilon)$. By letting $M\to\infty$, we then obtain $\lim_{M\to\infty}\limsup_{n\to\infty}E[|X_n|,|X_n|>M]=0$. As $E[|X_n|]\le E[|X_n-X|]+E[|X|]<\infty$, Proposition 3.3.1 gives the result.

□

## 3.4 Weak Law of Large Numbers

The proof of this result is a an immediate application of Chebychev's inequality. It is weak because it gives convergence in probability, not almost surely.

**Theorem 3.4.1** (Weak Law of Large Numbers)**.** *Let $(X_n:n\in\mathbb{N})$ be independent and uncorrelated with finite expectation $\mu$ and finite nonzero variance $\sigma^2$. For $n\in\mathbb{N}$, let $\bar{X}_n:=\frac{X_1+\cdots+X_n}{n}$ be the empirical mean. Then for every $\epsilon>0$,*

$$P(|\bar{X}_n-\mu|\ge\epsilon)\le\frac{\sigma^2}{n\epsilon^2}.$$

*In particular, $(\bar{X}_n:n\in\mathbb{N})$ converges in probability to $\mu$.*

*Proof.* Apply Chebychev to $\bar{X}_n$, noting that $E[\bar{X}_n]=\mu$ and $\mathrm{Var}(\bar{X}_n)=\frac{n\mathrm{Var}(X_1)}{n^2}$. □

Even with this very simple result, we can prove some intersecting things.

Let's take some continuous function $f:[0,1]\to\mathbb{R}$. Now let $(X_n:n\in\mathbb{N})$ be some IID sequence taking values in $[0,1]$ with mean $x$. Let's estimate $E[f(\bar{X}_n)]$, and see how far it is from $f(\mu)$. Because $f$ is continuous on a closed bounded interval it is bounded, $\|f\|:=\max_{x\in[0,1]}|f(x)|<\infty$ and it is uniformly continuous: for any $\epsilon>0$ there exists $\delta>$ such that if $|x-y|<\delta$, then $|f(x)-f(y)|<\epsilon$. This gives

$$|E[f(\bar{X}_n)-f(\mu)]|\le\epsilon P(|\bar{X}_n-\mu|<\delta)+2\|f\|_\infty P(|\bar{X}_n-\mu|\ge\delta).$$

Applying Chebychev's inequality, we have

$$|E[f(\bar{X}_n)]-f(\mu)|\le\epsilon+2\|f\|_\infty\frac{\mathrm{Var}(X_1)}{n\delta^2}. \tag{3.4.1}$$

Note that the lefthand side is independent of the choice of $\epsilon$. Now let $x\in[0,1]$ and pick $X_1\sim\mathrm{Bern}(x)$. Then $\mathrm{Var}(X_1)=x(1-x)\le\frac{1}{4}$.

Observe that

$$E[f(\bar{X}_n)] = \sum_{k=0}^{n} f(k/n) \binom{n}{k} x^k (1-x)^{n-k}.$$

Note that as a function of $x$, this is a polynomial, we denote it by $p_{f,n}(x)$. Therefore, (3.4.1) becomes

$$|p_{f,n}(x) - f(x)| \le \epsilon + 2\|f|_\infty \frac{1}{4n\delta^2}. \tag{3.4.2}$$

The righthand side is independent of $x$, and therefore

$$\limsup_{n\to\infty} \max_{x\in[0,1]} |p_{f,n}(x) - f(x)| \le \epsilon.$$

As $\epsilon$ was arbitrary, we proved the following:

**Theorem 3.4.2** (Bernstein's Polynomials)**.** *Let $f$ be a continuous function on $[0,1]$. For $n \in \mathbb{N}$, define the Bernstein polynomial $p_{f,n}(x) := \sum_{k=0}^n f(\frac{k}{n})\binom{n}{k}x^k(1-x)^{n-k}$. Then $(p_{f,n} : n \in \mathbb{N})$ converges uniformly to $f$.*

Let's get another classic result: confidence intervals. The problem here is to find a distribution through repeated sampling. Let's suppose that in a sequence of experiments we're looking at a certain feature (e.g. person sampled is left-handed), and that our resulting IID sequence is the respective indicators of that feature for the samples. Therefore, the distribution of $X_1 + \cdots + X_n$ is binomial with parameters $n$ and $p$, but $p$ is unknown. Now, we already know that $\text{Var}(X_1) = p(1-p) \le \frac{1}{4}$, a fact we used in our derivation of the result above. So Chebychev gives us:

$$P(|\bar{X}_n - p| \ge \epsilon) \le \frac{1}{4n\epsilon^2}.$$

Rewrite this as a lower bound on the probability of the complement gives the following:

**Theorem 3.4.3.** *Let $(X_n : n \in \mathbb{N})$ be IID Bern$(p)$. Fix $\epsilon > 0$. Then*

$$P(p \in (\bar{X}_n - \epsilon, \bar{X}_n + \epsilon)) \ge 1 - \frac{1}{4n\epsilon^2}.$$

The righthand side is independent of $p$. The meaning of this statement is this: Fix $\epsilon > 0$ and $n$ and consider the random interval with width $2\epsilon$, $\bar{X}_n - \epsilon, \bar{X}_n + \epsilon)$. This random interval is known as the confidence interval. Without any assumptions on $p$, it lies in the confidence interval, with probability larger or equal to the expression on the right and which is only a function of our pre-determined parameters $\epsilon$ and $n$. This probability is often referred to the confidence level (95% confidence = 0.95 probability). After we cover the strong law of large numbers we will be able this probability, you'll be convinced that the practical meaning of this probability is that if we repeat this procedure many many times, then the confidence level gives the proportion of times that the unknown parameter was within the confidence interval. In order words, we obtained a method for estimating the unknown $p$.

To give an illustration: the unknown parameter is within 5% (0.05) from our empirical mean with confidence level which is at least $1 - \frac{1}{4n0.05^2} = 1 - \frac{100}{n}$. It is enough to take 2000 samples to have a confidence level of at least 95%. This bound is pretty crude, but it shows that a relatively small sample can give a lot of information!

One can use Chebychev's inequality under less restrictive conditions. In fact, even when the RVs don't have a finite expectation. This is done through truncation. Here's one such result, which presents the idea. We will push this idea further in the next section.

**Theorem 3.4.4.** *Suppose that $(X_n : n \in \mathbb{N})$ are identically distributed and independent RVs satisfying*

$$\lim_{x\to\infty} xP(|X_1| > x) = 0.$$

*Let $\mu_n := E[X_1, |X_1| \le n]$. Then $(\bar{X}_n - \mu_n : n \in \mathbb{N})$ converges in probability to $0$.*

*Proof.* Let $X_{k,n} := X_k \mathbf{1}_{\{|X_k|\le n\}}$, and let $\bar{S}_n = \frac{1}{n}\sum_{k=1}^n X_{k,n}$. Then $E[\bar{S}_n] = \mu_n$ and $\text{Var}(\bar{S}_n) = \frac{\text{Var}(X_{1,n})}{n}$. Let $A_n := \{|\bar{S}_n - \mu_n| \ge \epsilon\}$, Chebychev's inequality gives

$$P(A_n) \le \text{Var}(X_{1,n})/(n\epsilon^2).$$

Now $\text{Var}(X_{1,n}) \le E[X_1^2, |X_1| \le n] \le \int_0^{n^2} P(X_1^2 \ge t)dt = 2\int_0^n tP(|X_1| \ge t)dt$. By assumption, this is $o(n)$ as $n \to \infty$.

Let $B_n := \{\bar{X}_n \ne \bar{S}_n\} \subseteq \cup_{k=1}^n \{|X_{k,n}| \ge n\}$. Therefore, $P(\bar{X}_n \ne \bar{S}_n) \le nP(|X_1| \ge n) \to 0$.

Clearly, $\{|\bar{X}_n - \mu_n| \ge \epsilon\} \subseteq (A_n \cap B_n^c) \cup B_n \subseteq A_n \cup B_n$. As $P(A_n) + P(B_n) \to 0$, the result follows. □

**Example 3.4.1.** *Lets assume that for $x \ge 0$, $P(X_1 \ge x) = \frac{e}{(e+x)\ln(e+x)}$ (note that in this case $X$ has a density. What is it?). Clearly, the condition in the theorem holds, but we also notice that the expectation of $X_1$ is infinite because $\int_0^\infty P(X_1 \ge t)dt = \infty$. Now*

$$\mu_n = \int_0^n P(X_1 \ge t)dt = e\ln\ln(e+n),$$

*and therefore $\bar{X}_n - e\ln\ln(e+n) \to 0$ in probability. Note that this implies, but is a stronger statement than $\frac{\bar{X}_n}{\ln\ln n} \to e$ in probability.*

## 3.5 Kolmogorov's Maximal Inequality

There is a nice and clever way to get a result much stronger than Chebychev's inequality, under similar assumptions.

**Theorem 3.5.1** (Kolmogorov's Maximal Inequality)**.** *Let $X_1, \dots, X_n$ be independent, each with zero expectation and finite variance. For $n \in \mathbb{N}$, let $S_n := \sum_{k=1}^n X_k$ be the partial sum. Then for every $x > 0$,*

$$P(\max_{k\le n} |S_k| \ge x) \le \frac{E[S_n^2]}{x^2}.$$

*Proof.* Let $A$ be the event in question:

$$A := \{\max_{k\le n} |S_k| \ge x\}.$$

We can write it as a disjoint union according to the first $k$ satisfying $S_k \ge x$. More precisely, let

$$A_k := \{|S_k| \ge x, \max_{l<k} |S_l| < x\}.$$

Then $A = \cup_{k=1}^n A_k$, and the union is disjoint. Now use domain additivity (or additivity of the integral to write):

$$E[S_n^2, A] = \sum_{k=1}^n E[S_n^2, A_k].$$

To estimate each summand on the righthand side, write

$$S_n^2 = (S_k + (S_n - S_k))^2 \ge S_k^2 + 2(S_n - S_k).$$

Now note that $A_k \in \sigma(X_1, \dots, X_k)$, and $S_n - S_k \in \sigma(X_{k+1}, \dots, X_n)$. These two $\sigma$-algebras are independent, and so $E[S_n - S_k, A_k] = E[S_n - S_k]P(A_k) = 0 * P(A_k)$. This gives

$$E[S_n^2] \ge \sum_{k=1}^n E[S_k^2, A_k] \ge x^2 \sum_{k=1}^n P(A_k) = x^2 P(A).$$

□

**Example 3.5.1.** *Let $(X_n : n \in \mathbb{N})$ be IID with mean $\mu$ and finite variance $\sigma^2 > 0$. Let $M_n = \max_{k \le n} |S_k - E[S_k]|$. From the theorem we have $P(M_{l^2} \ge \epsilon l) \le \frac{l^2\sigma^2}{\epsilon^2 l^4} = \frac{\sigma^2}{\epsilon^2 l^2}$. From Borel-Cantelli II we have $P(M_{l^2}/l^2 \ge \epsilon \text{ i.o.}) = 1$, or $\limsup \frac{M_{l^2}}{l^2} \le \epsilon$ a.s. Since this is true for all rational $\epsilon$, it follows that $\lim \frac{M_{l^2}}{l^2} = 0$ a.s. Now for every $n \in \mathbb{N}$ there's a unique $l \in \mathbb{N}$ such that $l^2 \le n < (l+1)^2$. We have*

$$\frac{M_{l^2}}{l^2} \times \frac{l^2}{(l+1)^2} \le \frac{M_n}{n} \le \frac{M_{(l+1)^2}}{(l+1)^2} \times \frac{(l+1)^2}{l^2}.$$

*Take $n \to \infty$ and apply the squeeze theorem to conclude that $M_n/n$ to$0$ a.s. Let's write this.*

$$\frac{\max_{k \le n} |S_k - k\mu|}{n} \to 0 \text{ a.s.}$$

*In particular, $\frac{S_n}{n} \to \mu$ a.s. (a statement we can prove by applying the argument directly to Chyebychev - no need for the maximal inequality). This gives us a strong law, under the assumption of finite variance. Much of work in the proof of the strong law of large numbers would be to get rid of this additional assumption.*

**Exercise 3.5.1.** *Repeat the argument in the example above, but now fixing some $\alpha > 0$ and looking at $M_{l^k}/l^{\frac{1}{2}(k(1+\alpha))}$, for $k$ satisfying $k\alpha > 1$, to conclude that $M_n/n^{1/2+\alpha} \to 0$ a.s.*

Here's another application.

**Theorem 3.5.2** (Kolmogorov's Two Series Theorem)**.** *Let $X_1, X_2, \dots$ be independent with expectation zero. If $\sum_{n=1}^\infty E[X_n^2] < \infty$ then $\sum_{n=1}^\infty X_n$ converges a.s.*

*Proof.* Define the partial sum $S_N := \sum_{n=1}^N X_n$. Now fix $M$ and $n$ and let's look at $P(\max_{k \le n} |S_{M+k} - S_M| \ge x)$. By the maximal inequality, this is bounded above by $E[(S_{M+n} - S_M)^2]/x^2 \le T_M/x^2$, where $T_M := \sum_{k \ge M} E[X_k^2]$. Use the continuity of probability measures from below to obtain

$$P(\sup_{k \in \mathbb{N}} |S_{M+k} - S_M| \ge x) \le T_M/x^2.$$

For $N, N' \ge M$, the triangle inequality gives $|S_N - S_{N'}| \le |S_N - S_M| + |S_M - S'_N|$, and therefore, $\sup_{N,N'} |S_N - S_{N'}| \le \sup_N |S_N - S_M| + \sup_{N'} |S'_N - S_M| \le 2 \sup |S_N - S_M|$. It follows that

$$P(\sup_{N,N' \ge M} |S_N - S_{N'}| \ge 2x) \le P(\sup_{N \ge M} |S_N - S_M| \ge x) \le \frac{T_M}{x^2}.$$

Take $x = \frac{1}{n}$ and observe that the supremum on the lefthand side is decreasing,

$$P(\sup_{N,N' \ge M} |S_N - S_{N'}| < \frac{2}{n}) \ge 1 - n^2 T_M.$$

The event on the left is decreasing as $M \to \infty$. Also, $\lim_{M \to \infty} T_M = 0$. We therefore have

$$P(\lim_{M \to \infty} \sup_{N,N' \ge M} |S_N - S_{N'} < \frac{2}{n}) = 1.$$

As this is true for all $n$, we have proved that $\lim_{M \to \infty} \sup_{N,N' \ge M} |S_N - S_{N'}| = 0$ a.s. That is $(S_N : N \in \mathbb{N})$ is a Cauchy sequence, a.s. As every Cauchy sequene in $\mathbb{R}$ is convergent, the result follows. □

The next example is a true classic. We all know that $\sum_{n=1}^\infty \frac{(-1)^n}{n}$ converges. What if instead of the periodic pattern of signs we will "sprinkle" signs randomly?

**Example 3.5.2.** *Let $(B_n : n \in \mathbb{N})$ be IID Ber$(p)$ for some $p \in (0,1)$. Consider the series $\sum_{n=1}^\infty \frac{(-1)^{B_n}}{n}$. Observe that $E[(-1)^{B_n}] = p - (1-p) = 2p - 1$. Let $X_n := \frac{(-1)^{B_n} - (2p-1)}{n}$. Then $E[X_n] = 0$ and $E[X_n^2] = \frac{1 - (p-(1-p))}{n^2} = \frac{2p(1-p)}{n^2}$. It follows from the theorem that $\sum_{n=1}^\infty X_n = \sum_{n=1}^\infty \left(\frac{(-1)^{B_n}}{n} - \frac{2p-1}{n}\right)$. converges a.s. Since the deterministic series $\sum_{n=1}^\infty \frac{2p-1}{n}$ is either $0$ or infinity according to whether $p = \frac{1}{2}$ or $p \ne \frac{1}{2}$, it follows that our original series converges a.s. if $p = \frac{1}{2}$ and diverges a.s. if $p \ne \frac{1}{2}$.*

## 3.6 Kolmogorov's Three Series Theorem

**Theorem 3.6.1.** *Suppose that* $(X_n : n \in \mathbb{N})$ *is a sequence of independent RVs. Let* $A > 0$ *and let* $Y_n := X_n \mathbf{1}_{\{|X_n| \le A\}}$*. Then the series* $\sum_{n=1}^\infty X_n$ *converges a.s. if and only if all of the following three series converge:*

1. $\sum_{n=1}^\infty P(|X_n| > A)$
2. $\sum_{n=1}^\infty E[Y_n]$.
3. $\sum_{n=1}^\infty E[Y_n^2]$.

*Proof.* Since $\{X_n \ne Y_n\} \subseteq \{|X_n| > A\}$, the convergence of the first series and the Borel-Cantelli lemma imply that $X_n = Y_n$ for all but finitely many $n$-s, a.s. In particular, $\sum_{n=1}^\infty X_n$ converges if and only if $\sum_{n=1}^\infty Y_n$ converges.

We will therefore assume the first series converges, and continue showing the sufficiency of the convergence of the remaining two series.

Let $Z_n = Y_n - E[Y_n]$. Then $E[Z_n] = 0$ and $\text{Var}(Z_n) \le E[Y_n^2]$. As the third series converges, we have that $\sum_{n=1}^\infty Z_n$ converges a.s. As $\sum_{n=1}^\infty E[Y_n]$ converges, we have that $\sum_{n=1}^\infty Y_n = \sum_{n=1}^\infty Z_n + \sum_{n=1}^\infty E[Y_n]$ converges a.s.

We will leave the necessity to a later stage.

□

**Example 3.6.1.** *Let's revisit Example 3.5.2, but in greater generality now. Let* $(a_n : n \in \mathbb{N})$ *be a sequence of positive numbers satisfying* $\lim_{n\to\infty} a_n = \infty$ *(we do not assume monotonicity), and let* $(B_n : n \in \mathbb{N})$ *be IID Bern*$(p)$ *for some* $p \in (0,1)$*. We want to consider the convergence of the random series* $\sum_{n=1}^\infty \frac{(-1)^{B_n}}{a_n}$*. Without loss of generality we can assume* $a_n \ge 1$ *for all* $n$*. We apply the three series theorem. By choosing* $A = 1$*, all the terms of the first series are zero. Also,* $Y_n = X_n$*. Now* $E[X_n] = \frac{1-2p}{a_n}$*, and* $E[X_n^2] = \frac{1}{a_n^2}$*. Let's continue case by case:*

1. *If* $p = \frac{1}{2}$*, then all three series converge a.s. if and only if* $\sum_{n=1}^\infty \frac{1}{a_n^2} < \infty$.
2. *If* $p \ne \frac{1}{2}$*. The second series convergence if and only if* $\sum_{n=1}^\infty \frac{1}{a_n}$ *converges, which then implies that* $\sum_{n=1}^\infty \frac{1}{a_n^2}$ *converges. Thus, in this case our series converges a.s. if and only if* $\sum_{n=1}^\infty \frac{1}{a_n}$ *converges. Or: in this case the random series converges if and only if it converges absolutely. Not really interesting.*

*Finally, due to Kolmogorov's zero-one law, if the series does not converge a.s. then it diverges a.s.*

## 3.7 Strong Law of Large Numbers

The difference from the results presented earlier is that the strong law gives convergence almost surely. We will present one proof, based on Kolmogorov's maximal inequality and truncation.

Let's begin with a calculus result.

**Lemma 3.7.1** (Kronecker's Lemma)**.** *Let* $\sum_{n=1}^\infty a_n$ *be a convergent sequence. Let* $0 \le b_1 \le b_2 \le \dots$ *satisfy* $\lim_{m\to\infty} b_n = \infty$*. Then* $\lim_{n\to\infty} \frac{1}{b_n} \sum_{k-1}^n b_k a_k = 0$.

*Proof.* Let $T_n := \sum_{k=1}^n a_k$. The convergence of the series is equivalent to $(T_n : n \in \mathbb{N})$ being a convergent sequence and in particular it is bounded by some constant $C$ and is a Cauchy sequence. That is, for any $\epsilon > 0$ there exists $N = N(\epsilon)$ such that $\sup_{N \le n' \le n} |T_n - T_{n'}| \le \epsilon$. Set $T_0 := 0$ and define $T_{n,n'} := T_n - T_{n'}$. Observe that

$$\begin{aligned}\sum_{k=1}^n b_k a_k &= b_1 T_{0,n} + (b_2 - b_1) T_{1,n} + \dots + (b_n - b_{n-1}) T_{n,n} \\ &= \sum_{k<N} (b_k - b_{k-1}) T_{k,n} + \sum_{k \ge N} (b_k - b_{k-1}) T_{k,n} = (*)\end{aligned}$$

Apply the triangle inequality to obtain

$$|(*)| \leq 2Cb_N + \epsilon b_n.$$

Divide by $b_n$ to obtain

$$\frac{(*)}{b_n} \leq \frac{2Cb_N}{b_n} + \epsilon.$$

Take $n \to \infty$ and then $\epsilon \to 0$. □

**Theorem 3.7.1.** *Suppose that $(X_n : n \in \mathbb{N})$ are independent RVs satisfying $\sum_{n=1}^\infty \frac{1}{n^2} Var(X_n) < \infty$. Let $\bar{X}_n := \frac{1}{n}\sum_{k=1}^n X_k$, and let $\mu_n := E[\bar{X}_n]$. Then*

$$\lim_{n\to\infty} \bar{X}_n - \mu_n = 0 \text{ a.s.}$$

*Proof.* Let $Z_n := X_n - E[X_n]$. Then $E[Z_n] = 0$, and we also have $\sum_{n=1}^\infty \text{Var}(\frac{Z_n}{n}) < \infty$. Therefore Theorem 3.5.2 gives $\sum_{n=1}^\infty \frac{1}{n} Z_n < \infty$ a.s. Use this and Kronecker lemma to conclude that $\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^n Z_k = 0$ a.s.. □

The conclusion does not have any reference to the variance. It was just a technical condition. We will get rid of it through truncation. For this we need yet another calculus lemma.

**Lemma 3.7.2.** *Suppose that $(a_n :\in \mathbb{N})$ is a nonnegative sequence satisfying $\sum_{m=1}^\infty a_m/m < \infty$. Then*

$$\sum_{n=1}^\infty n^{-2}\sum_{m=1}^n a_m < \infty.$$

*Proof.* Without loss of generality we may assume that $a_m \geq 0$ for all $m$. Write the inner sum as $\sum_{m=1}^\infty \mathbf{1}_{[0,n]}(m)a_m$, and move $n^{-2}$ inside this sum. Now use the monotone convergence theorem to conclude that the double sum is equal to

$$\sum_{m=1}^\infty \sum_{n\geq m} \frac{1}{n^2} a_m$$

Because $\sum_{n\geq m}\frac{1}{n^2} \leq \frac{C}{m}$, it follows that the entire sum is bounded above by $C\sum_{m=1}^\infty \frac{a_m}{m}$ □

**Corollary 3.7.1.** *Let $X$ be integrable. Then*

$$\sum_{n=1}^\infty n^{-2} E[|X|^2, |X| \leq n] < \infty.$$

*Proof.*

$$E[|X|^2, |X| \leq n] \leq \sum_{m=1}^n m^2 P(m-1 < |X| \leq m).$$

Let $a_m := m^2 P(m-1 < |X| \leq m)$. Then $a_m/m = mP(m-1 < |X| \leq m)$ and so $\sum_{m=1}^\infty a_m/m \leq E[|X|] + 1$. Therefore, the result follows from the Lemma. □

We are ready for the main result of this section.

**Theorem 3.7.2** (Strong Law of Large Numbers)**.** *Let $(X_n : n \in \mathbb{N})$ be IID and integrable RVs. Let $\mu := E[X]$. Then*

1. $\lim_{n\to\infty} \bar{X}_n = \mu$ *a.s.*
2. $\lim_{n\to\infty} E[|\bar{X}_n - \mu|] = 0.$

*Proof.* Let $Y_n := X_n \mathbf{1}_{\{|X_n|\le n\}}$. Then $\mathrm{Var}(Y_n) \le E[Y_n^2] = E[X_1^2, |X_1| \le n]$. It follows from the Corollary that $\sum_{n=1}^\infty \frac{1}{n^2}\mathrm{Var}(Y_n) < \infty$. Thus theorem 3.7.1 gives that $\bar{Y}_n - \frac{1}{n}\sum_{k=1}^n E[Y_n] \to 0$ a.s. Now $E[Y_n] = E[X_n, |X_n| \le n] = E[X_1, |X_1| \le n] \to E[X_1]$, by dominated convergence, and so $\frac{1}{n}\sum_{k=1}^n E[Y_n] - E[X_1] \to 0$. Next, $P(Y_n \ne X_n) \le P(|X_n| \ge n) = P(|X_1| \ge n)$. But

$$\sum_{n=1}^\infty P(|X_1| \ge n) \le \int_0^\infty P(|X_1| \ge t)dt = E[|X_1|] < \infty$$

, therefore $Y_n = X_n$ eventually a.s. and as a result $\bar{Y}_n - \bar{X}_n \to 0$ a.s. We therefore have $\bar{X}_n - E[X_1] \to 0$ a.s. This prove the first part.

To conclude, let $|\bar{X}|_n := |X_1| + \dots + |X_n|$. Then the first part gives $|\bar{X}|_n \to E[|X_1|]$ a.s. Now look at

$$Z_n := |\bar{X}|_n + E[|X_1|] - |\bar{X}_n - E[X_1]|.$$

By the triangle inequality, $Z_n \ge 0$. Apply Fatou's lemma to obtain

$$\liminf_{n\to\infty} E[Z_n] \ge 2E[|X_1|].$$

But $E[Z_n] = 2E[|X_1]] - E[|\bar{X}_n - E[|X_1|]$, and so $\liminf_{n\to\infty} E[Z_n] = 2E[|X_1|] - \limsup_{n\to\infty} E[|\bar{X}_n - E[X_1]|]$, and the result follows. □

An important application is the following:

**Theorem 3.7.3** (Glivenko-Cantelli)**.** *Let $(X_n : n \in \mathbb{N})$ be IID with distribution function $F$. For $x \in \mathbb{R}$, let $\bar{F}_n(x) = \frac{1}{n}\sum_{k=1}^n \mathbf{1}_{(-\infty,x]}(X_k)$. Then $\sup_{x\in\mathbb{R}} |\bar{F}_n(x) - F(x)| = 0$ a.s.*

Let's parse this: the random function $\bar{F}_n$ is called the empirical distribution function. For each $x$, it gives the proportion of the first samples which are $\le x$. It is a bonafide distribution function, yet random. The expectation of $\bar{F}_n(x)$ is $P(X \le x) = F(x)$. The theorem states that the sequence of empirical distribution functions converges uniformly to $F$, a.s. Or, the underlying distribution can be recovered through repeated samples.

*Proof.* We will give the proof in the special case of $(X_n : n \in \mathbb{N})$ are uniformly distributed, with the general case obtained as a corollary in Assignment #3. In this special case the CDF $F$ is given by

$$F(x) = \begin{cases} 0 & x < 0 \\ x & 0 \le x < 1 \\ 1 & 1 \le x \end{cases}$$

It is clearly uniformly continuous with $|F(x) - F(y)| \le |x - y|$. To prove the theorem in this setting it is enough to consider $x \in [0,1]$, as outside this interval $\bar{F}_n = F$.

Next, for every $l \in \mathbb{N}$ and $k \in \mathbb{Z}_+$ let $x_{l,k} := \frac{k}{l}$. Also define $T_l := \{x_{l,k} : k = 0, \dots, l\}$ and $T := \cup_{l=1}^\infty T_l$. By the strong law of large numbers for each $x \in \mathbb{R}$, $\bar{F}_n(x) \to \bar{F}(x)$ a.s. (this event may depend on $x$). Since $T$ is countable, it follows that $\bar{F}_n - F$ for all $x \in T$, a.s. (as the event in question is a countable intersection of events, each with probability 1). Denote this event by $A$.

Let $\omega \in A$ and $\epsilon > 0$. Pick $l > \frac{1}{\epsilon}$. Note then that the uniform continuity implies $F(x_{l,k}) - F(x_{l,k+1})| < \epsilon$. Next, pick $N = N(\omega, \epsilon)$ such that for all $n \ge N$, $|\bar{F}_n(x) - F(x)| < \epsilon$ for all $x \in T_l$. Now pick any $x \in [0,1]$. Clearly, there exists $k \in T_l$ such that $x_{l,k} \le x \le x_{l,k+1}$. Let's do the calculations:

$$\bar{F}_n(x) - F(x) \le \bar{F}_n(x_{l,k+1}) - F(x_{l,k}) = \bar{F}_n(x_{l,k+1}) - F(x_{l,k+1}) + F(x_{l,k+1}) - F(x_{l,k}) \le \epsilon + \epsilon = 2\epsilon,$$

with the first inequality obtained from the convergence on $T_l$ and the uniform continuity of $F$. A similar argument gives a lower bound of $-2\epsilon$. Therefore we have shown that for any $\omega \in A$ and $\epsilon > 0$, there exists $N = N(\omega, \epsilon) \in \mathbb{N}$ such that for all $n \ge N$, $|\bar{F}_n(x) - F(x)| \le 2\epsilon$ for all $x \in [0,1]$, hence for all $x \in \mathbb{R}$. This complete the proof. □

# Chapter 4

# Central Limit Theorems

## 4.1 Introduction

The law of large numbers gives the connection between the empirical means obtained by repeatedly and independently sampling from some distribution and the distribution itself. It states that the empirical means (or time averages) converge to the expectation, the sample space mean. The central limit theorem provides a correction, describing how far the empirical means are from the expectation. A hint on how to identify this correction is given by the following simple argument.

Suppose that $(X_n : n \in \mathbb{N})$ are IID with expectation $\mu$ and variance $\sigma^2$. Let $S_n$ be the partial sum $S_n := X_1 + \cdots + X_n$. Then $\mathrm{Var}(S_n/n^\alpha) = \sigma^2 n^{1-2\alpha}$. This tends to zero if $\alpha > \frac{1}{2}$ (including the case $\alpha = 1$), equal to 1 if $\alpha = \frac{1}{2}$, and to infinity otherwise. Therefore, $\alpha = \frac{1}{2}$ seems as the right scale to look for a "correction term" for the law of large numbers. Write $Y_n := \frac{S_n - n\mu}{\sqrt{n}} = \sqrt{n}(\bar{X}_n - \mu)$. Then the above argument shows that $\mathrm{Var}(Y_n) = \sigma^2$. By construction $E[Y_n] = 0$. We can therefore express $\bar{X}_n$ as

$$\bar{X}_n = \mu + \frac{1}{\sqrt{n}} Y_n,$$

now thinking of $Y_n$ as a RV with zero expectation and variance $\sigma^2$. The Central Limit Theorem describes the distribution of $Y_n$ as $n \to \infty$, asserting that for large $n$ it is nearly $N(0, \sigma^2)$, regardless of the underlying distribution. In this chapter we make this statement precise.

## 4.2 Weak Convergence

The first step towards the central limit theorem is a discussion on convergence of probability distributions. The standard notion of convergence that we will use in the statement and the proof of the Central Limit Theorem is called weak convergence. We will now describe it.

Recall that if $X$ is a RV its distribution is a Borel probability measure $P_X$ defined as the pushforward measure of $X$:

$$P_X(A) := P(X \in A).$$

Conversely, if $P$ is a Borel Probability measure on $\mathbb{R}$, and $X : \mathbb{R} \to \mathbb{R}$ is the identify function, then the distribution of $X$, $P_X$, is $P$.

The notion of the distribution of a RV allows us to standardize the discussion and instead of looking at probability measures on different and possible abstract spaces, restrict the discussion to Borel probability measures on the real line. As a concrete example, the indicator of Heads in a fair coin toss and the indicator that a uniform RV on $[0, 1]$ is $\leq \frac{1}{2}$ may be RVs on distinct probability spaces, but their distributions are the same.

In light of this, the notion of convergence we will define and study will be for Borel probability measures. Let's look at a simple example.

**Example 4.2.1.** *Let $P_n$ be the Borel probability measure which is uniform on $T_n := \{k/n : k \in \{0, \ldots, n-1\}\}$, equivalently, $P_n = \frac{1}{n}\sum_{k=0}^{n-1} \delta_{k/n}$. These measures have finite support. Letting $T = \cup_{n=1}^{\infty} T_n$. Then $P_n(T) = 1$. If $P$ is the uniform measure on $[0, 1]$, $P(T) = 0$ because $T$ is countable.*

*In particular, we do not have $P_n(A) \to P(A)$ for all Borel sets $A$. However, if we lift our gaze from the wild collection of Borel sets to integrals of "nice" functions, a connection between $P_n$ and $P$ will appear. Indeed, if we take any bounded and continuous function $f$, then $\int f dP_n = \frac{1}{n}\sum_{k=0}^{n-1} f(k/n) \to \int_0^1 f(t)dt = \int f dP$. Therefore, at least from the perspective of expectations of bounded continuous functions, the measures $(P_n : n \in \mathbb{N})$ do approach the measure $P$.*

This motivates the following definition:

**Definition 4.2.1.**

- *Let $(P_n : n \in \mathbb{N})$ and $P$ be Borel measures on $\mathbb{R}$. We say that the sequence converges to $P$ weakly if $\int f dP_n \to \int f dP$ for any continuous and bounded function $f$. We denote this convergence by $P_n \Rightarrow P$.*
- *We say that the distributions of the RVs $(X_n : n \in \mathbb{N})$ converge weakly to the distribution of the RV $X$ if $P_{X_n} \Rightarrow P_X$, or, equivalently*

$$\lim_{n\to\infty} E[f(X_n)] = E[f(X)]$$

*for every continuous and bounded $f$. Weak convergence of random variables is also known as convergence in distribution and convergence in law.*

We have the following Fatou-type characterization of weak convergence.

**Proposition 4.2.1.** *Let $(P_n : n \in \mathbb{N})$, $P$ be Borel probability measures on $\mathbb{R}$. Then $P_n \Rightarrow P$ if and only if $\liminf \int g dP_n \ge \int g dP$ for every nonnegative, bounded and continuous $g$.*

*Proof.* If $P_n \Rightarrow P$, the condition automatically holds. Assume that the condition holds then, and let $f$ be continuous and bounded. Then there exists some $M > 0$ such that $|f(x)| \le M$ for all $x \in \mathbb{R}$. Let $g := M - f$. Then $g$ is nonnegative, bounded and positive and therefore $\liminf \int g dP_n \ge \int g dP$. But the lefthand side is equal to $M - \limsup_{n\to\infty} \int f dP_n$ and the righthand side is $M - \int f dP$. Therefore, $\limsup \int f dP_n \le \int f dP$. To show the $\liminf \int f dP_n \ge \int f dP$ apply the same argument to the function $g := f + M$. □

Weak convergence follows from convergence in different modes we have mentioned before:

**Proposition 4.2.2.** *Suppose that $(X_n : n \in \mathbb{N})$ and $X$ are random variables.*

1. *If $X_n \to X$ in probability then the convergence also holds in distribution.*
2. *If $c$ is a constant and $X_n \Rightarrow c$, then $X_n \to X$ in probability.*

*Proof.*

1. Apply the dominated convergence theorem for sequences converging in probability.
2. Fix $\epsilon > 0$. Let $f_\epsilon$ be a function taking values in $[0,1]$ which is continuous, which is is equal to 1 outside the interval $(c-\epsilon, c+\epsilon)$ and equal to 0 at $c$. As $\{|X_n - c| \ge \epsilon\} \subseteq \{f_\epsilon(X_n) \ge 1\}$, and so $P(|X_n - c| \ge \epsilon) \le E[f_\epsilon(X_n)]$. But since $X_n \Rightarrow c$, it follows that $E[f_\epsilon(X_n)] \to \int f_\epsilon \delta_c = f_\epsilon(c) = 0$. □

We will now present several alternative characterizations of weak convergence.

**Theorem 4.2.1.** *Let $(P_n : n \in \mathbb{N})$ and $P$ be Borel probability measures on $\mathbb{R}$. Then the following are equivalent.*

1. *$P_n \Rightarrow P$.*
2. *$\limsup P_n(F) \le P(F)$ for all closed $F$.*
3. *$\liminf P_n(U) \ge P(U)$ for all open $U$.*
4. *$\lim_{n\to\infty} P_n(A) = P(A)$ for all $A$ with $P(\partial A) = 0$.*

Before proving the theorem, we present the following corollary which is used by some authors as the definition of weak convergence. Recall that the cumulative distribution function of a Borel probability measure $P$ is the function $F(x) := P(X \le x)$. We have the following, which some authors use as a definition for weak convergence.

**Corollary 4.2.1.** *Let $(P_n : n \in \mathbb{N})$ and $P$ be Borel probability measures, and let $(F_n : n \in \mathbb{N}), F$ be the respective distribution functions. Then $P_n \Rightarrow P$ if and only if $\lim_{n\to\infty} F_n(x) = F(x)$ at all continuity points of $F$.*

*Proof of Corollary 4.2.1.* For $x \in \mathbb{R}$, we let $A_x := (-\infty, x]$. If $P_n \Rightarrow P$ and $x$ is a continuity point of $F$, then $\partial A_x = \{x\}$, and since $P(\partial A_x) = F(x) - F(x-) = 0$, we have $F_n(x) = P_n(A_x) \to P(A_x) = F(x)$.

Conversely, let $U$ be any open set. Without loss of generality, $U = \cup_{i=1}^{\infty}(a_i, b_i)$ where the union is disjoint. Now $P_n((a_i, b_i)) = F_n(b_i-) - F_n(a_i)$. Since $F$ is nondecreasing, the set of points where it is discontinuous is countable. In particular, its complement in dense. Let $\epsilon > 0$. Then for each $i$ we can find continuity points of $F$ such that $(\bar a_i, \bar b_i] \subseteq (a_i, b_i)$ and $F(b_i-) - F(\bar b_i) + F(\bar a_i) - F(a_i) \le \epsilon/2^i$ (we use the fact that $F$ has left limit at $b_i$ and is right continuous at $a_i$). Thus,

$$P_n(U) \ge P_n(\cup_{i=1}^{\infty}(\bar a_i, \bar b_i]) = \sum_{i=1}^{\infty} P_n((\bar a_i, \bar b_i]) = \sum_{i=1}^{\infty}(F_n(\bar b_i) - F_n(\bar a_i)).$$

Apply Fatou's lemma to the righthand side to obtain

$$\liminf_{n\to\infty} P_n(U) \ge \sum_{i=1}^{\infty}(F(\bar b_i) - F(\bar a_i)) \ge \sum_{i=1}^{\infty}(F(b_i-) - F(a_i)) - \epsilon = P(U) - \epsilon.$$

The result now follows as $\epsilon > 0$ is arbitrary. □

*Proof of Theorem 4.2.1.*1 $\Rightarrow$ 2 If $F$ is empty, there's nothing to prove. Suppose then that $F$ is nonempty closed set. Let $d_F(x) := \inf\{|y - x| : y \in F\}$. Note that if $y \in F$, then $d_F(x) \le |y - x|$, and in particular, the infimum can be taken over $F \cap [x - y, x + y]$, a closed bound set. As a result, there exists a point $y_F \in F$, satisfying $d_F(x) = |y_F - x|$. Next we show that $d_F$ is continuous. For any $x, x'$ we have $d_F(x') \le |y_F - x'| \le d_F(x) + |x' - x|$. By switching the roles of $x$ and $x'$ we then have $|d_F(x) - d_F(x')| \le |x - x'|$. For $\epsilon > 0$, let $h_\epsilon(t)$ be the function which is equal to 1 on $[1, \infty)$, to zero on $(-\infty, 1 - \epsilon)$ and increases linearly on $(1 - \epsilon, 1)$. Clearly $h_\epsilon(1 - d_F \wedge 1)(x)$ is a continuous function which is equal to 1 on $F$, and is $< 1$ on $F^d$. Moreover, as $\epsilon \to 0$, $h_\epsilon \to \mathbf{1}_F$. Next,

$$\limsup_{n\to\infty} P_n(F) \le \lim_{n\to\infty} \int h_\epsilon dP_n = \int h_\epsilon dP,$$

where the equality follows from the assumption that $P_n \Rightarrow P$. Now take $\epsilon \to 0$ and use dominated convergence to conclude the proof.

2 $\Rightarrow$ 3 As the complement of every open set is a closed set, this implication follows immediately, and so does 3 $\Rightarrow$ 2.

3 $\Rightarrow$ 4 Let $A$ be any Borel set with $P(\partial A) = 0$. Recall that $A^\circ \subseteq A^\circ \cup \partial A$, where $A^\circ$ is the interior of $A$, which is open, and the righthand side is the closure of $A$ which is closed. Thus $P(A^\circ) \le P(A) \le P(A^\circ) + P(\partial A) = P(A^\circ)$, or $P(A) = P(A^\circ)$. Using this and part 3), we have

$$\liminf P_n(A) \ge \liminf P_n(A^\circ) \ge P(A^\circ) = P(A).$$

4 $\Rightarrow$ 1 To complete, suppose that $f$ is continuous, bounded and nonnegative function. For $t \ge 0$, let $A_t = \{x : f(x) \ge t\}$. We have $\partial A_t = \{x : f(x) = t\}$. Next, the function $t \to P(f \ge t)$ is nonincreasing and therefore has at most a countable number of discontinuities, all of which are jump discontinuities. The point $t$ is a discontinuity if and only if $P(A_t) > 0$. Therefore from 4) we have that $\lim_{n\to\infty} P_n(A_t) = P(A_t)$ for all but countably many $t$. In particular, the sequence of functions $t \to P_n(A_t)$ converges a.e. with respect to the Lebesgue measure on $\mathbb{R}$ to the function $t \to P(A_t)$. This and Fatou's lemma give

$$\liminf_{n\to\infty} \int f dP_n = \liminf_{n\to\infty} \int_{[0,\infty)} P_n(f \ge t) dm \ge \int_{[0,\infty)} P(f \ge t) dm = \int f dP,$$

which completes the proof due to Proposition 4.2.1.

□

## 4.3 Convergence to Poisson

We'll take a break from all the technical results characterizing weak convergence and present something different. We recall that a RV $X$ is Pois$(\lambda)$-distributed if $P(X=\ell)=e^{-\lambda}\frac{\lambda^\ell}{\ell!}$, $\ell\in\mathbb{Z}_+$. A simple calculation gives the following:

**Lemma 4.3.1.** *Let $X\sim Pois(\lambda)$ and $Y\sim Pois(\mu)$ be independent. Then $X+Y\sim Pois(\lambda+\mu)$*

**Exercise 4.3.1.** *Prove the lemma.*

**Proposition 4.3.1.** *Let $(X_k:k=1,\dots,n)$ be independent with $X_k\sim Bern(p_k)$ for some $p_k\in(0,1)$, and let $X:=\sum_{k=1}^n X_k$. Then there exists a probability space and on it random variables $\tilde X$ and $\tilde S$ such that:*

1. *$\tilde X$ has the same distribution as $X$.*
2. *$\tilde S\sim Pois(\sum_{k=1}^n p_k)$.*
3. *$P(\tilde X\neq\tilde S)\le\sum_{k=1}^n p_k^2$.*

This results is obtained by a technique we call coupling: we will construct random variables on the same space forcing them to be the same on events whose probability we will then estimate.

**Corollary 4.3.1** (Le Cam's Theorem)**.** *Let $(X_k:k=1,\dots,n)$ be independent with $X_k\sim Bern(p_k)$. Let $\lambda=\sum_{k=1}^n p_k$. Then*

$$\sum_{\ell\in\mathbb{Z}_+}|P(X=\ell)-P(Pois(\lambda)=\ell)|\le 2\sum_{k=1}^n p_k^2.$$

We first prove the Corollary:

*Proof.* Let $S\sim$ Pois$(\lambda)$ Then

$$\begin{aligned}|P(X=\ell)-P(S=\ell)|&=|P(\tilde X=\ell)-P(\tilde S=\ell)|=|E[\mathbf{1}_\ell(\tilde X)-\mathbf{1}_\ell(\tilde S)]|\\&\le E[\mathbf{1}_\ell(\tilde X)+\mathbf{1}_\ell(\tilde S),\tilde S\neq\tilde X]\\&=E[\mathbf{1}_\ell(\tilde X),\tilde S\neq\tilde X]+E[\mathbf{1}_\ell(\tilde S),\tilde S\neq\tilde X].\end{aligned}$$

Now sum over $\ell\in\mathbb{Z}_+$ to obtain the bound. □

Now let's prove the proposition.

*Proof of Proposition 4.3.1.* Let $(U_k:k=1,\dots,n)$ be IID U$[0,1]$. Define $\tilde X_k:=\mathbf{1}_{[1-p_k,1]}(U_k)$. Therefore $(\tilde X_k:k=1,\dots,n)$ have the same distribution as $(X_k:k=1,\dots,n)$. Let $\tilde X=\sum_{k=1}^n\tilde X_k$.

Next, we define a RV $\tilde Y_k\sim$ Pois$(p_k)$ as follows. Let $a_0=0$ and continue inductively letting $a_{j+1}=a_j+P(\text{Pois}(p_k)=j)$. Thus $(a_j:j\in\mathbb{Z}_+)$ is a strictly increasing sequence with $\lim_{j\to\infty}a_j=1$, and $P(U\in[a_j,a_{j+1}))=a_{j+1}-a_j=P(\text{Pois}(p_k)=j)$. Let $\tilde Y_k:=\sum_{j=0}^\infty j\mathbf{1}_{[a_j,a_{j+1})}(U_k)$. Then by construction, $\tilde Y_k\sim$ Pois$(p_k)$. Let $\tilde S=\sum_{k=1}^n\tilde Y_k$. Then $\tilde S\sim$ Pois$(\sum_{k=1}^n p_k)$.

We were able to define both $\tilde X_k$ and $\tilde Y_k$ as functions of the same RV $U_k$, and we will now examine the event on which they differ.

1. $\tilde X_k=0$ if and only if $U_k\in[0,1-p_k]$ and is 1 otherwise. Now $\tilde U_k=0$ if and only if $U_k\in[0,e^{-p_k})$. Since $e^{-p_k}>1-p_k$, $P(\tilde X_k=\tilde Y_k=0)=P(\tilde X_k=0)=P(U_k\in[0,1-p_k])=(1-p_k)$.
2. Now $\{\tilde Y_k=1\}=\{U_k\in[e^{-p_k},e^{-p_k}(1+p_k))\}\subset\{U_k\in[1-p_k,1]\}=\{\tilde X_k=1\}$. Therefore $P(\tilde X_k=\tilde Y_k=1)=P(\tilde Y_k=1)=e^{-p_k}p_k$.

Putting these together, $P(\tilde X_k=\tilde Y_k)=1-p_k(1-e^{-p_k})$. Therefore $P(\tilde X_k\neq\tilde Y_k)=p_k(1-e^{-p_k})\le p_k^2$. Finally,

$$P(\tilde X\neq\tilde S)\le P(\cup_{k=1}^n\{\tilde X_k\neq\tilde Y_k\})\le\sum_{k=1}^n P(\tilde X_k\neq\tilde Y_k)\le\sum_{k=1}^n p_k^2.$$

□

## 4.4 Tightness and Weak Convergence

Next we show that weak convergence is not that hard to attain. We need a definition.

**Definition 4.4.1.** *A sequence of Borel probability measures $(P_n : n \in \mathbb{N})$ is tight if for every $\epsilon > 0$, there exists $M > 0$ such that $\inf_n P_n((-M, M)) \geq 1 - \epsilon$.*

Tightness is simply the mathematical statement that no mass (probability) escapes to ±∞.

**Example 4.4.1.** *Let $(x_n : n \in \mathbb{N})$ be a sequence of real numbers. The sequence $(\delta_{x_n} : n \in \mathbb{N})$ is tight if and only if the sequence is bounded. Indeed, $\delta_{x_n}((-M, M)) = \mathbf{1}_{(-M,M)}(x_n)$. Therefore for any $\epsilon \in (0, 1)$, the infimum is $\geq 1 - \epsilon$ if and only if $|x_n| < M$ for all $n$, in which case the infimum is 1.*

**Exercise 4.4.1.** *Show that $(P_n : n \in \mathbb{N})$ are tight if and only if $\lim_{M\to\infty} \liminf_{n\to\infty} P_n((-M, M)) = 1$.*

**Exercise 4.4.2.** *Let $(X_n : n \in \mathbb{N})$ be RVs, and let $\varphi : \mathbb{R} \to [0, \infty]$ be Borel measurable satisfying $\lim_{|x|\to\infty} \varphi(x) = \infty$. Show that if $\limsup_{n\to\infty} E[\varphi(X_n)] < \infty$, then the distributions of $(X_n : n \in \mathbb{N})$ are tight.*

Tightness is a necessary condition for weak convergence.

**Lemma 4.4.1.** *$P_n \Rightarrow P$ implies that $(P_n : n \in \mathbb{N})$ are tight.*

*Proof.* Fix $\epsilon > 0$. Let $M_1$ be such that $P((-M_1, M_1)) \geq 1 - \epsilon/2$, and let $N = N(\epsilon)$ be such that $P_n((-M_1, M_1)) \geq P((-M_1, M_1)) - \epsilon/2$ for $n \geq N$. Now choose $M_2 \geq M_1$ such that $P_n((-M_2, M_2)) \geq 1 - \epsilon$ for all $n < N$. Thus, $P_n(-M_2, M_2)) \geq 1 - \epsilon$ for all $n$. □

The next result is a partial converse, and yet another justification for the notion of weak convergence.

**Theorem 4.4.1** (Helly-Bray Selection Theorem)**.** *Let $(P_n : n \in \mathbb{N})$ be Borel probability measures on $\mathbb{R}$ and $(F_n : n \in \mathbb{N})$ their respective distribution functions.*

1. *There exists a subsequence $(n_k : k \in \mathbb{N})$ and a right-continuous nondecreasing function $G : \mathbb{R} \to [0, 1]$ such that $\lim_{k\to\infty} F_{n_k}(x) = G(x)$ for every continuity point of $G$.*
2. *$(P_{n_k} : n \in \mathbb{N})$ is tight if and only if $\lim_{x\to-\infty} G(x) = 0$ and $\lim_{x\to\infty} G(x) = 1$.*

*Proof.*

1. By diagonalization, we can find a subsequence $(n_k : k \in \mathbb{N})$ such that $\lim_{k\to\infty} F_{n_k}(x)$ exists for every $x \in \mathbb{Q}$. Denote this limit by $\bar{G}$. Extend $\bar{G}$ to $\mathbb{R}$ by letting $G(x) = \inf\{\bar{G}(q) : q \in \mathbb{Q}, q > x\}$. Note that $G(q) \geq \bar{G}(q)$ for all $q \in \mathbb{Q}$. Clearly $G : \mathbb{R} \to [0, 1]$ and is nondecreasing. Next we show that it is right-continuous. Fix any $x$ and let $\epsilon > 0$. Then $G(x) + \epsilon$ is not a lower bound on $\{\bar{G}(q) : q \in \mathbb{Q}, q > x\}$ and therefore there exists $q > x$ such that $\bar{G}(q) \leq G(x) + \epsilon$. But then if $y \in [x, q)$, then $G(y) \leq \bar{G}(q) \leq G(x) + \epsilon$. Thus, for all $y \in [x, q)$, $G(x) \leq G(y) \leq G(x) + \epsilon$. This proves the right-continuity.

   Finally, we need to show convergence at all continuity points. Let $x \in \mathbb{R}$. By construction, there exists a rational $q > x$ such that $\bar{G}(q) \leq G(x) + \epsilon$, and therefore $\limsup_{k\to\infty} F_{n_k}(x) \leq \limsup_{k\to\infty} F_{n_k}(q) = \bar{G}(q) \leq G(x) + \epsilon$. This shows that $\limsup_{k\to\infty} F_{n_k}(x) \leq G(x)$. Now if $x$ is a continuity point, fix $\epsilon > 0$ and let $\delta$ be such that $G(x - \delta) \geq G(x) - \epsilon$. Pick any rational $q \in (x - \delta, x)$. Then
   $$\liminf_{k\to\infty} F_{n_k}(x) \geq \liminf_{k\to\infty} F_{n_k}(q) = \bar{G}(q).$$
   But since $q > x - \delta$, $G(x) - \epsilon \leq G(x - \delta) \leq \bar{G}(q)$, and so $\liminf_{k\to\infty} F_{n_k}(x) \geq G(x) - \epsilon$. Thus $\limsup_{k\to\infty} F_{n_k}(x) \geq G(x)$.

2. Suppose that the sequence is tight. Let $\epsilon > 0$. Pick continuity points $-M_1, M_2$ of $G$ such that $\inf_n P_n((-M_1, M_2)) \geq 1 - \epsilon$. Then $1 - \epsilon \leq \lim_{n\to\infty} F_{n_k}(M_2) - F_{n_k}(-M_1) = G(M_2) - G(-M_1)$. Take $M_1, M_2 \to \infty$ to obtain $\lim_{x\to\infty}(G(x) - G(-x)) \geq 1 - \epsilon$. Since $\epsilon > 0$ and $G$ takes values in $[0, 1]$, the result follows.

   Conversely, if $\lim_{x\to\infty} G(x) = 1$ and $\lim_{x\to-\infty} G(x) = 0$, then $G$ is the distribution function of a Borel probability measure, say $P$, and then the first part gives $P_{n_k} \Rightarrow P$. Lemma 4.4.1, completes the proof.

□

## 4.5 Characteristic Functions

Although we have multiple equivalent characterizations of weak convergence they are not very easy to work with, and do not offer a simple way to deal with sums of independent RVs. In this section we discuss yet another characterization, the all-mighty characteristic function, which also addresses these deficiencies. The characteristic function of a Borel probability measure is also known as the Fourier transform of the measure, so you may have seen similar objects elsewhere.

**Definition 4.5.1.** *Let $P$ be a Borel probability measure on $\mathbb{R}$. The characteristic function of $P$, $\varphi_P$ is defined as*

$$\varphi_P(t) = \int e^{itx} dP(x) = \int \cos(tx) dP + i \int \sin(tx) dP(x).$$

*If $X$ is a random variable with distribution $P$, then we will also write $\varphi_X$ for $\varphi_P$ and refer to it as the characteristic function of $X$. Then*

$$\varphi_X(t) = E[e^{itX}] = E[\cos(tX)] + iE[\sin(tX)].$$

As every Borel probability measure is the distribution of some RV. We will very often adopt the second approach, namely view the characteristic function as of the RV. Let's first observe some very basic properties of characteristic functions.

**Proposition 4.5.1.** *Let $\varphi$ be a characteristic function of some RV $X$. Then*

1. *$\varphi(0) = 1$ and $|\varphi(t)| \le 1$ for all $t \in \mathbb{R}$.*
2. *$\varphi$ is uniformly continuous.*
3. *If $X$ has a finite $n$-th moment. Then $\varphi$ is differentiable at $0$ up to the $n$-th order, with $\varphi^k(0) = i^k E[X^k]$, $k = 1, 2, dots, n$.*

We comment that the differentiability of the characteristic function at 0 does not automatically imply that $X$ has finite expectation. One such example is a RV satisfying $P(X = \pm n) = \frac{C}{n^2 \ln(n)}$ for $n = 2, 3, \dots$, where $C$ is a normalization constant. It is clear that $X$ is not integrable, yet, a straightforward calculation shows that characteristic function is differentiable at 0 wirth derivative equal to 0.

*Proof.* The first statement follows directly from the definition. The second from dominated convergence. Let's look at differentiation of $\varphi$ at 0.

$$\varphi(t) - \varphi(0) = E[\frac{e^{itX} - 1}{t}].$$

From the Fundamental Theorem of Calculus (written in complex form - view it as shorthand for the respective expression with cosines and sines):

$$e^{itX} - 1 = i \int_0^{tX} e^{is} ds.$$

Now $|e^{is}| = 1$ and therefore $|\frac{e^{itX}-1}{t}| \le |X|$. Since $\frac{d}{dt} e^{itX}|_{t=0} = iX$, it follows from dominated convergence that if $X$ is integrable, $\varphi'(0)$ exists and is equal to $iE[X]$. □

The behavior of the characteristic function near zero gives information about the tail of the RV. In most cases, this information is more useful for qualitative analysis than for quantitative analysis, as we will see below.

**Proposition 4.5.2.** *Let $\varphi$ be the characteristic function of $X$. Then for all $T > 0$,*

$$P(|X| > 2/\delta) \le \frac{1}{\delta} \int_{-\delta}^{\delta} 1 - \varphi(t) dt.$$

*Proof.* The function $x \to \frac{\sin x}{x}$ has a removable singularity at zero. Remove it to obtain a continuous function equal to 1 at 0. First, observe that $\frac{1}{2\delta}\int_{-\delta}^{\delta}(1-\cos(tX))dt = 1 - \frac{2\sin(\delta X)}{2\delta X}$. By taking Riemann sums and using dominated convergence,

$$\frac{1}{2\delta}E[\int_{-\delta}^{\delta} 1-\cos(tX)dt] = \frac{1}{2\delta}\int_{-\delta}^{\delta} 1-\varphi(t)dt.$$

Therefore

$$\frac{1}{2\delta}\int_{-\delta}^{\delta} 1-\varphi(t)dt = E[1-\frac{\sin(\delta X)}{\delta X}].$$

For all $|x|>2$, $|\frac{\sin x}{x}| \le \frac{\sin(2)}{2} \le \frac{1}{2}$. Therefore,

$$E[1-\frac{\sin(\delta X)}{\delta X}] \ge E[1-\frac{\sin(\delta X)}{\delta X}, \delta|X|\ge 2] \ge \frac{1}{2}P(\delta|X|\ge 2).$$

Thus,

$$P(|X|\ge \frac{2}{\delta}) \le \frac{1}{\delta}\int_{-\delta}^{\delta} 1-\varphi(t)dt.$$

□

**Corollary 4.5.1.** *Suppose $\varphi_X(t)=1$ for all $t$ in some open interval centered at 0. Then $X=0$ a.s.*

*Proof.* Suppose $\varphi$ is equal to 1 on on $(-\delta,\delta)$. Then $P(|X|>2/\delta)=0$. That is $X$ has bounded support and in particular has finite first and second moment. But then Proposition 4.5.1-3 gives that $0=\varphi'(0)=iE[X]$ and $0=\varphi''(0)=-E[X^2]$. Therefore $X$ has both expectation and variance equal to zero which implies $X\equiv 0$. □

The next step is to show that the characteristic function determines the distribution. This will be done through an inversion formula. What does "inversion" actually mean here? We need to recover a Borel probability measure from a characteristic function. Since any Borel probability measure is determined by the values it assigns to intervals, it is enough to find a formula that will output the measure of every interval.

Our goal is to express indicator of an interval as a linear function of complex exponentials. We will then take the expectation to obtain the probability of the interval. Easier said than done.

The starting point is the following calculus fact, which we won't prove here (nice exercise though):

$$\lim_{T\to\infty}\int_0^T \frac{\sin x}{x}dx = \frac{\pi}{2}.$$

This yields the following:

**Lemma 4.5.1.** *Let $\alpha<\beta$ be real numbers. Then*

$$\lim_{T\to\infty}\int_\alpha^\beta \frac{\sin(Tv)}{v}dv = \int_{\alpha T}^{\beta T}\frac{\sin u}{u}du = \begin{cases}\pi & \alpha<0<\beta\\ \frac{\pi}{2} & \alpha=0 \text{ or } \beta=0\\ 0 & \text{otherwise}\end{cases} \tag{4.5.1}$$

Now for some manipulation that will bring us to the promised land.

$$\frac{\sin(Tv)}{v} = \frac{e^{iTv}-e^{-iTv}}{2iv} = \frac{1}{2}\int_{-T}^{T} e^{itv}dt.$$

Therefore, the lefthand side of (4.5.1) is equal to

$$\begin{aligned}\lim_{T\to\infty}\frac{1}{2}\int_\alpha^\beta\int_{-T}^{T} e^{itv}dtdv &= \lim_{T\to\infty}\frac{1}{2}\int_{-T}^{T}\int_\alpha^\beta e^{itv}dvdt\\ &= \lim_{T\to\infty}\frac{1}{2}\int_{-T}^{T}\frac{e^{it\alpha}-e^{it\beta}}{it}dt\\ &= \lim_{T\to\infty}\frac{1}{2}\int_{-T}^{T}\frac{e^{-it\alpha}-e^{-it\beta}}{it}dt,\end{aligned}$$

where in the last step we changed variables $t \to -t$. Taking any reals $a < b$ and $x$, and letting $\alpha := a - x$ and $\beta := b - x$ we proved the following:

**Proposition 4.5.3.** *Fix real $a < b$ and $x$. Then*

$$\lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} e^{itx} dt = \mathbf{1}_{(a,b)}(x) + \frac{1}{2}\mathbf{1}_{\{a,b\}}(x)$$

All that's left is to use this to prove the inversion formula. That's the easiest step:

**Theorem 4.5.1.** *Let $X$ be a RV with characteristic function $\varphi$. Then*

$$P(X \in (a,b)) + \frac{1}{2} P(X \in \{a,b\}) = \lim_{T\to\infty} \frac{1}{2\pi} \int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} \varphi(t) dt.$$

*Proof.* From the proposition,

$$P(X \in (a,b)) + \frac{1}{2} P(X \in \{a,b\}) = \lim_{T\to\infty} \frac{1}{2\pi} E[\int_{-T}^{T} \frac{e^{-ita} - e^{-itb}}{it} e^{itX} dt].$$

We want to move the expectation inside the integral. To do that, observe that the integral is of the form $\int_{-T}^{T} f(t) e^{itX} dt$, where $f$ is a continuous function. This integral is the limit of Riemann sums. Specifically,

$$\frac{1}{n} \sum_{|kn| \le \lfloor T \rfloor} f(k/n) e^{ik/nX}.$$

These Riemann sums are bounded, and therefore it follows from the Dominated convergence theorem that their expectations converge to the expectation of their limit, the integral. However, as finite sums, and because of the linearity of the expectation, the expectation of each Riemann sum is equal to

$$\frac{1}{n} \sum_{|kn| \le \lfloor T \rfloor} f(k/n) \varphi(k/n).$$

Because $\varphi$ is continuous, this is a Riemann sum for the integral in the statement of the theorem. □

**Corollary 4.5.2.** *The characteristic function uniquely determines the distribution.*

*Proof.* We provide a simple algorithm utilizing the inversion formula and which gives the CDF.

1. $\frac{1}{2} P(X = x) = \lim_{n\to\infty} P(X \in (x - \frac{1}{n}, x)) + \frac{1}{2} P(\{x - \frac{1}{n}, x\})$. This is because the events $\{X \in (x - \frac{1}{n}, x)\}$ decrease to $\emptyset$ and $\sum_{n=1}^{\infty} P(X = x - \frac{1}{n})$ converges and therefore $\lim_{n\to\infty} P(X = x - \frac{1}{n}) = 0$.
2. Similarly, $P(X < x) + \frac{1}{2} P(X = x) = \lim_{n\to\infty} P(X \in (-n, x)) + \frac{1}{2} P(\{x - \frac{1}{n}, x\})$.
3. Now add the two.

□

So we know that a characteristic function determines a distribution. Next we show that weak convergence is determined by convergence of characteristic functions. This theorem is key to our proof of the Lindberg-Feller Central limit theorem in the next section.

**Theorem 4.5.2** (Levy's Continuity Theorem). *Let $(P_n : n \in \mathbb{N})$ be a sequence of Borel probability measures with respective characteristic functions $(\varphi_n : nin\mathbb{N})$. Then there exists a Borel probability measure $P$ such that $P_n \Rightarrow P$ if and only if there exists a function $\varphi(t)$ continuous at $0$ such that $\lim_{n\to\infty} \varphi_n(t) = \varphi(t)$ for all $t$.*

*Proof.* One direction is trivial. If $P_n \Rightarrow P$, then from the definition of weak convergence, for every $t \in \mathbb{R}$, $\varphi_n(t) \to \varphi(t)$, where $\varphi$ is the characteristic function of $P$, and as such, is continuous at 0.

The converse requires two steps. The first is to show that $(P_n : n \in \mathbb{N})$ (and therefore any subsequence) is tight. To do this we use Proposition 4.5.2. Fix $\epsilon > 0$ and pick $\delta > 0$, so that $\frac{1}{\delta} \int_{-\delta}^{\delta} (1 - \varphi(t) dt < \epsilon/2$. This is possible because $\varphi(0) = 1$ (why?) and $\varphi$ is continuous at 0. Next, pick

$N = N*(\epsilon,\delta)$ such that $frac1\delta\int_{-\delta}^{\delta} 1-\varphi_n(t)dt \le \frac{1}{\delta}\int_{-\delta}^{\delta} 1-\varphi(t)dt + \epsilon/2$ for all $n \ge N$. This is possible because of the pointwise convergence and the dominated convergence theorem as $|1-\varphi_n(t)| \le 2$ and we are integrating over a bounded interval. Putting both together, we have

$$\frac{1}{\delta}\int_{-\delta}^{\delta} 1-\varphi_n(t)dt \le \epsilon$$

for all $n \ge N$. By Proposition 4.5.2, $P_n((-2/\delta, 2/\delta)^c) \le \epsilon$ provided $n \ge N$. Decrease $\delta$ (and such decreasing the probabilities on the lefthand side) so that this also holds for $n \ge N$ and set $M := 2/\delta$. We proved that $P_n([-M,M]) \ge 1-\epsilon$ for all $n$, thus tightness holds.

It follows that every subsequence of $(P_n : n \in \mathbb{N})$ has a weakly convergent subsequence. But the assumed convergence of the seqeuence of characteristic functions to $\varphi$ implies that the limit along every convergent subsequence has $\varphi$ as its characteristic function, and as a result of Corollary 4.5.2, the limit is the same for all convergent subsequences. Call it $P$.

Finally, let $g$ be a bounded nonnegative and continuous function. Then $\liminf \int gdP_n$ is attained along some subsequence $(P_{n_k} : k \in \mathbb{N})$. Without loss of generality we may also assume that $P_{n_k} \Rightarrow P$. Therefore $\liminf \int gdP_n = \lim_{k\to\infty}\int gdP_{n_k} = \int gdP$, and the weak convergence is established due to Proposition 4.2.1. □

## 4.6 The Normal Distribution Revisited

Recall that a random variable $X$ is normal (Gaussian) with mean (expectation) $\mu$ and variance $\sigma^2$, denoted by $X \sim N(\mu,\sigma^2)$ if it has the density $f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-\mu)^2/(2\sigma^2)}$. The special case $\mu = 0$ and $\sigma^2 = 1$ is called the standard normal distribution.

**Exercise 4.6.1.** *Prove that if $X \sim N(\mu,\sigma^2)$, then the characteristic function of $X$, $\varphi_X$ is given by*

$$\varphi_X(t) = e^{i\mu t - \frac{\sigma^2 t^2}{2}}. \tag{4.6.1}$$

We will use characteristic functions to derive the following properties of normal distributions:

**Proposition 4.6.1.**
1. *Let $X \sim N(\mu,\sigma^2)$ and let $Z \sim N(0,1)$. Then the distribution of $X$ is equal to the distribution of $\sigma Z + \mu$. Conversely, the distribution of $Z$ is equal to the distribution of $(X-\mu)/\sigma$.*
2. *If $X_1 \sim N(\mu_1,\sigma_1^2)$ and $X_2 \sim N(\mu_2,\sigma_2^2)$, are independent, then $X_1 + X_2 \sim N(\mu_1+\mu_2, \sigma_1^2+\sigma_2^2)$.*

**Exercise 4.6.2.** *Prove the proposition using* (4.6.1).

There are a lot of interesting properties of normal distributions, particularly when considering their analogs in higher dimensions (normally distributed random vectors or fields).

## 4.7 Lindberg-Feller Central Limit Theorem

In this section we state and prove the central limit theorem. Before stating and proving the result, we two preliminary results that will be used in the proof. The main difficulty in the proof is establishing the existence of some limits of complex-valued objects.

**Lemma 4.7.1.** *Let $w_1,\dots,w_n$ and $z_1,\dots,z_n$ be complex numbers satisfying $|w_j|,|z_j| \le 1$ for all $j = 1,\dots,n$. Then $|\prod_{j=1}^n w_j - \prod_{j=1}^n z_j| \le \sum_{j=1}^n |w_j - z_j|$*

*Proof.* We prove the claim by induction on $n$. The base case $n = 1$ is trivial. We move to the induction step, assuming the inequality holds for $n$, we prove it for $n+1$:

$$\begin{aligned}\prod_{j=1}^{n+1} w_j - \prod_{j=1}^{n+1} z_j &= w_{n+1}\prod_{j=1}^{n} w_j - z_{n+1}\prod_{j=1}^{n} z_j\\ &= (w_{n+1}-z_{n+1})\prod_{j=1}^{n} w_j - z_{n+1}(\prod_{j=1}^{n} z_j - \prod_{j=1}^{n} w_j).\end{aligned}$$

Now take absolute value on both sides and apply the triangle inequality to separate the two summands on the righthand side. Since for any complex numbers $|\zeta_1\zeta_2| = |\zeta_1||\zeta_2|$, we obtain

$$\begin{aligned}|\prod_{j=1}^{n+1} w_j - \prod_{j=1}^{n+1} z_j| &\le |w_{n+1} - z_{n+1}||\prod_{j=1}^{n} w_j| + |z_{n+1}||\prod_{j=1}^{n} z_j - \prod_{j=1}^{n} w_j|\\ &\le |w_{n+1} - z_{n+1}| * 1 + |\prod_{j=1}^{n} z_j - \prod_{j=1}^{n} w_j|\\ &\le |w_{n+1} - z_{n+1}| + \sum_{j=1}^{n} |z_j - w_j|\end{aligned}$$

where the last line was obtained from the induction hypothesis. □

The next lemma gives us bounds on the error term in the Taylor expansion of $e^{ix}$.

**Lemma 4.7.2.** *Let $x \in \mathbb{R}$. Then*

$$|e^{ix} - \sum_{m=0}^{n} \frac{(ix)^m}{m!}| \le \frac{|x|^n}{n!} \min(\frac{|x|}{n+1}, 2).$$

*Proof.* The idea is iterative application of the fundamental theorem of calculus. This is one way to prove Taylor theorem. We will apply it specifically to the function $x \to e^{ix}$. Clearly, $e^{ix} = 1 + \int_0^x ie^{ix_1}dx_1$. Repeat to obtain

$$e^{ix} = 1 + \int_0^x i(1 + \int_0^{x_1} ie^{ix_2}dx_2)dx_1 = 1 + ix + \int_0^x \int_0^{x_1} i^2 e^{ix_2} dx_2 dx_1.$$

By induction,

$$e^{ix} = \sum_{m=0}^{n} \frac{(ix)^m}{m!} + \int_0^x \int_0^{x_1} \cdots \int_0^{x_n} i^{n+1} e^{ix_{n+1}} dx_{n+1} \cdots dx_1.$$

To complete the proof we need to bound the absolute value of the iterated integral on the right. We give two different bounds, which in turn give the desired result. Both utilizing the triangle inequality for integrals $|\int_0^x f(x)dx| \le |\int_0^x |f(x)|dx|$.

- The first bound is trivial: since $|i^{n+1}e^{ix_{n+1}}| = 1$, the absolute value of the iterated integral is bounded above by the absolute value of iterated integral of the constant function 1, which is equal to $|x|^{n+1}/(n+1)!$.
- The second is also trivial. we first integrate the innermost integral to obtain

  $$\int_0^x \int_0^{x_1} \cdots \int_0^{x_{n-1}} i^{n+1}(e^{ix_n} - 1)dx_n \cdots dx_1.$$

  Now, $|i^{n+1}(e^{ix_n} - 1)| = |\cos(x_n) + i\sin(x_n) - 1| \le 1 + 1 = 2$, and iterate this to obtain the bound $2|x|^n/n!$

□

We're finally ready to state and prove the theorem.

**Theorem 4.7.1.** *Suppose that for each $n \in \mathbb{N}$, $(X_{n,m} : m = 1, \dots, n)$ are independent Random variables satisfying*

1. $E[X_{n,m}] = 0$.
2. $\lim_{n\to\infty} \sum_{m=1}^n E[X_{n,m}^2] = \sigma^2 \in [0, \infty)$.
3. *For every $\epsilon > 0$, $\lim_{n\to\infty} \sum_{m=1}^n E[X_{n,m}^2, |X_{n,m}| > \epsilon] \to 0$.*

*Let $Z_n := \sum_{m=1}^n X_{n,m}$. Then $Z_{n,m} \Rightarrow N(0, \sigma^2)$.*

Before we prove the theorem we state the most useful corollary, the "standard" central limit theorem:

**Corollary 4.7.1.** *Let $(X_n : n \in \mathbb{N})$ be IID with expectation $\mu$ and variance $\sigma^2 \in [0,\infty)$. Let $S_n := \sum_{m=1}^n X_m$. Then*

$$\frac{S_n - n\mu}{\sqrt{n}} \Rightarrow N(0,\sigma^2).$$

I want to repeat the statement of the central limit theorem. If we denote the expression on the lefthand side by $Z_n$, then we observe that it is equal to $\sqrt{n}(\bar{X}_n - \mu)$, with $\bar{X}_n := \frac{\sum_{m=1}^n X_m}{n}$, the empirical mean. That is $\bar{X}_n = \mu + \frac{Z_n}{\sqrt{n}}$, or $\bar{X}_n$ is $\mu$ with a "correction" which is roughly $N(0,\sigma^2)/\sqrt{n}$ (or $N(0,\sigma^2/n)$).

*Proof of Corollary 4.7.1.* Let $X_{n,m} := \frac{X_m-\mu}{\sqrt{n}}$. Clearly the RVs $(X_{n,m} : m = 1,\dots,n)$ are independent. Also, $E[X_{n,m}] = 0$ and $E[X_{n,m}^2] = \frac{\sigma^2}{n}$. This gives the first two numbered conditions. As for the third, $E[X_{n,m}^2, |X_{n,m}| > \epsilon] = \frac{1}{n}E[(X_1-\mu)^2, |X_1-\mu| > \epsilon\sqrt{n}]$. Therefore $\sum_{m=1}^n E[X_{n,m}^2, |X_{n,m}| > \epsilon] = E[(X_1-\mu)^2, |X_1-\mu| > \epsilon\sqrt{n}] \to 0$ as $n \to \infty$ by dominated convergence. □

*Proof of Theorem* **??**. By Levy's continuity theorem it is enough to prove that the characteristic function of $Z_n$, $\varphi_n$, converges to that of $N(0,\sigma^2)$. We therefore need to show that for $t \in \mathbb{R}$,

$$|\varphi_n(t) - e^{-t\sigma_{n,m}^2/2}| \to 0,$$

where $\sigma_{m,n}^2 := E[X_{m,n}^2]$. By the triangle inequality, this will follow if we show

$$|\varphi_n(t) - \prod_{m=1}^n (1 - \sigma_{n,m}^2 t^2/2)| \to 0 \text{ and} \tag{4.7.1}$$

$$|\prod_{m=1}^n (1 - \sigma_{n,m}^2 t^2/2) - e^{-\sigma_{n,m}^2 t^2/2}| \to 0. \tag{4.7.2}$$

To prove (4.7.1) observe that its lefthand side is equal

$$\begin{aligned}|\prod_{m=1}^n E[e^{itX_{m,n}}] - \prod_{m=1}^n (1-\sigma_{n,m}^2 t^2/2)| &\le \sum_{m=1}^n |E[e^{itX_{m,n}}] - (1-\sigma_{n,m}^2 t^2/2)| \\ &\le \sum_{m=1}^n E[\frac{X_{n,m}^2}{2} \min(\frac{|X_{n,m}|}{3}, 2)].\end{aligned}$$

On the event $|X_{n,m}| \le \epsilon$. The minimum $\le \epsilon$. On the event $|X_{n,m}| > \epsilon$, the minimum is $\le 2$. Therefore

$$E[X_{n,m}^2 \min(\frac{|X_{n,m}|}{3}, 2)] \le \epsilon E[X_{n,m}^2] + E[X_{n,m}^2, |X_{n,m}| > \epsilon].$$

Taking the sum, we obtain the upper bound

$$\epsilon \sum_{m=1}^n E[X_{n,m}^2] + \sum_{m=1}^n E[X_{n,m}^2, |X_{n,m}| > \epsilon].$$

The second expression tends to zero by the third condition. The first tends to $\epsilon\sigma^2$. As $\epsilon$ is arbitrary, we established (4.7.1).

To prove (4.7.2), recall that for nonnegative $c$,

$$1 - c \le e^{-c} \le 1 - c + c^2/2.$$

This gives $|e^{-c} - 1 - c| \le c^2/2$. Taking $c = \sigma_{n,m}^2/2$ with this we obtain the following upper bounds on the lefthand side of (4.7.2)

$$\sum_{m=1}^n |(1 - \sigma_{n,m}^2 t^2/2) - e^{-\sigma_{n,m}^2 t^2/2}| \le \sum_{m=1}^n \sigma_{n,m}^4/4.$$

Now

$$\sigma_{n,m}^2 = E[X_{n,m}^2] \le \epsilon + E[X_{n,m}^2, |X_{n,m}| > \epsilon] \le \epsilon + \sum_{m=1}^n E[X_{n,m}^2, |X_{n,m}| > \epsilon].$$

For $n$ large enough the righthand side is bounded above by $2\epsilon$. Therefore

$$\limsup_{n\to\infty} \max_{m\le n} \sigma_{n,m}^2 \le 2\epsilon,$$

which in turn implies $\max_{m\le n} \sigma_{n,m}^2 \to 0$. Thus,

$$\sum_{m=1}^n \sigma_{n,m}^4/4 \le \max_{m\le n} \sigma_{n,m}^2 \sum_{m=1}^n \sigma_{n,m}^2 \to 0 \times \sigma^2.$$

We finished the proof. □

**Example 4.7.1.** *Let $(X_n : n \in \mathbb{N})$ be independent with $P(X_n = n) = P(X_n = -n) = \frac{1}{2}$. Then $E[X_m] = 0$ and $Var(X_m) = m^2$. As a result $Var(\sum_{m=1}^n X_m) = \sum_{m=1}^n m^2 \sim n^3/3$. Setting $X_{n,m} := X_m/n^{3/2}$, it follows that $E[X_{n,m}] = 0$ and $\lim_{n\to\infty} \sum_{m=1}^n E[X_{n,m}^2] = \frac{1}{3}$. In addition, $E[X_{n,m}^2, |X_{n,m}| > \epsilon] = \frac{1}{n^3} E[X_m^2, |X_m| > n^{3/2}]$. If $n \ge 2$, this is equal to zero for all $m = 1, \dots, n$. Therefore all conditions in the theorem hold and we have $\frac{\sum_{m=1}^n X_m}{n^{3/2}} \Rightarrow N(0, \frac{1}{3})$.*

### 4.7.1 Central Limit Theorem: The Swapping Method

With slight stronger assumptions we can get much more.

**Theorem 4.7.2.** *Let $(Y_m : m = 1, \dots, n)$ be independent, satisfying that for all $m = 1, \dots, n$*

1. *$E[Y_m] = 0$*
2. *$\sigma_m := E[Y_m^2] \in [0, \infty)$*
3. *$\rho_m := E[|Y_m|^3] < \infty$.*

*Let $\sigma^2 := \sum_{m=1}^n \sigma_m^2$, and let $S_n := \sum_{m=1}^n Y_m$. Then for every function $g : \mathbb{R} \to \mathbb{R}$ with bounded third derivative,*

$$\Big| E[g(S_n)] - E[g(N(0, \sigma^2))] \Big| \le \frac{\|g^{(3)}\|_\infty}{2} \sum_{m=1}^n \rho_m.$$

What does the theorem mean in the IID case? Let's suppose that $(X_n : n \in \mathbb{N})$ are IID with mean $\mu$, variance $\sigma^2 \in [0, \infty)$ and $\rho = E[|X_1|^3] \in [0, \infty)$. Now fix $n \in \mathbb{N}$ and set $Y_m := (X_m - \mu)/\sqrt{n}$. Then $E[Y_m] = 0$, $E[Y_m^2] = \sigma^2/n$ and $E[|X_m|^3] = \rho^3/n^{3/2}$. Therefore $\sigma^2$ in the theorem is $\sigma^2$ here (sorry about the notation!), and $\rho_m$ in the theorem is $\rho/n^{3/2}$. Rewriting the conclusion of the theorem:

**Corollary 4.7.2.** *Suppose that $(X_n : n \in \mathbb{N})$ are IID with mean $\mu$, variance $\sigma^2$ and $\rho := E[|X_1|^3] < \infty$. Let $g$ have bounded third derivative. Then for every $n \in \mathbb{N}$*

$$\Big| E[g(\frac{X_1 + \cdots + X_n - n\mu}{\sqrt{n}})] - E[g(N(0, \sigma^2))] \Big| \le \frac{\|g^{(3)}\|_\infty}{2} \frac{\rho}{\sqrt{n}}.$$

*Proof of Theorem 4.7.2.* The key is replacing each of the $Y_m$'s in the sum with a normal $N(0, \sigma + m^2)$ one by one and estimating the error thus created. Let $(Z_m : m = 1, \dots, n)$ be independent of the $Y_m$'s with $Z_m \sim N(0, \sigma_m^2)$, defined on the same probability space as $(Y_m : m = 1, \dots, n)$, and independent of the $Y_m's$ Define $S_{n,0} := \sum_{m=1}^n Y_m$, and set

$$S_{n,k+1} = S_{n,k} + (Z_{k+1} - Y_{k+1}).$$

In words, we replace $Y_k$ by $Z_k$ one by one: $S_{n,0} = \sum_{m=1}^n Y_m$ and $S_{n,n} = \sum_{m=1}^n Z_m \sim N(0, \sigma^2)$.

Clearly,

$$E[g(S_n)] - E[g(N(0, \sigma^2)] = E[g(S_{n,0})] - E[g(S_{n,n})].$$

This allows us to use a telescoping technique:

$$E[g(S_{n,0})] - E[g(S_{n,n})] = \sum_{k=0}^{n} E[g(S_{n,k}) - g(S_{n,k+1})].$$

The next step to estimate the differences inside this expectation. Let $\tilde{S}_{n,k} = S_{n,k} - Y_k$. Then

$$S_{n,k} = \tilde{S}_{n,k} + Y_k,\ S_{n,k+1} = \tilde{S}_{n,k} + Z_k.$$

Use Taylor expansion to obtain

$$g(S_{n,k}) = g(\tilde{S}_{n,k}) + g'(\tilde{S}_{n,k})Y_k + \frac{g''(\tilde{S}_{n,k})}{2}Y_k^2 + \frac{g^{(3)}(S'_{n,k})}{6}Y_k^3.$$

Similarly,

$$g(S_{n,k+1}) = g(\tilde{S}_{n,k}) + g'(\tilde{S}_{n,k})Z_k + \frac{g''(\tilde{S}_{n,k})}{2}Z_k^2 + \frac{g^{(3)}(S''_{n,k})}{6}Z_k^3.$$

Here $S'_{n,k}, S''_{n,k}$ are random points, obtained from Lagrange error term in Taylor's theorem. We therefore have

$$g(S_{n,k}) - g(S_{n,k+1}) = g'(\tilde{S}_{n,k})\frac{Y_k - Z_k}{+}g''(\tilde{S}_{n,k})(Y_k^2 - Z_k^2) + \frac{g^{(3)}(S'_{n,k})Y_k^3 - g^{(3)}(S''_{n,k})Z_k^3}{6}.$$

Taking expectations and using the fact that $\tilde{S}_{n,k}$ is independent of both $Y_k$ and $Z_k$, as well as the fact that $E[Y_k] = E[Z_k]$ and $E[Y_k^2] = E[Z_k^2]$, we have

$$|E[g(S_{n,k}) - g(S_{n,k+1})]| \le \frac{\|g^{(3)}\|_\infty}{6}E[|Y_k|^3 + |Z_k|^3].$$

A direct calculation shows that $E[|Z_k|^3] = 2\sigma_k^3$. Putting everything together we have

$$E[g(S_n)] - E[g(N(0,\sigma^2))] \le \frac{\|g^{(3)}\|_\infty}{6}(\sum_{k=1}^{n} E[|Y_k|^3] + \sum_{k=1}^{n} 2\sigma_k^3).$$

However, by Jensen's inequality, $\sigma_k^3 = E[Y_k^2]^{3/2} \le [|Y_k|^3]$, and therefore $2\sigma_k^3 \le 2E[|Y_k|^3]$, so the upper bound becomes $\frac{\|g^{(3)}\|_\infty}{2}\sum_{k=1}^{n}\rho_k$. □

The restriction of $g$ is the theorem and the corollary was necessary to obtain the error estimate. It does to pose any restriction when it comes to weak convergence. We record the following, which allows to conclude that under the assumptions of the corollary, weak convergence holds.

**Lemma 4.7.3.** *Let* $(Z_n : n \in \mathbb{N})$, $Z$ *be RVs satisfying* $\liminf_{n\to\infty} E[g(Z_n)] \to E[g(Z)]$ *for every nonnegative bounded* $g$ *with bounded third derivative. Then* $Z_n \Rightarrow Z$.

*Proof.* Let $U$ be any open set. Then $U$ is a countable union of disjoint open intervals, say $\cup_{i=1}^{\infty}(a_i,b_i)$. For every $N \in \mathbb{N}$, $P(Z_n \in U) \ge P(Z_n \in \cup_{i=1}^{N}(a_i,b_i))$. There exists a sequence of nonnegative continuous functions $(g_m : m \in \mathbb{N})$ each with bounded third derivative increasing to $\sum_{i=1}^{N}\mathbf{1}_{(a_i,b_i)}$. As a result

$$P(Z_n \in U) \ge P(Z_n \in \cup_{i=1}^{N}(a_i,b_i)) \ge E[g_m(Z_n)].$$

By by assumption $\liminf_{n\to\infty} E[g_m(Z_n)] \to E[g_m(Z)]$, and therefore, $\liminf P(Z_n \in U) \ge E[g_m(Z)]$. Now take $m \to \infty$ and apply dominated convergence to obtain $\liminf P(Z_n \in U) \ge E[P(Z \in \cup_{i=1}^{N}(a_i,b_i)]$. Now take $N \to \infty$, to complete the proof. □

# Chapter 5

# Martingales

Martingales are a simple mathematical model for a "fair" game of chance. They turn out to be very useful in numerous applications, and have a structure that allows to prove a rich collection of results beyond the IID setting.

## 5.1 Conditional Expectations

In this section we will assume that $(\Omega, \mathcal{F}, P)$ is a probability space. We will use the term sub-sigma algebra to describe a sigma algebra $\mathcal{G} \subset \mathcal{F}$. If $X$ is a random variable, we write $X \in \mathcal{G}$ (poor notation, yes) to mean that $X$ is measurable with respect to $\mathcal{G}$, namely

$$\{X \in B\} \in \mathcal{G}$$

for every Borel set $B$ (enough to check for $B$ of the form $(-\infty, x]$ with $x \in \mathbb{R}$ or even in $\mathbb{Q}$. Why?).

**Definition 5.1.1.** *Let $X$ be an integrable RV and let $\mathcal{G}$ be a sub-sigma algebra. The conditional expectation of $X$ with respect to $\mathcal{G}$ is a random variable $Y$ satisfying the following:*

1. *$Y \in \mathcal{G}$.*
2. *$E[Y, A] = E[X, A]$ for every event $A \in \mathcal{G}$.*

**Theorem 5.1.1.** *Let $X$ and $\mathcal{G}$ be as in Definition 5.1.1. Then the conditional expectation of $X$ with respect to $\mathcal{G}$ exists and is unique up to events with probability zero. We denote it by $E[X|\mathcal{G}]$.*

By taking $A = \Omega$, we conclude that $E[E[X|\mathcal{G}]] = E[X]$. What else? Can you show that $E[X|\mathcal{G}]$ is integrable?

The uniqueness statement simply means that if $Y$ and $Y'$ satisfy the conditions in the definition, then $Y = Y'$ a.s. So in fact the precise way to think of the conditional expectation is as of an equivalence class of RVs. We defer the existence proof in the theorem to the end of the chapter. The proof of the uniquness is very simple. Suppose $Y, Y'$ both satisfy the conditions, Let $A_+ = \{Y > Y'\} \in \mathcal{G}$. Then $E[Y, A_+] = E[Y', A_+]$ and therefore $E[Y - Y', A_+] = 0$. This implies $P(A_+) = 0$. Similarly $P(Y < Y') = 0$, and the uniqueness follows.

The uniqueness automatically gives us linearity. We record this as a result:

**Proposition 5.1.1.** *Let $X_1, X_2$ be integrable RVs and let $\mathcal{G}$ be a sub-sigma algebra. Then for every $c \in \mathbb{R}$,*

$$E[X_1 + cX_2|\mathcal{G}] = E[X_1|\mathcal{G}] + cE[X_2|\mathcal{G}_2].$$

Again, the precise statement here is the anything in the lefthand side is equal to anything on the righthand side.

*Proof.* Let $Y_j := E[X_j|\mathcal{G}]$, $j = 1, 2$. Clearly the RV $Y_1 + cY_2 \in \mathcal{G}$. Also, for every $A \in \mathcal{G}$, the linearity of the expectation gives $E[Y_1 + cY_2, A] = E[Y_1, A] + cE[Y_2, A]$. The righthand side is equal to $E[X_1, A] + cE[X_2, A] = E[X_1 + cX_2, A]$, again by linearity. Therefore, $Y_1 + cY_2$ satisfies both conditions in the definition of conditional expectation of $X_1 + cX_2$. □

**Example 5.1.1.** *Let $U \sim U[0,1]$, and let $p \in (0,1)$. Let $\mathcal{G}_p := \sigma(\{U \le p\})$. What is $E[U|\mathcal{G}_p]$? First $\mathcal{G}_p$ has four elements: $\{\emptyset, \{U \le p\}, \{U > p\}, \Omega\}$. Therefore, any $\mathcal{G}_p$-measurable RV $Y$ takes at most two distinct values and can be written as $Y = y_1 \mathbf{1}_{\{U \le p\}} + y_2 \mathbf{1}_{\{U > p\}}$. Since*

$$E[Y, U \le p] = y_1 P(U \le p) = y_1 p, \text{ and } E[Y, U > p] = y_2(1-p),$$

*and since*

$$E[U, U \le p] = \int_0^p u du = \frac{p^2}{2}, \ E[U, U > p] = \int_p^1 u du = \frac{1-p^2}{2},$$

*it follows that $Y = E[U|\mathcal{G}_p]$ if and only if $y_1 = E[U, U \le p] = E[U|U \le p]$ and $y_2 = E[U|U > p]$. The respective numerical values are $p/2$ and $(1-p)/2$, respectively but this is not so important. What is important? The value of $E[Y|\mathcal{G}_p]$ on each of the two disjoint events generating $\mathcal{G}_p$ is the expectation of $U$, conditioned to be on the event. Note, however, that the conditional expectation is a random variable.*

Let's record some basic properties of conditional expectation.

**Proposition 5.1.2.** *1. (positivity) If $X \ge 0$, $E[X|\mathcal{G}] \ge 0$.*

*2. (invariance) If $X \in \mathcal{G}$, $E[X|\mathcal{G}] = X$.*

*3. (independence) If $X$ is independent of $\mathcal{G}$, then $E[X|\mathcal{G}] = E[X]$.*

*4. Let $Z \in \mathcal{G}$. If $ZX$ is integrable, then $E[ZX|\mathcal{G}] = ZE[X|\mathcal{G}]$.*

*5. (tower) Let $\mathcal{H} \subseteq \mathcal{G}$ be a $\sigma$-algebra. Then $E[E[X|\mathcal{G}]|\mathcal{H}] = E[X|\mathcal{H}]$.*

Note that in particular if $c$ is a constant (viewed as a RV), $E[c|\mathcal{G}] = c$.

**Exercise 5.1.1.** *Prove the the first three items in the proposition.*

*Proof.* We only prove the last two statements. By linearity we may assume that both $Z$ and $X$ are nonnonegative, Let $(\varphi_n : n \in \mathbb{N})$ be simple and increasing to $Z$. Write $\varphi_n = \sum c_k \mathbf{1}_{A_k}$. Then

$$\begin{aligned} E[\varphi_n E[X|\mathcal{G}], A] &= \sum c_k E[X|\mathcal{G}, A \cap A_k] \\ &= \sum c_k E[X, A \cap A_k] = E[X \sum_k c_k \mathbf{1}_{A_k}, A] \\ &= E[X\varphi_n, A]. \end{aligned}$$

By monotone convergence, the righthand side converges to $E[XZ, A]$ and the lefthand side converges to $E[ZE[X|\mathcal{G}], A]$.

To prove the last property, Let $Y := E[X|\mathcal{G}]$. Then by definition, $E[Y, A] = E[X, A]$ for all $A \in \mathcal{G}$. As $\mathcal{H} \subseteq \mathcal{G}$, this holds for all $A \in \mathcal{H}$. For such $A$, the definition of $E[X|\mathcal{H}]$ gives, $E[X, A] = E[E[X|\mathcal{H}], A]$. Therefore reading the equalities from left to right we have

$$E[Y, A] = E[E[X|\mathcal{H}], A], \ A \in \mathcal{H}.$$

As $E[X|\mathcal{H}] \in \mathcal{H}$, we have proved that $E[Y|\mathcal{H}] = E[X|\mathcal{H}]]$. □

**Corollary 5.1.1.** *Let $X$ be integrable. Then $E[|X| \ |\mathcal{G}] \ge |E[X|\mathcal{G}]|$.*

*Proof.* The RV $|X| \pm X$ is nonnegative and therefore $E[|X| \pm X|\mathcal{G}] \ge 0$. Thus $E[|X| \ |\mathcal{G}] \ge \pm E[X|\mathcal{G}]$. □

Let's do another simple example.

**Example 5.1.2.** *Let $X_1, \dots, X_n$ be IID with finite expectaion and let $S_n := \sum_{k=1}^n X_k$. What is $E[X_1|S_n]$? For any statement on $S_n$, namely an event in $\sigma(S_n)$ we have that $E[X_1, A] = E[X_2, A] = \cdots = E[X_n, A]$. Since these add up to $E[S_n, A]$, and $E[S_n|\sigma(S_n)] = S_n$ (why?), it follows that $E[X_1, A] = \frac{1}{n} E[S_n, A]$. Therefore $E[X_1|S_n] = \frac{S_n}{n}$.*

*Next, what is $E[S_n|X_1]$ ? This is easier: $X_1 + (n-1)E[X_1]$. Why?*

If you follow closely, key properties that hold for the expectation as linearity, positivity and the triangle inequality carry over to conditional expectation. Let's get some more of these. This is just a sample.

**Proposition 5.1.3.** *Let $X$ be an integrable RV.*

1. *(Conditional Markov's Ineqaulity) Suppose that $X$ is nonnegative. Then for every $x > 0$, $E[\mathbf{1}_{\{X\ge x\}}|\mathcal{G}] \le \frac{E[X|\mathcal{G}]}{x}$.*
2. *(Conditional Jensen's inequality) Let $\varphi$ be convex. If $\varphi(X)$ is integrable, then $E[\varphi(X)|\mathcal{G}] \ge \varphi(E[X|\mathcal{G}])$*

*Proof.* As $x\mathbf{1}_{\{X\ge x\}} \le X$, the first statement follows from the positivity of the conditional expectation. As for the second, recall (see, Assignment #2 Problem #5), that for $\varphi(b) \ge \varphi(a) + (b-a)\varphi'(a+)$. So far so good. Now we can try to take conditional expectation of both sides of

$$\varphi(X) \ge \varphi(E[X|\mathcal{G}]) + (X - E[X|\mathcal{G}])\underbrace{\varphi'(E[X|\mathcal{G}]+)}_{=:Z},$$

alas, the righthand side has elements which may not be integrable and so we cannot use any of the previous results. So fix $M$ and let $X_M := X$ when $|X| \le M$ and $0$ otherwise. Now,

$$\varphi(X_M) \ge \varphi(E[X_M|\mathcal{G}]) + (X_M - E[X_M|\mathcal{G}])\underbrace{\varphi'(E[X_M|\mathcal{G}]+)}_{=:Z_M}.$$

As everything is bounded on both sides, we can take conditional expectations to and apply Proposition 5.1.2-1,4 to the first term and second term on the righthand side, respectively, to obtain

$$E[\varphi(X_M)|\mathcal{G}] \ge \varphi(E[X_M|\mathcal{G}]) + Z_M E[(X_M - E[X_M|\mathcal{G}])|\mathcal{G}] = \varphi(E[X_M|\mathcal{G}]) + Z_M * 0.$$

Note that we have used the first part of the proposition to obtain the zero in the last equality. Next,

$$|E[X|\mathcal{G}] - E[X_M|\mathcal{G}]| \le E[|X|\mathbf{1}_{\{|X|>M|\}}|\mathcal{G}].$$

Taking expectations on both sides and letting $M \to \infty$, we obtain $E[X_M|\mathcal{G}] \to E[X|\mathcal{G}]$ in probability as $M \to \infty$. Similarly $\lim_{M\to\infty} E[\varphi(X_M)|\mathcal{G}]$ converges in probability to $E[\varphi(X)|\mathcal{G}]$. Take a subsequence along which both converge a.s. and the result follows. □

We now discuss an important and concrete example of a conditional expectation in a case we can prove its existence trivially. We recall the "undergraduate" conditional expectation. If $G$ is any event with $P(G) > 0$, we define(d) $E[X|G]$ as $E[X,G]/P(G)$. In our language, this is simply the expectation of $X$ with respect to the conditional probability measure $P(\cdot\ |G)$ from Definition 1.2.3.

**Proposition 5.1.4.** *Suppose that $\Omega$ is the disjoint union of the events $(G_n : n \in \mathbb{N})$, and let $\mathcal{G} := \sigma((G_n : n \in \mathbb{N}))$ be the sub-sigma algebra generated by the sequence. Let $X$ be an integrable RV. Define*

$$y_n := \begin{cases} E[X|G_n] & P(G_n) > 0 \\ 0 & \textit{otherwise.} \end{cases}$$

*Then $E[X|\mathcal{G}] = \sum_{n=1}^{\infty} y_n \mathbf{1}_{G_n}$.*

*Proof.* Any event in $\mathcal{G}$ is a union of some (possibly none) elements in the sequence $(G_n : n \in \mathbb{N})$. Let $A$ be such an event. Then $A = \cup_{n\in I} G_n$ for some subset $I$ of $\mathbb{N}$. Let $I' = \{n \in I : P(G_n) > 0\}$. Then

$$\begin{aligned} E[X, A] &= \sum_{n\in I'} E[X, G_n] \\ &= \sum_{n\in I'} E[X|G_n]P(G_n) \\ &= \sum_{n\in I} E[y_n \mathbf{1} G_n] \\ &= E[\sum_{n=1}^{\infty} y_n \mathbf{1}_{G_n}, A]. \end{aligned}$$

This completes the proof. □

The next useful result is monotone convergence.

**Proposition 5.1.5** (Conditional Monotone Convergence)**.** *Let $(X_n : n \in \mathbb{N})$ be nonnegative integrable RVs increasing to $X$. Then $\lim_{n\to\infty} E[X_n|\mathcal{G}] = E[X|\mathcal{G}]$ a.s.*

*Proof.* Let $Y_n := E[X_n|\mathcal{G}]$. Then $(Y_n : n \in \mathbb{N})$ increases to some RV $Y$. Let $A \in \mathcal{G}$. By monotone convergence, $E[Y_n, A] \to E[Y, A]$. However, $E[Y_n, A] = E[X_n, A]$ and by monotone convergence the righthand side tends to $E[X, A]$. Therefore $E[Y, A] = E[X, A]$, and the proof is complete. □

## 5.2 Martingales

Throughout this section we fix a probability space $(\Omega, \mathcal{F}, P)$. In this context, a sub-sigma algebra $\mathcal{G}$ is any subset of $\mathcal{F}$ which is itself a sigma-algebra.

We need a bunch of definitions. At this stage it is good to recall that $\mathbb{N} = \{1, 2, 3, \dots\}$ is the set of (strictly) positive integers and $\mathbb{Z}_+ = \{0, 1, 2, \dots\}$ is the set of nonnegative integers.

**Definition 5.2.1.**

1. *A stochastic process (AKA process) $\mathbf{X}$ is a sequence of RVs indexed by $\mathbb{Z}_+$, $\mathbf{X} := (X_n : n \in \mathbb{Z}_+)$. We say that $\mathbf{X}$ is integrable if $E[|X_n|] < \infty$ for all $n \in \mathbb{Z}_+$.*

2. *A filtration $\mathfrak{F}$ is a nondecreasing sequence of sub-sigma algebras $\mathfrak{F} := (\mathcal{F}_n : n \in \mathbb{Z}_+)$. That is*

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \cdots \subseteq \mathcal{F}.$$

3. *The stochastic process $\mathbf{X}$ is adapted to the filtration $\mathfrak{F}$ if for all $n \in \mathbb{Z}_+$, $X_n \in \mathcal{F}_n$. That is $\{X_n \in B\} \in \mathcal{F}_n$ for all Borel sets $B$ and $n \in \mathbb{Z}_+$.*

When discussing stochastic processes it is very often a good idea to view them as describing evolution of some random system in time, where $X_n$ represents the state of the system at time $n \in \mathbb{Z}_+$. Let $\mathbf{X}$ be any process, and define $\mathcal{F}_n := \sigma(X_0, X_1, \dots, X_n)$. The resulting filtration $\mathfrak{F} := (\mathcal{F}_n : n \in \mathbb{Z}_+)$ is called the natural filtration for $\mathbf{X}$. We will often work with it.

**Definition 5.2.2.** *Let $\mathbf{M}$ be an integrable process adapted to the filtration $\mathfrak{F}$. We say that $\mathbf{M}$ is a martingale (resp. sub-martingale, super martingale) if $E[M_{n+1}|\mathcal{F}_n] = M_n$ (resp. $\geq$, $\leq$) for all $n \in \mathbb{Z}_+$.*

That is, a martingale has successive conditional expectations that are constant. When $\mathfrak{F}$ is the natural filtration, the martingale condition simply means that the expected value in the next time unit given the entire history is the current value of the process. One very simple example is when $M_n$ is the partial sum of IID RVs with expectation zero.

Before moving on, let's record some basic properties.

**Proposition 5.2.1.** *Let $\mathbf{M}$ be an integrable process, adapted with respect to $\mathfrak{F}$ and let $\varphi$ be a convex function, satisfying $\varphi(\mathbf{M}) := (\varphi(M_n) : n \in \mathbb{Z}_+)$ is integrable. Then $\varphi(\mathbf{M})$ is a sub-martingale if one of the following conditions holds:*

1. *$\mathbf{M}$ is a martingale.*

2. *$\mathbf{M}$ is a sub-martingale and $\varphi$ is nondecreasing.*

Note that the conditions listed are sufficient, not necessary.

**Exercise 5.2.1.** *Prove the proposition.*

**Example 5.2.1.** *Consider:*

1. *A martingale $\mathbf{M}$ with $E[M_n^2] < \infty$ for all $n \in \mathbb{Z}_+$. The $(M_n^2 : n \in \mathbb{Z}_+)$ is a sub-martingale.*

2. *a sub-martingale $\mathbf{M}$ and let $\varphi(x) := (x - a)_+ = \max(x - a, 0)$. This is a non-decreasing convex function and $|\varphi(x)| \leq |x| + |a|$, and therefore $((M_n - a)_+ : n \in \mathbb{Z}_+)$ is a sub-martingale.*

### 5.2.1 Some Concrete Martingales

#### Conditional Expectations

Let $X$ be an integrable random variable and let $\mathfrak{F}$ be a filtration. Then $M_n := E[X|\mathcal{F}_n]$ is a martingale. Indeed,

$$E[M_{n+1}|\mathcal{F}_n] = M_n$$

by the tower property. What can we say about $(M_n : n \in \mathbb{Z}_+)$? Does it converge? To what? We will answer this in Assignment #5.

#### Partial Sums

Let $(X_n : n \in \mathbb{N})$ be a sequence of RVs. Set $S_0$ to be a constant (hence measurable with respect to any sub-sigma algebra) and for $n \in \mathbb{N}$, let $S_n := S_0 + \sum_{k=1}^n X_k$. Consider the process $\mathbf{S} := (S_n : n \in \mathbb{Z}_+)$, equipped with the natural filtration.

- When is $\mathbf{S}$ a martingale? Clearly, $E[S_{n+1}|\mathcal{F}_n] = S_n + E[X_{n+1}|\mathcal{F}_n]$. Therefore $\mathbf{S}$ is a martingale if and only if $E[X_{n+1}|\mathcal{F}_n] = 0$ for all $n \in \mathbb{Z}_+$. For example, if $(X_n : n \in \mathbb{N})$ are all independent with mean zero.
- Similarly, $\mathbf{S}$, would be a sub-martingale if the $E[X_{n+1}|\mathcal{F}_n] \geq 0$ for all $n \in \mathbb{Z}_+$, for example if $(X_n : n \in \mathbb{N})$ are independent with nonnegative means (not necessarily all the same).

Let's go back to the general setup. There's a nice way to "center" $\mathbf{S}$ so we obtain a martingale by subtracting some process. Let's see how. Suppose that $\mathbf{D} = (D_n : n \in \mathbb{Z}_+)$ is a process such that $(S_n - D_n : n \in \mathbb{Z}_+)$ is a martingale. Then

$$E[S_{n+1} - D_{n+1}|\mathcal{F}_n] = S_n - D_n.$$

Equivalently,

$$E[S_{n+1} - S_n|\mathcal{F}_n] = E[D_{n+1}|\mathcal{F}_n] - D_n.$$

By construction, the lefthand side is equal to $E[X_{n+1}|\mathcal{F}_n]$, and to make the righthand side less ambiguous, let's assume for the moment that $D_{n+1} \in \mathcal{F}_n$. In this case, we necessarily have $D_{n+1} - D_n = E[X_{n+1}|\mathcal{F}_n]$, which results in $D_n = D_0 + \sum_{k=1}^n E[X_k|\mathcal{F}_{k-1}]$. If we let $D_0 := 0$, and define

$$M_n := S_n - D_n,\ n \in \mathbb{Z}_+,$$

then rewriting the equations above we conclude that $\mathbf{M} = (M_n : n \in \mathbb{Z}_+)$ is a martingale. We actually proved the following:

**Proposition 5.2.2.** *Let $(X_n : n \in \mathbb{N})$ be integrable RVs. Define the partial sums process $\mathbf{S}$ as follows. Set $S_0$ to be some constant, and define $S_n := S_0 + \sum_{k=1}^n X_k,\ n \in \mathbb{N}$. Let $\mathfrak{F}$ denote the natural filtration for $\mathbf{S}$. Then the process $\mathbf{D} = (D_n : n \in \mathbb{Z}_+)$, given by*

$$D_0 := 0,\ D_n := \sum_{k=1}^n E[X_k|\mathcal{F}_{k-1}],\ n \in \mathbb{N}.$$

*is the unique process satisfying all of the following conditions:*

1. $D_0 = 0$
2. $D_{n+1} \in \mathcal{F}_n$ *for all* $n \in \mathbb{Z}_+$.
3. $(S_n - D_n : n \in \mathbb{Z}_+)$ *is a martingale.*

As an example, let's suppose that $(X_n : n \in \mathbb{N})$ are IID with expectation $\mu$. Then $D_n = n\mu$ and so, $S_n - n\mu$ is a martingale. Ok, so what about $(S_n^2 : n \in \mathbb{N})$ in this case? Let's further assume $\mu = 0$, so that $\mathbf{S}$ is a martingale. We have

$$\begin{aligned} E[S_{n+1}^2|\mathcal{F}_n] &= E[(S_n + X_{n+1})^2|\mathcal{F}_n] \\ &= S_n^2 + 2S_n0 + E[X_{n+1}^2] \geq S_n^2. \end{aligned}$$

Therefore $(S_n^2 : n \in \mathbb{Z}_+)$ is a submartingale. Similar analysis reveals that the process $\mathbf{M} = (M_n : n \in \mathbb{Z}_+)$

$$M_n := S_n^2 - \sum_{1 \le k \le n} E[X_k^2],\ n \in \mathbb{Z}_+$$

is a martingale. In other words, the sub-matringale $(S_n^2 : n \in \mathbb{Z}_+)$ has a representation of the form

$$S_n^2 = M_n + \sum_{1 \le k \le n} E[X_k^2],$$

a martingale plus a non-decreasing process. As we will see later, any submartingale has a decomposition as a sum of a martingale and a non-decreasing process (both may be random!).

### "The" Martingale

Here's a game we're playing. Each round you bet some nonnegative amount $b$ on Heads. I toss a fair coin. If it lands Heads, I pay you $b$. If it lands Tails, you pay me $b$. We repeat. Is there a way for you to guarantee a win? Let's first analyze. If $b_n$ is the amount you bet in the $n$-th round, then your net gain after that round is $b_n$ or $-b_n$, each with probability $\frac{1}{2}$. Letting $(X_n : n \in \mathbb{N})$ be IID with $P(X_n = 1) = P(X_n = -1) = \frac{1}{2}$, then your net gain after $n$ rounds, $M_n$, is given by

$$M_0 = 0,\ M_n = \sum_{k=1}^n b_k X_k = \sum_{k=1}^n b_k (S_k - S_{k-1}), n \in \mathbb{N},$$

where $S_0 := 0$ and for $n \in \mathbb{N}$, $S_n := \sum_{k=1}^n X_k$. Let $\mathfrak{F}$ be the natural filtration for $\mathbf{S} = (S_n : n \in \mathbb{Z}_+)$. When is the process $\mathbf{M} := (M_n : n \in \mathbb{Z}_+)$ a martingale? Let's see.

$$E[M_{n+1}|\mathcal{F} = n] = M_n + E[b_{n+1} X_{n+1}|\mathcal{F}_n].$$

It we could "pull out" $b_{n+1}$, then the fact that $X_{n+1}$ is independent of $\mathcal{F}_n$, to conclude that $\mathbf{M}$ is a martingale. We can do that if $b_{n+1} \in \mathcal{F}_n$. Namely, the amount we bet depends only on the events prior to the coin toss following the bet. Very reasonable! We can only bet based on what we've seen, not what hasn't happened yet... One concrete example is $b_n = 1$ for all $n \in \mathbb{N}$. But there many other choices.

Thus, in what follows we will assume that $b_n \in \mathcal{F}_{n-1}$ for all $n \in \mathbb{N}$, in which case $\mathbf{M}$ is a martingale. In particular, $E[M_n] = 0$ for all $n \in \mathbb{N}$. Is it a fair game then?

Not quite. There is a way for you to pick $(b_n : n \in \mathbb{N})$ so that you guarantee a win no matter what. Pick $b_1 := 1$, and continue inductively,

$$b_{n+1} = \begin{cases} 2b_n & X_n = -1 \\ 0 & X_n = 1 \text{ or } M_n > 0. \end{cases}$$

In words: if you lose, you keep doubling the bet, until you win for the first time and then you don't bet anything, effectively exiting the game. This betting system is called the martingale (what martingales are named after).

Let's find your net gain with this betting system. There are two cases to consider:

1. $X_1 = 1$, so $M_1 = 1$, and so $b_2 = b_3 = \cdots = 0$, that is, you exit the game after one round with net gain 1.

2. $X_1 = -1$. Your keep losing until round $N$, the first satisfying $X_N = 1$. Your net gain at the conclusion of round $N$ is then $-1 - 2^1 - 2^{(N-2)}$ for rounds $1, \dots, N-1$ which you lost, plus $2^{N-1}$ for game $N$ which you won. Add these two together, to get a net gain of 1. After the $N$-th round you keep betting 0, effectively existing the game.

Bottom line: You can guarantee a net gain of 1. Note that there's some interesting discrepency. While $E[M_n] = 0$ for all $n$, $M_n = 1$ for all $n \ge N$. In particular $E[M_N] = 1$. So the game is not fair.

Not quite. To be able to implement this strategy, you need to bet a total of $1+2+\cdots+2^{N-1} = 2^N - 1$. This amount has to come from somewhere. The random variable $N$ is a Geometric RV with parameter $\frac{1}{2}$, and therefore the expected value of the amount you need to bet to guarantee a win is $E[2^N - 1] = \infty$. Why is this expectation relevant? If you want to replicate this game $n$ times, then the law of large numbers says that the amount you'll need to bet per game until a win will eventually exceed your fortune, so eventually... you'll go bankrupt playing the game.

### Polya's Urn

An urn initially contains $r$ red balls and $g$ green balls. Each unit of time we sample a ball from the urn uniformly, and return it along with an extra ball of the same color. Let $R_n$ denote the proportion of red balls in the urn after the $n$-th sample (meaning $R_0 := \frac{r}{r+g}$, as we haven't sampled yet). As usual, let's consider the natural filtration for the model. We will show that $(R_n : n \in \mathbb{Z}_+)$ is a martingale. Let $S_n$ denote the number of red balls in the urn after the $n$-th sample, then $S_n = r + \sum_{k=1}^n I_k$, where $I_k$ is the indicator of $k$-th sample was red. Now because $\mathcal{F}_k$ is a finite sigma-algebra, we can calculate the conditional expectation $E[I_k|\mathcal{F}_{k-1}]$ quite easily. Note: we're cheating and will calculate $E[I_k|S_{k-1}]$ instead. Let's find $E[I_k|S_{k-1} = \rho]$. Conditioning on $S_{k-1} = \rho$ ( note: number of balls in the urn at the same time is $\rho/(r + g + k - 1)$), the probability of $I_k = 1$ is $\rho/(r + g + (k - 1))$, and $I_k = 0$ with the remaining probability. Therefore, $E[I_k|S_{k-1} = \rho] = \rho/(r + g + k - 1)$. In other words, $S_n - \sum_{k=1}^n S_{k-1}/(r + g + (k - 1))$ is a martingale. In other words, $(r + g + n)R_n - \sum_{k=0}^{n-1} R_k$ is a martingale. Let's see what this means,

$$E[(r + g + n + 1)R_{n+1} - \sum_{k=0}^{n} R_k|\mathcal{F}_n] = (r + g + n)R_n - \sum_{k=0}^{n-1} R_k].$$

Or, after moving $\sum_{k=0}^n R_k$, which is $\mathcal{F}_n$-measurable to the righthand side, and taking care of the resulting cancellations:

$$E[(r + g + n + 1)R_{n+1}|\mathcal{F}_n] = (r + g + n)R_n + R_n,$$

Divide both sides by $(r + g + n)$ to obtain $E[R_{n+1}|\mathcal{F}_n] = R_n$. That is $(R_n : n \in \mathbb{Z}_+)$ is a martingale.

Note that $R_n$ is not a partial sum of RVs, despite the fact that we proved that it is a martingale by using techniques for partial sums.

**Exercise 5.2.2.** *Show directly that $(R_n : n \in \mathbb{Z}_+)$ is a martingale.*

### Branching Process

At time zero we have a population of, say, 1. We continue inductively. Each member of population at time $n$ gives birth to a random number of individuals with some fixed distribution, independent of anything else, and dies. The population at time $n + 1$ is the sum of all these individuals. More precisely,

$$Z_0 := 1,\ Z_{n+1} = \sum_{k=1}^{Z_n} B_{n+1,k},$$

where $(B_{m,k}, m, k \in \mathbb{N})$ are IID $\mathbb{Z}_+$-valued RVs. Note that if $Z_n = 0$, then $Z_{n+1} = 0$. That simple. Let's assume $E[B_{1,1}] = \mu \in (0, \infty)$. By conditioning on $Z_n$, we obtain

$$E[Z_{n+1}|\mathcal{F}_n] = \mu Z_n.$$

Therefore a quick calculation reveals that $(Z_n/\mu^n : n \in \mathbb{Z}_+)$ is a martingale. Clearly, $(Z_n : n \in \mathbb{Z}_+)$ is a martingale. Though the analysis is similar to what we would do for products of independent RVs, the structure is very different.

## 5.2.2 The Martingale Transform

We define a discrete version of an integral where the integration is with respect to a martingale, a sub-martingale or a super-martingale. One key property of this operation is that it allows to create new (sub- or super-) martingales from existing ones, and among other benefits, will allows us to control them. We already introduced a special version of it in the previous section when we described the gambling system.

**Definition 5.2.3.** *Let $\mathfrak{F} := (\mathcal{F}_n : n \in \mathbb{Z}_+)$ be a filtration, and $M := (M_n : n \in \mathbb{Z}_+)$ an adapted process, and $\mathbf{H} := (H_n : n \in \mathbb{N})$ a integrable process satisfying $H_n \in \mathcal{F}_{n-1}$ for all $n \in \mathbb{N}$. The integral transform of $H$ with respect to $M$ is the process $H \cdot M$ defined as follows:*

$$(H \cdot M)_0 := 0,\ (H \cdot M)_n := \sum_{k=1}^n H_k(M_k - M_{k-1}).$$

You can think of this as the analog of an integral of the form "$\int_0^n HdM$". An intuitive way to parse this integral transform is as follows. $M_n$ represents the value of a stock at the end of the $n$-the trading period, and $H_n$ represents the number of stocks purchased at the beginning of the $n$-th trading period, and therefore can only depend on information from the previous trading periods. If at the beginning of the $n$-the trading period we purchase $H_n$ units of stocks at the price $M_{n-1}$, then sell them at the end of the period, our net gain at the end of the period is $H_n(M_n - M_{n-1})$. Our net gain from the first $n$ periods is then $(H \cdot M)_n = \sum_{k=1}^n H_k(M_n - M_{n-1})$. The key observation is the following:

**Proposition 5.2.3.** *Let $\mathfrak{F}, \mathbf{M}$ and $\mathbf{H}$ be as in definition 5.2.3. Suppose that $\sup_n |H_n| < \infty$ a.s. Then*

1. *If $M$ is a martingale, $H \cdot M$ is a martingale.*
2. *If $H_n \geq 0$ for all $n \in \mathbb{N}$ and $M$ is a sub- or super-martingale then $H \cdot M$ is sub- or super-martingale, respectively.*

### 5.2.3 Stopping Times

Stopping times are $\mathbb{Z}_+ \cup \{\infty\}$-valued random variables which have an adaptability property. A stopping time should be viewed as a time representing the appearance of some event, like winning (or going bankrupt) in a game of chance. A key observation allowing to control martingale is the fact that a (sub- or super-) martingale frozen after a stopping time remains a martingale. This allows to control martingales and prove convergence and other results.

**Definition 5.2.4.** *Let $\mathfrak{F}$ be a filtration. A random variable $T$ taking values in $\mathbb{Z}_+ \cup \{\infty\}$ is called a stopping time if for all $n \in \mathbb{Z}_+$, $\{T \leq n\} \in \mathcal{F}_n$.*

**Example 5.2.2.** *Let $\mathbf{M}$ be a process equipped with its natural filtration.*

1. *$T := 10$ is a stopping time, because $\{T \leq n\}$ is either empty of $\Omega$, events which belong to any $\sigma$-algebra.*
2. *$T := \inf\{n \in \mathbb{Z}_+ : M_n \geq 0\}$, the first time $M_n$ is nonnegative, is also a stopping time. Indeed, $\{T \leq n\} = \{T > n\}^c = \cap_{k=0}^n \{M_k < 0\}$, a finite intersection of events in $\mathcal{F}_n$.*
3. *Let $T$ be as before and define $S := \inf\{n > T : M_n < 0\}$, the first time after $T$ the process is strictly negative, is also a stopping time. This requires a more elaborate proof. As before, we look at the complement of $\{S \leq n\}$.*

$$\{S > n\} = \{T > n\} \cup \{T \leq n, S > n\}.$$

*The first event is in $\mathcal{F}_n$. We turn to the second. We write it as a union.*

$$\{T \leq n, S > n\} = \cup_{k=0}^n \{T = k\} \cap (\cap_{l=k}^n \{M_l \geq 0\}).$$

*Again, this is a union of intersections of events in $\mathcal{F}_n$.*

4. *Let's suppose $M_0 = 0$ and let $T := \max\{n \leq 10 : M_n \leq 0\}$. That is, $T$ is the last time before time $n$ the process is nonnegative. The event $T = 1$ is then $\{M_1 \geq 0\} \cap (\cap_{k=2}^1 0 M_k < 0\}$. With the exception of some freak processes, this is not an event in $\mathcal{F}_1$. Think about it. By observing the process up to time 1 you cannot tell if that was the last time it was nonnegative. You simply have no sufficient information...*

Note the following simple observation.

**Proposition 5.2.4.** *Let $\mathfrak{F}$ be a filtration and let $T$ be a $\mathbb{Z}_+ \cup \{\infty\}$-valued random variable. Then $T$ is a stopping time if and only if $\{T = n\} \in \mathcal{F}_n$ for all $n \in \mathbb{Z}_+$.*

**Exercise 5.2.3.** *Prove the proposition.*

Here are some very easy and basic properties of hitting times.

**Proposition 5.2.5.** 1. *For any constant $c$, $T := c$ is a stopping time.*

2. *If $T$ and $S$ are stopping times, then $T+S$ is a stopping time.*

3. *If $(T_n : n\in\mathbb{N})$ are stopping times, then $\inf_n T_n$ and $\sup_n T_n$ are stopping times.*

**Exercise 5.2.4.** *Prove the Proposition.*

Suppose that $\mathbf{M}$ is adapted to the filtration $\mathfrak{F}$. For a borel subset $B$ of $\mathbb{R}$ define $\tau_B$, the hitting time of $B$ as follows:

$$\tau_B := \inf\{n\in\mathbb{Z}_+ : M_n\in B\}.$$

Then $\tau_B$ is a stopping time. Indeed, for every $n\in\mathbb{N}$, $\{\tau_B>n\} = \cap_{k=0}^n\{M_n\in B^c\}\in\mathcal{F}_n$.

### Stopped Processes

In this section we fix a filtration $\mathfrak{F}$, an adapted process $\mathbf{M}$ and an a.s. finite stopping time $T$

**Definition 5.2.5.** *The stopped process $\mathbf{M}^T := (M_n^T \in\mathbb{Z}_+)$ is the process defined by $M_n^T := M_{T\wedge n}$, $n\in\mathbb{Z}_+$.*

Thus $\mathbf{M}^T$ is basically the process $\mathbf{M}$, but frozen at time $T$. An example would be the following: $\mathbf{M}$ describes the net gain of a player in some game of chance and $T$ is the time the player quits the game (e.g. whenever my net gain is negative). So the net gain remains constant remains constant after the player quits the game.

**Proposition 5.2.6.** *If $\mathbf{M}$ is a (sub- or super-) martingale so is $\mathbf{M}^T$.*

The proof is a simple application of the integral transform.

*Proof.* Let $H_m := \mathbf{1}_{\{T\ge m\}}$. By Proposition 5.2.3 the process $M_0 + H\cdot M$ is then a (sub- or super-) martingale. However, this process is equal to

$$M_0 + \sum_{1\le k\le n} H_k(M_k - M_{k-1}) = M_0 + \sum_{1\le k\le T\wedge n}(M_k - M_{k-1}) = M_{T\wedge n}.$$

□

Let's put this into use.

**Example 5.2.3** (Simple Symmetric Random Walk)**.** *Let $X\in\mathbb{Z}$, $(X_n : n\in\mathbb{N})$ IID with $P(X_1 = 1) = P(X_1=-1) = \frac{1}{2}$. Define the simple symmetric random walk on $\mathbb{Z}$ starting from $x$ as the process $\mathbf{S} = (S_n : n\in\mathbb{Z}_+)$ with $S_n := x + \sum_{k\le n} X_k$. Then $\mathbf{S}$ is a martingale.*

*For $y\in\mathbb{Z}$, define*

$$T_y := \inf\{n\in\mathbb{Z}_+ : S_n = y\},$$

*the hitting time of $\{y\}$.*

*Now pick $a,b\in\mathbb{Z}$ with $a<x<b$, and define $T_{a,b} = T_a\wedge T_b$. We first show that $P(T_{a,b}<\infty) = 1$. Let $L>b-a$. Therefore for every $n$, on the event $A_{n,L} := \{X_n = X_{n+1}+\dots+X_{n+L} = 1\}$. Then on $A_{n,L}$, $S_{n+L}-S_n > (b-a)$. In particular, $\{T_{a,b}=\infty\}\cap\cup_{k=1}^\infty A_{kL,L} = \emptyset$. But $P(\cup_{k=1}^\infty A_{kL,L}) = 1 - P(\cap_{k=1}^\infty A_{kL,L}^c) = 1-(1-2^{-L})^\infty = 1$. Thereforfe, $P(T_{a,b}=\infty) = 0$.*

*With this let's consider the martingale $S^{T_{a,b}}$. As $T_{a,b}<\infty$ a.s. we have that $\lim_{n\to\infty} S_{T_{a,b}\wedge n} = S_{T_{a,b}}$ a.s. But since $a< S_{T_{a,b}}\le b$, it follows that $|S_{T_{a,b}}|\le |a|+|b|$, and it follows from dominated convergence that $\lim_{n\to\infty}E[S_{T_{a,b}\wedge n}] = E[S_{T_{a,b}}]$. Now, because of the martingale property, the lefthand side is equal to $E[S_0] = x$. In other words,*

$$x = E[S_{T_{a,b}}].$$

*But the righthand side is $aP(T_{a,b} = T_a) + bP(T_{a,b} = T_b)$. Let $\alpha := P(T_{a,b} = T_a) = P(T_a<T_b)$. Then have $x = \alpha a + (1-\alpha)b$. Doing the algebra, we obtain*

$$P(T_a<T_b) = \frac{b-x}{b-a}. \tag{5.2.1}$$

*Next, let's find $E[T_{a,b}]$. Clearly, $(S_n^2-n : n\in\mathbb{Z}_+($ is a martingale, Therefore so is $(S^2_{T_{a,b}\wedge n}-T_{a,b}\wedge n : n\in\mathbb{Z}_+)$. In particular,*

$$E[S^2_{T_{a,b}\wedge n}]-E[T_{a,b}\wedge n]=x^2,$$

*(the righthand side obtained by taking $n=0$). Rewriting, we have*

$$E[T_{a,b}\wedge n]=E[S^2_{T_{a,b}\wedge n}]-x^2.$$

*Now take $n\to\infty$, and apply monotone convergence to the lefthand side and dominated convergence to the righthande side ($S^2_{T_{a,b}\wedge n}\le a^2+b^2$), to conclude*

$$E[T_{a,b}]=E[S^2_{T_{a,b}}]-x^2.$$

*The expectation on the righthand side is equal to $a^2P(T_a<T_b)+b^2P(T_b<T_a)$, and therefore, using (5.2.1), we conclude that the righthand side is a quadratic polynomial of $x$ with leading coefficient equal to $-1$. Yet, if $x=a$ or $x=b$, the lefthand side is equal to $0$. Therefore we proved:*

$$E[T_{a,b}]=(x-a)(b-x) \tag{5.2.2}$$

*Long live martingales.*

## 5.3 Doob's Upcrossing and Convergence

The next topic we will discuss is a.s. convergence. This is all based on a clever observatrion and the integral transform. Let $\mathbf{M}$ be a process adapted to $\mathfrak{F}$. Let $a<b$ be real numbers. Define $T_0:=-1$ and define inductively a sequence of stopping times as follows:

$$T_{2k+1}:=\inf\{n>T_{2k}:M_n\le a\},\ k\in\mathbb{Z}_+T_{2k+2} \qquad :=\inf\{n>T_{2k+1}:M_n\ge b\},k\in\mathbb{Z}_+$$

That is, $T_1$ is the first time $\mathbf{M}$ is less than or equal to $a$, $T_2$ is the first time after $T_1$ that the process crosses the interval $[a,b]$, and so on. We define $U_n$, the number of upcorssings of $[a,b]$ up to time $n$ as

$$U_n:=\max(\sup\{k:T_{2k}\le n\},0). \tag{5.3.1}$$

Our first goal is get upper bounds on the expectation of $U_n$. This will be used control the oscillations of $\mathbf{M}$, and eventually resulting in a convergence result. We have the following:

**Theorem 5.3.1** (Doob's Upcrossing Inequality)**.** *Let $\mathbf{M}$ be a submartingale, $a<b$ real numbers. Let $U_n$ be as in (5.3.1). Then*

$$E[U_n]\le\frac{E[(M_n-a)_+]-E[(M_0-a)_+]}{b-a}.$$

In other words, the number of upcrossings can be controlled by the expected value of $(M_n-a)_+$. We will put that in use very soon, but let's first prove this beautiful result.

*Proof.* The first step is a reduction. The sequence of times $(T_l:l\in\mathbb{Z}_+)$ is not changed if we replace $\mathbf{M}$ and $a,b$ by $\mathbf{M}-a$ and $0,b-a$, respectively. The same applies if we replace the submartingale in the latter by the submartingale $(\mathbf{M}-a)_+$, Proposition 5.2.1. We will therefore proceed working with the latter, which wed denote by $\bar M$.

$$H_m:=\sum_{k=0}^\infty \mathbf{1}_{\{T_{2k+1}<m\le T_{2k+2}\}},\ m\in\mathbb{N}.$$

Each of the events in the indicators we sum over is an intersection of two events in $\mathcal{F}_{m-1}$. Therefore $H\cdot M$ is a sub-martingale. We have

$$(M_n-a)_+-(M_0-a)_+:=(H\cdot\bar M)_n+((1-H)\cdot\bar M)_n.$$

As both summands on the righthand side are sub-martingales equal to zero at $n=0$, taking expectation we have

$$E[(M_n-a)_+]-E[(M_0-a)_+]:=E[(H\cdot\bar M)_n]+E[((1-H)\cdot\bar M)_n]\ge E[(H\cdot\bar M)_n]+0.$$

Where the inequality was obtained by replaced the expectation at time $n$ with the expectation at tome 0. In addition, $(H\dot{}M)_n \geq (b-a)U_n$, because each completed upcrossing contributes at least $(b-a)$. For incomplete or between the completion of an upcorssing and before the beginning of a new one, the contribution is always nonngative, and so $E[(H \cdot \bar{M})_n] \geq (b-a)E[U_n]$, completing the proof. □

**Theorem 5.3.2** (Doob's Sub-Martingale Convergence Theorem)**.** *Let* $\mathbf{M}$ *be a sub-martingale satisfying* $\sup_{n\in\mathbb{Z}_+} E[(M_n)_+] < \infty$. *Then* $\lim_{n\to\infty} M_n$ *exists a.s. and is an integrable RV.*

*Proof.* $(M_n : n \in \mathbb{Z}_+)$ converges in the extended real line if and only if $\liminf M_n = \limsup M_n$. The negation of this statement is $\liminf M_n < \liminf M_n$. Equivalently, there exists some rationals $a < b$ such that $\liminf M_n < a$ and $\limsup M_n > b$. Implying that for some rational $a, b$, $\mathbf{M}$ upcrosses $[a, b]$ infinitely many times. We show that the latter event has probability zero. It is enough to show that if $a < b$ are rationals, the number of times the process upcrosses $[a, b]$ is finite a.s.

The proposition implies that $E[U_n] \leq E[(M_n - a)_+]/(b-a)$. However, $M_n - a \leq M_n + |a|$, and because the function $x \to x_+$ is nondecreasing and $x \leq x_+$ for all $x$, we have $(M_n - a)_+ \leq (M_n + |a|)_+ \leq (M_n)_+ + |a|$. Now let $U_\infty := \lim_{n\to\infty} U_n$, the number of upcrossings of the interval $[a, b]$. By monotone convergence $E[U_\infty] = \lim_{n\to\infty} E[U_n] \leq \sup \frac{E[(M_n)_+]+a}{b-a} < \infty$. Therefore $U_\infty < \infty$ a.s.

We therefore proved that $\lim_{n\to\infty} M_n$ exists a.s. in the extended sense. Denote the limit by $M_\infty$. By Fatou, $\liminf_{n\to\infty} E[(M_n)_+] \geq E[(M_\infty)_+]$. In addition, $E[M_n] = E[(M_n)_+] - E[(M_n)_-] \geq E[M_0]$. Therefore, $E[(M_n)_-] \leq E[(M_n)_+] - E[M_0]$, applying Fatou again, we conclude that $E[(M_\infty)_-] < \infty$. □

The first application is the following:

**Corollary 5.3.1.** *Let* $\mathbf{M}$ *be a nonnegative super-martingale. Then* $\lim_{n\to\infty} M_n$ *exists a.s.*

*Proof.* Let $-\mathbf{M}$ is a non-positive sub-martingale and hence its positive part is identically zero. The result follows from the Theorem. □

Now for some examples. The first two were introduced in Section 5.2.1

**Example 5.3.1** (Polya's Urn)**.** *We saw that* $R_n$*, the proportion of Red balls after* $n$ *samples, is a martingale. Since* $R_n \leq 1$*, it follows from the theorem that it converges a.s. One interesting observation is that unlike the law of large numbers, the limit is not a constant (unless initially all balls in the urn are of the same color).*

**Example 5.3.2** (Branching Processes)**.** *Recall that* $\mu$ *is the expected number of birth for an element of the population and that* $X_n$ *is the size of the population in generation* $n$*. We saw that* $X_n/\mu^n$ *is a martingale.*

*Suppose that* $\mu$ *is either* $< 1$ *(the critical regime) or* $= 1$ *(the subcritical regime). In the latter case we will exclude the case that a member of the population has exactly one birth with probability* $1$.

*Since* $X_n/\mu^n$ *is automatically a nonnegative super-martingale, it converges almost surely (this is true for all positive values of* $\mu$*). In the subcritical case, the denominator converges to zero, and therefore the ratio has a limit if and only if the numerator tends to* $0$*. But as* $X_n$ *is integer valued, it tends to zero if and only if it is eventually equal to zero. Therefore, the process hits* $0$ *a.s. (and stays there forever). In the critical case* $\lim_{n\to\infty} X_n$ *exists a.s. Since the process is integer valued, the convergence means that it remains constant eventually, or, there exits* $c \in \mathbb{Z}_+$ *and* $n \in \mathbb{Z}_+$ *such that* $X_{n+m} = c$ *for all* $m \in \mathbb{Z}_+$*. But for* $c \neq 0$*, and for any* $n \in \mathbb{N}$*, the event* $\cap_{m=0}^{\infty}\{X_n + m = c\}$ *implies that each of the* $c$ *elements in the* $n$*-th generation has exactly one birth, and then again for each of the* $c$ *elements in the* $n+1$*'th generation, etc. In other words it an infinite intersection of independent events, all with the same probability which is* $< 1$*. In other words, the event has probability zero. This leave* $c = 0$ *as the only candidate for the limit.*

## 5.4 Optional Stopping

Proposition 5.2.6 guarantees that a stopped (sub- or super-) martingale remains as such. Optional stopping goes one step beyond. Let's assume for the sake of the discussion that $\mathbf{M}$ is a martingale and so, the proposition implies that the stopped process $\mathbf{M}^T$ is also a martingale. What else can we ask for? A closure, intuitively, including the "last time" $n = \infty$ to our stopped martingale, and making the resulting process indexed by $\mathbb{Z}_+ \cup \{\infty\}$ a martingale. Specifically, a result of the form $E[M_T|\mathcal{F}_n] = M_{T\wedge n} =$. This condition clearly does not always hold. For example, if we consider the simple symmetric random walk of Example 5.2.3, starting from $x = 1$, then $P(T_1 < \infty) \geq P(T_0 < T_N) = \frac{N-1}{N} \to 1$. And so $T_0 < \infty$ a.s. As a result, $S_{T_0} = 0$, and $E[S_{T_0}|\mathcal{F}_n] = 0$ for all $n$. But $S_{T_0\wedge n}$ is a non-degenerate RV for all $n$. Specifically for $n = 1$ we have $S_{T_0\wedge n} = S_1$ which is equal to 0 and 2 with probability $\frac{1}{2}$ each. In fact, the fact that $S^{T_0}$ is a martingale implies $E[S_{T_0\wedge n}] = E[S_0] = 1$. Bottom line, we need some additional conditions...

Let $T$ be a stopping time. What are the events we can determine whether occurred or not by only observing the process up to time $T$? Mathematically, any event $A$ with the the property that $A \cap \{T = n\} \in \mathcal{F}_n$ for all $n$. That is, if $T = n$, then our information up to that time, namely $\mathcal{F}_n$, is sufficient to determine whether $A$ occurred or not. This leads to a definition.

**Definition 5.4.1.** *Let $T$ be a stopping time with respect to the filtration $mathfrakF$. The sigma-algebra of events determined by time $T$, $\mathcal{F}_T$, is defined as*

$$\mathcal{F}_T := \{A \in \mathcal{F} : A \cap \{T = n\} \in \mathcal{F}_n,\ \textit{for all } n \in \mathbb{N}\}.$$

**Exercise 5.4.1.** *Show that $\mathcal{F}_T$ defined above is a sigma-algebra.*

Note that when $T$ is deterministic, that is for some $n \in \mathbb{Z}_+$, $T = n$ a.s., $\mathcal{F}_T = \mathcal{F}_n$ (as mentioned a long time ago, all of our sigma-algebras are complete with respect to the underlying probability measure), so this notion really extends the filtration to indexing by stopping times. To state the optional stopping theorem, we need an additional assumption. Recall the definition of a uniformly integrable sequence of RVs, Definition 3.3.1. Though there we indexed the RVs by $\mathbb{N}$, the definition naturally extends to the index set $\mathbb{Z}_+$, which is the one we use here. A sub-martingale $\mathbf{M}$ which is also uniformly integrable automatically satisfies the conditions of the sub-martingale convergence theorem, Theorem 5.3.2 and therefore $M_\infty := \lim_{n\to\infty} M_n$ exists a.s. and $M_\infty$ is integrable. By Vitali's theorem, Theorem 3.3.1, the limit also holds in $L^1$, that is $\lim_{n\to\infty} E[|M_n - M_\infty|] = 0$. With this, we are ready to state the optional stopping theorem.

**Theorem 5.4.1.** *Let $\mathbf{M}$ be a uniformly integrable sub-martingale and let $S < \infty$ a.s. be a stopping time. Then $M_S$ is integrable and $E[M_\infty|\mathcal{F}_S] \geq M_S$ a.s.*

**Corollary 5.4.1.** *Suppose that $\mathbf{M}'$ is a sub-martingale and that $S \leq T < \infty$ a.s. are stopping times and ${M'}^T$ is uniformly integrable. Then $M'_S$ is integrable and $E[M'_T|\mathcal{F}_S] \geq M'_S$ a.s.*

Note that if $\mathbf{M}$ is a martingale, then it is a sub-martingale and $-\mathbf{M}$ is also a sub-martingale, and therefore, if we assume in Theorem 5.4.1 or Corollary 5.4.1 that $\mathbf{M}$ is a martingale, then the inequalities become equalities.

Word of warning: as we saw in the example opening this section, integrability of $M_T$ does not imply that $\mathbf{M}^T$ is uniformly integrable.

*Proof of Corollary 5.4.1.* Let $\mathbf{M} := {\mathbf{M}'}^T$. Then by assumption ${\mathbf{M}'}^T$ is UI and $M'_T = \lim_{n\to\infty} M_n$. Thus, the theorem gives $E[M_\infty|\mathcal{F}_S] \geq M_S$. The lefthand side is $E[M'_T|\mathcal{F}_S]$ and the righthand side is $M'_S$. □

*Proof of Theorem 5.4.1.* Let $A \in \mathcal{F}_S$ and let $n \in \mathbb{Z}_+$

$$E[M_n, A] = \sum_{k=0}^{n} E[M_n, A \cap \{S = k\}] + E[M_n, A \cap \{S > n\}].$$

For $k \leq n$,

$$E[M_n, A \cap \{S = k\}] \geq E[M_k, A \cap \{S = k\}] = E[M_S, A \cap \{S = k\}].$$

Therefore,

$$\sum_{k=0}^{n} E[M_n, A \cap \{S = k\}] = E[M_S, A \cap \{S \leq n\}].$$

Thus,

$$E[M_n, A] \geq E[M_S, A \cap \{S \leq n\}] + E[M_n, A \cap \{S > n\}]. \tag{5.4.1}$$

We will now squeeze (5.4.1) to obtain what we need to complete the proof.

1. Let $A := \{M_S > 0\}$, and then let $n \to \infty$ in (5.4.1). Then the uniform integrability (on lefthand side and second summand on the righthand side) and choice of $A$, and monotone convergence (right summand on righthand side) to conclude:

$$E[M_\infty, A] \ge E[(M_S)_+]$$

. Therefore $E[(M_S)_+] < \infty$.

2. Apply (5.4.1) to the sub-martingale $\mathbf{M}' := ((M_n)_+ : n \in \mathbb{Z}_+)$ and the stopping time $S' := S \wedge N$ for some $N \in \mathbb{Z}_+$. Thus, for $A \in \mathcal{F}_{S'}$ and $n = N$, (5.4.1) reads

$$E[(M_N)_+, A] \ge E[(M_{S'})_+, A].$$

In particular, letting $A := \Omega$, we have

$$E[(M_N)_+] \ge E[(M_{S\wedge N})_+].$$

We use these two facts to prove that $M_S$ is integrable. As the first gives $E[(M_S)_+] < \infty$, it remains to prove $E[(M_S)_-] < \infty$. Use the sub-martingale property of the stopped martingale $M^{S\wedge N}$ to conclude obtain $E[M_0] \le E[M_{S\wedge N}] = E[(M_{S\wedge N})_+] - E[(M_{S\wedge N})_-]$, and so

$$E[(M_{S\wedge N})_-] \le E[(M_{S\wedge N})_+] - E[M_0] \le E[(M_N)_+] - E[M_0],$$

where we have used the second fact to obtain the second inequality. Therefore, by Fatou,

$$E[(M_S)_-] \le \sup_N E[(M_N)_+] - E[M_0] < \infty.$$

Thus $M_S$ is integrable. Now return to (5.4.1) for the third time, and take the limit as $n \to \infty$. Vitali's convergence theorem, Theorem 3.3.1 gives that the lefthand side tends to $E[M_\infty, A]$. Also, UI and Proposition 3.3.3 gives that the second summand on the righthand side tends to zero. Finally, the integrability of $M_S$ and the dominated convergence theorem give that first summand on the righthand side tends to $E[M_S, A]$, completing the proof. □

We need some sufficient conditions for uniform integrability. Here is one result in this direction.

**Proposition 5.4.1.** *Let* $\mathbf{M}$ *be a sub-martingale and* $T < \infty$ *a.s. a stopping time. Let* $K > 0$. *Then each of the following implies that* $\mathbf{M}^T$ *is uniformly integrable.*

1. $T \le K$.
2. $E[T] < \infty$ *and* $|M_{n+1} - M_n| \le K$.

*Proof.* In the first case we have that $\mathbf{M}^T = (M_0, M_{T\wedge 1}, \dots, M_K, M_K, \dots)$. That is, $\mathbf{M}^T$ has at most $K+1$ distinct elements. Any finite collection of integrable RVs is uniformly integrable. In the second case, $|M_{T\wedge n}| \le TK$, an integrable RV and therefore $\mathbf{M}^T$ is uniformly integrable. □

It is worth also noting the following.

**Proposition 5.4.2.** *Let* $\mathbf{M}$ *be a UI sub-martingale and let* $T < \infty$ *a.s. be a stopping time. Then* $\mathbf{M}^T$ *is UI.*

*Proof.*

$$E[|M_{T\wedge n}|, |M_{T\wedge n}| > M] \le E[|M_{T\wedge n}| |\{T \le N\} \cap \{|M_{T\wedge n}| > M\}] + E[|M_n|, T > N].$$

Let $\epsilon > 0$. By Proposition 3.3.3, there exists $N' = N'(\epsilon)$ such that $\sup_n E[|M_n|, T > N] \le \epsilon/2$ provided $N > N'(\epsilon)$. As for the first term on the righthand side, it is bounded above by

$$\sum_{k=0}^{N} E[|M_{k\wedge n}|, |M_{k\wedge n}| > M].$$

As each of the RVs $M_{k\wedge n}$ is integrable, there exists some $M'=M'(N,\epsilon)$ such that this sum is bounded above by $\epsilon/2$, provided $M>M'$. Therefore, freezing $N>N'$, and $M>M'$ we have

$$\sup_n E[|M_{T\wedge n}|,||M_{T\wedge n}|>M]\le \epsilon/2+\epsilon/2.$$

In particular,

$$\lim_{M\to\infty}\sup_n E[|M_{T\wedge n}|,||M_{T\wedge n}|>M]\le \epsilon.$$

As $\epsilon$ is arbitrary, our work here is done. □

**Example 5.4.1.** *Suppose that $\mathbf{M}$ is a martingale, representing the net gain in some (fair) game of chance indexed by the number of rounds, so $M_0=0$. Let's also assume that the number of rounds cannot exceed some deterministic number $K$. The player is allowed to pick any exit strategy, that is, will quit the game at some stopping time $S\le N$. The optional stopping theorem then asserts that $E[M_K|\mathcal{F}_S]=M_S$. Therefore, $E[M_S]=E[M_K]=E[M_0]=0$. In other words, no strategy can improve the expected net gain. In other words, under this condition on the number of rounds (one can place more elaborate conditions, depending on the martingale), the game remains fair. One example where the uniform integrability is violated is of the "Martingale" of Section 5.2.1, where we found a strategy, namely a stopping time, which guarantees a win, despite the process being a martingale.*

## 5.5 Doob's Maximal Inequalities

The last section showed us that sub-martingales can be stopped. But here we say, not yet! They also exhibit some very powerful and useful maximal inequalities, inequalities which allow to control the running maxima of the process. We will present two types of maximal inequalities, one Doob's maximal inequality, is a Markov-type inequality, and the second, Doob's $L^p$ maximal inequality gives bounds on moments.

**Theorem 5.5.1.** *Let $\mathbf{M}$ be a submartingale and let $\lambda>0$. Then for every $n\in\mathbb{Z}_+$.*

$$\lambda P(\max_{0\le k\le n}M_k\ge\lambda)\le E[(M_n)_+,\max_{0\le k\le n}M_k\ge\lambda].$$

*Proof.* Let $A$ be the event in question, and let $S=\inf\{k\in\mathbb{N}:M_k\ge\lambda\}$. Then $A=\{S\le n\}$. First, $E[(M_S)_+,S\le n]\ge E[M_S,S\le n]\ge\lambda P(S\le n)=\lambda P(A)$. We now obtain an upper bound on the expression on the left. Use the sub-martingale property to obtain

$$E[(M_S)_+,S\le n]=\sum_{k=0}^n E[(M_k)_+,S=k]\le\sum_{k=0}^n E[(M_N)_+,S=k]=E[(M_N)_+,S\le n].$$

Putting it all together, we have $\lambda P(A)\le E[(M_n)_+,A]$, as desired. □

An easy corollary is Kolmogorov's Maximal inequality 3.5.1.

**Example 5.5.1.** *Set $S_0:=0$ and let $S_1,S_2,\dots$ be as in Theorem 3.5.1. Then the process $\mathbf{S}$ is a square integrable martingale, and therefore $\mathbf{S}^2:=(S_n^2:n\in\mathbb{Z}_+)$ is a nonnegative sub-martingale. Let $x>0$. Then Doob's maximal inequality with $\mathbf{M}:=\mathbf{S}^2$ and $\lambda:=x^2$ gives*

$$x^2P(\max_{0\le k\le n}S_k^2\ge x^2)\le E[S_n^2].$$

*But the event on the lefthand side is $\{\max_{k=0\dots,n}|S_k|\ge x\}$, and the result follows.*

We now turn to Doob's maximal $L^p$ inequality. Here is the statment

**Theorem 5.5.2.** *Let $\mathbf{M}$ be a submartingale and $p>1$ satisfying $E[(M_n)_+^p]<\infty$. Let $M_n^*:=\max_{0\le k\le n}(M_k)_+$. Then*

$$E[(M_n^*)^p]\le(\frac{p}{p-1})^pE[(M_n)_+^p].$$

**Example 5.5.2.** *Let* **S** *be as in Example 5.5.1. Then*

$$E[\max_{0\le k\le n} |S_k|^2] \le 4E[S_n^2].$$

*Proof.* The proof is an application of Doob's maximal inequality. We need take every bit of information from the righthand side in the inequality, in particular, the integrability of $(M_n)_+^p$. To simplify matters we observe that if we all Doob's inequality t the sub-martingale $\mathbf{M}_+$, it reads

$$P(M_n^* \ge \lambda) \le \lambda^{-1} E[(M_n)_+, M_n^* \ge \lambda]. \tag{5.5.1}$$

Next, we recall that for a nonnegative RV, $Z$, we have

$$E[Z^p] = \int_0^\infty P(Z^p \ge z)dz \overset{\lambda=z^{1/p}}{=} \int_0^\infty P(Z\ge \lambda)p\lambda^{p-1}dt.$$

Fix some $K>0$. Then taking $Z := M_n^* \wedge K$ and observeing that $P(Z>K)=0$, we have

$$E[(M_n^*\wedge K)^p] \le \int_0^K P(M_n^* \ge \lambda)d\lambda,$$

as for $\lambda \le K$, $M_n^* \wedge K \ge \lambda$ if and only if $M_n^* \ge \lambda$. We inserted $K$ because we do not know that $(M_n^*)^p$ is integrable, yet. Next, we plug (5.5.1) into this, and the rest is a general argument that has nothing to do with martingales, just convexity. Let's go.

$$E[(M_n^*\wedge K)^p] \le p\int_0^K \lambda^{p-2} E[(M_n)_+, M_n^* \ge \lambda]d\lambda.$$

Using Riemann sums, we can exchange the integral and expectation to rewrite the righthand side as

$$\begin{aligned} pE[(M_n)_+\int_0^K \mathbf{1}_{[\lambda,\infty)}(M_n^*)\lambda^{p-2}d\lambda] &= pE[(M_n)_+\int_0^{M_n^*\wedge K} \lambda^{p-2}d\lambda] \\ &= \frac{p}{p-1}E[(M_n)_+(M_n^*\wedge K)]. \end{aligned}$$

Therefore,

$$E[(M_n^*\wedge K)^p] \le \frac{p}{p-1}E[(M_n)_+(M_n^*\wedge K)^{p-1}].$$

To finish, we first apply the Holder inequality (Assignment #2, Problem 7, with $X := (M_n)_+$, $Y = (M_n^*\wedge K)$ and $\mu := P$) to obtain

$$E[(M_n)_+(M_n^*\wedge K)] \le E[(M_n)_+^p]^{1/p} E[(M_n^*\wedge K)^{(p-1)q}]^{\frac{1}{q}},$$

where $\frac{1}{q} = 1-\frac{1}{p} = \frac{p-1}{p}$. In particular, $q(p-1)=p$, and so

$$E[(M_n^*\wedge K)^p] \le \frac{p}{p-1}E[(M_n)_+^p]^{\frac{1}{p}} E[(M_n^*\wedge K)^p]^{1-\frac{1}{p}}.$$

If the lefthand side is zero for all $K$ there is really nothing left to prove, otherwise, pick $K$ large so that the lefthand side is strictly positive, to then obtain

$$E[(M_n^*\wedge K)^p]^{\frac{1}{p}} \le \frac{p}{1-p}E[(M_n)_+^p]^{\frac{1}{p}}.$$

Now raise both sides to the power $p$ and take $K\to\infty$ and monotone convergence to complete the proof. □

# Chapter 6

# Assignments

## 6.1 Assignment 1

1. Show that the Borel $\sigma$-algebra on $\mathbb{R}^2$, $\mathcal{B}(\mathbb{R}^2)$, is the product $\sigma$-algebra $\mathcal{B}(\mathbb{R}) \times \mathcal{B}(\mathbb{R})$.

2. Complete the details in the discussion below Proposition 1.2.5 showing that the product $\sigma$-algebra on $\{0,1\}^{\mathbb{Z}}$ is not the power set.

3. Let $(\mu_n : n \in \mathbb{N})$ be measures on a $\sigma$-algebra $\mathcal{F}$. Show that $\sum_{n=1}^{\infty} \mu_n$ is a measure. We'll return to this later. You'll appreciate the monotone convergence theorem...

4. Let $\mu$ be a Borel measure on $\mathbb{R}$. The support of $\mu$, $\mathrm{Supp}(\mu)$, is the set $\{x \in \mathbb{R} : \mu((x-\epsilon, x+\epsilon)) > 0 \text{ for all } \epsilon > 0\}$.

   (a) Show that $\mathrm{Supp}(\mu)$ is closed and that its complement is $\{x \in \mathbb{R} : \mu((x-\epsilon, x+\epsilon)) = 0 \text{ for some } \epsilon > 0\}$.

   (b) Find probability measures with the following supports: $\{0\}$, $\mathbb{Z}$ and $\{0\} \cup [1,2]$. All must be expressed in terms of objects we discussed in class or in this assignment.

5. Let $x \in \mathbb{R}$. Define $\delta_x : \mathcal{B}(\mathbb{R}) \to \{0,1\}$ by letting

$$\delta_x(A) := \begin{cases} 1 & x \in A \\ 0 & x \notin A \end{cases}$$

   (a) Show that $\delta_x$ is a probability measure. This measure is known as the Dirac delta measure.

   (b) Show that any Borel probability measure $P$ on $\mathbb{R}$ with finite support is a convex combination of delta measures. That is, there exist $n \in \mathbb{N}$, $c_1, \dots, c_n \geq 0$ with $\sum_{j=1}^{n} c_j = 1$ and $x_1, \dots, x_n \in \mathbb{R}$ so that $P = \sum_{j=1}^{n} c_j \delta_{x_j}$.

6. Recall that the completion of a measure space $(\Omega, \mathcal{F}, \mu)$ is a measure space $(\Omega, \bar{\mathcal{F}}, \bar{\mu})$, where $\bar{\mathcal{F}} := \{A \cup N' : A, N \in \mathcal{F}, \mu(N) = 0 \text{ and } N' \subseteq N\}$ and $\bar{\mu}(A \cup N) := \mu(A)$. Prove that $\bar{\mathcal{F}}$ is a $\sigma$-algebra and that $\bar{\mu}$ is well-defined and a measure on $\bar{\mathcal{F}}$.

7. Let $\mathcal{A}$ be an algebra and let $P$ be a probability measure on $\sigma(\mathcal{A})$. Prove the following "approximation" result: for every $B \in \sigma(\mathcal{A})$ and every $\epsilon > 0$ there exists $A \in \mathcal{A}$ such that $P(A \triangle B) < \epsilon$ (recall that $A \triangle B := (A - B) \cup (B - A)$).

8. Let $P$ be Borel probability measure on $\mathbb{R}$. Recall that the distribution function of $P$ is the function $F(x) = P((-\infty, x])$. Prove that the distribution function uniquely determines $P$ (that is, if $Q$ is a probability measure with distribution function $F$, then $Q = P$).

9. Let $P$ be a Borel probability measure on $\mathbb{R}$ and let $F$ be its distribution function.

   (a) Prove that

      i. $F$ is non-decreasing, right continuous and has left limits.

ii. $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$.

(b) For $< -\infty < a < b < \infty$, express the probability of each of the following intervals in terms of the distribution function: $\{a\}, (a, b], (a, b), [a, b), [a, b]$.

10. Suppose that $F : \mathbb{R} \to [0, 1]$ is a function satisfying the conditions listed in the previous question. For $y \in (0, 1)$, let $G(y) = \sup\{x : F(x) < y\}$. Show that the pushforward measure $m_G$ is a Borel probability measure on $\mathbb{R}$ with distribution function $F$. To do that, examine the intervals of the form $\{y : G(y) \leq x\}$.

## 6.2 Assignment 2

1. (a) Let $(a_{n,m} : n, m \in \mathbb{N})$ be an infinite array of nonnegative numbers. Use the MCT to prove

$$\sum_{n=1}^{\infty}\sum_{m=1}^{\infty} a_{n,m} = \sum_{m=1}^{\infty}\sum_{n=1}^{\infty} a_{n,m}.$$

Hint. Let $\mu$ be the counting measure on $\mathbb{N}$. For $n \in \mathbb{N}$, define the RV $f_N(m) = \sum_{n=1}^{N} a_{n,m}$. What is $\int f_N(m)d\mu$? What is $\lim_{N\to\infty}\int f_N(m)d\mu$? What is $\int \lim_{N\to\infty} f_N(m)d\mu$?

   (b) Use the result from the first part to redo Problem 3 from Assignment #1.

2. Prove Theorem 2.5.2.

3. Let $f(x) = \sum_{n=0}^{\infty} a_n x^n$ be a power series. Assume that the coefficients $(a_n : n \in \mathbb{N})$ are nonnegative and decreasing to 0 and that the radius of convergence of the power series is 1. Use the MCT to show that

$$\lim_{x\downarrow -1} f(x) = \sum_{n=0}^{\infty}(-1)^n a_n.$$

4. Let $f : [0,\infty) \to [0,\infty)$ be Riemann integrable, satisfying $\int_0^\infty f(t)dt = 1$, and suppose further that $\int_0^\infty xf(x)dx < \infty$. For $t \geq 0$, define

$$\phi(t) := \int_0^\infty e^{-tx} f(x)dx.$$

Prove, using the DCT, that $\phi'(0+)$ exists and is equal to $-\int xf(x)dx$. both terms on the righthand side are finite.

5. **Jensen's Inequality.** Let $U$ be an open interval in $\mathbb{R}$. Recall that a function $\varphi : U \to \mathbb{R}$ is called convex if for any $x, y \in U$ and $\lambda \in (0,1)$, $\phi(\lambda x + (1-\lambda)y) \leq \lambda\varphi(x) + (1-\lambda)\varphi(y)$. Among other properties, convex functions are continuous, and therefore Borel measurable.

Let $(\Omega, \mathcal{F}, P)$ be a probability space.

Jensen's inequality states the following. Let $\varphi$ be a convex function and let $X$ be an integrable RV. Then the expectation of $\varphi(X)$ is defined and satisfies

$$E[\varphi(X)] \geq \varphi(E[X]).$$

Prove this inequality and find necessary and sufficient conditions for an equality. Follow these steps.

   (a) Use the definition of convexity to show that for each $b, a \in \mathbb{R}$, $\varphi(b) \geq \varphi(a) + (b-a)\varphi'(a+)$, where $\varphi'(a+) := \lim_{c\downarrow a}\frac{\varphi(c)-\varphi(a)}{c-a} \in \mathbb{R}$.

   (b) Use the above inequality with $b = X$ and $a = E[X]$ to conclude that $E[\varphi(X)]$ is well defined and that the inequality holds.

6. **Applications of Jensen's ineqaulity**. You're expected to be familiar with characterization of convex functions. If you need help, let me know.

In all problems below use Jensen's inequality in an appropriate setting to prove the claim.

   (a) **Arithmetic Geometric Means Inequality (AGM).** Let $n \in \mathbb{N}$, $c_1, \dots, c_n > 0$ and let $\lambda_1, \dots, \lambda_n \in [0,1]$ with $\sum_{j=1}^{n}\lambda_j = 1$. Then

$$\prod_{j=1}^{n} c_j^{\lambda_j} \leq \sum_{j=1}^{n}\lambda_j c_j.$$

(This can be done directly from the definition - but also describe the relevant probability space).

In the next parts we assume that $(\Omega, \mathcal{F}, P)$ is a probability space and $X$ is a RV.

(b) If $P(X=0)<1$, $E[|X|]E[\frac{1}{|X|}]\geq 1$.

(c) For $p\in[1,\infty)$ define $\|X\|^p:=(E[|X|^p])^{1/p}$. Show that $p\to\|X\|^p$ is nondecreasing.

7. **Holder's Inequality.** Let $(\Omega,\mathcal{F},\mu)$ be a measure space. Let $p,q\in(1,\infty)$ satisfy $\frac{1}{p}+\frac{1}{q}=1$. Holder's inequality states that for RVs $X$ and $Y$:

$$|\int XY d\mu|\leq(\int|X|^p d\mu)^{\frac{1}{p}}(\int|Y|^q d\mu)^{\frac{1}{q}}.$$

whenever both terms on the righthand side are finite (this is not necessary, but we want to avoid $0*\infty$ on the righthand side, which is not really an issue because on this case the lefthand side is necessarily zero). Prove the inequality by following these steps.

(a) If one of the terms on the righthand side is zero, then the lefthand side is zero.

(b) Suppose both terms on the righthand side are finite and nonzero. Let $X'=X/(\int|X|^p d\mu)^{1/p}$ and $Y'=Y/(\int|Y|^q d\mu)^{1/q}$. Then the inequality is equivalent to $|\int X'Y'd\mu|\leq 1$. Now use the AGM to show that for $a,b\geq 0$, $ab\leq\frac{1}{p}a^p+\frac{1}{q}b^q$ to finish.

(c) State and prove a necessary and sufficient condition for an equality. It's hidden in your proof of the previous item.

The case $p=q=2$ is called the **Cauchy-Schwarz inequality**.

8. Consider the infinite product measure with $p=\frac{1}{2}$ and the coordinate mappings $(X_n:n\in\mathbb{N})$.

(a) Show that the RV $Y:=\sum_{n=1}^{\infty}2^{-n}X_n$ has U$[0,1]$ distribution by identifying its distribution function.

Hint. Given $n\in\mathbb{N}$ and $x_j\in\{0,1\}$, $j=1,\dots,n$ then there exists some $y\in[0,1]$ so that the event $\cap_{j=1}^n\{X_j=x_j\}$ is equal to the event $\{y\leq Y\leq y+2^{-n}\}$ almost surely. Use this to identify the CDF of $Y$.

(b) Show that you can modify the construction of $Y$ in part a) to obtain an infinite IID U$[0,1]$-distributed sequence (from the same sequence of coordinate mappings).

## 6.3 Assignment 3

1. Let $(X_n : n \in \mathbb{N})$ be independent satisfying $P(X_n = n) = P(X_n = -n) = \frac{1}{2n\ln(4n)}$, $P(X_n = 0) = 1 - \frac{1}{n\ln(4n)}$. Show that $\bar{X}_n \to 0$ in probability but not a.s.

2. Let $(A_n : n \in \mathbb{N})$ be a sequence of events satisfying $P(A_k \cap A_{k'}) = P(A_k)P(A_{k'})$ for $k < k'$ (AKA pairwise independent). Suppose that $\sum_{k=1}^{\infty} P(A_k) = \infty$.

   (a) Prove $\lim_{n\to\infty} \frac{\sum_{k=1}^n \mathbf{1}_{A_k}}{\sum_{k=1}^n P(A_k)} = 1$ a.s.

   Hint: denote the numerator by $S_n$. Let $n_k$ be the smallest index satisfying $E[S_{n_k}] \ge k^2$. Apply Chebychev and Borel-Cantelli along this sequence, and then complete for the full sequence.

   (b) Let $(X_n : n \in \mathbb{N})$ be IID continuous RVs (the CDF is continuous) and let $A_n := \{X_n = \max(X_1, \dots, X_n)\}$. That is, $X_n$ is a "record" value, or: a record is observed at time $n$. Prove that $P(A_n) = \frac{1}{n}$ and that $(A_n : n \in \mathbb{N})$ are pairwise independent. Let $R_n$ denote the time of the $n$-th record. Show that $\frac{\ln R_n}{n} \to 1$ a.s.

   Hint: If $X$ is a continuous RV and $Y$ is an independent RV then $P(X = Y) = 0$. You can use this without proof. This reduces all probability calculations into combinatorial calculations.

3. Complete the second part in the proof of the Strong Law of Large Numbers.

4. At time 0 your initial capital $X_0$ is 1. At time $n+1$, your capital is $X_{n+1} := R_{n+1}X_n$, where $(R_n : n \in \mathbb{N})$ are IID nonnegative RVs with finite expectation.

   (a) Show that $\frac{\ln X_n}{n}$ converges a.s. to some constant. Express it as an expectation of a function of $R_1$.

   (b) Give an example with $E[R_1] > 1$, yet $X_n \to 0$ a.s. When investing, average interest rate is not a correct measurement.

5. Let $(X_n : n \in \mathbb{N})$ be IID with zero expectation. Let $S_n := \sum_{k=1}^n X_k$ and let $M_n := \max_{k=1,\dots,n} |S_k|$ Prove $\lim_{n\to\infty} \frac{M_n}{n} = 0$ a.s. and $\lim_{n\to\infty} E[\frac{M_n}{n}] = 0$.

   Hint: Use the Strong Law of Large Numbers and Fatou's Lemma.

6. Complete the proof of the Glivenko-Cantelli Theorem by following these steps.

   (a) Let $F$ be any distribution function. Let $(U_n : n \in \mathbb{N})$ be IID $U[0,1]$. Let $G$ be the function from Assignment #2 Problem 10. Then the sequence $(X_n : n \in \mathbb{N})$ defined as $X_n := G(X_n)$ is an IID sequence with CDF $F$.

   (b) Let $F_U$ be the CDF of the uniform distribution on $[0,1]$, $\bar{F}_{U,n}$ the empirical distribution function for $U_1, \dots, U_n$, and $\bar{F}_n$ be the empirical distribution function for $X_1, \dots, X_n$. To conclude, express $\bar{F}_n - F$ in terms of $\bar{F}_{U,n} - F_U$.

7. The Strong Law of Large Numbers states that for an IID sequence $(X_n : n \in \mathbb{N})$, the integrability of $X_1$ ($X_1$ has finite expectation or, equivalently, $E[|X_1|] < \infty$) is sufficient for the a.s. convergence of the empirical means. We will now show that it is also necessary. Namely: if $E[|X_1|] = \infty$ then $P(\lim_{n\to\infty} \bar{X}_n$ exists and is in $\mathbb{R}) = 0$.

   To prove assume then that $E[|X_1|] = \infty$. Now follow these steps:

   (a) Show that $P(|X_n| \ge n \text{ i.o.}) = 1$.

   (b) Show that $|\bar{X}_{n+1} - \bar{X}_n| \ge \frac{|X_{n+1}|}{n+1} - \frac{|\bar{X}_n|}{n+1}$,

   (c) Now argue by contradiction assuming $\lim_{n\to\infty} \bar{X}_n$ exists and is in $\mathbb{R}$ on some event with positive probability, and use the two steps above to arrive to a contradiction.

8. **Some Information Theory.** Let $(X_n : n \in \mathbb{N})$ be IID with finite support $S$. You should think of the sequence of the RVs as data transmitted by some source over time (e.g. a rover on Mars transmitting status signals to earth, $S$ representing all possible signals: error codes, battery power levels, etc.). For a sequence $(s_1, \dots, s_n)$ of elements in $S$, define $p(s_1, \dots, s_n)$ as the probability of observing this particular sequence in the first $n$ units of time. That is $p(s_1, \dots, s_n) = P(X_1 = s_1, \dots, X_n = s_n)$. In particular, the random variable $p(X_1, \dots, X_n)$ gives the probability of the observed sequence.

   (a) Use the law of large numbers to show that

   $$-\frac{1}{n} \ln p(X_1, \dots, X_n)$$

   converges as $n \to \infty$ almost surely to a constant we denote by $H$, known as the entropy or information content of the distribution of $X_1$. Express the entropy $H$ as the expectation of some explicit function of $X_1$. In words, you showed that for $n$ large, with the exception of a set of "unlikely" sequences, the probability of a sequence of length $n$ is $e^{-n(H+o(1))}$. This property is called the equipartition property.

   (b) Use Jensen's inequality to show that $H \in [0, \ln |S|]$, and determine necessary and sufficient conditions for equality to 0 and to $|S|$.

   (c) For $\rho \in (0, 1]$ we say that the information content of $(X_n : n \in \mathbb{N})$ can be compressed at a ratio $\rho$ if for every $\epsilon > 0$ there exists $N = N(\epsilon, \rho)$ such that for each $n \geq N$, there exists a function $f_n : S^n \to S^{\lceil \rho n \rceil}$ and an event $A_n$ with $P(A_n) > 1 - \epsilon$ such that on $A_n$, $(X_1, \dots, X_n) \to f_n(X_1, \dots, X_n)$ is one-to-one. In other words, the bulk of the transmitted data can be expressed by using sequences which are shorter without loss of information.

   Use the equipartition property to prove **Shannon's compression theorem**: The information content of $(X_n : n \in \mathbb{N})$ can be compressed at any ratio larger than $H/\ln|S|$ and cannot be compressed at any ratio smaller than $H/\ln|S|$.

## 6.4 Assignment 5

Note: You must justify your all your steps in all problems below using techniques introduced in this course.

1. Let $\mathbf{X} := (X_n : n \in \mathbb{Z}_+)$ be a stochastic process. Suppose that $(A_k : k \in \mathbb{N})$ are Borel sets of $\mathbb{R}$. Define $T_1 := \inf\{n \in \mathbb{Z}_+ : X_n \in A_1\}$. Continue inductively, letting $T_{k+1} := \inf\{n > T_k : X_n \in A_{k+1}\}$. Show that $(T_k : k \in \mathbb{N})$ are all stopping times.

2. **Martingales with Bounded Increments**

   (a) Suppose that $\mathbf{M}$ is a martingale with the property that for some $C > 0$, $\sup_n |M_{n+1} - M_n| \le C$ a.s. Prove that that event $\{\lim_{n\to\infty} M_n \text{ exists in } \mathbb{R}\} \cup \{\liminf_{n\to\infty} M_n = -\infty, \limsup_{n\to\infty} M_n = \infty\}$ holds a.s.

   Hint: For $K \in \mathbb{N}$, let $\tau_{-K} = \inf\{n \in \mathbb{Z}_+ : M_n \le -K\}$. The martingale $(M_{\tau_{-K}\wedge n} : n \in \mathbb{Z}_+)$ is bounded below. Apply the convergence theorem to conclude that it converges a.s. Thus on $\{\liminf M_n > -\infty\}$, $\lim_{n\to\infty} M_n$ exists a.s. Prove an analogous statement regarding $\{\liminf M_n < \infty\}$. What is the complement of the union of these two events?

   Applications

   (b) Use the results above to prove that the simple symmetric random walk satisfies $\liminf_{n\to\infty} S_n = -\infty$ and $\limsup_{n\to\infty} S_n = \infty$, a.s.

   (c) Let $(A_n : n \in \mathbb{N})$ be a sequence of events. Let $\mathcal{F}_0 := \{\emptyset, \Omega\}$ and for $n \in \mathbb{N}$ define $\mathcal{F}_n := \sigma(A_1, \dots, A_n)$. Let $D := \{\sum_{n=1}^\infty E[\mathbf{1}_{A_n} | \mathcal{F}_{n-1}] = \infty\}$. Show that the following holds a.s. : on $D$, $\{A_n \text{ i.o.}\}$ and on $D^c$, $\{A_n \text{ f.o.}\}$. What two results we have extensively used in the past this extends?

3. **Biased Simple Random Walk**
   Let $x \in \mathbb{Z}$, $p \in (0, \frac{1}{2})$ and $(X_n : n \in \mathbb{N})$ IID with $P(X_n = 1) = p = 1 - P(X_n = -1)$. Let $S_0 := x$ and continue inductively, letting $S_{n+1} := S_n + X_{n+1}$. For any $y \in \mathbb{Z}$, let $T_y = \inf\{n \in \mathbb{Z}_+ : S_n = y\}$. Let $a, b \in \mathbb{Z}$ with $a < x < b$.

   (a) Find $\rho > 0$ such that $(\rho^{S_n} : n \in \mathbb{Z}_+)$ is a martingale.

   (b) Find $P(T_a < T_b)$.

   (c) Calculate $E[T_a \wedge T_b]$.

4. **Backward Martingale and LLN**
   Let $(X_n : n \in \mathbb{N})$ be IID with expectation $\mu$. Let $\mathbf{S} := (S_n : n \in \mathbb{N})$ be the sequence of partial sums $S_n := \sum_{k=1}^n X_k$. For $k \in \mathbb{N}$, define $\mathcal{G}_{-k} := \sigma(S_k, S_{k+1}, \dots)$. Thus $\mathcal{G}_{-k} \subseteq \mathcal{G}_{-k+1} \subseteq \cdots \subseteq \mathcal{G}_{-1} = \sigma(S_1, S_2, \dots)$, and let $M_{-k} := E[X_1 | \mathcal{G}_{-k}]$.

   (a) Show that $M_{-k} = S_k/k$.

   (b) Show that for each $N \in \mathbb{N}$, the process $(M_{-N}, M_{-N+1}, \dots, M_{-1})$ is a martingale with respect to the filtration $(\mathcal{G}_{-N}, \mathcal{G}_{-N+1}, \dots, \mathcal{G}_{-1})$.

   (c) Apply Doob's upcrossing inequality to show that for any $a < b$, the expected number of upcrossings of $[a, b]$ by the process above is bounded by some constant independent of $N$.

   (d) Conclude that $\lim_{k\to\infty} M_{-k}$ converges a.s. and that the limit is constant (a zero-one law).

   (e) Show that the constant of the last part is $E[X_1]$.

   Hint: show that you can assume without loss of generality that $X_1 \ge 0$ and under this assumption, $\lim_{M\to\infty} \sup_k E[(S_k/k), S_k/k > M] = 0$. Show how this leads to the desired conclusion.

   We've got ourselves another proof of the law of large numbers.

5. **Levy's Zero-One Law**
   Recall that a sequence of RVs $(X_n : n \in \mathbb{N})$ converges to an integrable RV $X_\infty$ in $L^1$ if $\lim_{n\to\infty} E[|X_n - X_\infty|] = 0$.

   Let $\mathfrak{F} = (\mathcal{F}_n : n \in \mathbb{Z}_+)$ be a filtration, and let $\mathcal{F}_\infty$ be the sigma-algebra generated by $\cup_{n\in\mathbb{N}} \mathcal{F}_n$. Suppose that $X$ is an integrable random variable measurable with respect to $\mathcal{F}_\infty$.

(a) Prove that $\lim_{n\to\infty} E[X|\mathcal{F}_n] = X$ a.s. and also in $L^1$.

**Applications**

(b) Suppose that $(X_n : n \in \mathbb{N})$ are RVs such that $\lim_{n\to\infty} X_n = X_\infty$ in $L^1$. Prove that $\lim_{n\to\infty} E[X_n|\mathcal{F}_n] = X$ in $L^1$.

(c) (Levy's Zero-One law) Conclude from (a) that for any event $A \in \mathcal{F}_\infty$, $E[\mathbf{1}_A|\mathcal{F}_n] \to \mathbf{1}_A$ a.s. and in $L^1$.

(d) Apply the last result to prove Kolmogorov's Zero-One law.

## 6.5 Assignment 6

The following refer to a single chapter from our lecture notes of your choice. Please fully cite all sources you have used.

1. Find at least 5 corrections or improvements to the presentation. Examples include: typos and other errors, missing or incorrect references, missing, incorrect, inconsistent notation.

2. Propose an additional result related to the chapter chosen and which has not been discussed or presented in the notes.

3. Propose two homework problems which have not appeared in our course and write their solution. The problems must be directly related to our course and the topic we covered, and should be above the level of a direct application of an existing result, generally the level of our homework problems. Math Stack Exchange is a great source of problems.

## 6.6 Topics for Essay 1

All references below are from the book: Probability: Theory and Examples, Version 5, 2019-01-11. You can freely use this book, but you must include at least one other reference. The first topic focuses on the foundations we covered in the beginning of the course, the second is closely related to laws of large numbers and the Poisson convergence, and the third is on weak convergence.

### Hewitt-Savage Zero-One Law, Sec. 2.5

Define and explain the notion of the exchangeable $\sigma$-algebra associated with an sequence of identically distributed RVs. Discuss its connection with the Tail $\sigma$-algebra, and give concrete examples of invariant events which are not tail events, at least one not in the book. State and Prove Hewitt-Savage Zero-One Law, Theorem 2.5.4, with concrete applications, at least one not in the book and which cannot be done with Kolmogorov's Zero-One Law.

### Occupation Problem, Sec. 2.2.2

Expand Example 2.2.10 to the following:

1. Weak limit for the the number of balls in each box.

2. Strong law of large numbers for the proportion of boxes with $k$ balls for any $k \in \mathbb{Z}_+$.

### The Method of Moments, Sec. 3.3.5

Describe the method of moments for weak convergence, including statement and proof of Theorem 3.3.26, at least one application not in the book, and an example showing two distinct distribution with identical moments and a discussion on why the conditions in the theorem do not hold.

## 6.7 Topics for Essay 2

As with the first essay, all references below are from the book: Probability: Theory and Examples, Version 5, 2019-01-11. You can freely use this book, but you must include at least two other references. I'll be happy to help with references and the work itself.

### Large Deviations, Sec. 2.7

State and prove Cramer's Theorem, Th. 2.7.7. Provide an example for a calculation of the rate function and give at least one application not in the book.

### De-Finetti's Theorem, Sec. 4.7

Define what an exchangeable sequence of random variables is. Provide some non trivial, not IID examples (e.g. Polya's Urn). State and prove De-Finetti's Theorem, Th. 4.7.9 and Th. 4.7.10. Show that the conclusion of the latter fails to hold for finite sequences. Find at least one application not in the book.

### Blackwell's Renewal Theorem, Sec. 2.6

State and prove Th. 2.6.4 and give at least one application not in the book.

### Berry Esseen Theorem, Sec. 3.4.4

Warning: A pretty technical project. State and Prove Th. 3.4.17. Present at least one application.

### Local Limit Theorems, Sec. 3.5

State and prove Th. 3.5.2, and either Th. 3.5.3 or 3.5.4. Provide at least one application.

### Stable Laws (Distributions), Sec. 3.8

State and prove Th. 3.8.2 and Th. 3.8.8, and give at least one example from the book.