CHAPTER 1

Probability Review.

1.1. Probability spaces. Sigma algebras.

We will use the notation from the measure theory $(\Omega, \mathscr{F}, \mathbf{P})$[1] for a probability space. Let us look to the constituent elements one at a time.

Let Ω is an abstract set. It is a set containing all the possible outcomes or results of a random experiment or phenomenon. I called it abstract because it could contain anything. For example if the experiment consists in tossing a coin once the space Ω could be represented as $\{Head, Tail\}$. However, it could just as well be represented as $\{Cap, Pajura\}$, these being the romanian equivalents of $Head$ and $Tail$. The space Ω could as well contain an infinite number of elements. For example measuring the diameter of a doughnut could result in possible numbers inside a whole range. Furthermore, measuring in inches or in centimeters would produce different albeit equivalent spaces.

We will use $\omega \in \Omega$ to denote a generic outcome or a sample point. We will use capital letters from the beginning of the alphabet A, B, C to denote events (any collection of outcomes).

We need to measure these events so we come to the next notion. The collection of events $\mathscr{F}$ represents the domain of definition for the function $\mathbf{P}$. We will need to provide internal consistencies when we define $\mathscr{F}$ to make sure that we are able to measure the information resulting from the experiment and any other event of possible interest to us. The mathematical structure for this purpose is the notion of σ-algebra (or σ-field). Before we define a σ-algebra, we will introduce a special collection of events:

$$\mathscr{P}(\Omega) = \text{The collection of all possible subsets of } \Omega = 2^{\Omega} \tag{1.1}$$

Exercise 1. Roll a die. Then $\Omega = \{1, 2, 3, 4, 5, 6\}$. An example of a event is $A = \{$ Roll an even number$\} = \{2, 4, 6\}$. Find the cardinality (number of elements of $\mathscr{P}(\Omega)$ in this case.

[1] Sometimes (specially in statistics) the whole setup is denoted with $(S, \Sigma, \mathbf{P})$

Having defined sets we can now define operations with them: *union, intersection, complement* and slightly less important *difference and symmetric difference.*

(1.2)

$$\begin{cases} A \cup B & = \text{set of elements that are } \textbf{either} \text{ in } A \textbf{ or} \text{ in } B \\ A \cap B & = AB = \text{set of elements that are } \textbf{both} \text{ in } A \textbf{ and} \text{ in } B \\ A^c & = \bar{A} = \text{set of elements that are in } \Omega \text{ but } \textbf{not} \text{ in } A \end{cases}$$

$$\begin{cases} A \setminus B = & \text{set of elements that are in } A \text{ but } \textbf{not} \text{ in } B \\ A \triangle B = & (A \setminus B) \cup (B \setminus A) \end{cases}$$

We can of course express every operation in terms of union and intersection. There are important relations between these operations, I will stop here with the mention of the De Morgan laws:

$$\begin{cases} (A \cup B)^c & = A^c \cap B^c \\ (A \cap B)^c & = A^c \cup B^c \end{cases} \tag{1.3}$$

Definition 1.1 (Algebra on Ω)**.** A collection $\mathscr{F}$ of events in Ω is called an algebra (or field) on Ω iff:

a) $\Omega \in \mathscr{F}$
b) Closed under complementarity: If $A \subseteq \mathscr{F}$ then $A^c \subseteq \mathscr{F}$
c) Closed under finite union: If $A, B \subseteq \mathscr{F}$ then $A \cup B \subseteq \mathscr{F}$

Remark 1.2. The first two properties imply that $\varnothing \in \mathscr{F}$. The third is equivalent by de Morgan laws (1.3) with $A \cap B \subseteq \mathscr{F}$

Definition 1.3 (σ-Algebra on Ω)**.** If $\mathscr{F}$ is an algebra on Ω and in addition it is closed under countable unions then it is a σ-algebra (or σ-field) on Ω

Note: Closed under countable unions means that the property c) in Definition 1.1 is replaced with: If $n \in \mathbb{N}$ is a natural number and $A_n \subseteq \mathscr{F}$ for all n then

$$\bigcup_{n \in \mathbb{N}} A_n \subseteq \mathscr{F}.$$

From b) and c) it of course follows that the σ-algebra is also closed under countable intersection. (via De Morgan's laws)

The σ-algebra provides an appropriate domain of definition for the probability function. However, it is such an abstract thing that it will be hard to work with it. This is the reason for the next definition, it will be much easier to work with the generators of a σ-algebra. *This will be a recurring theme in probability, in order to show a property for*

a big class we show the property for a small generating set of the class and then use standard arguments to extend to the whole class.

Definition 1.4 (σ algebra generated by a class of Ω)**.** Let $\mathscr{C}$ be a collection (class) of subsets of Ω. Then $\sigma(\mathscr{C})$ is the smallest σ-algebra on Ω that contains $\mathscr{C}$. The class $\mathscr{C}$ is called the generator of the σ-algebra.

Mathematically:

(a) $\mathscr{C} \subseteq \sigma(\mathscr{C})$
(b) $\sigma(\mathscr{C})$ is a σ-field
(c) If $\mathscr{C} \subseteq \mathscr{G}$ and $\mathscr{G}$ is a σ-field then $\sigma(\mathscr{C}) \subseteq \mathscr{G}$

Remark 1.5 (Properties of σ-algebras:)**.**

- $\mathscr{P}(\Omega)$ is a σ-algebra, the largest possible σ-algebra on Ω
- If $\mathscr{F}$ is already a σ-algebra then $\sigma(\mathscr{F}) = \mathscr{F}$
- If $\mathscr{F} = \{\varnothing\}$ or $\mathscr{F} = \{\Omega\}$ then $\sigma(\mathscr{F}) = \{\varnothing, \Omega\}$, the smallest possible σ-algebra on Ω
- If $\mathscr{F} \subseteq \mathscr{F}'$ then $\sigma(\mathscr{F}) \subseteq \sigma(\mathscr{F}')$
- If $\mathscr{F} \subseteq \mathscr{F}' \subseteq \sigma(\mathscr{F})$ then $\sigma(\mathscr{F}') = \sigma(\mathscr{F})$

Remark 1.6 (Finite space Ω)**.** When the sample space is finite, we can and typically will take the sigma algebra to be $\mathscr{P}(\Omega)$. Indeed, any event of a finite space can be trivially expressed in terms of individual outcomes. In fact, if the finite space Ω contains M possible outcomes, then the number of possible events is finite and is equal with 2^M.

1.2. An Example: Borel σ-algebra.

Let Ω be a topological space (think geometry exists in this space this assures us that the open subsets exist in this space).

Definition 1.7. We define:

$$\begin{aligned} \mathscr{B}(\Omega) &= \text{The Borel } \sigma\text{-algebra} \\ &= \sigma\text{-algebra generated by the class of open subsets of } \Omega \end{aligned} \tag{1.4}$$

In the special case when $\Omega = \mathbb{R}$ we denote $\mathscr{B} = \mathscr{B}(\mathbb{R})$. $\mathscr{B}$ is *the most important* σ-algebra. The reason for that is: most experiments can be brought to equivalence with $\mathbb{R}$. Thus, if we define a probability measure on $\mathscr{B}$, we have a way to calculate probabilities for most experiments.

Most subsets of $\mathbb{R}$ are in $\mathscr{B}$. However, it is possible (though very difficult) to construct a subset of $\mathbb{R}$ explicitly which is not in $\mathscr{B}$. See [**Bil95**] page 45 for such a construction in the case $\Omega = (0, 1]$.

There is nothing special about the open sets, except for the fact that they can be defined in any topological space. In $\mathbb{R}$ we have the alternate definition which you will have to prove:

Exercise 2. Show that the following classes all generate the Borel σ-algebra, or put differently show the equality of the following collections of sets:

$$\sigma\left((a,b) : a < b \in \mathbb{R}\right) = \sigma\left([a,b] : a < b \in \mathbb{R}\right) = \sigma\left((-\infty, b) : b \in \mathbb{R}\right) \\ = \sigma\left((-\infty, b) : b \in \mathbb{Q}\right),$$

where $\mathbb{Q}$ is the set of rational numbers.

1.3. Probability Measure

We are finally in the position to define a space on which we can introduce the probability measure.

Definition 1.8 (Measurable Space.)**.** A pair $(\Omega, \mathscr{F})$, where Ω is a set and $\mathscr{F}$ is a σ-algebra on Ω is called a *Measurable Space*.

Definition 1.9 (Probability measure. Probability space)**.** Given a measurable space $(\Omega, \mathscr{F})$, a probability measure is any function $\mathcal{P} : \mathscr{F} \to [0,1]$ with the following properties:

i) $\mathcal{P}(\Omega) = 1$

ii) (countable additivity) For any sequence $\{A_n\}_{n\in\mathbb{N}}$ of disjoint events in $\mathscr{F}$ (i.e. $A_i \cap A_j = \varnothing$, for all $i \neq j$):

$$\mathcal{P}\left(\bigcup_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mathcal{P}(A_n)$$

The triple $(\Omega, \mathscr{F}, \mathcal{P})$ is called a Probability Space.

The next two definitions are given for completeness, however we will use them later in this class. They are both defining more general notions than a probability measure and they will be used later in hypotheses of some theorems to show that the results apply to even more general measures than probability measures.

Definition 1.10 (Finite Measure)**.** Given a measurable space $(\Omega, \mathscr{F})$, a finite measure is a set function $\mu : \mathscr{F} \to [0,1]$ with the same countable additivity property as defined above and the measure of the space finite instead of one. More specifically the first property above is replaced with:

$$\mu(\Omega) < \infty$$

Definition 1.11 (σ-finite Measure)**.** A measure μ defined on a measurable space $(\Omega, \mathscr{F})$ is called σ-finite if it is countably additive and

there exist a partition[2] of the space Ω, $\{\Omega_i\}_{i\in I}$, and $\mu(\Omega_i) < \infty$ for all $i \in I$. Note that the index set I is allowed to be countable.

Example 1.12 (Uniform Distribution on (0,1))**.** As an example let $\Omega = (0,1)$ and $\mathscr{F} = \mathscr{B}((0,1))$. Define a probability measure U as follows: for any open interval $(a,b) \subseteq (0,1)$ let $U((a,b)) = b - a$ the length of the interval. For any other open interval O define $U(O) = U(O \cap (0,1))$.

Note that we did not specify $U(A)$ for all Borel sets A, rather only for the generators of the Borel σ-field. This illustrates the probabilistic concept presented above. In our specific situation, under very mild conditions on the generators of the σ-algebra any probability measure defined only on the generators can be uniquely extended to a probability measure on the whole σ-algebra (Carathèodory extension theorem). In particular when the generators are open sets these conditions are true and we can restrict the definition to the open sets only.

Proposition 1.13 (Elementary properties of Probability Measure)**.** *Let $(\Omega, \mathscr{F}, \mathcal{P})$ be a Probability Space. Then:*

(1) $\forall A, B \in \mathscr{F}$ with $A \subseteq B$ then $\mathbf{P}(A) \leq \mathbf{P}(B)$
(2) $\mathbf{P}(A \cup B) = \mathbf{P}(A) + \mathbf{P}(B) - \mathbf{P}(A \cap B)$, $\forall A, B \in \mathscr{F}$
(3) (General Inclusion-Exclusion formula, also named Poincaré formula):

$$\mathbf{P}(A_1 \cup A_2 \cup \cdots \cup A_n) = \sum_{i=1}^{n} \mathbf{P}(A_i) - \sum_{i<j\leq n} \mathbf{P}(A_i \cap A_j) + \sum_{i<j<k\leq n} \mathbf{P}(A_i \cap A_j \cap A_k) - \cdots + (-1)^n \mathbf{P}(A_1 \cap A_2 \cdots \cap A_n) \tag{1.5}$$

Successive partial sums are alternating between over-and-under estimating.
(4) (Finite subadditivity, sometimes called Boole's inequality):

$$\mathbf{P}\left(\bigcup_{i=1}^{n} A_i\right) \leq \sum_{i=1}^{n} \mathcal{P}(A_i), \quad \forall A_1, A_2, \ldots, A_n \in \mathscr{F}$$

1.3.1. Conditional Probability. Independence. Borel-Cantelli lemmas. Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a Probability Space. Then for $A, B \in \mathscr{F}$, with $\mathbf{P}(B) \neq 0$ we define the conditional probability of A given B as

[2] a partition of the set A is a collection of sets A_i, disjoint ($A_i \cap A_j = \emptyset$, if $i \neq j$) such that $\cup_i A_i = A$

usual by:

$$\mathbf{P}(A|\, B) = \frac{\mathbf{P}(A \cap B)}{\mathbf{P}(B)}.$$

We of course also have the *chain rule formulas*:

$$\mathbf{P}(A \cap B) = \mathbf{P}(A|\, B)\mathbf{P}(B),$$
$$\mathbf{P}(A \cap B \cap C) = \mathbf{P}(A|\, B \cap C)\mathbf{P}(B|\, C)\mathbf{P}(C), \quad \text{etc.}$$

Total probability formula: Given $A_1, A_2, \ldots, A_n$ a partition of Ω (i.e. the sets A_i are disjoint and $\Omega = \bigcup_{i=1}^n A_i$), then:

$$\mathbf{P}(B) = \sum_{i=1}^{n} \mathbf{P}(B|\, A_i)\mathbf{P}(A_i), \quad \forall B \in \mathscr{F} \tag{1.6}$$

Bayes Formula: If $A_1, A_2, \ldots, A_n$ form a partition of Ω:

$$\mathbf{P}(A_j|\, B) = \frac{\mathbf{P}(B|\, A_j)\mathbf{P}(A_j)}{\sum_{i=1}^n \mathbf{P}(B|\, A_i)\mathbf{P}(A_i)}, \quad \forall B \in \mathscr{F}. \tag{1.7}$$

Definition 1.14 (Independence)**.** The events $A_1, A_2, A_3, \ldots$ are called *mutually independent* (or sometimes simply independent) if for every subset J of $\{1, 2, 3, \ldots\}$ we have:

$$\mathbf{P}\left(\bigcup_{j \in J} A_j\right) = \prod_{j \in J} \mathbf{P}(A_j)$$

The events $A_1, A_2, A_3, \ldots$ are called *pairwise independent* (sometimes jointly independent) if:

$$\mathbf{P}\left(A_i \cup A_j\right) = \mathbf{P}(A_i)\mathbf{P}(A_j), \quad \forall i, j.$$

Note that jointly independent does not imply independence.

Two *sigma fields* $\mathscr{G}, \mathscr{H} \in \mathscr{F}$ *are* $\mathbf{P}$*–independent* if:

$$\mathbf{P}(G \cap H) = \mathbf{P}(G)\mathbf{P}(H), \quad \forall G \in \mathscr{G}, \forall H \in \mathscr{H}.$$

See [**Bil95**] for the definition of independence of $k \geq 2$ sigma-algebras.

1.3.2. Monotone Convergence properties of probability. Let $(\Omega, \mathscr{F}, \mathcal{P})$ be a Probability Space.

Lemma 1.15. *The following are true:*

(i) If $A_n, A \in \mathscr{F}$ and $A_n \uparrow A$ (i.e., $A_1 \subseteq A_2 \subseteq \ldots A_n \subseteq \ldots$ and $A = \bigcup_{n \geq 1} A_n$), then: $\mathbf{P}(A_n) \uparrow \mathbf{P}(A)$ as a sequence of numbers.

(ii) If $A_n, A \in \mathscr{F}$ and $A_n \downarrow A$ (i.e., $A_1 \supseteq A_2 \supseteq \dots A_n \supseteq \dots$ and $A = \bigcap_{n\geq 1} A_n$), then: $\mathbf{P}(A_n) \downarrow \mathbf{P}(A)$ as a sequence of numbers.

(iii) (Countable subadditivity) If $A_1, A_2, \dots,$ and $\bigcup_{i=1}^{\infty} A_n \in \mathscr{F}$, with A_i's not necessarily disjoint then:

$$\mathbf{P}\left(\bigcup_{n=1}^{\infty} A_n\right) \leq \sum_{n=1}^{\infty} \mathcal{P}(A_n)$$

PROOF. (i) Let $B_1 = A_1, B_2 = A_2 \setminus A_1, \dots, B_n = A_n \setminus A_{n-1}$. Because the sequence is increasing we have that the B_i's are disjoint thus from Proposition 1.13 we obtain:

$$\mathbf{P}(A_n) = \mathbf{P}(B_1 \cup B_2 \cup \dots \cup B_n) = \sum_{i=1}^{n} \mathbf{P}(B_i).$$

Thus using countable additivity:

$$\mathbf{P}\left(\bigcup_{n\geq 1} A_n\right) = \mathbf{P}\left(\bigcup_{n\geq 1} B_n\right) = \sum_{i=1}^{\infty} \mathbf{P}(B_i) = \lim_{n\to\infty} \sum_{i=1}^{n} \mathbf{P}(B_i) = \lim_{n\to\infty} \mathbf{P}(A_n)$$

(ii) Note that $A_n \downarrow A \Leftrightarrow {A_n}^c \uparrow A^c$ which from part (i) implies that $1 - \mathbf{P}(A_n) \uparrow 1 - \mathbf{P}(A)$.

(iii) Let $B_1 = A_1, B_2 = A_1 \cup A_2, \dots, B_n = A_1 \cup \dots \cup A_n, \dots$. From the finite subadditivity we have that $\mathbf{P}(B_n) = \mathbf{P}(A_1 \cup \dots \cup A_n) \leq \mathbf{P}(A_1) + \dots + \mathbf{P}(A_n)$.

$\{B_n\}_{n\geq 1}$ is an increasing sequence of events, thus from (i) we get that $\mathbf{P}(\bigcup_{n=1}^{\infty} B_n) = \lim_{n\to\infty} \mathbf{P}(B_n)$. Combining the two relations above we obtain:

$$\mathbf{P}(\bigcup_{n=1}^{\infty} A_n) = \mathbf{P}(\bigcup_{n=1}^{\infty} B_n) \leq \lim_{n\to\infty} (\mathbf{P}(A_1) + \dots + \mathbf{P}(A_n)) = \sum_{n=1}^{\infty} \mathbf{P}(A_n)$$

□

Lemma 1.16. *The union of a countable number of $\mathbf{P}$-null sets is a $\mathbf{P}$-null set*

Exercise 3. Prove the above Lemma 1.16

Next we state one of the most fundamental (and useful) results in probability theory the Borel-Cantelli lemmas:

Lemma 1.17. *[The Borel-Cantelli lemmas] Let $(\Omega, \mathscr{F}, \mathcal{P})$ be a Probability Space. Let $A_1, A_2, \dots, A_n, \dots$ a sequence of events.*

First Lemma: If $\sum_{i\geq 1}\mathbf{P}(A_i)<\infty$ *then:*

$$\mathbf{P}\left(\bigcap_{n=1}^{\infty}\bigcup_{i\geq n}A_i\right)=\mathbf{P}\left(A_i\text{'s are true infinitely often}\right)=0$$

Second Lemma: If $\sum_{i\geq 1}\mathbf{P}(A_i)=\infty$*, and in addition the events* $A_1,A_2,\ldots,A_n,\ldots$ *are independent then:*

$$\mathbf{P}\left(\bigcap_{n=1}^{\infty}\bigcup_{i\geq n}A_i\right)=\mathbf{P}\left(A_i\text{'s are true infinitely often}\right)=1$$

Let us clarify the notion of "infinitely often" and "eventually" a bit more. We say that an outcome ω happens infinitely often for the sequence $A_1,A_2,\ldots,A_n,\ldots$ if ω is in the set $\bigcap_{n=1}^{\infty}\bigcup_{i\geq n}A_i$. This means that for any n (no matter how big) there exist an $m\geq n$ and $\omega\in A_m$.

We say that an outcome ω happens eventually for the sequence $A_1,A_2,\ldots,A_n,\ldots$ if ω is in the set $\bigcup_{n=1}^{\infty}\bigcap_{i\geq n}A_i$. This means that there exist an n such that for any $m\geq n$, $\omega\in A_m$, so for an n on ω is in all such sets.

Why so complicate definitions? The basic intuition is obvious: say you roll a die infinitely many times, then it is obvious what it means for the outcome 1 to appear infinitely often. Also say the average of the rolls will eventually be arbitrarily close to 3.5. It is not so clear cut in general. The framework above provides a generalization to these notions.

Exercise 4. Show using the Cantelli lemma that when you roll a die the outcome $\{1\}$ will appear infinitely often. Also show that eventually the average of all rolls up to roll n will be within ε of 3.5 where $\varepsilon>0$ is any arbitrary real number.

1.4. Measurable Functions. Random Variables

All of these definitions with sets are consistent, however if we wish to calculate and compute numerical values related to abstract spaces we need to standardize the spaces. The first step is to give the following definitions:

Definition 1.18 (Measurable function (m.f.))**.** Let $(\Omega,\mathscr{F})$ and (S,Σ) be two measurable spaces. A function f is called measurable (function or m.f.) if and only if (notation iff) for every set $A\in\Sigma$ we have $f^{-1}(A)\in\mathscr{F}$.

1.4.1. Reduction to $\mathbb{R}$. Random variables.

Definition 1.19 (Random variable (r.v.))**.** Any measurable function with codomain $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ is called a random variable.

Traditionally, the random variables are denoted with capital letters from the end of the alphabet $X, Y, Z, \ldots$ and their values are denoted with corresponding small letters $x, y, z, \ldots$.

Definition 1.20 (The distribution of a random variable)**.** Assume that on the measurable space $(\Omega, \mathscr{F})$ we define a probability measure so that it becomes a probability space $(\Omega, \mathscr{F}, \mathbf{P})$. If a random variable $X : \Omega \to \mathbb{R}$ is defined then we call its distribution, the set function μ defined on the Borel sets of $\mathbb{R}$: $\mathscr{B}(\mathbb{R})$, with values in $[0, 1]$:

$$\mu(B) = \mathbf{P}\left(\{\omega : X(\omega) \in B\}\right) = \mathbf{P}\left(X^{-1}(B)\right) = \mathbf{P} \circ X^{-1}(B)$$

Remark 1.21. First note that the measure μ is defined on sets in $\mathbb{R}$ and takes values in the interval $[0, 1]$. Therefore, the random variable X allowed us to apparently eliminate the abstract space Ω. However, this is not the case since we still have to calculate probabilities using $\mathbf{P}$ in the definition of μ above.

There is one more simplification we can make. If we use the result of the exercise 2, we see that all borel sets are generated by the same type of sets. Using the same idea as before it is enough to describe how to calculate μ for the generators. We could of course specify any type of generating sets we wish (open sets, closed sets, etc) but it turns out the simplest way is to use sets of the form $(-\infty, x]$, since we only need to specify one end of the interval (the other is always $-\infty$).

Definition 1.22. [The distribution function of a random variable] The distribution function of a random variable X is $F : \mathbb{R} \to [0, 1]$ with:

$$F(x) = \mu(-\infty, x] = \mathbf{P}\left(\{\omega : X(\omega) \in (-\infty, x]\}\right) = \mathbf{P}\left(\{\omega : X(\omega) \leq x\}\right)$$

But wait a minute, this is exactly the definition of the cumulative distribution function (cdf) you see in any lower level probability classes. It is exactly the same thing except that in an effort to dumb down (in whomever opinion it was to teach the class that way) the meaning is lost and we cannot proceed with more complicated things. From the definition above we can deduce all the elementary properties of the cdf that you have learned (right-continuity, increasing, taking values between 0 and 1). In fact let me ask you to prove this.

Exercise 5. Show that the function F in Definition 1.22 is increasing, right continuous and taking values in the interval $[0, 1]$, using proposition 1.13.

Definition 1.23 (PDF, PMF)**.** In general the distribution function F is not necessarily derivable. If it is, we call its derivative $f(x)$ the probability density function (pdf) and notice that we have in this situation:

$$F(x) = \int_{-\infty}^{x} f(z)dz$$

Traditionally, a variable X with this property is called a continuous random variable.

Furthermore if F is piecewise constant (i.e., constant almost everywhere, or in other words there exist a countable sequence $\{a_1, a_2, \dots\}$ such that the function F is constant for every point except these a_i's) and we denote $p_i = F(a_i) - F(a_i-)$, then the collection of p_i's is the traditional probability mass function (pmf) that characterizes a discrete random variable. ($F(x-)$ is a notation for the left limit of function F at x or in a more traditional notation $\lim_{z\to x, z<x} F(z)$).

Also notice that traditional undergraduate textbooks segregate between discrete and continuous random variables. In fact there are many more and the definitions we used here cover all of them, likewise the treatment of random variables should be the same, which is now possible.

Important. So what is the point of all this? What did we just accomplish here? The answer is moving from the abstract space $(\Omega, \mathscr{F}, P)$ to something perfectly equivalent but defined on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Because of this fact we only need to define probability measures on $\mathbb{R}$ and show that things coming from our original abstract space are equivalent with these distributions on $\mathbb{R}$. We just constructed the first model for our problem.

Next we will define the simplest and one of the most important random variables.

Definition 1.24 (Indicator Function)**.** We define the indicator function of an event A as the function $\mathbf{1}_A : \Omega \to \{0, 1\}$,

$$\mathbf{1}_A(\omega) = \begin{cases} 1 & , \quad \text{if } \omega \in A \\ 0 & , \quad \text{if } \omega \notin A \end{cases}$$

Remember this definition, it is one of the most important ones in probability. We can build on it in the following way:

Furthermore, this variable is also called the Bernoulli random variable. Notice that the variable only takes values 0 and 1 and the probability that the variable is 1 we can calculate easily using the previous definitions as being:

$$\mathbf{P} \circ \mathbf{1}_A^{-1}(\{1\}) = \mathbf{P}\{\omega : \mathbf{1}_A(\omega) = 1\} = \mathbf{P}(A).$$

Therefore the variable is distributed as a Bernoulli random variable with parameter $p = \mathbf{P}(A)$.

Definition 1.25 (Elementary (Simple) Function)**.** An elementary function g is any linear combination of the indicator functions just introduced. More specifically, there exist sets $A_1, A_2, \ldots, A_n$ all in $\mathscr{F}$ and constants $a_1, a_2, \ldots, a_n$ in $\mathbb{R}$ such that:

$$g(\omega) = \sum_{i=1}^{n} a_i \mathbf{1}_{A_i}(\omega).$$

Note that the sets A_i do not have to be disjoint but an easy exercise shows that g could be written in terms of disjoint sets.

Exercise 6. Show that any simple function g can be written as $\sum_i b_i \mathbf{1}_{B_i}$ with B_i disjoint sets (i.e. $B_i \cap B_j = \emptyset$, if $i \neq j$).

Exercise 7. Let $f : (\Omega, \mathscr{F}) \to [0, \infty]$ be a non-negative and measurable function. For all $n \geq 1$, we define:

$$g_n(\omega) := \sum_{k=0}^{n2^n - 1} \frac{k}{2^n} \mathbf{1}_{\{\frac{k}{2^n} \leq f(\omega) < \}} + n\mathbf{1}_{\{f(\omega) \geq n\}} \tag{1.8}$$

(1) Show that g_n is a simple function on $(\Omega, \mathscr{F})$, for all $n \geq 1$.
(2) Show that equation (1.8) gives a partition, for all $n \geq 1$.
(3) Show that $g_n \leq g_{n+1} \leq f$, for all $n \geq 1$.
(4) Show that $g_n \uparrow f$ as $n \to \infty$[3].

1.4.2. Null element of $\mathscr{F}$. Almost sure (a.s.) statements. An event $N \in \mathscr{F}$ is called a null event if $P(N) = 0$.

Definition 1.26. A statement $\mathcal{S}$ about points $\omega \in \Omega$ is said to be true *almost surely* (a.s.), almost everywhere (a.e.) or with probability 1 (w.p.1) if the set N defined as:

$$N := \{\omega \in \Omega | \, \mathcal{S}(\omega) \text{ is true}\},$$

is in $\mathscr{F}$ and $\mathbf{P}(N) = 1$, (or N^c is a null set).

We will use the notions a.s., a.e., and w.p.1. interchangeably to denote the same thing – the definition above. For example we will say $X \geq 0$ a.s. and mean: $\mathbf{P}\{\omega | X(\omega) \geq 0\} = 1$ or equivalently $\mathbf{P}\{\omega | X(\omega) < 0\} = 0$. The notion of almost sure is a fundamental one in probability. Unlike in deterministic cases where something has

[3]This is not a.s., it is for all ω

to always be true no matter what, in probability we care about "the majority of the truth". In other words probability recognizes that some events may have extreme outcomes, but if they are extremely improbable then we do not care about them. Fundamentally, it is mathematics applied to reality.

1.4.3. Joint distribution, Random vectors. We talked about σ-algebras in the beginning and they kind of faded away after that. We will come back to them. It turns out, if there is any hope of rigorous introduction into probability and stochastic processes, they are *unavoidable*. Later, when we will talk about stochastic processes we will find out the *crucial* role they play in quantifying the information available up to a certain time. For now let us play a bit with them.

Definition 1.27 (σ-algebra generated by a random variable)**.** For a r.v. X we call the σ-algebra generated by X, denoted $\sigma(X)$ or sometime $\mathscr{F}_X$, the smallest σ-field $\mathscr{G}$ such that X is measurable on $(\Omega, \mathscr{G})$. It is the σ-algebra generated by the pre-images of Borel sets through X. Because of this we can easily show (remember that the Borel sets are generated by intervals of the type $(-\infty, \alpha]$):

$$\sigma(X) = \sigma(\{\omega | X(\omega) \leq \alpha\}, \text{ as } \alpha \text{ varies in } \mathbb{R}).$$

Similarly, given $X_1, X_2, \ldots, X_n$ random variables, we call the sigma algebra generated by them the smallest sigma algebra such that all are measurable with respect to it. It turns out we can show easily that it is the sigma algebra generated by the union of the individual sigma algebras or put more specifically $\sigma(X_i, i \leq n)$ is the smallest sigma algebra containing all $\sigma(X_i)$, for $i = 1, 2, \ldots, n$.

In the previous subsection we defined a random variable as a measurable function with codomain $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. A more specific case is when the random variable has also the domain equal to $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. In this case we talk about Borel functions.

Definition 1.28 (Borel measurable function)**.** A function $g : \mathbb{R} \to \mathbb{R}$ is called Borel (measurable) function if g is a measurable function from $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ into $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$.

Exercise 8. Show that any continuous function $g : \mathbb{R} \to \mathbb{R}$ is Borel measurable.

Hint: Look to what happens to the preimage of sets through a continuous function.

Exercise 9. Show that any piecewise constant function is Borel measurable. (see description of piecewise constant functions in Definition 1.22

In Section 1.2 we defined Borel sigma algebras corresponding to any space Ω. We presented the special case when $\Omega = \mathbb{R}$. It really is no big deal to consider $\Omega = \mathbb{R}^n$, for some integer n, and the Borel sigma algebra generated by it. This allows us to define a random vector on $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n), \mathbf{P})$ as $(X_1, X_2, \dots, X_n)$ where each X_i is a random variable. The probability $\mathbf{P}$ is defined on $\mathscr{B}(\mathbb{R}^n)$.

We can talk about its distribution (the "joint distribution" of the variables $(X_1, X_2, \dots, X_n)$) as the function:

$$\begin{aligned} F(x_1, x_2, \dots, x_n) &= \mathbf{P} \circ (X_1, X_2, \dots, X_n)^{-1} \left((-\infty, x_1] \times \cdots \times (-\infty, x_n]\right) \\ &= \mathbf{P}(X_1 \leq x_1, X_2 \leq x_2, \dots, X_n \leq x_n) \end{aligned}$$

We can introduce the notions of independence and joint independence using the definition in subsection 1.3.1, the probability measure $= \mathbf{P} \circ (X_1, X_2, \dots, X_n)^{-1}$ and any Borel sets. Writing more specifically it is transformed to:

Definition 1.29. The variables $(X_1, X_2, \dots, X_n, \dots)$ are independent if for every subset $J = \{j_1, j_2, \dots, j_k\}$ of $\{1, 2, 3, \dots\}$ we have:

$$\mathbf{P}\left(X_{j_1} \leq x_{j_1}, X_{j_2} \leq x_{j_2}, \dots, X_{j_k} \leq x_{j_k}\right) = \prod_{j \in J} \mathbf{P}(X_j \leq x_j)$$

1.5. Expectations of random variables.

We note that the distribution function $F(x)$ exists for any random variable. We can construct the integral with respect to F using the integration theory (details are omitted in this class) starting from indicators for which we have:

$$\mathbf{E}\left[\mathbf{1}_A\right] = \int_\Omega \mathbf{1}_A(\omega) dP(\omega) = \mathbf{P}(A)$$

In general we can construct the expectation of an integrable ($\mathbb{E}|X| < \infty$) random variable X as:

$$\mathbf{E}\left[X\right] = \int_\Omega X(\omega) dP(\omega) = \int_{-\infty}^{\infty} x dP \circ X^{-1}(x) = \int_{-\infty}^{\infty} x dF(x),$$

where we have used the transport formula (change of variable) which you can find in any graduate probability textbook. Furthermore, for any function $h : \mathbb{R} \to \mathbb{R}$ we of course can further define:

$$\mathbf{E}\left[h(X)\right] = \int_\Omega h(X(\omega)) dP(\omega) = \int_{-\infty}^{\infty} h(x) dF(x).$$

In the case when F is derivable with derivative $f(x)$ we can of course write: $dF(x) = f(x)dx$, therefore the formula reduces to the more familiar one from elementary probability classes. If F is piecewise constant then its derivative is zero a.e. and the integral reduces to a sum bringing back the formula for the expectation of a discrete random variable.

Exercise 10. Write the probability space $(\Omega, \mathscr{F}, \mathbf{P})$ for a random experiment which records the result of n independent rolls of a balanced six-sided die (including the order). Compute the expectation of the random variable $D(\omega)$ which counts the number of different sides of the die recorded during these n rolls.

The variance of a random variable X is the expectation of the function $h(x) = (x - \mu)^2$ where μ is a notation for $\mathbb{E}X$. The covariance of random variables X and Y is the expectation of the function $h(x, y) = (x - \mu_X)(y - \mu_Y)$ where again μ is a notation for the expectations of the respective random variables. The correlation is the ratio of covariance to the product of the square root of variations. More specifically:

$$\begin{aligned}
\mathbf{V}(X) &= \mathbb{E}[(x-\mu)^2] = \mathbb{E}X^2 - (\mathbb{E}X)^2 \\
\mathbf{Cov}(X,Y) &= \mathbb{E}\left[(X-\mu_X)(Y-\mu_Y)\right] = \mathbb{E}XY - \mathbb{E}X\mathbb{E}Y \\
\mathbf{Corr}(X,Y) &= \frac{\mathbf{Cov}(X,Y)}{\sqrt{\mathbf{V}(X)\mathbf{V}(Y)}}
\end{aligned}$$

The variable X and Y are called **uncorrelated** if the covariance (or equivalently the correlation) between them is zero. **Note** that this is not the same as the variables X and Y being independent. Independence implies that the variables are uncorrelated, however the converse is not true.

Exercise 11. Give an example of two variables X and Y which are uncorrelated but not independent.

Proposition 1.30 (Elementary properties of the expectation)**.** *The expectation has the following properties:*

(i) $\mathbf{E}[\mathbf{1}_A] = \mathbf{P}(A)$ *for any* $A \in \mathscr{F}$
(ii) If $g(\omega) = \sum_{i=1}^n a_i \mathbf{1}_{A_i}(\omega)$ *is an elementary function then* $\mathbb{E}[g] = \sum_{i=1}^n a_i \mathbf{P}(A_i)$.
(iii) If X *and* Y *are integrable r.v.'s then for any constants* α *and* β *the r.v.* $\alpha X + \beta Y$ *is integrable and* $\mathbb{E}[\alpha X + \beta Y] = \alpha\mathbb{E}X + \beta\mathbb{E}Y$.
(iv) If $X(\omega) = c$ *with probability 1 then* $\mathbb{E}X = c$.

(v) If $X \geq Y$ a.s. then $\mathbb{E}X \geq \mathbb{E}Y$. Furthermore, if $X \geq Y$ a.s. and $\mathbb{E}X = \mathbb{E}Y$ then $X = Y$ a.s.

We use the notation $L^1(\Omega)$ or sometimes just L^1 to denote the space of integrable random variables. In general:

$$L^p(\Omega) = \{X \text{ random variable s.t. } \mathbb{E}\,|X|^p < \infty\} \quad , \forall p \geq 1$$

We can make $L^p = L^p(\Omega) = L^p(\Omega, \mathscr{F}, \mathbf{P})$ a normed (metric) space by introducing the p-norm of an element (random variable) in L^p as:

$$\|X\|_p = \sqrt[p]{\mathbb{E}\left[X^p\right]}$$

1.6. Conditional Probability. Conditional Expectation.

Please read pages 5 to 9 for the definitions of conditional probability and expectation conditioned by the sigma algebra generated by a random variable.

Why do we need conditional expectation?

Conditional expectation is a fundamental concept in the theory of stochastic processes. The simple idea is the following: suppose we have no information about a certain variable then our best guess of it most of the time would be some sort of regular expectation. However, in real life it often happens that we have some partial information about the random variable (or in time we come to know more about it). Then what we should do is every time there is new information the sample space Ω or the σ-algebra $\mathscr{F}$ is changing so they need to be recalculated. That will in turn change the probability $\mathbf{P}$ which will change the expectation of the variable. The conditional expectation provides a way to recalculate the expectation of the random variable given any new "consistent" information without going through the trouble of recalculating $(\Omega, \mathscr{F}, \mathbf{P})$ every time.

It is also easy to reason that since we calculate with respect to more precise information it will be depending on this more precise information, thus it is going to be a random variable itself, "adapted" to this information.

Going back, to summarize the book notation, if X and Y are two random variables the authors define in the pages mentioned the expectation of X conditioned by the sigma-algebra generated by Y, $\sigma(Y)$ and they use the notation:

$$\mathbb{E}[X|Y] = \mathbb{E}[X|\sigma(Y)].$$

Note that the conditional expectation, unlike the regular expectation is a random variable measurable with respect to the sigma algebra

under which is conditioned, (in the above case with respect to $\sigma(Y)$). In general I will give you the following more general definition / theorem. We will skip the proof.

Theorem 1.31. *Let $(\Omega, \mathscr{F}, \mathbf{P})$ be a probability space, and let $\mathscr{K} \subseteq \mathscr{F}$ a sub-σ-algebra. Let X be a random variable on $(\Omega, \mathscr{F}, \mathbf{P})$ such that either X is positive or $X \in L^1(\Omega)$. Then there exist a random variable Y, measurable with respect to $\mathscr{K}$ with the property:*

$$\int_A Y dP = \int_A X dP \quad , \forall A \in \mathscr{K}$$

This Y is defined to be the conditional expectation of X with respect to $\mathscr{K}$ or using the notation $\mathbb{E}[X|\mathscr{K}]$.

Note that by construction Y is a $\mathscr{K}$-measurable random variable.

Proposition 1.32 (Properties of the Conditional Expectation)**.** *Let $(\Omega, \mathscr{F}, \mathbf{P})$ a probability space, and let $\mathscr{K}, \mathscr{K}_1, \mathscr{K}_2$ sub-σ-algebras. Let X and Y be random variables of the probability space. Then we have:*

(1) If $\mathscr{K} = \{\varnothing, \Omega\}$ then $\mathbb{E}[X|\mathscr{K}] = \mathbb{E}X = const.$
(2) $\mathbb{E}[\alpha X + \beta Y|\mathscr{K}] = \alpha\mathbb{E}[X|\mathscr{K}] + \beta\mathbb{E}[Y|\mathscr{K}]$ for α, β real constants.
(3) If $X \leq Y$ a.s. then $\mathbb{E}[X|\mathscr{K}] \leq \mathbb{E}[Y|\mathscr{K}]$
(4) $\mathbb{E}\left[\mathbb{E}[X|\mathscr{K}]\right] = \mathbb{E}X$
(5) If $\mathscr{K}_1 \subseteq \mathscr{K}_2$ then

$$\mathbb{E}\left[\mathbb{E}[X|\mathscr{K}_1]|\,\mathscr{K}_2\right] = \mathbb{E}\left[\mathbb{E}[X|\mathscr{K}_2]|\,\mathscr{K}_1\right] = \mathbb{E}[X|\mathscr{K}_1]$$

(6) If X is independent of $\mathscr{K}$ then

$$\mathbb{E}[X|\mathscr{K}] = \mathbb{E}[X]$$

(7) If Y is measurable with respect to $\mathscr{K}$ then

$$\mathbb{E}[XY|\mathscr{K}] = Y\mathbb{E}[X|\mathscr{K}]$$

Exercise 12. Using the Theorem-Definition 1.31 prove the seven properties of the conditional expectation in Proposition 1.32.

1.7. Generating Functions. Moment generating functions (Laplace Transform). Characteristic Function (Fourier transform)

Please read at a minimum the information in your textbook (pages 10-14) and supplement it with information from any probability textbook (including those referenced in the syllabus).

1.8. Identities. Inequalities. General Theorems

Proposition 1.33 (Jensen's Inequality)**.** *Suppose $f(\cdot)$ is a convex function, that means:*

$$f(\alpha x + (1-\alpha)y) \leq \alpha f(x) + (1-\alpha)f(y) \quad , \forall x, y \in \mathbb{R}, \forall \alpha \in [0,1].$$

Then for any integrable variable X such that $f(X) \in L^1$ we have:

$$f\left(\mathbb{E}X\right) \leq \mathbb{E}\left[f(X)\right]$$

Proof. skipped. The classical approach indicators $\rightarrow$ simple functions $\rightarrow$ positive measurable $\rightarrow$ measurable is a standard way to prove Jensen. □

Proposition 1.34 (Markov Inequality)**.** *Suppose that $g(\cdot)$ is a nondecreasing, positive measurable function. Then for any random variable X and any $\varepsilon > 0$ we have:*

$$\mathbf{P}\left(|X(\omega)| > \varepsilon\right) \leq \frac{\mathbb{E}\left[g\left(|X|\right)\right]}{g(\varepsilon)}.$$

Proof. Let $A = \{\omega : |X(\omega)| > \varepsilon\}$. We want to get to probability of A. We have using the fact that g is nonnegative:

$$\mathbb{E}\left[g\left(|X|\right)\right] = \mathbb{E}\left[g\left(|X|\right)\mathbf{1}_A\right] + \mathbb{E}\left[g\left(|X|\right)\mathbf{1}_{A^c}\right] \geq \mathbb{E}\left[g\left(|X|\right)\mathbf{1}_A\right].$$

On the set A the argument of g is greater than ε. Using this fact and that g is nondecreasing we have on A, $g\left(|X|\right) > g(\varepsilon)$. Thus we can continue:

$$\mathbb{E}\left[g\left(|X|\right)\mathbf{1}_A\right] \geq \mathbb{E}\left[g\left(\varepsilon\right)\mathbf{1}_A\right] = g(\varepsilon)\mathbf{P}(A).$$

Dividing with $g(\varepsilon)$ yields the desired result. □

Example 1.35 (Special cases of Markov Inequality)**.** These are the most common cases of the use of Markov's inequality.

(i) Take $X > 0$ a.s. and $g(x) = x$. Then we get:

$$\mathbf{P}\left(|X(\omega)| > \varepsilon\right) \leq \frac{\mathbb{E}X}{\varepsilon}$$

(ii) Take $g(x) = x^2$ and $X = Y - \mathbb{E}Y$, we then obtain:

$$\mathbf{P}\left(|Y - \mathbb{E}Y| > \varepsilon\right) \leq \frac{\mathbb{E}|Y - \mathbb{E}Y|^2}{\varepsilon^2} = \frac{\mathrm{Var}(Y)}{\varepsilon^2}.$$

A even more particular case of this is the Chebyshev's Inequality (taking $\varepsilon = k\sqrt{\mathrm{Var}(Y)} = k\sigma$).

(iii) Take $g(x) = e^{\theta x}$ for some $\theta > 0$. We get then

$$\mathbf{P}\left(X(\omega) > \varepsilon\right) \leq e^{-\theta\varepsilon}\mathbb{E}\left[e^{\theta X}\right].$$

This inequality states that the tail of the distribution decays exponentially in ε if X has finite exponential moments. With simple manipulations one can obtain Chernoff's inequality using it.

Lemma 1.36 (Cauchy-Bunyakovski-Schwarz inequality)**.** *If $X, Y \in L^2(\Omega)$, then $XY \in L^1(\Omega)$ and:*

$$|\mathbb{E}[XY]| \leq \mathbb{E}|XY| \leq \|X\|_2\|Y\|_2$$

More general we have:

Lemma 1.37 (Hölder inequality)**.** *If $1/p + 1/q = 1$, $X \in L^p(\Omega)$ and $Y \in L^q(\Omega)$ then $XY \in L^1(\Omega)$ and:*

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

1.9. Convergence of random variables.

Asymptotic behavior is a key issue in probability theory and in the study of the stochastic processes. Why do we even need to look at the asymptotic behavior? Most of the times we cannot work with the perfect variants of the variable under study. Most of the time we will construct an approximation of the random variables (the so called model) thus it is absolutely crucial to study the conditions under which the approximation converges to the real thing. In this section we will explore the varied notions of convergence characteristic to probability theory.

1.9.1. Almost sure (a.s.) convergence. Convergence in probability. The basic notion of convergence from analysis can be translated here as a everywhere convergence. That is a sequence X_n which converges to X everywhere on the Ω or $X_n(\omega) \to X(\omega)$ for all $\omega \in \Omega$. For example take $X_n(\omega) = (1 - 1/n)X(\omega)$. This sequence converges to X for every omega. In general this notion is not very useful. Note that in order to have everywhere convergence we need *everywhere convergence*. It is entirely possible that the sequence X_n will converge for almost all $\omega \in \Omega$ but not for some small subset N. The point is that if this subset N has a very small probability of happening we really do not care about it. The question is how small is the probability of N and that is what differentiate the a.s. convergence from convergence in probability.

Definition 1.38 (a.s. convergence)**.** We say that X_n converges to X almost surely denoted $X_n \xrightarrow{\text{a.s.}} X$, if there exist $N \in \mathscr{F}$ with $\mathbf{P}(N) = 0$ such that $X_n(\omega) \to X(\omega)$ as $n \to \infty$ for all $\omega \in N^c$, (where N^c is a notation for the complement of the set N).

Thus here the set of omega's for which we do not have convergence have to have probability zero. Similarly with the pointwise (everywhere) convergence, the a.s. convergence is invariant with respect to continuous functionals.

Exercise 13. Show that if $f : \mathbb{R} \to \mathbb{R}$ is a continuous function and $X_n \xrightarrow{\text{a.s.}} X$, then $f(X_n) \xrightarrow{\text{a.s.}} f(X)$ as well.

A technical point here is that starting with a sequence of random variables X_n, the limiting variable may not be a random variable itself ($\mathscr{B}(\mathbb{R})$-measurable). To avoid this technical problem if one assumes that the probability space is complete (as defined next) one will always obtain random variables as the limit of random sequences (if the limit exist of course). Throughout this course we will always assume that the probability space we work with is complete.

Definition 1.39 (Complete probability space)**.** We say that the probability space $(\Omega, \mathscr{F}, \mathbf{P})$ is complete if any subset of a probability zero set in $\mathscr{F}$ is also in $\mathscr{F}$. Mathematically: if $N \in \mathscr{F}$ with $\mathbf{P}(N) = 0$, then $\forall M \subset N$ we have $M \in \mathscr{F}$.

We can easily "complete" any probability space $(\Omega, \mathscr{F}, \mathbf{P})$ by adding to its sigma-algebra all the sets of probability zero.

So that was one type of convergence (a.s.). We can make it less restrictive by looking at the measure of N and requiring that this measure instead of being zero all the time to somehow converge to zero. This is the next definition (convergence in probability).

Definition 1.40 (Convergence in probability)**.** We say that X_n converges in probability to X denoted $X_n \xrightarrow{\text{p}} X$, if the sets $N_\varepsilon(n) = \{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\}$ have the property $P(N_\varepsilon(n)) \to 0$ as $n \to \infty$, for any fixed $\varepsilon > 0$.

Theorem 1.41 (Relation between a.s. convergence and convergence in probability)**.** *We have the following relations:*

(1) If $X_n \xrightarrow{a.s.} X$ then $X_n \xrightarrow{p} X$

(2) If $X_n \xrightarrow{p} X$ then there exist a subsequence n_k such that $X_{n_k} \xrightarrow{a.s.} X$ as $k \to \infty$

Proof. (a) Let $N^c = \{\omega : \lim |X_n(\omega) - X(\omega)| = 0\}$. We know form the definition of a.s. convergence that $P(N) = 0$.

Fix an $\varepsilon > 0$ and consider $N_\varepsilon(n) = \{\omega : |X_n(\omega) - X(\omega)| \geq \varepsilon\}$. Let now:

$$M_k = \left(\bigcup_{n \geq k} N_\varepsilon(n) \right)^c = \bigcap_{n \geq k} N_\varepsilon(n)^c \tag{1.9}$$

- M_k's are increasing sets ($M_k = N_\varepsilon(k)^c \cap M_{k+1}$ which implies $M_k \subseteq M_{k+1}$).

- If $\omega \in M_k$ this means that for all $n \geq k$, $\omega \in N_\varepsilon(n)^c$, or $|X_n(\omega) - X(\omega)| < \varepsilon$. By definition this means that the sequence is convergent at ω, therefore $M_k \subseteq N^c$, $\forall k$, thus $\cup M_k \subseteq N^c$.

I leave it as an easy exercise to take an $\omega \in N^c$ and to show that it must exist an k_0 such that $\omega \in M_{k_0}$, therefore we will easily obtain $N^c \subseteq \cup M_k$. This will imply that $\cup M_k = N^c$ and so $\mathbf{P}(\cup M_k) = 1$, by hypothesis.

Since the sets M_k are increasing this implies that $p(M_k) \to 1$ when $k \to \infty$. Looking at the definition of M_k in (1.9) this clearly implies that

$$\mathbf{P}\left(\bigcup_{n \geq k} N_\varepsilon(n) \right) \to 0 \quad , \text{ as } k \to \infty,$$

therefore $\mathbf{P}\left(N_\varepsilon(k)\right) \to 0$, as $k \to \infty$, which is the definition of the convergence in probability.

(b) For this part we will use the Borel-Cantelli lemmas (Lemma 1.17 on page 7). We will take ε in the definition of convergence in probability of the form $\varepsilon_k > 0$ and make it to go to zero when $k \to \infty$. By the definition of convergence in probability for every such ε_k we can find an n_k, such that $\mathbf{P}\{\omega : |X_n(\omega) - X(\omega)| > \varepsilon_k\} < 2^{-k}$, for every $n \geq n_k$. An easy process now will construct $m_k = \min(m_{k-1}, n_k)$ so that the subsequence is now increasing, while still having the above, desired property. Call:

$$N_k = \{\omega : |X_{m_k}(\omega) - X(\omega)| > \varepsilon_k\}.$$

Then from above $\mathbf{P}(N_k) < 2^{-k}$ which implies that $\sum_k \mathbf{P}(N_k) < \sum_k 2^{-k} < \infty$. Then by the first Borel-Cantelli lemma, the probability that N_k occurs infinitely often is zero. This means that with probability one N_k^c eventually. Or, the set of ω for which $\exists k_0$ and $|X_{m_k}(\omega) - X(\omega)| < \varepsilon_k\}$ for all $k \geq k_0$ has probability 1. Or the set $N := \{\omega : X_{m_k}(\omega) \to X(\omega)\}$ has probability $\mathbf{P}(N) = 1$. But this is exactly what we needed to prove. □

In general convergence in probability does not imply a.s. convergence.

Exercise 14 (Counterexample. $\xrightarrow{\text{p}}$ implies $\xrightarrow{\text{a.s.}}$)**.** You can construct your own counterexample. For instance take $\Omega = (0,1)$ with the Borel sets on it and the Lebesque measure (which is a probability measure for this Ω). Take now for every $n \in \mathbb{N}$ and $1 \leq m \leq 2^n$:

$$X_{n,m}(\omega) = \mathbf{1}_{\left[\frac{m-1}{2^n}, \frac{m}{2^n}\right]}(\omega).$$

Form a single subscript sequence by taking: $Y_1 = X_{0,1}$, $Y_2 = X_{1,1}, Y_3 = X_{1,2}$, $Y_4 = X_{2,1}, Y_5 = X_{2,2}, Y_6 = X_{2,3}, Y_7 = X_{2,4}$, etc. Draw these variables on a piece of paper for a better understanding of what is going on.

Prove that this sequence $\{Y_k\}$ has the property that $Y_k \xrightarrow{\text{p}} 0$ but $Y_k \nrightarrow Y$ a.s. In fact it does not converge for any $\omega \in \Omega$.

1.9.2. L^p convergence. Recall that we defined earlier the L^p spaces and the norm in L^p, for $p \geq 1$.

$$\|X\|_p = \sqrt[p]{\mathbb{E}\left[X^p\right]}$$

Definition 1.42. We say that the sequence X_n converges in L^p (or in the p-mean, denoted $X_n \xrightarrow{\text{L}^p(\Omega)} X$ if $X_n, X \in L^p$ and $\|X_n - X\|_p \to 0$ as $n \to \infty$ (or $\mathbb{E}(|X_n - X|^p) \to 0$ with n).

The particular case when $p = 2$ is detailed in your textbook and is called *convergence in quadratic mean*.

These L^p spaces form a complete normed vector space. This is interesting from the real analysis perspective. For our purposes the following is important:

Proposition 1.43. *Let X a random variable. Then the sequence of norms $\|X\|_p$ is non-decreasing (increasing) in p. This means that if a variable is in L^q for some q fixed then it also is in any L^r with $r \leq q$. Therefore we have (as spaces): $L^1(\Omega) \supseteq L^2(\Omega) \supseteq L^3(\Omega) \ldots$.*

PROOF. Let $p_1 > p_2$. Then the function $f(x) = |x|^{p_1/p_2}$ is convex (check this) and we can apply Jensen's inequality to the non-negative r.v. $Y = |X|^{p_2}$. The application immediately yields the desired result. □

Corollary 1.44. If $X_n \xrightarrow{\text{L}^p(\Omega)} X$ and $p \geq q$ then $X_n \xrightarrow{\text{L}^q(\Omega)} X$

PROOF. Exercise. □

Exercise 15. Show that if $X_n \xrightarrow{\text{L}^p(\Omega)} X$ then $\mathbb{E}|X_n|^p \to \mathbb{E}|X|^p$.

HINT: The $\|\cdot\|_p$ is a proper norm (recall the properties of a norm).

Next we will look into relations between the forms of convergence defined thus far.

Proposition 1.45. *If* $X_n \xrightarrow{\mathrm{L}^p(\Omega)} X$ *then* $X_n \xrightarrow{p} X$.

PROOF. This is an easy application of the Markov Inequality (Proposition 1.34). Take $g(x) = |x|^p$, and the random variable $X_n - X$. We obtain:

$$\mathbf{P}\left(|X_n - X|^p > \varepsilon\right) \leq \varepsilon^{-p}\mathbb{E}|X_n - X|^p.$$

Therefore, if $X_n \xrightarrow{\mathrm{L}^p(\Omega)} X$ then we necessarily have $X_n \xrightarrow{\mathrm{p}} X$ as well. □

Exercise 16. The converse of the previous result is not true in general. Consider the probability ensemble of Exercise 14.

Let $X_n(\omega) = n\mathbf{1}_{[0,\frac{1}{n}]}(\omega)$

Show that $X_n \xrightarrow{\mathrm{p}} X$ but $X_n \nrightarrow X$ in any L^p with $p \geq 1$.

What about convergence in L^p compared with convergence a.s.? It turns out that neither imply the other one. It is possible (easy) to come up with counterexamples for a.s. implies p-mean convergence and for p-mean convergence implies convergence a.s. However, what is true is that if both limits exist they must be the same.

Proposition 1.46. *If* $X_n \xrightarrow{\mathrm{L}^p(\Omega)} X$ *and* $X_n \xrightarrow{a.s.} Y$ *then* $X = Y$ *a.s.*

PROOF. (Sketch) We have already proven that both types of convergence imply convergence in probability. The proof then ends by showing a.s. the uniqueness of a limit in probability. □

1.9.3. Weak Convergence or Convergence in Distribution. All of the three modes of convergence discussed thus far are concerned with the case when all the variables X_n as well as their limit X are defined on the same probability space. In most applications the convergence is necessary only from the point of view of the distributions of X_n and X. I am going to stress this fact, though this is the weakest form of convergence in the sense that it is implied by all the others we are in fact discussing a totally different form of convergence.

Definition 1.47 (Convergence in Distribution – Convergence in Law – Weak-Convergence)**.** Consider a sequence of random variables X_n defined on probability spaces $(\Omega_n, \mathscr{F}_n, \mathbf{P}_n)$ (which might be all different) and a random variable X, defined on $(\Omega, \mathscr{F}, \mathbf{P})$. Let $F_n(t)$ and $F(t)$ be the corresponding distribution functions. X_n is said to converge to X in distribution (written $X_n \xrightarrow{\mathcal{D}} X$ or $F_n \Rightarrow F$) if for every point t at which F is continuous we have:

$$\lim F_n(t) = F(t).$$

Remark 1.48. There are many notations which are used interchangeably in various books, we mention $X_n \xrightarrow{\mathcal{L}} X$,$X_n \Rightarrow X$, $X_n \xrightarrow{\text{Distrib.}} X$, $X_n \xrightarrow{\text{d}} X$etc.

Remark 1.49. Why do we require t to be a continuity point of F? The simple answer is that in the discontinuity points weird things may happed even though we might have convergence everywhere else. I will give you a simple example that may illustrate this fact.

Let X_n be a $1/n$Bernoulli$(1/n)$ random variable. That is X_n takes value $1/n$ with probability $1/n$ and value 0 with probability $1 - 1/n$. Then:

$$F_n(t) = \begin{cases} 0 & \text{, if } t < \frac{1}{n} \\ 1 & \text{, if } t \geq \frac{1}{n}. \end{cases}$$

Looking at this it makes sense to say that the limit is $X = 0$ with probability 1 which has distribution function:

$$F(t) = \begin{cases} 0 & \text{, if } t < 0 \\ 1 & \text{, if } t \geq 0. \end{cases}$$

Yet, at the discontinuity point of F we have $F(0) = 1 \neq \lim F_n(0) = 0$. This is why we exclude these points from the definition.

There is one quantity where we do not care about these isolated points and that is the integral. That is why we have an alternate definition for convergence in distribution given by the next theorem. Note that it applies to random vectors X_n, X which are defined on $\mathbb{R}^d$.

Theorem 1.50. *Let X_n defined on probability spaces $(\Omega_n, \mathscr{F}_n, \mathbf{P}_n)$ and X, defined on $(\Omega, \mathscr{F}, \mathbf{P})$. Then $X_n \xrightarrow{\mathcal{D}} X$ if and only if for any bounded, continuous function on the range of X we have:*

$$\mathbb{E}[\phi(X_n)] \to \mathbb{E}[\phi(X)], \qquad \text{as } n \to \infty,$$

or equivalently:

$$\int \phi(t) dF_n(t) \to \int \phi(t) dF(t)$$

The following proposition states that (if possible to express) the convergence in probability will imply convergence in distribution. That is perhaps the reason for the name weak convergence.

Proposition 1.51. *Suppose that the sequence of random variables X_n and the random variable X are defined on the same probability space $(\Omega, \mathscr{F}, \mathbf{P})$. If $X_n \xrightarrow{p} X$ then $X_n \xrightarrow{\mathcal{D}} X$.*

Because of the fact that we are talking about apples and oranges when comparing weak convergence with anything else in general the converse of the previous theorem is false. However, there is one case where the converse is true, and that is when the limit X is a.s. a constant (notice that constants live in any probability space).

Proposition 1.52. *Let $X_n \xrightarrow{\mathcal{D}} X$ and X is a non-random constant (a.s.). Then $X_n \xrightarrow{p} X$.*

Furthermore, here is an interesting result:

Theorem 1.53 (Skorohod's representation theorem)**.** *Suppose $X_n \xrightarrow{\mathcal{D}} X$. There exists a probability space $(\Omega', \mathscr{F}', \mathbf{P}')$ and a sequence of random variables Y, Y_n on this new probability space, such that X_n has the same distribution as Y_n, X has the same distribution as Y, and $Y_n \to Y$ a.s. In other words, there is a representation of X_n and X on a single probability space, where the convergence occurs almost surely.*

Exercise 17. Write a statement explaining why the Skorohod's theorem does not contradict our earlier statement that convergence in distribution does not imply convergence a.s.

Finally, we will finish this section with the two main limit theorems from elementary probability: the law(s) of large numbers and the central limit theorem.

Theorem 1.54 (The Weak law of large numbers)**.** *Let X_n be a sequence of r.v.'s defined on probability spaces $(\Omega_n, \mathscr{F}_n, \mathbf{P}_n)$. Let us use the notations $S_n = X_1 + X_2 + \cdots + X_n$ for the sum and $\overline{X}_n = S_n/n$ for the average of the first n terms.*

Assume that X_n's are independent identically distributed (iid) with mean μ. Then $\overline{X}_n \xrightarrow{p} \mu$.

Note that the previous theorem says that this is equivalent with convergence in distribution, that is the reason for calling this result the weak law. The next result is stronger (it implies the weak law when the prob spaces are the same).

Theorem 1.55 (The Strong law of large numbers)**.** *Let X_n be a sequence of r.v.'s defined on the same probability space $(\Omega, \mathscr{F}, \mathbf{P})$. We will use the same notations from the Weak law.*

Assume that X_n's are independent identically distributed (iid) with mean μ. Then $X_n \to \mu$ a.s.

The next theorem talks about how the convergence to μ occurs.

Theorem 1.56 (The Central Limit Theorem (CLT))**.** *Let X_n be a sequence of r.v.'s defined on probability spaces $(\Omega_n, \mathscr{F}_n, \mathbf{P}_n)$. Assume as before that X_n's are iid and in addition that they have finite variance σ^2. We use the notations presented in the weak law and in addition we define the standardized variables:*

$$Z_n = \frac{\overline{X}_n - \mu}{\sigma/\sqrt{n}} = \frac{S_n - n\mu}{\sigma\sqrt{n}}$$

Let Z be a $N(0,1)$ random variable. Then we have:

$$Z_n \xrightarrow{\mathcal{D}} Z.$$

1.10. Uniform Integrability[4]

We have seen that convergence a.s. and convergence in L^p are generally not compatible. However, we will give next an integrability condition that together with convergence in probability will imply convergence in p-mean.

Definition 1.57 (Uniform Integrability criterion)**.** A collection of random variables $\{X_\alpha\}_{\alpha\in\mathcal{I}}$ is called uniform integrable (U.I.) if:

$$\lim_{M\to\infty} \sup_{\alpha} \mathbb{E}\left[|X_\alpha| \mathbf{1}_{\{|X_\alpha|>M\}}\right] = 0.$$

In other words the tails of the expectation converge to 0 uniformly for all the family.

Theorem 1.58. *If $X_n \xrightarrow{p} X$ and for a fixed $p \geq 1$ the family $\{|X_n|^p\}_{n\in\mathbb{N}}$ in U.I. then $X_n \xrightarrow{p} X$*

For the proof see [GS01, Theorem 7.10.3]
We will give a few more details about U.I.

Example 1.59. Examples of U.I. families:

- Any r.v. $X \in L^1$ is U.I.
  ($\mathbb{E}|X| < \infty$ implies immediately $\mathbb{E}\left[|X|\mathbf{1}_{\{|X|>M\}}\right] \xrightarrow[M\to\infty]{} 0$)
- Let the family X_α bounded by an integrable random variable i.e., $|X_\alpha| \leq Y$ and $Y \in L^1$ then X_α is U.I.
  Indeed, we have $\mathbb{E}\left[|X_\alpha|\mathbf{1}_{\{|X_\alpha|>M\}}\right] \leq \mathbb{E}\left[Y\mathbf{1}_{\{|Y|>M\}}\right]$, which does not depend on α and converges to 0 with M as in the previous example.

[4]Not normally taught in Ma611

- Any finite collection of r.v.'s in L^1 is U.I.
  This is just an application of the previous point. If $\{X_1, X_2, \ldots, X_n\}$ is the collection of integrable r.v.'s take for example $Y = |X_1| + |X_2| + \cdots + |X_n|$.
- The family $\{a_\alpha Y\}$ with $Y \in L^1$ and $a_\alpha \in [-1, 1]$, non-random constants is U.I.
- Any bounded collection of integrable r.v.'s is U.I.

Next we give a very useful criterion for U.I.

Proposition 1.60. *A family of r.v.'s $\{X_\alpha\}_{\alpha\in\mathcal{I}}$ is uniform integrable if $\mathbb{E}f(|X_\alpha|) \le C$ for some finite C and all α, where $f \ge 0$ is any function such that $f(x)/x \to \infty$ as $x \to \infty$.*

Here is an example of a family which is not U.I.

Example 1.61. Let us consider the probability space of all infinite sequences of coin tosses (we will see this space later on in reference to Bernoulli process). Assume that the coin is fair.

Let $X_n = \inf_{i>n}\{\text{toss } i \text{ is a } H\}$, the first toss after n where we obtain a head. Then for any M we can find $n \ge M$ therefore $X_n > n \ge M$ thus $\mathbb{E}\left[|X_n|\mathbf{1}_{\{|X_n|>M\}}\right] = \mathbb{E}[X_n] > n$, implying that X_n is not U.I.

1.11. Exchanging the order of limits and expectations

This is an important question. In many cases we need to put the limit under the integral sign, but are we doing it correctly?

There are 4 results that can help you with this question.

The first two results results basically require the sequence and the limit to be integrable.

Theorem 1.62 (Dominated Convergence)**.** *If there exists a random variable Y such that $\mathbb{E}Y < \infty$, $X_n \le Y$ for all n and if we have $X_n \xrightarrow{p} X$, then $\mathbb{E}X_n \to \mathbb{E}X$ as well.*

In the particular case when Y is non-random we obtain:

Corollary 1.63 (Bounded Convergence)**.** Suppose that $X_n \le C$, $\forall n$ for some finite constant C. If $X_n \xrightarrow{\text{p}} X$, then $\mathbb{E}X_n \to \mathbb{E}X$ as well.

In the case of monotone (increasing) convergence of non-negative r.v.'s we can exchange the limit and the expectation even if X is *non-integrable*.

Theorem 1.64 (Monotone Convergence)**.** *If $X_n \ge 0$ and $X_n(\omega) \uparrow X(\omega)$ a.s. then $\mathbb{E}X_n \uparrow \mathbb{E}X$. This is true even if $X(\omega) = \infty$ for some $\omega \in \Omega$*

Remark 1.65. You may think that as we have increasing convergence we must also have decreasing convergence. We indeed have but the result is not that useful. It requires the extra assumption $\mathbb{E}[X_1] < \infty$. But, if we make this assumption the exchange of limit and integral is true already from the dominated convergence theorem. If we wish to drop the extra assumption the result is no longer true as the next example demonstrates.

Example 1.66. Let Z be a random variable such that $\mathbb{E}Z = \infty$. Take $X_1 = Z$, and in general $X_n(\omega) = n^{-1}Z(\omega)$. Then we have that $\mathbb{E}X_n = \infty$, for any n but $X_n \downarrow 0$ wherever Z is finite.

Practice your understanding solving the following exercise:

Exercise 18. Let Y_n a sequence of non-negative random variables. Use the Monotone Convergence Theorem to show that:

$$\mathbb{E}\left[\sum_{n=1}^{\infty} Y_n\right] = \sum_{n=1}^{\infty} \mathbb{E}[Y_n].$$

Continue by showing that if $X \geq 0$ a.s. and A_n are disjoint sets with $\mathbf{P}(\cup_n A_n) = 1$ (partition of Ω), then:

$$\mathbb{E}[X] = \sum_{n=1}^{\infty} \mathbb{E}(X\mathbf{1}_{A_n}).$$

Furthermore, show that the result applies also when $X \in L^1$.

The last result presented bellow is the most useful in practice; we do not require the sequence or the limit to be integrable nor do we require a special (monotone) form of convergence. We only require the existence of a lower bound. However, the result is restrictive, it only allows exchange of the lim inf with the expectation.

Lemma 1.67 (Fatou's Lemma)**.** *Suppose that X_n is a sequence of random variables such that there exist a $Y \in L^1$ with $X_n > Y$ for all n. Then we have:*

$$\mathbb{E}\left[\liminf_{n\to\infty} X_n\right] \leq \liminf_{n\to\infty} \mathbb{E}[X_n]$$

Here:

$$\liminf_{n\to\infty} X_n = \lim_{n\to\infty} \left\{ \inf_{k\geq n} X_k \right\}.$$