Probability theory

Konstantin Izyurov

Contents

Chapter 1. Measure theoretic foundations. 4
1.1. Motivation: Probability meets measure theory. 4
1.2. Definitions and elementary facts from measure theory 6
1.3. Dynkin's $\pi - \lambda$ theorem and uniqueness of measures. 8
1.4. Caratheodory extension and existence of measures. 9
1.5. Expectation. 13
1.6. Direct products of measure spaces and Fubini's theorem 17
1.7. Infinite products of probability spaces and Kolmogorov extension theorem 20

Chapter 2. Sums of indepndent random variables 25
2.1. Independent events and variables 25
2.2. Gaussian random variables 27
2.3. Weak law of large numbers 28
2.4. Large deviations. 30
2.5. Strong law of large numbers 32
2.6. Kolmogorov's zero-one law 33
2.7. Various notions of convergence of random variables 34
2.8. More about convergence in distribution. 36
2.9. Characteristic functions 37
2.10. Explicit computations with characteristic functions 40
2.11. The Central limit theorem 41
2.12. Heavy tails and stable distributions 42
2.13. Multi-dimensional characteristic functions and Gaussian vectors 43
2.14. Random walks 46
2.15. Local central limit theorem. 48

Chapter 3. Markov chains and the Poisson process 53
3.1. Markov chains: key definitions 53
3.2. Examples of Markov chains 55
3.3. Stationary distributions 56
3.4. Aperiodicity and convergence results 59
3.5. Alternative proof of convergence (optional) 63
3.6. Poisson process 64

Chapter 4. Conditional expectations and martingales 69
4.1. Conditional expectation: motivation and definition 69
4.2. Examples and some properties of conditional expactation 71
4.3. $L^2(\Omega)$ and existence of conditional expectation 73
4.4. Regular conditional distribution 75
4.5. Martingales: simple properties and the optional stopping theorem 77
4.6. Almost sure convergence of supermartingales. 80
4.7. Doob's inequality and convergence in L^p for $p > 1$. 82
4.8. Unifrom integrability and convergence in L^1. 84

4.9. Backward martingales and the strong law of large numbers 87
4.10. Martingale proof of Radon-Nikodym theorem 88

CHAPTER 1

Measure theoretic foundations.

1.1. Motivation: Probability meets measure theory.

Probability theory aims at modeling real-life phenomena that are characterized by uncertainty[1]. This means that the outcome of an experiment (we will speak about "experiments" for simplicity, even though applicability of the theory is by no means limited by a laboratory setup) is not entirely determined by the initial conditions of that experiment. The prototypical examples are the experiments of "throwing a dye" and "flipping a coin": six (respectively, two) possible outcomes occur, and there is no way to predict in advance which one realizes.

Assume that a particular experiment A has a finite set Ω of possible outcomes, e. g. in the case of a dye, one has $\Omega = \{1, 2, 3, 4, 5, 6\}$, and for the coin flip, one has $\Omega = \{\text{heads}, \text{tails}\}$. Then, provided that the experiment can be repeated several times, one can compute frequencies of outcomes. Let $X_i \in \Omega$ denote the outcome of the i-th experiment. Put

$$f_N(\omega) := \frac{\#\{i \leq N : X_i = \omega\}}{N},$$

where N is the number of repetitions. Note that "series of N repetitions of an experiment A" may in itself be viewed as an experiment with an uncertain outcome, in particular, the frequencies $f_N(\omega)$ will be random numbers - meaning that if one performs two series of N repetitions of the experiment A, one will get, in general, different frequencies.

However, if the conditions of the experiment A are repeated precisely enough, and if N is large enough, the frequencies $f_N(\omega)$ are often experimentally observed to be close to certain numbers $p(\omega)$; loosely speaking,

$$f_N(\omega) \overset{N\to\infty}{\longrightarrow} p(\omega).$$

The numbers $p(\omega)$ are determined by the conditions of the experiment only, they are certain, "non-random" quantities. Thus, in the case of a fair dye, one has $p(\omega) = \frac{1}{6}$ for every outcome $\omega \in \Omega$, and for the coin flip, one has $p(\text{heads}) = p(\text{tails}) = \frac{1}{2}$.

Obviously, $f_N(\omega) \geq 0$ and $\sum_{\omega\in\Omega} f_N(\omega) = 1$. We naturally expect these properties to be inherited by the "idealized" frequencies p, which leads to the following definition.

Definition 1.1.1. A discrete probability space is a finite (or countably infinite) set Ω (called the set of outcomes), equipped with a function $p : \Omega \to \mathbb{R}_{\geq 0}$, such that $\sum_{\omega\in\Omega} p(\omega) = 1$. The quantity $p(\omega)$ is called the probability of an outcome ω.

While this definition is good enough for the case of finite (or countably infinite) set of outcomes, it does not cover the case of uncountable sets. Although "uncountable sets of outcomes" may sound fancy from the practical point of view, the following example shows that this setup arises quite naturally, even if we start from the most "finite" and "discrete" probabilistic models.

Example 1.1.2. Alice has 10 euros and Bob has 5 euros. They flip a fair coin, and if it comes up heads, Alice pays Bob 1 euro, otherwise Bob pays Alice 1 euro. They repeat the game until one of them loses all the money. What is the probability for Alice to win?

[1] here we follow what is close to the "propensity interpretation of Probability" to motivate the axioms. For other interpretations, see, e. g., http://plato.stanford.edu/entries/probability-interpret/.

Even before actually computing the probability, one must understand what does it mean, that is, what is the relevant probability space for this Alice-Bob game. If the number of coin flips were fixed in advance - say, N - the construction of probability space would be fairly easy. One would then take $\Omega := \{\text{heads}, \text{tails}\}^N$, that is, the set of all sequences $(\omega_1, \dots, \omega_N)$ of length N with $\omega_i \in \{\text{heads}, \text{tails}\}$, and the probability of each outcome would be 2^{-N}. Unfortunately, this is not enough for our purposes, since the duration of the Alice-Bob game is not fixed in advance. However large N we take, it is still possible that the game has no winner after N flips. In fact, the game might have no winner at all: Alice and Bob could play ad infinitum.

Taking this possibility into account, we take a different view on the Alice-Bob game. We may assume that they continue to flip the coin even after the winner is decided (the results of these flips are, of course, irrelevant). Or, in other words, we imagine that they first flip the coin an infinite number of times, record the resulting sequence of heads and tails, and only then examine this sequence in order to determine the winner. Thus, the natural choice of probability space for the game would be the space $\{\text{heads}, \text{tails}\}^{\mathbb{N}}$ of all infinite sequences $\omega = (\omega_1, \omega_2, \dots)$, with "equal probability assigned to each sequence". This is an uncountably infinite space, and the discrete probability space structure is inadequate here: an infinite amount of equal numbers cannot sum up to 1.

Still, we can meaningfully ascribe non-trivial probabilities to certain *sets of outcomes* - such sets are called *events*. For example, given a finite sequence[2] $\sigma = (\sigma_1, \dots, \sigma_N) \in \{0,1\}^N$, we can consider the set $\Omega_{\sigma_1,\dots,\sigma_N} := \{\omega \in \{0,1\}^{\mathbb{N}} : \omega_1 = \sigma_1, \dots, \omega_N = \sigma_N\}$. This is the set of all infinite sequences whose beginning agrees with the presecribed sequence $(\sigma_1, \dots, \sigma_N)$. In the probabilistic language, this corresponds to the event that the first N coin flips yield that particular sequence. Of course, the probability of this event should be 2^{-N}.

We can, furthermore, compute the probability of events like "Alice wins on move N". Indeed, these are just disjoint unions of events of the form $\Omega_{\sigma_1,\dots,\sigma_N}$, and the interpretation of probabilities in terms of frequencies suggests that if A, B are *disjoint* sets, then we must have the relation

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

A natural next step is to invoke the representation

$$\{\text{Alice wins}\} = \sqcup_{N=1}^{\infty} \{\text{Alice wins on move } N\}. \tag{1.1.1}$$

So, if we want the axiomatics of Probability to be rich enough to treat the Alice-Bob game, it should say something about the behavior of probabilities under *countable* disjoint unions. As Kolmogorov realised in 1930-s, measure theory provides exactly the right framework for Probability, as it studies functions of *sets* that are countably additive - exactly what is needed to pass from (1.1.1) to

$$\mathbb{P}\{\text{Alice wins}\} = \sum_{N=1}^{\infty} \mathbb{P}\{\text{Alice wins on move } N\}.$$

In this particular example, there is a way to further exhibit a direct relevance of measure theory - in fact, of the Lebesgue measure on the unit interval. Namely, given a finite sequence $\omega \in \{0,1\}^N$, or an infinite sequence $\omega \in \{0,1\}^{\mathbb{N}}$, define

$$R(\omega) := \sum_{i=1}^{N} \omega_i 2^{-i},$$

where $N = \infty$ for the case of an infinite sequence. This is just the real number in the interval $[0,1]$ whose binary representation consists of the digits in the sequence ω. Then, it is easy to see that $R(\Omega_{\sigma_1,\dots,\sigma_N}) = [R(\sigma), R(\sigma) + 2^{-N}]$, that is, the probability of an event $\Omega_{\sigma_1,\dots,\sigma_N}$ exactly equals the length of $R(\Omega_{\sigma_1\dots\sigma_N})$.

The mapping R is "almost a bijection" - the set of points with more than one preimage has Lebesgue measure 0 (Exercise!). Therefore, we can assign probability to more general sets - such as, e. g. "Alice beats Bob" - by putting

$$\mathbb{P}(A) := |R(A)|,$$

[2]here we switch the notation from $\{\text{heads}, \text{tails}\}$ to $\{0,1\}$

where $|\cdot|$ denotes the Lebesgue measure, provided that $R(A)$ is Lebesgue measurable. This way, probability is defined for a wide class of events - it is in fact a *measure* on $\{0,1\}^{\mathbb{N}}$.

1.2. Definitions and elementary facts from measure theory

We begin by recalling necessary definitions of measure theory. In what follows, 2^Ω denotes the set of all subsets of a set Ω.

DEFINITION 1.2.1. Suppose Ω is a set, and $\mathcal{F} \subset 2^\Omega$ a collection of its subsets. The collection $\mathcal{F}$ is called *a σ-algebra* if the following conditions are satisfied:

- $\emptyset \in \mathcal{F}$;
- if $A \in \mathcal{F}$, then $A^c := \Omega \setminus A \in \mathcal{F}$;
- if $A_1, A_2, \dots$ is a sequence of subsets of Ω such that $A_i \in \mathcal{F}$ for all i, then $\cup_{i=1}^\infty A_i \in \mathcal{F}$.

Since we could take $A_n = A_{n+1} = \dots = \emptyset$, the third condition is also satisfied for finite sequences of A_i. Also, by the first two conditions, $\Omega \in \mathcal{F}$, and by the second and third conditions, a σ-algebra is closed under countable intersectons: if $A_i \in \mathcal{F}$ for all i, then

$$\cap_{i=1}^\infty A_i = (\cup_{i=1}^\infty A_i^c)^c \in \mathcal{F}.$$

DEFINITION 1.2.2. A set equipped with a σ-algebra is called a *measurable space*, and the elements of the σ-algebra are called *measurable sets*.

DEFINITION 1.2.3. Given a measurable space $(\Omega, \mathcal{F})$, a function $\mu : \mathcal{F} \to \mathbb{R}_{\geq 0} \cup \{+\infty\}$ is called *a measure* if it satisfies the following properties:

- $\mu(\emptyset) = 0$;
- (σ-additivity or countable additivity) If $A_1, A_2, \dots$ is a sequence of disjoint sets (that is, $A_i \cap A_j = \emptyset$ for $i \neq j$), such that $A_i \in \mathcal{F}$ for all i, then

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

A measure μ is called *a probability measure* if $\mu(\Omega) = 1$.

The second condition in the above definition is equivalent to *finite additivity* (that is, $\mu(A \cup B) = \mu(A) + \mu(B)$ for disjoint measurable sets A, B) combined with *lower continuity*:

if $A_1 \subset A_2 \subset \dots$ are measurable sets, then

$$\mu(\cup_{i=1}^\infty A_i) = \lim_{i \to \infty} \mu(A_i),$$

In the case $\mu(\Omega) < \infty$, σ-additivity is also equivalent to finite additivity combined with *upper continuity*:

if $A_1 \supset A_2 \supset \dots$ are measurable sets, then

$$\mu(\cap_{i=1}^\infty A_i) = \lim_{i \to \infty} \mu(A_i).$$

A fundamental property of measures and σ-algebras is that they can be pushed forward by maps:

LEMMA 1.2.4. *If $(\Omega_1; \mathcal{F}_1)$ is a measurable space, and $f : \Omega_1 \to \Omega_2$ is a map, then the set*

$$\mathcal{F}_1 \circ f^{-1} := \{A \in 2^{\Omega_2} : f^{-1}(A) \in \mathcal{F}_1\}$$

is a σ-algebra on Ω_2. If, moreover, μ is a measure on $\mathcal{F}_1$, then $\mu \circ f^{-1}$, defined by

$$\mu \circ f^{-1}(A) = \mu(f^{-1}(A)),$$

is a measure on $\mathcal{F}_2$.

PROOF. This follows from the identities $f^{-1}(\emptyset) = \emptyset$, $f^{-1}(A^c) = \big(f^{-1}(A)\big)^c$, and $f^{-1}(\sqcup_{i=1}^\infty A_i) = \sqcup_{i=1}^\infty f^{-1}(A_i)$. □

DEFINITION 1.2.5. A map $f : \Omega_1 \to \Omega_2$ between two measurable spaces $(\Omega_1, \mathcal{F}_1)$ and $(\Omega_2, \mathcal{F}_2)$ is called *measurable* if $\mathcal{F}_2 \subset \mathcal{F}_1 \circ f^{-1}$, or, in other words, the preimage of any measurable set is measurable. When we want to be more explicit, we call it $\mathcal{F}_1$-to-$\mathcal{F}_2$ measurable.

How to construct σ-algebras? First, there is always the maximal one, i .e. the set 2^Ω of all subsets of Ω, and the minimal one, containing just two elements: $\emptyset$ and Ω. Second, an intersection of an arbitrary collection of σ-algebras is also a σ-algebra. Therefore, given an arbitrary collection $\mathcal{A}$ of subsets of Ω, we may define the smallest σ-algebra containing $\mathcal{A}$, denoted by $\sigma(\mathcal{A})$. This is just the intersection of all σ-algebras containing $\mathcal{A}$. When Ω bears a structure of a topological space, one may take $\mathcal{A}$ to be the set of all open subsets of Ω; the result is called *the Borel σ-algebra,* and denoted by $\mathcal{B}(\Omega)$. In the case $\Omega = \mathbb{R}$, the Borel σ-algebra coincides with $\sigma(\{(-\infty, a] : a \in \mathbb{R}\})$.

We are ready to give the key definitions of Probability:

DEFINITION 1.2.6. A *probability space* is a triple $(\Omega, \mathcal{F}, \mathbb{P})$, where Ω is a set, $\mathcal{F}$ is a σ-algebra on Ω, and $\mathbb{P}$ is a probability measure on $\mathbb{P}$.

REMARK 1.2.7. A discrete probability space is can be seen as a special case of a probability space, by taking $\mathcal{F} := 2^\Omega$, and $\mathbb{P}(A) := \sum_{\omega \in A} p(\omega)$.

DEFINITION 1.2.8. Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a measurable map from Ω to a measurable space $(\Omega', \mathcal{F}')$ is called a *random variable (with values in Ω').*

Usually, when speaking about random variables, the σ-algebra $\mathcal{F}'$ is not mentioned explicitly. If Ω' is a topological space, we assume $\mathcal{F}' = \mathcal{B}(\Omega)$ unless stated otherwise. In particular, if $\Omega' = \mathbb{R}$, then a random variable is a function $f : \Omega \to \mathbb{R}$ such that $f^{-1}((-\infty, a])$ is measurable for any $a \in \mathbb{R}$.

DEFINITION 1.2.9. If f is a random variable with values in Ω', then the measure on $\mathcal{F}'$ given by

$$\mu(A) := \mathbb{P}(f^{-1}(A)).$$

is called a *distribution of the random variable f.*

REMARK 1.2.10. One can write $\mathbb{P}(f^{-1}(A)) = \mathbb{P}(\{\omega \in \Omega : f(\omega) \in A\})$. It is customary in Probability texts to use capital latin letters for random variables (e. g., X instead of f) and abbreviate the last formula to something like $\mathbb{P}(X \in A)$.

A distribution of a scalar random variable X is thus a probability measure on $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. A function

$$F_X(a) = \mathbb{P}(X \in (-\infty, a])$$

is called the *probability distribution function*[3] of X.

LEMMA 1.2.11. *For any random variable X, $F_X(a)$ is non-decreasing, right-continuous (i. e., $F_X(a_i) \to F_X(a)$ whenever $a_i \searrow a$, and has limits $\lim_{a \to +\infty} F_X(a) = 1$, $\lim_{a \to -\infty} F_X(a) = 0$. Conversely, if F is any function with these properties, then there exists a probability measure μ on $\mathcal{B}(\mathbb{R})$ such that $F(a) = \mu((-\infty, a])$.*

PROOF. The properties of F_X follows from monotonicity and upper/lower continuity of measure; e. g., $\cap_{a_i \searrow a}(-\infty, a_i] = (-\infty, a]$ implies the right continuity.

The existence of measure μ is based on existence of Lebesgue measure[4]. Assume first that F is continuous and strictly increasing. Then it has an inverse $h := F^{-1} : (0, 1) \to \mathbb{R}$ which is also continuous and increasing, and $h^{-1}((-\infty; a]) = (0; F(a)]$. The pushforward of the Lebesgue measure by h is the desired measure on $\mathbb{R}$.

In the general case, define, for $x \in (0, 1)$,

$$h(x) := \inf\{y \in \mathbb{R} : F(y) \geq x\}.$$

[3] *or cumulative disctribution function*

[4] to be proved in Section 1.4

Since $F(a) \to 0$ as $a \to -\infty$ and $F(a) \to 1$ as $a \to +\infty$, this infimum is finite for all $x \in (0,1)$. Note that if $y_i \searrow y$ and $F(y_i) \geq x$, then, by right-continuity, $F(y) \geq x$. This means that the infimum is in fact a minimum, i. e. $F(h(x)) \geq x$. In plain words, $h(x)$ is determined by the following rules:

- if x has preimages, then we take $h(x)$ to be the left-most of them;
- if x has no preimages, then there exists a unique $y \in \mathbb{R}$ such that $\lim_{u \to y-} F(u) < x < F(y)$. In this case, we take $h(x) = y$.

Clealry, h is non-decreasing, therefore, $h^{-1}((-\infty, a])$ is an interval for each a. In particular, h is Borel-to-Borel measurable and we can define a measure μ on $\mathbb{R}$ by setting $\mu(A) := |h^{-1}(A)|$.

It remains to check that for all $a \in \mathbb{R}$, $h^{-1}(-\infty, a] = (0; F(a)]$. First,

$$h(F(a)) = \inf\{y \in \mathbb{R} : F(y) \geq F(a)\} \leq a,$$

because a belongs to the set $\{y \in \mathbb{R} : F(y) \geq F(a)\}$. Since h is increasing, this means that

$$h((0, F(a)] \subset (-\infty, a],$$

so $(0, F(a)] \subset h^{-1}(-\infty, a]$. To prove the opposite inclusion, pick $x > F(a)$; as noted above, $F(h(x)) \geq x > F(a)$, which means that $h(x) > a$, that is, $x \notin h^{-1}(0; a]$. □

One simple, but important operation with probability spaces is that of *restriction*: if $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, and $A \in \mathcal{F}$, then $\mathcal{F}|_A := \{A \cap B : B \in \mathcal{F}\}$ is a σ-algebra on A, and $\mu(B) := \mathbb{P}(A \cap B)$ is a measure on $\mathcal{F}|_A$. If $\mathbb{P}(A) = 0$, this measure identically zero; otherwise it could be normalized to be a probability measure on A. This probability measure is called *conditional probability,* and denoted by $\mathbb{P}(\cdot|A)$. So, by definition,

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

Exchanging A and B in the above definition, one arrives at *Bayes's formula:*

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

1.3. Dynkin's $\pi - \lambda$ theorem and uniqueness of measures.

In this section, we prove that if two probability measures on $\mathcal{B}(\mathbb{R})$ have the same probability distribution function - i. e., agree on all rays $(-\infty, a]$ - then they agree on $\mathcal{B}(\mathbb{R})$. It is very tempting to call this result "obvious": if two measures agree on a collection $\mathcal{A}$ of sets, surely they must agree on $\sigma(\mathcal{A})$! Unfortunately, this implication is badly wrong, as one of the exercises shows. We will develop an abstract tool that is useful to treat this and many other questions of the same flavour.

DEFINITION 1.3.1. A collection $\mathcal{A}$ of subsets of a set Ω is called a π-system if it is closed under intersections:

$$A, B \in \mathcal{A} \Rightarrow A \cap B \in \mathcal{A}.$$

DEFINITION 1.3.2. A collection $\mathcal{A}$ of subsets of a set Ω is called a λ-system[5] if

- $\Omega \in \mathcal{A}$;
- if $A, B \in \mathcal{A}$ and $A \subset B$, then $B \setminus A \in \mathcal{A}$;
- if $A_1 \subset A_2 \subset ...$ all belong to $\mathcal{A}$, then $\cup_{i=1}^{\infty} A_i \in \mathcal{A}$.

The λ-systems are a slight generalizations of σ-algebras. Indeed, every σ-algebra is a λ-system (because $B \setminus A = B \cap (\Omega \setminus A)$). On the other hand, the first two properties in the definition of a λ-system imply the first two properties in the definition of σ-algebra. Therefore, a λ-system that is closed under finite unions is a σ-algebra. By passing to complements, we see that a λ-system that is closed under finite intersections is also a σ-algebra.

[5]sometimes also a d-system, of Dynkin system.

EXAMPLE 1.3.3. Assume that μ and ν are two probability measures on the same σ-algebra $\mathcal{F}$. Then $\{A \in \mathcal{F} : \mu(A) = \nu(A)\}$ is a λ-system.

THEOREM 1.3.4. *(π-λ theorem of Dynkin's lemma) Let $\mathcal{A}$ be a π-system and $\mathcal{B}$ be a λ-system. Then*

$$\mathcal{A} \subset \mathcal{B} \Rightarrow \sigma(\mathcal{A}) \subset \mathcal{B}.$$

PROOF. Let $\lambda(\mathcal{A})$ denote the intersection of all λ-systems that contain $\mathcal{A}$; this is also a λ-system. Our goal is to prove that $\lambda(\mathcal{A})$ is closed under intersections. Denote $S(A) := \{B : A \cap B \in \lambda(\mathcal{A})\}$, and let us check that if $A \in \lambda(\mathcal{A})$, then $S(A)$ is a λ-system. Indeed, $\Omega \cap A = A \in \lambda(\mathcal{A})$. If $C \subset D$ and $C, D \in S(A)$, then

$$(D \setminus C) \cap A = (D \cap A) \setminus (C \cap A).$$

Both $(D \cap A)$ and $(C \cap A)$ belong to $\lambda(A)$, and since $\lambda(A)$ is a λ-system, so does their difference. Finally,

$$A \cap (\cup_{i=1}^{\infty} A_i) = \cup_{i=1}^{\infty} (A \cap A_i),$$

which shows that $\cup_{i=1}^{\infty} A_i$ belongs to $S(A)$ whenever all A_i do.

Now assume that $A \in \mathcal{A}$. Since $\mathcal{A}$ is a π-system, $\mathcal{A} \subset S(A)$. Since $S(A)$ is a λ-system, we have $\lambda(\mathcal{A}) \subset S(A)$. So, we have proved the following: if $A \in \mathcal{A}$ and $B \in \lambda(\mathcal{A})$, then $A \cap B \in \lambda(\mathcal{A})$. This means that $S(B)$ contains $\mathcal{A}$, and since it is a λ-system, it contains $\lambda(\mathcal{A})$. In other words, $\lambda(\mathcal{A})$ is closed under intersections.

Consequently, $\lambda(\mathcal{A})$ is a σ-algebra that contains $\mathcal{A}$, therefore $\sigma(\mathcal{A}) \subset \lambda(\mathcal{A}) \subset \mathcal{B}$. □

Collecting all the facts of this subsection together:

COROLLARY 1.3.5. *If two measures μ_1 and μ_2 such that $\mu_1(\Omega) = \mu_2(\Omega) < \infty$ agree on a π-system $\mathcal{A}$, they agree on $\sigma(\mathcal{A})$. In particular, a probability measure on $\mathbb{R}$ is uniquely determined by its p. d. f.*

PROOF. It only remains to notice that $\{(-\infty, a] : a \in \mathbb{R}\}$ is a π -system. □

1.4. Caratheodory extension and existence of measures.

In this section, we review a powerful tool to construct measures: the Caratheorody's extension theorem. This theorem is used to construct:

- Lebesgue measure on $\mathbb{R}$;
- Direct products of measure spaces;
- Non-direct products (projective limits) of measure spaces (Kolmogorov's extension theorem).

DEFINITION 1.4.1. A collection $\mathcal{R} \subset 2^{\Omega}$ of subsets of a set Ω is called a *semi-ring* if the following conditions are satisfied:

- $\emptyset \in \mathcal{R}$;
- it is a π-system, i. e., closed under intersections;
- if $A, B \in \mathcal{R}$, then there exists a finite collection of disjoint sets $A_1, \ldots, A_n \in \mathcal{R}$ such that

$$A \setminus B = \sqcup_{i=1}^{n} A_i.$$

EXAMPLE 1.4.2. The collection $\mathcal{I} := \{[a; b) : a, b \in \mathbb{R}\}$ is a semi-ring.

DEFINITION 1.4.3. We say that a function $\mu : \mathcal{R} \to \mathbb{R}_{\geq 0}$, is a pre-measure [6] if $\mu(\emptyset) = 0$, and the identity $A = \sqcup_{i=1}^{\infty} A_i$, where $A, A_i \in \mathcal{R}$, implies $\mu(A) = \sum_{i=1}^{\infty} \mu(A_i)$.

REMARK 1.4.4. Let μ be a *finitely additive* function on a semi-ring $\mathcal{R}$, that is, if $A, A_1, \ldots, A_N \in \mathcal{R}$ and $A = \sqcup_{i=1}^{N} A_i$, then $\mu(A) = \sum_{i=1}^{N} \mu(A_i)$. Assume that $A_1, \ldots, A_N \in \mathcal{R}$, $B_1, \ldots, B_M \in \mathcal{R}$, and$\sqcup_{i=1}^{N} A_i = \sqcup_{i=1}^{M} B_i$. Then

$$\sum_{i=1}^{N} \mu(A_i) = \sum_{i=1}^{N} \mu(\sqcup_{j=1}^{M} (A_i \cap B_j)) = \sum_{i=1}^{N} \sum_{j=1}^{M} \mu(A_i \cap B_j)) = \sum_{j=1}^{M} \mu(B_j).$$

[6] The conditions are the same as in the definition of measure. However, the term "measure" is reserved for functions defined on a σ-algebra.

This means that we can extend the function μ in a consistent way to *finite unions* of elements of R. Below we use this extension; e. g., for $A, B \in \mathcal{R}$, we write $\mu(A \setminus B)$ instead of $\sum_{i=1}^{n} \mu(A_i)$.

THEOREM 1.4.5. *(Caratheodory extension theorem). Let $\mathcal{R}$ be a semi-ring on a set Ω, and let $\mu : \mathcal{R} \to \mathbb{R}_{\geq 0}$ be a pre-measure. Then there exists a measure on $\sigma(\mathcal{R})$ that coincides with μ on $\mathcal{R}$.*

PROOF. The first step is to define the *outer measure* μ^* associated with μ. Namely, given $A \in 2^{\Omega}$, define

$$\mu^*(A) := \inf_{\substack{\cup_{i=1}^{\infty} A_i \supset A \\ A_i \in \mathcal{R}}} \sum_{i=1}^{\infty} \mu(A_i);$$

in words, the infimum is taken over all countable covers of A by elements of $\mathcal{R}$. The outer measure is defined for any subset of Ω. However, for it to be a measure, we need to restrict it to a smaller class of subsets.

The second step, the most ingenious one in the proof, is to describe that class explicitly. Namely, we say that a subset A of Ω is *well-splitting* if, for any $E \in 2^{\Omega}$, one has

$$\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \setminus A). \tag{1.4.1}$$

The rest of the proof boils down to a (rather straightforward) check of the following facts:

- sets in $\mathcal{R}$ are well-splitting;
- well-splitting sets form a σ-algebra;
- μ^*, when restricted to well-splitting sets, is a measure;
- if $A \in \mathcal{R}$, then $\mu^*(A) = \mu(A)$.

Before proceeding to the proof, we note that μ^* *is countably subadditive*:

$$\mu^*(\cup_{i=1}^{\infty} A_i) \leq \sum_{i=1}^{\infty} \mu^*(A_i). \tag{1.4.2}$$

To see this, fix $\varepsilon > 0$, and let, for each i, $\{A_{i,j}\}_{j=1}^{\infty}$ be a collection of elements of $\mathcal{R}$ that covers[7] A_i (that is, $A_i \subset \cup_{j=1}^{\infty} A_{i,j}$) such that $\sum_{j=1}^{\infty} \mu(A_{i,j}) \leq \mu^*(A_i) + \frac{\varepsilon}{2^i}$. Then the collection $\{A_{i,j}\}_{i,j=1}^{\infty}$ covers $\cup_{i=1}^{\infty} A_i$. Therefore,

$$\mu^*(\cup_{i=1}^{\infty} A_i) \leq \sum_{i,j=1}^{\infty} \mu(A_{i,j}) \leq \sum_{i=1}^{\infty} \mu^*(A_i) + \varepsilon \sum_{i=1}^{\infty} 2^{-i} = \sum_{i=1}^{\infty} \mu^*(A_i) + \varepsilon,$$

and (1.4.2) follows by letting $\varepsilon \to 0$. In view of the sub-additivity, the inequality

$$\mu^*(E) \geq \mu^*(E \cap A) + \mu^*(E \setminus A) \tag{1.4.3}$$

is sufficient for A to be well-splitting.

Sets in $\mathcal{R}$ are well-splitting. Suppose $E \subset \Omega$ and $A \in \mathcal{R}$. Let $\{A_i\}_{i=1}^{\infty}$, $A_i \in \mathcal{R}$ be a cover of E with $\sum_{i=1}^{\infty} \mu(A_i) \leq \mu^*(E) + \varepsilon$. We can decompose

$$A_i = (A_i \cap A) \sqcup (A_i \setminus A) \tag{1.4.4}$$

where all $A_{i,j} \in \mathcal{R}$; note that also $A_i \cap A \in \mathcal{R}$. Since μ is countably additive, we have[8]

$$\mu(A_i) = \mu(A_i \cap A) + \mu(A_i \setminus A), \tag{1.4.5}$$

or, after summing over i,

$$\mu^*(E) + \varepsilon \geq \sum_{i=1}^{\infty} \mu(A_i) = \sum_{i=1}^{\infty} \mu(A_i \cap A) + \sum_{i=1}^{\infty} \mu(A_i \setminus A) \geq \mu^*(E \cap A) + \mu^*(E \setminus A).$$

The last inequality follows from the fact that $\{A \cap A_i\}_{i=1}^{\infty}$ (respectively, $\{A_i \setminus A\}_{i=1}^{\infty}$) form a cover of $E \cap A$ (respectilvey, $E \setminus A$).

---

[7] If A_i is not covered by any countable collection of elements of $\mathcal{R}$, then $\mu^*(A_i) = +\infty$ by definition, and (1.4.2) is obviously true.

[8] note that in general; $A_i \setminus A \notin \mathcal{R}$. Here we use for the first time the convention of Remark 1.4.4.

If $A \in \mathcal{R}$, then $\mu(A) = \mu^*(A)$. Clearly, $\mu^*(A) \leq \mu(A)$, since A covers itself. If $\{A_i\}_{i=1}^\infty$ is any cover of A use the decomposition (1.4.4) and sum the identity (1.4.5) over i. By countable additivity of μ, we get

$$\sum_{i=1}^\infty \mu(A_i) = \sum_{i=1}^\infty \mu(A \cap A_i) + \sum_{i=1}^\infty \mu(A_i \setminus A) \geq \sum_{i=1}^\infty \mu(A \cap A_i).$$

Now, define

$$\tilde{A}_n := (A \cap A_n) \setminus \cup_{i=1}^{n-1}(A \cap A_i).$$

It is not hard to see that $\tilde{A}_n$ is a finite disjoint union of elements in $\mathcal{R}$, and that $\mu(\tilde{A}_n) \leq \mu(A \cap A_n)$. Moreover, $\tilde{A}_n$ are disjoint, and $\cup_{i=1}^\infty \tilde{A}_n = A$. This shows that

$$\sum_{i=1}^\infty \mu(A_i) \geq \sum_{i=1}^\infty \mu(A \cap A_i) \geq \sum_{i=1}^\infty \mu(\tilde{A}_n) = \mu(A)$$

Taking infimum over all covers gives $\mu^*(A) \geq \mu(A)$.

Well-splitting sets form a σ-algebra. It is obvious that Ω is well-splitting, and that a complement of a well-splitting set is well-splitting. Let us check that if A, B are well-splitting, then so is $A \cup B$. Indeed,

$$\mu^*(E) = \mu^*(E \cap A) + \mu^*(E \cap A^c) = \\ \mu^*(E \cap A \cap B) + \mu^*(E \cap A \cap B^c) + \mu^*(E \cap A^c \cap B) + \mu^*(E \cap A^c \cap B^c),$$

where in the first equality we used that A is well-splitting, and in the second one we used that B is well-splitting twice. Now, note that $A \cup B = (A \cap B) \cup (A \cap B^c) \cup (A^c \cap B)$. By the sub-additivity of μ^*, we can write

$$\mu^*(E \cap (A \cup B)) \leq \mu^*(E \cap A \cap B) + \mu^*(E \cap A \cap B^c) + \mu^*(E \cap A^c \cap B).$$

Plugging this into the above equality yields

$$\mu^*(E) \geq \mu^*(E \cap (A \cup B)) + \mu^*(E \cap A^c \cap B^c),$$

and since $A^c \cap B^c = (A \cup B)^c$, this means that $A \cup B$ is well-splitting.

It remains to check that if $A_1, A_2 \dots$ are well-splitting, then so is $A := \cup_{i=1}^\infty A_i$. We may assume without loss of generality that A_i are disjoint. Let $E \subset 2^\Omega$; we may assume $\mu^*(E) < \infty$, for otherwise (1.4.3) is vacuous. Let $A^{(N)} := \cup_{i=1}^N A_i$. Since $A^{(N)}$ is well-splitting, we can write

$$\mu^*(E) = \mu^*(E \cap A^{(N)}) + \mu^*(E \setminus A^{(N)}) \geq \mu^*(E \cap A^{(N)}) + \mu^*(E \setminus A).$$

Once again, since $A^{(N)}$ is well-splitting, we can write

$$\mu^*(E \cap A) = \mu^*(E \cap A \cap A^{(N)}) + \mu^*(E \cap A \setminus A^{(N)}) = \mu^*(E \cap A^{(N)}) + \mu^*(\cup_{i=N+1}^\infty (E \cap A_i)),$$

so, combining the last two formulas,

$$\mu^*(E) \geq \mu^*(E \cap A) + \mu^*(E \setminus A) - \mu^*(\cup_{i=N+1}^\infty (E \cap A_i)).$$

Since all A_i are well-splitting, $\sum_{i=1}^N \mu^*(E \cap A_i) = \mu^*(E \cap A^{(N)}) \leq \mu^*(E) < \infty$, therefore, the series $\sum_{i=1}^\infty \mu^*(E \cap A_i)$ converges. So, by sub-additivity, $\mu^*(\cup_{i=N+1}^\infty (E \cap A_i)) \leq \sum_{i=N+1}^\infty \mu^*(E \cap A_i) \to 0$ as $N \to \infty$.

When restricted to well-splitting sets, μ^* is a measure. Clearly, $\mu^*(\emptyset) = 0$. Let $A_1, A_2, \dots$ be disjoint well-splitting sets; then $\sum_{i=1}^N \mu^*(A_i) = \mu^*(\sqcup_{i=1}^N A_i) \leq \mu^*(\sqcup_{i=1}^\infty A_i)$ for all N, so, we have

$$\mu^*(\sqcup_{i=1}^\infty A_i) \geq \sum_{i=1}^\infty \mu^*(A_i).$$

The opposite inequality follows from subadditivity. □

PROPOSITION 1.4.6. *(Caratheodory extension theorem - uniqueness) If, in the conditions of the previous theorem, Ω can be written as a countable union of sets in $\mathcal{R}$, then the extension of μ to $\sigma(\mathcal{R})$ is unique.*

PROOF. Let $\Omega = \cup_{i=1}^{\infty} A_i$; where $A_i \in \mathcal{R}$; we may assume that they are disjoint. Suppose $\tilde{\mu}$ is another measure on $\sigma(\mathcal{R})$ that agrees with μ on $\mathcal{R}$. Let $E \in \sigma(\mathcal{R})$, By monotonicity of measures, any extension $\tilde{\mu}$ of μ must satisfy $\tilde{\mu}(E \cap A_i) \leq \mu^*(E \cap A_i) = \mu(E \cap A_i)$ and $\tilde{\mu}(A_i \setminus E) \leq \mu^*(A_i \setminus E) = \mu(E \cap A_i)$. Then,

$$\tilde{\mu}(A_i) = \tilde{\mu}(E \cap A_i) + \tilde{\mu}(A_i \setminus E) \leq \mu(E \cap A_i) + \mu(A_i \setminus E) = \mu(A_i),$$

and since $\mu(A_i) = \tilde{\mu}(A_i)$, all the inequalities are in fact equalities, i. e., $\mu(E \cap A_i) = \mu$. Consequently,

$$\tilde{\mu}(E) = \sum_{i=1}^{\infty} \tilde{\mu}(E \cap A_i) = \sum_{i=1}^{\infty} \mu^*(E \cap A_i) = \mu^*(E).$$

□

REMARK 1.4.7. Another proof of Proposition 1.4.6, not referring to the outer measure machinery, is based on the $\pi - \lambda$ theorem (Thm 1.3.4).

EXAMPLE 1.4.8. (Lebesgue measure.) There exists a unique measure λ on $\mathbb{R}$ such that for any $a < b$, $\lambda([a, b)) = b - a$.

PROOF. We apply Theorem 1.4.5 and Proposition 1.4.6 to the semi-ring $\{[a; b) : a, b \in \mathbb{R}\}$ with $\mu([a; b)) := b - a$. All we have to check is that if $I = \sqcup_{i=1}^{\infty} I_i$, then

$$\mu(I) = \sum_{i=1}^{\infty} \mu(I_i). \tag{1.4.6}$$

An elementary input is the finite additivity of μ: if $I = \sqcup_{i=1}^{N} I_i$, then

$$\mu(I) = \sum_{i=1}^{N} \mu(I_i). \tag{1.4.7}$$

To prove (1.4.7), let $I_i = [a_i; b_i)$ be numbered in the increasing order of their leftmost points[9], i. e. $a_1 \leq \cdots \leq a_N$. Since $I_i = [a_i; b_i)$ and $I_{i+1} = [a_{i+1}; b_{i+1})$ are disjoint, we must have $a_{i+1} \geq b_i$, that is, $a \leq a_1 \leq b_1 \leq a_2 \leq \cdots \leq b_N \leq b$. Since $I = \cup_{i=1}^{N} I_i$, we must have, in fact, $a = a_1$,$b_1 = a_2, \ldots, b_N = b$. So,

$$\sum_{i=1}^{N} \mu(I_i) = b_1 - a_1 + b_2 - a_2 + \cdots + b_N - a_N = b_N - a_1 = b - a = \mu(I).$$

This, in particular, implies that $\sum_{i=1}^{N} \mu(I_i) \leq \mu(I)$, and, by passing to the limit, $\sum_{i=1}^{\infty} \mu(I_i) \leq \mu(I)$. To prove the opposite inequality, assume by contradiction that $\varepsilon = \mu(I) - \sum_{i=1}^{\infty} \mu(I_i) > 0$. Then, each $I^{(N)} = I^{(N)} := I \setminus \left(\cup_{i=1}^{N} I_i\right)$ is a non-empty disjoint union of intervals. We can find a finite union $K^{(N)} \subset I^{(N)}$ of *compact* intervals such that[10] $\mu(I^{(N)}) - \mu(K^{(N)}) < \frac{\varepsilon}{2^{N+1}}$. Then,

$$\mu(I^{(N)} \setminus \cap_{i=1}^{N} K^{(i)}) = \mu\left(\cup_{i=1}^{N} (I^{(N)} \setminus K^{(i)})\right) \leq \sum_{i=1}^{N} \mu(I^{(N)} \setminus K^{(i)}) \leq \sum_{i=1}^{N} \mu(I^{(i)} \setminus K^{(i)}) \leq \sum_{i=1}^{N} \frac{\varepsilon}{2^{N+1}} \leq \frac{\varepsilon}{2}.$$

Since $\mu(I^{(N)}) \geq \varepsilon$, this implies that $\mu(\cap_{i=1}^{N} K^{(i)}) \geq \frac{\varepsilon}{2}$, in particular, it is non-empty. The intersection of a sequence of non-empty nested compact sets is non-empty, therefore, $I \setminus \cup_{i=1}^{\infty} I_i = \cap_{i=1}^{\infty} I^{(N)} \supset \cap_{i=1}^{\infty} K^{(N)} \neq \emptyset$, a contradiction. □

[9] observe that an attempt to prove (1.4.6) in a similar way would fail at this point. This is what makes (1.4.6) non-elementary as compared to (1.4.7).

[10] recall that we extend μ to finite unions of intervals according to Remark 1.4.4. Moreover, here we extend the definition of μ to finite unions of arbitrary intervals (e. g., open or closed); it is easy to see that (1.4.7) holds true with this extension.

1.5. Expectation.

An *expectation* of a real-valued random variable is just another name for the (Lebesgue) integral over a measure space:

$$\mathbb{E}(X) = \int_\Omega X d\mathbb{P}.$$

Recall that the integral $\int_\Omega h d\mu$, where $h : \Omega \to \mathbb{R}$, is defined in three steps. First, if $h = \sum_{i=1}^n a_i \mathbb{I}_{A_i}$, where $A_1, \dots, A_n$, are measurable of finite measure (such h are called *simple functions*), put

$$\int_\Omega h d\mu := \sum_{i=1}^n a_i \mu(A_i);$$

check that this is well defined, i. e., does not depend on the choice of representation of h in this form. Second, if h is a non-negative measurable function, put

$$\int_\Omega h d\mu := \sup_{\substack{g \le h \\ g \text{ simple}}} \int_\Omega g d\mu.$$

Finally, for a general h, put

$$\int_\Omega h d\mu := \int_\Omega h \mathbb{I}_{h \ge 0} d\mu - \int_\Omega (-h \mathbb{I}_{h<0})\, d\mu.$$

whenever at least one of the terms is finite, otherwise we say that the integral does not exist[11]. To get used to the notation, we formulate the following proposition in terms of expectations rather than integrals, even though all the statements are true for arbitrary measure spaces.

Proposition 1.5.1. *The expectation satisfies the following properties:*

- *(linearity) if $\alpha, \beta \in \mathbb{R}$, and $\mathbb{E}X$ and $\mathbb{E}Y$ exist, then $\mathbb{E}(\alpha X + \beta Y)$ exists, and*
$$\mathbb{E}(\alpha X + \beta Y) = \alpha \mathbb{E}(X) + \beta E(Y);$$
- *(monotonicity) if X, Y are measurable such that $0 \le X(\omega) \le Y(\omega)$ for all $\omega \in \Omega$ and $\mathbb{E}Y$ exists, then $\mathbb{E}(X) \le \mathbb{E}(Y)$.*
- *(monotone convergence theorem) if $X_i \ge 0$ are measurable and $X_i \nearrow X$ almost surely, then $\mathbb{E}(X_i) \to \mathbb{E}(X)$.*

Remark 1.5.2. The expression "$X_i \nearrow X$ almost surely" (and, in general, "Property $P = P(\omega)$ holds almost surely"), used above, means that $X_i(\omega) \nearrow X(\omega)$ for $\mathbb{P}$-almost every $\omega \in \Omega$, that is, for all $\omega \in \Omega$ except for a set of measure zero.

It is, in general, not true that $X_n \to X$ almost surely implies $\mathbb{E}X_n \to \mathbb{E}X$:

Example 1.5.3. (Growing bump) Let $\Omega = (0, 1)$ with Lebesgue measure λ, and $X_n = n\mathbb{I}_{(0;\frac{1}{n})}$. Then $X_n(\omega) \to 0$ for any $\omega \in (0, 1)$, but $\mathbb{E}X_i = n\lambda((0; \frac{1}{n})) \equiv 1$.

The following sufficient condition is of huge importance in practice:

Proposition 1.5.4. *(Dominated convergence theorem) if X_i are measurable, $X_i \to X$ almost surely, and there exists a random variable $Y \ge 0$ with $\mathbb{E}(|Y|) < \infty$ such that if $|X_i| \le |Y|$ almost surely for all i, then $\mathbb{E}(X_i) \to \mathbb{E}(X)$.*

Proof of Propositions 1.5.1 and 1.5.4. We refer the reader to the "Measure and integral" course. □

The integral is an important tool to construct new measures from existing ones.

[11] in probability, by saying that X has expectation, one means that it exists and is finite

LEMMA 1.5.5. *Let $f \geq 0$ be a measurable fuction on a measure space $(\Omega, \mathcal{F}, \mu)$ (not necessarily with finite measure). Then μ', defined on $\mathcal{F}$ by*

$$\mu'(A) := \int_A f d\mu = \int_\Omega (f \cdot \mathbb{I}_A) d\mu$$

is a measure on $\mathcal{F}$.

PROOF. If $A = \emptyset$, then $\mathbb{I}_A = 0$, so $\mu'(A) = 0$. If $A_1, A_2, \cdots \in \mathcal{F}$ are disjoint and $A = \sqcup_{i=1}^\infty A_i$, then

$$\mathbb{I}_A = \lim_{n\to\infty} \sum_{i=1}^n \mathbb{I}_{A_i}$$

pointwise, and

$$\mu'(A) = \int_\Omega (f \cdot \mathbb{I}_A) d\mu = \int_\Omega \left(\lim_{n\to\infty} f \cdot \sum_{i=1}^n \mathbb{I}_{A_i} \right) d\mu = \lim_{n\to\infty} \int_\Omega \left(\sum_{i=1}^n f \cdot \mathbb{I}_{A_i} \right) d\mu = \lim_{n\to\infty} \sum_{i=1}^n \mu'(A_i),$$

where we have used Monotone convergence theorem in the third identity and linearity of the integral in the last one. □

DEFINITION 1.5.6. If, for a measure μ', there exists a function f such that $\mu'(A) \equiv \int_A f d\mu$, then the function f is called the *Radon-Nikodym derivative*[12] of μ' with respect to μ, denoted $f = \frac{d\mu'}{d\mu}$, or $d\mu' = f d\mu$. In the special case when μ is the Lebesgue measure (on $\mathbb{R}$ or on $\mathbb{R}^n$), and μ' is a probability measure - e. g., when μ' is a distribution of a random variable - the function f is called *probability denisity* (of that random variable).

REMARK 1.5.7. If the probability disctibution function F_X of a scalar random variable has a continuous derivative[13], then F' is a probability density of X (because $\mathbb{P}(X \leq a) = F_X(a) = \int_{-\infty}^a F'(x)dx$ by Newton-Leibnitz).

REMARK 1.5.8. The Radon-Nikodym theorem asserts that if $\mu(A) = 0$ implies $\mu'(A) = 0$, then there exists a function f such that $d\mu' = f d\mu$.

PROPOSITION 1.5.9. *(abstract change of variable theorem) Let $(\Omega_1, \mathcal{F}_1, \mathbb{P})$ be a probability space, $(\Omega_2, \mathcal{F}_2)$ a measurable space, $X : \Omega_1 \to \Omega_2$ a random variable and $f : \Omega_2 \to \mathbb{R}$ a measurable function. Then,*

$$\mathbb{E}(f \circ X) = \int_{\Omega_2} f d\mu_X, \tag{1.5.1}$$

where μ_X denotes the distribution of the random varaible X.

PROOF. If f is a simple function assuming only the values $a_1, \ldots, a_n$, then $f \circ X$ is also simple, and

$$\int_{\Omega_2} f(x) d\mu(x) = \sum a_i \mu_X(f = a_i) = \sum a_i \mathbb{P}(X \in \{\omega \in \Omega_2 : f(\omega) = a_i\}) = \sum a_i \mathbb{P}(f \circ X = a_i) = \mathbb{E}(f \circ X).$$

If $f \geq 0$, then

$$\int_{\Omega_2} f d\mu_X \leq \mathbb{E}(f \circ X),$$

because if g is a simple function approximating f from below, then $g \circ X$ is a simple function approximating $f \circ X$ from below. Now we proceed to the proof of another inequality. Given $\varepsilon > 0$, denote $f_\varepsilon(\omega) := \max\{a \in \varepsilon\mathbb{N} : a \leq f(\omega)\}$, and for $T > 0$ (large), denote $f_\varepsilon^T := f_\varepsilon \mathbb{I}_{f_\varepsilon \leq T}$. Since f_ε^T is a simple function and $f_\varepsilon^T \leq f$, one has

$$\int_{\Omega_2} f d\mu_X \geq \int_{\Omega_2} f_\varepsilon^T d\mu_X = \mathbb{E}(f_\varepsilon^T \circ X).$$

[12] In Statistics, Radon-Nikodym derivarive is also calles "likehood ratio".

[13] This condition is not optimal. The right condition is called *absolute contnuity* of F.

By monotone convergence theorem, $\mathbb{E}(f_\varepsilon^T \circ X) \to \mathbb{E}(f_\varepsilon \circ X)$ as $T \to \infty$, so, passing to the limit in the above inequality gives

$$\int_{\Omega_2} f d\mu_X \geq \mathbb{E}(f_\varepsilon \circ X) \geq \mathbb{E}(f \circ X) - \varepsilon,$$

since $f \leq f_\varepsilon + \varepsilon$. Letting $\varepsilon \to 0$ gives the desired result.

Finally, the general case follows from linearity of both sides of (1.5.1) and the identities

$$(f \cdot \mathbb{I}_{f \geq 0}) \circ X = (f \circ X) \cdot \mathbb{I}_{f \circ X \geq 0},$$

$$(f \cdot \mathbb{I}_{f < 0}) \circ X = (f \circ X) \cdot \mathbb{I}_{f \circ X < 0}.$$

□

This theorem is useful[14], in particular, when $\Omega_2 = \mathbb{R}$. Then,

$$\mathbb{E}(f(X)) = \int_{\mathbb{R}} f(x) d\mu_X(x);$$

in particular, taking f to be the identity map:

$$\mathbb{E}(X) = \int_{\mathbb{R}} x d\mu_X(x).$$

This reduces the computation of an expectation as an integral over an abstract measure space to integration over $\mathbb{R}$, provided that the disctribution μ_X of X is known. In many practical cases (when μ_X has a density, see below), this boils down to integration over the Lebesgue measure, or to the familiar Riemann integrals from calculus courses.

Let us collect further important properties of the expectation.

PROPOSITION 1.5.10. *The expectation satisfies the following useful inequalities:*

- *(Cauchy-Schwarz)*

$$\mathbb{E}(XY) \leq \sqrt{\mathbb{E}(X^2)} \cdot \sqrt{\mathbb{E}(Y^2)}. \tag{1.5.2}$$

- *(Holder) for $p, q > 0$ such that $\frac{1}{p} + \frac{1}{q} = 1$,*

$$\mathbb{E}(XY) \leq (\mathbb{E}|X|^p)^{\frac{1}{p}} (\mathbb{E}|X|^q)^{\frac{1}{q}}. \tag{1.5.3}$$

- *(Jensen) if $f : \mathbb{R} \to \mathbb{R}$ is a convex function and $\mathbb{E}|X| < \infty$, $\mathbb{E}(|f(X)|) < \infty$, then*

$$f(\mathbb{E}(X)) \leq \mathbb{E}(f(X)). \tag{1.5.4}$$

  Particular useful cases are $|E(X)| \leq \mathbb{E}(|X|)$ and $(\mathbb{E}X)^2 \leq \mathbb{E}X^2$.

- *(Chebyshev) for a non-negative random variable X and $a > 0$, one has*

$$\mathbb{P}(X \geq a) \leq \frac{\mathbb{E}X}{a}. \tag{1.5.5}$$

  It is sometimes useful to apply this inequality to a function of a given random variable, e. g.

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

PROOF. Here, we only sketch the proofs. First, by linearity of the expectation, the inequality $\mathbb{E}(X + Y)^2 \geq 0$ can be rewritten as

$$\mathbb{E}XY \leq \frac{\mathbb{E}X^2 + \mathbb{E}Y^2}{2}.$$

Now note that the left-hand side is invariant under multiplication of X by λ and simultaneous multiplication of Y by $\frac{1}{\lambda}$, while the right-hand side is not; optimizing over such λ gives (1.5.2).

[14]Note a slight technicality: f is required to be a measurable map from $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$. By contrast, a measurable function as defined in analysis is a measurable map from $(\mathbb{R}, \mathcal{L}(\mathbb{R}))$ to $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, $\mathcal{L}$ being the Lebesgue σ-algebra. We will mostly apply (1.5.1) to continuous functions, which are always Borel-to-Borel measurable.

Elementary calculus allows one to prove Young's inequality

$$XY \leq \frac{|X|^p}{p} + \frac{|Y|^q}{q},$$

and taking expectations and pulling the same trick with λ gives (1.5.3).

To prove Jensen's inequality, we use the fact that for a convex function f, we have

$$f(x) = \sup(ax + b),$$

where the supremum is taken over all a, b such that $ay + b \leq f(y)$ for all $y \in \mathbb{R}$. For any such a, b, we have

$$\mathbb{E}(f(X)) \geq \mathbb{E}(aX + b) \geq a\mathbb{E}X + b.$$

Taking a supremum gives Jensen's inequality.

Finally, $\mathbb{E}X \geq \mathbb{E}(X\mathbb{I}_{X\geq a}) \geq \mathbb{E}(a\mathbb{I}_{x\geq a}) = a\mathbb{P}(X \geq a)$, which is (1.5.5). □

In probability, one often encounters integrals of functions that depend on a parameter; one often needs to differentiate the integral with respect to this parameter. This is usually amounts to the identity

$$\frac{\partial}{\partial x}\int_\Omega f(x,\omega)d\mu(\omega) = \int_\Omega \frac{\partial}{\partial x} f(x,\omega)d\mu(\omega). \tag{1.5.6}$$

This identity is, however, not true in general; the conterexample is given by a version of the "growing bump":

Example 1.5.11. Let, for $y > 0$ and $x \in \mathbb{R}$,

$$f(x,y) = \begin{cases} e^{-\frac{y^2}{x^2}}, & x \neq 0 \\ 0, & x = 0 \end{cases}.$$

Then, $\partial_x f(0,y) = 0$ for all $y > 0$ (indeed, for $y \neq 0$, $e^{-\frac{y^2}{x^2}}$ decays to zero faster than any polynomial as $x \to 0$). However, by the change of variables $w = \frac{y}{x}$,

$$\int_0^\infty e^{-\frac{y^2}{x^2}} d\lambda(y) = x\int_0^\infty e^{-w^2} dw,$$

so

$$\partial_x \int_{\mathbb{R}} f(x,y)dy\Big|_{x=0} = \int_{\mathbb{R}} e^{-w^2} dw \neq 0$$

. (The value of the integral is $\sqrt{\pi}/2$, as we will see soon, but it is not important for the conclusion).

We now formulate two sufficient conditions for the formula (1.5.6) to be true. The first one belongs to the realm of real analysis; it assumes that the parameter belongs to an interval; the regularity assumption needed to outrule the "growing bump" examples is formulated in terms of $\partial_x f(x,\omega)$. The second one assumes that the function $x \mapsto f(x,\omega)$ is analytic in some domain in the complex plane, but the regularity asumption is imposed on $f(x,\omega)$ itself, and the conclusion is much stronger: one gets all the derivatives at once.

Theorem 1.5.12. *(Differentiating an integral, real version) Let $I \subset \mathbb{R}$ be an open interval, and $(\Omega, \mathcal{F}, \mu)$ a measure space. Assume that a function $f : I \times \Omega \to \mathbb{R}$ satisfies the following properties:*

- *for every $x \in I$, the function $\omega \mapsto f(x,\omega)$ is integrable;*
- *for almost every ω and every $x \in I$, the derivative $\partial_x f(x,\omega)$ of the function $x \mapsto f(x,\omega)$ exists;*
- *there is a measurable function $h : \Omega \to \mathbb{R}_{\geq 0}$ such that $\int_\Omega h d\mu < \infty$ and $|\partial_x f(x,\omega)| \leq h(\omega)$ for all $x \in I$ and almost all $\omega \in \Omega$.*

Then, the derivative $\varphi'(x)$ of the function $\varphi(x) := \int_\Omega f(x,\omega)d\mu(\omega)$ exists at all $x \in I$, and $\varphi'(x) = \int_\Omega \partial_x f(x,\omega)d\mu(\omega)$.

PROOF. One has, for every $x \in \Omega$ and $\delta > 0$ small enough,

$$\frac{\varphi(x+\delta) - \varphi(x)}{\delta} = \int_\Omega \frac{f(x+\delta,\omega) - f(x,\omega)}{\delta} d\mu(\omega).$$

Note that the integrand tends to $f_x(x,\omega)$ for almost all $\omega \in \Omega$, along any subsequence of $\delta \to 0$. Also,

$$\left| \frac{f(x+\delta,\omega) - f(x,\omega)}{\delta} \right| \leq \sup_{[x,x+\delta]} |\partial_x f(\cdot,\omega)| \leq h(\omega),$$

so the Dominated Convergence theorem readily applies, and we conclude

$$\frac{\varphi(x+\delta) - \varphi(x)}{\delta} \to \int_\Omega \partial_x f(x,\omega) d\mu(\omega)$$

along any subsequence of δ. □

THEOREM 1.5.13. *(Differentiating an integral, complex version) Let $\Lambda \subset \mathbb{C}$ be an open set, $(\Omega, \mathcal{F}, \mu)$ a measure space, and assume that a function $f : \Lambda \times \Omega \to \mathbb{C}$ satisfies the following properties:*

- *for every $z \in \Lambda$, the function $\omega \mapsto f(z,\omega)$ is measurable;*
- *for almost every ω, the function $z \mapsto f(z,\omega)$ is analytic in Λ;*
- *there is a measurable function $h : \Omega \to \mathbb{R}_{\geq 0}$ such that $\int_\Omega h < \infty$ and $|f(z,\omega)| \leq h(\omega)$ for all $z \in \Lambda$ and almost every $\omega \in \Omega$.*

Then, the function $\varphi(z) := \int_\Omega f(z,\omega) d\mu(\omega)$ is analytic in Λ, and $\frac{\partial^n}{\partial z^n}\varphi(z) = \int_\Omega \frac{\partial^n}{\partial z^n} f(z,\omega) d\mu(\omega)$ for all $n = 1, 2, \ldots$ and for all $z \in \Lambda$.

PROOF. Note that the statement of the theorem is local, i. e., it is sufficient to prove the statement in a neighborhood of every point in Λ. Recall that the derivative of an analytic function can be estimated in terms of the function itself:

$$\left| \frac{\partial}{\partial z} f(z,\omega) \right| = \left| \frac{1}{2\pi i} \int_{|\zeta - z| = r} \frac{f(\zeta,\omega)}{(\zeta - z)^2} d\zeta \right| \leq \frac{1}{2\pi r} \sup_{z \in \Lambda} |f(z,\omega)| \leq \frac{1}{2\pi r} h(\omega),$$

where $r = r_z$ is small enough so that $\{\zeta : |\zeta - z| \leq r\} \subset \Lambda$. By applying exactly the same argument as in the real case to a small ball contained in Λ (so that there is a uniform lower bound on r_z over this ball), we see that the (complex!) derivative $\varphi'(z)$ exists and equals $\int_\Omega \frac{\partial}{\partial z} f(z,\omega) d\mu(\omega)$, which is finite. By repeating the same argument, we get higher derivatives. □

1.6. Direct products of measure spaces and Fubini's theorem

In this section, we construct direct products of measure spaces. This construction is crucial for the probabilitsic concept of independence. We will, however, go beyond the setup of probability spaces, to cover also other nice measure such as the Lebesgue measure.

DEFINITION 1.6.1. A measure space $(\Omega, \mathcal{F}, \mu)$ is called σ-finite if there is a sequence of sets $E_i \in \Omega$, such that $\mu(E_i) < \infty$, and $\Omega = \cup_{i=1}^\infty E_i$.

Finite measures and the Lebesgue measures on $\mathbb{R}$ and $\mathbb{R}^n$ are, of course, σ-finite. The example of a measure space which is not σ-finite is a counting measure ($\mu(E) = |\mathrm{E}|$) on an uncountable space (e. g., the real line). The σ-finiteness assumption is essential for all the statements in this section.

PROPOSITION 1.6.2. *(Direct product of measure spaces) Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ be two σ-finite measure spaces. Then, there exists a unique measure $\mu_1 \otimes \mu_2$ on $\sigma(\mathcal{F}_1 \times \mathcal{F}_2)$, such that for any $A_1 \in \mathcal{F}_1$ and $A_2 \in \mathcal{F}_2$ with $\mu_1(A_1) < \infty$ and $\mu_2(A_2) < \infty$, one has*

$$\mu_1 \otimes \mu_2(A_1 \times A_2) = \mu_1(A_1) \cdot \mu_2(A_2). \tag{1.6.1}$$

PROOF. Denote $\mathcal{F}_1^0 := \{A \in \mathcal{F}_1; \mu_1(A) < \infty\}$, and simiarly for $\mathcal{F}_2^0$. Let us check that $\mathcal{F}_1^0 \times \mathcal{F}_2^0 = \{A \times B : A \in \mathcal{F}_1^0, B \in \mathcal{F}_2^0\}$ is a semi-ring. If $A_1, B_1 \in \mathcal{F}_1^0$ and $A_2, B_2 \in \mathcal{F}_2^0$, then

$$(A_1 \times A_2) \cap (B_1 \times B_2) = (A_1 \cap B_1) \times (A_2 \cap B_2) \in \mathcal{F}_1^0 \times \mathcal{F}_2^0.$$

The set $(A_1 \times A_2) \setminus (B_1 \times B_2)$ is a discoint union of three sets: $(A_1 \setminus B_1) \times (A_2 \setminus B_2)$, $(A_1 \setminus B_1) \times (A_2 \cap B_2)$ and $(A_1 \cap B_1) \times (A_2 \setminus B_2)$; each of them belongs to $\mathcal{F}_1^0 \times \mathcal{F}_2^0$. Hence, indeed, $\mathcal{F}_1^0 \times \mathcal{F}_2^0$ is a semi-ring.

To apply Theorem 1.4.5, we have to show that if $A \times B = \sqcup_{i=1}^{\infty} A_i \times B_i$, where $A, A_i \in \mathcal{F}_1^0$ and $B, B_i \in \mathcal{F}_2^0$, then

$$\mu_1(A) \cdot \mu_2(B) = \sum_{i=1}^{\infty} \mu_1(A_i) \cdot \mu_2(B_i).$$

A convenient trick to do so is to use the monotone convergence theorem. Define functions $f, f_N : \Omega_1 \to \mathbb{R}$ by $f = \mu_2(B)\mathbb{I}_A$, and $f_N := \sum_{i=1}^{N} \mu_2(B_i)\mathbb{I}_{A_i}$. Note that f_N and f are measurable functions on $(\Omega_1, \mathcal{F}_1)$. Moreover, for any $\omega \in \Omega_1$, $f_N(\omega) = \mu_2(\omega' \in \Omega_2 : (\omega, \omega') \in \sqcup_{i=1}^{N} A_i \times B_i)$. Therefore, by the lower continuity of the measure μ_2,

$$\lim_{N\to\infty} f_N(\omega) = \mu_2(\omega' \in \Omega_2 : (\omega, \omega') \in \sqcup_{i=1}^{\infty} A_i \times B_i) = \mu_2(\omega' \in \Omega_2 : (\omega, \omega') \in A \times B) = f(\omega).$$

Now, the monotone convergence theorem implies that

$$\sum_{i=1}^{N} \mu_2(B_i)\mu_1(A_i) = \mathbb{E} f_N \overset{N\to\infty}{\longrightarrow} \mathbb{E} f = \mu_1(A) \cdot \mu_2(B).$$

The uniqueness follows for Proposition 1.4.6 and σ-finiteness. □

REMARK 1.6.3. Assume that $A \in \mathcal{F}_1$, $B \in \mathcal{F}_2^0$, and that $\mu_1(A) = +\infty$. Since μ_1 is σ-finite, we can write $\Omega_1 = \cup_{i=1}^{\infty} E_i$, where $E_1 \subset E_2 \subset \dots$ and $\mu_1(E_i) < \infty$. Consequently,

$$\mu_1 \otimes \mu_2(A \times B) = \lim_{i\to\infty} \mu_1 \otimes \mu_2(A \cap E_i \times B) = \lim_{i\to\infty} \mu(A \cap E_i) \times \mu(B) = \begin{cases} 0, & \mu_2(B) = 0 \\ +\infty, & \mu_2(B) > 0 \end{cases}$$

Consequently, we can extend the identity (1.6.1) to the whole $\mathcal{F}_1 \times \mathcal{F}_2$, if we stick to the convention that $0 \cdot \infty = 0$ and $a \cdot \infty = \infty$ for $a \neq 0$.

COROLLARY 1.6.4. *If $(\Omega_i, \mathcal{F}_i, \mu_i)$, $i = 1, \dots, n$ are σ-finite measure spaces, then there is a unique measure $\mu_1 \otimes \dots \otimes \mu_n$ on $\sigma(\mathcal{F}_1 \times \dots \times \mathcal{F}_n)$, such that for any $A_1 \in \mathcal{F}_1, \dots, A_n \in \mathcal{F}_n$, one has*

$$\mu_1 \otimes \dots \otimes \mu_n(A_1 \times \dots \times A_n) = \prod_{i=1}^{n} \mu_i(A_i). \tag{1.6.2}$$

PROOF. The existence follows from Proposition 1.6.2 by induction. For the uniqueness, assume first that all $\mu_i(\Omega_i)$ are finite. Then, $\mathcal{F}_1 \times \dots \times \mathcal{F}_n$ is a π-system, and the $\pi - \lambda$ theorem (more preciselly, Corollary 1.3.5) gives the result. The uniqueness in the σ-finite case follows by approximation; we leave it as an exercise. □

One of the most important facts about product measures is that a product measure of a set can be computed as an integral.

THEOREM 1.6.5. *(Cavalieri principle) Given σ-finite measure spaces $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$, let $E \in \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$. Then*

- *for all $\omega \in \Omega_1$, the set $E_\omega := \{\omega' \in \Omega_2 : (\omega, \omega') \in E\}$ is $\mathcal{F}_2$ measurable;*
- *the funtion $f_E : \Omega_1 \to \mathbb{R}$, defined by $f_E(\omega) := \mu_2(E_\omega)$ is $\mathcal{F}_1$ measurable[15], and*

$$\mu_1 \otimes \mu_2(E) = \int_{\Omega_1} f_E d\mu_1.$$

[15]more precisely, $\mathcal{F}_1$-to-$\mathcal{B}(\mathbb{R})$ measurable. Here and below, the omitted target σ-algebra is always Borel.

PROOF. We assume first that $\mu_1(\Omega_1) < \infty$ and $\mu_2(\Omega_2) < \infty$. Let $\mathcal{L}$ denote the set of all subsets of $\Omega_1 \times \Omega_2$ for which the conclusion of the theorem holds. Note that it contains $\mathcal{F}_1 \times \mathcal{F}_2$, which is a π-system. So, we only have to chek that $\mathcal{L}$ is a λ-system. If $E := E^{(2)} \setminus E^{(1)}$ with $E^{(1)}, E^{(2)} \in \mathcal{L}$ and $E^{(1)} \subset E^{(2)}$, then, for any $\omega \in \Omega_1$

$$E_\omega = E^{(2)}_\omega \setminus E^{(1)}_\omega$$

is $\mathcal{F}_2$-measurable; moreover,

$$f_E(\omega) = f_{E^{(2)}}(\omega) - f_{E^{(1)}}(\omega)$$

is $\mathcal{F}_1$-to-$\mathcal{B}(\mathbb{R})$ measurable as a sum of two measurable functions, and

$$\mu_1 \otimes \mu_2(E) = \mu_1 \otimes \mu_2(E^{(2)}) - \mu_1 \otimes \mu_2(E^{(1)}) = \int_{\Omega_1} f_{E^{(2)}} d\mu_1 - \int_{\Omega_1} f_{E^{(1)}} d\mu_1 = \int_{\Omega_1} f_E d\mu_1,$$

so $E \in \mathcal{L}$. If $E = \cup_{i=1}^\infty E^{(i)}$ with $E^{(1)} \subset E^{(2)} \subset \dots$ and $E^{(i)} \in \mathcal{L}$, then, for all $\omega \in \Omega_1$,

$$E_\omega = \cup_{i=1}^\infty E^{(i)}_\omega,$$

which is a measurable set, and, by the lower continuity of μ_2,

$$f_E(\omega) = \lim_{i\to\infty} f_{E^{(i)}}(\omega).$$

The monotone convergence theorem and the lower continuity of $\mu_1 \otimes \mu_2$ gives

$$\int_{\Omega_1} f_E d\mu_1 = \lim_{i\to\infty} \int_{\Omega_1} f_{E^{(i)}} d\mu_1 = \lim_{i\to\infty} \mu_1 \otimes \mu_2(E^{(i)}) = \mu_1 \otimes \mu_2(E).$$

In the σ-finite case, let $E^{(i)} \in \mathcal{F}_1 \times \mathcal{F}_2$ be a sequence of sets such that $E^{(i)} \subset E^{(i+1)}$ and $\cup_{i=1}^\infty E^{(i)} = \Omega_1 \times \Omega_2$. By the finite case, the theorem is true for each of the sets $E \cap E^{(i)}$, so,

$$\mu_1 \otimes \mu_2(E) = \lim_{i\to\infty} \mu_1 \otimes \mu_2(E \cap E^{(i)}) = \lim_{i\to\infty} \int_{\Omega_1} f_{E\cap E^{(i)}} d\mu_1.$$

However, by lower continuity of μ_2, one has $f_{E\cap E^{(i)}}(\omega) \nearrow f_E(\omega)$ for every $\omega \in \mu_1$, and the conclusion follows from the monotone convergence theorem. □

Before proceeding further, we need to clarify a technical point.

DEFINITION 1.6.6. Given a measure space $(\Omega, \mathcal{F}, \mu)$, a set $A \subset \Omega$ is called a *null-set* (w. r. t. μ) if it is contained in a measurable set of measure zero. The measure μ is called *complete* if all null-sets are $\mathcal{F}$-measurable. If μ is not complete, denote

$$\overline{\mathcal{F}} := \{A \cup B \subset \Omega : A \in \mathcal{F}, B \text{ is a null-set}\}.$$

Note that $\overline{\mathcal{F}}$ is a σ-algebra, and μ can be extended to a measure $\overline{\mu}$ on $\overline{\mathcal{F}}$ by $\mu(A) := \mu(A')$, where $A' \subset A$ is any $\mathcal{F}$-measurable set such that $A' \setminus A$ is a null-set. When extended to $\overline{\mathcal{F}}$, the measure μ is complete; we call this procedure a *completion* of the measure μ. (Exercise: prove all the claims in this paragraph).

Thus, we denote by $\mathcal{F}_1 \otimes \dots \otimes \mathcal{F}_n$ the completion of $\sigma(\mathcal{F}_1 \times \dots \times \mathcal{F}_n)$ with respect to the product measure $\mathbb{P}_1 \otimes \dots \otimes \mathbb{P}_n$, and usually assume the product measure to be extended to $\mathcal{F}_1 \otimes \dots \otimes \mathcal{F}_n$.

EXAMPLE 1.6.7. Let $A_1 \subset \Omega_1$ be a non-measurable set, and $A_2 \in \mathcal{F}_2$ be such that $\mathbb{P}_2(A_2) = 0$. Then $A_1 \times A_2$ is, in general, not measurable with respect to $\sigma(\mathcal{F}_1 \times \dots \times \mathcal{F}_n)$, but $A_1 \times A_2 \subset \Omega_1 \times A_2$, and $\mathbb{P}_1 \otimes \mathbb{P}_2(\Omega_1 \times A_2) = 1 \cdot 0 = 0$, so $A_1 \times A_2$ is a null-set, and, in paticular, it is $\mathcal{F}_1 \otimes \dots \otimes \mathcal{F}_n$ - measurable.

THEOREM 1.6.8. *(Cavalieri principle for completed spaces). Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ be complete σ-finite measure spaces. Assume that $E \subset \Omega_1 \times \Omega_2$ is $\mathcal{F}_1 \otimes \mathcal{F}_2$ measurable. Then*

- *for μ_1- almost all*[16] *$\omega \in \Omega_1$, the set $E_\omega := \{\omega' \in \Omega_2 : (\omega, \omega') \in E\}$ is $\mathcal{F}_2$ measurable;*
- *the fucntion $f_E(\omega) := \mu_2(E_\omega)$ is $\mathcal{F}_1$-to-$\mathcal{B}(\mathbb{R})$ (defined almost everywhere on Ω_1) is measurable;*

[16] this expression means "for all, except for a set of measure zero"

- *one has the identity*

$$\mu_1 \otimes \mu_2(E) = \int_{\Omega_1} f_E d\mu_1.$$

PROOF. Assume that $E = A \cup B$, where $A \in \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$ and B is a null-set. Then $E = A \sqcup (B \setminus A)$, where $B \setminus A$ is also null-set. The collection of all subsets of $\Omega_1 \times \Omega_2$ that satisfy the conclusion of the theorem is closed under disjoint unions. Therefore, in view of Theorem 1.6.5, it suffices to prove the result for E a null-set.

However, if $E \subset B \in \sigma(\mathcal{F}_1 \times \mathcal{F}_2)$ with $\mathbb{P}_1 \otimes \mathbb{P}_2(B) = 0$, then for all $\omega \in \Omega_1$, $E_\omega \subset B_\omega$. Applying Proposition 1.6.5 to B, we see that $\int_{\Omega_1} f_B d\mathbb{P}_1 = 0$. Since $f_B \geq 0$, actually $f_B = 0$ for almost every $\omega \in \Omega_1$. This means that for almost every $\omega \in \Omega_1$, E_ω is a null-set (and thus is $\mathcal{F}_2$-measurable, since $\mathbb{P}_2$ is complete), and $f_E = 0$ a. e. □

THEOREM 1.6.9. *(Tonnelli's theorem) Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ be complete σ-finite measure spaces. If $f : \Omega_1 \times \Omega_2 \to \mathbb{R}_{\geq 0}$ is $\mathcal{F}_1 \otimes \mathcal{F}_2$ measurable, then*

- *For a. e. $\omega \in \Omega_1$, the function $f_\omega(\cdot) := f(\omega, \cdot) : \Omega_2 \to \mathbb{R}$ is $\mathcal{F}_2$ measurable.*
- *The function $\omega \mapsto \int_{\Omega_2} f_\omega(\omega') d\mu_2(\omega')$, defined almost everywhere on Ω_1, is $\mathcal{F}_1$-measurable.*
- *The following identity holds:*

$$\int_{\Omega_1 \times \Omega_2} f d(\mu_1 \otimes \mu_2) = \int_{\Omega_1} \left(\int_{\Omega_2} f_\omega(\omega') d\mu_2(\omega') \right) d\mu_1(\omega). \tag{1.6.3}$$

PROOF. When $f = \mathbb{I}_E$, $E \in \mathcal{F}_1 \otimes \mathcal{F}_2$, the result follows directly from Theorem 1.6.8. The class of functions for which is holds is closed under linear combinations (hence contains all simple functions) and by monotone convergence (hence contains all non-negative measurable functions). □

COROLLARY 1.6.10. *(Fubini's theorem) Let $(\Omega_1, \mathcal{F}_1, \mu_1)$ and $(\Omega_2, \mathcal{F}_2, \mu_2)$ be complete σ-finite measure spaces. If $f : \Omega_1 \times \Omega_2 \to \mathbb{R}$ is $\mathcal{F}_1 \otimes \mathcal{F}_2$ measurable and such that*

$$\int_{\Omega_1 \times \Omega_2} |f| d(\mathbb{P}_1 \otimes \mathbb{P}_2) < \infty, \tag{1.6.4}$$

then the conclusion of Theorem 1.6.9 holds.

PROOF. Write $f = f\mathbb{I}_{f \geq 0} - (-f\mathbb{I}_{f<0})$ and apply Theorem 1.6.9 to each term. Note that (1.6.4) guarantees that

$$\int_{\Omega_1} \left(\int_{\Omega_2} f_\omega(\omega') \mathbb{I}_{f_\omega \geq 0} d\mu_2(\omega') \right) d\mu_1(\omega) = \int_{\Omega_1 \times \Omega_2} f\mathbb{I}_{f \geq 0} d(\mu_1 \otimes \mu_2) < \infty,$$

therefore, $\int_{\Omega_2} f_\omega(\omega') \mathbb{I}_{f_\omega \geq 0} d\mu_2(\omega') < \infty$ for a. e. $\omega \in \Omega_1$, and similarly for $\int_{\Omega_2} -f_\omega(\omega') \mathbb{I}_{f_\omega < 0} d\mu_2(\omega')$. Therefore, their difference is well defined almost everywhere on Ω_1, and we can apply linearity to finish the proof. □

REMARK 1.6.11. Notice that Theorem 1.6.9 states, in particular, that if one side of (1.6.3) is finite, then another one is also finite, and they are equal. By contrast, if the sign of f is not fixed, it might happen that the right-hand side of (1.6.3) is finite, but the left-hand side is not defined. In this case, it might happen that exchanging the order of integration changes the result.

1.7. Infinite products of probability spaces and Kolmogorov extension theorem

In this section, we construct infinite direct products of probability spaces, and also non-direct products (projective limits). We start with the countable case. Let $(\Omega_i; \mathcal{F}_i, \mathbb{P}_i)_{i=1}^\infty$ be probability spaces, and let $\Omega := \prod_{i=1}^\infty \Omega_i$, the set of all infinite sequences $(\omega_1, \omega_2, \dots)$, where $\omega_i \in \Omega_i$.

DEFINITION 1.7.1. Denote

$$\mathcal{C}_N := \{A_1 \times \ldots \times A_N \times \prod_{i=N+1}^{\infty} \Omega_i \subset \Omega\}$$

and $\mathcal{C} := \cup_{N=1}^{\infty} \mathcal{C}_N$. The elements of $\mathcal{C}$ are called *cylindrical sets*.

Note that the class of cylindrical sets forms a semi-ring. Indeed, $\emptyset \in \mathcal{C}$, and if $A, B \in \mathcal{C}$, then $A, B \in \mathcal{C}_N$ for some N. Assume that

$$A = A_1 \times \cdots \times A_N \times \Omega_{N+1} \times \dots, \quad B = B_1 \times \cdots \times B_N \times \Omega_{N+1} \times \dots$$

Then

$$A \cap B = A_1 \cap B_1 \times \cdots \times A_N \cap B_N \times \Omega_{N+1} \times \cdots \in \mathcal{C}$$

and

$$B \setminus A = \bigsqcup_{\substack{E_i = A_i \cap B_i \text{ or } B_i \setminus A_i \\ \exists i\colon E_i = B_i \setminus A_i}} \prod_{i=1}^{N} E_i \times \prod_{i=N+1}^{\infty} \Omega_i.$$

LEMMA 1.7.2. *Assume that $\mu : \mathcal{C} \to \mathbb{R}_{\geq 0}$ is such that*

- *$\mu|_{\mathcal{C}_N}$ is finitely additive for each N;*
- *μ is upper semicontinuous: if $B_1 \supset B_2 \supset \dots$ are finite disjoint unions[17] of sets in $\mathcal{C}$ and $\cap_{i=1}^{\infty} B_i = \emptyset$, then $\mu(B_n) \to 0$ as $n \to \infty$.*

Then μ is a pre-measure on $\mathcal{C}$.

PROOF. Let $A = \sqcup_{i=1}^{\infty} A_i$, where $A, A_i \in \mathcal{C}$; write $A = (\sqcup_{i=1}^{n} A_i) \sqcup \tilde{A}_n$. Note that there exists M>0 such that $A, A_1, \dots, A_n \in \mathcal{C}_M$. By additivity of μ on $\mathcal{C}_M$, this implies

$$\mu(A) = \sum_{i=1}^{n} \mu(A_i) + \mu(\tilde{A}_n),$$

and the result follows by applying the lower semicontinuity to $\tilde{A}_n$. □

THEOREM 1.7.3. *(Countable products of measure spaces) There is a unique probability measure $\mathbb{P}$ on $\sigma(\mathcal{C})$ satisfying*

$$\mathbb{P}\left(A_1 \times \cdots \times A_N \times \prod_{i=N+1}^{\infty} \Omega_i\right) = \mathbb{P}_1(A_1) \cdot \ldots \cdot \mathbb{P}_N(A_N) \tag{1.7.1}$$

for all N and all $A_1 \in \mathcal{F}_1, \dots, A_N \in \mathcal{F}_N$.

PROOF. Note that $\mathbb{P}$, defined by (1.7.1), extends to a measure on $\sigma(\mathcal{C}_N)$ for each N (essentially, the N-fold product measure); in particular, its restriction to $\mathcal{C}_N$ is finitely additive. In view of Caratheodory extension theorem and Lemma 1.7.2, it suffices to check the upper continuity of $\mathbb{P}$.

Let $B_1 \supset B_2 \supset \dots$ be finite disjoint unions of sets in $\mathcal{C}$ and $\cap_{i=1}^{\infty} B_i = \emptyset$, and assume that $\mathbb{P}(B_n) > \varepsilon > 0$ for all n. Given $\omega_1 \in \Omega_1$, define the sections

$$B_n^{\omega_1} := \{(\omega_2, \omega_3, \dots) \in \Omega_2 \times \Omega_3 \times \cdots : (\omega_1, \omega_2, \dots) \in B_n)\}.$$

Since $B_n \in \mathcal{C}_N$ for some $N = N(n)$, these sections have the form

$$B_n^{\omega_1} = (B_n^{\omega_1})' \times \Omega_{N+1} \times \Omega_{N+2} \times \dots$$

Let

$$h_n(\omega_1) := \mathbb{P}(\Omega_1 \times B_n^{\omega_1}) = \mathbb{P}_2 \otimes \cdots \otimes \mathbb{P}_N((B_n^{\omega_1})')$$

By Cavalieri's principle (applied to $\mathbb{P}_1 \otimes \cdots \otimes \mathbb{P}_N$), $h_n(\omega_1)$ is a measurable function and

$$\int_{\Omega_1} h_n d\mathbb{P}_1 = \mathbb{P}(B_n) > \varepsilon.$$

[17] recall that we extend μ to such sets according to Remark 1.4.4.

Now, h_n is a decreasing sequence of non-negative functions (bounded above by 1), therefore, by monotone convergence theorem,

$$\int_{\Omega_1} \lim_{n\to\infty} h_n = \lim \int_{\Omega_1} h_n \geq \varepsilon,$$

Therefore, there exists $\omega_1 \in \Omega_1$ such that $\lim_{n\to 0} h_n(\omega_1) > \frac{\varepsilon}{2}$, i. e., such that $\mathbb{P}(\Omega_1 \times B_n^{\omega_1}) > \frac{\varepsilon}{2}$ for all n.

We now iterate the procedure. By similar reasoning, we can find $\omega_2 \in \Omega_2$ such that for all n, $\mathbb{P}(\Omega_1 \times \Omega_2 \times B_n^{\omega_1\omega_2}) > \frac{\varepsilon}{4}$, where

$$B_n^{\omega_1\omega_2} := \{(\omega_3, \omega_4, \dots) \in \Omega_3 \times \Omega_4 \times \cdots : (\omega_1, \omega_2, \dots) \in B_n\}.$$

And so on. This way, we obtain an infinite sequence $\omega \in (\omega_1, \omega_2, \dots) \in \Omega$. We claim that ω belongs to each B_n. Indeed, if $\omega \notin B_n$, then, for N large enough, $(\omega_1, \dots, \omega_N, \omega'_{N+1}, \omega'_{N+2}, \dots) \notin B_n$ for any $\omega'_{N+1}, \omega'_{N+1}, \dots$. This means that $B_n^{\omega_1\omega_2\cdots\omega_N} = \emptyset$, a contradiction that shows that $\cap_{i=1}^{\infty} B_i \neq \emptyset$. □

We now switch to non-direct products (projective limits) of measures. Now, we consider a sequence of Let, for each n, μ_n be a probability measure on $\sigma(\mathcal{F}_1 \times \cdots \times \mathcal{F}_n)$.

Definition 1.7.4. Suppose $(\Omega_1, \mathcal{F}_1), (\Omega_2, \mathcal{F}_2), \dots$ are measurable spaces. A family $\{\mu_n\}$, where μ_n is a probability measure on $\sigma(\mathcal{F}_1 \times \cdots \times \mathcal{F}_n)$, is called *consistent* if, for every n and for every $A \in \sigma(\mathcal{F}_1 \times \cdots \times \mathcal{F}_n)$, one has

$$\mu_{n+1}(A \times \Omega_{n+1}) = \mu_n(A).$$

Given a consistent collection $\{\mu_n\}$, we can define a function $\mu : \mathcal{C} \to \mathbb{R}_{\geq 0}$ by

$$\mu(A' \times \Omega_{n+1} \times \Omega_{n+2} \times \dots) := \mu_n(A')$$

for every $A' \in \sigma(\mathcal{F}_1 \times \cdots \times \mathcal{F}_n)$. The consistency condition guarantees that μ is well defined, i. e. depends only on the set $A' \times \Omega_{n+1} \times \Omega_{n+2} \times \dots$ and not on the choice of n.

Example 1.7.5. (Direct products of measure spaces) If $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$ are probability spaces, then

$$\mu_n := \mathbb{P}_1 \otimes \cdots \otimes \mathbb{P}_n$$

is a consistent family of measures.

Clearly, μ is finitely additive on each $\mathcal{C}_n$. We would like to invoke Lemma 1.7.2 and Caratheodory extension theorem to extend μ to $\sigma(\mathcal{C})$. However, there are examples where such an extension does not exist. To avoid these patologies, we need to put additional structure of *metric space* on each Ω_i. We need the following preliminary result, very important on its own.

Theorem 1.7.6. *(Regularity of Borel measures). Let Ω be a complete separable metric space, and let μ be a finite measure on $\mathcal{B}(\Omega)$. Then, every set $A \in \mathcal{B}(\Omega)$ satisfies*

> *the approximation property: for every $\varepsilon > 0$, there exists a compact set $K_A \subset A$ and an open set $O_A \supset A$ such that $\mu(O_A) - \mu(K_A) < \varepsilon$.*

Proof. We first prove that every set $A \in \mathcal{B}(\Omega)$ *in any* metric space satisfies a weaker property:

> *The closed approximation property*: for every $\varepsilon > 0$, there exists a *closed* set $C_A \subset A$ and an open set $O_A \supset A$ such that $\mu(O_A) - \mu(C_A) < \varepsilon$.

Let us first see that closed sets satisfy the closed approximation property. Indeed, if A is closed, one can take $C_A := A$ and $O_A = O_\delta := \cup_{x\in A} B_\delta(x)$, where $B_\delta(x)$ is the open ball around x of radius δ. Since $\Omega \setminus A$ is open, for each $x \in \Omega \setminus A$ there exists $\delta_0 > 0$ such that $B_{\delta_0}(x) \subset \Omega \setminus A$. Then $x \notin O_\delta$ for $\delta < \delta_0$. This shows that $\cap_{n=1}^{\infty} O_{\frac{1}{n}} = A$, therefore $\lim_{n\to\infty} \mu(O_{\frac{1}{n}}) = \mu(A)$, which gives the desired result.

It remains to show that the sets satisfying the closed approximation property form a σ-algebra. Indeed, Ω does satisfy it, since it is closed. If O_A and C_A are approximations to A, then $O_A^c \subset A^c$ is closed, $C_A^c \supset A^c$ is open, and

$$\mu(C_A^c) - \mu(O_A^c) = \mu(\Omega) - \mu(C_A) - \mu(\Omega) + \mu(O_A) < \varepsilon,$$

so the approximation property is preserved under complements. If A_1, A_2 are approximated by O_1, C_1 and O_2, C_2 respecively, put $O = O_1 \cap O_2$ and $K = C_1 \cap C_2$. Then, $C \subset A_1 \cap A_2 \subset O$, and we have

$$O \setminus C = (O_1 \cap O_2 \cap C_1^c) \cup (O_1 \cap O_2 \cap C_2^c) \subset (O_1 \cap C_1^c) \cup (O_2 \cap C_2^c),$$

so

$$\mu(O \setminus C) \leq \mu(O_1 \setminus C_1) + \mu(O_2 \setminus C_2),$$

which shows that approximation property is preserved under finite interstctions. Finally, if A_i are disjoint and satify approximation property, let $O_1, O_2, \dots$ be open sets and $C_1, C_2, \dots$ be closed sets such that $C_i \subset A_i \subset O_i$ and $\mu(O_i \setminus C_i) < \frac{\varepsilon}{2^{i+1}}$. Since $\mu(\sqcup_{i=1}^{\infty} C_i) < \infty$, we can choose N so that $\sum_{i=N+1}^{\infty} \mu(C_i) < \frac{\varepsilon}{2}$. Then $\cup_{i=1}^{\infty} O_i \supset \cup_{i=1}^{\infty} A_i$ is open, $\sqcup_{i=1}^{N} C_i \subset \cup_{i=1}^{\infty} A_i$ is closed, and

$$\mu(\cup_{i=1}^{\infty} O_i) - \mu(\sqcup_{i=1}^{N} C_i) \leq \sum_{i=1}^{\infty} \mu(O_i) - \mu(\sqcup_{i=1}^{N} C_i) = \sum_{i=1}^{\infty} (\mu(O_i) - \mu(C_i) + \sum_{i=N+1}^{\infty} \mu(C_i) \leq \frac{\varepsilon}{2} + \frac{\varepsilon}{2},$$

concluding the proof of the closed approximation property.

Now we prove that if Ω is a complete separable metric space, then Ω itself satisfies the approximation property. Fix $\varepsilon > 0$, and let $y_1, y_2, \dots$ be a dense subsequence of Ω. Then, for each $\delta > 0$, we have $\cup_{n=1}^{\infty} \overline{B_\delta(y_n)} = \Omega$. By the lower continuity of a measure μ, this means that for any $m > 0$, there exists $n = n(m)$ such that

$$\mu\left(\cup_{n=1}^{n(m)} \overline{B_{\frac{1}{m}}(y_n)}\right) > \mu(\Omega) - \frac{\varepsilon}{2^m}.$$

Denote $C_m := \cup_{n=1}^{n(m)} \overline{B_{\frac{1}{m}}(y_n)}$, and take

$$K := \cap_{m=1}^{\infty} C_m;$$

we claim that $O_\Omega = \Omega$ and $K_\Omega := K$ satisfy the desired properties. First,

$$\mu(\Omega \setminus K) = \mu\left(\Omega \cap (\cup_{m=1}^{\infty} C_m^c)\right) = \mu\left(\cup_{m=1}^{\infty} (\Omega \setminus C_m)\right) \leq \sum_{m=1}^{\infty} \mu(\Omega \setminus C_m) < \sum_{m=1}^{\infty} \frac{\varepsilon}{2^m} = \varepsilon.$$

it remains to check that K is compact. Each C_m is closed, therefore K is closed, therefore K is a complete metric space. It suffices to show that for every $\varepsilon > 0$, K contains a finite ε-net, that is, a finite subset $x_1, \dots, x_{N(\varepsilon)}$ such that $K \subset \cup_{i=1}^{N(\varepsilon)} B_\varepsilon(x_i)$. But if m is such that $\frac{1}{m} < \varepsilon$, then $y_1, \dots, y_{n(m)}$ form an ε-net for C_m, and hence for K. Thus, indeed, Ω satisfies the approximation property.

Finally, it suffices to note that if O_A is open, C_A is closed with $\mu(O_A) - \mu(C_A) < \varepsilon$, then $K_A := C_A \cap K_\Omega$ is compact, and

$$\mu(O_A \setminus K_A) = \mu((O_A \setminus C_A) \cup (O_A \setminus K_\Omega)) \leq \mu(O_A \setminus C_A) + \mu(\Omega \setminus K_\Omega) \leq 2\varepsilon.$$

□

REMARK 1.7.7. The assumption for Ω to be complete and separable in the last theorem can be replaced with the assumption that Ω is a countable union of compact sets. (Indeed, if $K_1 \subset K_2 \subset \dots$ are compact and exhaust Ω, then $\mu(K_n) \to \mu(\Omega)$, and then the proof is finished as above.) The most important cases are $\Omega = \mathbb{R}$ and $\Omega = \mathbb{R}^n$, which are both complete separable and countable unions of compacts.

THEOREM 1.7.8. *(Kolmogorov extension theorem) Assume that $\{\mu_n\}$ is a consisent family of probability measures on $(\Omega_n, \mathcal{F}_n)$, where each Ω_n is a sigma-compact metric space and $\mathcal{F}_n = \mathcal{B}(\Omega_n)$. Then, there is a unique measure μ on $\sigma(\mathcal{C})$ such that*

$$\mu(A \times \Omega_{n+1} \times \Omega_{n+2} \times \dots) = \mu_n(A)$$

for any $A \in \sigma(\mathcal{F}_1 \times \dots \times \mathcal{F}_n)$.

We begin with a lemma. Denote $\Omega = \Omega_1 \times \Omega_2 = \dots$ We say that a set $A \subset \Omega$ is a *compact cylinder* if, for some N, $A = A_1 \times \dots \times A_N \times \prod_{i=N+1}^{\infty} \Omega_i$, where each A_i is either compact, or coincides with Ω_i.

LEMMA 1.7.9. *Assume that $K^{(1)} \supset K^{(2)} \supset \dots$ are finite disjoint unions of non-empty compact cylinders. Then $\cap_{i=1}^{\infty} K^{(i)}$ is non-empty.*

PROOF. *Case 1: each $K^{(i)}$ is a compact cylinder.* If $K^{(i)} = K_1^{(i)} \times K_2^{(i)} \times \dots$ and $K^{(i)}$ are nested and non-empty, then each $K_j^{(i)}$ is non-empty, and $K_j^{(1)} \supset K_j^{(2)} \supset \dots$ for every j. Since $K_j^{(i)}$ are either compact, or coincide with Ω_j, one has $\cap_{i=1}^\infty K_j^{(i)} \neq \emptyset$. Pick, for each j, $\omega_j \in \cap_{i=1}^\infty K_j^{(i)}$. Then $(\omega_1, \omega_2, \dots) \in \cap_{i=1}^\infty K^{(i)}$.

General case. Let $K^{(i)} = \sqcup_{j=1}^{N_i} K^{(i,j)}$, where $K^{(i,j)}$ are compact cylinders. We may assume that for each i, j, there is an index $\pi(j) \in \{1, \dots N_{i-1}\}$ such that $K^{(i,j)} \subset K^{(i-1,\pi(j))}$. If that is not the case, we can refine the partition of $K^{(i)}$ into compact cylinders: just replace, consequtively for each $i = 2, 3, \dots$, the collection $\{K^{(i,j)}\}_{j=1}^{N_i}$ by $\{K^{(i,j)} \cap K^{(i-1,j')}\}_{j,j'=1}^{N_i, N_{i-1}}$. This way, the collection $K^{(i,j)}$ can be given a tree structure, where $K^{(i-1,\pi(j))}$ is a parent of $K^{(i,j)}$ (since for each i, $K^{(i,j)}$ are disjoint, the parent is unique). Since each $K^{(i)}$ is non-empty, this tree must have an infinite branch $K^{(1,j_1)} \supset K^{(2,j_2)} \supset \dots$, where all $K^{(i,j_i)}$ are non-empty, and we apply Case 1 to this branch. □

PROOF OF THEOREM 1.7.8. By Lemma 1.7.2 and Caratheodory's extension theorem, it suffices to show that μ is upper semi-continuous. We first claim that for each cylindrical set A and for each $\varepsilon > 0$, we can find a compact cylinder $K_A \subset K$ sich that $\mu(A) - \mu(K_A) < \varepsilon$. Indeed, let $A = A' \times \Omega_N \times \Omega_{N+1} \times \dots$. By regularity theorem (Theorem 1.7.6) applied to A', we can find a compact subset $K_{A'}$ of $\Omega_1 \times \dots \times \Omega_N$ with $\mu_N(A') - \mu_N(K_{A'}) < \varepsilon$. Now take

$$K_A = \pi_1(K_{A'}) \times \dots \times \pi_N(K_{A'}) \times \Omega_{N+1} \times \dots,$$

where π_i are natural projections from $\Omega_1 \times \dots \times \Omega_N$ to Ω_i. Since π_i are continuous, each $\pi_i(K_{A'})$ is compact, and

$$K_{A'} \subset \pi_1(K_{A'}) \times \dots \times \pi_N(K_{A'}) \subset A',$$

so K_A has all the desired properties.

Now, if $B_1 \supset B_2 \supset \dots$ are finite disjoint unions of cylindrical sets such that $\mu(B_i) \geq \varepsilon > 0$ for all i, and let K_i be a finite disjoint union of compact cylinders such that $\mu(B_i) - \mu(K_i) < \frac{\varepsilon}{2^{i+1}}$. Then

$$\mu(B_n \setminus \cap_{i=1}^n K_i) = \mu\left(B_n \cap (\cup_{i=1}^n K_i^c)\right) = \mu\left(\cup_{i=1}^n \left(B_n \cap K_i^c\right)\right) \leq \sum_{i=1}^n \mu\left(B_n \setminus K_i\right) \leq \sum_{i=1}^n \mu\left(B_i \setminus K_i\right) \leq \frac{\varepsilon}{2}.$$

Therefore, $K^{(n)} = \cap_{i=1}^n K_i \subset B_i$ is non-empty; clearly, they are also finite unions of compact cylinders. Applying Lemma 1.7.9, we see that $\cap_{i=1}^\infty B_i \neq \emptyset$. □

REMARK 1.7.10. (Uncountable infinite products and Kolmogorov extension) Suppose T is any set and $(\Omega_t, \mathcal{F}_t, \mathbb{P}_t)$, $t \in T$ is a collection of σ-compact metric probability spaces with $\mathcal{F}_t = \mathcal{B}(\Omega)$. Assume that for each finite subset $\Lambda \subset T$, a probability measure μ_Λ is defined on the corresponding product $\Omega^\Lambda = \prod_{t \in \Lambda} \Omega_t$. (more precisely, on the σ-algebra $\mathcal{F}_\Lambda = \sigma(\prod_{t \in \Lambda} \mathcal{F}_t)$) For $\Lambda' \subset \Lambda$, denote by $\pi_{\Lambda \to \Lambda'}$ the natural projection from Ω^Λ to $\Omega^{\Lambda'}$. The family μ_Λ is called *consistent* if $\mu_\Lambda(\pi_{\Lambda \to \Lambda'}^{-1}(A)) = \mu_{\Lambda'}(A)$ for each measurable A. We call a set $A \subset \Omega^T$ cylindrical if A has the form $A = \pi_{T \to \Lambda}^{-1}(A')$ for some finite Λ and for $A' \in \mathcal{F}_\Lambda$; in this case, $\mu(A) = \mu_\Lambda(A')$ is well-defined on the semi-ring of cylindrical sets. One has the following result:

- μ extends uniquely to $\sigma(\cup_\Lambda \mathcal{F}_\Lambda)$;
- if μ_Λ are product measures, then the same holds true for arbitrary family $(\Omega_t, \mathcal{F}_t, \mathbb{P}_t), t \in T$, i. e., no metric structure is required.

The proof is based on the followins observations:

- then T is countable, the definitions coincide with the previous ones;
- define, for every countable subset $T' \subset T$, $\mathcal{F}_{T'} = \sigma(\cup_{\Lambda \subset T', \Lambda \text{ finite}} \mathcal{F}_\Lambda)$. Then $\cup_{T' \subset T, T' \text{ countable}} \mathcal{F}_\Lambda$ is a σ-algebra.

The details are left to the reader.

CHAPTER 2

Sums of indepndent random variables

2.1. Independent events and variables

In this section, all the events and random variables are defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$

DEFINITION 2.1.1. Two events $A, B \in \mathcal{F}$ are called *independent* if

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

Two random variables $f_1 : \Omega \to \Omega_1$ and $f_2 : \Omega \to \Omega_2$ are called independent if for any measurable sets $A_1 \in \Omega_1$ and $A_2 \in \Omega_2$, the events $f_1 \in A_1$ and $f_2 \in A_2$ are independent.

REMARK 2.1.2. If $\mathbb{P}(A) = 0$, then A, B are independent for any B. If $P(A) \neq 0$, then the definition is equivalent to

$$\mathbb{P}(B|A) = \mathbb{P}(B).$$

DEFINITION 2.1.3. A finite collection $A_1, \dots, A_n$ of events is called *independent* if, for any subset $1 \leq i_1 < \dots < i_k \leq n$, one has $\mathbb{P}(A_{i_1} \cap \dots \cap A_{i_k}) = \mathbb{P}(A_{i_1}) \cdot \dots \cdot \mathbb{P}(A_{i_k})$. A finite collection $X_1, \dots, X_n$ of random variables is independent if for any measurable sets $A_1, \dots, A_n$, the events $\{X_i \in A_i\}$, $i = 1, \dots, n$, are independent.

DEFINITION 2.1.4. An infinite collection of events (random variables) is called independent if all its finite subcollections are independent.

DEFINITION 2.1.5. A collection $\mathcal{A}_t \subset \mathcal{F}, t \in T$ of *families* of events (e. g., of σ-algebras) is called independent if every collection $A_t, t \in T$ of events such that $A_t \in \mathcal{A}_t$, is independent.

REMARK 2.1.6. (pairwise independent vs. independent) Events $A_1, \dots, A_n$ are called *pairwise independent* if A_i is independent of A_j for any $i \neq j$. Pairwise independence is *strictly weaker* than independence. As an example, let $\Omega = \{1, 2, 3, 4\}$ equipped with uniform measure ($\mathbb{P}(\{1\}) = \mathbb{P}(\{2\}) = \mathbb{P}(\{3\}) = \mathbb{P}(\{4\}) = \frac{1}{4}$), and consider the events $A_1 = \{1, 2\}$, $A_2 = \{1; 3\}$, $A_3 = \{1; 4\}$. They are pairwise independent: e. g.,

$$\mathbb{P}(A_1 \cap A_2) = \mathbb{P}(\{1\}) = \frac{1}{4} = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2).$$

However, $\mathbb{P}(A_1 \cap A_2 \cap A_3) = \mathbb{P}(\{1\}) = \frac{1}{4} \neq \frac{1}{8} = \mathbb{P}(A_1) \cdot \mathbb{P}(A_2) \cdot \mathbb{P}(A_3)$.

The following proposition shows that the study of independent random variables boils down to the study of product measures.

PROPOSITION 2.1.7. *Let $X_1, X_2, \dots$ be random variables on the same space Ω, with values in measurable spaces $(\Omega_1, \mathcal{F}_1), (\Omega_2, \mathcal{F}_2), \dots$ respectively. Then, the variable $X = (X_1, X_2, \dots)$ with values in $\prod_{i=1}^{\infty} \Omega_i$ is measurable with respect to the product σ-algebra. If, moreover, $X_1, X_2, \dots$ are independent, then the distribution μ_X of X coincides with the direct product $\prod_{i=1}^{\infty} \mu_i$, where μ_i is the disctribution of X_i.*

PROOF. The product σ-algebra is generated by cylinders, and the preimage of $A_1 \times \dots \times A_N \times \prod_{i=N+1}^{\infty} \Omega_i$, where $A_i \in \mathcal{F}_i$, is

$$\{\omega \in \Omega : X_1 \in A_1, \dots X_N \in A_N\} = \cap_{i=1}^{N} \{\omega \in \Omega : X_i(\omega) \in A_i\},$$

which is measurable because each X_i is measurable. Since cylinders form a π-system, π-λ theorem (more precisely, Corollary 1.3.5) implies that it suffices to check that μ_X argees with $\prod_{i=1}^{\infty} \mu_i$ on cylinders, which is exactly the definition of independence. □

We need some practical criteria to check independence. Those are provided by the following observation:

PROPOSITION 2.1.8. *Let $X_1, X_2, \dots$ be random variables with values in $(\Omega_1, \mathcal{F}_1), (\Omega_2, \mathcal{F}_2), \dots$, and let $\mathcal{A}_1, \mathcal{A}_2, \dots$ be π-systems such that $\sigma(\mathcal{A}_i) = \mathcal{F}_i$. If, for any $A_1 \in \mathcal{A}_1, A_2 \in \mathcal{A}_2, \dots$, the events $X_i \in A_i$ are independent, then the random variables $X_1, X_2, \dots$ are independent.*

PROOF. The condition says that the distribution μ of the random variable $X = (X_1, X_2, \dots)$ agrees with the product measure on the π-system of all sets of the form $A_1 \times \dots \times A_N \times \prod_{i=N+1}^{\infty} \Omega_i$, where $A_i \in \mathcal{A}_i$ or $A_i = \Omega_i$. Since this π-system generates the product σ-algebra, the result follows from Corollary 1.3.5. □

COROLLARY 2.1.9. *Scalar random variables $X_1, \dots, X_N$ with probability distribution functions $F_{X_1}, \dots, F_{X_N}$ are independent if an only if, for any $a_1, \dots, a_N \in \mathbb{R}$,*

$$\mathbb{P}(X_1 \leq a_1; \dots; X_N \leq a_N) = F_{X_1}(a_1) \cdot \dots \cdot F_{X_N}(a_N).$$

PROOF. This follows directly from the previous Proposition, since the sets $\{(\infty; a] : a \in \mathbb{R}\}$ form a π-system that generates Borel σ-algebra. □

In the next two criteria, we will use the following very special case of Fubuni's theorem:

LEMMA 2.1.10. *Let $(\Omega_1; \mathcal{F}_1; \mu_1), \dots, (\Omega_N; \mathcal{F}_N; \mu_N)$ be σ-finite measure spaces, and let $f_i : \Omega_i \to \mathbb{R}$ be integrable functions. Then, the function $f : \Omega \to \mathbb{R}$, where $\Omega = \prod_{i=1}^{N} \Omega_i$, defined by*

$$f(\omega_1, \dots, \omega_N) = f_1(\omega_1) \cdot \dots \cdot f_N(\omega_N)$$

is integrable w. r. t. $\mu = \mu_1 \otimes \dots \otimes \mu_N$, and

$$\int_{\Omega} f d\mu = \prod_{i=1}^{N} \int_{\Omega_i} f_i d\mu_i. \tag{2.1.1}$$

PROOF. Let $\pi_i : \Omega \to \Omega_i$ denote the natural projection, $\pi_i((\omega_1, \dots, \omega_N)) = \omega_i$. They are measurable functions, hence, $f_i \circ \pi_i$ are also measurable functions on Ω, so, their product is measurable. Denote $\Omega' = \Omega_2 \times \dots \times \Omega_N$ and $\mu' = d\mu_2 \otimes \dots \otimes d\mu_N$, and $\omega' = (\omega_2, \dots, \omega_N)$. We have

$$\int_{\Omega} |f_1(\omega_1)| \cdot \dots \cdot |f_N(\omega_N)| d\mu = \int_{\Omega_1} \left(\int_{\Omega'} |f_1(\omega_1)| \cdot \dots \cdot |f_N(\omega_N)| d\mu'(\omega') \right) d\mu_1(\omega_1) =$$
$$\int_{\Omega_1} |f_1(\omega_1)| \left(\int_{\Omega'} |f_2(\omega_2)| \cdot \dots \cdot |f_N(\omega_N)| d\mu'(\omega') \right) d\mu_1(\omega_1) =$$
$$\left(\int_{\Omega_1} |f_1(\omega_1)| d\mu_1(\omega_1) \right) \cdot \left(\int_{\Omega'} |f_2(\omega_2)| \cdot \dots \cdot |f_N(\omega_N)| d\mu'(\omega') \right),$$

where the first identity is Tonnelli's theorem, in the second one we use the $|f_1(\omega_1)|$ does not depend on ω', and hence is just a constant that can be taken out of the integral, and in the third one we note that the inner integral does not depend on ω_1. Iterating, we conclude that

$$\int_{\Omega} |f| d\mu = \prod_{i=1}^{N} \int_{\Omega_i} |f_i| d\mu_i < \infty$$

therefore, using Fubini instead of Tonnelli, we can repeat the same computation without absolute values, which gives (2.1.1). □

PROPOSITION 2.1.11. *Scalar random variables $X_1, \dots, X_N$ are independent if and only if, for any measurable functions $f_i : \mathbb{R} \to \mathbb{R}$ such that $\mathbb{E} f_i(X_i)$ exists, one has*

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

PROOF. The "if" direction follows from Corollary 2.1.9 by taking $f_i = \mathbb{I}_{(-\infty;a_i]}$. For the "only if" direction, let μ_X be a the distribution of the vector $X = (X_1, \dots, X_N) \in \mathbb{R}^N$. Note that by Proposition 2.1.7, we have $\mu_X = \mu_{X_1} \otimes \cdots \otimes \mu_{X_N}$. Then, with the notation $x = (x_1, \dots, x_n)$,

$$\mathbb{E}\left(\prod_{i=1}^N f_i(X_i)\right) = \int_{\mathbb{R}^N} \prod_{i=1}^N f_i(x_i) d\mu_X(x) = \prod_{i=1}^N \int_{\mathbb{R}} f_i(x_i) d\mu_{X_i}(x_i) = \prod_{i=1}^N \mathbb{E} f_i(X_i),$$

where the first identity is the by abstract change of variable (Proposition 1.5.9), the second one is Lemma 2.1.10, and the third one is again the abstract change of variable. □

COROLLARY 2.1.12. *If independent scalar random variables $X_1, \dots, X_N$ have densities $f_1, \dots, f_N$, then the random vector $X = (X_1, \dots, X_N)$ has density*

$$f(x_1, \dots x_n) = f_1(x_1) \cdot \dots \cdot f_N(x_N). \tag{2.1.2}$$

with respect to the N-dimensional Lebesgue measure λ^N. Conversely, if the random vector $X = (X_1, \dots, X_N)$ has a density f, and there exist integrable functions $f_i : \mathbb{R} \to \mathbb{R}_{\geq 0}$ such that (2.1.2) holds almost everywhere, then $X_1, \dots, X_N$ are independent.

PROOF. Since the variables are independent, for every set $A \subset \mathbb{R}^N$ of the form

$$A = [a_1; b_1) \times \cdots \times [a_N; b_N), \tag{2.1.3}$$

one has

$$\mathbb{P}(X \in A) = \prod_{i=1}^N \mathbb{P}(X_i \in [a_i; b_i)) = \prod_{i=1}^N \int_{[a_i;b_i]} f(x_i) dx_i = \int_A f d\lambda^N,$$

where the last equality follows from Lemma 2.1.10. Consequently, the distribution μ_X of the random vector $X \in \mathbb{R}^N$ agrees with the measure $f d\lambda^N$ on the π-system of the sets of the form (2.1.3) that generates $\mathcal{B}(\mathbb{R}^N)$. Once again, Corollary 1.3.5 shows that the two measures coincide.

Conversely, assume that the density of X has the form (2.1.2). First, Lemma 2.1.10 shows that

$$1 = \int_{\mathbb{R}^N} f d\lambda^N = \prod_{i=1}^N \int_{\mathbb{R}} f_i d\lambda =: \prod_{i=1}^N Q_i.$$

Therefore, replacing each f_i with f_i/Q_i, we may assume that $\int_{\mathbb{R}} f_i d\lambda = 1$ for every i. Then, one more application of Lemma 2.1.10 shows that

$$\mathbb{P}(X_i \leq a_i) = \int_{\mathbb{R}^{i-1} \times (-\infty;a_i] \times \mathbb{R}^{N-i}} f d\lambda^N = \int_{(-\infty;a_i]} f_i d\lambda.$$

Finally, again by Lemma 2.1.10,

$$\mathbb{P}(X_1 \leq a_1; \dots; X_N \leq a_N) = \int_{(-\infty;a_1] \times \cdots \times (-\infty;a_N]} f d\lambda^N = \prod_{i=1}^N \int_{(-\infty;a_i]} f_i d\lambda = \prod_{i=1}^N \mathbb{P}(X_i \leq a_i),$$

and we apply Corollary 2.1.9 to conclude. □

2.2. Gaussian random variables

We begin by computing the Euler-Poisson integral.

LEMMA 2.2.1. *One has*

$$\int_{\mathbb{R}} e^{-\frac{x^2}{2}} dx = \sqrt{2\pi}.$$

PROOF. The trick is to use Fubini's theorem and a bit of multi-dimensional calculus. Denote the integral by I; then we can write, by Lemma 2.1.10,

$$I^2 = \int_{\mathbb{R}} e^{-\frac{x_1^2}{2}} dx_1 \int_{\mathbb{R}} e^{-\frac{x_2^2}{2}} dx_2 = \int_{\mathbb{R}^2} e^{-\frac{(x_1^2+x_2^2)}{2}} d\lambda^2(x),$$

where $x = (x_1; x_2) \in \mathbb{R}^2$. We now evaluate the integral in polar coordinates. Recall the multi-dimensional change of variables formula: if Λ and Λ' are open subsets of $\mathbb{R}^n$, $\psi : \Lambda \to \Lambda'$ is a diffeomorphism, and $f : \Lambda' \to \mathbb{R}$ an integrable function, then

$$\int_{\Lambda'} f d\lambda^n = \int_{\Lambda} f \circ \psi |\det \psi'| d\lambda^n,$$

where $|\det \psi'|(x)$ denotes the Jacobian of ψ at x (i. e., the determinant of the matrix composed of partial derivatives). In our case, $\Lambda' = \mathbb{R}^2 \setminus [0; \infty)$, $\Lambda = (0; \infty) \times (0; 2\pi)$, and ψ is defined by

$$\psi(r, \varphi) = \begin{bmatrix} r\cos\varphi \\ r\sin\varphi \end{bmatrix}.$$

We compute the differential

$$\psi'(r, \varphi) = \begin{bmatrix} \cos\varphi & -r\sin\varphi \\ \sin\varphi & r\cos\varphi \end{bmatrix},$$

and the Jacobian $|\det \psi'(r, \varphi)| = r\cos^2\varphi + r\sin^2\varphi = r$. Consequently,

$$I^2 = \int_{(0;\infty)\times(0;2\pi)} re^{-r^2/2} d\lambda^2((r, \varphi)) = 2\pi \int_{(0;\infty)} re^{-\frac{r^2}{2}} dr = -2\pi\, e^{-\frac{r^2}{2}}\Big|_{r=0}^{\infty} = 2\pi.$$

□

DEFINITION 2.2.2. A scalar random variable with probability density function

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

is called *standard Gaussian*. A scalar random variable X is called Gaussian if there exist $\sigma > 0$ and $\mu \in \mathbb{R}$ such that $X' = \frac{1}{\sigma}(X - \mu)$ is a standard Gaussian.

REMARK 2.2.3. It is easy to see that if a random variable X has density $f(x)$, then $X - \mu$ has density $f(x + \mu)$, and αX has density $\frac{1}{\alpha} f(\frac{x}{\alpha})$. Consequently, a Gaussian random variable has distribution

$$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}.$$

This distribution is denoted by $\mathcal{N}(\mu, \sigma)$.

DEFINITION 2.2.4. A random variable X with values in $\mathbb{R}^N$ is called Gaussian (or Gaussian random vector) if, for any linear function $l : \mathbb{R}^N \to \mathbb{R}$, $l(X)$ is a Gaussian.

In other words, if $(\cdot, \cdot)$ denotes a scalar product in $\mathbb{R}^N$, the definition says that (X, V) is a Gaussian for every fixed $V \in \mathbb{R}^N$.

2.3. Weak law of large numbers

THEOREM 2.3.1. *Let $X_1, \dots, X_n$ be i. i. d. (independent, identically distributed) random variables such that $\mathbb{E}|X_1| < \infty$. Denote $S_n := X_1 + \cdots + X_n$ and $\mu := \mathbb{E}X_1$. Then, any $\varepsilon > 0$,*

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \overset{n\to\infty}{\longrightarrow} 0.$$

We begin with a simple, but fundamental estimate.

PROPOSITION 2.3.2. *Let $X_1, \dots, X_n$ be independent*[1] *random variables with $\sigma^2 = Var X_1 < \infty$, then*

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \le \frac{\sigma^2}{\varepsilon^2 n}.$$

PROOF. Assume without loss of generality that $\mu = 0$[2]; otherwise consider $\tilde{X}_i = X_i - \mu$ and note that $\left|\frac{S_n}{n} - \mu\right| > \varepsilon$ if and only if $\left|\frac{1}{n}\sum_{i=1}^n \tilde{X}_n\right| > \varepsilon$. One has

$$\mathbb{E}\left(\frac{S_n}{n}\right)^2 = \frac{1}{n^2}\mathbb{E}\left(\sum_{i=1}^n X_i\right)^2 = \frac{1}{n^2}\sum_{i=1}^n \mathbb{E}X_i^2 + \frac{1}{n^2}\sum_{i\neq j}\mathbb{E}(X_i X_j) = \frac{\sigma^2}{n},$$

since for $i \neq j$, $\mathbb{E}X_i X_j = \mathbb{E}X_i \mathbb{E}X_j = 0$. The proposition now follows by Chebyshev inequality. □

We will also need the following elementary result.

LEMMA 2.3.3. *Assume that $X \ge 0$ be a random variable with $\mathbb{E}X < \infty$. Then*

$$a\mathbb{P}(X \ge a) \overset{a\to\infty}{\longrightarrow} 0. \tag{2.3.1}$$

PROOF. By Chebyshev's inequality, for all $a > 0$,

$$a\mathbb{P}(X \ge a) \le \mathbb{E}\left(X\mathbb{I}_{X\ge a}\right).$$

The latter quantity tends to zero as $a \to \infty$ by Dominated convergence theorem: $X\mathbb{I}_{X\ge a} \to 0$ pointwise, and X is the integrable majorant. □

PROOF OF THEOREM 2.3.1. As above, we assume without loss of generality that $\mu = 0$. The strategy to prove Theorem 2.3.1 is to truncate the random variables. We write, for $L > 0$,

$$X_1 + \dots + X_n = \sum_{i=1}^n X_i \mathbb{I}_{|X_i|\le L} + \sum_{i=1}^n X_i \mathbb{I}_{|X_i|>L} =: S_n^{(\le L)} + S_n^{(>L)}.$$

the number L will be chosen later. One has

$$\mathbb{P}\left(\left|\frac{S_n}{n} - \mu\right| > \varepsilon\right) \le \mathbb{P}\left(\left|\frac{S_n^{(\le L)}}{n} - \mu\right| > \varepsilon\right) + \mathbb{P}(\exists 1 \le i \le n : X_i > L). \tag{2.3.2}$$

Let $\mu_L := \mathbb{E}X_1\mathbb{I}_{|X_1|>L}$ and $\sigma_L^2 = \mathrm{Var}\left(X_1\mathbb{I}_{|X_1|\le L}\right)$. One has

$$\sigma_L^2 \le \mathbb{E}(X_1\mathbb{I}_{|X_1|\le L})^2 \le \mathbb{E}\left(L \cdot |X_1|\right) = L\mathbb{E}|X_1|.$$

Therefore, by Proposition 2.3.2, we have

$$\mathbb{P}\left(\left|\frac{S_n^{(\le L)}}{n} - \mu_L\right| > \frac{\varepsilon}{2}\right) \le \frac{4L\mathbb{E}|X_1|}{n\varepsilon^2}.$$

The dominated convergence theorem gives $\mu_L \to \mu$ as $L \to \infty$, therefore

$$\mathbb{P}\left(\left|\frac{S_n^{(\le L)}}{n}\right| > \varepsilon\right) \le \frac{4L\mathbb{E}|X_1|}{n\varepsilon^2} \tag{2.3.3}$$

for L large enough. On the other hand, by the union bound,

$$\mathbb{P}(\exists 1 \le i \le n : |X_i| > L) \le \sum_{i=1}^n \mathbb{P}(|X_i| > L) \le n\mathbb{P}(|X_1| > L). \tag{2.3.4}$$

Taking $L = \alpha n$, where $\alpha > 0$ is a small parameter, and combining (2.3.2), (2.3.3) and (2.3.4), we get

$$\mathbb{P}\left(\left|\frac{S_n}{n}\right| > \varepsilon\right) \le \alpha \cdot \frac{4\mathbb{E}|X_1|}{\varepsilon^2} + \frac{1}{\alpha} \cdot (\alpha n \mathbb{P}(|X_1| > \alpha n))$$

[1] In fact, we only use that they are pairwise uncorrelated.

[2] Random variables with $\mathbb{E}X = 0$ are called *centered*

for n large enough. The first term can be made as small as we please by the choice of α, and the second one goes to zero as $n \to \infty$ (2.3.1). Hence, the right-hand side tends to zero. □

2.4. Large deviations.

The previous section presents a qualitative result - the probability for averages of i. i. d. random variables to deviate from their mean tends to zero. In practice, however, one is usually interested in explicit estimates on those probabilities. Proposition 2.3.2 provides a polynomial ($O(n^{-1})$) rate of convergence in the finite variance case. As we will see, under stronger, but fairly general assumptions, the convergence is in fact exponential.

THEOREM 2.4.1. *Let $X_1, \dots, X_n$ be centered i. i. d. such that $\mathbb{E}e^{\theta X_1} < \infty$ for some $\theta > 0$. Then, for every $a > 0$ and every n,*

$$\mathbb{P}\left(\sum_{i=1}^{n} X_i \geq na\right) \leq e^{-\gamma(a)n}, \tag{2.4.1}$$

where

$$\gamma(a) = \sup_{\theta>0}(a\theta - \varphi(\theta)), \tag{2.4.2}$$

and $\varphi(\theta) = \log \mathbb{E}e^{\theta X_1}$.

PROOF. One has, for any $\theta > 0$,

$$\mathbb{P}(\sum_{i=1}^{n} X_i > na) = \mathbb{P}(e^{\theta \sum_{i=1}^{n} X_i} > e^{\theta na}) \leq e^{-\theta na}\mathbb{E}(e^{\theta \sum_{i=1}^{n} X_i}) =$$

$$e^{-\theta na}\prod_{i=1}^{n}\mathbb{E}(e^{\theta X_i}) = e^{-\theta na}\left(\mathbb{E}(e^{\theta X_i}\right)^n = e^{-n(\theta a - \varphi(\theta))};$$

here we have applied Chebyshev inequality. Taking supremum over θ gives the result. □

This theorem already contains everything one needs to start working out examples. But in order to draw some general conclusions, we need to study general properties of the function φ.

LEMMA 2.4.2. *Let X be any scalar random variable. Then*

- *The set $I := \{\theta \in \mathbb{R} : \mathbb{E}e^{\theta X} < \infty\}$ is an interval;*
- *The function $f = \mathbb{E}e^{\theta X}$ is analytic inside the strip $I + i\mathbb{R}$, and its derivatives at any inner point of the strip are given by $f^{(n)}(\theta) = \mathbb{E}X^n e^{\theta X} < \infty$;*
- *The function $\varphi(\theta) = \log \mathbb{E}e^{\theta X}$ is convex on I.*

PROOF. Clearly, $0 \in I$. If $\theta_{1,2} \in I$ and $\theta_1 < \mathfrak{Re}\theta < \theta_2$, then

$$\mathbb{E}|e^{\theta X}| = \mathbb{E}\left(e^{\mathfrak{Re}\theta X}\mathbb{I}_{X\geq 0}\right) + \mathbb{E}\left(e^{\mathfrak{Re}\theta X}\mathbb{I}_{X<0}\right) \leq \mathbb{E}\left(e^{\theta_2 X}\mathbb{I}_{x\geq 0}\right) + \mathbb{E}\left(e^{\theta_1 X}\mathbb{I}_{X<0}\right) < \infty,$$

which implies the first assertion, and also the second one by Theorem 1.5.13 . If $\theta_1, \theta_2 \in I$, and $0 < \lambda < 1$, then

$$\lambda\varphi(\theta_1) + (1-\lambda)\varphi(\theta_2) = \log(\mathbb{E}e^{\theta_1 X})^{\lambda}(\mathbb{E}e^{\theta_2 X})^{1-\lambda} \geq \log \mathbb{E}e^{\lambda\theta_1 X}\cdot e^{(1-\lambda)\theta_2 X} = \varphi(\lambda\theta_1 + (1-\lambda)\theta_2).$$

The inequality is Holder's inequality with $p = \frac{1}{\lambda}$, $q = \frac{1}{1-\lambda}$. □

Now, if 0 is an interior point of I, then Lemma 2.4.2 implies that $\varphi'(0) = \mathbb{E}X = 0$ and $\varphi''(0) = \mathbb{E}X^2 =: \sigma^2 > 0$. Therefore, $\gamma(a) = \sup_{\theta>0}(a\theta - \varphi(\theta)) > 0$ for any $a > 0$ (take θ small enough). This means that the bound (2.4.1) indeed gives exponential decay (for fixed a). Moreover, for every $\varepsilon > 0$, we have $\varphi(\theta) < (\sigma^2 + \varepsilon)\frac{\theta^2}{2}$ for θ small enough; this means that for all a small enough, one has

$$\gamma(a) \geq \sup_{\theta>0}\left(a\theta - (\sigma^2 + \varepsilon)(\frac{\theta^2}{2})\right).$$

The supremum is attained at $\theta = \frac{a}{\sigma^2+\varepsilon}$ and equals $\frac{a^2}{2(\sigma^2+\varepsilon)}$. Therefore, (2.4.1) becomes

$$\mathbb{P}(\sum_{i=1}^{n} X_i > na) \leq e^{-\frac{na^2}{2(\sigma^2+\varepsilon)}}.$$

for all a small enough. This is small for $a \gg n^{-\frac{1}{2}}$, and is of order $e^{-\frac{\alpha^2}{2(\sigma^2+\varepsilon)}}$ for $a \sim \alpha n^{-\frac{1}{2}}$, - a glimpse of the future central limit theorem!

Finally, we show that for a fixed $a > 0$, the estimate (2.4.1) is, in a sence, sharp. This is done by a useful techniques called "tilting".

LEMMA 2.4.3. *Assume that $a > 0$ is such that the supremum in (2.4.2) is attained at an interior point of I. Then*

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}(\sum_{i=1}^{n} X_i \geq na) \to -\gamma(a).$$

PROOF. Note first of all that by (2.4.1), we have

$$\frac{1}{n} \log \mathbb{P}\left(\sum_{i=1}^{n} X_i \geq na\right) \leq -\gamma(a).$$

Let $\sup_{\theta>0}(a\theta - \varphi(\theta)) = a\beta - \varphi(\beta)$, where $\beta = \beta_a$ is an interior point of I. Define the random variable

$$Y = e^{\beta X_1} \cdot \ldots \cdot e^{\beta X_n},$$

by our assumptions, $\mathbb{E}Y < \infty$. Define a new probability measure $\mathbb{P}_\beta$ on the same probability space Ω by

$$\frac{d\mathbb{P}_\beta}{d\mathbb{P}} = \frac{Y}{\mathbb{E}Y};$$

recall (Lemma 1.5.5) that this means that $\mathbb{P}_\beta(A) = \mathbb{E}\left(\frac{Y}{\mathbb{E}Y}\mathbb{I}_A\right)$; denote by $\mathbb{E}_\beta$ the expectation with respect to this measure. Note that for any scalrar random variable X defined on Ω, one has

$$\mathbb{E}_\beta X = \mathbb{E}\left(\frac{Y}{\mathbb{E}Y}X\right).$$

This is true for indicator functions by definition; for simple functions by linearity of both sides; for non-negative functions by monotone convergence theorem (applied on each side), and for general function by linearity again.

In the special case $X = f(X_i)$, one has

$$\mathbb{E}_\beta f(X_i) = \mathbb{E}\left(\frac{Y}{\mathbb{E}Y}X\right) = \frac{1}{\prod_{i=1}^n \mathbb{E}e^{\beta X_i}} \mathbb{E}(f(X_i)e^{\beta X_i}) \prod_{j\neq i} \mathbb{E}e^{\beta X_j} = \frac{\mathbb{E}(f(X_i)e^{\beta X_i})}{\mathbb{E}e^{\beta X_i}}.$$

In particular, using Lemma 2.4.2, we compute

$$\mathbb{E}_\beta X_i = \frac{\mathbb{E}(X_i e^{\beta X_i})}{\mathbb{E}e^{\beta X_i}} = \varphi'(\beta) = a,$$

because the derivative of $a\theta - \varphi(\theta)$ vanishes at $\theta = \beta$. Also, X_i are independent under the new probability measure $\mathbb{P}_\beta$. Indeed,

$$\mathbb{E}_\beta\left(\prod_{i=1}^{n} f_i(X_i)\right) = \frac{1}{\mathbb{E}Y}\mathbb{E}\left(\prod_{i=1}^{n} e^{\beta X_i} f_i(X_i)\right) = \frac{1}{\mathbb{E}Y}\prod_{i=1}^{n} \mathbb{E}e^{\beta X_i} f_i(X_i) = \prod_{i=1}^{n} \mathbb{E}_\beta(f_i(X_i)).$$

Therefore, we can apply weak law of large number to conclude that for any $\varepsilon > 0$,

$$\mathbb{P}_\beta(I_{\varepsilon,n}) := \mathbb{P}_\beta\left(\left|\sum_{i=1}^{n} X_i - an\right| < \varepsilon n\right) \overset{n\to\infty}{\longrightarrow} 1.$$

On the other hand, on the event $I_{\varepsilon,n}$, one has $Y \le e^{\beta(a+\varepsilon)n}$. Therefore,

$$\mathbb{P}_\beta\left(I_{\varepsilon,n}\right) = \mathbb{E}\left(\frac{Y}{\mathbb{E}Y} I_{\varepsilon,n}\right) \le \frac{e^{\beta(a+\varepsilon)n}}{e^{n\varphi(\beta)}} \mathbb{P}\left(I_{\varepsilon,n}\right),$$

which implies that for any $\varepsilon > 0$,

$$\frac{1}{n}\log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i \ge a-\varepsilon\right) \ge \frac{1}{n}\log \mathbb{P}\left(I_{\varepsilon,n}\right) \ge \varphi(\beta) - \beta a - \beta\varepsilon + \frac{\log \mathbb{P}_\beta\left(I_{\varepsilon,n}\right)}{n} \ge -\gamma(a) - (\beta+1)\varepsilon$$

for n large enough. Taking ε to zero, we infer that for any $a' < a$,

$$\liminf_{n\to\infty} \frac{1}{n}\log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i > a'\right) \ge -\gamma(a).$$

Now, note that since $\varphi(\theta)$ is analytic and convex, it is either linear, or strictly convex (i. e., with φ' strictly increasing). The linear case is outruled by quadratic behaviour at 0 discussed above (see Exercises for justification in full generality). Therefore, for $\beta_1 \in I$ such that $\beta_1 > \beta$, one has $\varphi'(\beta_1) > a$. Then, for $a' \in (a, \varphi'(\beta_1))$, the supremum $\gamma(a') = \sup(\theta a' - \varphi(\theta))$ is attainded in the interval (β, β_1), that is, at the inner point of I. So, we can apply the above argument to get that

$$\liminf_{n\to\infty} \frac{1}{n}\log \mathbb{P}\left(\frac{1}{n}\sum_{i=1}^{n} X_i > a\right) \ge -\gamma(a').$$

We now take $\beta_1 \to \beta$, then $a' \to a$, and, by continuity of φ, $\gamma(a') \to \gamma(a)$. □

REMARK 2.4.4. With some work, using convexity of φ, one can relax the assumption that the supremum is attained in an interior point of I.

2.5. Strong law of large numbers

The weak law of large numbers is, in a sense, a question about the behaviour of *distributions of n-tuples* of random variables. In spirit of a our approach to Alice-Bob problem, we might ask a different question: first sample the the whole sequence $(X_1, X_2, \dots)$, and then ask what happens with averages $S_n/n = \sum_{i=1}^n X_i/n$ with X_i in *this particular sequence.*

THEOREM 2.5.1. *(Strong[3] law of large numbers.) Assume that X_i are i. i. d. scalar random variables with expectation μ, such that $\mathbb{E}X_1^4 < \infty$. Then, with probability 1,*

$$\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu.$$

To prove this theorem, we need the following fundamental result:

THEOREM 2.5.2. *(Borel-Cantelli lemma) Assume that $A_1, A_2, \dots$ are events on the same probability spaces such that*

$$\sum_{i=1}^{\infty} P(A_i) < \infty. \tag{2.5.1}$$

Let $N(\omega) := \#\{i : \omega \in A_i\}$. Then $\mathbb{P}(N = \infty) = 0$.

In words, if the sequence of events satisfies (2.5.1), then with probability 1, only finitely many of them occur.

[3]the use of adjective "strong", as compared to "weak", will be clarified in Section 2.7

PROOF. We have

$$\infty > \sum_{i=1}^{\infty} \mathbb{P}(A_i) = \sum_{i=1}^{\infty} \mathbb{E}\mathbb{I}_{A_i} = \mathbb{E}\left(\sum_{i=1}^{\infty} \mathbb{I}_{A_i}\right) = \mathbb{E}N$$

Therefore,$\mathbb{P}(N = \infty) = 0$. □

PROOF OF THE STRONG LAW OF LARGE NUMBERS. We assume w. l. o. g. that X_i are centered. Fix $\varepsilon > 0$. Note that

$$\mathbb{E}\left(\sum_{i=1}^{n} X_i\right)^4 = \sum_{i_1,i_2,i_3,i_4=1}^{n} \mathbb{E}(X_{i_1}X_{i_2}X_{i_3}X_{i_4}).$$

Now, we claim that the majority of terms in this sum are equal to zero. Indeed, if $i_1 \notin \{i_2, i_3, i_4\}$, then $\mathbb{E}(X_{i_1}X_{i_2}X_{i_3}X_{i_4}) = \mathbb{E}X_{i_1} \cdot \mathbb{E}(X_{i_2}X_{i_3}X_{i_4}) = 0$, and similarly for the cases $i_2 \notin \{i_1, i_3, i_4\}$ etc. Therefore, the following are are only cases where $\mathbb{E}(X_{i_1}X_{i_2}X_{i_3}X_{i_4}) \neq 0$:

- $i_1 = i_2 = i_3 = i_4$. In that case, $\mathbb{E}(X_{i_1}X_{i_2}X_{i_3}X_{i_4}) = \mathbb{E}X_1^4$; there are in total n such terms;
- $i_1 = i_2 \neq i_3 = i_4$. In that case, $\mathbb{E}(X_{i_1}X_{i_2}X_{i_3}X_{i_4}) = \left(\mathbb{E}X_1^2\right)^2 \leq \mathbb{E}X_1^4$ and there are $n(n-1)$ such terms: we have n ways to choose i_1 and $n-1$ ways to choose $i_3 \neq i_1$.
- $i_1 = i_3 \neq i_2 = i_4$ or $i_1 = i_4 \neq i_2 = i_3$; the same computation applies.

This discussion leads to

$$\mathbb{E}\left(\sum_{i=1}^{n} X_i\right)^4 = (n + 3n(n-1))\mathbb{E}X_1^4 \leq 3 \cdot \mathbb{E}X_1^4 \cdot n^2,$$

Applying Chebyshev inequality,

$$\mathbb{P}\left(\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^4 \geq \varepsilon\right) = \mathbb{P}\left(\left(\sum_{i=1}^{n} X_i\right)^4 \leq \varepsilon n^4\right) \leq \frac{\mathbb{E}\left(\sum_{i=1}^{n} X_i\right)^4}{\varepsilon n^4} \leq \frac{3 \cdot \mathbb{E}X_1^4}{\varepsilon n^2}.$$

Since $\sum_{n=1}^{\infty} \frac{1}{n^2}$ converges, we conclude by Borel-Cantelli that for every fixed ε, with probability 1, $\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)^4 < \varepsilon$ for all but finitely many n. Then, with probability 1, the same holds true for all ε:

$$\mathbb{P}\left(\bigcup_{m=1}^{\infty}\left\{\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| > \frac{1}{m} \text{ for infinitely many n}\right\}\right) \leq \sum_{m=1}^{\infty} \mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n} X_i\right| > \frac{1}{m} \text{ for infinitely many n}\right) = 0.$$

In other words, $\frac{1}{n}\sum_{i=1}^{n} X_i \to 0$ almost surely. □

2.6. Kolmogorov's zero-one law

In this section, we consider an infinite direct product $\Omega = \prod_{i=1}^{\infty} \Omega_i$ of probability spaces $(\Omega_i, \mathcal{F}_i, \mathbb{P}_i)$. This product is equipped with the product σ-algebra $\mathcal{F}$ and the product measure $\mathbb{P}$.

Let $\mathcal{F}_{>N}$ be the σ-algebra generated by the events of the form

$$\prod_{i=1}^{N} \Omega_i \times A \times \prod_{i=M+1}^{\infty} \Omega_i,$$

where $A \in \mathcal{F}_{N+1} \times \cdots \times \mathcal{F}_M$ is a cylindrical set.

DEFINITION 2.6.1. The σ-algebra $\mathcal{F}_\infty = \cap_{N=1}^{\infty} \mathcal{F}_{>N}$ is called the *tail σ-algebra*.

EXAMPLE 2.6.2. Let $X_1, X_2, \ldots$ be independent scalara random variables, and $\mu \in \mathbb{R}$. Then, the event $A = \{\frac{1}{n}\sum_{i=1}^{n} X_i \to \mu\}$ belongs to the tail σ-algebra.

PROOF. Let us first show that this event is measurable. Put $A_{m,n} = \{|\frac{S_n}{n} - \mu| < \frac{1}{m}\}$; this measurable w. r. t. $\sigma(\mathcal{F}_1 \times \cdots \times \mathcal{F}_n)$ and therefore measurable w. r. t. product measure. But

$$A = \cap_{m=1}^{\infty} \cup_{n=1}^{\infty} \cap_{i \geq n} A_{m,i},$$

therefore is is measurable. Demystifying this formula:

- $\cap_{i\geq n} A_{m,i}$ is the event that for every $i \geq n$ one has $|\frac{S_n}{n} - \mu| < \frac{1}{m}$.
- $\cup_{n=1}^{\infty} \cap_{i\geq n} A_{m,i}$ is the event that the latter property is satified for some n;
- Finally, $\cap_{m=1}^{\infty} \cup_{n=1}^{\infty} \cap_{i\geq n} A_{m,i}$ is the event that for every m, there exists n such that for $i \geq n$, one has $|\frac{S_n}{n} - \mu| < \frac{1}{m}$. This is exactly the definition of the limit.

Now note that for every n, the event A can be written in the form

$$A = \prod_{i=1}^{n} \Omega_i \times A_{n+1},$$

because the event A is not affected by changing the values of finitely many of X_i. By the same argument, A_{N+1} is measurable w. r. t. product σ-algebra on $\Omega_{N+1} \times \Omega_{N+2} \times \dots$, that is, A is $\mathcal{F}_{>N}$-measurable. □

THEOREM 2.6.3. *Any $\mathcal{F}_\infty$-measurable event has probability* 0 *or* 1.

PROOF. If A has the form $A_1 \times \dots \times A_N \times \prod_{i=N+1}^{\infty} \Omega_i$, then it is independent on any cylindrical event in $\mathcal{F}_{\geq N+1}$. Cylindrical events in $\mathcal{F}_{\geq N+1}$ form a π-system, and events independent of a given event form a λ-system (Exercise!). Therefore, A is independent on any event in $\mathcal{F}_{N+1\geq\infty}$. This implies that if A is cylindrical, then it is independent on any event $A' \in \mathcal{F}_\infty$. Since cylindrical events form a π-system and, once again, the events independent on A' form a λ-system, this shows that any event A in the product σ-algebra is independent of any event in $A' \in \mathcal{F}_\infty$. In particular, A' is independent of itself:

$$\mathbb{P}(A) = \mathbb{P}(A \cap A) = \mathbb{P}(A)^2.$$

□

2.7. Various notions of convergence of random variables

We have already seen at least two different ways to assert that two random variables are close to each other - that of almost sure convergence of the Strong law of large numbers, and Weak law of large numbers. In this sections, we systematically study these, notions of convergence, and introduce some new ones.

DEFINITION 2.7.1. Let $X, X_1, X_2, \dots$ be scalar random variables defined on the same probability space Ω. We say that

- $X_i \to X$ a. s.(X_i convergence to X *almost surely*) if there is an event E of probability 1 such that $X_i(\omega) \to X(\omega)$ for each $\omega \in E$;
- $X_i \xrightarrow{\mathcal{P}} X$ (X_i converges to X in probability), if for any $\varepsilon > 0$, $\mathbb{P}(|X_i - X| > \varepsilon) \to 0$;
- $X_i \xrightarrow{L^p} X$, (X_i converges to X in L^p)where $p \geq 1$, if $\mathbb{E}|X_i - X|^p \to 0$. The most common cases are $p = 1$ (convergence in mean) and $p = 2$ (mean-square convergence).

REMARK 2.7.2. The notions of almost sure convergence and convergence in probability makes sense for random variables with values in metric spaces.

DEFINITION 2.7.3. Let $X, X_1, X_2, \dots$ be random variables with values in the same metric space M (but not necessarily defined on the same probability space). We say that a sequence of random variables $X_1, X_2, \dots$ converges to a random variable in distribution, denoted $X_i \xrightarrow{\mathcal{D}} X$, if for any bounded continous function $f : M \to \mathbb{R}$, one has

$$\mathbb{E}f(X_i) \to \mathbb{E}f(X).$$

REMARK 2.7.4. By the abstract change of variable theorem (Proposition 1.5.9), $\mathbb{E}f(X_i) = \int_M f d\mu_{X_i}$, so, in fact, only distributions μ_{X_i} and μ_X are involved in above definition. Therefore, convergence in distribution may be viewed as a notion of convergence for (Borel) *measures* on metric spaces. It is also called weak convergence or vague convergence.

PROPOSITION 2.7.5. *(Implications between notions of convergence) There are the following implications between notions of convergence:*

- *a. s. convergence implies convergence in probability;*

- *convergence in L^p for any $p \geq 1$ implies convergence in probability;*
- *convergence in probability implies convergence in distribution.*
- *convergence in L^p implies convergence in L^q if $p > q$.*

REMARK 2.7.6. There are, in general, no other implications, see Exercises.

PROOF. *$X_i \to X$ a. s. implies $X_i \xrightarrow{\mathcal{P}} X$.* Given $\varepsilon > 0$, denote $E_{i,\varepsilon} := \{|X_i - X| < \varepsilon\}$. Then, $\cap_{i=n}^{\infty} E_{i,\varepsilon}$ is the event that the inequality $|X_i - X| < \varepsilon$ holds for all $i \geq n$, and $\cup_{n=1}^{\infty} \cap_{i=n}^{\infty} E_{i,\varepsilon}$ is contained in the event that $X_i \to X$. Therefore,

$$\mathbb{P}\left(\cup_{n=1}^{\infty} \cap_{i=n}^{\infty} E_{i,\varepsilon}\right) = 1.$$

By lower continuity of probability, this means that for any $\varepsilon > 0$

$$\mathbb{P}(E_{n,\varepsilon}) \geq \mathbb{P}(\cap_{i=n}^{\infty} E_{i,\varepsilon}) \to 1,$$

which is exactly the definition of convergence in probability.

$X_i \xrightarrow{L^p} X$ implies $X_i \xrightarrow{\mathcal{P}} X$. This is a direct consequence of Chebyshev's inequality:

$$\mathbb{P}(|X_i - X| \geq \varepsilon) \leq \frac{\mathbb{E}|X - X_i|^p}{\varepsilon^p}.$$

$X_i \xrightarrow{\mathcal{P}} X$ implies $X_i \xrightarrow{\mathcal{D}} X$. We first note that $X_i \xrightarrow{\text{a.s.}} X$ implies $X_i \xrightarrow{\mathcal{D}} X$. Indeed, if f is continuous and bounded, then $f(X_i(\omega))$ converges to $f(X(\omega))$ for almost every ω, and $f \circ X_i$ are uniformly bounded. So, Dominated convergence theorem implies $\mathbb{E}(f(X_i)) \to \mathbb{E}f(X)$.

Now we assume by contradiction that $X_i \xrightarrow{\mathcal{P}} X$, but there is a bounded continuous function f such that $\mathbb{E}f(X_i) \nrightarrow \mathbb{E}f(X)$. By passing to a subsequence, we may assume that $|\mathbb{E}f(X_i) - \mathbb{E}f(X)| > \varepsilon$ for some $\varepsilon > 0$ and all i. But now, by Proposition 2.7.7 below, we can extract further subsequence i_k such that $X_{i_k} \xrightarrow{\text{a.s.}} X$, and thus $\mathbb{E}(f(X_i)) \to \mathbb{E}f(X)$, which is a contradiction.

$X_i \xrightarrow{L^p} X$ implies $X_i \xrightarrow{L^q} X$ if $p > q$. One has

$$\mathbb{E}|X_i - X|^q = \mathbb{E}\left(|X_i - X|^q \mathbb{I}_{|X_i - X|^q \geq 1}\right) + \mathbb{E}\left(|X_i - X|^q \mathbb{I}_{|X_i - X|^q < 1}\right).$$

The first term tends to zero, because $a^q \leq a^p$ for every $a \geq 1$. For the second term, note that if $0 \leq Y \leq 1$, then for any $\varepsilon > 0$,

$$\varepsilon^p \mathbb{P}(Y > \varepsilon) \leq \mathbb{E}Y^p \leq \mathbb{E}Y^p \mathbb{I}_{Y_i \geq \varepsilon} + \mathbb{E}Y^p \mathbb{I}_{Y_i < \varepsilon} \leq \mathbb{P}(Y > \varepsilon) + \varepsilon^p.$$

Therefore, if $0 \leq Y_i \leq 1$, then

$$Y_i \xrightarrow{L^p} 0 \text{ if and only if } Y_i \xrightarrow{\mathcal{P}} 0,$$

and the last condition does not depend on p. □

The end of the proof shows that under additional conditions there might be additional implications. Let's have some more examples.

PROPOSITION 2.7.7. *If $X_i \xrightarrow{\mathcal{P}} X$, then there is a subsequence X_{i_k} such that $X_{i_k} \to X$ almost surely.*

PROOF. Let $\varepsilon_n = \frac{1}{n}$. Since $\mathbb{P}(|X_i - X| > \varepsilon_1) \to 0$, we can choose a subsequence $i_k^{(1)}$ of integers such that $\sum_{i=1}^{\infty} \mathbb{P}(|X_{i_k^{(1)}} - X| > \varepsilon_1) < \infty$; Borel-Cantelli then implies that almost surely, $|X_{i_k^{(1)}} - X| < \varepsilon_1$ for all k large enough. Similarly, from this subsequence, we can choose further subsequence $i_k^{(2)}$ such that almost surely, $|X_{i_k^{(2)}} - X| < \varepsilon_2$ for all k large enough, etc. Then, almost surely, the diagonal sequence $i_k^{(k)}$ satisfies $|X_{i_k^{(2)}} - X| < \varepsilon_n$ for all k large enough, i. e., tends to zero. □

2.8. More about convergence in distribution.

We start by giving an alternative definition of convergence in distribution for scalar random variables.

DEFINITION 2.8.1. We say that a sequence X_i of scalar random variables converges in distribution to X if

$$F_{X_i}(a) \to F(a)$$

for all $a \in \mathbb{R}$ such that F_X is continuous at a.

Before proving the equivalence of two definition, we state a useful result.

THEOREM 2.8.2. *(Skorokhod representation theorem) Suppose $X_1, X_2, \dots$ are scalar random variables such that $X_i \to X$ in distribution in the sense of Definition 2.8.1. Then there exist random variables $Y, Y_1, Y_2, \dots$, defined on a common probability space, that agree in distribution with $X, X_1, X_2, \dots$ respectively, such that*

$$Y_i \to Y \text{ a. s.}$$

PROOF. We use the construction of random variables with prescribed p. d. f., as in the proof of Lemma 1.2.11. Let $\Omega := (0,1)$ with Lebesgue measure, and put, for $\omega \in \Omega$

$$Y_i(\omega) = \inf\{x \in \mathbb{R} : F_i(x) \geq \omega\}; \quad Y(\omega) = \inf\{x \in \mathbb{R} : F_X(x) \geq \omega\},$$

where $F_i = F_{X_i}$. By Lemma 1.2.11, $Y, Y_1, Y_2, \dots$ agrees with $X, X_1, X_2, \dots$, respectively, in disctribution. We claim that if $\omega \in (0,1)$ has no more than one preimage under F_X, then $Y_i(\omega) \to Y(\omega)$.

To prove the claim, note that if ω has no more than one preimage under F_X, then $F_X(y) > \omega$ for each $y > Y(\omega)$, and $F_X(y) < \omega$ for each $y < Y(\omega)$. Since the set of discontinuity poitns of a non-decreasing function is at most countable[4], for each $\varepsilon > 0$, we can find continuity points $x_-, x_+ \in (Y(\omega) - \varepsilon; Y(\omega) + \varepsilon)$ such that $F_X(x_-) < \omega < F_X(x_+)$. Then, we have $F_i(x_-) < \omega < F_i(x_+)$ for i large enough. By definition, this means that $Y_i(\omega) \in [x_-; x_+]$ for i large enough. Since ε is arbitrary, this means that $Y_i(\omega) \to Y(\varepsilon)$.

It remailns to note that if the preimage of ω contains more than one point, then it is actually an interval. Therefore, there are at most countably many such ω. □

PROPOSITION 2.8.3. *Definitions 2.7.3 and 2.8.1 of convergence in distribution of scalar random variables are equivalent.*

PROOF. Assume that $X_i \xrightarrow{\mathcal{D}} X$ in the sense of Definition 2.8.1. Then, by Theorem 2.8.2, we may assume that they are defined on the same probability space and converge a. s.. But then

$$\mathbb{E}f(X_i) \to \mathbb{E}f(X)$$

by dominated convergence theorem. Indeed, f is continuous, therefore $f(X_i) \to f(X)$ whenever $X_i \to X$, that is, almost surely, and f is bounded, therefore $f \circ X_i$ are uniformly bounded.

Conversely, assume that $X_i \xrightarrow{\mathcal{D}} X$ in the sense of Definition 2.7.3. For $\delta > 0$, and $y \in \mathbb{R}$, define a bounded continuous function f_δ by

$$f_y(x) = \begin{cases} 1, & x \in (-\infty; y] \\ 0, & x \geq [y + \delta; \infty) \\ \text{linear}, & x \in [y; y + \delta] \end{cases}.$$

We have $f_{y-\delta} \leq \mathbb{I}_{(-\infty,y]} \leq f_y$ and ; therefore,

$$\mathbb{E}f_{y-\delta}(X_i) \leq F_{X_i}(y) \leq \mathbb{E}f_y(X_i). \tag{2.8.1}$$

Similarly,

$$F_X(y - \delta) \leq \mathbb{E}f_{y-\delta}(X) \leq F_X(y) \leq \mathbb{E}f_y(X) \leq F_X(y + \delta). \tag{2.8.2}$$

[4]Indeed, if F_X is discontinuous at $x_0 \in \mathbb{R}$, then $F(x_0 - 0) := \lim_{x \nearrow x_0} F_X(x_0) < F_X(x_0)$. Since F_X is non-decreasing, the open intervals $(F_X(x_0 - 0); F_X(x_0)) \subset (0,1)$ are disjoint for different discontinuity points x_0; since each interval contains a rational, there are at most countably many of them.

Given $\varepsilon > 0$, denote

$$I_\varepsilon := (F_X(y) - \varepsilon; F_X(y) + \varepsilon).$$

Since F_X is continuous at y, we can choose δ so small that $F_X(y-\delta) \in I_\varepsilon$ and $F_X(y+\delta) \in I_\varepsilon$. Then, by (2.8.2), $\mathbb{E}f_{y-\delta}(X) \in I_\varepsilon$ and $\mathbb{E}f_y(X) \in I_\varepsilon$. Since $\mathbb{E}f_{y-\delta}(X_i) \to \mathbb{E}f_{y-\delta}(X)$ and $\mathbb{E}f_{y-\delta}(X_i) \to \mathbb{E}f_{y-\delta}(X)$, this means that $\mathbb{E}f_{y-\delta}(X_i) \in I_\varepsilon$ and $\mathbb{E}f_y(X_i) \in I_\varepsilon$ for i large enough. Combining this with (2.8.1), we get that $F_{X_i}(y) \in I_\varepsilon$ for i large enough, which means that $F_{X_i}(y) \to F_X(y)$. □

Even if a sequence of probability distribution functions converges at every point, the limit may not be a probability distribution function (take, e. g., $X_i = i$ with probability 1). To outrule this situation, the following definition is useful:

Definition 2.8.4. A sequience X_i of scalar random variables is called *tight* if for any $\varepsilon > 0$, there exists $R > 0$ such that for any i,

$$\mathbb{P}(X_i \in [-R; R]) > 1 - \varepsilon.$$

Theorem 2.8.5. *(Helly's selection theorem)*

- *If X_i is any sequence of scalar random variables, then there is a subsequence i_k and a right-continuous non-decreasing function $F : \mathbb{R} \to (0,1)$ such that $F_{X_i}(a) \to F(a)$ for all a at which F is continuous.*
- *If, in addition, X_i is tight, then F is a disctribution function of a random variable X (and thus $X_i \xrightarrow{\mathcal{D}} X$).*

Proof. Denote $F_i := F_{X_i}$. We can construct a subsequence i_k such that $\lim_{k\to\infty} F_{i_k}(a) =: H(a) \in [0,1]$ exists for all $a \in \mathbb{Q}$[5]. Clearly, $H(a)$ is non-decreasing, but might not be right-continuous. To fix that, we put

$$F(a) := \inf_{x>a} H(x),$$

Clearly, F is non-decreasing and right-continuous. Let a be a point of continuity of F. Given $\varepsilon > 0$, let δ be such that $F(a+\delta) \le F(a) + \varepsilon$ and $F(a-\delta) \le F(a) - \varepsilon$. Then, for a rational number $q_+ \in (a; a+\delta)$, we have $\lim_{k\to\infty} F_{i_k}(q_+) = H(q_+) \le F(q_+) \le F(a) + \varepsilon$; similarly, for a rational $q_- \in (a-\delta, a)$, we have $\lim_{k\to\infty} F_{i_k}(q_-) \ge F(a) - \varepsilon$. Consequently, for k large enough,

$$F_{i_k}(a) \in (F_{i_k}(q_-); F_{i_k}(q_+)) \subset (F(a) - \varepsilon; F(a) + \varepsilon),$$

that is, $F_{i_k}(a) \to F(a)$.

If, in addition, the sequence X_i is tight, then, for any $\varepsilon > 0$, we can find $R > 0$ such that $F_{i_k}(a) < \varepsilon$ for all $a < -R$. Then $H(a) \le \varepsilon$ for all $a < -R$, therefore $F(a) \le \varepsilon$ for all $a < -R$. That is to say, $F(a) \to 0$ as $a \to -\infty$. Similarly, $F(a) \to 1$ as $a \to +\infty$. □

Remark 2.8.6. The above theorem says that the set of probability measures on $\mathbb{R}$ is compact with respect to the topology of convergence in distribution. A similar result for complete separable metric spaces is known as Prokhorov's theorem, and for compact metric spaces - as Banach-Alaoglu's theorem.

2.9. Characteristic functions

In this section, we switch gears and develop an important concept of a characteristic function of a random variable.

Definition 2.9.1. If X is a scalar random variable, then the characteristic function of X is defined as

$$\varphi_X(t) = \mathbb{E}e^{itX}, \quad t \in \mathbb{R}.$$

[5]Cantor's diagonal argument: let $a_1, a_2, \dots$ be enumeration of $\mathbb{Q}$; since $F_i(a_1) \in [0,1]$ for all i, we can choose a subsequence $i_k^{(1)}$ such that $\lim_{k\to\infty} F_{i_k^{(1)}}(a_1)$ exists; from this subsequence, extract further subsequence $i_k^{(2)}$ such that $\lim_{k\to\infty} F_{i_k^{(2)}}(a_2)$ exists, etc. Then $i_k := i_k^{(k)}$ is the required subsequence.

Since $|e^{itX}| = 1$, the expectation always exists.

The relevance of characteristic functions to the study of independent random variables is revealed by in the following proposition:

PROPOSITION 2.9.2. *If X and Y are independent, then*

$$\varphi_{X+Y}(t) \equiv \varphi_X(t)\varphi_Y(t).$$

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

The logic behind the use of this formula is that if one knows distributions of $X_1, \dots, X_n$, one can compute their characteristic functions and multiply them together to obtain the characteristic function of $S_n = X_1 + \cdots + X_n$, and then recover the distribution of S_n. Our next goal is to show that the distribution of X is uniquely determined by φ_X.

When written in terms of integral over $\mathbb{R}$, characteristic function reads

$$\varphi_X(t) = \int_{\mathbb{R}} e^{itx} d\mu_X(x).$$

Indeed, "characteristic function" is just a probabilistic name for the Fourier transform; it is convenient at this point to switch to the analysis terminology for a while.

DEFINITION 2.9.3. Given an integrable function $f : \mathbb{R} \to \mathbb{C}$, the Fourier transform $\mathfrak{F}(f)$ is defined by

$$\mathfrak{F}f(t) = \int_{\mathbb{R}} e^{itx} f(x) dx.$$

We begin with an important lemma.

LEMMA 2.9.4. *(Riemann-Lebesgue lemma) If $f : \mathbb{R} \to \mathbb{C}$ is an integrable function, then*

$$\mathfrak{F}f(t) \to 0 \text{ as } t \to \pm\infty.$$

PROOF. We first prove the lemma for continuous, compactly supported functions f. In that case, since

$$\int_{\mathbb{R}} f(x)e^{itx}dx = -\int_{\mathbb{R}} f(x)e^{it(x+\frac{\pi}{t})}dx = -\int_{\mathbb{R}} f(x-\frac{\pi}{t})e^{itx}dx,$$

we can write

$$\mathfrak{F}f(t) = \frac{1}{2}\int_{\mathbb{R}} \left(f(x) - f(x-\frac{\pi}{t})\right) e^{itx}dx.$$

Since f is compactly supported, we can find $R > 0$ such that the last integral is actually over $[-R; R]$. Since f is continuous, and therefore uniformly continuous, $|\mathfrak{F}f(t)| \leq 2R \sup_{[-R;R]} |f(x) - f(x - \frac{\pi}{t})| \to 0$ as $t \to \pm\infty$.

Now, we use the following

> Claim: for every integrable function f and for every $\varepsilon > 0$, there exists a continous, compactly supported function h such that $\int_{\mathbb{R}} |f - h| < \varepsilon$.

The lemma follows immediately, since

$$|\mathfrak{F}f(t)| = |\mathfrak{F}(f(t) - h(t)) + \mathfrak{F}h(t)| \leq \int_{\mathbb{R}} |f(t) - h(t)| + |\mathfrak{F}h(t)|,$$

where the first term can be chosen as small as we please, and the second one tends to zero.

To prove the claim, note that we can find an $R > 0$ such that

$$\int_{\mathbb{R}\setminus[-R;R]} |f| < \frac{\varepsilon}{2},$$

therefore, we may assume that $\operatorname{supp} f \subset [-R; R]$ for some $R > 0$. If suffices to prove the claim for $f = \mathbb{I}_A$, with $A \subset [-R; R]$ a measurable set, because an integrable f can be approximated by linear combinations of such functions as in the proof of Proposition 1.5.9. But for such A, the claim follows from regularity of finite Borel measures (Theorem 1.7.6): given $K \subset O$ with K compact and O open, we can define a continuous

function $0 \leq h \leq 1$ such that $h \equiv 1$ on K and $h \equiv 0$ on $\mathbb{R} \setminus O$[6]. If $K_A \subset A \subset O_A$ with $\lambda(O_A) - \lambda(K_A) < \varepsilon$, then

$$\int_{\mathbb{R}} |\mathbb{I}_A - h| = \int_{O_A \setminus K_A} |\mathbb{I}_A - h| \leq \varepsilon,$$

and the claim follows. □

PROPOSITION 2.9.5. *(Fourier inversion formula) If f is integrable and continuously differentiable, then*[7]

$$\mathfrak{F}\mathfrak{F}f(t) = 2\pi f(-t).$$

PROOF. We can write

$$\mathfrak{F}\mathfrak{F}f(-t) = \lim_{R\to\infty} \int_{[-R;R]} e^{-it\theta} \left(\int_{\mathbb{R}} e^{i\theta x} f(x) dx \right) d\theta =: \lim_{R\to\infty} I_R$$

Since $\int_{[-R;R]\times\mathbb{R}} |f(x)| dy dx \leq 2R \int_{\mathbb{R}} |f(x)| dx < \infty$, Fubini's theorem readily applies:

$$I_R = \int_{\mathbb{R}} \left(f(x) \int_{[-R;R]} e^{-it\theta} e^{i\theta x} d\theta \right) dx = \int_{\mathbb{R}} f(x) \frac{e^{i(x-t)R} - e^{-i(x-t)R}}{i(x-t)} dx = \int_{\mathbb{R}} f(x) \frac{2\sin((x-t)R)}{x-t} dx.$$

We can decompose this integral as

$$I_R = \int_{\mathbb{R}\setminus[t-1;t+1]} f(x) \frac{2\sin((x-t)R)}{x-t} dx + \int_{[t-1;t+1]} f(t) \frac{2\sin((x-t)R)}{x-t} + \int_{[t-1;t+1]} \frac{f(x)-f(t)}{x-t} 2\sin((x-t)R).$$

We claim that the first and the third integrals tend to zero as $R \to \infty$. Indeed, since $h(x) := (f(x) - f(t))/(x-t)$ is a continuous function on $[t-1; t+1]$, we have

$$\int_{[t-1;t+1]} \frac{f(x)-f(t)}{x-t} 2\sin((x-t)R) = \int_{\mathbb{R}} h(x) \mathbb{I}_{[t-1;t+1]} 2\sin((x-t)R) = -ie^{-itR} \mathfrak{F}\left(h\mathbb{I}_{[t-1;t+1]}\right)(R) + ie^{itR} \mathfrak{F}\left(h\mathbb{I}_{[t-1;t+1]}\right)(-R),$$

which tends to zero by Riemann-Lebesgue lemma. Similarly, since $|f(x)\mathbb{I}_{\mathbb{R}\setminus[t-1;t+1]}/(x-t)| \leq |f(x)|$ is integrable, Riemann-Lebesgue lemma shows that the first integral tends to zero. Therefore,

$$I_R = f(t) \int_{[t-1;t+1]} \frac{2\sin((x-t)R)}{x-t} dx + o(1) = 2f(t) \int_{-R}^{R} \frac{\sin y}{y} dy + o(1),$$

where we have made a change of variable $y = (x-t)R$. The integral $\int_{-\infty}^{\infty} \frac{\sin y}{y} dy =: C_F$ converges, therefore

$$\lim_{R\to\infty} I_R = 2C_F f(t).$$

□

REMARK 2.9.6. The value of the constant $C_F = \pi$ can be computed in a number of ways. We prefer to derive it from a computation with Gaussian densities (Example 2.10.1).

COROLLARY 2.9.7. *The distribution of a scalar random variable is uniquely determined by its characteristic function.*

---

[6] E. g., take $h(x) := 1 - \delta^{-1} \min(\text{dist}(x, K); \delta)$ for $0 < \delta < \text{dist}(K; O^c)$. Note that the latter distance is positive since K is compact and O^c is closed.

[7] If $\mathfrak{F}f$ happens to be non-integrable, its Fourier transform is defined to be the limit in the first line of the proof.

PROOF. Let f be a twice continuously differentiable, compactly supported function. By Proposition 2.9.5, we can write

$$f(x) = \frac{1}{2\pi}\int_{\mathbb{R}} e^{-itx} g(t)dt,$$

where $g = \mathfrak{F}f$. Assuming the exchange of $\mathbb{E}$ and $\int$ to be legitimate,

$$\mathbb{E}f(X) = \mathbb{E}\left(\frac{1}{2\pi}\int_{\mathbb{R}} e^{-itX} g(t)dt\right) = \frac{1}{2\pi}\int_{\mathbb{R}} \mathbb{E}e^{-itX} g(t)dt = \frac{1}{2\pi}\int_{\mathbb{R}} \varphi_X(t)g(t)dt; \tag{2.9.1}$$

this means that $\mathbb{E}f(X)$ is uniquely determined by $\varphi_X(t)$. However, for every interval $[a;b]$ and every $\varepsilon > 0$ we can finde a C^2-smooth function f_ε supported on $(a-\varepsilon, b+\varepsilon)$ such that $0 \le f_\varepsilon \le 1$ and $f_\varepsilon \equiv 1$ on $[a,b]$. Then

$$\mathbb{P}(X \in [a,b]) \le \mathbb{E}f_\varepsilon(X) \le \mathbb{P}(X \in (a-\varepsilon, b+\varepsilon)) \xrightarrow{\varepsilon\to 0} \mathbb{P}(X \in [a,b]).$$

Hence $\mathbb{P}(X \in [a,b]) = \lim_{\varepsilon\to 0} \mathbb{E}f_\varepsilon(X)$ is uniquely determined by φ_X.

It remains to justify (2.9.1). To this end, it suffices to show that $\left|e^{-itX} g(t)\right| = |g(t)|$ is integrable over $\mathbb{R}\times\Omega$, that is, that g is integrable over $\mathbb{R}$. However,

$$\mathfrak{F}(f'')(t) = \int_{\mathbb{R}} e^{itx} f''(x)dx = -\int_{\mathbb{R}} (it)\, e^{itx} f'(x)dx = (it)^2 \int_{\mathbb{R}} e^{itx} f(x)dx = (it)^2 g(t).$$

Therefore, Riemann-Lebesgue lemma implies that $t^2 g(t) \to 0$ as $t \to \pm\infty$, that is, $g(t)$ is integrable. □

REMARK 2.9.8. The above proof indicates an explicit way to compute $\mathbb{E}f(X)$ in terms of φ_X and $\mathfrak{F}f$. Although $f = \mathbb{I}_{[a,b]}$ is not twice continuously differentiable, one might still try to plug it into the fromula, and then justify the result directly by adapting the proof of Proposition 2.9.5. This is indeed possibe, albeit with additional technicalities, and leads to Lévy's inversion formula; see Willams' or Durrett's books.

COROLLARY 2.9.9. *If X is a scalar random variable, and $X_1, X_2, \dots$ is a tight sequence of scalar random variables such that $\varphi_{X_n}(t) \to \varphi_X(t)$ for all $t \in \mathbb{R}$, then $X_n \xrightarrow{\mathcal{D}} X$.*

PROOF. Assume, on the contrary, that for some $\varepsilon > 0$, there is a bounded continuous function f and a subsequence n_k of integeres such that for all k,$|\mathbb{E}f(X_{n_k}) - \mathbb{E}f(X)| > \varepsilon$. By Helly's theorem (Theorem 2.8.5), by passing to further subsequence, we may assume that $X_{n_k} \xrightarrow{\mathcal{D}} Y$ for some random variable Y. Then $\mathbb{E}f(Y) \ne \mathbb{E}f(X)$, i. e., Y does not agree in distribution with X.

On the other hand, $\varphi_Y(t) = \mathbb{E}\exp(itY) = \lim_{k\to\infty} \mathbb{E}\exp(itX_{n_k}) = \varphi_X(t)$, that is, X and Y have the same characteristic functions, and thus they agree in distribution by Corollary 2.9.7. □

REMARK 2.9.10. If X_n is a tight sequence and the limit $\varphi(t) = \lim_{n\to\infty} \varphi_{X_n}(t)$ exists for any t, then authomatically $\varphi = \varphi_X$ for some random variable X: just choose a convergent subsequence and take X to be its limit. A slightly more subtle result (Lévy's criterion) asserts that if the limit $\varphi(t)$ exists and is continous at $t = 0$, then the sequence is tight.

2.10. Explicit computations with characteristic functions

EXAMPLE 2.10.1. If $f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$ is the Gaussian density, then

$$\mathfrak{F}f(t) = \sqrt{2\pi} e^{-\frac{t^2}{2}}.$$

Hence $C_F = \pi$.

PROOF. We have

$$\int_{\mathbb{R}} e^{-\frac{x^2}{2}+itx} dx = e^{-\frac{t^2}{2}} \int_{\mathbb{R}} e^{-\frac{(x-it)^2}{2}} dx.$$

For *purely imaginary* t, the last term is just an integral of a shifted Gaussian density, and hence equals $\sqrt{2\pi}$. We need to extend this result to *real* t. Note that $|e^{-\frac{(x-it)^2}{2}}| = e^{-\Re\mathfrak{e}(x-it)^2/2} = e^{\Re\mathfrak{e}(it+t^2/2)} e^{-x^2/2}$. Since $e^{\Re\mathfrak{e}(it+t^2/2)}$ is bounded in a neighborhood of any $t_0 \in \mathbb{C}$, Theorem 1.5.13 implies that $h : t \mapsto \int_{\mathbb{R}} e^{-\frac{(x-it)^2}{2}} dx$

is an analytic function. As noted above, this function is equal to $\sqrt{2\pi}$ on the imaginary axis; therefore it is equal to $\sqrt{2\pi}$ everywhere. □

COROLLARY 2.10.2. *Let X and Y be independent random variables distributed as $\mathcal{N}(\mu_1, \sigma_1)$ and $\mathcal{N}(\mu_2, \sigma_2)$ respectively. Then $X + Y$ is distributed as $\mathcal{N}(\mu_1 + \mu_2, \sqrt{\sigma_1^2 + \sigma_2^2})$.*

PROOF. We first note some simple properties of φ_X First, for a constant a,

$$\varphi_{X+\mu}(t) = \mathbb{E}e^{it(X+\mu)} = e^{it\mu}\varphi_X(t).$$

Second,

$$\varphi_{\sigma X}(t) = \mathbb{E}e^{it\sigma X} = \varphi_X(t\sigma).$$

Since $X = \sigma_1 X' + \mu_1$ and $Y = \sigma_1 Y' + \mu_2$, where X', Y' are standard Gaussians, we get

$$\varphi_{X+Y}(t) = \varphi_X(t)\varphi_Y(t) = e^{it(\mu_1+\mu_2)}e^{-\frac{t^2\left(\sigma_1^2+\sigma_2^2\right)}{2}},$$

and the result follows. □

2.11. The Central limit theorem

THEOREM 2.11.1. *(The Central Limit Theorem)Let $X_1, \ldots, X_n$ be independent, identically distributed scalar random variables such that $\mathbb{E}X_1 = 0$ and $\mathbb{E}X_1^2 = \sigma^2 < \infty$. Then*

$$\frac{S_n}{\sqrt{n}} := \frac{\sum_{i=1}^n X_i}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma).$$

The heart of the matter is the followsing estimate.

LEMMA 2.11.2. *If $\mathbb{E}X^2 < \infty$, then*

$$\varphi_X(t) = e^{-\frac{t^2\sigma^2}{2}+o(t^2)} \quad \text{as } t \to 0.$$

PROOF. Let us compute the first two derivatives of $\varphi_X(t)$. Formally,

$$\varphi'_X(t) = \frac{\partial}{\partial t}\mathbb{E}e^{itX} = i\mathbb{E}Xe^{itX} \text{ and } \varphi''_X(t) = \frac{\partial}{\partial t}i\mathbb{E}Xe^{itX} = -\mathbb{E}X^2e^{itX}.$$

Since for all t, $|Xe^{itX}| \le |X|$ and $|X^2e^{itX}| \le |X^2|$, which are both integrable, Proposition 1.5.12 validates the computations. It follows by Taylor expansion that

$$\varphi_X(t) = \varphi_X(0) - \frac{t^2}{2}\varphi''_X(0) + o(t^2) = 1 - \frac{t^2\sigma^2}{2} + o(t^2) = e^{\log(1-\frac{t^2\sigma^2}{2}+o(t^2))} = e^{-\frac{t^2\sigma^2}{2}+o(t^2)}.$$

□

Another small input is

LEMMA 2.11.3. *The sequence $\frac{S_n}{\sqrt{n}}$ is tight.*

PROOF. This is true for any sequence Y_n of random variables such that $\operatorname{Var} Y_n$ is bonded, Indeed, given $\varepsilon > 0$, Chebyshev's inequality implies $\mathbb{P}(|Y_n| > R) \le \frac{\operatorname{Var} Y_n}{R^2} < \varepsilon$, provided that R is large enough. This is exactly the definition of a tight sequence.

It remains to notice that

$$\operatorname{Var}\left(\frac{S_n}{\sqrt{n}}\right) = \frac{\sum_{i=1}^n \mathbb{E}\left(X_i^2\right)}{n} = \frac{n\sigma^2}{n} = \sigma^2.$$

□

PROOF OF THE CENTRAL LIMIT THEOREM. One has,

$$\varphi_{S_n/\sqrt{n}}(t) = \left(\varphi_{X_1}\left(\frac{t}{\sqrt{n}}\right)\right)^n = e^{-\frac{t^2\sigma^2}{2}+o(1)} \to \varphi_{\mathcal{N}(0,\sigma)}(t).$$

Since the sequence $S_n/\sqrt{n}$ is tight, Corollary 2.9.9 implies that $S_n/\sqrt{n} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma)$. □

2.12. Heavy tails and stable distributions

In this section, we address the question as to whether one can drop the assumption $\mathbb{E}X^2 < \infty$ in the Central Limit Theorem. The answer will be in the negative; moreover, we will discover an infinite family of laws - the stable laws - that may replace the Gaussian in the statement of CLT. We will not formulate or prove general statements, but will instead focus on a particular family of examples, which exhibit most features we are interested in.

Let $\alpha \in (0,2)$, and put

$$f_\alpha(x) = \begin{cases} \frac{1}{2}\alpha|x|^{-\alpha-1}, & x \in \mathbb{R} \setminus [-1;1] \\ 0, & x \in [-1;1]. \end{cases}$$

It is easy to see that $\int_{\mathbb{R}} f_\alpha = 1$, that is, f_α is a probaility density of a random variable. Note that $\int_{\mathbb{R}} x^2 f_\alpha(x)dx = \infty$; this is due the the fact that the "tails" $\int_R^\infty f_\alpha$ decay too slowly - in Probability jargon, they are "heavy".

PROPOSITION 2.12.1. *Assume that $X_1, X_2, \dots$ are i. i. d. with density f_α. Then,*

$$\frac{S_n}{n^{1/\alpha}} \xrightarrow{\mathcal{D}} Y,$$

where Y is a random variable with characteristic function $\varphi_Y(t) = \exp(-C|t|^\alpha)$ for some $C > 0$.

PROOF. Let us start by computing the characteristic function of X_1. We have, for $t > 0$,

$$\varphi_{X_1}(t) = \int_{-\infty}^{-1} e^{itx} f_\alpha(x)dx + \int_1^\infty e^{itx} f_\alpha(x)dx = \int_1^\infty \alpha|x|^{-\alpha-1}\cos(tx)dx = \alpha t^\alpha \int_t^\infty |y|^{-\alpha-1}\cos(y)dy,$$

where the last identity is the change of variable $tx \mapsto y$. As in the proof of the Central Limit theorem, we are interested in the small t behavior of $\varphi_X(t)$. Note that

$$\int_t^\infty |y|^{-\alpha-1}\cos(y) = \int_t^1 |y|^{-\alpha-1}\cos(y) + \int_1^\infty |y|^{-\alpha-1}\cos y dy =: I_1(t) + I_2,$$

and, because $\cos y - 1 = O(y^2)$ as $y \to 0$,

$$I_1(t) = \int_t^1 |y|^{-\alpha-1} + \int_t^1 |y|^{-\alpha-1}(\cos y - 1)dy = \frac{1}{\alpha}t^{-\alpha} - \frac{1}{\alpha} + \int_0^1 |y|^{-\alpha-1}(\cos y - 1)dy + o(1) = \frac{t^{-\alpha}}{\alpha} + I_3 + o(1).$$

Putting everything together, and using that $\varphi_X(-t) = \varphi_{-X}(t) = \varphi_X(t)$ for a symmetric random variabe, we get

$$\varphi_{X_1}(t) = 1 + \alpha(I_2 + I_3)|t|^\alpha + o(|t|^\alpha) = e^{\log(1+\alpha(I_2+I_3)|t|^\alpha + o(|t|^\alpha))} = e^{\alpha(I_2+I_3)|t|^\alpha + o(|t|^\alpha)} =: e^{-C|t|^\alpha + o(|t|^\alpha)}$$

Therefore, for each t,

$$\varphi_{S_n/n^{\frac{1}{\alpha}}}(t) = \exp\left(-n\left(C\left|\frac{t}{n^{\frac{1}{\alpha}}}\right|^\alpha + o\left(\left|\frac{t}{n^{\frac{1}{\alpha}}}\right|^\alpha\right)\right)\right) = \exp\left(-C\,|t|^\alpha + o\,(1)\right).$$

Therefore, in view of Corollary 2.9.9 and Remark 2.9.10, it remains to check that the random variables S_n are tight[8]. As in the proof of the weak law of large numbers, we use truncation:

$$\mathbb{P}\left(\left|n^{-\frac{1}{\alpha}}S_n\right| > R\right) \le \mathbb{P}\left(\left|\sum_{i=1}^n X_i \mathbb{I}_{|X_i| \le n^{\frac{1}{\alpha}}R}\right| > n^{\frac{1}{\alpha}}R\right) + \mathbb{P}\left(\max_{1\le i\le n}|X_i| > n^{\frac{1}{\alpha}}R\right) =: \mathbb{P}_1 + \mathbb{P}_2$$

By the union bound,

$$\mathbb{P}_2 \le \sum_{i=1}^n \mathbb{P}(|X_i| > n^{\frac{1}{\alpha}}R) = n\alpha \int_{n^{\frac{1}{\alpha}}R}^\infty |x|^{-1-\alpha} = R^{-\alpha}.$$

[8] in fact, tightness is authomatic from Lévy's criterion, but we choose not to prove it here.

For the first term, we can calculate

$$\mathrm{Var}\left(X_1\mathbb{I}_{|X_1|\le n^{\frac{1}{\alpha}}R}\right) = \alpha\int_1^{n^{\frac{1}{\alpha}}R} x^2x^{-\alpha-1}dx \le \frac{\alpha}{2-\alpha}n^{\frac{2}{\alpha}-1}R^{2-\alpha}.$$

Therefore, Chebyshev's inequality gives

$$\mathbb{P}\left(\left|\sum_{i=1}^n X_i\mathbb{I}_{|X_i|\le n^{\frac{1}{\alpha}}R}\right| > n^{\frac{1}{\alpha}}R\right) \le \frac{\frac{\alpha}{2-\alpha}n^{\frac{2}{\alpha}}R^{2-\alpha}}{n^{\frac{2}{\alpha}}R^2} \le \frac{\alpha}{2-\alpha}R^{-\alpha},$$

and we get

$$\mathbb{P}_1+\mathbb{P}_2 \le \left(1+\frac{\alpha}{2-\alpha}\right)R^{-\alpha} \to 0$$

uniformly in n as $R\to\infty$, which shows that $n^{-\frac{1}{\alpha}}S_n$ are tight. □

Remark 2.12.2. The distribution of a random variable with characteristic function $\varphi_Y(t) = \exp(-C|t|^\alpha)$ is called *symmetric stable distribution* with parameter α. The only symmetric stable distribution for which a density is known explicitly is the Cauchy distribution, corresponding to $\alpha = 1$, and the density given by $x \mapsto \frac{\pi^{-1}}{1+x^2}$.

Remark 2.12.3. The behaviour of sums of heavy-tailed random variables is very different from the finite variance case, in that the maximum of $|X_1|,\dots,|X_n|$ is not small compared to $\sum_{i=1}^n X_i$. For $\alpha<1$, we have $n^{\frac{1}{\alpha}} \gg n$, that is, the typical order of magnitude of the sum grows faster than n; in fact, in that case the biggest term deteremines most of the sum. See exercise sheet 7 for further details.

2.13. Multi-dimensional characteristic functions and Gaussian vectors

The theory of characteristic functions of random vectors (that is, of d- dimensional random variables) more or less repeats the theory for scalar random variables.

Definition 2.13.1. Let X be a random variable with values in $\mathbb{R}^d$. Its characteristic function $\varphi_X : \mathbb{R}^d \to \mathbb{C}$ is defined as

$$\varphi_X(v) := \mathbb{E}e^{i(v;X)},$$

where $(v;X)$ denotes the scalar product of the vectors v and X.

Definition 2.13.2. If $f:\mathbb{R}^d\to\mathbb{C}$ is an integrable function, then its Fourier transorm is defined as

$$(\mathcal{F}f)(v) = \int_{\mathbb{R}^d} f(x)e^{i(v;x)}dx.$$

Theorem 2.13.3. *(Fourier inversion formula) If f is integrable, smooth function, then $\mathcal{F}f$ is integrable, and*

$$(\mathcal{F}\mathcal{F}f)(v) = (2\pi)^d f(-v).$$

Proof. The proof is similar to the one-dimensional case an will be omited. □

Corollary 2.13.4. *If X and Y are random variables with values in $\mathbb{R}^d$ such that $\varphi_X \equiv \varphi_Y$, then X and Y agree in distribution.*

Proof. Once again, the proof is similar to the one-dimensional case. □

Remark 2.13.5. Note that φ_X is completely determined by the distributions of random variables $(v;X)$, which is their turn are completely determined by the collection $\{\mathbb{P}((v;X)\le a, v\in\mathbb{R}^d,\ a\in\mathbb{R}\}$, i. e., a probability measure on $\mathbb{R}^d$ is completely determined by its values on half-planes - a result that is not so easy to prove directly.

Definition 2.13.6. A random variable X with values in $\mathbb{R}^d$ is called Gaussian (or Gaussian random vector) if, for any linear function $l:\mathbb{R}^d\to\mathbb{R}$, $l(X)$ is a Gaussian.

In other words, if $(\cdot,\cdot)$ denotes a scalar product in $\mathbb{R}^d$, the definition says that (X, v) is a Gaussian for every fixed vector $v \in \mathbb{R}^d$. In even more details, a random vector $(X_1, \dots, X_d)$ is called Gaussian if, for any $v_1, \dots, v_d \in \mathbb{R}$, the variable $v_1X_1 + \dots + v_dX_d$ is Gaussian. In that case, the variables $X_1, \dots, X_d$ are sometimes called *jointly Gaussians*.

REMARK 2.13.7. By linearity of expectation, we have $\mathbb{E}l(X) = l(\mathbb{E}X)$, therefore, if we subtract $\mathbb{E}X$ from X, all $l(X)$ become centered Gaussians.

EXAMPLE 2.13.8. If $X_1, \dots, X_d$ are i. i. d. standard Gaussians, then $(X_1, \dots, X_d)$ is a Gaussian vector.

PROOF. We calculate the characteristic function:

$$\varphi_{v_1X_1+\dots+v_dX_d}(t) = \varphi_{v_1X_1}(t) \cdot \dots \cdot \varphi_{v_dX_d}(t)$$
$$= \varphi_{X_1}(v_1t) \cdot \dots \cdot \varphi_{X_d}(v_dt) = e^{-\frac{v_1^2t^2}{2}} \cdot \dots \cdot e^{-\frac{v_d^2t^2}{2}} = e^{-\frac{(v_1^2+\dots+v_d^2)t^2}{2}},$$

hence $v_1X_1 + \dots + v_dX_d$ is a Gaussian with variance $v_1^2 + \dots + v_d^2$. □

LEMMA 2.13.9. *If X is an $\mathbb{R}^d$-valued Gaussian vector, and $A : \mathbb{R}^d \to \mathbb{R}^{d'}$ is a linear map, then AX is a Gaussian vector.*

PROOF. For any $v \in \mathbb{R}^{d'}$, $(AX; v) = (X; A^*v)$, where $A^* : \mathbb{R}^{d'} \to \mathbb{R}^d$ is the adjoint[9] operator. Therefore, $(AX; v)$ is a Gaussian. □

DEFINITION 2.13.10. (Covariance form) Let X be a random variable with values in $\mathbb{R}^d$ such that $\mathbb{E}|X|^2 < \infty$. If X is centered (i. e., $\mathbb{E}X = 0$), then its covariance form $\mathrm{Cov}_X : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ is defined by

$$\mathrm{Cov}_X(v_1, v_2) = \mathbb{E}\left((X; v_1) \cdot (X; v_2)\right).$$

In general, if X is non-centered, we define $\mathrm{Cov}_X := \mathrm{Cov}_{X-\mathbb{E}X}$.

REMARK 2.13.11. If X is a random varaible with values in a finitely-dimensional vector space V, it is more natural to define the covariance form as a bilinear form on the conjugate space V^*, given by

$$\mathrm{Cov}_X(l_1, l_2) = \mathbb{E}(l_1(X - \mathbb{E}X) \cdot l_2(X - \mathbb{E}X))$$

for any pair l_1, l_2 of linear functions on V. This way, the definition does not depend on the choice of scalar product in V. We will, however, assume the scalar product to be fixed.

A general result in linear algebra states that to the bilinear form Cov_X one can associate a linear operator $\Sigma_X : \mathbb{R}^d \to \mathbb{R}^d$ such that

$$\mathrm{Cov}_X(v_1, v_2) = (\Sigma_X v_1, v_2).$$

The operator Σ_X is called the *covariance operator*, and its matrix is called a *covariance matrix*. Plugging in the basis vectors as v_1, v_2 in the definition, we get an explicit expression for the matrix elements: if $\mathbb{E}X = 0$, then

$$\Sigma_X = \begin{pmatrix} \mathbb{E}(X_1X_1) & \dots & \mathbb{E}(X_1X_n) \\ \vdots & \ddots & \vdots \\ \mathbb{E}(X_nX_1) & \dots & \mathbb{E}(X_nX_n) \end{pmatrix},$$

where $X = (X_1, \dots, X_n)$. In general,

$$\Sigma_X = \begin{pmatrix} \mathbb{E}((X_1 - \mu_1)(X_1 - \mu_1)) & \dots & \mathbb{E}((X_1 - \mu_1)(X_n - \mu_n)) \\ \vdots & \ddots & \vdots \\ \mathbb{E}((X_n - \mu_n)(X_1 - \mu_1)) & \dots & \mathbb{E}((X_n - \mu_n)(X_n - \mu_n)) \end{pmatrix},$$

[9]Recall that by definition, the adjoint operator is defined by condition $(Av; w) \equiv (v; A^*w)$. Its matrix is just the transpose of the matrix of A.

where $\mu = \mathbb{E}X$. As in the one-dimensional case, we can calculate

$$\begin{aligned}\mathrm{Cov}_X(v_1, v_2) = \mathbb{E}\left((X-\mu; v_1)\cdot(X-\mu; v_2)\right) = \\ \mathbb{E}\left((X; v_1)\cdot(X; v_2)\right) - \mathbb{E}(X; v_1)\cdot(\mu; v_2) - \mathbb{E}(\mu; v_1)(X; v_2) + \mathbb{E}(\mu; v_1)\cdot(\mu; v_2) = \\ \mathbb{E}\left((X; v_1)\cdot(X; v_2)\right) - (\mu; v_1)\cdot(\mu; v_2).\end{aligned}$$

We note that for any random variable X, the covariance operator Σ_X is self-adjoint (that is, $\Sigma_X^* = \Sigma_X$) and non-negative definite (that is, $(\Sigma_X v; v) \geq 0$ for any $v \in \mathbb{R}^n$. Equivalently, all the eighenvalues of Σ are non-negative).

PROPOSITION 2.13.12. *(Classification of Gaussian vectors)*

(1) *Given a non-negative self-adjoint operator Σ, there exist a centered Gaussian vector X such that $\Sigma_X = \Sigma$;*
(2) *If X, Y are two centered Gaussian vectors with $\Sigma_X = \Sigma_Y$, then the vectors X and Y have the same distribution;*
(3) *If X is a centered Gaussian with values in $\mathbb{R}^d$, such that $\Sigma := \Sigma_X$ is positive definite*[10]*, then X has a density with respect to the d-dimensional Lebesgue measure given by*

$$f_\Sigma(v) := \frac{1}{(2\pi)^{\frac{d}{2}}\left(\det\Sigma\right)^{\frac{1}{2}}} e^{-(\Sigma^{-1}v; v)/2} \tag{2.13.1}$$

PROOF. If Σ is a non-negative self-adjoint operator, then it has a square root $\Sigma^{\frac{1}{2}}$, a non-negative self-adjoint operator such that $(\Sigma^{\frac{1}{2}})^2 = \Sigma$. The square root can be constructed as follows: since Σ is self-adjoint, it can be diagonalized by an orthogonal transformation, that is, there is a matrix C such that $C^{-1} = C^*$ and

$$\Sigma = C^{-1}DC,$$

where $D = \mathrm{diag}(\lambda_1, \dots, \lambda_d)$ is a diagonal matrix. Since Σ is non-negative, all λ_i are non-negative. Therefore, we can put $D^{\frac{1}{2}} = \mathrm{diag}(\lambda_1^{\frac{1}{2}}, \dots, \lambda_d^{\frac{1}{2}})$ and $\Sigma^{\frac{1}{2}} = C^{-1}D^{\frac{1}{2}}C$, in which case

$$\Sigma^{\frac{1}{2}}\cdot\Sigma^{\frac{1}{2}} = C^{-1}D^{\frac{1}{2}}C\cdot C^{-1}D^{\frac{1}{2}}C = C^{-1}DC = \Sigma,$$

as required.

Now, let $Y = (Y_1, \dots, Y_d)$ be a Gaussian vector whose components are i. i. d. standard Gaussians, and put $X = \Sigma^{\frac{1}{2}}Y$. Then, by Lemma 2.13.9, X is a Gaussian vector. Also, by the linearity of the expectation, $\mathbb{E}X = \mathbb{E}\Sigma^{\frac{1}{2}}Y = \Sigma^{\frac{1}{2}}\mathbb{E}Y = 0$. We compute

$$\mathrm{Cov}_X(v_1, v_2) = \mathbb{E}((X, v_1)\cdot(X; v_2)) = \mathbb{E}((Y; \Sigma^{\frac{1}{2}}v_1)\cdot(Y; \Sigma^{\frac{1}{2}}v_2)) = (\Sigma^{\frac{1}{2}}v_1, \Sigma^{\frac{1}{2}}v_2) = (\Sigma v_1; v_2),$$

where we used that $\Sigma^{\frac{1}{2}}$ is self-adjoint, and that the covariance operator of Y is the identity. This concludes the proof of the first part.

For the second part, note that for every fixed $v \in \mathbb{R}^d$, $(X; v)$ and $(Y; v)$ have the same distribution, since they are both centered Gaussians with variance $(\Sigma_X v; v) = (\Sigma_Y v, v)$. Therefore, $\varphi_X \equiv \varphi_Y$, and by Corollary 2.13.4, they agree in distribution.

For the third part, note that if Σ_X is positive, then it is invertible (since all eigenvalues are strictly positive). In particular, $\Sigma^{\frac{1}{2}}$ is a diffeomorphism of $\mathbb{R}^d$ onto itself, with Jacobian identically equal to $\det\Sigma^{\frac{1}{2}}$. By the second part, X agrees in distribution with $\Sigma^{\frac{1}{2}}Y$, where $Y = (Y_1, \dots, Y_d)$ and $Y_1, \dots, Y_d$ are i. i. d. standard Gaussian. Note that by Corollary 2.1.12, Y as the density

$$f_Y(y) = \frac{1}{\sqrt{2\pi}}e^{-\frac{y_1^2}{2}}\cdot\ldots\cdot\frac{1}{\sqrt{2\pi}}e^{-\frac{y_d^2}{2}} = \frac{1}{(2\pi)^{\frac{d}{2}}}e^{-\frac{(y;y)}{2}}.$$

Therefore, for every continuous function g with bounded support,

$$\mathbb{E}g(X) = \mathbb{E}g(\Sigma^{\frac{1}{2}}Y) = \frac{1}{(2\pi)^{\frac{d}{2}}}\int_{\mathbb{R}^d} g(\Sigma^{\frac{1}{2}}y)e^{-\frac{(y;y)}{2}}d\lambda^d(y) = \frac{1}{(2\pi)^{\frac{d}{2}}}\int_{\mathbb{R}^d} g(x)e^{-(\Sigma^{-\frac{1}{2}}x;\Sigma^{-\frac{1}{2}}x)/2}\det\Sigma^{-\frac{1}{2}}d\lambda^n(x),$$

[10] that is, $(\Sigma v; v) = 0$ implies $v = 0$

where we have used the multi-dimensional change of variables $y = \Sigma^{-\frac{1}{2}}x$. This implies that

$$f(x) = \frac{1}{(2\pi)^{\frac{d}{2}}} \det \Sigma^{-\frac{1}{2}} e^{-(\Sigma^{-\frac{1}{2}}x;\Sigma^{-\frac{1}{2}}x)/2} = \frac{1}{(2\pi)^{\frac{d}{2}} (\det \Sigma)^{\frac{1}{2}}} e^{-(\Sigma^{-1}v;v)/2},$$

is the density of X, as required. □

REMARK 2.13.13. This, in particular, implies that if $(X_1, \dots, X_d)$ is a Gaussian vector, and X_i are uncorrelated, then they are independent. Warning: it is not enough to assume that they are individually Gaussians, one does need the "jointly Gaussian" assumption.

2.14. Random walks

Let $X_1, X_2, \dots$ be i. i. d. random variables with values in $\mathbb{Z}^d$. One can view the sum $S_n = \sum_{i=1}^n X_i$ as a discrete time stochastic process: S_n is a position of a particle at time n, and the next moment of time it moves to $S_n + X_{n+1}$. In this context, one says that S_n is a *random walk* on $\mathbb{Z}^d$. The particular case when $\mathbb{P}(X_i = \pm e_j) = \frac{1}{2d}$, where e_j are unit orts, is called the *simple random walk*. In this case, at each step, the particle jumps into a lattice neighbor of its current position, chosen uniformly at random among all $\frac{1}{2d}$ neighbors.

One of the basic questions one asks about random walks is that of recurrence. Given $x \in \mathbb{Z}^d$, define τ_x to be the (random) time of the first visit of the walk to x:

$$\tau_x := \min\{n : S_n = x\}.$$

DEFINITION 2.14.1. A random walks is called *recurrent* if $\mathbb{P}(\tau_0 < \infty) = 1$, and *transient* otherwise.

LEMMA 2.14.2. *A random walks is recurrent if and only if $\#\{n : S_n = 0\} = \infty$ almost surely. A random walk is transient if and only if $|S_n| \to \infty$ almost surely.*

PROOF. The argument is based on the simple observation that from the time τ_x on, we can view the random walk S_n as a new random walk launched from x, independent of its past. Let us make a rigorous proof.

Clearly, if $\#\{n : S_n = 0\} > 0$, then $\tau_0 < \infty$. If $\mathbb{P}(\tau_0 < \infty) = 1$, define $\tau^{(1)} := \tau_0$, and, inductively, $\tau^{(k)} = \min\{n > \tau^{(k-1)} : S_n = 0\}$ the time of k-th visit to the origin. We have, for all $n \in \mathbb{N}$

$$\mathbb{P}(\tau^{(2)} - \tau^{(1)} = n) = \sum_{m=1}^{\infty} \mathbb{P}(\tau^{(2)} - \tau^{(1)} = n, \tau^{(1)} = m) = \sum_{m=1}^{\infty} \mathbb{P}\left(\tau^{(1)} = n\right) \mathbb{P}(\tau^{(1)} = m) = \mathbb{P}(\tau^{(1)} = n)$$

In the second identity, we have used that we can write

$$\{\tau^{(2)} - \tau^{(1)} = n, \tau^{(1)} = m\} = \{A, \tau^{(1)} = m\},$$

where A is the event that $\sum_{i=m+1}^{m+n} X_i = 0$, but $\sum_{i=m+1}^{m+n'} X_i \neq 0$ for any $1 \leq n' < n$. This event depends only on $X_{m+1}, \dots, X_{m+n}$, whereas the event $\tau^{(1)} = m$ depends only on $X_1, \dots, X_m$, hence they are independent. Also, $\mathbb{P}(A) = \mathbb{P}(\tau^{(1)} = n)$.

The conclusion of this computation is that $\tau^{(2)} - \tau^{(1)}$ has the same distribution as $\tau^{(1)}$, in particular, $\tau^{(2)}$ is almost surely finite. Repeating the same argument, we get that, inductively, $\tau^{(k+1)} - \tau^{(k)}$ all have the same distribution[11], in particular, $\tau^{(k)} < \infty$ for all k almost surely, that is, $\#\{n : S_n = 0\} = \infty$ almost surely.

For the second assertion, first note that if $|S_n| \to \infty$ almost surely, then $\#\{n : S_n = 0\}$ is finite almost surely, that is, the random walk is not recurrent. Assume now that S_n is transient. Then, for every $x \in \mathbb{Z}^d$

[11] a closer examination of the argument shows that they are also independent

and every $m = 1, 2, \dots$, and using the same independence observation as above,

$$\mathbb{P}(\#\{n : S_n = x\} = m) = \sum_{i=1}^{\infty} \mathbb{P}(\#\{n : S_n = x\} = m, \tau_x = i) =$$

$$= \mathbb{P}(\#\{n : S_n = 0\} = m-1) \sum_{i=1}^{\infty} \mathbb{P}(\tau_x = i) = \mathbb{P}(\#\{n : S_n = 0\} = m-1)\mathbb{P}(\tau_x < \infty).$$

Summing over $m = 0, 1, \dots$ we get that

$$\mathbb{P}(\#\{n : S_n = x\} < \infty) = \mathbb{P}(\tau_x = \infty) + \mathbb{P}(\tau_x < \infty)\mathbb{P}(\#\{n : S_n = 0\} < \infty) = \mathbb{P}(\tau_x = \infty) + \mathbb{P}(\tau_x < \infty) = 1.$$

This means that almost surely, every lattice cite is visited by S_n only finitely many times, that is, $|S_n| \to \infty$ almost surely. □

LEMMA 2.14.3. *If $\mu := \mathbb{E}X_i \neq 0$, then the random walk S_n is transient*[12].

PROOF. If $d = 1$, then, by the Strong law of large numbers,

$$|\frac{1}{n}S_n - \mu| > \frac{|\mu|}{2}$$

for finitely many n only. This implies that $S_n = 0$ for finitely many n only, that is, S_n is transient.

If $d > 1$, then we may view each coordinate of S_n as a 1-dimensional random walk, and at least one of those walks has steps with non-zero mean. □

In the case $\mu = 0$, the random walk is transient in dimension $d \geq 3$ and recurrent in dimensions $d = 1, 2$. This result, knowns as Polya's theorem, will follow (under certain technical conditions) from the next Lemma and the Local Central limit theorem of the next chapter.

LEMMA 2.14.4. *The random walk S_n is recurrent if and only if*

$$\sum_{n=1}^{\infty} \mathbb{P}(S_n = 0) = +\infty.$$

PROOF. Denote $p_n := \mathbb{P}(\tau_0 = n)$ and $q_n := \mathbb{P}(S_n = 0)$. We define the *generating functions* of these sequence by

$$P(t) := \sum_{n=1}^{\infty} p_n t^n, \quad Q(t) := \sum_{n=0}^{\infty} q_n t^n.$$

Since the coefficients are bounded, the radius of convergence of both series is at least 1, that is, they are well-defined at least for $|t| < 1$. Now, we compute, for $n \geq 1$,

$$q_n = \mathbb{P}(S_n = 0) = \sum_{i=1}^{n} \mathbb{P}(S_n = 0, \tau_0 = i) = \sum_{i=1}^{n} \mathbb{P}\left(\sum_{j=i+1}^{n} X_j = 0, \tau_0 = i\right)$$

$$= \sum_{i=1}^{n} \mathbb{P}\left(\sum_{j=i+1}^{n} X_j = 0\right) \mathbb{P}(\tau_0 = i) = \sum_{i=1}^{n} p_{n-i} q_i,$$

where we once again used the fact that the event $\tau_0 = i$ is determined by the values of $X_1, \dots, X_i$ and is thus independent of the event $\sum_{j=i+1}^{n} X_j = 0$. We multiply the last identity by t^n and sum over n. Taking into account that $q_0 = 0$, we get

$$Q(t) - 1 = \sum_{n=1}^{\infty} q_n t^n = \sum_{n=1}^{\infty} t^n \left(\sum_{i=1}^{n} p_{n-i} q_i\right) = \sum_{m=1}^{\infty} \sum_{n=0}^{\infty} p_m q_n t^{n+m} = P(t)Q(t),$$

[12]Formally, we should assume $\mathbb{E}|X_i|^4 < \infty$, since the Strong law of large numbers was proven under this assumption only. This assumption is not necesssary.

that is,

$$P(t) = 1 - \frac{1}{Q(t)}, \quad |t| < 1.$$

Also, monotone convergence theorem implies that

$$\mathbb{P}(\tau_0 < \infty) = \sum_{n=1}^{\infty} p_n = P(1) = \lim_{t \nearrow 1} P(t)$$

and

$$\sum_{n=o}^{\infty} \mathbb{P}(S_n = 0) = \sum_{n=0}^{\infty} q_n = Q(1) = \lim_{t \nearrow 1} Q(t).$$

Therefore, $Q(1) = \infty$ if and only if $P(1) = 1$, as required. □

With this lemma, Polya's theorem can be explained as follows. If the variance $\operatorname{Var}|X_i|$ is finite, then $\operatorname{Var}|S_n| \sim C \cdot n$, that is, roughly speaking, the distribution of S_n is supported on the area of diameter $\approx \sqrt{n}$ around the origin. There are $\approx n^{\frac{d}{2}}$ lattice points in this area, and if one believes that S_n is more ore less uniformly distributed over its support, we should have $\mathbb{P}(S_n = 0) \approx n^{-\frac{d}{2}}$. Thus, the series in Lemma 2.14.4 indeed should converge for $d \geq 3$ and diverge for $d = 1, 2$. In the next section, we make this heuristics precise.

2.15. Local central limit theorem.

Let $X_1, X_2, \dots$ be i. i. d., centered scalar random variables taking integer values, such that $\mathbb{E}X_i = 0$ and $\mathbb{E}X_i^2 =: \sigma^2 < \infty$. The Central limit theorem asserts that, with the notation $S_n = \sum_{i=1}^n X_i$,

$$\frac{S_n}{\sqrt{n}} \xrightarrow{\mathcal{D}} \mathcal{N}(0, \sigma).$$

In other words, for any $a < b$,

$$\sum_{a\sqrt{n} \leq m \leq b\sqrt{n}} \mathbb{P}(S_n = m) \overset{n\to\infty}{\longrightarrow} \frac{1}{\sqrt{2\pi}\sigma} \int_a^b e^{-\frac{x^2}{2\sigma^2}} dx \tag{2.15.1}$$

This is a statement about collective behaviour of probabilities of the form $\mathbb{P}(S_n = m)$ (i. e., behaviours of sums of $\sim \sqrt{n}$ of these probabilities). It is natural to look into the individual behaviour of each of these probabilities. What can be said about $\mathbb{P}(S_n = m)$? Assuming that (2.15.1) is a Riemann sum type approximation of the integral, one could expect

$$\mathbb{P}(S_n = m) \sim \frac{1}{\sqrt{2\pi n}\sigma} e^{-\frac{m^2}{2n\sigma^2}}, \quad n \to \infty.$$

The following simple example shows that the reality is slightly more complicated:

Example 2.15.1. Let $X_1, X_2, \dots$ be i. i. d. with $\mathbb{P}(X_i = 1) = \mathbb{P}(X_i = -1) = \frac{1}{2}$. Then

$$\mathbb{P}(S_n = 0) = 0$$

for odd n.

Indeed, the sum of an odd number of ± 1 is odd, and thus cannot be equal to zero. The following theorem states that the algebraic obstructions of these type are the only possible obstructions. For a random variable X with values in $\mathbb{Z}^d$, denote

$$\operatorname{supp}_X := \{x \in \mathbb{Z}^d : \mathbb{P}(X = x) \neq 0\}.$$

Theorem 2.15.2. *Let , $X, X_1, X_2, \dots$ be i. i. d. with values in $\mathbb{Z}^d$. Denote by $\mu := \mathbb{E}X$ their expectation and by $\Sigma = \Sigma_X$ the covariance operator. Assume that $\mathbb{E}|X|^3 < \infty$, and that the set $\{x - y : x, y \in supp_X\}$ generates $\mathbb{Z}^d$ as an abelian group. Then, for any sequence $z_1, z_2, \dots$ of points of $\mathbb{Z}^d$ such that*

$$z_n = \mu n + a\sqrt{n} + o(\sqrt{n}), \quad n \to \infty$$

for some $a \in \mathbb{R}^d$, we have

$$n^{\frac{d}{2}} \cdot \mathbb{P}(S_n = z_n) \overset{n\to\infty}{\longrightarrow} f_\Sigma(a),$$

where f_Σ is the d-dimensional Gaussian density with covariance Σ (see (2.13.1)).

COROLLARY 2.15.3. *(Polya's theorem) If S_n is a random walk whose steps X_i satisfy the conditions of Theorem 2.15.2 with $\mu = 0$, then S_n is recurrent in dimensions $d = 1, 2$ and transient in dimensions $d \geq 3$.*

PROOF. This follows directly from Lemma 2.14.4 and the fact that $\sum_{n=1}^{\infty} n^{-\frac{d}{2}}$ diverges for $d = 1, 2$ and converges for $d \geq 3$. □

REMARK 2.15.4. The assumption on supp_X, in particular, implies that Σ is positive definite, that is, that f_Σ is well defined. Indeed, assume that there is a non-zero vector $v \in \mathbb{R}^d$ such that $(\Sigma v; v) = 0$. This means that $\mathbb{E}(X - \mu; v)^2 = 0$, that is, $(X - \mu; v) = 0$ almost surely. Then for every $x, y \in \mathrm{supp}_X$, we have $(x - y; v) = 0$. In that case, the abelian group generated by $\{x - y : x, y \in \mathrm{supp}_X\}$ is contained in the hyperplane $\{w : (w; v) = 0\}$.

We start the proof of Theorem 2.15.2 with an observation concerning characteristic functions of random variables with values in $\mathbb{Z}^d$.

LEMMA 2.15.5. *(Fourier inversion for discrete random variables) Let X be a random variable with values in $\mathbb{Z}^d$. Then φ_X is 2π-periodic with respect to each coordinate, that is, $\varphi_X(v) = \varphi_X(v + 2\pi w)$ for any $v \in \mathbb{R}^d$ and $w \in \mathbb{Z}^d$. Moreover,*

$$\mathbb{P}(X = x) = \frac{1}{(2\pi)^d} \int_{[-\pi;\pi]^d} \varphi_X(v) e^{-i(v;x)} d\lambda^d(v).$$

PROOF. We have

$$\varphi_X(v + 2\pi w) = \mathbb{E} e^{i(v+2\pi w;X)} = \mathbb{E}\left(e^{i(v;X)} \cdot e^{2\pi i(w;X)}\right) = \mathbb{E}\left(e^{i(v;X)}\right) = \varphi_X(v),$$

since both X and w are in $\mathbb{Z}^d$ and hence $(\omega; X) \in \mathbb{Z}$. We have

$$e^{-i(v;x)} \varphi_X(v) = e^{-i(v;x)} \sum_{y\in\mathbb{Z}^d} e^{i(v;y)} \mathbb{P}(X = y) = \sum_{y\in\mathbb{Z}^d} e^{i(v;y-x)} \mathbb{P}(X = y).$$

We are going to integrate this with respect to v over $[-\pi; \pi]^d$. In order to justify the interchange of the sum and the integral, consider the function $g : [-\pi; \pi]^d \times \mathbb{Z}^d$ defined by $g(v, y) = e^{i(v;y-x)} \mathbb{P}(X = y)$, and note that

$$\int_{[-\pi;\pi]^d \times \mathbb{Z}^d} |g| = \int_{v\in[-\pi;\pi]^d} \sum_{y\in\mathbb{Z}^d} |g(v, y)| = \int_{[-\pi;\pi]^d} \sum_{y\in\mathbb{Z}^d} \mathbb{P}(X = y) = \int_{[-\pi;\pi]^d} 1 = (2\pi)^d < \infty,$$

therefore, Fubini's theorem readily applies, and we get

$$\int_{[-\pi;\pi]^d} e^{-(v;x)} \varphi_X(v) d\lambda^d(v) = \sum_{y\in\mathbb{Z}^d} \mathbb{P}(X = y) \int_{[-\pi;\pi]^d} e^{i(v;y-x)} d\lambda^d(v).$$

It remains to notice that

$$\int_{[-\pi;\pi]^d} e^{i(v;z)} d\lambda^d(v) = \prod_{m=1}^{d} \int_{-\pi}^{\pi} e^{iv_m z_m} dv_m = \begin{cases} (2\pi)^d, & z = 0 \\ 0, & \text{otherwise.} \end{cases}$$

□

From this, we deduce, in particular, that

$$\mathbb{P}(S_n = z_n) = \frac{1}{(2\pi)^d} \int_{[-\pi;\pi]^d} \varphi_{S_n}(v) e^{-(v;x)} d\lambda^d(v) = \frac{1}{(2\pi)^d} \int_{[-\pi;\pi]^d} (\varphi_{X_1}(v))^n \, e^{-i(v;z_n)} d\lambda^d(v). \tag{2.15.2}$$

Thus, we are left with the asymptotics of this integral. First, we transform the algebraic assumption on supp_X into an analytic condition on φ_X.

LEMMA 2.15.6. *If the set $\{x-y : x,y \in supp_X\}$ generates $\mathbb{Z}^d$ as an abelian group, then $|\varphi_X(v)| < 1$ for all $v \notin 2\pi\mathbb{Z}^d$.*

PROOF. Indeed, assume that

$$|\varphi_X(v)| = \left| \sum_{x\in\mathbb{Z}^d} e^{i(v;x)}\mathbb{P}(X=x) \right| = 1.$$

However, in general,

$$\left| \sum_{x\in\mathbb{Z}^d} e^{i(v;x)}\mathbb{P}(X=x) \right| \le \sum_{x\in\mathbb{Z}^d} |e^{i(v;x)}\mathbb{P}(X=x)| = 1,$$

and the equality is possible only if for all $x \in \operatorname{supp}_X$, the arguments of $e^{i(v;x)}$ are the same. This means that $e^{i(v;x-y)} = 1$ for all $x,y \in \operatorname{supp}_X$, that is, $(v;x-y) \in 2\pi\mathbb{Z}$ for all $x,y \in \operatorname{supp}_X$. Then, we have $(v;\alpha) \in 2\pi\mathbb{Z}$ for all $\alpha \in \mathbb{Z}^d$, and, taking α to be unit orts, we conclude that $v_1 \in 2\pi\mathbb{Z}, \dots, v_d \in 2\pi\mathbb{Z}$, i. e., $v \in 2\pi\mathbb{Z}^d$. □

The asymptotics is done by what's called Laplace's method. The above Lemma, in particular, implies that

$$\left| \int_{[-\pi;\pi]^d\setminus(-\varepsilon;\varepsilon)^d} (\varphi_X(v))^n e^{-i(v;z_n)} d\lambda^d(v) \right| \le (2\pi)^d \alpha^n,$$

where $\alpha = \sup_{[-\pi;\pi]^d\setminus(-\varepsilon;\varepsilon)^d} |\varphi_X| < 1$. Since this decays exponentially fast (and we are aiming at much slower power-law decay in the statement of the theorem), only the integral over an (arbitrary small) vicinity of $v = 0$ matters. The idea of Laplace's method is to make this vicinity shrink with n (that is, to allow $\varepsilon = \varepsilon_n$ depend on n), and use Taylor approximation of φ_X. We start by working out the approximation.

LEMMA 2.15.7. *Under the condition of Theorem 2.15.2, we have*

$$\varphi_X(v) = e^{i(v;\mu)-(\Sigma v;v)/2+\Theta(v)},$$

where the error term $\Theta(v)$ satisfies $|\Theta(v)| \le C|v|^3$ for all v with $|v|$ small enough.

PROOF. Fix a vector v with $|v| = 1$, and consider the function $\psi_v(t) := \varphi_X(tv)$, where $t \in \mathbb{R}$. Let us calculate the derivatives of ψ_v. We have

$$\psi_v'(t) = \frac{\partial}{\partial t}\mathbb{E}e^{it(v;X)} = \mathbb{E}\frac{\partial}{\partial t}e^{it(v;X)} = \mathbb{E}\left(i(v;X)e^{it(v;X)}\right),$$

where the interchange of the differentiation and the expectation is justified Theorem 1.5.12

$$\left|i(v;X)e^{it(v;X)}\right| = |(v;X)| \le |v|\cdot|X| = |X|,$$

which is integrable and does not depend on t. Similarly,

$$\psi_v''(t) = -\mathbb{E}\left((v;X)^2 e^{it(v;X)}\right).$$

and

$$\psi_v'''(t) = -\mathbb{E}\left(i(v;X)^3 e^{it(v;X)}\right),$$

justified similarly. Note that the first three derivatives of ψ_v are bounded by $\mathbb{E}|X|$, $\mathbb{E}|X|^2$ and $\mathbb{E}|X|^3$, respectively. Since $\psi_v(0) = 1$, the bound on the first derivartives implies that $|\psi_v(t)| > \frac{1}{2}$ for $|t| < \varepsilon$, where $\varepsilon = \frac{1}{2\mathbb{E}|X|}$. It follows that for $|t| < \varepsilon$, the logarithm $\log\psi_v(t)$ is well-defined, and its derivatives

$$(\log\psi_v)' = \frac{\psi_v'}{\psi_v}; \quad (\log\psi_v)'' = \frac{\psi_v''\psi_v - (\psi_v')^2}{\psi_v^2}, \quad (\log\psi_v''') = \frac{\text{Polynomial}(\psi_v,\psi_v',\psi_v'',\psi_v''')}{\psi_v^4}$$

are bounded by some constant (that can be expressed in terms of $\mathbb{E}|X|$, $\mathbb{E}|X|^2$ and $\mathbb{E}|X|^3$). By Taylor formula, this implies that, for $|t| < \varepsilon$,

$$\log\psi_v(t) = \log\psi_v(0) + t\,(\log\psi_v)'(0) + \frac{t^2}{2}(\log\psi_v)''(0) + \Theta(t),$$

where $\Theta(t) \le C|t^3|$ with $C = \frac{1}{6}\sup_{|t|<\varepsilon}(\log\psi_v)'''$. It remains to do the computations:

$$\log\psi_v(0) = \log 1 = 0;$$

$$(\log\psi_v)'(0) = \frac{\psi_v'(0)}{\psi_v(0)} = \mathbb{E}\left(i(v;X)\right) = i(v;\mu);$$

$$(\log\psi_v)''(0) = \frac{\psi_v''\psi_v - (\psi_v')^2}{\psi_v^2}(0) = -\mathbb{E}(v;X)^2 + (v;\mu)^2 = -\mathbb{E}(v;X-\mu)^2 = -(\Sigma v;v).$$

Therefore, for all $v \in \mathbb{R}^d$ with $|v| = 1$ and $t \in \mathbb{R}$ with $|t| < \varepsilon$,

$$\varphi_X(tv) = e^{it(\mu;v) - \frac{(\Sigma v;v)}{2}t^2 + \Theta(t)} = e^{i(\mu;tv) - \frac{(\Sigma tv;tv)}{2} + \Theta(t)},$$

as required. □

We are now in the position to prove Theorem 2.15.2.

Proof of Theorem 2.15.2. . We first note there is a constant $c > 0$ such that

$$\sup_{[-\pi;\pi]^d \setminus (-\varepsilon;\varepsilon)^d} |\varphi_X| \le e^{-c\varepsilon^2} \tag{2.15.3}$$

for all ε small enough. Indeed, since Σ is positive definite, we can find $c > 0$ such that $(\Sigma v; v) \ge 4c|v|^2$ for all $v \in \mathbb{R}^d$. We then can find $\varepsilon' > 0$ such that $|\Theta(v)| \le c|v|^2$ for all $v \in (-\varepsilon';\varepsilon')^d$. Denote

$$M := \sup_{[-\pi;\pi]^d \setminus (-\varepsilon';\varepsilon')^d} |\varphi_X|;$$

we have $M < 1$ by Lemma 2.15.6. If $\varepsilon < \varepsilon'$, then for all $v \in [-\varepsilon';\varepsilon']^d \setminus (-\varepsilon;\varepsilon)^d$, we have

$$|\varphi_X(v)| = e^{-\frac{(\Sigma v;v)}{2} + \Re\mathfrak{e}\,\Theta(v)} \le e^{-2c|v|^2 + c|v|^2} \le e^{-c|v|^2} \le e^{-c\varepsilon^2}.$$

If ε is so small that

$$e^{-c\varepsilon^2} > M,$$

then $|\varphi_X(v)| < e^{-c\varepsilon^2}$ for all $v \in [-\pi;\pi]^d \setminus (-\varepsilon;\varepsilon)^d$, which proves (2.15.3).

This gives us a hint for the choice of the sequence ε_n. We have, for ε small enough,

$$e^{-i(v;z_n)}\left(\varphi_X(v)\right)^n = e^{-i(v;z_n) + i(n\mu;v) - n\frac{(\Sigma v;v)}{2} + n\Theta(v)}$$

for $v \in (-\varepsilon;\varepsilon)^d$, and

$$\left|e^{-i(v;z_n)}\left(\varphi_X(v)\right)^n\right| \le e^{-cn\varepsilon^2}$$

for $v \notin (-\varepsilon;\varepsilon)^d$. We want both the error term $|n\Theta(v)| \le Cn\varepsilon^3$ and the last bound to be small. Therefore, we should choose ε_n so that $n\varepsilon_n^3 \to 0$, but $n\varepsilon_n^2 \to \infty$. We put $\varepsilon_n := n^{-\frac{2}{5}}$. Write

$$\int_{[-\pi;\pi]^d} \left(\varphi_X(v)\right)^n e^{-i(v;z_n)} d\lambda^d(v) = \int_{[-\varepsilon_n;\varepsilon_n]^d} \left(\varphi_X(v)\right)^n e^{-i(v;z_n)} d\lambda^d(v) + \int_{[-\pi;\pi]^d \setminus [-\varepsilon_n;\varepsilon_n]^d} \left(\varphi_X(v)\right)^n e^{-i(v;z_n)} d\lambda^d(v)$$

The second integral is bounded above by $(2\pi)^d e^{-cn\varepsilon_n^2} = (2\pi)^d e^{-cn^{\frac{1}{5}}}$, which decays smaller that any power of n, and thus is negligible for our analysis. The first integral is

$$\int_{[-\varepsilon_n;\varepsilon_n]^d} \left(\varphi_X(v)\right)^n e^{-i(v;z_n)} d\lambda^d(v) = \int_{[-\varepsilon_n;\varepsilon_n]^d} e^{-i(v;z_n) + i(n\mu;v) - n\frac{(\Sigma v;v)}{2} + n\Theta(v)} d\lambda^d(v) = \int_{[-\varepsilon_n;\varepsilon_n]^d} e^{-i(z_n';\sqrt{n}v) - \frac{(\Sigma(\sqrt{n}v);(\sqrt{n}v))}{2} + n\Theta(v)} d\lambda^d(v),$$

where $z_n' = (z_n - \mu n)/\sqrt{n} = a + o(1)$. Now, by the change of variable $\sqrt{n}v := w$, the last integral is equal to

$$n^{-\frac{d}{2}} \cdot \int_{[-\sqrt{n}\varepsilon_n;\sqrt{n}\varepsilon_n]^d} e^{-i(a+o(1);w)-\frac{(\Sigma w;w)}{2}+n\Theta(\frac{w}{\sqrt{n}})} d\lambda^d(w).$$

Since $|n\Theta(\frac{w}{\sqrt{n}})| \leq Cn\varepsilon_n^3 \leq Cn^{-\frac{1}{5}} \to 0$, the function under the integral tends pointwise to

$$e^{-i(a;w)-\frac{(\Sigma w;w)}{2}};$$

also, for n large enough (so that $|e^{Cn^{-\frac{1}{5}}}| < 2$), its absolute value is bounded from above by

$$2e^{-(\Sigma w,w)/2}.$$

Therefore, we apply the dominated convergence theorem to conclude that

$$n^{\frac{d}{2}}\mathbb{P}(S_n = z_n) \to \frac{1}{(2\pi)^d} \int_{\mathbb{R}^d} e^{-i(a;w)-\frac{(\Sigma w;w)}{2}} d\lambda^n(w).$$

Making a change of variable $u := \Sigma^{\frac{1}{2}} w$, the integral can be computed:

$$\begin{aligned}\int_{\mathbb{R}^d} e^{-i(a;w)-\frac{(\Sigma w;w)}{2}} d\lambda^n(w) &= \frac{1}{(\det\Sigma)^{\frac{1}{2}}} \int_{\mathbb{R}^d} e^{-i(\Sigma^{-\frac{1}{2}}a;u)} e^{-\frac{(u;u)}{2}} d\lambda^d(u) \\ &= \frac{1}{(\det\Sigma)^{\frac{1}{2}}} \prod_{j=1}^d \int_{\mathbb{R}} e^{-i(\Sigma^{-\frac{1}{2}}a)_j u_j - \frac{u_j^2}{2}} = \frac{(2\pi)^{\frac{d}{2}}}{(\det\Sigma)^{\frac{1}{2}}} \prod_{j=1}^d e^{-(\Sigma^{-\frac{1}{2}}a)_j^2/2} \\ &= \frac{(2\pi)^{\frac{d}{2}}}{(\det\Sigma)^{\frac{1}{2}}} e^{-(\Sigma^{-\frac{1}{2}}a;\Sigma^{-\frac{1}{2}}a)/2},\end{aligned}$$

as claimed. □

CHAPTER 3

Markov chains and the Poisson process

3.1. Markov chains: key definitions

DEFINITION 3.1.1. A discrete (respectively, continuous) time *stochastic process* is a sequence of random variables X_t, $t \in \mathbb{Z}_{\geq 0}$ (respectively, $t \in \mathbb{R}_{\geq 0}$)[1] defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and taking value in the same measurable space S. The space S is calles the *space of states* of the process X_t.

A stochastic process can be thought of as a random quantity evolving in time. We will restrict our attention to the case when

S is a finite or countable space

(with a σ-algebra 2^S); such stochastic processes are called *discrete*.

Little can be said about stochastic processes in general. Markov chains form a simple yet very general class.

DEFINITION 3.1.2. A discrete time stochastic process X_t is called a *Markov chain* if for any $x_0, \ldots, x_{t+1} \in S$ such that

$$\mathbb{P}(X_{t+1} = x_{t+1} | X_t = x_t, \ldots, X_0 = x_0) = \mathbb{P}(X_{t+1} = x_{t+1} | X_t = x_t)$$

for every $x_0, \ldots, x_{t+1} \in S$ such that $\mathbb{P}(X_t = x_t, \ldots, X_0 = x_0) > 0$.

The last condition allows one to treat the conditional expectations in the elementary way; recall that by definition

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

The intuitive meaning of the definition is that the distribution of a Markov chain in the future (that is, on step $t+1$) only depends on its present state (i. e., on its state at step t) and not on its past.

We let $\mu^{(t)}$, $t = 0, 1, \ldots$ denote the distribution of X_t, that is[2],

$$\mu^{(t)}(x) := \mathbb{P}(X_t = x), \quad \text{for all } x \in S.$$

The most important quantity pertaining to a Markov chain is the transition matrix

DEFINITION 3.1.3. A transition matrix of a Markov process at time $t = 1, 2, \ldots$ is defined as

$$P_{xy}^{(t)} := \mathbb{P}(X_t = y | X_{t-1} = x), \quad x, y \in S.$$

Thus, the transition matrix is an $|S| \times |S|$ matrix whose rows and columns are indexed by elements of S.

A transition matrix satisfies the following obvious properties: first,

$$P_{xy}^{(t)} \geq 0 \quad \text{for all } x, y \in S, \tag{3.1.1}$$

because the entries of the matrix are probabilities, and second,

$$\sum_{y \in S} P_{xy}^{(t)} = 1, \quad \text{for all } x \in S, \tag{3.1.2}$$

simply because the event $X_t \in S$ has probability 1 and is a disjoint union of the events $\{X_t = y\}$, $y \in S$.

[1] sometimes $t \in \mathbb{Z}$ or $t \in \mathbb{R}$

[2] naturally identifying probability measures on S with functions $\mu : S \to \mathbb{R}_{\geq 0}$ such that $\sum_{x \in S} \mu(x) = 1$

The transition matrices allow one to compute $\mu^{(t)}$ inductively.

LEMMA 3.1.4. *(Kolmogorov-Chapman equations) We have, for all $t \geq 1$,*

$$\mu^{(t)}(y) = \sum_{x \in S} \mu^{(t-1)}(x) P^{(t)}_{xy} \quad \text{for all } x \in S,$$

or, in a matrix form,

$$\mu^{(t)} = \mu^{(t-1)} P^{(t)}, \tag{3.1.3}$$

where $\mu^{(t)}$ is viewed as a row *vector. More generally,*

$$\mu^{(t)} = \mu^{(0)} P^{(1)} \cdot \dots \cdot P^{(t)}. \tag{3.1.4}$$

PROOF. We have, as required,

$$\mu^{(t)}(y) = \mathbb{P}(X_t = y) = \sum_{x \in S} \mathbb{P}(X_t = y \text{ and } X_{t-1} = x) = \sum_{x \in S} \mathbb{P}(X_t = y | X_{t-1} = x) \mathbb{P}(X_{t-1} = x)$$
$$= \sum_{x \in S} \mu^{(t-1)}(x) P^{(t)}_{xy}.$$

The equation (3.1.4) is obtained from (3.1.3) by iteration. □

This shows that all $\mu^{(t)}$ are determined by $\mu^{(0)}$ and the transition matrices. In fact, the same data determine uniquely the entire law of the Markov chain:

PROPOSITION 3.1.5. *For every $x_0, \dots, x_t \in S$, we have*

$$\mathbb{P}(X_0 = x_0, \dots, X_t = x_t) = \mu^{(0)}(x_0) P^{(1)}_{x_0 x_1} \cdot \dots \cdot P^{(t)}_{x_{t-1} x_t},$$

in particular,[3]*, the law of $(X_0, X_1, \dots, X_t)$ is uniquely determined by $P^{(1)}, \dots, P^{(t)}$ and $\mu^{(0)}$. Conversely, given a probability measure μ on S and a sequence $P^{(t)}$ of $S \times S$ matrices satisfying (3.1.1) and (3.1.2), there is a Markov chain $X_0, X_1, \dots$ such that $\mathbb{P}(X_0 = x) = \mu(x)$ and $\mathbb{P}(X_t = y | X_{t-1} = x) = P^{(t)}_{xy}$ for all $x, y \in S$ and $t \geq 1$.*

PROOF. We have, as required,

$$\mathbb{P}(X_t = x_t, \dots, X_0 = x_0) = \mathbb{P}(X_t = x_t | X_{t-1} = x_{t-1}, \dots, X_0 = x_0) \mathbb{P}(X_{t-1} = x_{t-1}, \dots, X_0 = x_0)$$
$$\overset{\text{Markov property}}{=} \mathbb{P}(X_t = x_t | X_{t-1} = x_{t-1}) \mathbb{P}(X_{t-1} = x_{t-1}, \dots, X_0 = x_0)$$
$$= P^{(t)}_{x_{t-1} x_t} \mathbb{P}(X_{t-1} = x_{t-1}, \dots, X_0 = x_0) = \dots = \mathbb{P}(X_0 = x_0) P^{(1)}_{x_0 x_1} \cdot \dots \cdot P^{(t)}_{x_{t-1} x_t}.$$

For the converse part, define a measure $\mu^{[t]}$ on $S^{t+1} = \{(x_0, \dots, x_t) : x_i \in S\}$ by

$$\mu^{[t]}(\{(x_0, \dots, x_t)\}) = \mu(x_0) P^{(1)}_{x_0 x_1} \cdot \dots \cdot P^{(t)}_{x_{t-1} x_t}.$$

Conditions (3.1.1) and (3.1.2) ensure that this is a probability measure, and, moreover, that $\mu^{[t]}$ is a consistent family of measures: indeed, $\mu^{[t]}$ is positive, and for all $A \subset S^t$,

$$\mu^{[t+1]}(A \times S) = \sum_{(x_0, \dots, x_t) \in A} \sum_{x_{t+1} \in S} \mu(x_0) P^{(1)}_{x_0 x_1} \cdot \dots \cdot P^{(t)}_{x_t x_{t+1}} = \sum_{(x_0, \dots, x_t) \in A} \mu(x_0) P^{(1)}_{x_0 x_1} \cdot \dots \cdot P^{(t)}_{x_{t-1} x_t} = \mu^{[t]}(A).$$

Therefore, by Kolmogorov extension theorem, there exists a probability measure $\mu^{[\infty]}$ on the set of infinite sequences $S^{\mathbb{Z}_{\geq 0}} := \{x = (x_0, x_1, \dots) : x_i \in S\}$ that coincides with $\mu^{[t]}$ on cylindrical sets. Taking $X_t : S^{\mathbb{Z}_{\geq 0}} \to S$ to be the the coordinate function $X_t(x) = x_t$ provides the required Markov chain. □

DEFINITION 3.1.6. A Markov chain is called *homogeneous* if the transition matrix $P = P^{(t)}$ does not depend on t.

[3]This implies that the distribution of the infinite vector $(X_0, X_1, \dots)$, restricted to the product σ-algebra, is uniquely determined by $\mu^{(0)}$ and $P^{(t)}$.

From now on, we will restrict our attention to homogeneous Markov chains. In that case, Kolmogorov-Chapman equations (3.1.4) take a particularly simple form:

$$\mu^{(n)} = \mu^{(0)} P^n. \tag{3.1.5}$$

We will often consider Markov chains with the same transition matrix P, but different initial conditions. In that case, we will abuse terminology and say "the Markov chain P (with initial conditions $\mu^{(0)}$)". If the initial conditions are such that $\mathbb{P}(\mu^{(0)} = x) = 1$ for some x (that is, the chain starts from a certain non-random state x), we will write $\mathbb{E}_x$ and $\mathbb{P}_x$ for expectation and probability pertaining to a Markov chain P with initial state $\mu^{(0)} = \delta(x = 0)$.

3.2. Examples of Markov chains

The first example of a Markov chain has already been discussed: a random walk on $\mathbb{Z}^d$. Indeed, if $X_1, X_2, \dots$ are i. i. d. with values in $\mathbb{Z}^d$, then $S_n = \sum_{i=1}^n X_i$ is a $\mathbb{Z}^d$-valued stochastic process. Let us check from the definition that it is Markov:

$$\begin{aligned}\mathbb{P}(S_{n+1} = x_{n+1} | S_n = x_n, \dots, S_0 = x_0) &= \mathbb{P}(X_{n+1} = x_{n+1} - x_n | S_n = x_n, \dots, S_0 = x_0) \\ &\overset{\text{independence}}{=} \mathbb{P}(X_{n+1} = x_{n+1} - x_n) = \mathbb{P}(S_{n+1} = x_{n+1} | S_n = x_n).\end{aligned}$$

Remark 3.2.1. This argument can be generalized as follows. A Markov chain in a *random mapping representation* consists of a measurable space Λ, a measurable map $f : \Lambda \times S \to S$, a sequence of i. i. d. random variables $\xi_1, \xi_2, \dots$ with values in Λ, and a random variable X_0 with values in S, independent of $\xi_1, \xi_2, \dots$. Define, inductively, $X_n := f(\xi_n, X_{n-1})$. Then

$$\begin{aligned}\mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n, \dots, X_0 = x_0) &= \mathbb{P}(f(\xi_{n+1}, X_n) = x_{n+1} | X_n = x_n, \dots, X_0 = x_0) \\ = \mathbb{P}(f(\xi_{n+1}; x_n) = x_{n+1} | X_n = x_n, \dots, X_0 = x_0) &\overset{\text{independence}}{=} \mathbb{P}(f(\xi_{n+1}; x_n) = x_{n+1}) \\ &= \mathbb{P}(X_{n+1} = x_{n+1} | X_n = x_n),\end{aligned}$$

that is, X_i is a Markov chain. Conversely, let P be any transition matrix, and let ξ_i be i. i. d. uniformly distributed in $(0; 1)$. Choose, for each $x \in S$, a partition $(0; 1) = \sqcup_{y \in S} A_{xy}$ with $\lambda(A_{xy}) = P_{xy}$, and define $f(\xi, x) := y$ if $\xi \in A_{xy}$. Then,

$$\mathbb{P}(f(\xi_{n+1}, X_n) = x_{n+1} | X_n = x_{n+1}) = \mathbb{P}(f(\xi_{n+1}, x_n) = x_{n+1}) = \lambda(A_{xy}) = P_{xy},$$

that is, X_n has transition matrix P. In practice, random mapping representations are used to simulate Markov chains.

The next two famous Markov chains turn out also to be examples of random walks, this time with values in groups.

Example 3.2.2. (Card shuffling) Let μ be any probability measure on the set π_n of transpositions on n elements. Let $\sigma_1, \sigma_2, \dots$ be independent random variables with values in π_n, distributed according to μ. Then $X_n := \sigma_n \cdot \ldots \cdot \sigma_1$ is a π_n -valued Markov chain. Each σ_i corresponds to one iteration of shuffling, and its distribution μ depends on the shuffling method. One (rather inefficient) way to toss the deck is to take the top card and put it at a uniform random position in the deck, then repeat. In this case, we have

$$\mu(123 \dots n) = \mu(213 \dots n) = \mu(231 \dots n) = \dots = \mu(234 \dots n1) = \frac{1}{n}.$$

Example 3.2.3. (Ehrenfest diffusion model). Consider N particles divided into two chambers. There is a small hole between the chambers, and at each discrete instant of time, one particle, chosen uniformly among all of them, diffuses from its current chamber to the other one. If the left chamber has k particles, then the probability that the particle will jump from left to right is $\frac{k}{N}$. The state space of the chain is

$\{0, 1, \dots, N\}$, and the transition matrix is given by

$$P_{xy} = \begin{cases} \frac{x}{N}, & y = x - 1; \\ \frac{N-x}{N}, & y = x + 1; \\ 0, & \text{else.} \end{cases}$$

The Ehrenfest model can also be constructed from random walk on $(\mathbb{Z}/2\mathbb{Z})^N$. The elements of $(\mathbb{Z}/2\mathbb{Z})^N$ are strings of bits (e.g, $(1, 0, 1, 1, 0, \dots, 1)$). Let $\sigma_1, \sigma_2, \dots$ be i. i. d. elements of $(\mathbb{Z}/2\mathbb{Z})^N$ with distribution

$$\mu((1, 0, 0, \dots, 0)) = \mu((0, 1, 0, \dots, 0)) = \dots = \mu((0, 0, 0, \dots, 1)) = \frac{1}{N},$$

and consider $X_n = \sigma_1 \oplus \sigma_2 \oplus \dots \oplus \sigma_n$, where $\oplus$ is the addition in $(\mathbb{Z}/2\mathbb{Z})^N$ (or bit-wise XOR). The passage from X_n to X_{n+1} amounts to choosing a bit uniformly at random and switching it, so, if we only keep track of the number of 1's, we are back to the Ehrenfest's diffusion model.

Example 3.2.4. (Branching processes aka Galton-Watson processes) This is an (oversimplified) model for dynamics of population. We start with one individual, who produces a random number of children (distributed according to some distribution μ on $\mathbb{Z}_{\geq 0}$). Each child then produces her own offspring, distributed according to μ and independently of each other (and of the number of her sisters). The process continues in the same way; we are interested in the number of individuals in n-th generation. Formally, consider a double array of i. i. d. random variables

$$\begin{matrix} \xi_1^{(1)} & \xi_2^{(1)} & \dots \\ \xi_1^{(2)} & \xi_2^{(2)} & \dots \\ \vdots & \vdots & \ddots \end{matrix}$$

distributed according to μ, and put $X_1 := 1$, and $X_{n+1} = \xi_1^{(n)} + \dots + \xi_{X_n}^{(n)}$ for $n = 1, 2, \dots$

The state space of the chain is given by $\mathbb{Z}_{\geq 0}$, and we have

$$P_{xy} = \mathbb{P}(\sum_{i=1}^{x} \xi_i = y),$$

where $\xi_1, \dots, \xi_x$ are i. i. d. distributed according to μ.

Example 3.2.5. (Queing processes) In this example, the state space is $\mathbb{Z}_{\geq 0}$, and the states correspond to the number of customers in a queue. At each instant of time, a new customer arrives with probability p, and, independently of that, the first customer in the queue (should there be any) is served and leaves with probability q. The transition matrix is given by

$$P_{xy} = \begin{cases} p(1-q), & y = x + 1; \\ q(1-p), & y = x - 1; \\ pq + (1-p)(1-q), & y = x; \\ 0, & \text{else,} \end{cases} \quad \text{for } x > 0, \text{ and } P_{0y} = \begin{cases} 1 - p, & y = 0 \\ p, & y = 1 \\ 0 & \text{else} \end{cases}$$

3.3. Stationary distributions

The main question about the Markov chains is the asymptotic behaviour of the measures $\mu^{(n)}$, that is, the existence of the limit $\mu = \lim_{n\to\infty} \mu^{(n)}$. If such a limit exists, then, from the condition $\mu^{(n)} = \mu^{(n-1)}P$, we must have

$$\mu = \mu P.$$

Definition 3.3.1. A measure μ on S satisfying the equation $\mu = \mu P$ is called *a stationary measure.* If μ is a probability measure, then it is called a *stationary distribution.*

Our first task will be to study stationary distributions. We d

DEFINITION 3.3.2. A Markov chain P is called *irreducible* if, for every $x, y \in S$, there is a number $n \in \mathbb{N}$ such that $(P^n)_{xy} > 0$.

Since we have $(P^n)_{xy} = \mathbb{P}(X_{t+n} = y | X_t = x)$, a Markov chain is irreducible if and only if, starting from any state, it has positive probability to pass through any other state.

THEOREM 3.3.3. *Let P be a finite*[4] *Markov chain. Then*

(1) *P has a stationary distribution μ;*
(2) *If P is irreducible, then μ is unique. In this case,*

$$\mu(x) = \frac{1}{\mathbb{E}_x \tau_x},$$

where $\tau_x = \min\{t > 0 : X_t = x\}$ is the time of the first return of the chain to 0.

The intuition behind the last equation is as follows. Since we expect μ to be the limit of $\mu^{(n)}$ for large n, we can think of $\mu(x)$ as the proportion of time that the chain, eventually, spends at the state x. It is then natural to expect that this proportion, multiplied by the average time between visits to x, should give 1.

A TOPOLOGICAL PROOF OF (1). Denote by $\mathcal{P}(S)$ the set of all probability measures on S, that is, the set of all function $\mu : S \to \mathbb{R}$ such that $\mu(x) \geq 0$ for all $x \in S$ and $\sum_{x \in S} \mu(x) = 1$. Let $\mu \in \mathcal{P}(S)$. Then, $(\mu P)(y) = \sum_{x \in S} \mu(x) P_{xy} \geq 0$ for all y. Also,

$$\sum_{y \in S} (\mu P)(y) = \sum_{y \in S} \sum_{x \in S} \mu(x) P_{xy} = \sum_{x \in S} \mu(x) \sum_{y \in S} P_{xy} \overset{(3.1.2)}{=} \sum_{x \in S} \mu(x) = 1.$$

This means that the map $\psi : \mu \mapsto \mu P$ maps the set $\mathcal{P}(S)$ into itself. The set $\mathcal{P}(S)$ is a closed $|S| - 1$ dimensional simplex, homeomorphic to a closed ball. Since ψ is linear, it is continuous. Therefore, by Brouwer's theorem, ψ has a fixed point, as required. □

A PROBABILISTIC PROOF OF (1). Take any state $z \in S$, and let $\rho(x)$ be the average time[5] the chain spends at x between two concequtive visits to z:

$$\rho(x) := \mathbb{E}(\#\{t < \tau_z : X_t \in x\}) = \mathbb{E}\left(\sum_{t=0}^{\infty} \mathbb{I}_{\{t < \tau_z \text{ and } X_t \in x\}} \right) = \sum_{t=0}^{\infty} \mathbb{P}(t < \tau_z \text{ and } X_t \in x),$$

where $\mathbb{E} = \mathbb{E}_z$ and $\mathbb{P} = \mathbb{P}_z$ (that is, we are assuming that the initial state of the chain is z). We claim that $\rho = \rho P$, i. e., that ρ is a stationary measure. Indeed,

$$\sum_{x \in S} \rho(x) P_{xy} = \sum_{x \in S} \sum_{t=0}^{\infty} \mathbb{P}(t < \tau_z \text{ and } X_t \in x) \mathbb{P}(X_{t+1} = y | X_t = x).$$

Since the condition $t < \tau_z$ is determined by values of $X_0, \ldots, X_t$, the Markov property allows us to insert it into the last probability:

$$\begin{aligned} \mathbb{P}(t < \tau_z \text{ and } X_t \in x) \mathbb{P}(X_{t+1} = y | X_t = x) & \\ = \mathbb{P}(X_{t+1} = y | X_t = x \text{ and } t < \tau_z) \mathbb{P}(t < \tau_z \text{ and } X_t \in x) & \\ = \mathbb{P}(X_{t+1} = y, X_t = x \text{ and } t < \tau_z). & \end{aligned}$$

[4] i. e., with finite state space S

[5] possibly infinite. Our choice of z in the end of the proof will guarantee that $\rho(x) < \infty$ for all x.

Therefore,

$$\sum_{x\in S}\rho(x)P_{xy}=\sum_{t=0}^{\infty}\sum_{x\in S}\mathbb{P}(X_{t+1}=y,X_t=x \text{ and } t<\tau_z)=\sum_{t=0}^{\infty}\mathbb{P}(X_{t+1}=y \text{ and } t<\tau_z)$$
$$=\sum_{t=1}^{\infty}\mathbb{P}(X_t=y \text{ and } t\le\tau_z)=\mathbb{E}(\#\{t\in\{1,\dots,\tau_z\}:X_t\in y\}).$$

Since $X_0=X_{\tau_z}=z$, we have, for all $y\in S$,

$$\#\{t\in\{1,\dots,\tau_z\}:X_t\in y\}=\#\{t\in\{0,\dots,\tau_z-1\}:X_t\in y\}\quad\text{almost surely,}$$

i. e., $\rho P=\rho$.

We now wish to normalize ρ to make it a probability measure. We have,

$$\sum_{x\in S}\rho(x)=\mathbb{E}_z\left(\sum_{x\in S}\#\{t<\tau_z:X_t\in x\}\right)=\mathbb{E}_z(\#\{t:t<\tau_z\})=\mathbb{E}_z(\tau_z).$$

Now, by Lemma 3.3.7 below, we can choose choose z so that $\mathbb{E}_z\tau_z<\infty$, so, putting $\mu=\frac{\rho}{\mathbb{E}(\tau_z)}$ concludes the proof.

Motivated by the proof above, we introduce the following terminology. □

Definition 3.3.4. A state x of a chain S is called *recurrent* if $\mathbb{P}_x(\tau_x<\infty)=1$. A state is called *positive recurrent* or *non-null recurrent* if $\mathbb{E}_x(\tau_x)<\infty$.

We now observe that in the Probabilistic proof of (1), we never used that S is finite. What we have in fact proven can be summarized as follows:

Proposition 3.3.5. *If a Markov chain has at least one positive recurrent state, then it has a stationary distribution.*

Remark 3.3.6. It is possible to show that the converse is also true. Simple random walk on $\mathbb{Z}$ (or $\mathbb{Z}^2$) provides an example of a chain in which all states are recurrent, but not positive recurrent.

The final piece, referred to in the proof above, is the following:

Lemma 3.3.7. *Any finite Markov chain has a positive recurrent state. If a finite chain is irreducible, then, in fact, all states are positive recurrent.*

Proof. We start with the second assertion. Let $T:=\max_{x,y\in S}\min\{m:(P^m)_{xy}>0\}+1$; since the chain is irreducible, the minimum is finite for every x,y, and since the chain is finite, the maximum is finite. Then, there exists a number $p>0$ such that

$$\mathbb{P}_x(\tau_y\le T)>p.$$

Indeed, for any x and y, there is a positive probability to get from x to y in no more than T steps, so, we take p to be the minimum of these probabilities.

This, in particular, means that $\mathbb{P}_x(\tau_y>T)<1-p$, for all x,y. If this event occurs, then we may think of the future evolution of the chain as a new chain (started from a random state X_R) and deduce that $\mathbb{P}(\tau_y>2T)\le(1-p)^2$, and so on. Formally, for any $k\in\mathbb{N}$ and $R\in\mathbb{N}$, we have, for all z and y,

$$\mathbb{P}_z(\tau_y>R+T)=\mathbb{P}_z(\tau_y>R \text{ and } X_{R+1}\ne y,\dots,X_{R+T}\le y)=$$
$$=\sum_{x\in S}\mathbb{P}_z(X_{R+1}\ne y,\dots,X_{R+T}\ne y \text{ and } \tau_y>R \text{ and } X_R=x)$$
$$=\sum_{x\in S}\mathbb{P}_z(X_{R+1}\ne y,\dots,X_{R+T}\ne y|X_R=x)\mathbb{P}(\tau_y>R \text{ and } X_R=x)$$
$$\le(1-p)\sum_{x\in S}\mathbb{P}(\tau_y>R \text{ and } X_R=x)\le(1-p)\mathbb{P}(\tau_y>R).$$

In the third equality, we used once again that conditioning on $\tau_z > R$ and $X_R = x$ is the same as conditioning on $X_R = x$ alone, by Markov property. Iterating, we get $\mathbb{P}_z(\tau_y > kT) \le (1-p)^k$, and

$$\mathbb{E}_z(\tau_y) = \sum_{t=1}^{\infty} t\mathbb{P}_z(\tau_y = t) = \sum_{k=0}^{\infty} \sum_{t=kT+1}^{(k+1)T} t\mathbb{P}_z(\tau_y = t) \le \sum_{k=0}^{\infty} (k+1)T\mathbb{P}_z(\tau_y \ge kT+1) \le \sum_{k=0}^{\infty} (k+1)T(1-p)^k < \infty.$$

To extend the proof to the general case, consider the oriented graph whose vertices are states of S, and two vertices x, y are connected by an edge if and only if $P_{xy} > 0$. Let us say that a vertex x communicates with y if $(P^n)_{xy} > 0$ for some $n \ge 0$, or, equivalently, that we can pass from x to y by following edge in the graph. Note that "x communcates with y and y communicates with x" is an equivalence relation. Identifying equivalent vertices, we obtain a factor-graph. Note that the factor-graph has a node without outgoing edges: just take any node and follow the arrows; since we cannot form a loop (otherwise the nodes along the loop would be identified), the process must terminate.

So, let $S' \subset S$ be the set of equivalent vertices in S that form a node of the factor-graph without outgoing edges. Then, for every $x \in S'$, we have $\sum_{y \in S'} P_{xy} = \sum_{y \in S} P_{xy} = 1$. This means that we can define a Markov chain with state space S' and the transition matrix P_{xy} restricted to S' (that is to say, simply, that if the original chain starts in S', it will remain there forever); this new chain is irreducible, and hence $\mathbb{E}_z \tau_z < \infty$ for all $z \in S'$. □

Proof of (2). Note that if z is any state and ρ is as in the proof of (1), then $\mu(z) = \rho(z)/\mathbb{E}_z(\tau_z)$, and $\rho(z) = \mathbb{E}(\#\{t < \tau_z : X_t = z\}) = 1$. So, the identity follows once we prove uniqueness.

We call a function $f : S \to \mathbb{R}$ *harmonic* if $Pf = f$, where f is viewed as a *column* vertor, that is,

$$\sum_{y \in S} P_{xy} f(y) = f(x) \quad \text{for all } x \in S.$$

We claim that for a finite Markov chain, f is harmonic if and only if f is constant. Indeed, let f be harmonic, and x be a vertex where f attains its maximal value. We have

$$f(x) = \sum_{y \in S} P_{xy} f(y) \le \sum_{y \in S} P_{xy} f(x) = f(x) \sum_{y \in S} P_{xy} = f(x),$$

therefore, the inequality in the middle is in fact equality, and we conclude that $f(y) = f(x)$ for all y such that $P_{xy} > 0$. Applying the same agrument to these y instead of x, we show that $f(z) = f(x)$ for all z such that $P_{xy} > 0$ and $P_{yz} > 0$ for some y. Continuing this way, we will eventually prove that $f(w) = f(x)$ for all w such that x communicates with w, that is, for all $w \in S$, since the chain is irreducible.

Now, what we have proven is

$$\operatorname{rank}(P - I) = |S| - \dim \ker(P - I) = |S| - 1.$$

Since the rank of a matrix is the same as the rank of its transpose, this implies that

$$\dim\{\mu : \mu(P - I) = 0\} = 1,$$

as required. □

Remark 3.3.8. The necessary and sufficient condition for the uniqueness of the stationary measure is that the factor-graph discussed in the proof of Lemma 3.3.7 has only one node without outgoing edges.

3.4. Aperiodicity and convergence results

The measure $\mu^{(n)}$ may fail to have a limit because of the periodicity phenomenon: consider the chain with just two states x, y, with transition matrix given by $P_{xy} = P_{yx} = 1$ and $P_{xx} = P_{yy} = 0$. Then

$$\mathbb{P}_x(X_t = x) = \begin{cases} 1, & t \text{ is even}, \\ 0, & t \text{ is odd}, \end{cases}$$

which does not converge to any limit as $t \to \infty$. To exclude such a situation, we adopt the following definition.

DEFINITION 3.4.1. Denote $\mathcal{T}(x) := \{t > 0 : (P^t)_{xx} > 0\} = \{t > 0 : \mathbb{P}_x(X_t = x) > 0\}$ the set of possible return times to x. We call a state x aperiodic if $\gcd(\mathcal{T}(x)) = 1$. A chain is called aperiodic if all its states are aperiodic.

We will use the following elementary number theoretic fact:

LEMMA 3.4.2. *If the state x is aperiodic, then $\mathcal{T}(x)$ contains all but finitely many natural numbers.*

PROOF. We shall prove that any set $\mathcal{T} \subset \mathbb{N}$ closed under addition and with $\gcd(\mathcal{T}) = 1$ contains all but finitely many natural numbers.

First, note that we can choose finitely many $t_1, \dots, t_M \in \mathcal{T}$ such that $\gcd(t_1, \dots, t_M) = 1$. Indeed, take $t_1 = \min \mathcal{T}(x)$, and let

$$t_1 = q_1^{k_1} \cdots q_l^{k_l}$$

be its factorization into primes. For each $j = 1, \dots, l$, we can choose a $t_\text{j} \in \mathcal{T}$ that is not divisible by q_j; then $\gcd(t_1, \dots, t_l) = 1$, and it remains to get rid of repetitions.

Now, basic number theory (Euclid's algorithm) says that there are integer number $\beta_1, \dots, \beta_M$ such that

$$\beta_1 t_1 + \cdots + \beta_M t_M = 1.$$

Therefore, for every $n \in \mathbb{N}$, we can find $\alpha_1, \dots, \alpha_M \in \mathbb{Z}$ such that

$$\alpha_1 t_1 + \cdots + \alpha_M t_M = n.$$

We claim that if $n > M t_1 \cdot \cdots \cdot t_M$, then we can choose $\alpha_j \geq 0$ for all j, and then $n \in \mathcal{T}$ since $t_1, \dots, t_M \in \mathcal{T}$ and $\mathcal{T}$ is closed under addition. Assume that there is an i such that $\alpha_i < 0$. We can find j such that $\alpha_j t_j \geq \frac{n}{M}$ (otherwise $\sum \alpha_j t_j < n$). Replace $\alpha_i \mapsto \alpha_i + t_j$ and $\alpha_j \mapsto \alpha_j - t_i$; the sum $\alpha_1 t_1 + \cdots + \alpha_M t_M$ will not change. But $\alpha_j t_j \geq t_1 \cdot \cdots \cdot t_M \geq t_i t_j$, therefore, $\alpha_j - t_i > 0$, so, this operation does not create new negative coefficients. Iterating, we will eventually make all the coefficient non-negative.

Clearly, $\mathcal{T}(x)$ is closed under addition (since $(P^{t_1+t_2})_{xx} \geq P^{t_1}_{xx} P^{t_2}_{xx} > 0$ whenever $P^{t_1}_{xx} > 0$ and $P^{t_1}_{xx} > 0$), hence the Lemma follows. □

To state and prove the convergence result, we need an appropriate notion of convergence. Given a function $\nu : S \to \mathbb{R}$, we define its l_1 norm as

$$||\nu||_1 := \sum_{x \in S} |\nu(x)|.$$

This norm is also called a *total variation* norm, especially when used to measure distance between two probability measures:

$$d_{TV}(\nu_1, v_2) = ||\nu_1 - \nu_2||_1.$$

THEOREM 3.4.3. *Let P be an aperiodic, irreducible Markov chain that has at least one positive recurrent state, and let μ be its stationary measure. Then, for any initial state $\mu^{(0)}$*

$$||\mu^{(n)} - \mu||_1 \overset{n \to \infty}{\longrightarrow} 0.$$

Informally speaking, the proof goes as follows. We run simultateously two copies X_n and Y_n of the Markov chain, one with initial distribution $\mu^{(0)}$, and another one with initial distribution μ. They evolve independently until the first time τ they visit the same state at the same time, and from that point on, they evolve together. At time n, the distribution of the second chain is μ, and on the event $\tau < n$, we have $X_n = Y_n$. So, if $\tau < n$ with high probability, then the distribution of X_n is close to that of Y_n, i. e., to μ. This construction is called a *coupling* of two Markov chains. For technical reasons, we will not actually construct a coupling, but instead will work with two independent copies of the chain all the way through.

PROOF. Consider the Markov chain $Z_n = (X_n; X'_n)$ with state space $S \times S$, initial state $\nu(x_1, x_2) := \mu^{(0)}(x_1)\mu(x_2)$, and transition matrix $\tilde{P}_{(x,x')(y,y')} = P_{xy} P_{x'y'}$. This is indeed a transition matrix:

$$\sum_{(y,y') \in S \times S} P_{xy} P_{x'y'} = \sum_{y \in S} \sum_{y' \in S} P_{xy} P_{x'y'} = \sum_{y \in S} P_{xy} \sum_{y' \in S} P_{x'y'} = 1.$$

Since

$$\mathbb{P}(X_n = x_n, \dots, X_0 = x_0, X'_n = x_n, \dots, X'_0 = x'_0) = \\ = \mu^{(0)}(x_0)\mu(x'_0)P_{x_0x_1}P_{x'_0x'_1}\dots P_{x_{n-1}x_n}P_{x'_{n-1}x'_n} = \\ \left(\mu^{(0)}(x_0)P_{x_0x_1}\dots P_{x_{n-1}x_n}\right)\left(\mu(x'_0)P_{x'_0x'_1}\dots P_{x'_{n-1}x'_n}\right),$$

we see that X_n and Y_n are Markov chains with transition matrix P. The heart of the matter is the following two Lemmas:

LEMMA 3.4.4. *We have, for all $y \in S$ and all $n \in \mathbb{N}$,*

$$\mathbb{P}(X_n = y, \tau \leq n) = \mathbb{P}(X'_n = y, \tau \leq n).$$

PROOF. Indeed,

$$\mathbb{P}(X_n = y, \tau \leq n) = \sum_{t=1}^{n} \mathbb{P}(X_n = y, \tau = t) = \sum_{t=1}^{n}\sum_{x\in S} \mathbb{P}(X_n = y, X_t = x, \tau = t) \\ = \sum_{t=1}^{n}\sum_{x\in S} \mathbb{P}(X_n = y | X_t = x, \tau = t)\mathbb{P}(X_t = x, \tau = t).$$

By Markov property, we can remove $\tau = t$ from the conditioning. Also, on the event $\tau = t$, we have $X_t = X'_t$. Therefore,

$$\mathbb{P}(X_n = y, \tau \leq n) = \sum_{t=1}^{n}\sum_{x\in S} \mathbb{P}(X_n = y | X_t = x)\mathbb{P}(X'_t = y, \tau = t) = \sum_{t=1}^{n}\sum_{x\in S} (P^{n-t})_{xy}\mathbb{P}(X'_t = x, \tau = t) \\ = \sum_{t=1}^{n}\sum_{x\in S} \mathbb{P}(X'_n = y | X'_t = x)\mathbb{P}(X'_t = x, \tau = t) = \mathbb{P}(X'_n = y, \tau \leq n).$$

□

LEMMA 3.4.5. *We have*

$$||\mu^{(n)} - \mu||_1 \leq 2\mathbb{P}(\tau > n).$$

PROOF. Indeed,

$$||\mu^{(n)} - \mu||_1 = \sum_{x\in S} |\mathbb{P}(X_n = x) - \mathbb{P}(X'_n = x)| \leq \sum_{x\in S} |\mathbb{P}(X_n = x, \tau > n) - \mathbb{P}(X'_n = x, \tau > n)| \\ \leq \sum_{x\in S} \mathbb{P}(X_n = x, \tau > n) + \sum_{x\in S} \mathbb{P}(X'_n = x, \tau > n) = 2\mathbb{P}(\tau > n),$$

as required. □

With this Lemma, it remains to prove that $\mathbb{P}(\tau \geq n) \to 0$ as $n \to \infty$, that is, that τ is finite almost surely. The first step is to prove that Z_n is irreducible. Indeed, for each x, x' and y, y', there exist $n_{xx'}$ and $n_{yy'}$ such that

$$(P^{n_{xy}})_{xy} > 0 \quad \text{and} \quad (P^{n_{x'y'}})_{x'y'} > 0.$$

By Lemma 3.4.2, there exist a number T such that $(P^{T-n_{xy}})_{xx} > 0$ and $(P^{T-n_{x'y'}})_{x'x'} > 0$. Then,

$$(P^T)_{xy}(P^T)_{x'y'} \geq (P^{T-n_{xy}})_{xx}(P^{n_{xy}})_{xy}(P^{T-n_{x'y'}})_{x'x'}(P^{n_{x'y'}})_{x'y'} > 0,$$

which shows that Z_n is irreducible.

Now, let us note that for a finite chain, the proof is already complete, since we have shown in the course of the proof of Lemma 3.3.7 that for a finite irreducible chain,

$$\mathbb{P}_x(\tau_y > kT) \leq \alpha^k$$

for some $\alpha < 1$. This also shows that in this case, convergence of $\mu^{(n)}$ to μ is exponential.

In the inifinite case, we have to work a bit harder. We first note that the chain Z_n has a stationary distribution, given by $\overline{\mu}(x,y) = \mu(x)\mu(y)$, and $\overline{\mu}(x,x) > 0$ for any x with $\mu(x) > 0$. The result then follows from Lemma 3.4.6 below. □

LEMMA 3.4.6. *If an irreducible chain X_n has a stationary measure μ, then*

$$\mathbb{P}_x(\tau_y < \infty) = 1$$

for all x and all y with $\mu(y) > 0$, where $\tau_y = \min\{n > 0 : X_n = y\}$.

PROOF. We first show that $\mathbb{P}_y(\tau_y < \infty) = 1$ for any $y \in S$ with $\mu(y) > 0$. Indeed, let μ be a stationary measure, and consider the Markov chain with initial distribution μ. Then, if $\mu(y) > 0$, then the expected number of visits to y is infinite, because

$$\mathbb{E}_\mu(\#\{t \le N : X_t = y\}) = \sum_{x\in S} \mu(x) \sum_{t=1}^{N} P_{xy}^t = \sum_{n=1}^{N} (\mu P^t)_y = N\mu(y).$$

Now, assume that $\mathbb{P}_y(\tau_y < \infty) = \alpha < 1$, and let $\tau_y^{(k)}$ be the time of k-th return to y:

$$\tau_y^{(k)} = \min\{n > \tau_y^{(k-1)} : X_n = y\}.$$

Then, the probability the it is finite is exponentially small in k:

$$\begin{aligned}
\mathbb{P}(\tau_y^{(k)} < \infty) &= \sum_{t=1}^{\infty} \mathbb{P}(\tau_y^{(k-1)} = t, X_t = y, \exists t' > t : X_{t'} = y) \\
&= \sum_{t=1}^{\infty} \mathbb{P}(\exists t' > t : X_{t'} = y | X_t = y, \tau_y^{(k-1)} = t) \mathbb{P}(\tau_y^{(k-1)} = t, X_t = y) \\
&= \sum_{t=1}^{\infty} \mathbb{P}(\exists t' > t : X_{t'} = y | X_t = y) \mathbb{P}(\tau_y^{(k-1)} = t) \\
&= (1-\alpha)\mathbb{P}(\tau_y^{(k-1)} = t) = \cdots \le (1-\alpha)^k.
\end{aligned}$$

This means that

$$\mathbb{E}_\mu(\#\{t \le N : X_t = y\}) = \mathbb{E}_\mu(\max\{k : \tau_y^{(k)} < \infty\}) < \infty,$$

a contradiction. So, $\mathbb{P}_y(\tau_y < \infty) = 1$.

Now, assume that there is an $x \in S$ such that $\mathbb{P}_x(\tau_y > 0) < 1$. Informally speaking, because of irreducibility, we have a positive chance to get to x before the first return to y, and once we are there, there is a positive chance to never return to y, a contradiction. Formally,

$$\begin{aligned}
\mathbb{P}_y(\tau_x < \tau_y) = \mathbb{P}_y(\tau_y < \infty, \tau_x < \tau_y) &= \sum_{t=1}^{\infty} \mathbb{P}_y(\exists t' > t : X_{t'} = y, \tau_x = t, \tau_y > t, X_t = x) \\
&= \sum_{t=1}^{\infty} \mathbb{P}_y(\exists t' > t : X_{t'} = y | X_t = x) \mathbb{P}(\tau_x = t, \tau_y > t) \le \mathbb{P}_x(\tau_y > 0)\mathbb{P}(\tau_x < \tau_y),
\end{aligned}$$

from which $\mathbb{P}(\tau_x < \tau_y) = 0$. Similarly, one can show that $\mathbb{P}(\tau_x < \tau^{(k)}) = 0$ for all k, and taking the limit $k \to \infty$, we get that $\mathbb{P}(\tau_x < \infty) = 0$, a contradiction with the irreducibility. □

REMARK 3.4.7. In this secion, we did not pusrue a complete treatment of the subject, instead taking a shortest possible path to the convergence theorem. In fact, it is possible to show that for an irreducible chain the following are true:

(1) A stationary distribution, if exists, is unique;
(2) If a stationary distribution μ exists, then all states are positive recurrent and $\mu(x) > 0$ for any state x;
(3) Either all states are transient, or all are recurrent.
(4) If the stationary measure does not exist, then $\mu^{(n)}(x) \to 0$ for all x.

3.5. Alternative proof of convergence (optional)

In this optional section, we give another proof of convergence for finite chains, based on the fact that a transition matrix is non-expanding in l_1 norm.

COROLLARY 3.5.1. *If a finite Markov chain is irreducible and aperiodic, then there is a number $N > 0$ such that*

$$(P^N)_{xy} > 0$$

for all $x, y \in S$.

PROOF. By irreducibility, for each x, y, we can choose $M_{xy} \in \mathbb{N}$ such that $(P^{M_{xy}})_{xy} > 0$. Take $M := \max M_{xy}$. By Lemma 3.4.2, we can find $R \in N$ such that $(P^r)_{xx} > 0$ for all $r \geq R$. Then we take $N := M + R$. We have

$$(P^N)_{xy} \geq (P^{N-M_{xy}})_{xx}(P^{M_{xy}})_{xy} > 0,$$

since $N - M_{xy} \geq N - M = R$. □

The $||\cdot||_1$ norm is expecially useful in the study of Markov chains because of the following Lemma.

LEMMA 3.5.2. *Assume that Q is an $S \times S$ matrix such that $Q_{xy} \geq 0$ for all $x, y \in S$, and $\sum_{y \in S} Q_{xy} \leq r$ for all $x \in S$. Then,*

$$||\nu Q||_1 \leq r||\nu||_1$$

for any row vector ν.

PROOF. If $\nu(x) \geq 0$ for all $x \in S$, then also $(\nu Q)(x) \geq 0$ for all $x \in S$. Then,

$$||\nu Q||_1 = \sum_{x \in S}(\nu Q)(x) = \sum_{x \in S}\sum_{y \in S}\nu(y)Q_{yx} = \sum_{y \in S}\nu(y)\sum_{x \in S}Q_{yx} \leq r\sum_{y \in S}\nu(y) = r||\nu||_1.$$

Generally, let ν_+ and $\nu-$ be the positive and negative parts of ν, that is, $\nu_+(x) := \max(\nu(x); 0)$ and $\nu_-(x) := \min(\nu(x); 0)$. Then,

$$||\nu Q||_1 = ||(\nu_+ + \nu_-)Q||_1 \leq ||\nu_+ Q||_1 + ||\nu_- Q||_1 \leq r(||\nu_+||_1 + ||\nu_-||_1) = r||\nu||_1.$$

□

PROOF OF THEOREM 3.4.3 FOR FINITE CHAINS. Let N be such as in Lemma 3.5.1, and denote $\tilde{P} := P^N$. We define a matrix M by

$$M_{xy} = \mu(y)$$

for all x. Then, for any probability measure ν on S, we have

$$(\nu M)(y) = \sum_{x \in S}\nu(x)M_{xy} = \mu(y),$$

that is, $\nu M = \mu$. By Lemma 3.5.1, we can find a number $c > 0$ such that

$$\tilde{P}_{xy} - cM_{xy} > 0.$$

for all $x, y \in S$. We can write, for any probability measure ν,

$$\nu\tilde{P} - \mu = (\nu - \mu)\tilde{P} = (\nu - \mu)(\tilde{P} - cM + cM) = (\nu - \mu)(\tilde{P} - cM).$$

Plugging in $\nu = \mu^{(0)}\tilde{P}^{n-1}$ and iterating, we get

$$\mu^{(0)}\tilde{P}^n - \mu = (\mu^{(0)}\tilde{P}^{n-1} - \mu)(\tilde{P} - cM) = \cdots = (\mu^{(0)} - \mu)(\tilde{P} - cM)^n.$$

We now can apply Lemma 3.5.2 to get

$$||\mu^{(0)}\tilde{P}^n - \mu||_1 \leq (1 - c)^n||\mu^{(0)} - \mu||_1.$$

Generally, if $n = kN + r$, where $0 \leq r < N$, we have

$$||\mu^{(n)} - \mu||_1 = ||\mu^{(0)}P^r\tilde{P}^k - \mu||_1 \leq (1-c)^k||\mu^{(0)}P^r - \mu||_1 = (1-c)^k||(\mu^{(0)} - \mu)P^r||_1 \leq (1-c)^k||(\mu^{(0)} - \mu)||_1.$$

If $n > N$, then $k > 0$, and we write

$$(1-c)^k = \left((1-c)^{\frac{k}{kN+r}}\right)^n \leq \left((1-c)^{\frac{1}{2N}}\right)^n =: \alpha^n,$$

which completes the proof. □

3.6. Poisson process

The Poisson process is a prime example of a Markov process in continuous time. Poisson process is ubiquitious in Probability and its applications. It is used to model, e. g., a number of clicks or the Geyger's counter, the number of customers arriving in a shop by time t, etc. We start by discussing exponential random variables:

Definition 3.6.1. An exponential random variable with parameter λ is a scalar random variable that has density

$$\lambda e^{-\lambda x}\mathbb{I}_{x>0}.$$

Equivalently, it is a random variable X satrisfying $\mathbb{P}(X > x) = e^{-\lambda x}$, $x \geq 0$.

The importance of exponential random variables is revealed in the following proposition:

Proposition 3.6.2. *(Lack of memory of exponential random variables) Let X be a random variable with values in $\mathbb{R}_{\geq 0}$. Then, the following are equivalent:*

1. *X is an exponential random variable;*
2. *for every $t > 0$ such that $\mathbb{P}(X > t) > 0$, conditionally on $X > t$, the variable $X - t$ has the same distribution as X. In formulas,*

$$\mathbb{P}(X > t + s | X > t) = \mathbb{P}(X > s), \quad s, t > 0.$$

3. *for any random variable $Y \geq 0$ independent of X, and any $s > 0$,*

$$\mathbb{P}(X > Y + s | X > Y) = \mathbb{P}(X > s)$$

Proof. (3) $\Longrightarrow$ (2) is trivial (take $Y = t$ almost surely). To prove (2) $\Longrightarrow$ (1), denote $G(t) = 1 - F_X(t)$. Note that since $\mathbb{P}(X > 2\varepsilon | X > \varepsilon) = \mathbb{P}(X > \varepsilon)$, we have, inductively, $\mathbb{P}(X > k\varepsilon) > 0$ for all $k \in \mathbb{N}$ and ε such that $\mathbb{P}(x > \varepsilon) > 0$. Thus, either $X = 0$ a. s., or $G(t) > 0$ for all $t > 0$. For all $t, s \geq 0$, G satisfies the following functional equation:

$$G(t+s) = \mathbb{P}(X > t+s) = \mathbb{P}(X > t+s, X > t) = \mathbb{P}(X > t+s | X > t)\mathbb{P}(X > t) = G(t)G(s).$$

Taking logarithms, we arrive at

$$\log G(t+s) = \log G(t) + \log G(s).$$

Then, $n \log G(\frac{1}{n}) = \log G(1)$, and $\log G(\frac{k}{n}) = \frac{k}{n} \log G(1)$ for all $k, n \in \mathbb{N}$. Since $\log G(x)$ is decreasing, this implies

$$\log G(x) = x \log G(1)$$

for all $x \in \mathbb{R}$, or $G(x) = \mathbb{P}(X > x) = e^{-\lambda x}$, where $\lambda = \log G(1)$.

Finally, we prove (1) $\Longrightarrow$ (3). Let μ_Y denote the distribution of Y; then the distribution of $(X; Y)$ is given by the direct product of measures $\lambda e^{-\lambda x}\mathbb{I}_{x>0} d\lambda(x) \otimes \mu_Y$. We have, for any $s \geq 0$,

$$\mathbb{P}(X > Y + s, X > Y) = \mathbb{P}(X > Y + s) = \int_{\mathbb{R}^2} \mathbb{I}_{x>y+s} \lambda e^{-\lambda x} d\lambda(x) \otimes d\mu_Y(y)$$

$$\overset{\text{Fubini}}{=} \int_{\mathbb{R}} \left(\int_{y+s}^{\infty} \lambda e^{-\lambda x} d\lambda(x) \right) d\mu_Y(y) = e^{-\lambda s} \int_{\mathbb{R}} e^{-\lambda y} d\mu_Y(y) = \mathbb{P}(X > x)\mathbb{P}(X > Y),$$

as required. □

If we think of an exponential random variable as a waiting time before some event happens (e. g., a radioactive atom decays, a first customer of the day walks into a shop, etc.), the lack of memory property means that the distribution of the *remaining* waiting time does not depend on the time that we have already waited.

DEFINITION 3.6.3. The *Poisson process* with intensity λ is a stochastic process on $\mathbb{R}_{\geq 0}$ with values in $\mathbb{Z}_{\geq 0}$, defined as

$$X_t := \max\{n : \xi_1 + \cdots + \xi_n \leq t\},$$

where $\xi_1, \xi_2, \dots$ are i. i. d. exponential random variables with parameter λ.

This way, Poission process is a collection of integer-valued random variables X_t (almost surely increasing etc.). Another way to think about it is that of a *random function* $t \mapsto X_t$, that is, a random variable with values in a space of functions. (For a moment, we ignore the questions about the target space of functions, its σ-algebra etc.). Yet another useful way to view it is that of a *point process*, that is, a random collection of points $\xi_1, \xi_1 + \xi_2, \xi_1 + \xi_2 + \xi_3, \cdots \in \mathbb{R}$, where the above-mentioned function has jumps.

This point process is stationary, in that its restriction to a ray $[a; +\infty)$ looks the same as the process itself:

PROPOSITION 3.6.4. *If X_t is a Poisson process, then, for every $a > 0$, the process $Y_t := X_{a+t} - X_a$ is again a Poisson process. Moreover, for each $t > 0$ and each $a_1, \dots, a_k \leq a$, Y_t is independent of $X_{a_1}, \dots, X_{a_k}$.*

PROOF. Denote $S_n := \sum_{i=1}^n \xi_i$. Fix $t > 0$ and $n, m \in \mathbb{Z}_{\geq 0}$, and let us look at the event

$$\{Y_t = n, X_a = m\} = \{S_m \leq a, S_{m+1} > a, S_{n+m} \leq t + a, S_{n+m+1} > t + a\}.$$

Denote $\tilde{\xi}_{m+1} := S_{m+1} - a = \xi_{m+1} + S_m - a$. Then, we rewrite the above event as

$$\{S_m \leq a, \xi_{m+1} > (a - S_m), \tilde{\xi}_{m+1} + \xi_{m+2} + \cdots + \xi_{m+n} \leq t, \tilde{\xi}_{m+1} + \cdots + \xi_{m+n+1} > t\}.$$

We first remark that $\tilde{\xi}_{m+1}$ is independent of $\xi_{m+2}, \dots, \xi_{m+n+1}$. Also, by Proposition 3.6.2 (applied to $X = \xi_{m+1}$ and $Y = (a - S_m)\mathbb{I}_{a - S_m \geq 0}$, which is non-negative and independent of X), conditionally on the events $S_m \leq a$ and $\xi_{m+1} > (a - S_m)$, the variable $\tilde{\xi}_{k_a+1}$ is exponentially distributed (and, of course, since these events are formulated in terms of $\xi_1, \dots, \xi_{m+1}$, they are indepentent of $\xi_{m+2}, \dots$). Therefore,

$$\begin{aligned}\mathbb{P}(Y_t = n, X_a = m) = \\ = \mathbb{P}(\tilde{\xi}_{m+1} + \xi_{m+2} + \cdots + \xi_{m+n} \leq t, \tilde{\xi}_{m+1} + \cdots + \xi_{m+n+1} > t | S_m \leq a, \xi_{m+1} > a - S_m) \times \\ \times \mathbb{P}(S_m \leq a, \xi_{m+1} > a - S_m) \\ = \mathbb{P}(X_t = n)\mathbb{P}(X_a = m).\end{aligned}$$

This shows that Y_t is a Poisson process independent of X_a. In fact, assuming that $X_a = m$, the events of the form $X_{a_i} = m_i$, where $a_i \leq a$ and $m_i \leq m$, can be expressed in terms of $\xi_i, i \leq m$. Thus, they can be inserted into the above probability, without changing the conditional distribution of $\tilde{\xi}_{m+1}, \dots, \xi_{m+n}$. This proves the "moreover" claim. □

REMARK 3.6.5. We have shown, in particular, that for any $m_1 \leq \cdots \leq m_k$ and $0 = a_0 \leq a_1 \leq \cdots \leq a_k$, we have

$$\mathbb{P}(X_{a_k} = m_k, \dots, X_{a_1} = m_1) = \mathbb{P}(X_{a_k} - X_{a_{k-1}} = m_k - m_{k-1})\mathbb{P}(X_{a_{k-1}} = m_{k-1}, \dots, X_{a_1} = m_1).$$

Iterating, we conclude that the increments $X_{a_i} - X_{a_{i-1}}$ are independent random variables.

Moreover, we can strengthen the independence statement a little bit. Observe that for a fixed a, the events of the form

$$X_{a_1} = m_1 \text{ and } \dots \text{ and } X_{a_k} = m_k$$

with varying $a_1 \leq \cdots \leq a_k \leq a$, form a π- system. This π-system generates a σ-algebra

$$\mathcal{F}_{\leq a} := \sigma(X_t : t \leq a) = \sigma(\{X_t = m\} : t \leq a, m \in \mathbb{Z}_{\geq 0}).$$

Since, as discussed in the proof of Kolmogorov's 0-1 law, the events independent of a given event form a λ- system, we conclude that any event of the form $Y_t = m$ is independent of (any event in) $\mathcal{F}_{\leq a}$. Arguing similarly, we conclude that any event in

$$\mathcal{F}^Y := \sigma(X_{a+t} - X_a : t \in \mathbb{R}_{\geq 0})$$

is independent of any event in $\mathcal{F}_{\leq a}$.

EXAMPLE 3.6.6. Consider the event

$$E_{(0;1)} := \{\exists a, b \in (0;1) : b - a > \frac{1}{2}, X_a = X_b\}$$

that in $(0;1)$, there is a subinterval of length greater than $\frac{1}{2}$ without jumps. Consider a similar event

$$E_{(1;2)} := \{\exists a, b \in (1;2) : b - a > \frac{1}{2}, X_a = X_b\}.$$

The above considerations show that $E_{(0,1)}$ and $E_{(1;2)}$ are independent. We leave the details (check that these events indeed belong to relevant σ-algebrae) to the reader.

We proceed by the study of distributions of the increments of the Poisson process.

PROPOSITION 3.6.7. *Let X_t be a Poisson process. Then, its increments have Poisson distribution with mean $\lambda(b-a)$:*

$$\mathbb{P}(X_b - X_a = m) = e^{-\lambda(b-a)} \frac{(\lambda(b-a))^m}{m!}.$$

PROOF. By Proposition 3.6.4, it is enough to consider the case $a = 0$. Fix $b > 0$. Although we could compute the distribution of X_t directly using properties of Gamma distributions, we prefer another approach. First, note that

$$\mathbb{P}(X_t \geq 2) = \mathbb{P}(\xi_1 + \xi_2 \leq t) \leq \mathbb{P}(\xi_1 \leq t)\mathbb{P}(\xi_2 \leq t) = (1 - e^{-\lambda t})^2 = O(t^2), \quad t \to 0.$$

$$\mathbb{P}(X_t = 1) = \mathbb{P}(\xi_1 < t) - \mathbb{P}(X_t \geq 2) = 1 - e^{-\lambda t} + O(t^2) = \lambda t + O(t^2), \quad t \to 0.$$

Partition the interval $(0; b)$ into n equal sub-intervals, and denote by $Y_i := X_{b\frac{i}{n}} - X_{b\frac{i-1}{n}}$ the corresponding increments of the process. Then, Y_i are indepdendent random variables. Denote

$$\hat{Y}_i := \min(Y_i; 1) = \begin{cases} 1, & Y_i \geq 1; \\ 0, & Y_i = 0, \end{cases}$$

then

$$X_b = \sum_{i=1}^n Y_i = \sum_{i=1}^n \hat{Y}_i + \sum_{i=1}^n (Y_i - \hat{Y}_i).$$

Note that $\hat{Y}_i$ are Bernoulli random variables: $p_n := \mathbb{P}(\hat{Y}_i = 1) = 1 - \mathbb{P}(\hat{Y}_i = 0) = \lambda\frac{b}{n} + O\left(\frac{1}{n^2}\right)$. Also,

$$\mathbb{P}\left(\sum_{i=1}^n (Y_i - \hat{Y}_i) \neq 0\right) \leq \mathbb{P}(\exists i : Y_i > 1) \leq \sum_{i=1}^n \mathbb{P}(Y_i > 1) = O\left(\frac{1}{n}\right).$$

Therefore,

$$\mathbb{P}(X_b = m) = \mathbb{P}\left(\sum_{i=1}^n \hat{Y}_i = m\right) + O(n^{-1}),$$

so, the question is reduced to one about independent Bernoulli random variables. Consider, for $y > 0$, the generating function

$$\sum_{m=1}^n y^m \mathbb{P}\left(\sum_{i=1}^n \hat{Y}_i = m\right) = \mathbb{E}y^{\sum_{i=1}^n \hat{Y}_i} = \mathbb{E}y^{\hat{Y}_1} \cdot \ldots \cdot \mathbb{E}y^{\hat{Y}_n} = ((1 - p_n) + p_n y)^n =: \psi_n(y)$$

From here, we can compute

$$m!\mathbb{P}\left(\sum_{i=1}^n \hat{Y}_i = m\right) = \left(\frac{\partial^m}{\partial y^m}\psi_n\right)(0) = (p_n \cdot n) \ldots (p_n(n-m))(1-p_n)^{n-m},$$

Because $p_n \sim \frac{\lambda b}{n}$, we see that each of the terms $p_n n, \ldots, p_n(n-m)$ tends to λb, and

$$(1 - p_n)^{n-m} = \left((1-p_n)^{\frac{1}{p_n}}\right)^{(n-m)p_n} \to e^{-\lambda b}.$$

So, plugging everuthing together, we get

$$\mathbb{P}\left(\sum_{i=1}^{n}\hat{Y}_i = m\right) \to e^{-\lambda b}\frac{(\lambda b)^m}{m!},$$

as required. □

REMARK 3.6.8. It follows from this result that the sum of indepdenent Poisson random variables with parameters λ_1 and λ_2 is Poisson with parameter $\lambda_1 + \lambda_2$.

REMARK 3.6.9. The statement about Bernoulli random variables that we have proven along the way is called *Poisson limit theorem*, aka "weak law of small numbers" or "law of rare events". The argument above, actually, indicates a way to refine and generalize this result, as follows. Let $Y_1', Y_2', \dots, Y_n'$ be independent Bernoulli, such that $\mathbb{P}(Y_i' = 1) = p_i$. Let X_t be a Poisson process with parameter 1, and this time, we *choose* a partitioin $t_1 \le t_2 \le \dots \le t_n$ so that $\hat{Y}_i := \min(Y_i; 1)$ is distributed as Y_i', where $Y_i := X_{t_i} - X_{t_{i-1}}$. Since we know that Y_i are Poisson with parameter $\varepsilon_i := t_i - t_{i-1}$, we infer that ε_i should be chosen from the condition

$$\mathbb{P}(\hat{Y}_i = 0) = e^{-\varepsilon_i} = 1 - p_i,$$

or $\varepsilon_i = -\log(1-p_i)$. If this condition is met, then, if the denote the distribution of $\hat{S}_n := \sum_{i=1}^n \hat{Y}_n$ by μ, and the distribution of $S_n = \sum_{i=1}^n Y_n$ by Poisson(λ), where $\lambda := -\sum_{i=1}^n \log(1-p_i)$, we have

$$\begin{aligned}
||\mu - \text{Poisson}(\lambda)||_1 &= \sum_{m=1}^{\infty}\left|\mathbb{P}\left(S_n = m\right) - \mathbb{P}\left(\sum_{i=1}^{n}\hat{S}_n = m\right)\right| = \\
&= \sum_{m=1}^{\infty}\left|\mathbb{P}\left(S_n = m, S_n \neq \hat{S}_n\right) - \mathbb{P}\left(\hat{S}_n = m, S_n \neq \hat{S}_n\right)\right| \\
&\le 2\mathbb{P}(S_n \neq \hat{S}_n) \le 2\sum_{i=1}^{n}\mathbb{P}(Y_n > 1) \\
&\le 2\sum_{i=1}^{n}(1 - e^{-\varepsilon_i} - \varepsilon_i e^{-\varepsilon_i}) = 2\sum_{i=1}^{n}(p_i + (1-p_i)\log(1-p_i)) \le 2\sum_{i=1}^{n}p_i^2
\end{aligned}$$

since $\log(1-p) \le -p$. This inequality is called Le Cam's (1960) inequality[6]

We now investigate the following question: let X_t be a Poisson process, and assume that we know X_a, that is, the number of points in $[0; a]$. What is the conditional distribution of these points? A rather striking answer is the they are *indepdent of each other, uniform* on $[0; 1]$. We prefer to put is other way, namely:

PROPOSITION 3.6.10. *Let $N, Y_1, Y_2, \dots$ be independent random variables, where N is Poisson with parameter λa, and $Y_1, Y_2, \dots$ are uniformly distributed on $(0; a)$. Denote*

$$X_t := \#\{i \le N : Y_i \le t\}.$$

Then, X_t is a Poisson process with intensity λ.

PROOF. We will check that for $0 = a_0 < a_1 < a_2 < \dots < a_n = a$, the variables $X_{a_i} - X_{a_{i-1}}$ are indepdendent Poisson random variables. To illustrate the computation, we start with the case $n = 2$. Let $m_1, m_2 \in \mathbb{N}$, and denote $m = m_1 + m_2$, $I = \{1, \dots, m\}$. We have

$$p_{m_1,m_2} = \mathbb{P}(X_{a_1} - X_{a_0} = m_1, X_{a_2} - X_{a_1} = m_2) = \mathbb{P}(N = m, \text{ exacty } m_1 \text{ of } Y_1, \dots, Y_m \text{ belong to } (a_0; a_1)).$$

[6]it is possible to replace $\lambda = -\sum_{i=1}^n \log(1-p_i)$ by $\sum_{i=1}^n p_i$ in the statement, by choosing a slightly different coupling that has no direct relation to the Poisson process.

Now, we can break this probability into a sum according to different possibilities as to which of Y_i belong to (a_0, a_1); there are $C_m^{m_1} = \frac{m!}{m_1!m_2!}$ possible choces. We encode these choices by functions $\sigma : \{1, \dots m\} \to \{1, 2\}$ and denote $J_1 := (a_0; a_1)$ and $J_2 = (a_1; a_2)$. Then

$$p_{m_1,m_2} = \sum_{\substack{\sigma: I \to \{1,2\} \\ |\sigma^{-1}(1)| = m_1}} \mathbb{P}(N = m, Y_1 \in J_{\sigma(1)}, \dots, Y_n \in J_{\sigma(n)})$$

$$\overset{\text{independence}}{=} \frac{m!}{m_1!m_2!} e^{-\lambda a} \frac{(\lambda a)^m}{m!} \left(\frac{|J_1|}{a}\right)^{m_1} \left(\frac{|J_1|}{a}\right)^{m_2} = \frac{(\lambda|J_1|)^{m_1}}{m_1!} e^{-\lambda|J_1|} \frac{(\lambda|J_2|)^{m_2}}{m_2!} e^{-\lambda|J_2|},$$

which shows that $X_{a_2} - X_{a_1}$ and $X_{a_1} - X_{a_0}$ are independent Poisson with parameters $\lambda(a_2 - a_1)$ and $\lambda(a_1 - a_0)$, respectively.

For general n, the proof similar, except that now one has to compute the number of ways to decompose I as $I = I_1 \sqcup \dots \sqcup I_n$ with $|I_i| = m_i$ for all i. There are $C_m^{m_1}$ ways to choose I_1, given that choice, there are $C_{m-m_1}^{m_2}$ ways to choose a subset I_2 of $I \setminus I_1$, etc. All in all, the number of terms is

$$C_m^{m_1} \cdot C_{m-m_1}^{m_2} \cdot \dots \cdot C_{m-m_1-\dots-m_{n-1}}^{m_n} = \frac{m!}{m_1!(m-m_1)!} \cdot \frac{(m-m_1)!}{m_2!(m-m_1-m_2)!} \cdot \dots \cdot \frac{(m-\dots-m_{n-1})!}{m_n!0!}$$

$$= \frac{m!}{m_1! \dots m_n!}.$$

(this number is called *multinomial coefficient*). We conclude that

$$\mathbb{P}(X_{a_1} - X_{a_0} - m_1, \dots, X_{a_n} - X_{a_{n-1}} = m_n) =$$

$$\frac{m!}{m_1! \dots m_n!} e^{-\lambda a} \frac{(\lambda a)^m}{m!} \prod_{i=1}^n \left(\frac{|J_i|}{a}\right)^{m_i} = \prod_{i=1}^n \frac{(\lambda|J_i|)^{m_i}}{m_i!} e^{-m_i\lambda}.$$

Now, if we consider $\xi_i := \min\{t : X_t = i\} - \min\{t : X_i = i-1\}$, the events of the form $a_1 \le \xi_1 \le b_1, \dots, a_1 \le \xi_n \le b_n$ can be expressed in terms of the values of X_t for finitely many t, therefore, their probabilities are the same as that for the gaps in the Poisson process, which shows that ξ_i are independent exponentials. □

REMARK 3.6.11. The above Proposition suggests an alternative definition of the Poisson process, that generalises nicely to the notion of Poisson processes on arbitrary (σ-finite) measure spaces.

CHAPTER 4

Conditional expectations and martingales

4.1. Conditional expectation: motivation and definition

To motivate what follows, suppose one wants to extend the theory of Markov process to the case of *uncountable* state space S, for example, $S = \mathbb{R}$. One runs into an immediate problem: the Markov property

$$\mathbb{P}(X_{t+1} = x_{t+1} | X_t = x_t, \dots, X_0 = x_0) = \mathbb{P}(X_{t+1} = x_{t+1} | X_t = x_t)$$

does not make sense any more, because the event $X_t = x_t$ typically has zero probability for a continuous random variable X_t. Therefore, we need a more advanced theory of conditional expectation (and probability).

First of all, let us give (or recall) the definition of elementary conditional expectation. Given events A, B with $\mathbb{P}(A) > 0$, the conditional probability was defined as

$$\mathbb{P}(B|A) = \frac{\mathbb{P}(A \cap B)}{\mathbb{P}(A)} = \frac{\mathbb{E}(\mathbb{I}_A \mathbb{I}_B)}{\mathbb{P}(A)}.$$

The general definition of elementary conditional expectation is just replacing $\mathbb{I}_B$ by arbitrary random variable

$$\mathbb{E}(X|A) = \frac{\mathbb{E}(X\mathbb{I}_A)}{\mathbb{P}(A)}.$$

In other words, this is just the expectation with respect to the original probability measure, restricted to A and normalized to be a probability measure again.

Let Y be a scalar random variable (e. g., $\mathbb{I}_A$ for an event A), and let X be a random variable with values in a finite or countable space S. When studying Markov chain, conditional expactations typically arose in formulae like this:

$$\mathbb{E}(Y) = \sum_{x \in S} \mathbb{E}(Y, X = x) = \sum_{x \in S} \mathbb{E}(Y|X = x)\mathbb{P}(X = x).$$

This suggests that we are interested in the whole collection $\{\mathbb{E}(Y|X = s)\}_{s \in S}$ rather than in individual numbers $\mathbb{E}(Y|X = s)$. Moreover, the expression in the right-hand side of the above formula looks like an expectation of a random variable: $\mathbb{E}(h(X)) = \sum_{x \in S} h(x)\mathbb{P}(X = x)$. We can *define* $\mathbb{E}(Y|X)$ to be that random variable; the last formula becomes

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

An intuitive way to think about $\mathbb{E}(Y|X)$ is that after an experiment is performed (that is, some outcome $\omega \in \Omega$ has realized), we are told the value of $X(\omega)$, and we are trying to guess the value of $Y(\omega)$; $\mathbb{E}(Y|X)$ is our best guess. Of course, our best guess depends on the outcome ω (through the value $X(\omega)$ that we are told), that is, it in itself is a random variable.

Note that the random variable $\mathbb{E}(Y|X)$ is a function of X; equivalently, it is constant on each set $\{X = x\}$. Yet another way to spell out this property is as follows. Define $\sigma(X) \subset \mathcal{F}$ to be the smallest σ-algebra $\mathcal{G}$ such that X is $\mathcal{G}$-to-2^S-measurable. Then, we have the following

Claim. A scalar random variable Z is a function of X if and only if it is $\sigma(X)$-to-$\mathcal{B}(\mathbb{R})$-measurable.

PROOF. Indeed, the "only if" condition follows from the fact that any function is 2^S-to-anything measurable, and composition of two measurable functions is measurable. For the "if" part, note that $\sigma(X)$ has the following structure:

$$\sigma(X) = \{\{\omega : X(\omega) \in S'\}\}_{S' \subset S}.$$

(All "atoms", or "level sets of X" $\{\omega : X(\omega) = x\}$ belong to $\sigma(X)$; therefore, all (disjoint) unions of such atoms belong to $\sigma(X)$, since S is at most countable). Let $t \in \mathbb{R}$, and consider $Z^{-1}(\{t\}) \in \sigma(X)$. Since $Z^{-1}(\{t\}) \cap \{\omega : X(\omega) = x\}$ is measurable for any x, it is either empty, or coincides with $\{\omega : X(\omega) = x\}$. In other words, Z is constant on each atom $\{X = x\}$, therefore, $h(x) := Z(\omega_x)$, where ω_x is any element of $\{\omega : X(\omega) = x\}$, is well-defined and satisfies $h(X) \equiv Z$. □

The conclusion of this discussion is that

$$\mathbb{E}(Y|X) \text{ is } \sigma(X)\text{-measurable.} \tag{4.1.1}$$

This condition encompasses the property that $\mathbb{E}(Y|X)$ is a constant on $\{X = x\}$, but doesn't yet take into account which constant it is. In fact, we know by definition that

$$\mathbb{I}_{X=x}\mathbb{E}(Y|X) = \mathbb{I}_{X=x}\mathbb{E}(Y|X = x) = \mathbb{I}_{X=x}\frac{\mathbb{E}(Y\mathbb{I}_{X=x})}{\mathbb{P}(X = x)}.$$

Taking expectation, we get

$$\mathbb{E}(\mathbb{I}_{X=x}\mathbb{E}(Y|X)) = \mathbb{E}(\mathbb{I}_{X=x}Y).$$

By linearity, we can replace $\mathbb{I}_{X=x}$ with $\mathbb{I}_{X\in S'} = \sum_{x\in S'} \mathbb{I}_{X=x}$ in the above formula. Therefore,

$$\mathbb{E}(\mathbb{I}_A\mathbb{E}(Y|X)) = \mathbb{E}(\mathbb{I}_A Y) \text{ for every } A \in \sigma(X). \tag{4.1.2}$$

In fact, we will see in a moment that the properties (4.1.1) and (4.1.2) characterize the conditional expectation uniquely[1]. For now, we remark that *these two conditions only depend on $\sigma(X)$ and not on X.* Therefore, in our general definition, we will define the conditional expectation conditionally on a σ-algebra rather than on a random variable.

DEFINITION 4.1.1. Suppose $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability space, X a random variable satisfying $\mathbb{E}|X| < \infty$, and $\mathcal{G} \subset \mathcal{F}$ is a σ-algebra. Then, $\mathbb{E}(X|\mathcal{G})$ is any random variable that is $\mathcal{G}$ -measurable and statisfies

$$\mathbb{E}(\mathbb{I}_A\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(\mathbb{I}_A X)$$

for any $A \in \mathcal{G}$. We also write $\mathbb{E}(Y|X)$ for $\mathbb{E}(Y|\sigma(X))$.

REMARK 4.1.2. Taking $A = \Omega$ in the definition, we note that by definition, $\mathbb{E}(X|\mathcal{G})$ is integrable ($\mathbb{E}|\mathbb{E}(X|\mathcal{G})| < \infty$).

Once again, intuitively, the σ-algebra $\mathcal{G}$ is supposed to represent the knowledge given to us (for every $A \in \mathcal{G}$, we know whether A happened or not), while $\mathbb{E}(X|\mathcal{G})$ is our best guess given that knowledge.

Before proving existence and uniqueness, we collect the following facts about conditional expectation (similar to the properties of the usual expectation, cf. Proposition 1.5.1)

PROPOSITION 4.1.3. *The conditional expectation satisfies the following properties:*

1. *(Linearity) If $\mathbb{E}(X|\mathcal{G})$ exists and $\mathbb{E}(Y|\mathcal{G})$ exists, then $\mathbb{E}(\alpha X + \beta Y|\mathcal{G})$ exists and equals $\alpha\mathbb{E}(X|G) + \beta\mathbb{E}(Y|\mathcal{G})$ almost surely;*
2. *(Monotonicity) If $X \geq 0$ almost surely, then $\mathbb{E}(X|\mathcal{G}) \geq 0$ almost surely;*
3. *(Monotone convergence) If $X_i \geq 0$, $X_i \nearrow X$ almost surely, $\mathbb{E}(X_i|\mathcal{G})$ exists for all i, and $\mathbb{E}(|X|) < \infty$, then $\mathbb{E}(X|\mathcal{G})$ exists and is equal to $\lim_{i\to\infty}\mathbb{E}(X_i|\mathcal{G})$ almost surely.*

[1] up to set of probability 0

PROOF. (1) Clearly, $\alpha\mathbb{E}(X|\mathcal{G}) + \beta\mathbb{E}(Y|\mathcal{G})$ is $\mathcal{G}$-measurable, and for any $A \in \mathcal{G}$,

$$\mathbb{E}(\mathbb{I}_A(\alpha\mathbb{E}(X|G)+\beta\mathbb{E}(Y|\mathcal{G}))) = \alpha\mathbb{E}(\mathbb{I}_A\mathbb{E}(X|\mathcal{G}))+\beta\mathbb{E}(\mathbb{I}_A\mathbb{E}(Y|\mathcal{G})) = \alpha\mathbb{E}(\mathbb{I}_A X)+\beta\mathbb{E}(\mathbb{I}_A Y) = \mathbb{E}(\mathbb{I}_A(\alpha X+\beta Y)).$$

(2) Let $A_\varepsilon := \{\mathbb{E}(X|\mathcal{G}) < \varepsilon < 0\}$, then $A \in \mathcal{G}$, and

$$-\varepsilon\mathbb{P}(A_\varepsilon) \geq \mathbb{E}(\mathbb{I}_{A_\varepsilon}\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(\mathbb{I}_{A_\varepsilon}X) \geq 0,$$

which means that $\mathbb{P}(A_\varepsilon) = 0$ for any $\varepsilon > 0$, and hence $\mathbb{P}(\mathbb{E}(X|\mathcal{G}) < 0) = \mathbb{P}(\cup_{n\in\mathbb{N}} A_{\frac{1}{n}}) = 0$.

(3) By (1) and (2), the sequence $\mathbb{E}(X_i|\mathcal{G})$ is almost surely increasing. Define $Y := \limsup \mathbb{E}(X_i|\mathcal{G})$, then Y is $\mathcal{G}$-measurable, and $\mathbb{E}(X_i|\mathcal{G}) \nearrow Y$ almost surely. For any $A \in \mathcal{G}$, monotone convergence theorem, applied twice, implies

$$\mathbb{E}(\mathbb{I}_A Y) = \lim_{i\to\infty} \mathbb{E}(\mathbb{I}_A\mathbb{E}(X_i|\mathcal{G})) = \lim_{i\to\infty} \mathbb{E}(\mathbb{I}_A X_i) = \mathbb{E}(\mathbb{I}_A X),$$

as required. □

COROLLARY 4.1.4. *Suppose that Y and Y' are two random variables satisfying the definition of $\mathbb{E}(X|\mathcal{G})$. Then, $Y = Y'$ almost surely.*

PROOF. Since $X \geq X$, monotonicity implies $Y \geq Y'$ almost surely and $Y' \geq Y$ almost surely. □

REMARK 4.1.5. Since conditional expectation, in general, is only defined up to a set of measure 0, one usually says that a variable Y satisfying Definition 4.1.1 is a version of the conditional expectation of X, or $Y = \mathbb{E}(X|\mathcal{G})$ almost surely.

4.2. Examples and some properties of conditional expactation

To get accustomed with the definitions, let us consider some simple examples.

First, consider $\mathcal{G} = \mathcal{F}$. Informally, this means that we are given all possible information about the outcome ω of the experiment, which means that, in particular, we know $X(\omega)$. Clearly, in this case our best guess about the value of X must be X itself. And indeed, since $\mathcal{G} = \mathcal{F}$, X is $\mathcal{G}$-measurable, and it clearly satisfies the identity $\mathbb{E}(\mathbb{I}_A X) = \mathbb{E}(\mathbb{I}_A X)$. Therefore, $\mathbb{E}(X|\mathcal{F}) = X$. The same applies whenever X is $\mathcal{G}$-measurable.

Second, consder the case $\mathcal{G} = \{\emptyset, \Omega\}$ (we are given no information about the outcome of the experiment). Then the only functions that are measurable w. r. t. $\mathcal{G}$ are constants. Among those constants, $\mathbb{E}X$ should be our best guess about the value of X. And indeed, we have

$$\mathbb{E}(\mathbb{I}_\emptyset\mathbb{E}X) = 0 = \mathbb{E}(\mathbb{I}_\emptyset X); \quad \mathbb{E}(\mathbb{I}_\Omega\mathbb{E}X) = \mathbb{E}(\mathbb{E}X) = \mathbb{E}X = \mathbb{E}(\mathbb{I}_\Omega X),$$

so that $\mathbb{E}(X|\{\emptyset;\Omega\}) = \mathbb{E}X$.

This example can be generalized. Assume that $\mathcal{G}$ is independent of X, that is, $\mathbb{I}_A$ and X are independent for any $A \in \mathcal{G}$. We claim that in this case also $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}X$. (This is very natural: if the information we are given is independent of what we are interested in, it is as if we are given no information at all.) Since $\mathbb{E}X$ is constant, it is $\mathcal{G}$-measurable for any $\mathcal{G}$. Then, if $\mathbb{I}_A$ is independent of X, then

$$\mathbb{E}(\mathbb{I}_A\mathbb{E}X) = \mathbb{E}X\mathbb{E}\mathbb{I}_A = \mathbb{E}(\mathbb{I}_A X),$$

as required.

We collect these and some more properties in the following proposition:

PROPOSITION 4.2.1. *The conditional expectation satisfies the following properties:*

1. *If X is $\mathcal{G}$-measurable, then $\mathbb{E}(X|\mathcal{G}) = X$ a. s.*
2. *If X is independent of $\mathcal{G}$, then $\mathbb{E}(X|\mathcal{G}) = \mathbb{E}X$ a. s.*
3. *(Tower property) If $\mathcal{G}_1 \subset \mathcal{G}_2$, then*

$$\mathbb{E}(\mathbb{E}(X|\mathcal{G}_2)|\mathcal{G}_1) = \mathbb{E}(X|\mathcal{G}_1).$$

In particular, $\mathbb{E}(\mathbb{E}(X|\mathcal{G})) = \mathbb{E}X$.

(4) *(Taking out what is known) If Y is a bounded, $\mathcal{G}$-measurable random variable and $\mathbb{E}(X|\mathcal{G})$ exists*[2]*, then*

$$\mathbb{E}(YX|\mathcal{G}) = Y\mathbb{E}(X|\mathcal{G}).$$

The same is true Y is unbounded, but $\mathbb{E}X^2 < \infty$ and $\mathbb{E}Y^2 < \infty$.

Proof. The first two properties are already proven above. For the fourth property, observe first that if Z is bounded and $\mathcal{G}$-measurable, then

$$\mathbb{E}(Z\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(ZX). \tag{4.2.1}$$

Indeed, for $Z = \mathbb{I}_A$ with $A \in \mathcal{G}$, this follows from the definition; by linearity, it is true for all simple Z, and by monotone convergence, for all bounded Z. Note that if Z is bounded and $\mathbb{E}X$ exists, then $\mathbb{E}(ZX)$ exists.

Using this identity, we check that $Y\mathbb{E}(X|\mathcal{G})$ satisfies the definition of $\mathbb{E}(XY|\mathcal{G})$. Indeed, it is $\mathcal{G}$-measurable as a product of two $\mathcal{G}$-measurable functions, and for any $A \in \mathcal{G}$, applying (4.2.1) to $Z = \mathbb{I}_A Y$, we get

$$\mathbb{E}(\mathbb{I}_A Y\mathbb{E}(X|\mathcal{G})) = \mathbb{E}(\mathbb{I}_A YX),$$

as required. The boundedness of Y was only used to ensure existence of $\mathbb{E}(XY)$, which also holds true whenever $\mathbb{E}X^2 < \infty$ and $\mathbb{E}Y^2 < \infty$.

To deduce the tower property, we first note that the "in particular" claim follows readily by taking $A = \Omega \in \mathcal{G}$ in the definition of conditional expectation. Further, by definition, $\mathbb{E}(\mathbb{E}(X|\mathcal{G}_2)|\mathcal{G}_1)$ is $\mathcal{G}_1$-measurable, and

$$\mathbb{E}(\mathbb{I}_A\mathbb{E}(\mathbb{E}(X|\mathcal{G}_2)|\mathcal{G}_1)) \overset{(1)}{=} \mathbb{E}(\mathbb{E}(\mathbb{I}_A\mathbb{E}(X|\mathcal{G}_2)|\mathcal{G}_1)) \overset{(2)}{=} \mathbb{E}(\mathbb{I}_A\mathbb{E}(X|\mathcal{G}_2)) \overset{(3)}{=} \mathbb{E}(\mathbb{I}_A X).$$

In (1), we used that $\mathbb{I}_A\mathbb{E}(Y|\mathcal{G}) = \mathbb{E}(\mathbb{I}_A Y|\mathcal{G})$ for a $\mathcal{G}$-measurable A, applied to $Y = \mathbb{E}(X|\mathcal{G}_2)$ and $\mathcal{G} = \mathcal{G}_1$ ("taking what is known out of conditional expectation"). In (2), we use that $\mathbb{E}(\mathbb{E}(Y|\mathcal{G})) = \mathbb{E}Y$, applied to $Y = \mathbb{I}_A\mathbb{E}(X|\mathcal{G}_2)$ and $\mathcal{G} = \mathcal{G}_1$. In (3), we use the definition of $\mathbb{E}(X|\mathcal{G}_2)$; note that since $A \in \mathcal{G}_1$, also $A \in \mathcal{G}_2$. □

We now connect the our definition of conditional expectation with the classical one. One approach to defining the conditional expectation would be to use a limiting procedure. Assume that X, Y are scalar random variables, so that $(X; Y)$ is a random vector with a (nice enough) density $f(x, y)$. For a measurable function $h : \mathbb{R} \to \mathbb{R}$, we can then try to define the conditional expectation as a limit:

$$\mathbb{E}(h(X)|Y = t) := \lim_{\varepsilon\to 0} \mathbb{E}(h(X)|Y \in I_\varepsilon),$$

where $I_\varepsilon = (t - \varepsilon, t + \varepsilon)$. Using the formula for the density of (X, Y), we obtain

$$\mathbb{E}(h(X)|Y \in I_\varepsilon) = \frac{\mathbb{E}(h(X)\mathbb{I}_{Y\in I_\varepsilon})}{\mathbb{P}(Y \in I_\varepsilon)} = \frac{\int_{\mathbb{R}^2} \mathbb{I}_{y\in I_\varepsilon} h(x)f(x, y)d\lambda^2(x, y)}{\int_{\mathbb{R}^2} \mathbb{I}_{y\in I_\varepsilon} f(x, y)d\lambda^2(x, y)} \overset{\text{Fubini}}{=} \frac{\int_{\mathbb{R}} h(x)\left(\int_{t-\varepsilon}^{t+\varepsilon} f(x, y)dy\right)dx}{\int_{\mathbb{R}^2}\left(\int_{t-\varepsilon}^{t+\varepsilon} f(x, y)dy\right)dx}.$$

If f is continuous, then $\frac{1}{\varepsilon}\int_{t-\varepsilon}^{t+\varepsilon} f(x, y)dy \to f(x, t)$. Therefore, multiplying the numerator and the denominator by $\frac{1}{\varepsilon}$, we get

$$\lim_{\varepsilon\to 0} \frac{\int_{\mathbb{R}} h(x)\left(\int_{t-\varepsilon}^{t+\varepsilon} f(x, y)dy\right)dx}{\int_{\mathbb{R}^2}\left(\int_{t-\varepsilon}^{t+\varepsilon} f(x, y)dy\right)dx} = \frac{\int_{\mathbb{R}} h(x)\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left(\int_{t-\varepsilon}^{t+\varepsilon} f(x, y)dy\right)dx}{\int_{\mathbb{R}}\lim_{\varepsilon\to 0}\frac{1}{\varepsilon}\left(\int_{t-\varepsilon}^{t+\varepsilon} f(x, y)dy\right)dx} = \frac{\int_{\mathbb{R}} h(x)f(x, t)dx}{\int_{\mathbb{R}} f(x, t)dx},$$

provided that $\int_{\mathbb{R}} f(x, t) \neq 0$. (Exercise: justify the exchange of the integral and the limit!). So, we arrive at a putative formula

$$\mathbb{E}(h(X)|Y = t) \overset{?}{=} \frac{\int_{\mathbb{R}} h(x)f(x, t)dx}{\int_{\mathbb{R}} f(x, t)dx}.$$

Let us check that this formula indeed gives the conditional expectation in the sense of our general definition above.

[2] in the next section, we will prove that it always exists

EXAMPLE 4.2.2. Suppose a random vector (X, Y) has a Borel-measurable density $f(x, y)$, and let $h : \mathbb{R} \to \mathbb{R}$ be a Borel measurable function such that $\mathbb{E}|h(X)| < \infty$. Define

$$H(t) := \begin{cases} \frac{\int_{\mathbb{R}} h(x)f(x,t)dx}{\int_{\mathbb{R}} f(x,t)dx}, & \text{if } \int_{\mathbb{R}} f(x,t)dx > 0 \\ 0, & \text{else.} \end{cases}$$

Then, almost surely,

$$\mathbb{E}(h(X)|\sigma(Y)) = H(Y).$$

To check this, we need the following technical lemma:

LEMMA 4.2.3. *H is Borel measurable.*

Taking this lemma for granted, we first note that $H(Y)$ is $\sigma(Y)$-measurable. Then, we note that

$$\sigma(Y) = \{\{\omega : Y(\omega) \in A\} | A \in \mathcal{B}(\mathbb{R})\}$$

(check this!) Therefore, thr following computation justifies the result:

$$\begin{aligned} \mathbb{E}(\mathbb{I}_{Y\in A}H(Y)) = \int_{(x,y)\in\mathbb{R}\times A} H(y)f(x,y)d\lambda^2(x,y) &= \int_A H(y)\left(\int_{\mathbb{R}} f(x,y)dx\right)dy \\ &= \int_A \left(\int_{\mathbb{R}} h(x)f(x,y)dx\right)dy = \mathbb{E}(\mathbb{I}_{Y\in A}h(X)). \end{aligned}$$

PROOF OF LEMMA 4.2.3. It suffices to do the case $h \geq 0$ (otherwise use linearity). Since h and f are assumed Borel measurable, hf is also Borel measurable, and the subgraph

$$S = \{(t; x; y) : 0 \leq t < h(x)f(x, y)\} = \cup_{q\in\mathbb{Q}_{\geq 0}}[0; q) \times \{(x, y) : h(x)f(x, y) > q\}$$

is a Borel measurable set. Therefore, Cavalieri's principle (non-complete version) implies that the function

$$y \mapsto \int_{\mathbb{R}} h(x)f(x, y)dx = \lambda^2(\{(t, x) : (t; x; y) \in S\})$$

is Borel measurable. Similarly, $y \mapsto \int_{\mathbb{R}} f(x, y)dx$ is Borel measurable, and, from this it is not hard to check from definition that if g is a measurable function, then

$$y \mapsto \begin{cases} \frac{1}{g(y)}, & g(y) \neq 0 \\ 0, & \text{otherwise,} \end{cases}$$

is also measurable. Since product of two measurable functions is measurable, this implies the desired result. □

4.3. $L^2(\Omega)$ and existence of conditional expectation

We will first construct conditional expectation for square-integrable random variables. Denote

$$\mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P}) := \{X \in \mathcal{F}, \mathbb{E}X^2 < \infty\}.$$

Hereinafter the notation $X \in \mathcal{F}$ means that X is $\mathcal{F}$-to-$\mathcal{B}(\mathbb{R})$-measurable.

The idea behing the construction of the conditional expectation can be illustrated on the trivial case $\mathcal{G} = \{\emptyset, \Omega\}$. In that case, Y is $\mathcal{G}$-measurable if and only if Y is constant, and the conditional expectation is just the usual expectation (Exercise: check that!).

Let $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$. Then, $\mathbb{E}(X - c)^2 = \mathbb{E}X^2 - 2c\mathbb{E}X + c^2$ is a quadratic function of c that attains its minimum at $c = \mathbb{E}X$. Therefore, we can characterize $\mathbb{E}X$ as the function $Y \in \mathcal{G}$ such that $\mathbb{E}(X - Y)^2$ is minimal among all $\mathcal{G}$-measurable functions Y.

This argument extends to arbitrary $\mathcal{G}$:

LEMMA 4.3.1. *Let $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$, and assume that $Y \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$ is such that*

$$\mathbb{E}(X - Y)^2 \leq \mathbb{E}(X - Y')^2$$

for any $\mathcal{G} \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$. Then, Y is a version of $\mathbb{E}(X|\mathcal{G})$.

PROOF. Let $A \in \mathcal{G}$. Consider the quadratic function

$$c \mapsto \mathbb{E}(X - Y - c\mathbb{I}_A)^2 = \mathbb{E}(X-Y)^2 - 2c\mathbb{E}(\mathbb{I}_A(X-Y)) + c^2\mathbb{P}(A).$$

Since $Y + c\mathbb{I}_A$is $\mathcal{G}$-measurable, and $\mathbb{E}(Y + c\mathbb{I}_A)^2 < \infty$, this quadratic function chould have a minimum at $c = 0$, that is,

$$\mathbb{E}(\mathbb{I}_A(X-Y)) = 0,$$

as required. □

It remains to assure that the $\min_{Y \in \mathcal{L}^2(\Omega,\mathcal{G},\mathbb{P})} \mathbb{E}(X-Y)^2$ is always attained. To this end, we need a bit of functional analysis. First, we remark that $\mathcal{L}^2$ has a scalar product

$$\mathbb{E}(XY).$$

By Cauchy-Scwarz inequality, this is always $\leq (\mathbb{E}X^2)^{\frac{1}{2}}(\mathbb{E}Y^2)^{\frac{1}{2}}$ (and, in particular, it is finite for all $\mathcal{L}^2(\Omega,\mathcal{G},\mathbb{P})$). This scalar product defines a semi-norm $||X||_2 := \sqrt{\mathbb{E}X^2}$; $||X||_2 = 0$ if and only if $X = 0$ almost surely. This is indeed a semi-norm:

$$||X+Y||_2^2 = \mathbb{E}(X+Y)^2 = \mathbb{E}X^2 + 2\mathbb{E}XY + \mathbb{E}Y^2 \overset{\text{Cauchy-Schwarz}}{\leq} \mathbb{E}X^2 + 2\sqrt{\mathbb{E}X^2}\sqrt{\mathbb{E}Y^2} + \mathbb{E}Y^2 = (||X||_2 + ||Y||_2)^2.$$

This semi-norm can be made into a norm by identifying functions that agree almost surely; however, we will not do it here for technical reasons related to the fact that underlying σ-algebras need not be complete.

Anyway, recall that, given $X \in \mathcal{L}^2(\Omega,\mathcal{F},\mathbb{P})$ and a sequence $X_i \in \mathcal{L}^2(\Omega,\mathcal{G},\mathbb{P})$, we say that

$$X_i \to X \text{ in } \mathcal{L}^2$$

if $||X - X_i||_2 \to 0$ as $i \to \infty$. We say that a sequence X_i is Cauchy if for every $\varepsilon > 0$, there exists N such that $||X_i - X_j||_2 < \varepsilon$ for all $i, j \geq N$. The following important Proposition shows that Cauchy sequences converge:

PROPOSITION 4.3.2. *(Completeness of $\mathcal{L}^2$) Let $X_i \in \mathcal{L}^2(\Omega,\mathcal{F},\mathbb{P})$ be a Cauchy sequence. Then, there exists $X \in \mathcal{L}^2(\Omega,\mathcal{F},\mathbb{P})$ such that $X_i \to X$ in $\mathcal{L}^2$.*

PROOF. First, we prove that if Y_i is such that $\sum_{i=1}^{\infty} ||Y_i||_2 < \infty$, then there exists $Y \in \mathcal{L}^2(\Omega,\mathcal{F},\mathbb{P})$ such that

$$\sum_{i=1}^{N} Y_i \to Y \text{ in } \mathcal{L}^2.$$

Indeed, consider instead the sequence $\sum_{i=1}^{N} |Y_i|$. We have, by monotone convergence theorem,

$$\left\| \lim_{N\to\infty} \sum_{i=1}^{N} |Y_i| \right\|_2 = \lim_{N\to\infty} \left\| \sum_{i=1}^{N} |Y_i| \right\|_2 \leq \lim_{N\to\infty} \sum_{i=1}^{N} ||Y_i||_2 = \sum_{i=1}^{\infty} ||Y_i||_2 < \infty,$$

in particular, the series $\sum_{i=1}^{\infty} |Y_i(\omega)|$ converges to a finite limit for almost all ω. Then, the series $\sum_{i=1}^{\infty} Y_i(\omega)$ converges to a finite limit for almost all ω. Now, if we define $Y := \limsup Y_i$, then Y is $\mathcal{F}$-measurable,[3] and $Y_i - Y \to 0$ almost surely. Moreover, almost surely,

$$\left| \sum_{i=1}^{N} Y_i - Y \right| \leq \left| \sum_{i=1}^{N} Y_i \right| + |Y| \leq 2 \sum_{i=1}^{\infty} |Y_i| \in \mathcal{L}^2,$$

which means that, by Dominated convergence theorem, $|| \sum_{i=1}^{N} Y_i - Y||_2 \to 0$, as required.

To deduce the Proposition, we choose a subsequence i_n such that $||X_i - X_j||_2 \leq 2^{-n}$ for $i, j \geq i_n$. Then, put $Y_n := X_{i_{n+1}} - X_{i_n}$; certainly, $\sum_{n=1}^{\infty} ||Y_n||_2 \leq \sum_{n=1}^{\infty} 2^{-n} < \infty$, hence

$$X_{i_{n+1}} = X_{i_0} + \sum_{i=1}^{N} Y_i \to X_{i_0} + Y =: X \text{ in } \mathcal{L}^2$$

[3] because $Y < c$ if and only if $\exists N : Y_i < c \forall i \geq N$, that is, $\{\omega \in \Omega : Y(\omega) < c\} = \cup_{N\in\mathbb{N}} \cap_{i=N}^{\infty} \{Y_i < c\}$, which is measurable.

Moreover, if $r > i_n$, then

$$||X_r - X||_2 \leq ||X_r - X_{i_n}||_2 + ||X_{i_n} - X||_2 \leq 2^{-n} + ||X_{i_n} - X||_2 \to 0,$$

as $n \to \infty$, that is, $X_i \to X$ in $\mathcal{L}^2$. □

PROPOSITION 4.3.3. *Let $X \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$. Then, for any σ-algebra $\mathcal{G} \subset \mathcal{F}$, there exists $\mathbb{E}(X|\mathcal{G}) \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$.*

REMARK 4.3.4. Note that $\mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P}) \subset \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$ only differ in that they contain functions measurable with respect to different σ-algebras.

PROOF. Let $Y_1, Y_2, \cdots \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$ be such that $||X - Y_i||_2 \to \inf_{Y \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})} ||X - Y_i||_2 =: M$. Let us check that Y_i is a Cauchy sequence. To this end, we use polarization identity, proved simply by expanding the squares:

$$||F + G||_2^2 + ||F - G||_2^2 = 2||F||_2^2 + 2||G||_2^2$$

for all $F, G \in \mathcal{L}^2(\Omega, \mathcal{F}, \mathbb{P})$. Given $\varepsilon > 0$, we take N such that $||X - Y_i||_2 \leq M + \varepsilon$ for all $i \geq N$. Then, plugging $F = X - Y_i$ and $G = X - Y_j$ with $i, j \geq N$, we obtain

$$\frac{1}{2}||Y_i - Y_j||_2 = ||X - Y_i||_2 + ||X - Y_j||_2 - 2\left\|X - \frac{Y_i + Y_j}{2}\right\|_2 \leq 2(M + \varepsilon)^2 - 2M^2 \leq 4\varepsilon M + 2\varepsilon^2.$$

Here we used that $\frac{Y_i + Y_j}{2} \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$, and hence $\left\|X - \frac{Y_i + Y_j}{2}\right\|_2 \geq M$ by definition of M. Therefore, Y_i is indeed a Cauchy sequence, and hence by Proposition 4.3.2 it converges in $\mathcal{L}^2$ to $Y \in \mathcal{L}^2(\Omega, \mathcal{G}, \mathbb{P})$. Then, also $||X - Y_i||_2 \to ||X - Y||_2$, so $||X - Y||_2 = M$, and hence $Y = \mathbb{E}(X|\mathcal{G})$ by Lemma 4.3.1. □

From this, it is not hard to derive existence of conditional expectations in the general case.

THEOREM 4.3.5. *Let X be a random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, such that $\mathbb{E}|X| < \infty$, and let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra. Then, a conditional expectation $\mathbb{E}(X|\mathcal{G})$ exists.*

PROOF. Since we can write $X = X\mathbb{I}_{X \geq 0} + X\mathbb{I}_{X \leq 0}$, Proposition 4.1.3 (1) implies that we may assume $X \geq 0$. Then, $X_N := X\mathbb{I}_{X \leq N} \nearrow X$, and $\mathbb{E}(X_N^2) \leq N^2 < \infty$. Therefore, by Proposition 4.3.3, $\mathbb{E}(X_N|\mathcal{G})$ exist, and Proposition 4.1.3 allows to conclude. □

4.4. Regular conditional distribution

Given a random variable X and a sub-sigma-algebra $\mathcal{G} \subset \mathcal{F}$, the existence of conditional expectation allows one to define $\mathbb{E}(h(X)|\mathcal{G})$ for every nice enough (say, bounded and measurable) function h. In this section, we explore the question as to whether it makes sense to speak about a *conditional distribution* of a random variable X given $\mathcal{G}$, that encapsulates these conditional expectations for all h simultaneously.

DEFINITION 4.4.1. Given a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, a measurable space $(\Omega', \mathcal{F}')$, a random variable $X : \Omega \to \Omega'$ and a sigma-algebra $\mathcal{G} \subset \mathcal{F}$, we say that a function $\mu : \Omega \times \mathcal{F}' \mapsto \mathbb{R}$ is a *regular conditional distribution* (RCD) of X given $\mathcal{G}$ if

(1) for every fixed $A \in \mathcal{F}'$, we have that $\mu(\cdot, A) = \mathbb{E}(\mathbb{I}_{X \in A}|\mathcal{G})$ almost surely;
(2) for almost every fixed $\omega \in \Omega$, we have that $\mu(\omega, \cdot)$ is a probability measure on $\mathcal{F}'$.

REMARK 4.4.2. The second condition asserts, in particular, that

> Almost surely, for every sequence $A_1, A_2, \cdots \in \mathcal{F}'$ of disjoint sets, $\mu(\omega, \sqcup_{i=1}^\infty A_i) = \sum_{i=1}^\infty \mu(\omega, A_i)$.

This is not the same as

> For every sequence $A_1, A_2, \cdots \in \mathcal{F}'$ of disjoint sets, almost surely $\mu(\omega, \sqcup_{i=1}^\infty A_i) = \sum_{i=1}^\infty \mu(\omega, A_i)$.

The two properties are not the same because the set of sequences $A_1, A_2, \dots$ is (usually) uncountable. Note that the latter statement would follow readily from $\mu(\cdot, A) = \mathbb{E}(\mathbb{I}_{X\in A}|\mathcal{G})$ a. s. and the properties of conditional expectation (Proposition 4.1.3):

$$\mathbb{E}(\mathbb{I}_{X\in\sqcup_{i=1}^{\infty}A_i}|\mathcal{G}) = \mathbb{E}(\lim_{N\to\infty}\mathbb{I}_{X\in\sqcup_{i=1}^{N}A_i}|\mathcal{G}) \overset{\text{Monotone convergence}}{=} \lim_{N\to\infty}\mathbb{E}(\mathbb{I}_{X\in\sqcup_{i=1}^{N}A_i}|\mathcal{G}) = \lim_{N\to\infty}\sum_{i=1}^{\infty}\mathbb{E}(\mathbb{I}_{X\in A_i}|\mathcal{G})$$

almost surely.

The R. C. D. do not always exist, but they do exist in most practical cases. Here we prove that they do exist for scalar random variables.

THEOREM 4.4.3. *If $(\Omega', \mathcal{F}') = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$, then the regular conditional distribution exists, and it is unique in the following sense: if μ and ν are two R. C. D., then, for almost every ω, almost surely, $\mu(\omega, \cdot) \equiv \nu(\omega, \cdot)$.*

PROOF. Define, for every $q \in \mathbb{Q}$, a random variable $G_q : \Omega \mapsto \mathbb{R}_{\geq 0}$ by

$$G_q = \mathbb{E}(\mathbb{I}_{x\in(-\infty,q]}|\mathcal{G}).$$

Further, for every $x \in \mathbb{R}$, define

$$F(x, \omega) = \inf_{q\in\mathbb{Q}, q>x} G_q(\omega).$$

Note that $F(x, \cdot)$ is measurable because as an infimum of countably many measurable random variables. By monotonicity of conditional expectation (Proposition 4.1.3), we have that for all $q_1 < q_2$, we have

$$G_{q_1} \leq G_{q_2} \text{ almost surely.}$$

Since there are only countably many pairs $(q_1, q_2) : q_{1,2} \in \mathbb{Q}$, from this we deduce that for almost all ω, $G_q(\omega)$ is increasing in q. Also, by monotone convergence theorem for conditional expectations, almost surely

$$\lim_{q\to+\infty} G_q = \lim_{q\to+\infty}\mathbb{E}(\mathbb{I}_{x\in(-\infty,q]}|\mathcal{G}) = \mathbb{E}(\lim_{q\to+\infty}\mathbb{I}_{x\in(-\infty,q]}|\mathcal{G}) = \mathbb{E}(1|\mathcal{G}) = 1,$$

and, similarly, $\lim_{q\to-\infty} G_q = 0$.

From this, we see that for almost every $\omega \in \Omega$, $F(\cdot, \omega)$ is increasing, right-continuous, and has limit 1 (resp. 0) at plus infinity (resp., minus infinity).Therefore, it is a distribution function of a Borel probability measure on $\mathbb{R}$. We define

$$\mu(\omega, \cdot)$$

to be that probability measure. Then, the second condition in the definition of the R. C. D. is satisfied.

To check the first one, define

$$\mathcal{A} := \{A \in \mathcal{B}(\mathbb{R}) : \mu(\cdot, A) = \mathbb{E}(\mathbb{I}_A|\mathcal{G}) \text{ almost surely}\}.$$

It follows that from the properties of conditional expectation (Proposition 4.1.3) that $\mathcal{A}$ is a λ-system. However, almost surely,

$$\mu(\cdot, (-\infty; x]) = F(x, \cdot) = \lim_{q\searrow x} G_q(\cdot) = \lim_{q\searrow x}\mathbb{E}(\mathbb{I}_{X\in(-\infty;q]}|\mathcal{G}) = \mathbb{E}(\lim_{q\searrow x}\mathbb{I}_{X\in(-\infty;q]}|\mathcal{G}) = \mathbb{E}(\mathbb{I}_{X\in(-\infty;x]}|\mathcal{G}).$$

Therefore, $\mathcal{A}$ contains the π-system $\{(\infty, x] : x \in \mathbb{R}\}$ that generates $\mathcal{B}(\mathbb{R})$, and so, by π-λ theorem, $\mathcal{A} \supset \mathcal{B}(\mathbb{R})$.

To see the uniqueness, note that by the uniqueness of conditional expectation, for every $q \in \mathbb{Q}$, we have $\mu(\cdot, (-\infty, q]) = \nu(\cdot(-\infty, q])$ almost surely. Since $\mathbb{Q}$ is countable, this implies that almost surely, $\mu(\cdot, (-\infty, q]) = \nu(\cdot, (-\infty, q])$ for all $q \in \mathbb{Q}$. Since $\{(-\infty; q] : q \in \mathbb{Q}\}$ is a π-system that generates $\mathcal{B}(\mathbb{R})$, this implies that for almost all $\omega \in \Omega$, $\mu(\omega, \cdot) \equiv \nu(\omega, \cdot)$ on $\mathcal{B}(\mathbb{R})$. □

REMARK 4.4.4. Informally, a way to think about the regular conditional distribution is that it is a random variable with values in the space of probability measures on $\mathcal{F}'$: of an experiment is performed and we are told the information contained in $\mathcal{G}$, we view X as distributed according to a certain law, which in itself is random since it depends on the information we are told. Note, however, that to make this point of view rigorous, one would need to show that this measure-valued function is measurable in a suitable sense, which may not be true in general.

REMARK 4.4.5. Two measurable spaces $(\Omega', \mathcal{F}')$ and $(\Omega'', \mathcal{F}'')$ are said to be *(Borel-)isomprphic* if there is a measurable bijection $f : \Omega' \to \Omega''$ with a measurable inverse. It is clear that R. C. D. exists also whenever $(\Omega', \mathcal{F}')$ is isomorphic to $(\mathbb{R}; \mathcal{B}(\mathbb{R}))$. It turns out that in fact, if M is any uncountable[4] compete separable metric space, then $(M; \mathcal{B}(M))$ is isomorphic to $(\mathbb{R}; \mathcal{B}(\mathbb{R}))$.

Finally, we connect the RCD with conditional expectations:

LEMMA 4.4.6. *If X is a scalar random variable and $h : \mathbb{R} \mapsto \mathbb{R}$ is any measurable function such that $\mathbb{E}|h(X)| < \infty$, then, almost surely,*

$$\mathbb{E}(h(X)|\mathcal{G}) = \mathbb{E}_{X|\mathcal{G}}(h(X)) = \int_{\mathbb{R}} h(x) d\mu(\cdot, x),$$

where μ is the R. C. D. of X given $\mathcal{G}$.

PROOF. Indeed, for $h = \mathbb{I}_A$, $A \in \mathcal{B}(\mathbb{R})$, the identity holds true by the definition of R. C. D. Therefore, by linearity, it holds for all simple functions; by monotone convergence theorem, it holds for any non-negative h with $\mathbb{E}h(X) < \infty$, and by linearity again, it holds for any measurable h such that $\mathbb{E}|h(X)| < \infty$. □

This lemma explains in a natural way why many properties of expectation (e. g., Holder, Chebyshev, Jensen inequalities) extend to conditional expectations. Let us do in detail the Jensen case:

PROPOSITION 4.4.7. *(Conditional Jensen inequality) If h is any convex function, and X is any scalar random variable such that $\mathbb{E}|X| < \infty$ and $\mathbb{E}|h(X)| < \infty$, then*

$$\mathbb{E}(h(X)|\mathcal{G}) \geq h(\mathbb{E}(X|\mathcal{G})) \text{ almost surely.}$$

PROOF. Since $\mathbb{E}|X| = \mathbb{E}(\mathbb{E}(|X||\mathcal{G})) < \infty$, we infer that $\mathbb{E}(|X||\mathcal{G})$ is finite almost surely. Similarly, $\mathbb{E}(|h(X)|\mathcal{G}) < \infty$ almost surely. Therefore, applying Jensen inequality to the probability measure $\mu(\omega, \cdot)$, we get

$$\mathbb{E}(h(X)|\mathcal{G}) = \mathbb{E}_{X|\mathcal{G}}(h(X)) \geq h(\mathbb{E}_{X|\mathcal{G}}(X)) = h\mathbb{E}(X|\mathcal{G})$$

almost surely, as required. □

4.5. Martingales: simple properties and the optional stopping theorem

Let $\xi_1, \xi_2, \ldots$ be indepdenent scalar random variables, and put $X_n = \xi_1 + \cdots + \xi_n$. We can consider this as a model for betting on several games in succession; ξ_i is a gain at game i, and X_n is a net gain after n games. If $\mathbb{E}\xi_i \geq 0$ (resp. $\mathbb{E}\xi_i \leq 0$) for all i, then each individual game is favourable (resp. unfavourable) for us; linearity of expectation then implies that the whole gamble will be favourable (resp., unfavourable).

In this setup, ξ_n is independent of $\xi_1, \ldots, \xi_{n-1}$, that is, the game we are playing at stage n is chosen in advance and does not depend on the outcomes of the previous games. We would like to generalize this setup to allow the player to have a *gambling system,* that is, to *choose* the game at stage n based on the outcomes of the previous games; still preserving the condition that all the games available are favourable (resp., unfavourable or even). This leads naturally to the definition of martingales and sub/supermartingales.

DEFINITION 4.5.1. A *filtered probability space* is a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ equipped with a *filtration*, that is, with an increasing sequence $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \cdots \subset \mathcal{F}$ of σ-algebras. A distrete time stochastic process $X_0, X_1, \ldots$ is called *adapted* to a filtration $\{\mathcal{F}_n\}$ if X_n is $\mathcal{F}_n$-measurable for all $n = 0, 1, \ldots$

Informally, $\mathcal{F}_n$ represents the information available at time n; the amount of information increases with n. If $X_n$is a stochastic process, then $\mathcal{F}_n = \sigma(X_0, \ldots, X_n)$ is called the *natural filtration* of X_n; it is the minimal filtration to which $X_n$is adapted. In this case, the only "sourse of information" is the observation the process X_n.

DEFINITION 4.5.2. Let X_n be a stochastic process adapted to a filtration $\{\mathcal{F}_n\}$, such that $\mathbb{E}|X_n| < \infty$ for all n. We say that

- X_n is a submartingale if $\mathbb{E}(X_{n+1}|\mathcal{F}_n) \geq X_n$ a. s. for all $n = 0, 1, \ldots$;

[4] in the case of a countable or finite M, the existence of RCD is fairly easy to show.

- X_n is a supermartingale if $\mathbb{E}(X_{n+1}|\mathcal{F}_n) \leq X_n$ a. s. for all $n = 0, 1, \dots$;
- X_n is a martingale if $\mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n$ a. s. for all $n = 0, 1, \dots$.

EXAMPLE 4.5.3. If $\xi_1, \xi_2, \dots$ are indepdenent scalar random variables such that $\mathbb{E}|\xi_i| < \infty$ for all i, put $X_n = \xi_1 + \dots + \xi_n$, $\mathcal{F}_n = \sigma(\xi_1, \dots, \xi_n) = \sigma(X_0, X_1, \dots, X_n)$. Then, using linearity and Proposition 4.2.1

$$\mathbb{E}(X_{n+1}|\mathcal{F}_n) = \mathbb{E}(X_n|\mathcal{F}_n) + \mathbb{E}(\xi_{n+1}|\mathcal{F}_n) = X_n + \mathbb{E}(\xi_{n+1}).$$

Therefore, if $\mathbb{E}\xi_i \geq 0$ for all i (respectively, $\mathbb{E}\xi_i \leq 0$, $\mathbb{E}\xi_i = 0$), then X_n is a submartingale (respectively, supermartingale, martingale)

REMARK 4.5.4. Since X_n is adapted to $\mathcal{F}_n$, we have $\mathbb{E}(X_n|\mathcal{F}_n) = X_n$. Therefore, the condition for being a submartingale (resp., supermartingale) is equivelent to

$$\mathbb{E}(X_{n+1} - X_n|\mathcal{F}_n) \geq 0 \quad (\text{ respectively, } \mathbb{E}(X_{n+1} - X_n|\mathcal{F}_n) \leq 0).$$

REMARK 4.5.5. Clearly, X_n is a submartingale iff $-X_n$ is a supermartingale. Therefore, many result stated below for submartingales have natural counterparts of supermartingales and martingales.

REMARK 4.5.6. If X_n is a (sub-)martingale and a is a constant, then $X_n + a$ is a (sub-)martingale. If $X_1, X_2, \dots$ is a (sub-)martingale, then $\mathbb{E}X_1, X_1, X_2$ is also a (sub-)martingale. Therefore, one typically does not lose generality by considering only (sub-)martingales with $X_0 = 0$ almost surely.

We will prove several versions of the statement that on average, a person betting on a submartingale does not lose. The simplest one is as follows:

LEMMA 4.5.7. *If X_n is a submartingale (respectively, martingale), then, for any $m > n$,*

$$\mathbb{E}(X_m|\mathcal{F}_n) \geq X_n \text{ almost surely,}$$

respectively,

$$\mathbb{E}(X_m|\mathcal{F}_n) = X_n \text{ almost surely.}$$

In particular, $\mathbb{E}X_m \geq \mathbb{E}X_0$ (resp., $\mathbb{E}X_m = \mathbb{E}X_0$) for all $m \geq 0$.

PROOF. Indeed, by tower property and monotonicity of conditional expectation (Propositions 4.2.1 and 4.1.3), we have

$$\mathbb{E}(X_m|\mathcal{F}_n) = \mathbb{E}(\mathbb{E}(X_m|\mathcal{F}_{m-1})|\mathcal{F}_n) \geq \mathbb{E}(X_{m-1}|\mathcal{F}_n) \geq \dots \geq \mathbb{E}(X_n|\mathcal{F}_n) = X_n,$$

and in the martingale case we have equalities instead of inequalities. For the "in particular" statement, put $n = 0$ and take the expectation. □

We will now extend this result slightly.

DEFINITION 4.5.8. A process H_n, $n = 1, 2, \dots$, is called predictable (w. r. t. a filtration $\mathcal{F}_n$) if for all $n = 1, 2, \dots$ H_n is $\mathcal{F}_{n-1}$ -measurable. Given a stochastic process X_n and a predictable process H_n, define

$$(H \bullet X)_n := \sum_{i=1}^{n} H_i(X_i - X_{i-1}).$$

Informally, assume that every day at noon, the price X_n of a stock is announced, after which we are allowed to buy (or sell) any amount of stock at this price. The quantity H_n then corresponds to the amount of stock we choose to possess in the interval from day $n-1$ to day n, and $H_n(X_n - X_{n-1})$ is our monetary gain during that interval. The next theorem shows that if the overall trend is upwards, on average we will gain money whatever we do.

PROPOSITION 4.5.9. *(Gambling systems) If H_n is any bounded predictable process and X_n is a matirtingale, then $(H \bullet X)_n$ is a martingale. If H_n is bounded and non-negative (respectively, non-positive) predictable process and X_n is a submartingale, then $(H \bullet X)_n$ is a submartingale (respectively, supermartingale).*

PROOF. We have

$$\mathbb{E}((H \bullet X)_{n+1}|\mathcal{F}_n) = \mathbb{E}((H \bullet X)_n|\mathcal{F}_n) + \mathbb{E}(H_{n+1}(X_{n+1} - X_n)|\mathcal{F}_n).$$

Since $(H \bullet X)_n$ is $\mathcal{F}_n$ -measurable, the first term is just $(H \bullet X)_n$. Since H_{n+1} is $\mathcal{F}_n$ -measurable and bounded, it can be taken out of the conditional expectation. Therefore,

$$\mathbb{E}((H \bullet X)_{n+1}|\mathcal{F}_n) = (H \bullet X)_n + H_{n+1}\mathbb{E}(X_{n+1} - X_n|\mathcal{F}_n),$$

from which all the cases readily follow. □

DEFINITION 4.5.10. A random variable $\tau \in \mathbb{Z}_{\geq 0} \cup \{\infty\}$ is called a *stopping time* if for each n, the event $\tau \leq n$ is $\mathcal{F}_n$-measurable.

Informally, the stopping time is the moment we may choose to stop gambling. The decision as to whether to do it at time n must be based on the information available at that time, hence the definition.

PROPOSITION 4.5.11. *If X_n is a (sub-)martingale and τ is a stopping time, then $X_{\tau \wedge n}$ is a (sub-)martingale.*

Here and below we use the notation $a \wedge b = \min(a, b)$.

PROOF. We define a process

$$H_n := \begin{cases} 1, & \tau \geq n \\ 0, & \text{otherwise.} \end{cases}$$

Since the event $\tau < n$ is $\mathcal{F}_{n-1}$-measurable, H_n is indeed predictable. We have

$$(H \bullet X)_n = \begin{cases} X_n - X_0, & \tau > n \\ X_\tau - X_0, & \tau \leq n \end{cases} = X_{\tau \wedge n} - X_0.$$

Since X_0 is $\mathcal{F}_n$ -measurable for all n, we have $\mathbb{E}(X_0|\mathcal{F}_n) = X_0$, therefore, the result follows from Proposition 4.5.9. □

COROLLARY 4.5.12. *If X_n is a submartingale and a stopping time τ is almost surely bounded, then $\mathbb{E}X_\tau \geq \mathbb{E}X_0$.*

PROOF. If $\tau \leq N$ for some N, then $\tau \wedge N = \tau$ almost surely, so, the result follows from Lemma 4.5.7 and Proposition 4.5.11. □

The most interesting stopping time, however, are unbounded. For example, let X_n be a simple random walk on $\mathbb{Z}$, that is, $X_n = \xi_1 + \cdots + \xi_n$, where ξ_i are independent and $\mathbb{P}(\xi_i = 1) = \mathbb{P}(\xi_i = -1) = \frac{1}{2}$. Then, for $b \in \mathbb{Z}_{>0}$ and $-a \in \mathbb{Z}_{<0}$, we can define

$$\tau := \min\{n : X_n = -a \text{ or } X_n = b\}.$$

If we were able to extend the above Corollary to τ (which we eventually will), its application would imply that

$$a\mathbb{P}(X_\tau = a) + b\mathbb{P}(X_\tau = b) = \mathbb{E}X_\tau = \mathbb{E}X_0 = 0,$$

or

$$\mathbb{P}(X_\tau = a) = \frac{b}{b + a},$$

a result obtaind in a different way in the exercises. Moreover, it is also easy to see (see Exercises) that $X_n^2 - n$ is also a martingale, which implies that

$$0 = \mathbb{E}X_0^2 - \mathbb{E}0 = \mathbb{E}X_\tau^2 - \mathbb{E}\tau = \frac{ba^2}{b + a} + \frac{ab^2}{b + a} = ab,$$

also a result obtained in the exercises.

REMARK 4.5.13. Corollary 4.5.12 cannot hold true for *any* stopping time (even almost surely finite). Let X_n be a simple random walk, and define $\tau = \min\{n : X_n = -1\}$. Then, as it follows from recurrence, $\tau < \infty$ almost surely. But

$$\mathbb{E}X_\tau = -1 \neq 0 = \mathbb{E}X_0.$$

THEOREM 4.5.14. *(Optional stopping theorem)Let X_n be a submartingale, and τ be an almost surely finite stopping time. Assume that*

- *either X_n is almost surely bounded (that is, there is a constant $C > 0$ such that $|X_n| \leq C$ for all n almost surely), or*
- *$\mathbb{E}\tau < \infty$ and there is a constant $C > 0$ such that $|X_{n+1} - X_n| < C$ almost surely.*

Then, $\mathbb{E}|X_\tau| < \infty$, and

$$\mathbb{E}X_\tau \geq \mathbb{E}X_0.$$

Applying the theorem to $-X_n$, we see that it also holds for supermartingales with the inequality reversed, and for martingales with equality instead for inequality.

PROOF. Note that since $\tau < \infty$ almost surely, we have $X_{\tau\wedge n} \to X_\tau$ almost surely. Proposition 4.5.11 and Lemma 4.5.7 imply that $\mathbb{E}X_{\tau\wedge n} \geq \mathbb{E}X_0$. Therefore, the theorem follows once we justify the exchange of the expectation and the limit:

$$\mathbb{E}X_\tau = \mathbb{E}\lim_{n\to\infty} X_{\tau\wedge n} \stackrel{?}{=} \lim_{n\to\infty} \mathbb{E}X_{\tau\wedge n} \geq \mathbb{E}X_0.$$

If X_n are bounded, then $|X_{\tau\wedge n}|$ are bounded, this is just the Dominated convergence theorem. If the second contition holds true, we note that

$$|X_{\tau\wedge n}| \leq |X_0| + \sum_{i=1}^{\tau\wedge n} |X_i - X_{i-1}| \leq |X_0| + C\tau,$$

which is integrable. So, the dominated convergence theorem applies. □

REMARK 4.5.15. Since the condition $|X_{n+1} - X_n| < C$ is satisfied if X_n is a random walk with bounded increments, we must have $\mathbb{E}\tau = \infty$ for $\tau = \min\{n : X_n = -1\}$ (See solution to Exercise set I for a different proof.) To see that the condition $|X_n - X_{n-1}| < C$ is needed, consider betting on a coin flip, starting from a unit stake, doubling the stakes as long as we lose, and stopping after the first win. Then, τ is geometrically distributed ($\mathbb{P}(\tau = k) = 2^{-k}$), hence $\mathbb{E}\tau < \infty$, but $\mathbb{E}(X_\tau) = 1 \neq 0 = \mathbb{E}X_0$.

4.6. Almost sure convergence of supermartingales.

We start with some definition. Given an adapted process X_n, and $a < b \in \mathbb{R}$, define inductively the stopping times $\tau_1, \tau_2, \ldots$, by

$$\tau_1 = \min\{n \geq 0 : X_n \leq a\} \quad \tau_2 = \min\{n > \tau_1 : X_n \geq b\}$$
$$\tau_3 = \min\{n \geq \tau_2 : X_n \leq a\} \quad \tau_4 = \min\{n > \tau_3 : X_n \geq b\}$$
$$\ldots\ldots$$

An interval $[\tau_{2k-1}, \tau_{2k})$ is called an upcrossing of the strip (a, b). We are interested in extimating the number of upcrossing completed up to time n: define

$$U_n^{a,b} = \max\{k : \tau_{2k} \leq n\}.$$

LEMMA 4.6.1. *(Upcrossing lemma) Let X_n be a supermartingale. Then, for all $a < b \in \mathbb{R}$,*

$$(b-a)\mathbb{E}U_n^{a,b} \leq \mathbb{E}\left((a - X_n)\mathbb{I}_{a\geq X_n}\right).$$

PROOF. The proof is based on the gambling system theorem. Viewing X_n as a stock price, consider the following strategy: wait until X_n gets below a, buy a unit amount of stock, wait until X_n gets above b, sell, repeat. Each completed upcrossing of (a, b) by $X_n$will give us a profit of at least $b - a$. Proposition 4.5.9 tells us that however natural this strategy looks, the game will still be unfavourable. The reason is the last incompleted upcrossing (at some point, one may enter a losing streak one does not survive.)

Formally, define a proceess H_n by

$$H_{n+1} := \begin{cases} 1, & \tau_{2k-1} \leq n < \tau_{2k} \text{ for some } k \\ 0, & \text{else.} \end{cases}$$

Observe that the event $\{H_{n+1} = 1\}$ is determined by $X_0, \ldots, X_n$, so, H_n is predictable. If we denote $U = U_n^{a,b}$, then

$$(H \bullet X)_n = \sum_{i=1}^{n} H_i(X_i - X_{i-1}) = \sum_{k=1}^{U} (X_{\tau_{2k}} - X_{\tau_{2k-1}}) + (X_n - X_{\tau_{2U+1}})\mathbb{I}_{\tau_{2U+1} \leq n} \geq (b-a)U - (a - X_n)\mathbb{I}_{a \geq X_n}.$$

The logic behind the last inequality is as follows: the last uncompleted upcrossing (if it has started by time n) starts with some value $X_{\tau_{2U+1}} < a$ and ends with X_n. So, if $X_n \leq a$, we lose at most $(a - X_n)$ during that upcrossing, and if $X_n > a$, we actually win (so we extimate the loss from above by 0).

Proposition 4.5.9 implies that

$$\mathbb{E}(H \bullet X)_n \leq \mathbb{E}(H \bullet X)_0 = 0,$$

so,

$$(b-a)\mathbb{E}U_n^{a,b} - \mathbb{E}((a - X_n)\mathbb{I}_{a \geq X_n}) \leq 0,$$

as required. □

THEOREM 4.6.2. *Let X_n be a supermartingale which is bounded in L^1, that is, $\sup_n \mathbb{E}|X_n| < \infty$. Then there exists a scalar random variable X with $\mathbb{E}|X| < \infty$, such that $X_n \to X$ almost surely.*

Before going into the proof, we prove Fatou's lemma:

LEMMA 4.6.3. *(Fatou) If $X_n \geq 0$, then*

$$\liminf \mathbb{E}X_n \geq \mathbb{E} \liminf X_n.$$

PROOF. Observe that $Y_n := \inf_{m \geq n}\{X_m\}$ is a non-decreasing sequence of non-negative random variables, therefore, by monotone convergence theorem,

$$\mathbb{E} \liminf X_n = \mathbb{E} \lim_{n\to\infty} Y_n = \lim_{n\to\infty} \mathbb{E}Y_n = \liminf \mathbb{E}Y_n$$

On the other hand, for any n, we have $X_n \geq Y_n$, hence $\mathbb{E}X_n \geq \mathbb{E}Y_n$, and therefore,

$$\liminf \mathbb{E}X_n \geq \liminf \mathbb{E}Y_n.$$

□

PROOF OF THEOREM 4.6.2. We define $X(\omega) := \liminf X_n(\omega)$, which is measurable as a $\liminf$ of countably many measurable functions. Now, note that for all $a \in \mathbb{R}$,

$$\mathbb{E}(a - X_n)\mathbb{I}_{a \geq X_n} \leq \mathbb{E}|a - X_n| \leq \mathbb{E}|a| + \mathbb{E}|X_n| \leq C < \infty,$$

where $C = |a| + \sup_n |X_n|$ does not depend on n. Therefore, for all $a, b \in \mathbb{R}$, if we define $U^{a,b} := \lim_{n\to\infty} U_n^{a,b}$, then, by monotone convergence theorem and the upcrossing lemma,

$$\mathbb{E}U^{a,b} = \lim_{n\to\infty} \mathbb{E}U_n^{a,b} \leq C < \infty.$$

In particular, for any fixed $a < b \in \mathbb{R}$, we have $U^{a,b} < \infty$ almost surely. Since the number of pairs $(a; b) : a, b \in \mathbb{Q}$ is countable, we deduce that almost surely, $U^{a,b} < \infty$ for any $a < b \in \mathbb{Q}$.

On the other hand, $X_n(\omega) \nrightarrow X(\omega)$ if and only if $\liminf X_n(\omega) < \limsup X_n(\omega)$. This happends if and only if there exist $a, b \in \mathbb{Q}$ such that $\liminf X_n < a < b < \limsup X_n$, which happends if and only if there exist $a < b \in \mathbb{Q}$ such that $U^{a,b} = \infty$. Thus, $\mathbb{P}(X_n(\omega) \nrightarrow X(\omega)) = 0$.

Now, Fatou's lemma implies that

$$\mathbb{E}|X| = \mathbb{E}\liminf |X_n| \leq \liminf \mathbb{E}|X_n| \leq \sup_n \mathbb{E}|X_n| < \infty$$

in particular, X is almost surely finite. □

COROLLARY 4.6.4. *Let $X_n \geq 0$ be a supermartingale. Then X_n converges almost surely.*

PROOF. In that case, $\mathbb{E}|X_n| = \mathbb{E}X_n \leq \mathbb{E}X_0$, so, the conditions of the above theorem are satisfied. □

4.7. Doob's inequality and convergence in L^p for $p > 1$.

Theorem 4.6.2, although very general, leaves some questions unanswered. Namely, we would often like to have a generalization of the optional stopping theorem: say, for a martingale X_n, we would like to conclude that

$$\mathbb{E}(\lim X_n) = \mathbb{E}X_0$$

(Optional stopping theorem is a particular case of this setup, $X_{\tau \wedge n}$ being a (sub-)martingale converging almost surely to X_τ).

One particularly nice class of martingales for which the assertion is true is martingales and non-negative submartingales bounded in L^p.

THEOREM 4.7.1. *Assume that $p > 1$, and let X_n be a martingale (or a non-negative submartingale) such that $\sup \mathbb{E}|X_n|^p < \infty$. Then there exists a random variable X with $\mathbb{E}|X|^p < \infty$ such that $X_n \to X$ almost surely and in L^p.*

REMARK 4.7.2. Clearly, $\mathbb{E}|X|^p < \infty$ implies that X is almost surely finite, and$\mathbb{E}|X| < \infty$. Also, convergence in L^p for $p > 1$ implies convergence in L^1 (see Proposition 2.7.5), therefore, $|\mathbb{E}X - \mathbb{E}X_n| \leq \mathbb{E}|X - X_n| \to 0$. Thus, if X_n is a martingale bounded in L^p, then $\mathbb{E}X = \mathbb{E}X_n = \mathbb{E}X_0$ (respectively, $\mathbb{E}X \geq \mathbb{E}X_0$ if X_n is a non-negative submartingale). So, for martingales bounded in L^p, $p > 1$, the above questions are answered in the positive.

We start with a lemma revealing the relation between martingales, submartingales, and convex functions.

PROPOSITION 4.7.3. *Let h be a convex function, and assume that $\mathbb{E}|h(X_n)| < \infty$ for all n. If*

- *either X_n is a martingale,*
- *or X_n is a submartingale and h is non-decreasing,*

then $h(X_n)$ is a submartingale.

PROOF. The proposition follows easily from conditional Jensen's inequality (Proposition 4.4.7):

$$\mathbb{E}(h(X_n)|\mathcal{F}_{n-1}) \geq h(\mathbb{E}(X_n|\mathcal{F}_{n-1})). \tag{4.7.1}$$

If X_n is a martingale, then $\mathbb{E}(X_n|\mathcal{F}_{n-1}) = X_{n-1}$, so the right-hand side is equal to $h(X_{n-1})$. If X_n is a submartingale, then $\mathbb{E}(X_n|\mathcal{F}_{n-1}) \geq X_{n-1}$, so, when h is non-decreasing, the right-hand side of (4.7.1) is $\geq h(X_{n-1})$. □

REMARK 4.7.4. If $h \geq 0$ is convex and $\mathbb{E}|h(X_N)| < \infty$ for *some* N, then the above proposition still holds true, with the conclusion that $h(X_0), \ldots, h(X_N)$ is a submartingale. Indeed, the same proof as above shows that $h(X_{N-1}) \leq \mathbb{E}(h(X_N)|\mathcal{F}_{N-1})$, and, taking expectations, we infer that $\mathbb{E}h(X_{N-1}) \leq \mathbb{E}(h(X_N))$. Repeating, we show by induction that $\mathbb{E}(h(X_n)) \leq \mathbb{E}h(x_N) < \infty$ for all $n \leq N$.

EXAMPLE 4.7.5. Let $p \geq 1$, and let X_n be a martingale or a non-negative submartingale such that $\mathbb{E}|X_n|^p < \infty$ for all n. Then, $|X_n|^p$ is a submartingale. If X_n is a submartingale, then, for any $a \in \mathbb{R}$, $\max\{a; X_n\}$ is a submartingale.

PROOF. Apply Proposition 4.7.3 to the convex function $h(x) = |x|^p$, non-decreasing convex function $h(x) = x^p \mathbb{I}_{x\geq 0}$, and non-decreasing convex function $h(x) = \max\{a; x\}$. □

PROPOSITION 4.7.6. *(Doob's inequality) Let X_n be a non-negative submartingale, and denote $\bar{X}_n := \max\{X_0, X_1, \dots X_n\}$. Then, for any $a > 0$,*

$$\mathbb{P}(\bar{X}_n > a) \leq \frac{\mathbb{E}(X_n \mathbb{I}_{\bar{X}_n > a})}{a} \leq \frac{\mathbb{E}(X_n)}{a}.$$

REMARK 4.7.7. Observe that the inequality looks similar to Chebyshev's inequality, but is stronger (because $X_n \geq a$ implies $\bar{X}_n \geq a$).

PROOF. Denote $\tau = \min\{n : X_n \geq a\}$. Then $\bar{X}_n > a$ if and only if $\tau \leq n$. We have

$$\mathbb{E}(X_n \mathbb{I}_{\tau\leq n}) = \sum_{k=0}^{n} \mathbb{E}(X_n \mathbb{I}_{\tau=k}) \overset{(1)}{=} \sum_{k=0}^{n} \mathbb{E}(\mathbb{E}(X_n \mathbb{I}_{\tau=k}|\mathcal{F}_k))$$
$$\overset{(2)}{=} \sum_{k=0}^{n} \mathbb{E}(\mathbb{I}_{\tau=k}\mathbb{E}(X_n|\mathcal{F}_k)) \overset{(3)}{\geq} \sum_{k=0}^{n} \mathbb{E}(\mathbb{I}_{\tau=k} X_k) \overset{(4)}{\geq} a \sum_{k=0}^{n} \mathbb{E}(\mathbb{I}_{\tau=k}) = a\mathbb{P}(\tau \leq n),$$

as required. Above, (1) is $\mathbb{E}\mathbb{E}(Y|\mathcal{G}) = \mathbb{E}Y$, (2) is "taking out what is known" (see Proposition 4.2.1, we use that τ is a stopping time and hence $\{\tau = k\}$ is $\mathcal{F}_k$-measurable), (3) uses that X_n is a submartingale, and (4) uses that if $\tau = k$, then $X_k \geq a$. □

COROLLARY 4.7.8. *(Kolmogorov's inequality) If $\xi_1, \dots, \xi_n$ are independent centered random variables with $\mathbb{E}\xi_i^2 < \infty$, and $S_n = \xi_1 + \cdots + \xi_n$, then*

$$\mathbb{P}(\max\{|S_1|, \dots |S_n|\} > a) \leq \frac{Var\, S_n}{a^2}.$$

PROOF. Since S_n is a martingale and x^2 is convex, ${S_n}^2$ is a submartingale by Proposition 4.7.3. Applying Doob's inequality to that submartingale give the result. □

THEOREM 4.7.9. *(Maximal inequality for L^p) Let X_n be a non-negative submartingale, and let $p > 1$. Then*

$$\mathbb{E}(\bar{X}_n^p) \leq \left(\frac{p}{p-1}\right)^p \mathbb{E}X_n^p.$$

PROOF. We assume $\mathbb{E}X_n^p < \infty$, for otherwise there is nothing to prove. We have

$$\mathbb{E}(\bar{X}_n^p) = \mathbb{E}\left(\int_{\mathbb{R}} \mathbb{I}_{0\leq a\leq \bar{X}_n^p} da\right) \overset{\text{Tonnelli}}{=} \int_0^\infty \left(\mathbb{E}\mathbb{I}_{0\leq a\leq \bar{X}_n^p}\right) da = \int_0^\infty \mathbb{P}(\bar{X}_n^p > a) da \overset{a=\lambda^p}{=} \int_0^\infty p\lambda^{p-1}\mathbb{P}(\bar{X}_n > \lambda) d\lambda$$
$$\overset{\text{Doob}}{\leq} \int_0^\infty p\lambda^{p-2}\mathbb{E}(X_n \mathbb{I}_{\bar{X}_n>\lambda}) d\lambda \overset{\text{Tonnelli}}{=} \mathbb{E}(X_n \int_0^\infty p\lambda^{p-2}\mathbb{I}_{\lambda<\bar{X}_n} d\lambda) = \frac{p}{p-1}\mathbb{E}\left(X_n \bar{X}_n^{p-1}\right).$$

Applying Hölder's inequality with parameters p and $q = \frac{p}{p-1}$, we get from this that

$$\mathbb{E}(\bar{X}^p) \leq \frac{p}{p-1}\left(\mathbb{E}X_n^p\right)^{\frac{1}{p}} \left(\mathbb{E}\left(\bar{X}_n^p\right)\right)^{1-\frac{1}{p}}.$$

Multiplying both parts by $\left(\mathbb{E}\left(\bar{X}_n^p\right)\right)^{\frac{1}{p}-1}$ and taking to the power p gives the desired result, *once we know that* $\mathbb{E}\left(\bar{X}_n^p\right) < \infty$. But since $x^p\mathbb{I}_{x>0}$ is convex and increasing, Proposition 4.7.3 and Remark 4.7.4 imply that $X_0^p, X_1^p, \dots, X_n^p$ is a non-negative submartingale, in particular,

$$\mathbb{E}\bar{X}_n^p \leq \mathbb{E}(X_1^p + \cdots + X_n^p) \leq n\mathbb{E}X_n^p < \infty.$$

□

We are now in a position to prove Theorem 4.7.1.

PROOF OF THEOREM 4.7.1. Since by Jensen's inequality, $|\mathbb{E}|X_n||^p \le \mathbb{E}|X_n|^p$, we have that $\sup \mathbb{E}|X_n| < \infty$, and therefore, by Theorem 4.6.2, $X_n \to X$ almost surely. Now, note that for each ω, $|\bar{X}|_n^p(\omega)$ is increasing. Since $|X_n|$ is a submartingale, it follows from the maximal inequality and the monotone convergence theorem that

$$\mathbb{E}\sup_{n\ge 0}|X_n|^p = \mathbb{E}\lim_{n\to\infty}|\bar{X}|_n^p = \lim_{n\to\infty}\mathbb{E}|\bar{X}|_n^p \le \left(\frac{p}{p-1}\right)^p \sup_{n\ge 0}\mathbb{E}|X_n|^p < \infty.$$

Therefore, $\mathbb{E}|X|^p = \mathbb{E}\limsup|X|_n^p \le \mathbb{E}\sup|X|_n^p < \infty$. Moreover, almost surely, $|X - X_n^p| \le ||X| + |X_n||^p \le 2\sup|X_n|^p$, which has finite expectation. Therefore, by dominated convergence theorem,

$$\mathbb{E}|X - X_n^p| \to 0.$$

□

4.8. Unifrom integrability and convergence in L^1.

We start by a remark that theorem 4.7.1 is false for $p = 1$. The example is provided by the "doubling the stakes" betting strategy as described in Remark 4.5.15. Indeed, in that case,

$$X_{\tau\wedge n} = \begin{cases} 1 - 2^n, & \tau > n, \\ 1, & \tau \le n, \end{cases}$$

and the probability that $\tau > n$ is 2^{-n}. Therefore, $\mathbb{E}|X_{\tau\wedge n}| = 1 - 2^{-n} + 1\cdot(1 - 2^{-n}) = 2 - 2^{-n-1}$, which is bounded. However, we have seen that $X_{\tau\wedge n} \nrightarrow X_\tau$ in L^1.

The notion that captures the difference between almost sure convergence and convergence in L^1 is the *uniform integrablity*. Note that a single random variable X is integrable if and only if

$$\mathbb{E}|X|\mathbb{I}_{|X|>M} \to 0$$

as $M \to \infty$: the "if" part follows from the identity

$$\mathbb{E}|X| = \mathbb{E}|X|\mathbb{I}_{|X|>M} + \mathbb{E}|X|\mathbb{I}_{|X|\le M} \le \mathbb{E}|X|\mathbb{I}_{|X|>M} + M,$$

and the "only if " part follows from dominated convergence theorem, for

DEFINITION 4.8.1. A family $\{X_\alpha\}_{\alpha\in\mathcal{A}}$ of random variables (indexed by arbitrary set $\mathcal{A}$) is called *uniformly integrable (UI)* if $\mathbb{E}\left(|X_\alpha|\mathbb{I}_{|X_\alpha|>M}\right) \to 0$ as $M \to \infty$ uniformly over $\alpha \in \mathcal{A}$. That is,

$$\sup_{\alpha\in\mathcal{A}}\mathbb{E}\left(|X_\alpha|\mathbb{I}_{|X_\alpha|>M}\right) \overset{M\to\infty}{\longrightarrow} 0.$$

THEOREM 4.8.2. *Assume that $X_n \to X$ in probability. Then, the following are equivalent:*

(1) *$X_n \to X$ in L^1;*
(2) *X_n are uniformly integrable;*
(3) *$\mathbb{E}|X_n| \to \mathbb{E}|X|$ and $\mathbb{E}|X| < \infty$.*

PROOF OF THEOREM . (2) $\Longrightarrow$ (1) Assume that $\mathbb{E}|X_n - X| \nrightarrow 0$. Then, by passing to a subsequence, we may assume that $\mathbb{E}|X_n - X| > \varepsilon > 0$ for all n. Using Proposition 2.7.7, we may, by passing to a further subsequence, assume that $X_n \to X$ almost surely. We first show that X is integrable. Indeed, there is an M such that $\mathbb{E}|X_n|\mathbb{I}_{|X_n|>M} < C$ for all n. By Fatou's lemma (Lemma 4.6.3),

$$\mathbb{E}|X|\mathbb{I}_{|X|>M} = \mathbb{E}\liminf|X_n|\mathbb{I}_{|X_n|>M} \le \liminf\mathbb{E}|X_n|\mathbb{I}_{|X_n|>M} \le C < \infty,$$

so that X is integrable.

Choose M so that $\mathbb{E}|X_n|\mathbb{I}_{|X_n|\ge M} < \frac{\varepsilon}{4}$ for all n and $\mathbb{E}|X|\mathbb{I}_{|X|\ge M} < \frac{\varepsilon}{4}$. Then,

$$\mathbb{E}|X_n - X| \le \mathbb{E}\left|X_n\mathbb{I}_{|X_n|<M} - X\mathbb{I}_{|X|<M}\right| + \mathbb{E}|X_n|\mathbb{I}_{|X_n|\ge M} + \mathbb{E}|X|\mathbb{I}_{|X|\ge M} \le \mathbb{E}\left|X_n\mathbb{I}_{|X_n|<M} - X\mathbb{I}_{|X|<M}\right| + \frac{\varepsilon}{2} \to \frac{\varepsilon}{2},$$

by dominated convergence theorem, which is a contradiction.

(1) $\Longrightarrow$ (3) We have $|\mathbb{E}|X_n| - \mathbb{E}|X|| = |\mathbb{E}(|X_n| - |X|)| \le \mathbb{E}|X_n - X|$.

(3) $\Longrightarrow$ (2). Put $Y_{n,M} := |\mathrm{X_n}|\mathbb{I}_{|X_n|\ge M}$. Assume that X_n are not uniformly integrable, that is, $\sup_{n\ge 0}\mathbb{E}Y_{n,M} > \varepsilon > 0$ for all M. In that case, for every $k = 1, 2, \ldots$, there is an index n_k such that

$\mathbb{E}(Y_{n_k,k}) > \varepsilon$. Since $Y_{n_{k+m},k} \geq Y_{n_{k+m},k+m} > \varepsilon$, the property $\mathbb{E}(Y_{n_k,k}) > \varepsilon$ for all k is preserved under passing to subsequences, if $k(i)$ is a further subsequence, then $k_i \geq i$ and hence $\mathbb{E}(Y_{n_{k(i)},i}) > \varepsilon$. By Proposition 2.7.7, we can pick such a subsequence so that the convergence is almost sure. We will drop a subscript k and assume that $X_n \to X$ almost surely and $\mathbb{E}Y_{n,n} > \varepsilon$ for all n.

Now, fix an integer M such that $\mathbb{E}|X|\mathbb{I}_{|X|\geq M} < \frac{\varepsilon}{2}$. Then, we have

$$\mathbb{E}|X_n|\mathbb{I}_{|X_n|\geq M} = \mathbb{E}(|X_n| - |X|) - \mathbb{E}(|X_n|\mathbb{I}_{|X_n|<M} - |X|\mathbb{I}_{|X|\leq M}) - \mathbb{E}|X|\mathbb{I}_{|X|\geq M}$$

By dominated convergence theorem, the second term converges to zero as $n \to \infty$. Therefore, the right-hand side is less than ε for n large enough. But if $n > M$, then $\mathbb{E}|X_n|\mathbb{I}_{|X_n|\geq M} \geq \mathbb{E}|X_n|\mathbb{I}_{|X_n|\geq n} > \varepsilon$, which is a desired contradiction. □

REMARK 4.8.3. Informally, if a sequence X_n converges to X in probability, and $\mathbb{E}|X| < \infty$, then the only obstruction to L^1 convergence is "mass escaping to infinity". The proof of the implication (3) $\Longrightarrow$ (2) shows that if this happends, then the limit $|X|$ of $|X_n|$ will necessarily have a "deficiency of mass".

REMARK 4.8.4. Note that since $\mathbb{E}|X_n|\mathbb{I}_{|X_n|>M} \geq M\mathbb{P}(|X_n| \geq M)$, the uniform integrability condition is stronger than the tightness condition.

We now explore some examples of UI families.

EXAMPLE 4.8.5. Assume that for all α, $|X_\alpha| \leq Y$, where Y is an integrable random variable. Then, $\{X_\alpha\}$ is UI.

PROOF. Indeed, $\sup \mathbb{E}|X_\alpha|\mathbb{I}_{|X_\alpha|>M} \leq \mathbb{E}Y\mathbb{I}_{Y>M} \to 0$. □

Thus, Theorem 4.8.2 in fact generalized dominated convergence theorem.

EXAMPLE 4.8.6. Assume that there exist a function $\varphi : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ with $\varphi(x)/x \to \infty$ as $x \to \infty$ and a constant $C > 0$ such that $\mathbb{E}\varphi(|X_\alpha|) \leq C$ for all α. Then, $\{X_\alpha\}$ is UI. In particualr, if X_α are uniformly bounded in L^p for $p > 1$, then $\{X_\alpha\}$ is UI.

PROOF. For every α, we have

$$\mathbb{E}|X_\alpha|\mathbb{I}_{|X_\alpha|>M} \leq \mathbb{E}\left(\varphi(|X_\alpha|)\frac{|X_\alpha|}{\varphi(|X_\alpha|)}\mathbb{I}_{|X_\alpha|>M}\right) \leq \sup_{x\geq M}(x/\varphi(x))\mathbb{E}\varphi(|X_\alpha|) \leq C \sup_{x\geq M}(x/\varphi(x)) \to 0.$$

□

The next example is especially important in the context of martingales.

PROPOSITION 4.8.7. *Let X be an integrable random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and define*

$$X_\mathcal{G} := \mathbb{E}(X|\mathcal{G}),$$

a family indexed by the set of all sigma-algebras $\mathcal{G} \subset \mathcal{F}$. Then, $\{X_\mathcal{G}\}$ is UI.

PROOF. We can write, using conditional Jensen's inequality,

$$\mathbb{E}(|\mathbb{E}(X|\mathcal{G})|\mathbb{I}_{|X_\mathcal{G}|\geq M}) \leq \mathbb{E}\mathbb{E}(|X||\mathcal{G})|\mathbb{I}_{|X_\mathcal{G}|\geq M}).$$

Since $X_\mathcal{G}$ is $\mathcal{G}$-measurable, by definition of conditional expectation the right-hand side is equal to $\mathbb{E}(|X|\mathbb{I}_{A(\mathcal{G},M)})$, where $A(\mathcal{G}, M) = \{|X_\mathcal{G}| > M\}$. We can easily estimate the probability of $A(\mathcal{G}, M)$ from above by Chebyshev's inequality:

$$\mathbb{P}(A(\mathcal{G}, M)) \leq \frac{\mathbb{E}(|\mathbb{E}(X|\mathcal{G})|)}{M} \overset{\text{Jensen}}{\leq} \frac{\mathbb{E}\mathbb{E}(|X||\mathcal{G})}{M} = \frac{\mathbb{E}|X|}{M}.$$

We conclude that

$$\mathbb{E}|X_\mathcal{G}|\mathbb{I}_{|X_\mathcal{G}|\geq M} \leq \sup_{A:\mathbb{P}(A)<\delta_M} \mathbb{E}(|X|\mathbb{I}_A),$$

where $\delta_M = \frac{\mathbb{E}|X|}{M}$. The right-hand side does not depend on $\mathcal{G}$, so we only need to check that it goes to zero as δ_M goes to zero. We do this as a separate lemma 4.8.8 below. □

For future references, the lemma will be stated in a more general form. Let μ, ν be two measures on the same measurable space $(\Omega, \mathcal{F})$. Recall that μ is called absolutely continuous with respect to ν (written $\mu \preceq \nu$) if $\nu(A) = 0$ implies $\mu(A) = 0$. Clearly, if $\mathbb{P}$ is a probability measure and X is an integrable random variable, then $\mu(A) = \mathbb{E}(\mathbb{I}_A|X|)$ defines a measure such that $\mu \preceq \nu$.

LEMMA 4.8.8. *("Absolute continuity") Let ν be a probability measure, and $\mu \preceq \nu$ for every $\varepsilon > 0$, there exists $\delta > 0$ such that $\nu(A) < \delta$ implies $\mu(A) < \varepsilon$.*

PROOF. Assume that this is not the case. Then, there exists $\varepsilon > 0$ and a sequence $A_1, A_2, \dots$ of sets with $\nu(A_n) \to 0$, but $\mu(A_n) \geq \varepsilon$ for all n. By passing to a subsequence, we may assume that $\nu(A_n) < 2^{-n}$. Define $B_n := \cup_{i=n}^{\infty} A_i$. Then $B_1 \supset B_2 \supset \dots$, $\mu(B_n) \geq \mu(A_n) > \varepsilon$ and $\nu(B_n) \leq \sum_{i=n}^{\infty} \mu(A_n) \leq 2^{-n+1}$, therefore, by lower continuity of measures, $\mu(\cap_{n=1}^{\infty} B_n) \geq \varepsilon > 0$ and $\nu(\cap_{n=1}^{\infty} B_n) = 0$, which is a impossible since $\mu \preceq \nu$. □

DEFINITION 4.8.9. (Lévy martingale a. k. a. Doob martingale) Let Y be a random variable wnth $\mathbb{E}|Y| < \infty$, and let $\{\mathcal{F}_n\}$ be a filtration. Then,

$$Y_n = \mathbb{E}(Y|\mathcal{F}_n)$$

is called a Lévy martingale (associated to Y).

That this is a martingale is an easy consequence of the tower property. Proposition 4.8.7 shows that this martingale is in fact UI. A remarkable fact is that in fact, all UI martingales have this form:

THEOREM 4.8.10. *Let X_n be a uniformly integrable martingale, and let $X = \lim_{n\to\infty} X_n$ almost surely*[5]*. Then, $X_n = \mathbb{E}(X|\mathcal{F}_n)$. Conversely, if $X_n = \mathbb{E}(X|\mathcal{F}_n)$ and X is measurable with respect to $\mathcal{F}_\infty := \sigma(\cup_{n=0}^{\infty} \mathcal{F}_n)$, then $X = \lim_{n\to\infty} X_n$ almost surely and in L^1.*

REMARK 4.8.11. In general, if X_n is a Lévy martingale associated to X, then, by tower property,

$$\mathbb{E}((\mathbb{E}(X|\mathcal{F}_\infty)|\mathcal{F}_n) = \mathbb{E}(X|\mathcal{F}_n) = X_n,$$

so that X_n is also the Lévy martingale associated to $\mathbb{E}(X|\mathcal{F}_\infty)$. The latter is $\mathcal{F}_\infty$ measurable, therefore,

$$\lim_{n\to\infty} X_n = \mathbb{E}(X|\mathcal{F}_\infty).$$

PROOF. Let us check by the definition of conditional expectation that $X_n = \mathbb{E}(X|\mathcal{F}_n)$. Clearly, $X_n \in \mathcal{F}_n$. Pick $A \in \mathcal{F}_n$, and note that $\mathbb{I}_A X_n, \mathbb{I}_A X_{n+1}, \dots$ is a martingale: A is measurable with respect to $\mathcal{F}_r$, $r \geq n$, and thus can be pulled out of all the conditional expectations. Therefore,

$$\mathbb{E}(\mathbb{I}_A X_n) = \mathbb{E}(\mathbb{I}_A X_r), \quad r > n.$$

However, since X_n are UI, $X_n \to X$ in L^1, therefore,

$$\mathbb{E}(\mathbb{I}_A X) = \mathbb{E}(\mathbb{I}_A(X - X_r)) + \mathbb{E}(\mathbb{I}_A X_n) \stackrel{r\to\infty}{\longrightarrow} \mathbb{E}(\mathbb{I}_A X_n).$$

This completes the proof that $X_n = \mathbb{E}(X|\mathcal{F}_n)$.

For the "conversely" part, we know that X_n is a UI martingale. Define $X' = \lim_{n\to\infty} X_n$. Since $X = \mathbb{E}(X|\mathcal{F}_\infty)$, it suffices to check that $X' = \mathbb{E}(X|\mathcal{F}_\infty)$ almost surely. By construction, X' is $\mathcal{F}_\infty$-measurable. If $A \in \mathcal{F}_n$ for some n, then the first part shows that

$$\mathbb{E}(\mathbb{I}_A X') = \mathbb{E}(\mathbb{I}_A X_n) = \mathbb{E}(\mathbb{I}_A \mathbb{E}(X|\mathcal{F}_n)) = \mathbb{E}(\mathbb{I}_A X).$$

The collection $\{A : \mathbb{E}(\mathbb{I}_A X) = \mathbb{E}(\mathbb{I}_A X')\}$ is a λ-system, and we just have showed that it contains the π-system $\cup_n \mathcal{F}_n$ that generates $\mathcal{F}_\infty$. Hence, it contains $\mathcal{F}_\infty$. □

[5] and hence in L^1 since X_n is UI

4.9. Backward martingales and the strong law of large numbers

DEFINITION 4.9.1. A *backward martingale* is a martingale indexed by negative integers. That is, given a sequence of σ-algebras $\cdots \subset \mathcal{G}_{-2} \subset \mathcal{G}_{-1} \subset \mathcal{G}_0$, a backward martingale is a sequence $\dots, X_{-2}, X_{-1}, X_0$ of integrable random variables such that X_n is $\mathcal{F}_n$-measurable and $\mathbb{E}(X_{n+1}|\mathcal{F}_n) = X_n$ for all $n = -1, -2, \dots$

The theory of backward martingales is much simpler than that of martingales. While martingales can "blow up" as $n \to \infty$, backward martingales are always controlled by X_0. For example, Proposition 4.7.3 implies that for any $p \geq 1$, $|X_{-n}|^p, \dots, |X_0|^p$ is a submartingale, thus $\mathbb{E}|X_{-n}|^p \leq \mathbb{E}|X_0|^p$ for all $n \in \mathbb{N}$, that is, backward martingales are always bounded in L^1, and if $\mathbb{E}|X_0|^p < \infty$, they are bounded in L^p. Moreover, applying the definition several times, we see that $\mathbb{E}(X_0|\mathcal{F}_{-n}) = X_{-n}$ for all $n \in \mathbb{N}$, that is, $\{X_{-n}\}$ are always uniformly integrable. Also, if $U_{-n}^{a,b}$ denotes the number of upcrossings of (a, b) in the interval $[-n, 0]$, the upcrossing lemma gives the bound

$$(b-a)\mathbb{E}U_{-n}^{a,b} \leq \mathbb{E}\max(a - X_0, 0) \leq |a| + \mathbb{E}|X_0|.$$

This means that almost surely, the expected total number of upcrossings of any (a, b) with $(a, b) \in \mathbb{Q}$ is finite, so that X_{-n} converges almost surely as $n \to \infty$. Since, as we have mentioned, X_{-n} are UI, the convergence also holds true in L^1. Finally, if $\mathbb{E}|X_0|^p < \infty$, then the maximal inequality (theorem 4.7.9)

$$\mathbb{E}\sup_n |X_{-n}|^p = \mathbb{E}\left(\lim_{n\to\infty} \sup\{|X_{-n}|^p, \dots, |X_0|^p\}\right) = \lim_{n\to\infty} \mathbb{E}\left(\sup\{|X_{-n}|^p, \dots, |X_0|^p\}\right) \leq C \cdot \mathbb{E}|X_0|^p,$$

so that $\mathbb{E}|X|^p < \infty$, where $X = \lim_{n\to\infty} X_{-n}$, and $|X - X_n|^p$ is dominated by an integrable function. So, the convergence actually holds in L^p. We collect this discussion in the following theorem:

THEOREM 4.9.2. *If X_{-n} is an inverse martingale, then there is a random variable X such that $X_n \to X$ almost surely and in L^1. If, in addition, $\mathbb{E}|X_0|^p < \infty$ is bounded, then the convergence holds in L^p.*

A striking application of this result is the strong law of large numbers (without moment assumptions):

THEOREM 4.9.3. *Let $X_1, X_2, \dots$ be i. i. d. scalar random variables with $\mathbb{E}X_i = 0$. Denote $S_n := X_1 + \cdots + X_n$. Then, $n^{-1}S_n \to 0$ almost surely and in L^1.*

PROOF. Define $\mathcal{F}_{-n} := \sigma(S_n, X_{n+1}, X_{n+2}, \dots)$. Since $S_{n+1} = S_n + X_{n+1}$, we have that S_{n+1} is measurable with respect to $\sigma(\mathcal{F}_{-n})$, i. e $\mathcal{F}_{-n-1} \subset \mathcal{F}_{-n}$ for all n. The key observation is that $\frac{1}{n}S_n$ is a backward martingale with respect to this (inverse) filtration. Indeed,

$$\mathbb{E}(X_1|S_n) + \cdots + \mathbb{E}(X_n|S_n) = \mathbb{E}(X_1 + \cdots + X_n|S_n) = \mathbb{E}(S_n|S_n) = S_n.$$

By symmetry, the expectations of $\mathbb{E}(X_1|S_n), \dots, \mathbb{E}(X_n|S_n)$ must be all equal. (Formally,

$$\mathbb{E}(X_2|X_1 + X_2 \cdots + X_n) = \mathbb{E}(X_1|X_2 + X_1 + \cdots + X_n) = \mathbb{E}(X_1|X_1 + X_2 + \cdots + X_n),$$

where the first identity is just reindexing). Therefore, we conclude that

$$\mathbb{E}(X_i|S_n) = n^{-1}S_n, \quad i = 1, \dots, n$$

In fact, this result does not change if we add information about $X_{n+1}, \dots$:

$$\mathbb{E}(X_i|\mathcal{F}_{-n}) = n^{-1}S_n, \quad i = 1, \dots, n.$$

To check it rigirously, observe that $n^{-1}S_n$ is $\mathcal{F}_{-n}$ measurable, and for any $A \in \sigma(X_{n+1}, \dots)$, we do have, by independence,

$$\mathbb{E}(X_i\mathbb{I}_A) = \mathbb{I}_A\mathbb{E}(X_i) = \mathbb{I}_A\mathbb{E}(n^{-1}S_n) = \mathbb{E}(n^{-1}S_n\mathbb{I}_A).$$

Therefore, this identity holds for every A in the π-system $\sigma(S_n) \cup \sigma(X_{n+1}, \dots)$, therefore, by π-λ theorem, it holds true for all $A \in \mathcal{F}_{-n}$.

From this, we compute

$$\mathbb{E}((n-1)^{-1}S_{n-1}|\mathcal{F}_{-n}) = \frac{n-1}{(n-1)n}S_n = \frac{1}{n}S_n,$$

that is, $\frac{1}{n}S_n$ is a backward martingale. Therefore, there is a random variable S such that $\frac{1}{n}S_n \to S$ almost surely and in L^1 (and therefore in probability). But we know from weak law of large numbers that $\frac{1}{n}S_n \to 0$ in probability, therefore, $S = 0$. □

4.10. Martingale proof of Radon-Nikodym theorem

Let μ, ν be two measures on the same measurable space $(\Omega, \mathcal{F})$. Recall that μ is called absolutely continuous with respect to ν (written $\mu \preceq \nu$) if $\nu(A) = 0$ implies $\mu(A) = 0$. Clearly, if $d\mu = f d\nu$ (that is, by definition, $\mu(A) = \int \mathbb{I}_A f d\nu$ for any $A \in \mathcal{F}$) with $f \geq 0$, $\int f d\nu < \infty$, then $\mu \preceq \nu$. The following theorem shows that the converse is also true:

THEOREM 4.10.1. *(Radon-Nikodym) If $(\Omega, \mathcal{F}, \mathbb{P})$ is a probability measure and μ is a finite measure on $\mathcal{F}$ such that $\mu \preceq \mathbb{P}$, then there exists an $\mathcal{F}$-measurable function $X \geq 0$ with $\mathbb{E}X < \infty$ such that $d\mu = X d\mathbb{P}$.*

PROOF. *Step 1: assume that $\mathcal{F} = \sigma(A_1, \dots, A_n)$ for some sets $A_1, \dots, A_n$.* In that case, the theorem is elementary. Indeed, Ω can be partitioned into a disjoint union of sets $\Omega = \Omega_1 \sqcup \dots \sqcup \Omega_{2^n}$, where each Ω_i has the form $\Omega_i = B_1 \cap \dots \cap B_n$ with $B_i = A_i$ or $B_i = A_i^c$. In fact, $\mathcal{F}$ will be then an atomic σ-algebra[6] with atoms $\Omega_1, \dots, \Omega_{2^n}$, that is,

$$\mathcal{F} = \{\sqcup_{i \in I'} \Omega_i : I' \subset I\}.$$

The $\mathcal{F}$-measurable functions are just functions that are constants on atoms. Given $\omega \in \Omega$, let $i(\omega)$ be such that $\omega \in \Omega_i$. If we now *define* by

$$X(\omega) = \begin{cases} \frac{\mu(\Omega_{i(\omega)})}{\mathbb{P}(\Omega_{i(\omega)})}, & \text{if } \nu(\Omega_{i(\omega)}) \neq 0, \\ 0, & \text{else.} \end{cases}$$

Then, $\mu(A) = \mathbb{E}(X\mathbb{I}_A)$ when $A = \Omega_i$ for some i, and, therefore, by linearity, for any $A \in \mathcal{F}$.

Step 2: assume that $\mathcal{F} = \sigma(A_1, A_2, \dots)$ for a sequence $A_1, A_2, \dots$ of sets. This is where martingale theory enters the game. Define $\mathcal{F}_n = \sigma(A_1, \dots, A_n)$, and let X_n be the corresponding Radon-Nikodym derivative. Then, for any $A \in \mathcal{F}_n$, we have

$$\mathbb{E}(X_{n+1}\mathbb{I}_A) = \mu(A) = \mathbb{E}(X_n \mathbb{I}_A).$$

that is,

$$X_n = \mathbb{E}(X_{n+1}|\mathcal{F}_n),$$

that is, f_n is a martingale. In order to apply the martingale convergence theorem, we will check that X_n is UI. This is done much as in the proof of Proposition 4.8.7: first,

$$\mathbb{E}(X_n \mathbb{I}_{X_n > M}) = \mu(\{X_n > M\}).$$

Second, by Chebyshev's inequality,

$$\mathbb{P}(X_n > M) \leq M^{-1}\mathbb{E}X_n = M^{-1}\mu(\Omega).$$

Therefore, uniform integrability follows from the absolute continuity (Lemma 4.8.8).

We conclude that there is a random variable X such that $X_n \to X$ almost surely and in L^1, and it follows from Theorem 4.8.10 that $X_n = \mathbb{E}(X|\mathcal{F}_n)$. Therefore, if $A \in \mathcal{F}_n$ for some n, then $\mathbb{E}(\mathbb{I}_A X) = \mathbb{E}(\mathbb{I}_A X_n) = \mu(A)$. Therefore, the measures μ and $X d\mathbb{P}$ agree on the π-system $\cup \mathcal{F}_n$, so they agree on $\sigma(\cup \mathcal{F}_n)$ by Corollary 1.3.5.

Step 3: arbitrary $\mathcal{F}$. This part hinges on the theory of *nets* (a. k. a. generalized sequences, or Moore-Smith sequences). We come back to the proof after recalling the key notions of that theory. □

DEFINITION 4.10.2. A *directed set* is a partially ordered set $(\mathcal{S}, \preceq)$ such that any two elements have a common upper bound (that is, for any $a, b \in \mathcal{S}$, there exist $c \in \mathcal{S}$ such that $a \preceq c$ and $b \preceq c$). A *net* is a map $f : \mathcal{S} \mapsto M$, where $\mathcal{S}$ is a directed set and M is a topological space. A net is said to *converge* to an element $x \in M$, if for any neighborhood U of x, there exists $a \in \mathcal{S}$ such that $f(b) \in U$ for all $b \succeq a$.

EXAMPLE 4.10.3. The following are examples of nets:

[6] see Exercise set 1 of Probability theory I course

(1) A usual sequence is a net with $(\mathcal{S}, \preceq) = (\mathbb{N}, \leq)$, and the convergence is the usual convergence of sequences.
(2) A function $f : \mathbb{R} \to \mathbb{R}$ can be viewed as a net with $(\mathcal{S}, \preceq) = (\mathbb{R}, \leq)$, and the convergence of this net is equivalent to the usual convergence of f as $t \to +\infty$. If we define the directed set to be $\mathcal{S} = \mathbb{R} \setminus \{a\}$ the order by $x \preceq y$ if and only if $|x - a| \geq |y - a|$, then the convergence of this net is equivalent to convergence of $f(x)$ as $x \to a$.
(3) Given an interval $[a, b]$ and a function $f : [a, b] :\mapsto \mathbb{R}$, define

$$\mathcal{S} := \{a = t_0 \leq \xi_1 \leq t_1 \leq \xi_2 \leq \cdots \leq t_k = b, k \in \mathbb{N}\},$$

the set of partitions of $[a, b]$ into finitely many intervals $[t_0, t_1), \dots, [t_{k-1}, t_2)$, where each partition is equipped with points $\xi_1 \in [t_0, t_1], \dots \xi_k \in [t_{k-1}, t_k]$. For two such partitions T and T', we say that $T \preceq T'$ if $\{t_0, \dots, t_k\} \subset \{t'_0, \dots, t'_{k'}\}$ (that is, T' is a refinement of T. The points ξ_i are irrelevant for the order.) Define

$$f_T = \sum_{i=1}^{k} f(\xi_i)(t_i - t_{i-1}).$$

This is a net, and its limit is the Riemann integral of f.

The main motivation behing the study of nets is as follows: in the setting of metric spaces, certain key notions can be expressed in terms of sequences: e. g., a function is continuous iff it maps convergence sequences to convergent sequences, a space is compact if any sequence contain convergent subsequence, etc. For topological spaces, this is no more true, but the analogous statements are true if the sequences are replaced by nets.

Let us say that a σ-algebra $\mathcal{G}$ is *separable* if it is as in Step 2 of the proof of theorem 4.10.1, and denote the set of all such σ-algebras by $\mathcal{S}$. Note that if $\mathcal{G} = \sigma(A_1, A_2, \dots)$ and $\mathcal{G}' = \sigma(A'_1, A'_2, \dots)$ are separable, then $\sigma(\mathcal{G}_1, \mathcal{G}_2) = \sigma(A_1, A'_1, A_2, \dots)$ is also separable. That is to say, $(\mathcal{S}, \subset)$ is a directed set. The uniformly integrable collection $X_{\mathcal{G}}$ of the Radon-Nikodym derivatives, constructed in Step 2, is a net with values in $\mathcal{L}^1(\mathbb{P})$. Moreover, if $\mathcal{G}_1 \subset \mathcal{G}_2$, then, for every $A \in \mathcal{G}_1$, we have

$$\mathbb{E}(\mathbb{I}_A X_{\mathcal{G}_1}) = \mu(A) = \mathbb{E}(\mathbb{I}_A X_{\mathcal{G}_2}),$$

so $X_{\mathcal{G}_1} = \mathbb{E}(X_{\mathcal{G}_2}|\mathcal{G}_1)$. It is natural to call this structure *a uniformly integrable martingale net.*

PROPOSITION 4.10.4. *Let $(\mathcal{S}, \subset)$ be a directed set of σ-algebras, and let $X_{\mathcal{G}}, \mathcal{G} \in \mathcal{S}$, be a UI martingale net. Then there exists a random variable $X \in \mathcal{L}^1$ such that $X_{\mathcal{G}} \to X$ in $\mathcal{L}^1$.*

PROOF. We first claim that for every $\varepsilon > 0$, there exists a σ-algebra $\mathcal{G} \in \mathcal{S}$ such that if $\mathcal{G} \subset \mathcal{G}'$ then $\mathbb{E}|X_{\mathcal{G}} - X_{\mathcal{G}'}| < \varepsilon$ ("the net is a Cauchy net"). Indeed, assume the contrary. Then, there is an $\varepsilon > 0$ and a sequence of σ-algebras $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \dots$ such that $\mathbb{E}|X_{\mathcal{G}_k} - X_{\mathcal{G}_{k-1}}| > \varepsilon$. But we know that $X_{\mathcal{G}_1}, X_{\mathcal{G}_2}, \dots$ is a UI martingale, therefore, it converges in $\mathcal{L}^1$, which is a contradiction.

Now, let $\mathcal{G}_1, \mathcal{G}_2, \dots$ be σ-algebrae as above, corresponding to $\varepsilon_1 = \frac{1}{1}$, $\varepsilon_2 = \frac{1}{2}$ etc., that is, for every $\mathcal{G} \supset \mathcal{G}_k$, we have $\mathbb{E}(|X_{\mathcal{G}} - X_{\mathcal{G}_k}|) \leq \frac{1}{k}$. By replacing $\mathcal{G}_k$ with $\sigma(\mathcal{G}_1, \dots, \mathcal{G}_k)$, we may assume that $\mathcal{G}_1 \subset \mathcal{G}_2 \subset \dots$. Then $X_{\mathcal{G}_1}, X_{\mathcal{G}_2}, \dots$ is a UI martingale, and let X denote its limit. If $\mathcal{G}' \supset \mathcal{G}_k$, then

$$\mathbb{E}|X - X_{\mathcal{G}'}| \leq \mathbb{E}|X_{\mathcal{G}_k} - X_{\mathcal{G}'}| + \mathbb{E}|X_{\mathcal{G}_k} - X_{\mathcal{G}}| \leq \frac{2}{k}.$$

That is to say, $X_{\mathcal{G}} \to X$ in the sense of convergence of nets. □

REMARK 4.10.5. What we have essentially proved here is that if f is a net with values in a metric space, then f converges if and only if the sequence $f(a_1), f(a_2), \dots$ converges for every $a_1 \preceq a_2 \preceq \dots$.

PROOF OF STEP 3 OF THEOREM 4.10.1. Let X be the limit of the net $X_{\mathcal{G}}$, $\mathcal{G} \in \mathcal{S}$. We have to show that for any $A \in \mathcal{F}$, we have $\mu(A) = \mathbb{E}(\mathbb{I}_A X)$. Pick $\varepsilon > 0$ and let $\mathcal{G} \in \mathcal{S}$ be such that $\mathbb{E}(|X_{\mathcal{G}'} - X| < \varepsilon)$ for all $\mathcal{G}' \supset \mathcal{G}$. Define $\mathcal{G}' := \sigma(A, \mathcal{G})$; clearly, $\mathcal{G}' \in \mathcal{S}$, and

$$|\mathbb{E}(\mathbb{I}_A X) - \mu(A)| = \mathbb{E}(|X - X_{\mathcal{G}'}| + |\mathbb{E}(\mathbb{I}_A X_{\mathcal{G}'}) - \mu(A)| < \varepsilon,$$

since $A \in \mathcal{G}'$ and hence $\mathbb{E}(\mathbb{I}_A X_{\mathcal{G}'}) = \mu(A)$. Since ε is arbitrary, we are done. $\square$