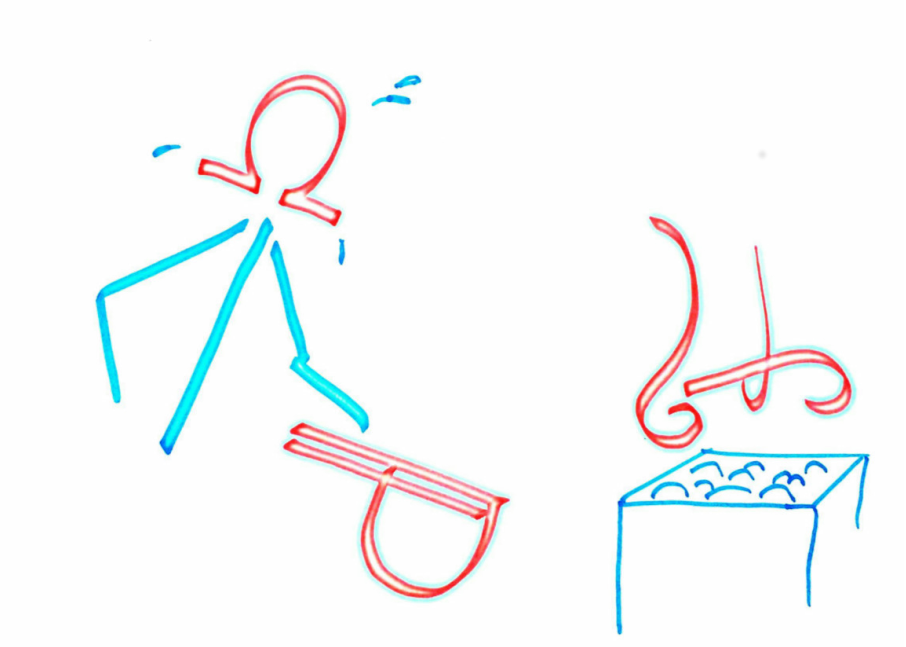


Probability Theory

Kalle Kytölä

Contents

Foreword

These lecture notes are primarily intended for the regular M.Sc. level course **MS-E1600** *Probability Theory* at Aalto University.

The principal aim of the course is to familiarize the students with the mathematical foundations of randomness. The reasons why one should study such a theoretical formalism vary according to the ambitions of the individual. The development of a logically solid theory of random phenomena should perhaps be seen as worthwhile in its own right. We will stick strictly to the fundamentals, so this course offers quite ideal practice on precise mathematical reasoning, including formulating proofs. A more pragmatic motivation might be that these theoretical foundations are necessary for following many subsequent courses in probability and statistics, and for understanding more advanced topics. In any case, whether one plans to work in statistics, machine learning, or pure mathematics, relevant research literature often requires familiarity of this theory as a language, e.g., being able to distinguish between convergence in probability, convergence in distribution, convergence almost surely, or other notions of convergence of random variables. The present course attempts to provide just enough of the core mathematical theory to develop an appreciation of such differences.

The course in its current format is very concise: 12 lectures and six sets of exercises during a six weeks period. One of the regrettable consequences is that there is almost no time to enter any of the interesting applications of the theory that is being developed. There are other courses devoted to more specific topics in probability which build on the theoretical foundations of the present course and come closer to actual applications.

In order to be prepared to internalize the theory during this concise course, the student must have a little bit of mathematical maturity to begin with. Besides some calculus of infinite series, differentiation, and integration, it is crucial to have a working knowledge of set theory, especially the notion of countability, and a little bit of familiarity with continuous functions in the context of metric spaces or topological spaces, say. Appendices A and B serve as quick reminders of such prerequisites, and before engaging in this text beyond the introduction, one should make sure to grasp their content.

The material in the chapters which correspond to the 12 lectures has been kept to minimum. A number of basic results that do not fit in these are left to Appendices C – F. The material in the main chapters can be considered as the minimum of what the students are expected to internalize during the short course, while the material in these appendices is something that one can expect to encounter soon after this basic course, and it can then be quickly picked up with a little bit of further effort.

There is already a vast number of textbooks in probability theory, and the contents of mathematics courses at advanced B.Sc. or early M.Sc. level have become quite standard. For the students of the present course we recommend in particular [**JP04**], because it is a very concise account of probability theory quickly covering very much the same topics as the present course. A slightly more challenging alternative is [**Wil91**], which is a remarkably well-written, mathematically elegant account of probability that manages incorporate fascinating and important probabilistic insights into a brief text. Both the theory and a significant number of interesting

and relevant examples and applications are covered in [**Dur10**]. The present lecture notes borrow shamelessly from all of the above sources. And the purpose of these notes is not to replace the best textbooks on the subject, but rather to provide the students an account that follows the structure and scope of the concise six weeks probability theory course as closely as possible.

The structure of these notes is largely based on an earlier version of the course taught by Lasse Leskelä, and on parts of the textbook [**Wil91**]. I have received very valuable comments, especially by Joona Karjalainen, Alex Karrila, and Niko Lietzén, as well as many students, which have lead to improvements to the notes. I am, of course, responsible for all remaining mistakes. Still, you could help me — and perhaps more importantly the students who will use this material — by sending comments about mistakes, misprints, needs for clarification, etc., to me by email (`kalle.kytola@aalto.fi`).

Glossary of notations

For convenience, we provide here a list of some of the used mathematical notation and abbreviations, together with brief explanations or references to the appropriate definitions during the course.

Numbers

$\mathbb{Z}$	the set of integers	$\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\}$
$\mathbb{Z}_{\geq 0}$	the set of non-negative integers	$\mathbb{Z}_{\geq 0} = \{0, 1, 2, 3, \ldots\}$
$\mathbb{N}$	the set of natural numbers	$\mathbb{N} = \{1, 2, 3, 4, \ldots\}$
$\mathbb{Q}$	the set of rational numbers	$\mathbb{Q} = \left\{ \frac{n}{m} \;\middle\vert\; n, m \in \mathbb{Z},\ m \neq 0 \right\}$
$\mathbb{R}$	the set of real numbers	
$\mathbb{C}$	the set of complex numbers	$\mathbb{C} = \left\{ x + \mathbb{i}y \;\middle\vert\; x, y \in \mathbb{R} \right\}$
$\mathbb{i}$	imaginary unit	$\mathbb{i} = \sqrt{-1} \in \mathbb{C}$
(a, b)	open interval	$(a, b) = \left\{ x \in \mathbb{R} \;\middle\vert\; a < x < b \right\}$
$[a, b]$	closed interval	$[a, b] = \left\{ x \in \mathbb{R} \;\middle\vert\; a \leq x \leq b \right\}$
$(a, b]$		$(a, b] = \left\{ x \in \mathbb{R} \;\middle\vert\; a < x \leq b \right\}$
$[a, b)$		$[a, b) = \left\{ x \in \mathbb{R} \;\middle\vert\; a \leq x < b \right\}$

Logical notation

$\Rightarrow$	logical implication ("only if")	$P \Rightarrow Q$ means: if P is true then also Q is true
$\Leftarrow$	reverse logical implication ("if")	$P \Leftarrow Q$ means: if Q is true then also P is true
$\Leftrightarrow$	logical equivalence	$P \Leftrightarrow Q$ means: P is true if and only if Q is true
$\forall$	"for all" (logical quantifier)	
$\exists$	"there exists" (logical quantifier)	
s.t.	such that	
iff	if and only if	

Set operations

$s \in S$	s is an element of set S	
$\#S$	number of elements in the set S	
$\subset$	subset relation	$A \subset B$ means: if $x \in A$ then also $x \in B$
$\emptyset$	the empty set	
$\mathscr{P}$	power set / collection of all subsets	$\mathscr{P}(S) = \{A \subset S\}$
$\cup, \bigcup$	union	$A \cup B = \{x \mid x \in A \text{ or } x \in B\}$
		$\bigcup_j A_j = \{x \mid x \in A_j \text{ for some } j\}$
$\cap, \bigcap$	intersection	$A \cap B = \{x \mid x \in A \text{ and } x \in B\}$
		$\bigcap_j A_j = \{x \mid x \in A_j \text{ for all } j\}$
$\times$	Cartesian product	$A \times B = \{(a, b) \mid a \in A,\ b \in B\}$
$\setminus$	set difference	$A \setminus B = \{x \mid x \in A,\ x \notin B\}$
$(\cdots)^{\mathbf{c}}$	complement of $\cdots$	
lim sup		$\limsup_n A_n := \bigcap_{m \in \mathbb{N}} \bigcup_{n \geq m} A_n$
lim inf		$\liminf_n A_n := \bigcup_{m \in \mathbb{N}} \bigcap_{n \geq m} A_n$

Probability and measure theory

Ω	sample space / the set of outcomes	
ω	an outcome	$\omega \in \Omega$
$\mathscr{F}$	sigma-algebra / the collection of events	see Lecture I
P	probability measure	see Lecture II
E	expected value	see Lecture VII
$\mathbb{I}_A$	indicator of event A	
$\perp\!\!\!\perp$	independent	see Lecture V
a.s.	almost surely / with probability one	Remark II.6

Specific sigma algebras

$\mathscr{S}$	generic sigma-algebra on a set S	Definition I.1
$\mathscr{F}$	generic sigma-algebra on the sample space Ω	
$\mathscr{B}(\mathfrak{X})$	Borel sigma-algebra on a topological space $\mathfrak{X}$	Definition I.10
$\mathscr{B} = \mathscr{B}(\mathbb{R})$	Borel sigma-algebra on the real line $\mathbb{R}$	Section I.3.1
$\mathscr{T}_\infty$	tail sigma-algebra	Equation (VI.4)
$\sigma(\cdots)$	the sigma-algebra generated by $\cdots$	Definitions I.6 and IV.1

Specific measures

μ	generic measure on measurable space $(S, \mathscr{S})$	Definition II.4
$\mu_\#$	counting measure (on some set)	Example II.10
P	generic probability measure on sample space Ω	Definition II.5
Λ	Lebesgue measure on the real line $\mathbb{R}$	Example II.12

Spaces of random variables

$\mathrm{m}\mathscr{F}$	random variables measurable w.r.t. $\mathscr{F}$
$\mathrm{m}\mathscr{F}^+$	non-negative random variables measurable w.r.t. $\mathscr{F}$
$\mathrm{s}\mathscr{F}$	simple random variables measurable w.r.t. $\mathscr{F}$
$\mathrm{s}\mathscr{F}^+$	non-negative simple random variables measurable w.r.t. $\mathscr{F}$
$\mathrm{b}\mathscr{F}$	bounded random variables measurable w.r.t. $\mathscr{F}$
$\mathrm{b}\mathscr{F}^+$	non-negative bounded random variables measurable w.r.t. $\mathscr{F}$
$\mathcal{L}^p$	p-integrable random variables

Notions of convergence

$\xrightarrow{\text{a.s.}}$	convergence almost surely	Definition XI.1
$\xrightarrow{\mathsf{P}}$	convergence in probability	Definition XI.2
$\xrightarrow{\mathcal{L}^1}$	convergence in $\mathcal{L}^1(\mathsf{P})$	Definition XI.8
$\xrightarrow{\text{law}}$	convergence in distribution (law)	Definition XII.10

Lecture O

Introduction

O.1. What are the basic objects of probability theory?

Probability theory forms the mathematically precise and powerful foundations for the study of randomness. Its most basic objects — defined and studied in the rest of this course — are:

Ω — **Outcomes (of a random experiment)**
: An *outcome* ω of a random experiment represents a single *realization* of the randomness involved. The *sample space* Ω is the set consisting of all possible outcomes.

$\mathscr{F}$ — **Events**[1]
: An *event* E is a subset $E \subset \Omega$ of the set of possible outcomes. The event E is said to occur if the randomly realized outcome $\omega \in \Omega$ belongs to this subset, i.e., if $\omega \in E$. Generally we can not allow all subsets of Ω as events, but instead we have to select a suitable collection $\mathscr{F}$ of subsets on which it is possible to have consistent rules of probability.

P — **Probability (measure)**[2]
: To each event E we assign the *probability* $\mathsf{P}[E]$ of the event, which is a real number between 0 and 1.

In addition to the three basic objects $(\Omega, \mathscr{F}, \mathsf{P})$ above, the following two fundamental notions will also be indispensable:

Random variable[3]
: Random variables are the quantities of interest in our probabilistic model. A *random variable* is a suitable function $X\colon \Omega \to S$, associating to each possible outcome $\omega \in \Omega$ a value $X(\omega) \in S$. You may think of the *Goddess of Chance* choosing the outcome ω at random, and the chosen outcome subsequently determining the value $X(\omega)$ of any quantity of interest.

Expected value[4]
: The *expected value* $\mathsf{E}[X]$ of a real-valued quantity of interest X, i.e., a random variable $X\colon \Omega \to S \subset \mathbb{R}$, represents an average of the possible values of X over all randomness, weighted according to probabilities P. The expected value is an integral with respect to the probability measure P in the sense of Lebesgue, and we will correspondingly use the

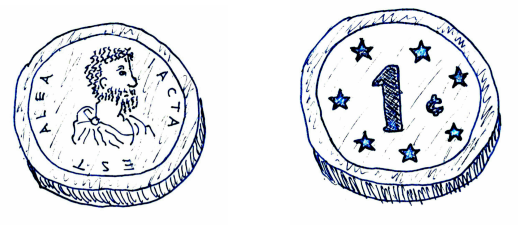
[1] We will address the precise axioms required of the collection $\mathscr{F}$ of events in Lecture I.
[2] We will address the precise axioms required of the probability measure P in Lecture II.
[3] Random variables will be defined precisely in Lecture III.
[4] Expected values will be defined precisely in Lecture VII.

following notations interchangeably

$$\mathsf{E}\big[X\big] = \int_\Omega X(\omega)\, \mathrm{d}\mathsf{P}(\omega).$$

The purpose of this course is to make precise mathematical sense of the above notions. Before rushing into the theory, however, we continue with a brief informal introduction. For the informal examples below, it suffices to have an intuitive idea of the above notions.

O.2. Informal examples of the basic objects in random phenomena



Example O.1 (One coin toss)**.**
The possible outcomes of a single coin toss are "heads" and "tails", abbreviated H and T, respectively. The sample space of a single coin toss experiment would thus be

$$\Omega = \{\mathrm{H}, \mathrm{T}\}.$$

As events, we can in this case allow all subsets of Ω, so the collection of events is

$$\mathscr{F} = \Big\{\ \emptyset,\ \{\mathrm{H}\},\ \{\mathrm{T}\},\ \{\mathrm{H}, \mathrm{T}\}\ \Big\},$$

with interpretations of the events:

$\{\mathrm{H}\}$	the event that the coin toss results in "heads"
$\{\mathrm{T}\}$	the event that the coin toss results in "tails"
$\emptyset$	the event that the coin toss results in neither "heads" nor "tails"
$\{\mathrm{H}, \mathrm{T}\}$	the event that the coin toss results in either "heads" or "tails"

The last two events may appear perplexingly trivial, but we want to allow them as events, because logical reasoning with other events may result in impossibilities or certainties. In fact, $\emptyset \subset \Omega$ always corresponds to the impossible event, which is not realized by any possible outcome $\omega \in \Omega$, whereas $\Omega \subset \Omega$ always corresponds to the sure event which is realized by any outcome $\omega \in \Omega$ of the randomness.

A fair coin toss is considered equally likely to result in "heads" or "tails", and a single fair coin toss is thus unsurprisingly governed by the probability measure P which assigns the following probabilities to the above events:

$$\mathsf{P}\big[\,\{\mathrm{H}\}\,\big] = \frac{1}{2}, \qquad \mathsf{P}\big[\,\{\mathrm{T}\}\,\big] = \frac{1}{2}, \qquad \mathsf{P}\big[\emptyset\big] = 0, \qquad \mathsf{P}\big[\,\{\mathrm{H}, \mathrm{T}\}\,\big] = 1.$$

This example is not overly exciting, but the distinct roles of the three basic objects Ω, $\mathscr{F}$, and P should be recognized here!

Example O.2 (Repeated coin tossing)**.**
In an experiment where coin tosses are repeated ad infinitum, the possible outcomes are all possible sequences of "heads" and "tails", i.e., functions from $\mathbb{N}$ to $\{\mathrm{H}, \mathrm{T}\}$. The sample space of such a repeated coin tossing experiment would be the space of all such functions

$$\Omega = \{\mathrm{H}, \mathrm{T}\}^{\mathbb{N}} = \Big\{\omega \colon \mathbb{N} \to \{\mathrm{H}, \mathrm{T}\}\Big\},$$

which is uncountably infinite (it can be identified with the set of infinite binary sequences, Example A.16). This uncountable cardinality in a rather innocent probabilistic model can be taken as the first warning that some care is needed in a proper mathematical treatment of probability.

Let X_n denote the relative frequency of heads in the first n coin tosses. The relative frequency is a function of the outcome ω (as random variables generally are!), given by the ratio

$$X_n(\omega) = \frac{\#\left\{j \mid j \leq n \text{ and } \omega(j) = \mathrm{H}\right\}}{n}.$$

We may ask whether the frequency X_n tends to $\frac{1}{2}$ in the long run, as $n \to \infty$. This is certainly not true for all $\omega \in \Omega = \{\mathrm{H}, \mathrm{T}\}^{\mathbb{N}}$. For example for the sequence $\omega' = (\mathrm{H}, \mathrm{H}, \mathrm{H}, \mathrm{H}, \ldots)$ of all heads, we have $X_n(\omega') = 1$ for all n and therefore $\lim_{n\to\infty} X_n(\omega') = 1 \neq \frac{1}{2}$. Even worse, for the sequence

$$\omega'' = (\mathrm{H}, \underbrace{\mathrm{T}, \mathrm{T}, \mathrm{T}}_{3 \text{ times}}, \underbrace{\mathrm{H}, \mathrm{H}, \mathrm{H}, \mathrm{H}, \mathrm{H}, \mathrm{H}, \mathrm{H}, \mathrm{H}, \mathrm{H}}_{3^2 \text{ times}}, \underbrace{\mathrm{T}, \ldots, \mathrm{T}}_{3^3 \text{ times}}, \ldots)$$

the limit $\lim_{n\to\infty} X_n(\omega'')$ does not even exists. In view of these "counterexamples", it should be clear that any statement about X_n tending to $\frac{1}{2}$ in the long run must be probabilistic in nature (somehow referring to P), rather than pointwise on the entire sample space Ω.

There are various possible probabilistic notions of X_n tending to $\frac{1}{2}$ as $n \to \infty$. Compare for example the following two statements:

(a) For any $\varepsilon > 0$ we have

$$\lim_{n\to\infty} \mathsf{P}\Big[|X_n - \frac{1}{2}| > \varepsilon\Big] = 0.$$

(b) We have

$$\mathsf{P}\Big[\lim_{n\to\infty} X_n = \frac{1}{2}\Big] = 1.$$

As we develop the mathematical foundations of probability, we should ask, for example:

- Can we make precise mathematical sense of statements such as (a) and (b) above?
- Does one of these two statements imply the other?
- Is either of the statements actually true?
  (say for the usual probability P governing fair, repeated coin tossing)

The next example further illustrates the types of questions one may want to answer with the mathematical theory of probability. It concerns a branching population model known as Galton-Watson process.[5] In the footnotes we indicate which parts of the present course are relevant for the questions that arise, but the reader interested in the detailed solutions should look them up in books on stochastic processes.

[5]The idea of using this branching process to illustrate the applicability of probability theory is borrowed from the excellent textbook [**Wil91**].

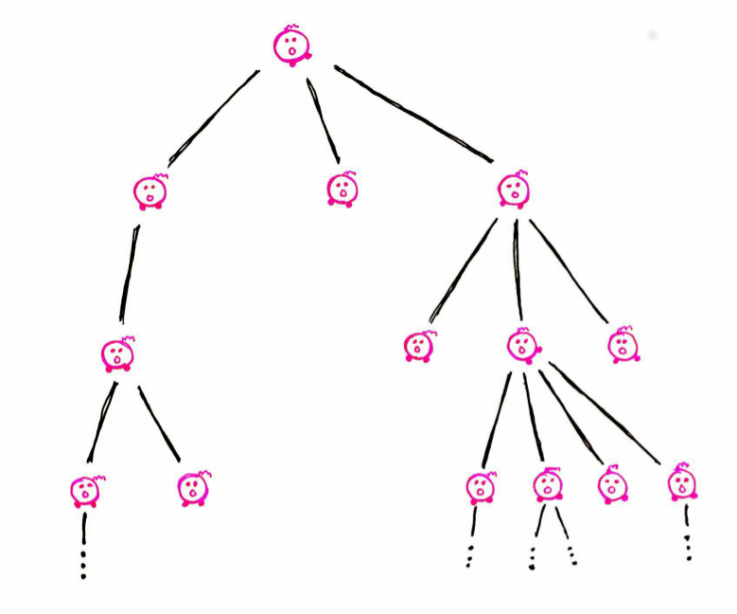
Example O.3 (Branching process)**.**

Consider a population producing offspring randomly as follows (this population model is known as the Galton-Watson process).

The population is started from a single ancestor, who has a random number of descendants according to some given distribution. These descendants of the ancestor form the first generation in the population, and we denote the random number of individuals in the first generation by Z_1.

Each of the Z_1 individuals in the first generation then independently of each other and the ancestor has a random number of its own descendants according to the same distribution. These descendants of the descendants of the ancestor form the second generation, and we denote the number of individuals in the second generation by Z_2.

The process continues branching in the same way, with all individuals in each generation having descendants randomly, which together form the next generation. The numbers of descendants of each individual are assumed to follow the same distribution and to be independent of each other. The random number of individuals in the n:th generation is denoted by Z_n.



The first question that this model poses in the foundations of probability theory is:

Is there a well defined mathematical model of this branching process?

In particular, what are the sample space Ω, the collection of events $\mathscr{F}$, and the probability measure P describing the process?[6] At least each Z_n should be a random variable, i.e., a function of the realization $\omega \in \Omega$ of all involved randomness.

Admitting that the model can be defined, we can start asking more interesting questions about the model itself. One such question concerns the survival of the population, or conversely the extinction of the population:

Does the lineage of the ancestor ever terminate?

The lineage terminates, i.e., the population becomes extinct, if in some generation $n \in \mathbb{N}$ there are no individuals, $Z_n = 0$. The event E_n that the generation n contains no individuals consists of those outcomes ω of the randomness for which $Z_n(\omega) = 0$, i.e.,

$$E_n = \Big\{ \omega \in \Omega \;\Big|\; Z_n(\omega) = 0 \Big\}.$$

[6]It turns out that countable product spaces of rather simple discrete probability spaces are sufficient to precisely construct the model, cf. Lecture IX.

We may observe that if the generation n contains no individuals, then the next generation $n+1$ can not contain any either, so

$$Z_n(\omega) = 0 \quad \Longrightarrow \quad Z_{n+1}(\omega) = 0.$$

Equivalently, for the corresponding events we have the inclusion

$$E_n \subset E_{n+1}.$$

These events thus form an increasing sequence (of subsets of Ω)

$$E_1 \subset E_2 \subset E_3 \subset \cdots.$$

The probabilities of the events should increase correspondingly,

$$\mathsf{P}[E_1] \;\leq\; \mathsf{P}[E_2] \;\leq\; \mathsf{P}[E_3] \;\leq\; \cdots$$

and as a bounded increasing sequence of real numbers, these probabilities have a limit

$$\lim_{n\to\infty} \mathsf{P}[E_n].$$

One natural question is: can we use the events E_n to construct the event E that extinction occurs eventually?[7] And is its probability $\mathsf{P}[E]$ equal to the above limit $\lim_{n\to\infty} \mathsf{P}[E_n]$?[8] A more ambitious version of the question is: can we concretely calculate the probability $\mathsf{P}[E]$ of eventual extinction?[9]

The expected size $\mathsf{E}[Z_n]$ of generation n turns out to be d^n, where d is the expected number of descendants of one individual. Imagine then that we define the "renormalized size" of generation n as $R_n = d^{-n}\, Z_n$, so as to have expected value one, $\mathsf{E}[R_n] = 1$. By techniques that go just a little bit beyond the present course (martingale convergence theorem) one can show that there exists a limit of the random variables R_n as $n \to \infty$. The limit

$$\lim_{n\to\infty} R_n$$

is itself a random variable, which can be interpreted to describe the asymptotic long term size of the population in units that make the expected sizes equal to one. Given this, it is natural to wonder whether the expected value of this asymptotic quantity can be calculated by interchanging the limit and the expected value[10]

$$\mathsf{E}\Big[\lim_{n\to\infty} R_n\Big] \;\stackrel{?}{=}\; \lim_{n\to\infty} \underbrace{\mathsf{E}\Big[R_n\Big]}_{=1} \;=\; 1.$$

Probability theory provides the means to make sense of and answer the many questions that arise in this branching process example — as well as in other interesting models.

O.3. Probability theory vs. measure theory

The present course may appear to involve not so much of random phenomena themselves, but more of dry and formal measure theory instead. Our justification for this is that virtually all advanced probability and statistics builds on the measure theoretical foundations covered in the present course. Frequently used measure theoretical

[7] It turns out that σ-algebras, studied in Lecture I, permit just flexible enough logical operations to allow such a construction.

[8] This indeed turns out to be true, by monotonicity properties of probability measures established in Lecture II.

[9] This can be done using generating functions — a close cousin of the characteristic functions studied in Lecture XII.

[10] Whether this can be done turns out to be subtle. In Lectures VII, VIII, and IX we will learn under which conditions one can interchange the order of limits and expected values (and integrals and sums, etc.). Such interchanges of order of operations are tremendously useful in many calculations in practice.

tools in stochastics include Dominated Convergence Theorem, Monotone Convergence Theorem, Fubini's theorem, L^p-spaces, etc. The formal foundations also serve as a common and well-defined language across different branches of stochastics. For example, various different notions of convergence of random variables used in mathematical statistics are what we will be ready to introduce in the last two of our lectures. The role of the present course is to develop the mathematical foundations of probability mainly for future use!

Since a large number of basic definitions and results in measure theory and probability theory are literally identical, anyone who has already studied measure theory will recognize many familiar notions. The overlap may raise the question: are measure theory and probability theory really separate topics in their own right, and is it necessary to study them separately? It would, in fact, be possible to combine measure theory and probability theory in one extended and coherent course, but the scope of that could easily become daunting. As the topics are currently taught in separate courses, it does not matter much whether one first studies measure theory and later learns about its applicability to probability, or if one proceeds in the opposite order.

Even concerning abstract measure theoretic notions and results, probability theory in fact offers very interesting and useful interpretations. In this course, such interpretations include, e.g.,

- product measures interpreted as probabilistic independence
- push-forward measures interpreted as laws of random variables
- (sub-)sigma-algebras interpreted as (partial) information.

At its best, probabilistic thinking leads to entirely new techniques in mathematics, such as

- coupling arguments
- existence proofs relying on random choice.

And finally, probability theory includes inherently stochastic results which do not belong to the domain of measure theory. Some such results covered in the present course are:

- zero-one laws (Borel-Cantelli lemmas, Kolmogorov's 0-1 law)
- laws of large numbers
- central limit theorems.

And despite the similarities, measure theory and probability theory are ultimately concerned with different questions. To highlight just one difference in emphasis, note that the identification of the law of a random variable occupies a much more central place in probability theory than the corresponding question does in analysis. In developments beyond this first theoretical course, it becomes even more apparent that probability theory is not just a subset of measure theory[11] — consider, e.g., martingales, ergodic theory, large deviations, stochastic calculus, optimal stopping, etc. It is also in such further studies, which build on the present foundational course, that the advantages of the theory will become clearer.

[11]Vice versa, of course, there are topics covered in courses of measure theory such as [**Kin16**], which are not covered in this course of probability theory, and there are aspects of the theory into which one gains valuable insights from analysis. Therefore, especially for serious mathematicians, it is highly recommended to study both topics!

We hope that these reassurances and the occasional genuinely probabilistic interpretations and results included in the lectures provide a sufficient motivation to seriously study also the formal (measure theoretical) aspects of probability!

Lecture I

Structure of event spaces

I.1. Set operations on events

Recall that an event is a subset $E \subset \Omega$ of the set of all possible outcomes, and the event is said to occur if the outcome $\omega \in \Omega$ that is realized belongs to this subset, $\omega \in E$. We then have the following interpretations of set operations on events:

		interpretation
the whole sample space	Ω	sure event (contains all possible outcomes)
the empty set	$\emptyset$	impossible event (contains no outcomes)
intersection	$E_1 \cap E_2$	"events E_1 and E_2 both occur"
union	$E_1 \cup E_2$	"event E_1 or event E_2 occurs"
complement	$E^{\mathbf{c}} = \Omega \setminus E$	"event E does not occur"
subset	$E_1 \subset E_2$	"occurrence of E_1 implies E_2"

In other words, set theoretic operations enable logical reasoning with events. The logical operations `and`, `or`, and `not` are implemented by intersections $\cap$, unions $\cup$, and complements $(\cdots)^{\mathbf{c}}$, respectively.[1] We therefore at least want that the collection $\mathscr{F}$ of events is stable under such operations, i.e.,

- $\Omega \in \mathscr{F}$ and $\emptyset \in \mathscr{F}$
- if $E_1, E_2 \in \mathscr{F}$ then $E_1 \cap E_2 \in \mathscr{F}$ and $E_1 \cup E_2 \in \mathscr{F}$
- if $E \in \mathscr{F}$ then $E^{\mathbf{c}} = \Omega \setminus E \in \mathscr{F}$.

In fact, for a meaningful mathematical theory, we need to be able to form also countably infinite intersections and unions of events. This is the reason for the following definition.

[1] Intersection $\cap$ is indeed the equivalent of the logical quantifier "for all", $\forall$, and union $\cup$ is the equivalent of the logical quantifier "for some", i.e., "there exists", $\exists$.

I.2. Definition of sigma algebra

Definition I.1 (Sigma algebra)**.**
A collection $\mathscr{F} \subset \mathscr{P}(\Omega)$ of subsets of a set Ω is a *σ-algebra* on Ω if

$$
\begin{aligned}
&(\Sigma\text{-}1): && \Omega \in \mathscr{F} \\
&(\Sigma\text{-}\mathbf{c}): && \text{if } E \in \mathscr{F} \text{ then } E^{\mathbf{c}} = \Omega \setminus E \in \mathscr{F} \\
&(\Sigma\text{-}\cup): && \text{if } E_1, E_2, \ldots \in \mathscr{F} \text{ then } \bigcup_{n\in\mathbb{N}} E_n \in \mathscr{F}.
\end{aligned}
$$

Remark I.2 (A sigma algebra is stable under countable set operations)**.**
Note that properties (Σ-1) and (Σ-$\mathbf{c}$) imply that $\emptyset \in \mathscr{F}$, since $\emptyset = \Omega^{\mathbf{c}}$.

Likewise, properties (Σ-$\cup$) and (Σ-$\mathbf{c}$) imply that if $E_1, E_2, \ldots \in \mathscr{F}$ then also $\bigcap_{n=1}^{\infty} E_n \in \mathscr{F}$, since $\bigcap_{n=1}^{\infty} E_n = \big(\bigcup_{n=1}^{\infty} {E_n}^{\mathbf{c}}\big)^{\mathbf{c}}$ by De Morgan's laws, Proposition A.1.

Also, since $\emptyset \in \mathscr{F}$, we can extend any finite sequence $E_1, E_2, \ldots, E_k \in \mathscr{F}$ of members of the collection to an infinite sequence by setting $E_n = \emptyset$ for all $n > k$, and we thus deduce from (Σ-$\cup$) that the finite union $E_1 \cup \cdots \cup E_n \in \mathscr{F}$ also belongs to the collection. Similarly, finite intersections $E_1 \cap \cdots \cap E_n$ of members of the collection remain in the collection.

In view of the definition and remark above, σ-algebras are stable under countable set operations. Since we will always assume the collection of events to be a σ-algebra, we are thus allowed to perform rather flexible logical constructions with events.

Example I.3 (Examples and counterexamples of sigma algebras)**.**

(i) $\mathscr{F} = \{\emptyset, \Omega\}$ is a σ-algebra on Ω, albeit not a very interesting one: it only contains the impossible event $\emptyset$ and the sure event Ω.
(ii) $\mathscr{F} = \mathscr{P}(\Omega)$, the collection of all subsets of Ω, is a σ-algebra on Ω. However, when Ω is uncountably infinite, consistent rules of probability can typically only be given on a smaller collection of events.
(iii) $\mathscr{F} = \mathscr{T}(\mathbb{R})$, the collection of all open subsets of $\mathbb{R}$, is <u>not</u> a σ-algebra on $\mathbb{R}$! You should recall that arbitrary unions of open sets are open, and finite intersections of open sets are also open. However, for example the countable intersection $\bigcap_{n=1}^{\infty}(-\frac{1}{n}, \frac{1}{n}) = \{0\}$ of open intervals consists of a single point and is not open. The collection in fact satisfies (Σ-1) and (Σ-$\cup$), but fails to satisfy (Σ-$\mathbf{c}$).

Exercise I.1 (Sigma algebras on small finite sets)**.**
Let a, b, c be three distinct points.

(a) Write down all σ-algebras on $\Omega = \{a, b\}$.
(b) Write down all σ-algebras on $\Omega = \{a, b, c\}$.
(c) Give an explicit counterexample which shows that the union of two σ-algebras is not necessarily a σ-algebra.

In probability theory, we require the collection of events $\mathscr{F}$ to be a σ-algebra on the sample space Ω. The next examples illustrate what sorts of countable set operations we might encounter in practice. These examples also give a fair idea of the expressive power of such operations, when used iteratively.

Example I.4 (The event of branching process extinction)**.**
Let us revisit Example O.3 about the branching process. Let Z_n denote the random size of the population in generation $n \in \mathbb{N}$, which, as any random variable, depends on the outcome ω of the underlying randomness. Consider first an event defined by the condition

that the generation n contains no individuals

$$E_n = \Big\{ \omega \in \Omega \;\Big|\; Z_n(\omega) = 0 \Big\}.$$

Extinction happens in some generation in the future if there exists some $n \in \mathbb{N}$ such that $Z_n = 0$. The corresponding event is

$$\begin{aligned} E &= \Big\{ \omega \in \Omega \;\Big|\; \exists n \in \mathbb{N} : Z_n(\omega) = 0 \Big\} \\ &= \bigcup_{n\in\mathbb{N}} \Big\{ \omega \in \Omega \;\Big|\; Z_n(\omega) = 0 \Big\} = \bigcup_{n\in\mathbb{N}} E_n. \end{aligned}$$

We see that the event E of eventual extinction is the union over the countably many generations $n \in \mathbb{N}$ of the events E_n that generation n is already extinct.

Countable set operations came to our rescue!

Example I.5 (Long term frequency of heads in coin tossing)**.**
Let us revisit Example O.2 about repeated coin tossing. Let X_n be the relative frequency of heads in the first n coin tosses. Observe that the following are logically equivalent ways of expressing the property that the frequency tends to $\frac{1}{2}$ in the long run:

$$\begin{aligned} \lim_{n\to\infty} X_n = \frac{1}{2} \quad &\iff \quad \forall \varepsilon > 0 \; \exists k \in \mathbb{N} \text{ such that } \forall n \geq k \text{ we have } \Big|X_n - \frac{1}{2}\Big| < \varepsilon \\ &\iff \quad \forall m \in \mathbb{N} \; \exists k \in \mathbb{N} \text{ such that } \forall n \geq k \text{ we have } \Big|X_n - \frac{1}{2}\Big| < \frac{1}{m}. \end{aligned}$$

The last expression requires only quantifiers over countable collections, and is therefore good for our purposes. Consider first an event by the condition $|X_n - \frac{1}{2}| < \frac{1}{m}$,

$$E_n^{(m)} := \Big\{ \omega \in \{\mathrm{H},\mathrm{T}\}^{\mathbb{N}} \;\Big|\; \frac{1}{2} - \frac{1}{m} < X_n(\omega) < \frac{1}{2} + \frac{1}{m} \Big\},$$

we can now express the event E that the frequency tends to $\frac{1}{2}$ in the long run by the following countable set operations:

$$E = \bigcap_{m\in\mathbb{N}} \bigcup_{k\in\mathbb{N}} \bigcap_{n\geq k} E_n^{(m)}.$$

So at least provided that each $E_n^{(m)}$ belongs to the collection $\mathscr{F}$ of admissible events (seems reasonable) the properties of σ-algebras allow us to construct the more complicated but much more interesting event $E \in \mathscr{F}$, which contains precise information about the long term behavior of the frequencies of heads.

Iteratively constructed countable set operations came to our rescue!

I.3. Generating sigma algebras

Definition I.6 (Sigma algebra generated by a collection of subsets)**.**
Let $\mathscr{C} \subset \mathscr{P}(\Omega)$ be a collection of subsets of Ω. Then we define $\sigma(\mathscr{C})$ as the smallest σ-algebra on Ω which contains the collection $\mathscr{C}$. We call $\sigma(\mathscr{C})$ the *σ-algebra generated by the collection $\mathscr{C}$*.

Remark I.7. The language of the above definition is intended to be as accessible as possible, but let us make sure that the precise meanings are clear as well:

- We say that a σ-algebra $\mathscr{F}$ contains the collection $\mathscr{C}$, if each member of $\mathscr{C}$ is a member in $\mathscr{F}$, i.e., if we have the inclusion $\mathscr{C} \subset \mathscr{F}$ of the collections of sets.
- For two σ-algebras $\mathscr{F}_1$ and $\mathscr{F}_2$, we say that $\mathscr{F}_1$ is smaller than $\mathscr{F}_2$ if $\mathscr{F}_1 \subset \mathscr{F}_2$.

With these clarifications, the meaning should be unambiguous, but one still has to verify that $\sigma(\mathscr{C})$ becomes well defined. If we want to define $\sigma(\mathscr{C})$ as the smallest σ-algebra containing $\mathscr{C}$,

we first need to know that such a σ-algebra exists and that it is unique![2] These concerns will be settled in Corollary I.9 below.

One standard use of generated σ-algebras is the following. If $\mathscr{C}$ is a collection of some basic events that we want to be able to discuss, in our definition of a probabilistic model we could set $\mathscr{F} = \sigma(\mathscr{C})$, which is exactly the smallest possible collection of events that contains the basic events and behaves well under countable operations. Isn't this convenient!

In Lecture IV we discuss the interpretation of σ-algebras as describing information. We will realize that the notion of generated σ-algebras corresponds to what information can be deduced from some initially given pieces of information (the knowledge about events in the generating collection $\mathscr{C}$).

Finally, generated σ-algebras can also be used as a technical tool. It is often very difficult to describe explicitly all members of even very common and reasonable σ-algebras. Working with suitably chosen generating collections can bring about significant simplifications.

Having thus motivated the notion of generated σ-algebras, let us finally address their well-definedness. The key observation is that intersections of σ-algebras are themselves σ-algebras.[3]

Lemma I.8. *Suppose that $(\mathscr{F}_\alpha)_{\alpha\in I}$ is a non-empty collection (indexed by $I \neq \emptyset$) of σ-algebras $\mathscr{F}_\alpha$ on Ω. Then also the intersection $\mathscr{F} = \bigcap_{\alpha\in I} \mathscr{F}_\alpha$ is a σ-algebra on Ω.*

Proof. By requiring the collection to be non-empty, we ensured that the intersection is well-defined.

We need to verify that the intersection $\bigcap_{\alpha\in I} \mathscr{F}_\alpha$ satisfies the three properties in Definition I.1. Note that for a subset $E \subset \Omega$, we have $E \in \mathscr{F} = \bigcap_{\alpha\in I} \mathscr{F}_\alpha$ if and only if $E \in \mathscr{F}_\alpha$ for all $\alpha \in I$.

We have $\Omega \in \mathscr{F}_\alpha$ for all $\alpha \in I$, and therefore $\Omega \in \mathscr{F}$. Thus condition (Σ-1) holds for $\mathscr{F}$.

Suppose that $E \in \mathscr{F}$. Then for all $\alpha \in I$ we have $E \in \mathscr{F}_\alpha$. By property (Σ-**c**) for the σ-algebra $\mathscr{F}_\alpha$ we get that $E^{\mathbf{c}} \in \mathscr{F}_\alpha$. Since this holds for all α, we conclude $E^{\mathbf{c}} \in \mathscr{F}$. Thus condition ($\Sigma$-**c**) holds for $\mathscr{F}$.

Suppose that $E_1, E_2, \ldots \in \mathscr{F}$. Then for all $\alpha \in I$ we have $E_1, E_2, \ldots \in \mathscr{F}_\alpha$. By property ($\Sigma$-$\cup$) for the σ-algebra $\mathscr{F}_\alpha$ we get that $\bigcup_{n=1}^\infty E_n \in \mathscr{F}_\alpha$. Since this holds for all α, we conclude $\bigcup_{n=1}^\infty E_n \in \mathscr{F}$. Thus condition ($\Sigma$-$\cup$) holds for $\mathscr{F}$. □

We are now ready to conclude that Definition I.1 indeed made sense.

Corollary I.9 (Well-definedness of the generated sigma algebra)**.** *Let $\mathscr{C} \subset \mathscr{P}(\Omega)$ be a collection of subsets of Ω. Then the smallest σ-algebra on Ω which contains the collection $\mathscr{C}$ exists and is unique.*

[2]Such issues must be taken seriously. To illustrate the existence issue, imagine trying to define $s > 0$ as the smallest real number which is strictly positive: no such thing exists, and if we disregard that fact, we will soon run into logical contradictions. To illustrate the uniqueness issue, suppose that a, b, c are three distinct elements, and imagine trying to define $S \subset \{a, b, c\}$ as the smallest subset which contains an odd number of elements: any of the three singleton subsets $\{a\}, \{b\}, \{c\} \subset \{a, b, c\}$ are equally small, so which one should S be?

[3]By contrast, in Exercise I.1(c) you showed that the union of σ-algebras may fail to be a σ-algebra.

Proof. The uniqueness part is usual abstract nonsense. Suppose we had two different smallest σ-algebras which contain the collection $\mathscr{C}$, say $\mathscr{F}_1$ and $\mathscr{F}_2$. Then we would have $\mathscr{F}_1 \subset \mathscr{F}_2$ because $\mathscr{F}_1$ is smallest and $\mathscr{F}_2 \subset \mathscr{F}_1$ because $\mathscr{F}_2$ is smallest, so we get that $\mathscr{F}_1 = \mathscr{F}_2$.

It remains to show that a smallest σ-algebra which contains the collection $\mathscr{C}$ exists. Consider the collection $\mathcal{S}$ of all σ-algebras $\mathscr{F}$ on Ω such that $\mathscr{C} \subset \mathscr{F}$. This collection $\mathcal{S}$ is non-empty, because as in Example I.3(ii), the power set of Ω is such a σ-algebra, i.e., $\mathscr{P}(\Omega) \in \mathcal{S}$. Let us define $\sigma(\mathscr{C})$ as the intersection

$$\sigma(\mathscr{C}) := \bigcap_{\mathscr{F} \in \mathcal{S}} \mathscr{F}$$

of all these σ-algebras. By Lemma I.8, $\sigma(\mathscr{C})$ is itself a σ-algebra. Since for each $\mathscr{F} \in \mathcal{S}$ we have $\mathscr{C} \subset \mathscr{F}$, the intersection also has this property, $\mathscr{C} \subset \sigma(\mathscr{C})$. If $\mathscr{F}$ is any σ-algebra which contains $\mathscr{C}$, then clearly $\sigma(\mathscr{C})$ is smaller, since $\mathscr{F} \in \mathcal{S}$ appears in the intersection and thus $\sigma(\mathscr{C}) \subset \mathscr{F}$. Thus the intersection $\sigma(\mathscr{C})$ is smallest. □

I.3.1. Borel sigma algebra

Definition I.10 (Borel sigma algebra)**.**
For a topological space $\mathfrak{X}$, the *Borel σ-algebra* on $\mathfrak{X}$ is the σ-algebra $\mathscr{B}(\mathfrak{X})$ generated by the collection $\mathscr{T}(\mathfrak{X})$ of open sets in $\mathfrak{X}$.

Arguably the most important σ-algebra in all of probability theory is the Borel σ-algebra on the real line $\mathbb{R}$, because it is needed whenever we consider real valued random variables. We denote simply $\mathscr{B} = \mathscr{B}(\mathbb{R})$. By definition, $\mathscr{B}$ is the smallest σ-algebra on $\mathbb{R}$ which contains all opens sets $V \subset \mathbb{R}$. The following proposition establishes that $\mathscr{B}$ can alternatively be generated by various convenient collections of subsets of the real line.

Proposition I.11 (Generating the Borel sigma algebra on the real line)**.**
The Borel σ-algebra $\mathscr{B}$ on the real line $\mathbb{R}$ is generated by any of the following collections of subsets of $\mathbb{R}$:

$$\text{(i)}: \quad \mathscr{C} = \left\{ (-\infty, x] \;\middle|\; x \in \mathbb{R} \right\}, \qquad \text{(iii)}: \quad \mathscr{C} = \left\{ (x, y) \;\middle|\; x, y \in \mathbb{R},\ x < y \right\},$$

$$\text{(ii)}: \quad \mathscr{C} = \left\{ [x, y] \;\middle|\; x, y \in \mathbb{R},\ x \leq y \right\}, \quad \text{(iv)}: \quad \mathscr{C} = \left\{ (x, y] \;\middle|\; x, y \in \mathbb{R},\ x < y \right\}.$$

Remark I.12. The reader can certainly imagine further variations of generating collections of intervals, and is invited to think about the modifications needed in the proof below.

Proof of Proposition I.11. We will only explicitly check that the collection (i) generates $\mathscr{B}$, the other cases are similar.

So for the case (i), let $\mathscr{C}$ be the collection of all intervals of the form $(-\infty, x]$, with $x \in \mathbb{R}$. In order to show that $\sigma(\mathscr{C}) = \mathscr{B}$, we will separately check the two converse inclusions $\sigma(\mathscr{C}) \subset \mathscr{B}$ and $\sigma(\mathscr{C}) \supset \mathscr{B}$.

inclusion $\sigma(\mathscr{C}) \subset \mathscr{B}$: To show that $\sigma(\mathscr{C}) \subset \mathscr{B}$, it is sufficient to show that $\mathscr{B}$ contains all intervals of the form $(-\infty, x]$, because $\sigma(\mathscr{C})$ is by definition the smallest such σ-algebra.

Note that the set $(x, +\infty)$ is open, and thus is contained in the Borel σ-algebra by definition. The complement of it is $\big((x, +\infty)\big)^{\mathbf{c}} = \mathbb{R} \setminus (x, +\infty) = (-\infty, x]$. Since $\mathscr{B}$ is a σ-algebra on $\mathbb{R}$ which contains $(x, +\infty)$, by property (Σ-**c**) it contains also the complement $(-\infty, x]$. The inclusion $\sigma(\mathscr{C}) \subset \mathscr{B}$ follows.

inclusion $\sigma(\mathscr{C}) \supset \mathscr{B}$: To show that $\sigma(\mathscr{C}) \supset \mathscr{B}$, it is sufficient to show that $\sigma(\mathscr{C})$ contains all open sets $V \subset \mathbb{R}$, because $\mathscr{B}$ is by definition the smallest such σ-algebra. Let us show step by step that $\sigma(\mathscr{C})$ contains all sets of the forms

(a) semi-open intervals intervals $(x, y]$, for $x, y \in \mathbb{R}$
(b) open intervals (x, z), for $x, z \in \mathbb{R}$
(c) open sets $V \subset \mathbb{R}$.

For case (a), note that

$$(x, y] = (-\infty, y] \setminus (-\infty, x] = (-\infty, y] \cap (-\infty, x]^{\mathbf{c}},$$

so $(x, y]$ is obtained from members of the collection $\mathscr{C}$ by countable (in fact finite) intersections and complements. Therefore we have $(x, y] \in \sigma(\mathscr{C})$.

For case (b), note that

$$(x, z) = \bigcup_{n=1}^{\infty} \big(x, z - \frac{1}{n}\big],$$

so the open interval (x, z) is obtained from intervals of type (a) by a countable union. Since intervals of type (a) are already known to belong to $\sigma(\mathscr{C})$, we get that $(x, z) \in \sigma(\mathscr{C})$.

Finally, for case (c), note that any open set $V \subset \mathbb{R}$ is a countable union of open intervals — see Proposition B.5. Since open intervals are already known to belong to $\sigma(\mathscr{C})$ by case (b), we also get $V \in \sigma(\mathscr{C})$. This concludes the proof. □

The Borel σ-algebra on $\mathbb{R}$ will be needed in particular for real valued random variables. Likewise, for vector valued random variables, we will need the Borel σ-algebra $\mathscr{B}(\mathbb{R}^d)$ on the vector spaces $\mathbb{R}^d$. Recall that by definition $\mathscr{B}(\mathbb{R}^d)$ is the smallest σ-algebra on $\mathbb{R}^d$ which contains all open sets $V \subset \mathbb{R}^d$ of the d-dimensional Euclidean space. The next exercise gives a concrete and useful generating collection for this σ-algebra, in the case $d = 2$.

Exercise I.2 (Borel σ-algebra on the two-dimensional Euclidean space)**.**
Denote by

$$\mathscr{C} = \Big\{ (-\infty, x] \times (-\infty, y] \;\Big|\; x \in \mathbb{R},\ y \in \mathbb{R} \Big\}.$$

the collection of closed south-west quadrants in $\mathbb{R}^2$. Prove that the collection $\mathscr{C}$ of closed south-west quadrants generates $\mathscr{B}(\mathbb{R}^2)$, that is, show that $\mathscr{B}(\mathbb{R}^2) = \sigma(\mathscr{C})$.

Hint: *Compare with the proof of Proposition I.11. You may use the fact that every open set in $\mathbb{R}^2$ can be written as a countable union $\bigcup_{n=1}^{\infty} R_n$ of open rectangles of the form $R_n = (a_n, b_n) \times (a'_n, b'_n)$.*

Lecture II

Measures and probability measures

Recall that the basic objects of probability theory are:

Ω — the set of all possible outcomes (sample space)
$\mathscr{F}$ — the collection of all events
P — the probability (measure).

The sample space Ω can be any (non-empty) set, which we in our probabilistic modelling deem representative of the possible outcomes of the randomness involved.

In the previous lecture we explained why the collection $\mathscr{F}$ of events should be stable under countable set operations, i.e., why it must be a σ-algebra on Ω.

In this lecture we examine the last remaining basic object, P, the probability itself. We give the axiomatic properties that P is required to satisfy, and we begin studying the consequences. By the axioms, P is a special case of a mathematical object called a measure, so there is a large amount of overlap between probability theory and measure theory. We in fact choose to first develop measure theory in the general setup up to some point, because even in stochastics we make use of also other measures besides just probability measures.

II.1. Measurable spaces

In the previous lecture, we emphasized the importance of being able to perform countable set operations. This merits a definition in its own right.

Definition II.1 (Measurable space)**.**
If S is a set and $\mathscr{S}$ is a σ-algebra on S, then we call the pair $(S, \mathscr{S})$ a *measurable space*. A subset $A \subset S$ is called *measurable* if $A \in \mathscr{S}$.

Think of measurable spaces as spaces which are ready to accommodate measures. They come equipped with a good collection $\mathscr{S}$ of subsets, which behaves well under set operations as discussed in Lecture I, and a measure will assign to each of these good subsets a numerical value appropriately quantifying the size of the subset.

For convenience, let us once more unravel the definition and summarize what a measurable space is:

- S is a set
- $\mathscr{S} \subset \mathscr{P}(S)$ is a collection of subsets which satisfies
  (Σ-1): $S \in \mathscr{S}$
  (Σ-$\mathbf{c}$): if $A \in \mathscr{S}$ then $A^{\mathbf{c}} = S \setminus A \in \mathscr{S}$.
  (Σ-$\cup$): if $A_1, A_2, \ldots \in \mathscr{S}$ then $\bigcup_{n=1}^{\infty} A_n \in \mathscr{S}$.

Let us then give some examples of measurable spaces.

The following simple example is primarily relevant when S is finite or countably infinite.

Example II.2 (Measurable spaces where all subsets are measurable)**.**
If S is any set and $\mathscr{P}(S)$ is the collection of all subsets of S, then by Example I.3(ii), $\mathscr{P}(S)$ is a σ-algebra on S. Thus the pair $(S, \mathscr{P}(S))$ is a measurable space.

The following example is extremely important: in integration theory of real valued functions (or real valued random variables) one needs the set of real numbers $\mathbb{R}$ to carry the structure of a measurable space.

Example II.3 (Real line as a measurable space)**.**
Consider $S = \mathbb{R}$ and let $\mathscr{S} = \mathscr{B}$ be the Borel σ-algebra on $\mathbb{R}$ as in Section I.3.1. Then the pair $(\mathbb{R}, \mathscr{B})$ is a measurable space.

The measurable space of Example II.3 will in particular accommodate the usual measure of length on the real line $\mathbb{R}$ (cf. Example II.12 below).

Finally, the class of examples most relevant for probability theory: any pair $(\Omega, \mathscr{F})$ of a sample space Ω together with the collection of events $\mathscr{F}$ on it has to be a measurable space — ready to accommodate a probability measure in the next step!

II.2. Definition of measures and probability measures

The final basic object of probability theory is the probability measure P. In fact, even in probability theory we actually very often use also measures which are not necessarily probability measures. For example, counting measures are used when handling infinite sums, and the (intuitively) familiar measures on $\mathbb{R}$ and $\mathbb{R}^d$ are used as references when talking about densities of real valued or vector valued random variables, respectively.

Let us therefore first define measures in general.

Definition II.4 (Measure)**.**
Let $(S, \mathscr{S})$ be a measurable space. A *measure* μ on $(S, \mathscr{S})$ is a function

$$\mu\colon \mathscr{S} \to [0, +\infty]$$

such that

$$\mu\big[\emptyset\big] = 0 \tag{M-$\emptyset$}$$

and if $A_1, A_2, \ldots \in \mathscr{S}$ are disjoint, then

$$\mu\Big[\bigcup_{n=1}^{\infty} A_n\Big] = \sum_{n=1}^{\infty} \mu\big[A_n\big]. \tag{M-$\cup$}$$

A probability measure has just one further requirement added: that the total probability must be equal to one.

Definition II.5 (Probability measure)**.**
Let $(\Omega, \mathscr{F})$ be a measurable space. A *probability measure* P on $(\Omega, \mathscr{F})$ is a measure on $(\Omega, \mathscr{F})$ such that $\mathsf{P}\big[\Omega\big] = 1$.

Remark II.6 (Sure vs. almost sure)**.**
The event Ω is the *sure event*: it contains all possible outcomes. The additional requirement in the above definition merely says that the probability of the sure event is one: $\mathsf{P}[\Omega] = 1$.

It is worth noting that there may be also other events $E \in \mathscr{F}$, $E \subsetneq \Omega$, which have probability one, $\mathsf{P}[E] = 1$. We say that such an event E is *almost sure*, or alternatively we say that the event E occurs *almost surely*. The notion of *sure* only depends on the sample space Ω itself, but the notion of *almost sure* depends on the probability measure P as well, so occasionally it is appropriate to use the more specific terminology P-*almost sure* and P-*almost surely*.

Remark II.7 (Abuse of terminology)**.**
Often the σ-algebra of the underlying measurable space is clear from the context. Then, rather than saying that μ is a measure on $(S, \mathscr{S})$, we simply say that μ is a measure on S. Likewise, rather than saying that P is a probability measure on $(\Omega, \mathscr{F})$, we simply say that P is a probability measure on Ω.

If μ is a measure on $(S, \mathscr{S})$, then we call the triple $(S, \mathscr{S}, \mu)$ a *measure space*. Likewise, if P is a probability measure on $(\Omega, \mathscr{F})$, then we call the triple $(\Omega, \mathscr{F}, \mathsf{P})$ a *probability space*. In particular, any probability space is a measure space, and all results for measure spaces can be used for probability spaces. For this reason, especially in Section II.3 below, we content ourselves to stating basic properties only for measure spaces in general. There are some results which are valid for probability measures, but not for general measures. Many such results would actually hold under a milder assumption, described below.

Definition II.8 (Total mass)**.**
Let μ be a measure on $(S, \mathscr{S})$. The value $\mu[S] \in [0, +\infty]$ that the measure assigns to the whole space S is called the *total mass* of μ.

Definition II.9 (Finite measure)**.**
We say that the measure μ on $(S, \mathscr{S})$ is *finite* if its total mass is finite, $\mu[S] < +\infty$. We then also say that the corresponding measure space $(S, \mathscr{S}, \mu)$ is finite.

Probability measures, in particular, are finite measures (since $\mathsf{P}[\Omega] = 1 < +\infty$).

Let us now give a few examples of measures and probability measures.

Example II.10 (Counting measure)**.**
Let S be any set. Equip S with the σ-algebra $\mathscr{P}(S)$ consisting of all subsets of S. Then the *counting measure* on S is the measure $\mu_{\#}$, which associates to any subset $A \subset S$ the number of elements $\#A$ in the subset,

$$\mu_{\#}\big[A\big] = \#A.$$

In particular, for all infinite subsets $A \subset S$ we have $\mu_{\#}\big[A\big] = +\infty$. Property (M-$\emptyset$) holds for $\mu_{\#}$, since the empty set has no elements. Property (M-$\cup$) holds since the number of elements in a disjoint union of sets is obtained by adding up the numbers of elements in each set.

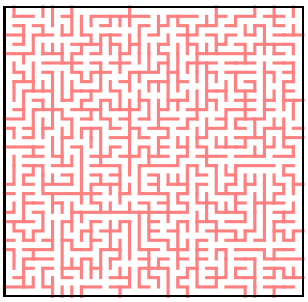
The counting measure $\mu_\#$ is a finite measure if and only if the underlying set S is a finite set.

Example II.11 (Discrete uniform probability measure)**.**
Let Ω be any finite non-empty set. Equip it with the σ-algebra $\mathscr{P}(\Omega)$ consisting of all subsets of Ω. Then the *(discrete) uniform probability measure* on Ω is the measure P_{unif} given by

$$\mathsf{P}_{\text{unif}}\big[E\big] = \frac{\#E}{\#\Omega} \qquad \text{for all } E \subset \Omega.$$

In other words, the (discrete) uniform probability measure is just the counting measure normalized to have total mass one, $\mathsf{P}_{\text{unif}} = \frac{1}{\#\Omega}\,\mu_\#$.

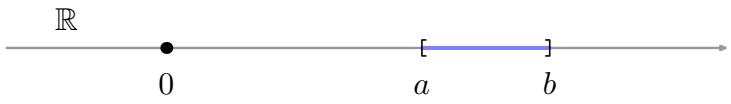
The figure on the right shows a uniform random sample from the finite set of all 30×30 labyrinths, illustrating that despite its simplicity, the discrete uniform probability measure can give rise to intricate behavior.



Example II.12 (Lebesgue measure on the real line)**.**
The natural notion of "length" on the real line $\mathbb{R}$ corresponds to the *Lebesgue measure* Λ on $(\mathbb{R}, \mathscr{B})$. For instance a closed interval $[a, b] \subset \mathbb{R}$, with $a \le b$, has measure

$$\Lambda\big[[a, b]\big] = b - a$$

equal to the length of the interval, and this property in fact is sufficient to characterize the measure Λ. The length of the entire real axis, on the other hand, is infinite: $\Lambda[\mathbb{R}] = +\infty$.

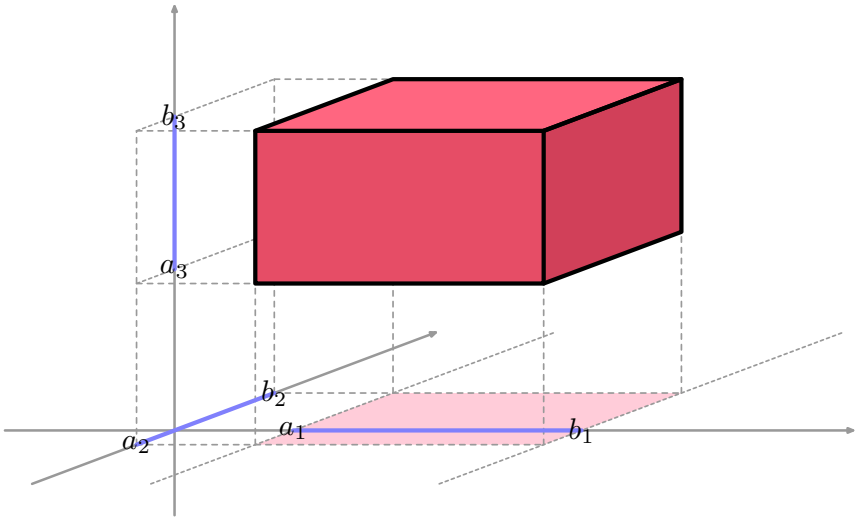


Example II.13 (Higher dimensional Lebesgue measure)**.**
Example II.12 on the one-dimensional space $\mathbb{R}$ has a d-dimensional generalization — a measure on the Euclidean space $\mathbb{R}^d$. The cases $d = 1$, $d = 2$, and $d = 3$ have the interpretation of "length on the line $\mathbb{R}$", "area in the plane $\mathbb{R}^2$", and "volume in the space $\mathbb{R}^3$", respectively.

The Euclidean space $\mathbb{R}^d$ is equipped with the Borel σ-algebra $\mathscr{B}(\mathbb{R}^d)$ (see Definition I.10). The d-dimensional Lebesgue measure Λ^d is a measure on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$, which is characterized by the property that any rectangular box $[a_1, b_1] \times \cdots \times [a_d, b_d] \subset \mathbb{R}^d$ has measure

$$\Lambda^d\Big[[a_1, b_1] \times \cdots \times [a_d, b_d]\Big] = \prod_{j=1}^{d} (b_j - a_j)$$

given by the product of the side lengths of the box.



Any course in traditional measure theory covers the particular example of Lebesgue measures Λ and Λ^d in great detail, so we choose not to elaborate on them too extensively here.

Exercise II.1 (Truncation of measures and conditioning of probability measures)**.**

(a) Let μ be a measure on $(S, \mathscr{S})$ and let $B \in \mathscr{S}$. Show that also $A \mapsto \mu[A \cap B]$ defines a measure on $(S, \mathscr{S})$.

(b) Let P be a probability measure on $(\Omega, \mathscr{F})$, and let $B \in \mathscr{F}$ be an event such that $\mathsf{P}[B] > 0$. Show that the conditional probability $A \mapsto \mathsf{P}\big[A \,\big|\, B\big] := \frac{\mathsf{P}[A \cap B]}{\mathsf{P}[B]}$ is a probability measure on $(\Omega, \mathscr{F})$.

Example II.14 (Uniform probability measure on an interval)**.**

Consider the truncation of the Lebesgue measure to the unit interval $[0, 1] \subset \mathbb{R}$. Let Λ be the Lebesgue measure on $\mathbb{R}$ as in Example II.12. Define a new measure P on $\mathbb{R}$ by truncation as in part (a) of Exercise II.1:

$$\mathsf{P}[A] = \Lambda\Big[A \cap [0, 1]\Big] \qquad \text{for all } A \in \mathscr{B}.$$

Then we have $\mathsf{P}[\mathbb{R}] = \Lambda\big[[0, 1]\big] = 1$, so P is a probability measure on $\mathbb{R}$. For subsets $A \subset [0, 1]$ of the unit interval, P coincides with the Lebesgue measure, $\mathsf{P}[A] = \Lambda[A]$. For subsets outside the unit interval, on the other hand, we have $A \cap [0, 1] = \emptyset$ and thus these sets carry no probability mass: $\mathsf{P}[A] = \Lambda[\emptyset] = 0$. We call P the *uniform probability measure* on the unit interval.

More generally, the uniform probability measure on any interval $[a, b] \subset \mathbb{R}$ of positive length is defined by the formula $A \mapsto \frac{1}{b-a}\Lambda\big[A \cap [a, b]\big]$, where we first truncate the Lebesgue measure to $[a, b]$ and then normalize the total mass by the length $b - a$ of the interval.

Before starting to examine general results about measures, we still look into one further class of examples of probability measures which is relatively easy, yet important in practice.

II.2.1. Probability distributions on countable spaces

Many probabilistic models concern distributions on the natural numbers $\mathbb{N}$, the integers $\mathbb{Z}$, or other countable sets. It turns out that on such countable spaces,

we can characterize probabilty measures in an intuitive way using probability mass functions.

For the rest of this section we therefore assume that Ω is a non-empty countable set. Then there exists an enumeration[1] $\Omega = \{\omega_1, \omega_2, \omega_3, \ldots\}$ with distinct elements $\omega_1, \omega_2, \omega_3, \ldots \in \Omega$. Summation over Ω can be defined using the enumeration,

$$\sum_{\omega\in\Omega} a(\omega) := \sum_j a(\omega_j),$$

and if the terms of the sum are non-negative, $a(\omega) \geq 0$, then the result of the sum is independent of the chosen enumeration.

Definition II.15 (Probability mass function)**.**
A *probability mass function* (p.m.f.) on Ω is a function

$$p\colon \Omega \to [0,1]$$

such that

$$\sum_{\omega\in\Omega} p(\omega) = 1.$$

To a probability mass function p it is natural to associate a measure defined by

$$\mathsf{P}[E] = \sum_{\omega\in E} p(\omega) \qquad \text{for all } E \subset \Omega, \tag{II.1}$$

and conversely to a probability measure P on $(\Omega, \mathscr{P}(\Omega))$ it is natural to associate masses of singleton events $\{\omega\} \subset \Omega$

$$p(\omega) = \mathsf{P}\big[\,\{\omega\}\,\big] \qquad \text{for all } \omega \in \Omega. \tag{II.2}$$

The following exercise shows that on countable spaces, probability mass functions are in one-to-one correspondence with probability measures via the above formulas.

Exercise II.2 (Probability distributions on countable spaces)**.**
Let Ω be a finite or a countably infinite set, and denote by $\mathscr{P}(\Omega)$ the collection of all subsets of Ω.

(a) Show that if p is a probability mass function on Ω, then the set function P defined by (II.1) is a probability measure on $(\Omega, \mathscr{P}(\Omega))$.
(b) Show that if P is a probability measure on $(\Omega, \mathscr{P}(\Omega))$, then the function p defined by (II.2) is a probability mass function on Ω.

Example II.16 (Poisson distribution)**.**
Let $\lambda > 0$. Recalling the power series $\sum_{k=0}^{\infty} \frac{1}{k!}\lambda^k = e^\lambda$ of the exponential function, it is easy to see that the function p given by

$$p(k) = e^{-\lambda}\,\frac{\lambda^k}{k!} \qquad \text{for } k \in \mathbb{Z}_{\geq 0} = \{0, 1, 2, \ldots\} \tag{II.3}$$

is a probability mass function on $\mathbb{Z}_{\geq 0}$.

The Poisson distribution with parameter λ is the probability measure on $\mathbb{Z}_{\geq 0}$ with the above probability mass function.

[1] If Ω is finite, the enumeration terminates, $\Omega = \{\omega_1, \omega_2, \ldots, \omega_n\}$. The more interesting case is if Ω is countably infinite.

Example II.17 (Geometric distribution)**.**
Let $q \in (0,1)$. Using the geometric series $\sum_{k=0}^{\infty} r^k = \frac{1}{1-r}$ with $r = 1-q$, it is easy to see that the function p given by

$$p(k) = (1-q)^{k-1} q \qquad \text{for } k \in \mathbb{N} = \{1, 2, 3, \ldots\} \tag{II.4}$$

is a probability mass function on $\mathbb{N}$.

The geometric distribution with parameter q is the probability measure on $\mathbb{N}$ with the above probability mass function.

Example II.18 (Binomial distribution)**.**
Let $n \in \mathbb{N}$ and $q \in (0,1)$. Using the binomial formula $\sum_{k=0}^{n} \binom{n}{k} a^j b^{n-j} = (a+b)^n$ with $a = q$ and $b = 1-q$ it is easy to see that the function p given by

$$p(k) = \binom{n}{k} q^k (1-q)^{n-k} \qquad \text{for } k \in \{0, 1, \ldots, n-1, n\} \tag{II.5}$$

is a probability mass function on the finite set $\{0, 1, \ldots, n-1, n\}$.

The binomial distribution with parameters n and q is the probability measure on the finite set $\{0, 1, \ldots, n-1, n\}$ with the above probability mass function.

II.3. Properties of measures and probability measures

Let us now discuss some of the first properties of measures and probability measures.

Subadditivity of measures and the union bound

For repeated later use, we start by proving the following additivity properties for measures of disjoint sets, subadditivity properties of measures of (not necessarily disjoint) sets, as well as related monotonicity and monotone convergence properties of measures.

Lemma II.19 (First properties of measures)**.**
Let μ be a measure on a measurable space $(S, \mathscr{S})$. Then we have the following:

(a) Finite additivity: *If $A_1, \ldots, A_n \in \mathscr{S}$ are disjoint measurable sets, then we have:*

$$\mu\Big[A_1 \cup \cdots \cup A_n\Big] \;=\; \mu[A_1] + \cdots + \mu[A_n]. \tag{II.6}$$

(b) Monotonicity: *If $A, B \in \mathscr{S}$ and $A \subset B$, then we have:*

$$\mu[A] \;\leq\; \mu[B]. \tag{II.7}$$

(c) Finite subadditivity: *If $A_1, \ldots, A_n \in \mathscr{S}$ are any measurable sets, then we have:*

$$\mu\Big[A_1 \cup \cdots \cup A_n\Big] \;\leq\; \mu[A_1] + \cdots + \mu[A_n]. \tag{II.8}$$

(d) Monotone convergence of measures: *Let $A_1 \subset A_2 \subset \cdots$ be an increasing sequence of measurable sets, $A_n \in \mathscr{S}$ for all $n \in \mathbb{N}$. Then the measures of the increasing limit $A_n \uparrow A = \bigcup_{j=1}^{\infty} A_j$ of sets constitute the increasing limit*

$$\mu[A_n] \;\uparrow\; \mu[A]. \tag{II.9}$$

(e) Countable subadditivity: *If $A_1, A_2, \ldots \in \mathscr{S}$ is a sequence of measurable sets (not necessarily disjoint), then we have:*

$$\mu\Big[\bigcup_{j=1}^{\infty} A_j\Big] \;\leq\; \sum_{j=1}^{\infty} \mu[A_j]. \tag{II.10}$$

Proof of (a): Given the disjoint measurable sets $A_1, \ldots, A_n \in \mathscr{S}$, let us extend this finite sequence by empty sets: define $A_{n+1} = A_{n+2} = \cdots = \emptyset$. Since $\emptyset \in \mathscr{S}$ by properties of σ-algebras, we thus obtain an infinite sequence $A_1, A_2, \ldots \in \mathscr{S}$ of measurable sets. This sequence of sets is disjoint (the newly added empty sets do not have common elements with the already disjoint $A_1, \ldots, A_n$). Thus from axiom (M-$\cup$) it follows that

$$\mu\Big[\bigcup_{j=1}^{\infty} A_j\Big] = \sum_{j=1}^{\infty} \mu\big[A_j\big].$$

But on the left hand side, the union is simply $\bigcup_{j=1}^{\infty} A_j = A_1 \cup \cdots \cup A_n$, because the empty sets do not contribute to the union. On the right hand side, the sum is $\sum_{j=1}^{\infty} \mu[A_j] = \mu[A_1] + \cdots + \mu[A_n]$, because the measures of the empty sets $\mu[\emptyset] = 0$ do not contribute to the sum. Assertion (a) follows.

Proof of (b): Assume that $A, B \in \mathscr{S}$. Note that then also $B \setminus A = B \cap A^{\mathbf{c}} \in \mathscr{S}$ by properties of σ-algebras. If $A \subset B$, then $B = A \cup (B \setminus A)$ is a disjoint union. For these two disjoint sets, we can use part (a) to get

$$\mu[B] = \mu\Big[A \cup (B \setminus A)\Big] \overset{\text{(a)}}{=} \mu[A] + \mu\big[B \setminus A\big].$$

Since $\mu\big[B \setminus A\big] \geq 0$ by properties of measures, the assertion $\mu[B] \geq \mu[A]$ follows.

Proof of (c): We will prove the inequality

$$\mu\Big[A_1 \cup \cdots \cup A_n\Big] \;\leq\; \mu[A_1] + \cdots + \mu[A_n].$$

for all $A_1, \ldots, A_n \in \mathscr{S}$ by induction on the number n of sets in the union.

The case $n = 1$ is clear — the two sides of the inequality are in fact equal. Now assume the inequality for unions of n sets, and consider $A_1, \ldots, A_{n+1} \in \mathscr{S}$. Define $A = A_1 \cup \cdots \cup A_n$ and $B = A_{n+1} \setminus A$. Then we have $A_1 \cup \cdots \cup A_{n+1} = A \cup B$, where the sets A and B are disjoint. Thus by part (a) we get

$$\mu\big[A \cup B\big] = \mu[A] + \mu[B].$$

The first term on the right hand side is

$$\mu[A] = \mu\big[A_1 \cup \cdots \cup A_n\big] \leq \mu[A_1] + \cdots + \mu[A_n]$$

by induction assumption. The second term on the right hand side is $\mu[B] \leq \mu[A_{n+1}]$ by monotonicity proven in part (b), since $B \subset A_{n+1}$. Notice that the right hand side $\mu\big[A \cup B\big] = \mu\big[A_1 \cup \cdots \cup A_{n+1}\big]$ is the measure we are interested in. Therefore, by combining the observations, we conclude

$$\mu\Big[A_1 \cup \cdots \cup A_{n+1}\Big] \;\leq\; \Big(\mu[A_1] + \cdots + \mu[A_n]\Big) + \mu[A_{n+1}],$$

which finishes the proof of assertion (c) by induction.

Proof of (d): Suppose that $A_1 \subset A_2 \subset \cdots$ is an increasing sequence of measurable sets, and denote its limit by $A = \bigcup_{j=1}^{\infty} A_j$. Then A is also measurable by properties of σ-algebras. Now write first $B_1 = A_1$, and then $B_2 = A_2 \setminus A_1$, …, $B_n = A_n \setminus A_{n-1}$, …. These sets $B_1, B_2, \ldots$ are disjoint and $A_n = B_1 \cup \cdots \cup B_n$ for all $n \in \mathbb{N}$. From part (a) we get

$$\mu[A_n] = \mu\big[B_1 \cup \cdots \cup B_n\big] = \sum_{j=1}^{n} \mu[B_j].$$

The right hand sides are the partial sums of an infinite sum with non-negative terms, so they form a sequence increasing to that infinite sum, and we conclude

$$\mu[A_n] \uparrow \sum_{j=1}^{\infty} \mu[B_j] \qquad \text{as } n \to \infty.$$

On the other hand, by disjointness of $B_1, B_2, \ldots$ and axiom (M-$\cup$), this infinite sum equals

$$\sum_{j=1}^{\infty} \mu[B_j] = \mu\Big[\bigcup_{j=1}^{\infty} B_j\Big] = \mu\Big[\bigcup_{j=1}^{\infty} A_j\Big] = \mu[A].$$

This proves the assertion (d), $\mu[A_n] \uparrow \mu[A]$ as $n \to \infty$.

Proof of (e): Let $A_1, A_2, \ldots \in \mathscr{S}$ be a sequence of measurable sets. Form their finite unions $C_n = A_1 \cup \cdots \cup A_n$, for all $n \in \mathbb{N}$. Then, as a union of measurable sets, each C_n is also measurable. This sequence is clearly increasing $C_1 \subset C_2 \subset \cdots$, and its limit is the countably infinite union $C = \bigcup_{j=1}^{\infty} A_j$. We can therefore apply part (d) to get

$$\mu[C_n] \uparrow \mu[C] \qquad \text{as } n \to \infty. \tag{II.11}$$

On the other hand, by part (c) we have for all $n \in \mathbb{N}$

$$\mu[C_n] = \mu\big[A_1 \cup \cdots \cup A_n\big] \le \sum_{j=1}^{n} \mu[A_j] \le \sum_{j=1}^{\infty} \mu[A_j].$$

If we denote the value of the infinite sum by $c := \sum_{j=1}^{\infty} \mu[A_j]$, then this bound $\mu[C_n] \le c$ for all n implies that for the limit (II.11) we also have $\mu[C] \le c$. Recalling what C and c are, we have now obtained

$$\mu\Big[\bigcup_{j=1}^{\infty} A_j\Big] = \mu[C] \le c = \sum_{j=1}^{\infty} \mu[A_j].$$

This finishes the proof. □

Especially part (e) of the above lemma is, despite its simplicity, so useful in probability theory that it has been given an affectionate name: "the union bound". Because of its importance, we record this fact once more in the probabilistic context.

Theorem II.20 (The union bound)**.**
Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space and let $E_1, E_2, \ldots \in \mathscr{F}$ be a sequence of events. Then we have

$$\mathsf{P}\Big[\bigcup_{j=1}^{\infty} E_j\Big] \le \sum_{j=1}^{\infty} \mathsf{P}[E_j]. \tag{II.12}$$

In other words, the probability that at least one event in a sequence occurs can not exceed the sum of the probabilities of the events in the sequence.

Probability measures enjoy some properties that may not be valid for measures of infinite total mass. The following exercise gives a few of them.

Exercise II.3 (Properties specific to probability measures)**.**
Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space.

(a) Show that for any event $E \in \mathscr{F}$ we have

$$\mathsf{P}[E^{\mathbf{c}}] = 1 - \mathsf{P}[E].$$

(b) Show that for any two events $E_1, E_2 \in \mathscr{F}$ we have

$$\mathsf{P}[E_1 \cup E_2] = \mathsf{P}[E_1] + \mathsf{P}[E_2] - \mathsf{P}[E_1 \cap E_2].$$

Monotone convergence of probability measures

Part (d) of Lemma II.19 is a monotone convergence statement of measures for increasing sequences of sets. For general measures we do not have the corresponding monotone convergence for decreasing sequences of sets, as the following counterexample shows.

Example II.21. (Decreasing monotone convergence of measures can fail in general)
Consider the set $\mathbb{N} = \{1, 2, 3, \ldots\}$ of natural numbers with the counting measure $\mu_\#$, as defined in Example II.10:

$$\mu_\#[A] = \#A \qquad \text{for all } A \subset \mathbb{N}.$$

Consider the subsets $A_n = \{n, n+1, n+2, \ldots\} \subset \mathbb{N}$. Each of these is an infinite set, so their counting measures are infinite, $\mu_\#[A_n] = +\infty$. These sets form a decreasing sequence,

$$A_1 \supset A_2 \supset A_3 \supset \cdots,$$

and the limit is the intersection

$$A = \bigcap_{n\in\mathbb{N}} A_n.$$

But no natural number $m \in \mathbb{N}$ belongs to all of A_n, $n \in \mathbb{N}$, (indeed, $m \notin A_n$ as soon as $n > m$). Therefore the intersection is empty, $A = \emptyset$. The number of elements in this empty set is zero, $\#A = 0$. In particular, the sequence of counting measures $\mu_\#[A_n] = +\infty$ does *not* tend to the counting measure $\mu_\#[A] = 0$ of the decreasing limit set A.

Exercise II.4. Construct a similar counterexample with the Lebesgue measure Λ on $\mathbb{R}$.

For probability measures, however, monotone convergence of measures holds for both increasing and decreasing sequences of events.

Theorem II.22 (Monotone convergence of probability measures)**.**
Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space.

(a) *If $E_1 \subset E_2 \subset \cdots$ is an increasing sequence of events with limit $E = \bigcup_{n\in\mathbb{N}} E_n$, then we have $\mathsf{P}[E_n] \uparrow \mathsf{P}[E]$ as $n \to \infty$.*
(b) *If $E_1 \supset E_2 \supset \cdots$ is a decreasing sequence of events with limit $E = \bigcap_{n\in\mathbb{N}} E_n$, then we have $\mathsf{P}[E_n] \downarrow \mathsf{P}[E]$ as $n \to \infty$.*

Proof. Part (a) follows from part (d) of Lemma II.19, since any probability measure is a measure. Part (b) is left as an exercise. □

Exercise II.5. Prove part (b) of Theorem II.22 above.

II.4. Identification and construction of measures

We now turn to the following questions:

- Does a measure with some desired properties exist?
  (How can we construct measures?)
- How can we check whether two measures are the same?
  (What does one need to know to uniquely identify a measure?)

As we noted in Section I.3, it can be complicated to work with σ-algebras. For the purposes of the above two questions, in particular, we often prefer to work with simpler collections. For the identification part, the appropriate simpler collections are called π-systems (Definition II.23 below)

Courses on traditional measure theory focus quite a lot on the construction of measures. We refer the interested reader to such measure theory courses for details — the dedicated reader will find for example Carathéodory's extension theorem, with which it is possible to construct, e.g., the Lebesgue measure Λ on $\mathbb{R}$ (Example II.12) and its multi-dimensional analogue Λ^d on $\mathbb{R}^d$ (Example II.13).

Instead, identification of probability measures is of practical relevance in stochastics and statistics, so we focus more on this latter question. The proof of the main identification result (Dynkin's identification theorem, Theorem II.26), is not given immediately. It can be found in Appendix C, and may be most natural to study together with the topics of Lecture IV.

Identification of probability measures

Collections of the following type are good for the purposes of identification of measures.

Definition II.23 (Pi-system)**.**
A collection $\mathscr{J}$ of subsets of S is called a *π-system* if the following holds:

$$(\Pi\text{-}\cap): \qquad \text{if } A, B \in \mathscr{J}, \text{ then also } A \cap B \in \mathscr{J}.$$

Remark II.24. Any σ-algebra $\mathscr{S}$ is also a π-system (since the intersection of two sets in a σ-algebra belongs to the σ-algebra). However, not every π-system is a σ-algebra, as the following example shows.

Example II.25 (A pi-system of semi-infinite intervals)**.**
Consider the collection

$$\mathscr{J}(\mathbb{R}) := \Big\{(-\infty, x] \;\Big|\; x \in \mathbb{R}\Big\} \tag{II.13}$$

of semi-infinite intervals $(-\infty, x] \subset \mathbb{R}$. Then $\mathscr{J}(\mathbb{R})$ is a π-system on $\mathbb{R}$: indeed it is clearly a non-empty collection of subsets of $\mathbb{R}$, and given any two intervals from the collection, $(-\infty, x_1]$ and $(-\infty, x_2]$, their intersection is the interval

$$(-\infty, x_1] \,\cap\, (-\infty, x_2] \;=\; (-\infty, z], \qquad \text{where } z = \min\{x_1, x_2\}$$

which itself belongs to the collection $\mathscr{J}(\mathbb{R})$. Thus property $(\Pi\text{-}\cap)$ holds for $\mathscr{J}(\mathbb{R})$.

In Proposition I.11 we saw that $\mathscr{J}(\mathbb{R})$ generates the Borel σ-algebra $\mathscr{B}$ on $\mathbb{R}$, i.e., $\sigma(\mathscr{J}(\mathbb{R})) = \mathscr{B}$. It is one of the simplest such π-systems, and for this reason we will use $\mathscr{J}(\mathbb{R})$ over and over again, especially when dealing with real valued random variables.

The main result which is used for identification of measures is the following.

Theorem II.26 (Dynkin's identification theorem)**.**
Let P_1 and P_2 be two probability measures on a measurable space $(\Omega, \mathscr{F})$. Assume that $\mathscr{J}$ is a π-system on Ω such that the σ-algebra $\sigma(\mathscr{J})$ generated by it coincides with the σ-algebra $\mathscr{F}$ of measurable sets in the measurable space, i.e., $\sigma(\mathscr{J}) = \mathscr{F}$. Then the following are equivalent:

(i) $\mathsf{P}_1[E] = \mathsf{P}_2[E]$ *for all* $E \in \mathscr{J}$
(ii) *the two probability measures are equal,* $\mathsf{P}_1 = \mathsf{P}_2$.

Remark II.27. Condition (ii) above is clearly stronger than condition (i). Namely, the equality of probability measures $\mathsf{P}_1 = \mathsf{P}_2$ means that $\mathsf{P}_1[E] = \mathsf{P}_2[E]$ for all $E \in \mathscr{F}$, and there are in general more sets E in this σ-algebra $\mathscr{F}$ than in the π-system $\mathscr{J}$. The nontrivial part of the proof is therefore that condition (i) implies (ii). This will be proven in Appendix C.3.

Cumulative distribution function

In the following we consider a probability measure on the real axis $\mathbb{R}$. Typically such a probability measure could appear as the law of a real-valued random variable, as we will discuss later on in the course. In order to have a suitable notation for such common situations, let us denote the probability measure in this case by ν instead of P.

Definition II.28 (Cumulative distribution function)**.**
If ν is a probability measure on $(\mathbb{R}, \mathscr{B})$, then the *cumulative distribution function* (c.d.f.) of ν is the function $F\colon \mathbb{R} \to [0,1]$ defined by

$$F(x) := \nu\Big[(-\infty, x]\Big].$$

A simple but important application of Dynkin's identification theorem is the following. This case is applicable, e.g., to the identification of the laws of real-valued random variables.

Corollary II.29 (Cumulative distribution function identify distributions)**.**
Let ν_1 *and* ν_2 *be two probability measures on* $(\mathbb{R}, \mathscr{B})$*, and* F_1 *and* F_2 *their cumulative distribution functions, reprectively. Then the following are equivalent:*

(i) *The cumulative distribution functions are equal,* $F_1 = F_2$.
(ii) *The probability measures are equal,* $\nu_1 = \nu_2$.

Proof: Equivalence if proved by establishing both implications (i) $\Rightarrow$ (ii) and (ii) $\Rightarrow$ (i).

proof of (ii) $\Rightarrow$ *(i):* Assuming the probability measures are equal, $\nu_1 = \nu_2$, we get for any $x \in \mathbb{R}$

$$F_1(x) := \nu_1\big[(-\infty, x]\big] = \nu_2\big[(-\infty, x]\big] =: F_2(x).$$

proof of (i) $\Rightarrow$ *(ii):* Assume the equality $F_1 = F_2$ of cumulative distribution functions, i.e., that $F_1(x) = F_2(x)$ for all $x \in \mathbb{R}$. Consider the π-system $\mathscr{J}(\mathbb{R})$ of Example II.25. A set $A = \mathscr{J}(\mathbb{R})$ of this π-system is by definition of the form $A = (-\infty, x]$ for some $x \in \mathbb{R}$. For such a set, we get

$$\nu_1\big[(-\infty, x]\big] =: F_1(x) = F_2(x) := \nu_1\big[(-\infty, x]\big],$$

so we have that ν_1 and ν_2 coincide on $\mathscr{J}(\mathbb{R})$. The σ-algebra $\sigma\big(\mathscr{J}(\mathbb{R})\big)$ generated by the π-system $\mathscr{J}(\mathbb{R})$ coincides with the Borel σ-algebra $\mathscr{B}$ by Proposition I.11(i). Theorem II.26 then guarantees that ν_1 and ν_2 coincide on the entire Borel σ-algebra $\mathscr{B}$. □

Since cumulative distribution functions characterize probability measures on $(\mathbb{R}, \mathscr{B})$ by Corollary II.29 above, it is natural to next ask which functions F can qualify as cumulative distribution functions. There is indeed a rather explicit characterization.

Deriving the necessary conditions below is an instructive application of the basic properties of measures which we established in Lemma II.19.

Proposition II.30 (Properties of cumulative distribution functions)**.** *If $F\colon \mathbb{R} \to [0,1]$ is the cumulative distribution function of a probability measure ν on $(\mathbb{R}, \mathscr{B})$, then it satisfies the following properties:*

(a) *F is increasing: if $x \leq y$ then $F(x) \leq F(y)$*
(b) *F is right-continuous: if $x_n \downarrow x \in \mathbb{R}$ as $n \to \infty$, then $F(x_n) \downarrow F(x)$*
(c) *$\lim_{x\to+\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$.*

Exercise II.6. Prove Proposition II.30.
Hint: *Use appropriate parts of Lemma II.19.*

Lecture III

Random variables

Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space, i.e.,

Ω — the set of all possible outcomes
$\mathscr{F}$ — the collection of events (see Lecture I)
P — the probability measure (see Lecture II).

The key idea of a random variable is the following two step procedure by which randomness is thought to have effect:

1.) "Chance determines the random outcome $\omega \in \Omega$."
2.) "The outcome ω determines various quantities of interest." (random variables)

Therefore, a random variable will be a function X defined on Ω, which to an outcome $\omega \in \Omega$ associates the value $X(\omega)$ of some quantity of interest. The function

$$X \colon \Omega \to S'$$

takes values in a suitable set S' of possible values of the quantity of interest (some examples are given below, in Example III.4). Crucially, this function has to be sufficiently well-behaved so that we can talk about probabilities with which the quantity assumes certain values. So whenever $A' \subset S'$ is a reasonable enough subset of the possible values, the set of outcomes ω for which $X(\omega)$ belongs to A' should constitute an event, i.e.,

$$\big\{\omega \in \Omega \;\big|\; X(\omega) \in A'\big\} \in \mathscr{F}. \tag{III.1}$$

This requirement of well-behavedness of the function $X \colon \Omega \to S'$ is what the notion of *measurable function* captures (cf. Definition III.1 below).

The set in (III.1) above is just the preimage of A' under the function $X \colon \Omega \to S'$: indeed by definition we have

$$X^{-1}(A') = \big\{\omega \in \Omega \;\big|\; X(\omega) \in A'\big\} \subset \Omega.$$

For simplicity, we will often abbreviate this just as

$$\{X \in A'\} \subset \Omega.$$

This last slight abuse of notation is not only shorter, but it also has the advantage that the probabilistic interpretation

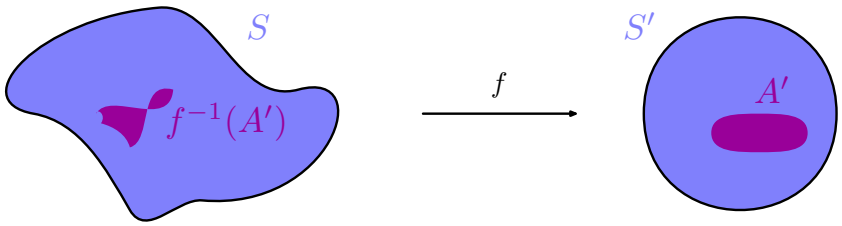
"the value of our (random) quantity of interest X lies in A'"

of the event becomes apparent at a glance.

III.1. Measurable functions and random variables

Let $(S, \mathscr{S})$ and $(S', \mathscr{S}')$ be two measurable spaces, i.e., S and S' are two sets and $\mathscr{S}$ and $\mathscr{S}'$ are σ-algebras on these two respectively.

Definition III.1 (Measurable function)**.**
A function $f\colon S \to S'$ is called *$\mathscr{S}/\mathscr{S}'$-measurable*[1] if for all $A' \in \mathscr{S}'$ we have $f^{-1}(A') \in \mathscr{S}$.



Let now $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space, and $(S', \mathscr{S}')$ a measurable space.

Definition III.2 (Random variable)**.**
A *random variable* with values in S' is a $\mathscr{F}/\mathscr{S}'$-measurable function $X\colon \Omega \to S'$.

Remark III.3. This definition precisely requires that $\{X \in A'\} \subset \Omega$ is an event whenever the subset $A' \subset S'$ is $\mathscr{S}'$-measurable.

Example III.4 (Examples of types of random variables)**.**
Depending on our quantity of interest, the set S' of allowed values of the random variable can be for example one of the following:

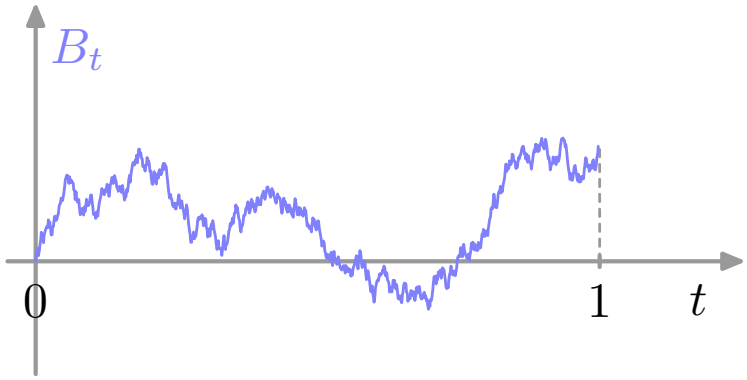
random numbers: The case $S' = \mathbb{R}$, $\mathscr{S}' = \mathscr{B}$ (Borel σ-algebra on the real line) corresponds to real-valued random variables, i.e., *random numbers*.
random vectors: e.g., $S' = \mathbb{R}^d$, $\mathscr{S}' = \mathscr{B}(\mathbb{R}^d)$ (Borel sigma-algebra on $\mathbb{R}^d$)
random matrices: e.g., $S' = \mathbb{R}^{m\times n}$, $\mathscr{S}' = \mathscr{B}(\mathbb{R}^{m\times n})$ (Borel sigma-algebra on $\mathbb{R}^{m\times n}$)
random graphs: S' some set of graphs, $\mathscr{S}'$ a suitably chosen σ-algebra (often simply the power set $\mathscr{P}(S')$)
etc.: ...

Usually S and S' are at least topological spaces, so we can and will equip them with their Borel σ-algebras $\mathscr{S} = \mathscr{B}(S)$ and $\mathscr{S}' = \mathscr{B}(S')$, generated by their open subsets (see Definition I.10). A $\mathscr{B}(S)/\mathscr{B}(S')$-measurable function $f\colon S \to S'$ will be called a *Borel-measurable function* or simply a *Borel function*. Let us give one interesting example of a random variable with values in a topological space which is not as simple as the finite dimensional Euclidean spaces in the previous example.

Example III.5 (Brownian motion as a random variable)**.**
Brownian motion is one of the most important stochastic processes: it is a continuous real valued Markov process in continuous time, which is used to model many things from thermal motion of microscopic particles to stock prices in finance. Mathematically, the Brownian

[1] When the two σ-algebras are clear from the context, we usually just say that the function is *measurable*.

motion on the unit time interval is a certain random variable taking values in the space $S' = \mathcal{C}([0,1])$ of continuous functions $h\colon [0,1] \to \mathbb{R}$, with the topology induced by the uniform norm $\|h\|_\infty = \sup_{t\in[0,1]} |h(t)|$ and the corresponding Borel σ-algebra $\mathscr{B}\big(\mathcal{C}([0,1])\big)$.



The law of a random variable

Suppose that $X\colon \Omega \to S'$ is a random variable. Then there is a probability measure on S' which describes how the values of the random variable are distributed.

Definition III.6 (The law of a random variable)**.**
The *law* (or the *distribution*) of the random variable $X\colon \Omega \to S'$ is the probability measure P_X on $(S', \mathscr{S}')$ defined by

$$P_X[A'] = \mathsf{P}\big[X^{-1}(A')\big], \qquad \text{for } A' \in \mathscr{S}'. \tag{III.2}$$

Exercise III.1. Verify that P_X given by (III.2) is indeed a probability measure on $(S', \mathscr{S}')$.

With a slight abuse of notation, we usually write the defining equation (III.2) of the law of X in the more descriptive form

$$P_X[A'] = \mathsf{P}\big[X \in A'\big].$$

If we were to insist on carefully following the notation that was introduced in Lecture II, then instead of "$X \in A'$" we should in principle write "$\big\{\omega \in \Omega \;\big|\; X(\omega) \in A'\big\}$" (this subset of the sample space is the event whose probability concerns us). But it is clear that doing so would become really cumbersome in actual practice, so shorthand notations of the above kind are commonplace in probabilistic literature.

III.2. Indicator random variables

Constant functions provide trivial examples of random variables (verify the measurability directly from the definition!). They in fact correctly model situations when a quantity of interest contains no randomness — such random variables are usually called "deterministic".

Constant functions have only one possible value (the constant in question). Arguably the next simplest example of a random variable would be one which assumes one of two possible values (depending on the random outcome). It is convenient to take 0

and 1 as those two values, in which case one speaks of an *indicator random variable*. For $E \subset \Omega$ a subset, we define the *indicator function* $\mathbb{I}_E \colon \Omega \to \mathbb{R}$ of E by

$$\mathbb{I}_E(\omega) = \begin{cases} 1 & \text{if } \omega \in E \\ 0 & \text{if } \omega \notin E. \end{cases} \tag{III.3}$$

Exercise III.2 (Measurability of indicators)**.**
Prove that $\mathbb{I}_E$ is $\mathscr{F}$-measurable if and only if $E \in \mathscr{F}$, i.e., if $E \subset \Omega$ is an event.

When $\mathbb{I}_E$ is $\mathscr{F}$-measurable, we call it the *indicator random variable* of the event E. It "indicates" the occurrence of the event E in the sense that it takes the value 1 if the event E occurs and value 0 otherwise.

Exercise III.3 (Indicators of intersections and unions)**.**
Let $A, B \subset \Omega$.

(a) Show that $\mathbb{I}_{A\cap B} = \mathbb{I}_A \mathbb{I}_B$.
(b) When is it true that $\mathbb{I}_{A\cup B} = \mathbb{I}_A + \mathbb{I}_B$?

III.3. Constructing random variables

So far we have given the definition of random variables, but we have hardly addressed the issue of constructing them — we have only left it as an exercise to characterize the measurability of random variables with values 0 and 1. At this stage one might therefore still worry that perhaps the requirements of a measurable function are too stringent for any interesting examples to exist... Fortunately, this is not the case: almost all functions that you ever encounter turn out to be measurable. The rest of this lecture is devoted to understanding why.

A very easy case

The example here concerns one case in which we do not have to worry about the existence of random variables at all — when all subsets of the sample space Ω are events, then every function on Ω is measurable.

Example III.7 (A case when all functions are measurable)**.**
Suppose that $(\Omega, \mathscr{F}, \mathsf{P})$ is a probability space in which all subsets of Ω are events $\mathscr{F} = \mathscr{P}(\Omega)$. Then any function $X \colon \Omega \to \mathbb{R}$ is a random variable. Indeed, for any $B \in \mathscr{B}$, the preimage $X^{-1}(B) \subset \Omega$ is a subset, and therefore $X^{-1}(B) \in \mathscr{P}(\Omega)$.

As we have mentioned before, we can usually only take $\mathscr{F} = \mathscr{P}(\Omega)$ when Ω is countable, so Example III.7 only reassures us of the existence of plenty of random variables on countable sample spaces. It remains to convince ourselves that interesting random variables exist in other common situations.

Composition of measurable functions

Most of the time in practice, we construct more complicated random variables from simpler ones. The measurability of the simpler building blocks should ideally directly imply the measurability of the more complicated construction.

Composition of functions is one such construction.

Proposition III.8 (Composition of measurable functions is measurable)**.** *Suppose that $(S, \mathscr{S})$, $(S', \mathscr{S}')$, and $(S'', \mathscr{S}'')$ are three measurable spaces and that we are given two measurable functions:*

$$\begin{aligned} f\colon\ & S \to S' && \text{is } \mathscr{S}/\mathscr{S}'\text{-measurable} \\ g\colon\ & S' \to S'' && \text{is } \mathscr{S}'/\mathscr{S}''\text{-measurable.} \end{aligned}$$

Then the composite function $s \mapsto g(f(s))$ is also measurable:

$$g \circ f\colon\ S \to S'' \qquad \text{is } \mathscr{S}/\mathscr{S}''\text{-measurable.}$$

Proof. We check the measurability of $g \circ f$ directly using the definition. Let $A'' \in \mathscr{S}''$ be a measurable subset of S''. Then first, by the $\mathscr{S}'/\mathscr{S}''$-measurability of g, the preimage under g of A'' is measurable: we have $g^{-1}(A'') \in \mathscr{S}'$. Using the $\mathscr{S}/\mathscr{S}'$-measurability of f, it then follows that also the preimage under f of the measurable set $g^{-1}(A'')$ is measurable: we have $f^{-1}\big(g^{-1}(A'')\big) \in \mathscr{S}$. It is an easy fact about preimages of composite functions that $f^{-1}\big(g^{-1}(A'')\big) = (g \circ f)^{-1}(A'')$. Therefore we have checked the measurability of the preimage $(g \circ f)^{-1}(A'')$ of any measurable subset $A'' \subset S''$ under the composite function $g \circ f\colon S \to S''$, which by definition says that the composite function $g \circ f$ is measurable. □

We postpone examples until after we have some more tools at our disposal.

Practical verification of measurability

To verify measurability, it is in fact enough to check the defining condition for just some collection of subsets that generates the σ-algebra on the target space.

Lemma III.9 (A sufficient condition for measurability)**.** *Let $\mathscr{C}' \subset \mathscr{P}(S')$ be a collection of subsets of S' that generates the σ-algebra $\mathscr{S}'$, i.e., $\sigma(\mathscr{C}') = \mathscr{S}'$. Then a function $f\colon S \to S'$ is $\mathscr{S}/\mathscr{S}'$-measurable if and only if $f^{-1}(C') \in \mathscr{S}$ for all $C' \in \mathscr{C}'$.*

Proof. The condition is clearly necessary for measurability of f: if $C' \in \mathscr{C}' \subset \sigma(\mathscr{C}') = \mathscr{S}'$ then the definition of measurability requires that $f^{-1}(C') \in \mathscr{S}$. It therefore remains only to prove that the condition is also sufficient.

Assume now that $f^{-1}(C') \in \mathscr{S}$ for all $C' \in \mathscr{C}'$. We must prove that then the function f is $\mathscr{S}/\mathscr{S}'$-measurable. Define

$$\mathscr{G}' = \Big\{ G' \in \mathscr{S}' \;\Big|\; f^{-1}(G') \in \mathscr{S} \Big\}$$

as the collection of the "good" subsets G' of S', whose preimages are measurable. By assumption we have $\mathscr{C}' \subset \mathscr{G}'$. Now since preimages satisfy the properties

$$f^{-1}(S') = S$$
$$f^{-1}\big((G')^{\mathbf{c}}\big) = \big(f^{-1}(G')\big)^{\mathbf{c}}$$
$$f^{-1}\Big(\bigcup_{n=1}^{\infty} G'_n\Big) = \bigcup_{n=1}^{\infty} f^{-1}(G'_n)$$

(see Exercise A.2) and $\mathscr{S}$ is a σ-algebra (on S), we see that the collection $\mathscr{G}'$ of subsets of S' is a σ-algebra (on S').

The fact that the σ-algebra $\mathscr{G}'$ contains the collection $\mathscr{C}'$ implies that it contains also the σ-algebra generated by that collection, i.e., $\sigma(\mathscr{C}') \subset \mathscr{G}'$. But we have assumed that $\sigma(\mathscr{C}') = \mathscr{S}'$, so we conclude that $\mathscr{S}' \subset \mathscr{G}'$. By definition of $\mathscr{G}'$ this means that every $\mathscr{S}'$-measurable subset A' has the property that $f^{-1}(A') \in \mathscr{S}$. This shows the measurability of the function f. □

As an application, we see that all continuous functions are good for our purposes. This gives us quite a lot of measurable functions already.

Corollary III.10 (Continuous functions are Borel-measurable)**.**
Let $\mathfrak{X}$ and $\mathfrak{X}'$ be two topological spaces (e.g., metric spaces). Then any continuous function $f\colon \mathfrak{X} \to \mathfrak{X}'$ is Borel-measurable.

Proof. Let $f\colon \mathfrak{X} \to \mathfrak{X}'$ be a continuous function and let $\mathscr{B}(\mathfrak{X})$ and $\mathscr{B}(\mathfrak{X}')$ be the Borel σ-algebras on $\mathfrak{X}$ and $\mathfrak{X}'$, respectively. Recall that Borel-measurability of the function f means that f is $\mathscr{B}(\mathfrak{X})/\mathscr{B}(\mathfrak{X}')$-measurable.

Recall also that the Borel σ-algebra $\mathscr{B}(\mathfrak{X}')$ is generated by the collection $\mathscr{T}(\mathfrak{X}')$ of all open subsets $V' \subset \mathfrak{X}'$. If $f\colon \mathfrak{X} \to \mathfrak{X}'$ is continuous, then for any open subset $V' \subset \mathfrak{X}'$ the preimage $f^{-1}(V') \subset \mathfrak{X}$ is also open: $f^{-1}(V') \in \mathscr{T}(\mathfrak{X})$. As an open set, the preimage is in particular Borel-measurable: $f^{-1}(V') \in \mathscr{B}(\mathfrak{X})$. Therefore we have checked that $f^{-1}(V') \in \mathscr{B}(\mathfrak{X})$ holds for all V' in the collection $\mathscr{T}(\mathfrak{X}')$ which generates $\mathscr{B}(\mathfrak{X}')$. By Lemma III.9, this is sufficient to show that the function f is $\mathscr{B}(\mathfrak{X})/\mathscr{B}(\mathfrak{X}')$-measurable. □

For the case $S' = \mathbb{R}$ (e.g., random numbers), we can in fact use even simpler collections that generate the Borel σ-algebra.

Corollary III.11 (Measurability of real valued functions)**.**
A function $f\colon S \to \mathbb{R}$ is $\mathscr{S}/\mathscr{B}$-measurable if and only if for all $c \in \mathbb{R}$ we have $\{f \leq c\} \in \mathscr{S}$.

Remark III.12. Recall that the notation $\{f \leq c\}$ is shorthand for

$$\{f \leq c\} := \big\{s \in S \;\big|\; f(s) \leq c\big\} \subset S,$$

which is also the preimage $f^{-1}\big((-\infty, c]\big)$ of the interval $(-\infty, c]$.

Proof of Corollary III.11. The collection $\mathscr{C} = \Big\{(-\infty, c] \;\Big|\; c \in \mathbb{R}\Big\}$ generates the Borel σ-algebra $\mathscr{B}$ on $\mathbb{R}$ by Proposition I.11, so the assertion follows from Lemma III.9 above. □

Pointwise operations on measurable functions

Since real numbers have addition and multiplication, we can perform such operations on real-valued functions pointwise. If $f_1, f_2\colon S \to \mathbb{R}$ are two functions, then the

pointwise sum $f_1 + f_2 \colon S \to \mathbb{R}$ is defined by

$$(f_1 + f_2)(s) = f_1(s) + f_2(s) \qquad \forall s \in S,$$

the *pointwise product* $f_1 f_2 \colon S \to \mathbb{R}$ by

$$(f_1 f_2)(s) = f_1(s)\, f_2(s) \qquad \forall s \in S,$$

and if $f \colon S \to \mathbb{R}$ is a function and $\lambda \in \mathbb{R}$ is a scalar, then the *pointwise scalar multiple* $\lambda f \colon S \to \mathbb{R}$ is defined by

$$(\lambda f)(s) = \lambda\, f(s) \qquad \forall s \in S.$$

Let us denote by $\mathrm{m}\mathscr{S}$ the set of all $\mathscr{S}/\mathscr{B}$-measurable functions $S \to \mathbb{R}$. The pointwise operations allow us to construct new measurable functions from old ones.

Proposition III.13 (Pointwise sums and products preserve measurability)**.** *Let $(S, \mathscr{S})$ be a measurable space. Then the set $\mathrm{m}\mathscr{S}$ of all measurable real valued functions on it is stable under taking pointwise sums, pointwise products, and pointwise scalar multiples, i.e., the following hold:*

(i) $f \in \mathrm{m}\mathscr{S},\ \lambda \in \mathbb{R} \implies \lambda f \in \mathrm{m}\mathscr{S}$
(ii) $f_1, f_2 \in \mathrm{m}\mathscr{S} \implies f_1 + f_2 \in \mathrm{m}\mathscr{S}$
(iii) $f_1, f_2 \in \mathrm{m}\mathscr{S} \implies f_1 f_2 \in \mathrm{m}\mathscr{S}$.

Proof: We will prove part (ii), and leave parts (i) and (iii) as exercises.

proof of (ii): Suppose that $f_1, f_2 \in \mathrm{m}\mathscr{S}$. Note that for $c \in \mathbb{R}$ and $s \in S$, the condition $f_1(s) + f_2(s) > c$ holds if and only if there exists a rational number $q \in \mathbb{Q}$ such that we have $f_1(s) > q$ and $f_2(s) > c - q$. In other words, the following two subsets of S are equal

$$\{f_1 + f_2 > c\} = \bigcup_{q \in \mathbb{Q}} \Big(\{f_1 > q\} \cap \{f_2 > c - q\} \Big).$$

By measurability of the functions f_1 and f_2, the subsets $\{f_1 > q\}, \{f_2 > c - q\} \subset S$ are measurable. By properties of σ-algebras, then, the set on the right hand side above is measurable (note that the set $\mathbb{Q}$ of rational numbers is countable). We have thus shown that $\{f_1 + f_2 > c\} \in \mathscr{S}$, and by taking complements we get

$$\{f_1 + f_2 \le c\} = \{f_1 + f_2 > c\}^{\mathbf{c}} \in \mathscr{S}.$$

By Corollary III.11, this is sufficient to show the measurability of the pointwise sum function $f_1 + f_2$. □

Exercise III.4. Prove the assertions (i) and (iii) in Proposition III.13.

We can also use more subtle pointwise operations. If we are given a sequence $f_1, f_2, \ldots$ of functions $f_n \colon S \to \mathbb{R}$ for $n \in \mathbb{N}$, then we can define functions $\sup_n f_n$ and $\inf_n f_n$ on S by the pointwise supremum and infimum

$$(\sup_{n \in \mathbb{N}} f_n)(s) = \sup_{n \in \mathbb{N}} f_n(s) \qquad \text{and} \qquad (\inf_{n \in \mathbb{N}} f_n)(s) = \inf_{n \in \mathbb{N}} f_n(s).$$

These functions, however, may assume values $+\infty$ and $-\infty$ even if each f_n takes only finite real values. Thus the functions

$$\sup_{n \in \mathbb{N}} f_n \colon S \to [-\infty, +\infty] \qquad \text{and} \qquad \inf_{n \in \mathbb{N}} f_n \colon S \to [-\infty, +\infty]$$

are defined so that their allowed range of values is the *extended real line*

$$[-\infty, +\infty] = \mathbb{R} \cup \{+\infty\} \cup \{-\infty\}.$$

The extended real line is equipped with the topology of a closed interval, so that for example the tangent function

$$\tan\colon \big[-\frac{\pi}{2},\frac{\pi}{2}\big] \to [-\infty,+\infty]$$

is a homeomorphism, when we interpret $\tan(\pi/2) = +\infty$ and $\tan(-\pi/2) = -\infty$ (see Example B.19). From this topology, we get the Borel σ-algebra $\mathscr{B}([-\infty,+\infty])$ on the extended real line, in the usual way (Definition I.10). It is so convenient to occasionally allow $+\infty$ and $-\infty$ as possible values that we will slightly abuse the notation, and write $f \in \mathrm{m}\mathscr{S}$ also for functions

$$f\colon S \to [-\infty,+\infty]$$

that are $\mathscr{S}/\mathscr{B}([-\infty,+\infty])$-measurable.

Likewise, for a sequence $f_1, f_2, \dots$ of functions $S \to \mathbb{R}$ we can define the pointwise lim sup and lim inf,

$$(\limsup_n f_n)(s) = \limsup_n f_n(s) \qquad \text{and} \qquad (\liminf_n f_n)(s) = \liminf_n f_n(s).$$

Proposition III.14 (Pointwise supremum and infimum preserve measurability)**.** *Let $(S,\mathscr{S})$ be a measurable space. Suppose that $f_1, f_2, \dots \in \mathrm{m}\mathscr{S}$. Then we have:*

(i) $\sup_n f_n \in \mathrm{m}\mathscr{S}$
(ii) $\inf_n f_n \in \mathrm{m}\mathscr{S}$
(iii) $\limsup_n f_n \in \mathrm{m}\mathscr{S}$
(iv) $\liminf_n f_n \in \mathrm{m}\mathscr{S}$.

Proof: We will prove only parts (i) and (iii) — parts (ii) and (iv) are entirely similar.[2]

proof of part (i). Note that we have $\sup_n f_n(s) \le c$ if and only if $f_n(s) \le c$ for all $n \in \mathbb{N}$. Therefore we can write

$$\Big\{\sup_n f_n \le c\Big\} = \bigcap_{n\in\mathbb{N}} \{f_n \le c\}\,.$$

For each $n \in \mathbb{N}$ we have $\{f_n \le c\} = f_n^{-1}\big((-\infty,c]\big) \in \mathscr{S}$, since $f_n \in \mathrm{m}\mathscr{S}$. Therefore we have $\{\sup_n f_n \le c\} \in \mathscr{S}$ as a countable intersection of sets in the σ-algebra $\mathscr{S}$. By Corollary III.11 we conclude that $\sup_n f_n$ is a $\mathscr{S}$-measurable function.

proof of part (iii). Recall that we have

$$\limsup_n f_n(s) = \inf_{n\in\mathbb{N}} \big(\sup_{k\ge n} f_k(s)\big),$$

because the sequence of functions $g_1, g_2, \dots$ defined by

$$g_n(s) = \sup_{k\ge n} f_k(s)$$

is decreasing (for larger n the supremum contains fewer terms)

$$g_1(s) \ge g_2(s) \ge g_3(s) \ge \cdots.$$

The function g_n is a pointwise supremum of the measurable functions $f_k \in \mathrm{m}\mathscr{S}$, $k \ge n$, and thus itself $\mathscr{S}$-measurable by part (i). Consequently, $\limsup_n f_n$ is the pointwise infimum of the measurable functions $g_n \in \mathrm{m}\mathscr{S}$, $n \in \mathbb{N}$, and thus itself $\mathscr{S}$-measurable by part (ii). □

[2] In the case when values $\pm\infty$ do not appear, parts (ii) and (iv) in fact directly follow from parts (i) and (iii) and Proposition III.13(i) by observing that $\inf_n f_n = -\sup_n(-f_n)$ and $\liminf_n f_n = -\limsup_n(-f_n)$.

With the above we can conclude the extremely useful fact that pointwise limits of measurable functions are measurable.

Corollary III.15 (Pointwise limits of measurable functions are measurable)**.** *Let $(S, \mathscr{S})$ be a measurable space. Suppose that $f_1, f_2, \ldots \in \mathrm{m}\mathscr{S}$, and suppose that for all $s \in S$ the limit*

$$\lim_{n\to\infty} f_n(s)$$

exists. Then the pointwise limit function is measurable: $\lim_{n\to\infty} f_n \in \mathrm{m}\mathscr{S}$.

Proof. When the assumed limit exists, it coincides with both the lim sup and lim inf (cf. Proposition B.4),

$$\liminf_n f_n(s) = \lim_{n\to\infty} f_n(s) = \limsup_n f_n(s).$$

Therefore the assertion follows from Proposition III.14, part (iii) or (iv). □

Let us now revisit earlier examples about repeated coin tossing. The following illustrates how the operations we studied above allow us to construct rather nontrivial random variables starting from very basic ones.

Example III.16 (The question of existence of limit frequency in coin tossing)**.** Let $\Omega = \{\mathrm{H}, \mathrm{T}\}^{\mathbb{N}}$ be the sample space for repeated coin tossing as in Examples O.2 and I.5. Let $\mathscr{F}$ be the σ-algebra on Ω generated by the events

$$E_j := \Big\{\omega \in \Omega \;\Big|\; \omega(j) = \mathrm{H}\Big\} = \text{“the } j\text{:th coin toss is heads”}$$

for $j \in \mathbb{N}$, i.e., $\mathscr{F} = \sigma\big(\{E_j \,|\, j \in \mathbb{N}\}\big)$. Then the indicator random variable of E_j,

$$\mathbb{I}_{E_j}(\omega) = \begin{cases} 1 & \text{if } \omega(j) = \mathrm{H} \\ 0 & \text{if } \omega(j) = \mathrm{T}, \end{cases}$$

is $\mathscr{F}$-measurable by Exercise III.2. The relative frequency of heads in the first n coin tosses,

$$X_n(\omega) = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}_{E_j}(\omega),$$

is then also $\mathscr{F}$-measurable by Proposition III.13 (the relative frequency X_n is constructed from the indicators $\mathbb{I}_{E_j}$ by finite sums and a scalar multiple). Therefore also the upper and lower limits

$$L^+(\omega) = \limsup_n X_n(\omega) \qquad \text{and} \qquad L^-(\omega) = \liminf_n X_n(\omega)$$

are random variables, by Proposition III.14. Knowing that these are random variables, do we learn something interesting about events?

For example, for any $r \in [0, 1]$, we should hope to be able to form the event

$$\text{“relative frequencies of heads tend to } r\text{”} \;=\; \Big\{\lim_{n\to\infty} X_n = r\Big\}$$

which in more careful notation is the subset $\big\{\omega \in \Omega \;\big|\; \lim_{n\to\infty} X_n(\omega) = r\big\} \subset \Omega$. Note that we have

$$\Big\{\lim_{n\to\infty} X_n = r\Big\} = \big\{L^+ = r\big\} \cap \big\{L^- = r\big\},$$

and since L^+ and L^- are $\mathscr{F}$-measurable random variables, the two preimages

$$\big\{L^+ = r\big\} = (L^+)^{-1}(\{r\}) \qquad \text{and} \qquad \big\{L^- = r\big\} = (L^-)^{-1}(\{r\})$$

are $\mathscr{F}$-measurable events. Thus also their intersection is indeed an event,

$$\Big\{\lim_{n\to\infty} X_n = r\Big\} \in \mathscr{F},$$

and so it will at least be meaningful to talk about the probability (once we introduce a probability measure P) of the events that the relative frequency tends to a particular limiting value $r \in [0,1]$.

In Example I.5 we in fact already reached essentially the same conclusion by direct manipulations with events, without using random variables. The present approach, however, has some advantages. Consider for instance the slight variation:

$$\text{“relative frequencies have a limit”} \;=\; \Big\{\exists \lim_{n\to\infty} X_n\Big\}.$$

To show that this is an event, we may now just notice that the limit exists if and only if the corresponding upper and lower limits coincide, $L^+ = L^-$, i.e. if their difference vanishes

$$\Big\{\exists \lim_{n\to\infty} X_n\Big\} = \big\{L^+ - L^- = 0\big\}.$$

This is the preimage of $\{0\} \subset \mathbb{R}$ under the pointwise difference $L^+ - L^-$ of random variables, and as such it is measurable,

$$\Big\{\exists \lim_{n\to\infty} X_n\Big\} \in \mathscr{F}.$$

It would be more cumbersome to try to conclude the same directly by manipulating events using countable set operations in the spirit of Example I.5.

III.4. Simple functions

In Section III.2 we discussed the measurability of functions assuming two possible values, the indicator functions of subsets $A \subset S$ defined by

$$\mathbb{I}_A \colon S \to \mathbb{R}, \qquad \mathbb{I}_A(s) = \begin{cases} 1 & \text{if } s \in A \\ 0 & \text{if } s \notin A. \end{cases}$$

In Exercise III.2 the measurability of the indicator functions was characterized: we have $\mathbb{I}_A \in \mathrm{m}\mathscr{S}$ if and only if $A \in \mathscr{S}$.

From two possible values we next proceed modestly to finitely many possible values.

Definition III.17 (Simple function)**.**
A real valued measurable function assuming only finitely many different values is called a *simple function* (or a *simple random variable* in the probabilistic context).

Any finite linear combination of indicator functions

$$f(s) = \sum_{k=1}^{m} a_k \, \mathbb{I}_{A_k} \tag{III.4}$$

of indicators of measurable sets $A_1, \dots, A_m \in \mathscr{S}$ is a simple function: it is measurable by Proposition III.13, and there are only finitely many real numbers that can be expressed as a sum of some of the coefficients $a_1, \dots, a_m \in \mathbb{R}$. In fact, if a simple function $f \colon S \to \mathbb{R}$ can only assume m different real values $a_1, \dots, a_m$, then we can write it as a linear combination (III.4) with the disjoint measurable sets $A_k := f^{-1}(\{a_k\})$. Occasionally, choosing this minimal linear combination is very convenient, but at other times we might not want to insist on disjointness.

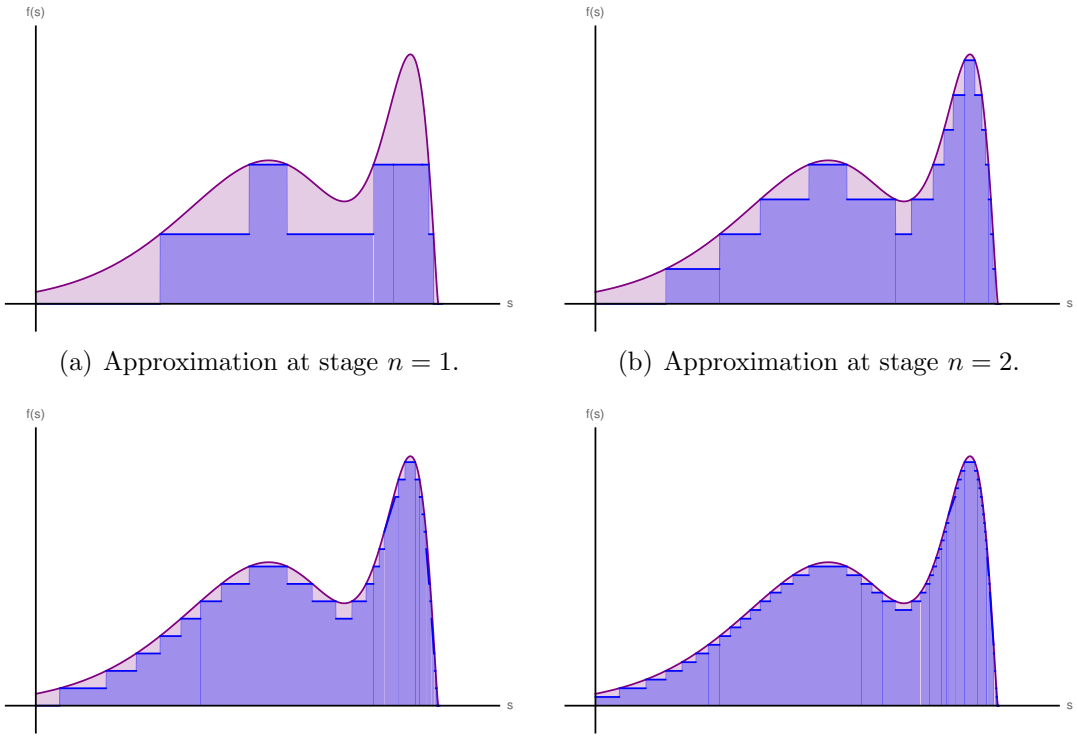
In Corollary III.15 we saw that any pointwise limit of measurable functions is measurable. An important version of the converse statement is also true: any measurable function can be obtained as a pointwise limit of simple functions!

When $(S, \mathscr{S})$ is a measurable space, denote by $\mathrm{m}\mathscr{S}^+$ the set of all functions

$$f\colon S \to [0, +\infty]$$

which are $\mathscr{S}/\mathscr{B}([0, +\infty])$ measurable.

Lemma III.18 (Approximation of non-negative measurable functions)**.** *Let $f \in \mathrm{m}\mathscr{S}^+$. Then there exists a sequence $f_1, f_2, \ldots : S \to [0, +\infty)$ of non-negative simple functions such that $f_n \uparrow f$ pointwise as $n \to \infty$.*



(a) Approximation at stage $n = 1$.

(b) Approximation at stage $n = 2$.

(c) Approximation at stage $n = 3$.

(d) Approximation at stage $n = 4$.

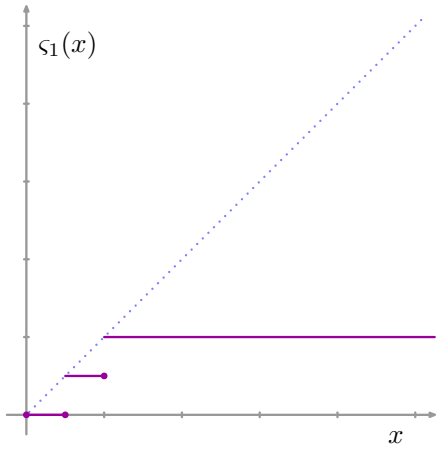
FIGURE III.1. Pointwise increasing approximation of non-negative measurable functions by simple functions.

An illustration of the approximation of a non-negative measurable function is given in Figure III.1. To prove the above approximation lemma, we will construct the approximating sequence explicitly. The idea is that at the n:th stage of approximation, we truncate the values that exceed level n to exactly n, and we replace values below level n by the nearby values on a grid of mesh 2^{-n}. As $n \to \infty$, the truncation is done ever further away, and the grid becomes ever finer. The truncation and discretization at the n:th stage are achieved with the *staircase functions*

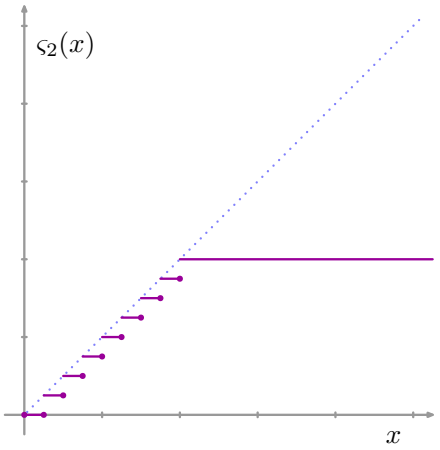
$$\varsigma_n\colon [0, +\infty] \to [0, n]$$

illustrated in Figure III.2 and defined piecewise by the formula

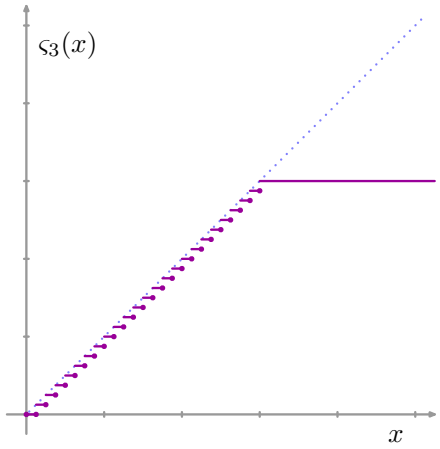
$$\varsigma_n(x) = \begin{cases} 0 & \text{if } 0 \leq x \leq \frac{1}{2^n} \\ \frac{1}{2^n} & \text{if } \frac{1}{2^n} < x \leq \frac{2}{2^n} \\ \frac{2}{2^n} & \text{if } \frac{2}{2^n} < x \leq \frac{3}{2^n} \\ & \vdots \\ \frac{n2^n-1}{2^n} & \text{if } \frac{n2^n-1}{2^n} < x \leq n \\ n & \text{if } n < x. \end{cases} \tag{III.5}$$



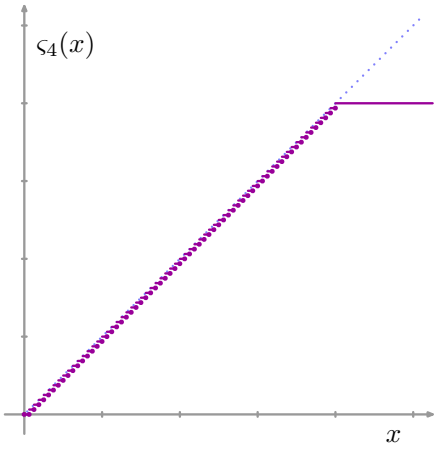
(a) The staircase function ς_n for $n = 1$.



(b) The staircase function ς_n for $n = 2$.



(c) The staircase function ς_n for $n = 3$.



(d) The staircase function ς_n for $n = 4$.

FIGURE III.2. The staircase functions $\varsigma_n \colon [0, +\infty] \to [0, n]$.

We first check that these staircase functions provide a good approximation of the identity function of $[0, +\infty]$.

Lemma III.19 (Properties of the staircase functions)**.**

(a) *The staircase functions ς_n are simple, Borel-measurable, and left-continuous.*
(b) *For every $x \in [0,+\infty]$, we have $\varsigma_n(x) \uparrow x$ as $n \to \infty$.*

Proof. From the definition (III.5) it is clear that the set of possible values of ς_n is the finite set $\{j\,2^{-n} \mid j \in \{0,1,\dots,n2^{-n}\}\}$. Each of the preimages $\varsigma_n^{-1}(\{j2^{-n}\})$ is an interval, therefore a Borel set, so ς_n is Borel-measurable and simple. Left continuity

$$\lim_{x' \uparrow x} \varsigma_n(x') = \varsigma_n(x) \qquad \forall x \in [0,+\infty] \tag{III.6}$$

is also clear from the definition (III.5). This proves part (a).

For part (b), we consider the cases $x = +\infty$ and $x \in [0,+\infty)$ separately.

Consider first $x = +\infty$. For any $n \in \mathbb{N}$, we have $\varsigma_n(+\infty) = n$ by definition (III.5). These values form an increasing sequence tending to infinity, i.e., we have $\varsigma_n(+\infty) \uparrow +\infty$ as claimed.

Consider then $x \in [0,+\infty)$. For any $n > x$, we have $|\varsigma_n(x) - x| \leq 2^{-n}$ by definition (III.5). This shows that $\varsigma_n(x) \to x$ as $n \to \infty$. It is also easy to see that the sequence of values is increasing, $\varsigma_1(x) \leq \varsigma_1(x) \leq \cdots$. Thus we have $\varsigma_n(x) \uparrow x$ as claimed. This finishes the proof of part (b). □

With this, we are ready to prove the approximation lemma.

Proof of Lemma III.18 Let $f \in \mathrm{m}\mathscr{S}^+$ be a non-negative measurable function. For each $n \in \mathbb{N}$, define the function $f_n = \varsigma_n \circ f$, i.e.,

$$f_n(s) = \varsigma_n\big(f(s)\big) \qquad \text{for } s \in S.$$

This f_n is measurable as the composition of the measurable function $f\colon S \to [0,+\infty]$ and the Borel function $\varsigma_n\colon [0,+\infty] \to [0,n]$. The possible values of f_n are contained in the finite set of possible values of ς_n. We conclude that f_n is a non-negative simple function.

At any $s \in S$ we have $f(s) \in [0,+\infty]$, and therefore

$$f_n(s) = \varsigma_n\big(f(s)\big) \uparrow f(s) \qquad \text{as } n \to \infty$$

by part (b) of the previous lemma. We have thus constructed the desired sequence $f_1, f_2, \dots$ of simple functions, which approximates f pointwise in a monotone increasing way. □

Remark III.20 (Approximating approximations)**.**

For the proof of the approximation lemma itself, it was not really important whether we made the staircase functions left-continuous or not. Where left-continuity is actually convenient is the following situation, which appears in particular in the proof of the Monotone convergence theorem (Theorem VII.8) in Appendix D.

If $g_1, g_2, \dots \in \mathrm{m}\mathscr{S}^+$ is any sequence of non-negative functions such that $g_n \uparrow g$ as $n \to \infty$, then we may construct the simple approximations $g_n^{(r)} := \varsigma_r \circ g_n$, $r \in \mathbb{N}$, of each g_n,

$$g_n^{(r)} \uparrow g_n \qquad \text{as } r \to \infty$$

as well as the simple approximations $g^{(r)} := \varsigma_r \circ g$, $r \in \mathbb{N}$, of the limit function g,

$$g^{(r)} \uparrow g \qquad \text{as } r \to \infty.$$

In this setup, the approximations of the limit function g are the limits of the approximations of each g_n: for every $s \in S$ we have by assumption $g_n(s) \uparrow g(s)$ as $n \to \infty$, and therefore by left-continuity (III.6) of ς_n we get, for any $r \in \mathbb{N}$,

$$g_n^{(r)}(s) = \varsigma_r\big(g_n(s)\big) \uparrow \varsigma_r\big(g(s)\big) = g^{(r)}(s) \qquad \text{as } n \to \infty.$$

Lecture IV

Information generated by random variables

Probability theory offers an important interpretation of σ-algebras: they describe information.

Let us first mention a few contexts in which this notion of information is used in stochastics. In this course we will first use of the notion in the next Lecture V in relation to independence: the elementary notion of independence of events is generalized to the notion of independence of information. Another common and fruitful use of information (see Appendix E) is conditional expected value, which represents the best estimate of a random number given some (partial) information about it. Finally, stochastic processes are random time-dependent phenomena, and it is often relevant to model how information accumulates as we observe the phenomenon over a period of time — the mathematical notion suitable for this is refining collections of σ-algebras indexed by time known as filtrations. There would be yet other contexts, but it is in fact useful to interpret all σ-algebras as describing information, and relate the notion of measurability of functions to this interpretation. So let us start with an informal description of the ideas and then proceed to precise definition and properties.

To describe the idea informally, suppose that $(\Omega, \mathscr{F}, \mathsf{P})$ is a probability space and we have an indexed collection $(Y_\gamma)_{\gamma\in\Gamma}$ of random variables $Y_\gamma\colon \Omega \to \mathbb{R}$ on it.[1] To understand what information is contained in these random variables, recall the two step procedure by which randomness is thought to arise:

1.) "Chance determines the random outcome $\omega \in \Omega$."
2.) "The outcome ω determines the values $Y_\gamma(\omega)$ of quantities of interest Y_γ."

The motivating question about information is then:

> "If you do not know the outcome ω of all randomness, but someone tells you the values of the quantities of interest Y_γ for all $\gamma \in \Gamma$, then for which events $E \in \mathscr{F}$ are you able to decide whether E occurs or not?"

Formulated in this way it makes sense that the information contained in the collection $(Y_\gamma)_{\gamma\in\Gamma}$ of random variables is some collection of events — namely those events whose occurrence can be decided based on the random variables. Evidently, any event of type $\{Y_\gamma \in A'\}$ concerning the value of any one of the random variables Y_γ can be decided, and thus belongs to the collection. However, in deciding about

[1] There is no fundamental reason to require that the random variables are real-valued, but we assume this for the sake of concreteness — and in order to avoid the very awkward notation that would arise if each random variable $Y_\gamma\colon \Omega \to S_\gamma$ in the collection would have a different set S_γ of allowed values (necessarily then also equipped with its own σ-algebra $\mathscr{S}_\gamma$).

events you are furthermore allowed to use logical reasoning (e.g., the logical operations described in Section I.1), so the collection of events that can be decided should itself be a σ-algebra. This motivates the following definition.

IV.1. Definition of σ-algebra generated by random variables

Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space and $(Y_\gamma)_{\gamma\in\Gamma}$ an indexed collection of random variables $Y_\gamma\colon \Omega \to \mathbb{R}$ on it.

Definition IV.1 (Sigma-algebra generated by random variables)**.**
The *σ-algebra generated by the collection $(Y_\gamma)_{\gamma\in\Gamma}$ of random variables* is the smallest σ-algebra $\mathscr{Y}$ on Ω such that for each $\gamma \in \Gamma$ the random variable Y_γ is $\mathscr{Y}$-measurable. We denote the σ-algebra generated by the collection by $\mathscr{Y} = \sigma\big((Y_\gamma)_{\gamma\in\Gamma}\big)$.

Remark IV.2. Similarly to Section I.3, the smallest σ-algebra with the above property exists and is unique: it is the intersection of all σ-algebras satisfying the property.

Remark IV.3. Since each random variable Y_γ is by definition at least $\mathscr{F}$-measurable, we obviously have $\sigma\big((Y_\gamma)_{\gamma\in\Gamma}\big) \subset \mathscr{F}$. According to the information interpretation, $\mathscr{F}$ represents "full information" (all events on our probability space), so no amount of random variables could contain more information than that.

Remark IV.4. Although we use the notation $\sigma(\cdots)$ both for the σ-algebra generated by a collection of subsets (Definition I.6) and for the σ-algebra generated by a collection of random variables (Definition IV.1), we trust that there is no risk of confusion: it should always be clear from the context which of these two closely related notions is meant.

An already interesting special case is a collection which contains only one random variable: the σ-algebra generated by $Y\colon \Omega \to \mathbb{R}$ is the smallest σ-algebra with respect to which Y is measurable. We denote it simply by $\sigma(Y)$.

Exercise IV.1 (The σ-algebra generated by a random number)**.**
Let Y be a real-valued random variable defined on a probability space $(\Omega, \mathscr{F}, \mathsf{P})$.

(a) Show that the σ-algebra $\sigma(Y)$ generated by the random variable Y coincides with the σ-algebra $\sigma\big(Y^{-1}(\mathscr{B})\big)$ generated by the collection of events $Y^{-1}(\mathscr{B}) = \{Y^{-1}(B) \,\big|\, B \in \mathscr{B}\}$.
(b) Show that we in fact have the equality $\sigma(Y) = Y^{-1}(\mathscr{B})$.

Exercise IV.2 (A π-system to generate the σ-algebra generated by a random number)**.**
Let Y be a real-valued random variable defined on a probability space $(\Omega, \mathscr{F}, \mathsf{P})$. Let $\mathscr{J}(\mathbb{R}) = \big\{(-\infty, x] \,\big|\, x \in \mathbb{R}\big\}$ be the π-system on $\mathbb{R}$ as in Example II.25 and define

$$\mathscr{I} = Y^{-1}\big(\mathscr{J}(\mathbb{R})\big) = \Big\{Y^{-1}\big((-\infty, x]\big) \,\Big|\, x \in \mathbb{R}\Big\}.$$

Show that $\mathscr{I}$ is a π-system on Ω which generates the σ-algebra $\sigma(Y)$ generated by the random variable Y, i.e., $\sigma(\mathscr{I}) = \sigma(Y)$.

IV.2. Doob's representation theorem

The definition of information contained in random variables may seem abstract. We next state and prove a theorem, which offers a good interpretation for the information contained in one real-valued random variable.

Theorem IV.5 (Doob's representation theorem)**.**
Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space and $Y\colon \Omega \to \mathbb{R}$ and $Z\colon \Omega \to \mathbb{R}$ two real valued random variables on it. Then Z is $\sigma(Y)$-measurable[2] *if and only if there exists a Borel function $f\colon \mathbb{R} \to \mathbb{R}$ such that*

$$Z = f(Y).$$

Remark IV.6. The precise meaning of

$$Z = f(Y)$$

is that for all possible outcomes $\omega \in \Omega$ we have

$$Z(\omega) = f(Y(\omega)).$$

In other words, keeping in mind that random variables are ultimately just functions on the sample space Ω, the function $Z\colon \Omega \to \mathbb{R}$ is the composition

$$Z = f \circ Y$$

of functions $Y\colon \Omega \to \mathbb{R}$ and $f\colon \mathbb{R} \to \mathbb{R}$

$$\Omega \xrightarrow{Y} \mathbb{R} \xrightarrow{f} \mathbb{R}\,, \qquad \Omega \xrightarrow[Z=f\circ Y]{} \mathbb{R}\,.$$

Theorem IV.5 gives the following interpretation: to say that Z is measurable with respect to the information contained in Y means that the value of Z could be obtained from the value of Y by applying some deterministic function f.

In the proof we use the Monotone class theorem, which can be found in Appendix C.

Proof of Theorem IV.5. The "if" direction of the statement is easy. Namely, if we have $Z = f \circ Y$ for some $\mathscr{B}/\mathscr{B}$-measurable function $f\colon \mathbb{R} \to \mathbb{R}$, then since $Y\colon \Omega \to \mathbb{R}$ is by definition $\sigma(Y)/\mathscr{B}$-measurable, it follows from Proposition III.8 that the composition $Z = f \circ Y\colon \Omega \to \mathbb{R}$ is $\sigma(Y)/\mathscr{B}$-measurable. It therefore remains to prove the "only if" direction. We first prove the "only if" direction assuming that Z is bounded, and afterwards extend to full generality.

bounded Z: Let us define $\mathscr{H}$ as the collection of all those bounded functions $Z\colon \Omega \to \mathbb{R}$ which can be written as $Z = f \circ Y$ for some bounded Borel function f. Our goal is to show that the collection $\mathscr{H}$ contains all $\sigma(Y)$-measurable bounded functions.

Denote $\mathscr{J} = \sigma(Y)$. We start by checking that at least the indicator $\mathbb{I}_E$ of any event $E \in \mathscr{J}$ belongs to $\mathscr{H}$. Recall the fact $\mathscr{J} = \sigma(Y) = \{Y^{-1}(B) \,|\, B \in \mathscr{B}\}$ from Exercise IV.1: any $E \in \mathscr{J}$ is of the form $E = Y^{-1}(B)$ for some $B \in \mathscr{B}$. Therefore the indicator of E is the function

$$\mathbb{I}_E(\omega) = \mathbb{I}_{Y^{-1}(B)}(\omega) = \begin{cases} 1 & \text{if } \omega \in Y^{-1}(B) \\ 0 & \text{otherwise.} \end{cases} \tag{IV.1}$$

In contrast, the indicator function of $B \subset \mathbb{R}$ is a function defined on $\mathbb{R}$: in fact, $\mathbb{I}_B\colon \mathbb{R} \to \mathbb{R}$ is a bounded Borel function. We notice that

$$\mathbb{I}_B\big(Y(\omega)\big) = \begin{cases} 1 & \text{if } Y(\omega) \in B \\ 0 & \text{otherwise.} \end{cases} \tag{IV.2}$$

[2]More precisely, $\sigma(Y)/\mathscr{B}$-measurable.

Comparing (IV.1) and (IV.2) we see that

$$\mathbb{I}_E(\omega) = \mathbb{I}_B\big(Y(\omega)\big).$$

This conclusion $\mathbb{I}_E = \mathbb{I}_B \circ Y$ shows that the indicator of any $E \in \mathscr{J}$ belongs to the collection we are considering, $\mathbb{I}_E \in \mathscr{H}$.

We now show that $\mathscr{H}$ is a *monotone class* as defined in Appendix C (see Definition C.1), i.e., it satisfies the three properties (MC-1), (MC-$\mathbb{R}$), and (MC-$\uparrow$).

The first of these properties is obvious: if we take f to be the constant function $f(x) = 1$ for all $x \in \mathbb{R}$ then $Z = f \circ Y$ is the constant random variable $Z(\omega) = f(Y(\omega)) = 1$ for all $\omega \in \Omega$. Thus indeed the constant function 1 on Ω belongs to $\mathscr{H}$. This is property (MC-1) for $\mathscr{H}$.

The second property is easy, too. If we have $Z_1, Z_2 \in \mathscr{H}$, then we can write $Z_1 = f_1 \circ Y$ and $Z_2 = f_2 \circ Y$ for some bounded Borel-measurable functions $f_1, f_2 \colon \mathbb{R} \to \mathbb{R}$. Then for $c_1, c_2 \in \mathbb{R}$ we have $c_1 Z_1 + c_2 Z_2 = f \circ Y$, where $f = c_1 f_1 + c_2 f_2$ is a pointwise linear combination function $\mathbb{R} \to \mathbb{R}$, which is also bounded and Borel-measurable by Proposition III.13. This is property (MC-$\mathbb{R}$) for $\mathscr{H}$.

The last property is checked as follows. Suppose that $Z_n \uparrow Z$ as $n \to \infty$, and that $Z_n \in \mathscr{H}$ for all $n \in \mathbb{N}$, and that $0 \le Z_n(\omega) \le K$ for all $\omega \in \Omega$ and some constant $K < \infty$. Then we have $Z_n = f_n \circ Y$ for some bounded Borel-measurable functions $f_n \colon \mathbb{R}$. We may assume that $0 \le f_n \le K$ pointwise[3]. Now define $f = \limsup_n f_n$. Then $f \colon \mathbb{R} \to \mathbb{R}$ is Borel measurable by Proposition III.14, and it is also bounded: $0 \le f \le K$. Moreover, from our assumptions it now follows that

$$Z = \lim_{n\to\infty} Z_n = \lim_{n\to\infty} f_n \circ Y = \limsup_n \; f_n \circ Y = f \circ Y.$$

We thus obtain that $Z \in \mathscr{H}$. This is property (MC-$\uparrow$) for $\mathscr{H}$.

We have shown that $\mathscr{H}$ is a monotone class which contains the indicator functions of all sets E in the collection $\mathscr{J} = \sigma(Y)$. The collection $\mathscr{J} = \sigma(Y)$ is a σ-algebra and thus in particular a π-system (Remark II.24). The Monotone class theorem (Theorem C.2) therefore guarantees that $\mathscr{H}$ contains all bounded $\sigma(Y)$-measurable functions. In other words, every bounded $\sigma(Y)$-measurable $Z \colon \Omega \to \mathbb{R}$ is of the form $Z = f \circ Y$ for some bounded Borel function $f \colon \mathbb{R} \to \mathbb{R}$.

unbounded Z: Consider now the general case, where $Z \colon \Omega \to \mathbb{R}$ can be unbounded. In that case we can apply the function

$$\arctan \colon \mathbb{R} \to \big(-\frac{\pi}{2}, \frac{\pi}{2}\big)$$

and get the bounded random variable

$$\widetilde{Z} = \arctan \circ Z.$$

By the composition property, if Z is $\sigma(Y)$-measurable, then $\widetilde{Z}$ is also (arctan is continuous and therefore Borel). From the first step of the proof, we thus know that $\widetilde{Z} = \widetilde{f} \circ Y$ for some bounded Borel function $\widetilde{f} \colon \mathbb{R} \to \mathbb{R}$. Then we have

$$Z = \tan \circ \widetilde{Z} = \tan \circ \widetilde{f} \circ Y = f \circ Y,$$

where $f = \tan \circ \widetilde{f}$. This f is Borel measurable and the proof is complete. □

[3]If f_n do not already satisfy this property, we may truncate them. Using the functions

$$\tilde{f}_n(x) := \begin{cases} 0 & \text{when } f_n(x) < 0 \\ f_n(x) & \text{when } 0 \le f_n(x) \le K \\ K & \text{when } K < f_n(x) \end{cases}$$

instead, we still have $Z_n = \tilde{f}_n \circ Y$.

Lecture V

Independence

In the previous lecture (Lecture IV) we saw that probability theory offers an important interpretation of σ-algebras: they describe information. One of the first uses of the notion of information pertains to probabilistic independence: we will generalize the elementary notion of independence of events to the notion of independence of information.

The intuitive interpretation of probabilistic independence should be familiar from basic courses in probability and statistics, so we recall the idea only very briefly. Suppose that A and B are two events, and assume also that $\mathsf{P}[B] > 0$ so that the conditional probability $\mathsf{P}[A|B] = \frac{\mathsf{P}[A\cap B]}{\mathsf{P}[B]}$ of the event A given the occurrence of B can be defined. If the probabilities of the two events satisfy

$$\mathsf{P}[A \cap B] = \mathsf{P}[A]\ \mathsf{P}[B],$$

then we get

$$\mathsf{P}[A|B] = \frac{\mathsf{P}[A \cap B]}{\mathsf{P}[B]} = \frac{\mathsf{P}[A]\ \mathsf{P}[B]}{\mathsf{P}[B]} = \mathsf{P}[A],$$

i.e., the conditional probability of A given B is just the probability of A. We interpret this as saying that the knowledge of the occurrence of B does not reveal anything that could be used to improve our estimate about the occurrence of A, and we therefore consider the event A independent of the event B.[1]

We start this lecture by introducing the abstract and general notion of probabilistic independence, and we then show its relation to the more familiar elementary notion which was also used in the interpretation above. After the definitions and basic properties, we also discuss the first profound techniques related to independence: the Borel-Cantelli lemmas.

V.1. Definition of independence

Throughout, let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space.

Let $(\mathscr{G}_j)_{j\in J}$ be a collection of σ-algebras, $\mathscr{G}_j \subset \mathscr{F}$ for all $j \in J$. For each j, we think of $\mathscr{G}_j$ representing some information (for example available to a person j). Whether these informations, for different j, are independent of each other with respect to the underlying probability P is captured by the following definition.

Definition V.1 (Independence of sigma-algebras)**.**
The collection $(\mathscr{G}_j)_{j\in J}$ of σ-algebras is *independent* if for any distinct $j_1, \ldots, j_n \in$

[1] Since the condition $\mathsf{P}[A \cap B] = \mathsf{P}[A]\ \mathsf{P}[B]$ is symmetric under interchange of A and B, we also consider the event B independent of A. It is, in fact, better to use symmetric terminology and say that events A and B are independent.

J and any events $A_{j_1} \in \mathscr{G}_{j_1}, \dots, A_{j_n} \in \mathscr{G}_{j_n}$ we have

$$\mathsf{P}\big[A_{j_1} \cap A_{j_2} \cap \cdots \cap A_{j_n}\big] = \mathsf{P}\big[A_{j_1}\big]\, \mathsf{P}\big[A_{j_2}\big] \cdots \mathsf{P}\big[A_{j_n}\big]. \tag{V.1}$$

This is the abstract and general notion of probabilistic independence, on which other notions of independence are founded. It is important to note that while the collection of σ-algebras here may be infinite, the formula (V.1) is only ever required for intersections of finitely many events.

To get from the abstract and general notion to a more concrete case, let now $(X_j)_{j\in J}$ be a collection of random variables. We need to define what it means for these random variables to be independent. For this purpose, we consider the σ-algebra $\sigma(X_j)$ generated by X_j, for each $j \in J$ separately (see Definition IV.1).

Definition V.2 (Independence of random variables)**.**
The collection $(X_j)_{j\in J}$ of random variables is *independent* if the collection $\big(\sigma(X_j)\big)_{j\in J}$ of σ-algebras generated by them is independent in the sense of Definition V.1.

Finally, let $(E_j)_{j\in J}$ be a collection of events, $E_j \in \mathscr{F}$. To define what it means for these events to be independent, we consider the indicator random variables $\mathbb{I}_{E_j} \colon \Omega \to \mathbb{R}$ of the events E_j, for $j \in J$ (see Equation (III.3)).

Definition V.3 (Independence of events)**.**
The collection $(E_j)_{j\in J}$ of events is *independent* if the collection $\big(\mathbb{I}_{E_j}\big)_{j\in J}$ of the corresponding indicator random variables is independent in the sense of Definition V.2.

Remark V.4 (The elementary notion of independence of events)**.**
The σ-algebra generated by the indicator random variable $\mathbb{I}_E \colon \Omega \to \mathbb{R}$ of an event $E \subset \Omega$ is

$$\sigma(\mathbb{I}_E) = \{\emptyset, E, E^{\mathbf{c}}, \Omega\}\,.$$

Thus events $(E_j)_{j\in J}$ are independent if and only if the σ-algebras $\big(\{\emptyset, E_j, E_j^{\mathbf{c}}, \Omega\}\big)_{j\in J}$ are independent. In view of Definition V.1, this amounts to the equalities

$$\mathsf{P}\big[E_{j_1}^\star \cap E_{j_2}^\star \cap \cdots \cap E_{j_n}^\star\big] = \mathsf{P}\big[E_{j_1}^\star\big]\, \mathsf{P}\big[E_{j_2}^\star\big] \cdots \mathsf{P}\big[E_{j_n}^\star\big], \tag{V.2}$$

where each $E_{j_k}^\star$ stands for either E_{j_k} or its complement $E_{j_k}^{\mathbf{c}}$ (note that including impossible events $\emptyset$ in the intersection is unnecessary, since both sides of the equation would then vanish automatically, and including sure events Ω just amounts to having fewer terms in both the intersection and the product on the two sides of the equation).

For the sake of concreteness, consider now just two events, E_1 and E_2. If E_1 and E_2 are independent, then as a special case of (V.2) we at least have the familiar defining equality

$$\mathsf{P}[E_1 \cap E_2] = \mathsf{P}[E_1]\, \mathsf{P}[E_2]. \tag{V.3}$$

But also conversely, if (V.3) holds then one can derive the equations (V.2) involving possible complements — the reader is invited to directly derive at least the equality $\mathsf{P}[E_1 \cap E_2^{\mathbf{c}}] = \mathsf{P}[E_1]\,\mathsf{P}[E_2^{\mathbf{c}}]$ from (V.3). Of course, case by case checking the equations (V.2) would be impractical as well as inelegant, so we instead develop systematic tools in Section V.2 below.

Exercise V.1. Show that the equality $\mathsf{P}[E_1 \cap E_2^{\mathbf{c}}] = \mathsf{P}[E_1]\,\mathsf{P}[E_2^{\mathbf{c}}]$ follows from (V.3).

Exercise V.2 (Independence is preserved under measurable functions)**.**
Let X and Y be two $\mathbb{R}$-valued random variables defined on a common probability space $(\Omega, \mathscr{F}, \mathsf{P})$.

(a) Show that if X and Y are independent, then $f(X)$ and $g(Y)$ are independent for any Borel-measurable functions $f, g\colon \mathbb{R} \to \mathbb{R}$.
(b) Show that if $f(X)$ and $g(Y)$ are independent for all Borel-measurable functions $f, g\colon \mathbb{R} \to \mathbb{R}$, then X and Y are independent.

Exercise V.3. Assume that X and Y are two $\mathbb{R}$-valued random variables such that we have $\mathsf{P}\big[X + Y = 42\big] = 1$. Is it possible that X and Y are independent?

Exercise V.4 (Calculations with two independent geometrically distributed numbers)**.**
Let $X, Y\colon \Omega \to \mathbb{N}$ be two independent random variables with

$$\mathsf{P}[X = j] = \mathsf{P}[Y = j] = \frac{1}{2^j} \qquad \text{for all } j \in \mathbb{N} = \{1, 2, \ldots\}.$$

(a) Show that $\mathsf{P}[Y > n] = \frac{1}{2^n}$ for any $n \in \mathbb{N}$.

Calculate the following probabilities

(b): $\mathsf{P}[X = Y]$ (c): $\mathsf{P}[\min(X, Y) \le k]$, where $k \in \mathbb{N}$
(d): $\mathsf{P}[Y > X]$ (e): $\mathsf{P}[X > kY]$, where $k \in \mathbb{N}$
(f): $\mathsf{P}[X \text{ divides } Y]$

Hint: *The correct final results are among the following:*

$$\frac{1}{3}, \qquad \frac{1}{2^{k+1} - 1}, \qquad 1 - \frac{1}{4^k}, \qquad \sum_{k=1}^{\infty} \frac{1}{2^{k+1} - 1} = \sum_{k=1}^{\infty} \frac{1}{4^k - 2^k}.$$

Notation

We abbreviate independence by the symbol $\perp\!\!\!\perp$. We thus denote:

$(\mathscr{G}_j)_{j\in J} \perp\!\!\!\perp$ if the collection $(\mathscr{G}_j)_{j\in J}$ of σ-algebras is independent
$(X_j)_{j\in J} \perp\!\!\!\perp$ if the collection $(X_j)_{j\in J}$ of random variables is independent
$(E_j)_{j\in J} \perp\!\!\!\perp$ if the collection $(E_j)_{j\in J}$ of events is independent.

In the case of enumerated (countable) collections, we use the notation:

$\mathscr{G}_1, \mathscr{G}_2, \ldots \perp\!\!\!\perp$ if the collection $(\mathscr{G}_j)_{j\in\mathbb{N}}$ of σ-algebras is independent
$X_1, X_2, \ldots \perp\!\!\!\perp$ if the collection $(X_j)_{j\in\mathbb{N}}$ of random variables is independent
$E_1, E_2, \ldots \perp\!\!\!\perp$ if the collection $(E_j)_{j\in\mathbb{N}}$ of events is independent.

In the case of collections of just two members, we use notation:

$\mathscr{G}_1 \perp\!\!\!\perp \mathscr{G}_2$ if the collection $(\mathscr{G}_j)_{j\in\{1,2\}}$ of σ-algebras is independent
$X_1 \perp\!\!\!\perp X_2$ if the collection $(X_j)_{j\in\{1,2\}}$ of random variables is independent
$E_1 \perp\!\!\!\perp E_2$ if the collection $(E_j)_{j\in\{1,2\}}$ of events is independent.

Exercise V.5 (Pairwise independence does not imply independence)**.**
Construct an example in which three σ-algebras $\mathscr{G}_1$, $\mathscr{G}_2$, $\mathscr{G}_3$ are pairwise independent,

$$\mathscr{G}_1 \perp\!\!\!\perp \mathscr{G}_2, \qquad \mathscr{G}_1 \perp\!\!\!\perp \mathscr{G}_3, \qquad \mathscr{G}_2 \perp\!\!\!\perp \mathscr{G}_3,$$

but the collection $\big(\mathscr{G}_j\big)_{j\in\{1,2,3\}}$ of all three is not independent.

V.2. Verifying independence

To check independence, one usually does not want to work directly with the general definition, but rather use an easier sufficient condition.

Proposition V.5 (A sufficient condition for independence)**.** *Suppose that $\mathscr{I}_1$ and $\mathscr{I}_2$ are two π-systems on Ω and let $\mathscr{G}_1 = \sigma(\mathscr{I}_1)$ and $\mathscr{G}_2 = \sigma(\mathscr{I}_2)$ be the σ-algebras generated by them. Then the following conditions are equivalent:*

(i) $\mathscr{G}_1 \perp\!\!\!\perp \mathscr{G}_2$
(ii) *For all $I_1 \in \mathscr{I}_1$ and $I_2 \in \mathscr{I}_2$, we have $\mathsf{P}[I_1 \cap I_2] = \mathsf{P}[I_1]\,\mathsf{P}[I_2]$.*

Proof. Clearly condition (i) implies (ii), because $\mathscr{I}_1 \subset \sigma(\mathscr{I}_1) = \mathscr{G}_1$ and $\mathscr{I}_2 \subset \sigma(\mathscr{I}_2) = \mathscr{G}_2$. σ-algebras It remains to show that (ii) implies (i). Let us therefore assume (ii). According to Definition V.1, we must prove that then

$$\mathsf{P}[E_1 \cap E_2] = \mathsf{P}[E_1]\,\mathsf{P}[E_2]$$

holds for all $E_1 \in \mathscr{G}_1$, $E_2 \in \mathscr{G}_2$. We do this in two steps: first assuming that E_1 is in the π-system $\mathscr{I}_1$, and then for a general $E_1 \in \mathscr{G}_1$.

step 1: Let $I_1 \in \mathscr{I}_1$. In this first step we seek to prove that

$$\mathsf{P}[I_1 \cap E_2] = \mathsf{P}[I_1]\,\mathsf{P}[E_2] \tag{V.4}$$

holds for all $E_2 \in \mathscr{G}_2$. If $\mathsf{P}[I_1] = 0$, then equations (V.4) hold trivially, because both sides vanish. We may therefore assume that $\mathsf{P}[I_1] > 0$. Then we define a new probability measure $\widetilde{\mathsf{P}}_2$ on $(\Omega, \mathscr{G}_2)$ by the formula

$$\widetilde{\mathsf{P}}_2[E_2] = \frac{\mathsf{P}[I_1 \cap E_2]}{\mathsf{P}[I_1]} \qquad \text{for } E_2 \in \mathscr{G}_2 \tag{V.5}$$

(this is indeed a probability measure by Exercise II.1). From assumption (ii), it follows that the two probability measures P and $\widetilde{\mathsf{P}}_2$ coincide on the π-system $\mathscr{I}_2$: if $I_2 \in \mathscr{I}_2$ then

$$\widetilde{\mathsf{P}}_2[I_2] = \frac{\mathsf{P}[I_1 \cap I_2]}{\mathsf{P}[I_1]} \overset{\text{(ii)}}{=} \frac{\mathsf{P}[I_1]\,\mathsf{P}[I_2]}{\mathsf{P}[I_1]} = \mathsf{P}[I_2].$$

Dynkin's identification theorem (Theorem II.26) thus implies that they coincide on the entire σ-algebra $\mathscr{G}_2$ generated by this π-system, i.e. $\widetilde{\mathsf{P}}_2[E_2] = \mathsf{P}[E_2]$ for all $E_2 \in \mathscr{G}_2$. This shows that

$$\mathsf{P}[E_2] = \widetilde{\mathsf{P}}_2[E_2] = \frac{\mathsf{P}[I_1 \cap E_2]}{\mathsf{P}[I_1]},$$

which upon multiplying by $\mathsf{P}[I_1]$ gives the desired equality (V.4).

step 2: Let $E_2 \in \mathscr{G}_2$. In this last step we seek to prove, with a method analogous to step 1, that

$$\mathsf{P}[E_1 \cap E_2] = \mathsf{P}[E_1]\,\mathsf{P}[E_2] \tag{V.6}$$

holds for all $E_1 \in \mathscr{G}_1$. This will conclude the proof. We may assume that $\mathsf{P}[E_2] > 0$, because otherwise both sides of (V.6) vanish. Then we again define a new probability measure $\widetilde{\mathsf{P}}_1$, this time on $(\Omega, \mathscr{G}_1)$, by the formula

$$\widetilde{\mathsf{P}}_1[E_1] = \frac{\mathsf{P}[E_1 \cap E_2]}{\mathsf{P}[E_2]} \qquad \text{for } E_1 \in \mathscr{G}_1. \tag{V.7}$$

By step 1, the two probability measures P and $\widetilde{\mathsf{P}}_1$ coincide on the π-system $\mathscr{I}_1$. From Dynkin's identification theorem (Theorem II.26) it thus follows that they coincide on the σ-algebra $\mathscr{G}_1$ generated by this π-system. This shows that for all $E_1 \in \mathscr{G}_1$ we have

$$\mathsf{P}[E_1] = \widetilde{\mathsf{P}}_1[E_1] = \frac{\mathsf{P}[E_1 \cap E_2]}{\mathsf{P}[E_2]}.$$

Upon multiplying by $\mathsf{P}[E_2]$ we get the desired equality (V.6), which shows that the σ-algebras $\mathscr{G}_1$ and $\mathscr{G}_2$ are independent. This establishes the implication (ii) $\Rightarrow$ (i), and concludes the proof. □

A typical example application of this is the following (familiar and) practical criterion for the independence of two random numbers using their cumulative distribution functions.

Corollary V.6 (Independence using cumulative distribution functions)**.**
Suppose that $X_1, X_2 \colon \Omega \to \mathbb{R}$ are two random variables. Then we have $X_1 \perp\!\!\!\perp X_2$ if and only if

$$\mathsf{P}\big[X_1 \leq x_1,\ X_2 \leq x_2\big] = \mathsf{P}\big[X_1 \leq x_1\big]\ \mathsf{P}\big[X_2 \leq x_2\big] \tag{V.8}$$

for all $x_1, x_2 \in \mathbb{R}$.

Proof. Let $\mathscr{J}(\mathbb{R}) = \big\{(-\infty, x] \;\big|\; x \in \mathbb{R}\big\}$ be the π-system on $\mathbb{R}$ as in Example II.25. Let

$$\mathscr{I}_1 = X_1^{-1}\big(\mathscr{J}(\mathbb{R})\big) = \left\{ \big\{\omega \in \Omega \;\big|\; X_1(\omega) \leq x\big\} \;\middle|\; x \in \mathbb{R} \right\}.$$

Then $\mathscr{I}_1$ is a π-system on Ω which generates $\sigma(X_1)$ (see Exercise IV.2). Similarly $\mathscr{I}_2 = X_2^{-1}\big(\mathscr{J}(\mathbb{R})\big)$ is a π-system on Ω which generates $\sigma(X_2)$. The assumption (V.8) exactly says that $\mathsf{P}[A_1 \cap A_2] = \mathsf{P}[A_1]\ \mathsf{P}[A_2]$ for all $A_1 \in \mathscr{I}_1$ and $A_2 \in \mathscr{I}_2$. The statement therefore follows from Proposition V.5 □

Similar statements hold for finite collections of π-systems and finite collections of random numbers. To indicate how to deal with more than two σ-algebras, consider the following exercise.

Exercise V.6 (Independence of three sigma algebras)**.**
Let $\mathscr{G}_1, \mathscr{G}_2, \mathscr{G}_3 \subset \mathscr{F}$ be sigma-algebras on Ω. Assume that $\mathscr{G}_k$ is generated by a π-system $\mathscr{I}_k$ which contains Ω.

(a) Show that $\mathscr{G}_1, \mathscr{G}_2, \mathscr{G}_3$ are independent if and only if

$$\mathsf{P}\big[I_1 \cap I_2 \cap I_3\big] = \mathsf{P}[I_1]\ \mathsf{P}[I_2]\ \mathsf{P}[I_3]$$

for all $I_1 \in \mathscr{I}_1$, $I_2 \in \mathscr{I}_2$, $I_3 \in \mathscr{I}_3$.

(b) Why did we here require that $\mathscr{I}_k$ contains Ω?
Hint: *Consider, e.g., the finite set $\Omega = \{1, 2, \ldots, 8\}$ and the discrete uniform probability measure P on it, and three π-systems $\mathscr{I}_1$, $\mathscr{I}_2$, $\mathscr{I}_3$ each consisting of a single event, suitably chosen.*

V.3. Borel – Cantelli lemmas

Given a sequence $E_1, E_2, \ldots \in \mathscr{F}$ of events we occasionally care about whether infinitely many of them occur, or whether only finitely many of them occur. Borel – Cantelli lemmas are examples of what are called 0-1 laws in probability theory — they state that under some mild conditions that are often relatively simple to verify, the probability that we are interested in is trivial: either 0 or 1.

Occurrence infinitely often and occurrence eventually

We use the following definitions for a sequence $E_1, E_2, \ldots \in \mathscr{F}$ of events:

- *"E_n occurs infinitely often"*, abbreviated "E_n i.o.", is the event

$$\begin{aligned}\text{"}E_n \text{ i.o."} &= \bigcap_{m\in\mathbb{N}} \bigcup_{n\geq m} E_n \;=\; \limsup_n E_n \\ &= \big\{\omega\in\Omega \;\big|\; \omega\in E_n \text{ for infinitely many indices } n\big\}.\end{aligned}$$

- *"E_n occurs eventually"*, abbreviated "E_n ev.", is the event

$$\begin{aligned}\text{"}E_n \text{ ev."} &= \bigcup_{m\in\mathbb{N}} \bigcap_{n\geq m} E_n \;=\; \liminf_n E_n \\ &= \big\{\omega\in\Omega \;\big|\; \omega\in E_n \text{ for all except finitely many } n\big\}.\end{aligned}$$

These two are related by taking complements: by De Morgan's laws, we have

$$\Big(\bigcap_{m\in\mathbb{N}} \bigcup_{n\geq m} E_n\Big)^{\mathbf{c}} = \bigcup_{m\in\mathbb{N}} \bigcap_{n\geq m} E_n^{\mathbf{c}}, \qquad \text{i.e.,}$$

$$\Big(\text{"}E_n \text{ occurs infinitely often"}\Big)^{\mathbf{c}} = \text{"}E_n^{\mathbf{c}} \text{ occurs eventually"}.$$

Exercise V.7 (Indicators and upper and lower limits)**.**
Consider a sequence $E_1, E_2, \ldots \subset \Omega$ of subsets. Show that for all $\omega\in\Omega$ we have

$$\limsup_n \mathbb{I}_{E_n}(\omega) = \mathbb{I}_{\limsup E_n}(\omega) \qquad \text{and} \qquad \liminf_n \mathbb{I}_{E_n}(\omega) = \mathbb{I}_{\liminf E_n}(\omega).$$

Use this to show that

$$\liminf_n E_n \subset \limsup_n E_n.$$

The two Borel – Cantelli lemmas

The first Borel – Cantelli lemma says that whenever the probabilities of the events E_n decay fast enough, it is (almost surely) impossible for the events to occur infinitely often.

Lemma V.7 (Borel–Cantelli lemma: convergence part)**.**
Suppose that $E_1, E_2, \ldots \in \mathscr{F}$ are such that $\sum_{n=1}^{\infty} \mathsf{P}[E_n] < +\infty$. Then we have

$$\mathsf{P}\big[\text{"}E_n \text{ occurs infinitely often"}\big] = 0.$$

Proof. Denote $G_m = \bigcup_{n=m}^{\infty} E_n$, so that the event $E = \limsup_n E_n$ of interest is the decreasing limit $G_m \downarrow E$ as $m\to\infty$. By monotone convergence for probability measures, Theorem II.22, we have

$$\mathsf{P}\big[E\big] = \lim_{m\to\infty} \mathsf{P}\big[G_m\big].$$

Now we can use the union bound, Theorem II.20, to estimate

$$0 \leq \mathsf{P}\big[G_m\big] = \mathsf{P}\Big[\bigcup_{n=m}^{\infty} E_n\Big] \leq \sum_{n=m}^{\infty} \mathsf{P}[E_n].$$

Since the series $\sum_n \mathsf{P}[E_n]$ is convergent, its tail goes to zero: $\sum_{n=m}^{\infty} \mathsf{P}[E_n] \to 0$ as $m\to\infty$. This shows that $\lim_{m\to\infty} \mathsf{P}[G_m] = 0$ and thus $\mathsf{P}[E] = 0$. □

The second Borel–Cantelli lemma says that if the probabilities of the events E_n do not decay fast and if the events are in addition independent, then the events must (almost surely) occur infinitely often.

Lemma V.8 (Borel – Cantelli lemma: divergence part)**.**
Suppose that $E_1, E_2, \ldots \in \mathscr{F}$ are independent and $\sum_{n=1}^{\infty} \mathsf{P}[E_n] = +\infty$. Then we have

$$\mathsf{P}\big[\text{“}E_n \text{ occurs infinitely often”}\big] = 1.$$

Proof. Instead of the event $E = \bigcap_{m\in\mathbb{N}} \bigcup_{n\geq m} E_n$ of interest, consider first its complement $E^{\mathbf{c}} = \bigcup_{m\in\mathbb{N}} \bigcap_{n\geq m} E_n^{\mathbf{c}}$. Denote the members in this union by $F_m = \bigcap_{n\geq m} E_n^{\mathbf{c}}$. The event F_m is an intersection — but in order to use independence we first need to arrange things to finite intersections. To achieve this, define $F_{m,\ell} = \bigcap_{n=m}^{\ell} E_n^{\mathbf{c}}$ and note that these finite intersections decrease to the infinite intersection $F_{m,\ell} \downarrow F_m$ as $\ell \to \infty$. Now by independence we have

$$\mathsf{P}\big[F_{m,\ell}\big] = \mathsf{P}\Big[\bigcap_{n=m}^{\ell} E_n^{\mathbf{c}}\Big] = \prod_{n=m}^{\ell} \mathsf{P}\big[E_n^{\mathbf{c}}\big].$$

Denote $p_n = \mathsf{P}[E_n]$ and use the estimate $1 - p_n \leq e^{-p_n}$ to get

$$\mathsf{P}\big[F_{m,\ell}\big] = \prod_{n=m}^{\ell} (1 - p_n) \leq \prod_{n=m}^{\ell} e^{-p_n} = \exp\Big(-\sum_{n=m}^{\ell} p_n\Big).$$

By the assumption $\sum_n p_n = +\infty$ we have $\exp\big(-\sum_{n=m}^{\ell} p_n\big) \to 0$ as $\ell \to \infty$. Therefore from $F_{m,\ell} \downarrow F_m$ and monotone convergence for probability measures, Theorem II.22, we get

$$\mathsf{P}\big[F_m\big] = \lim_{\ell\to\infty} \mathsf{P}\big[F_{m,\ell}\big] = 0.$$

Now for the complement $E^{\mathbf{c}} = \bigcup_m F_m$ use the union bound, Theorem II.20, to get

$$\mathsf{P}\big[E^{\mathbf{c}}\big] = \mathsf{P}\Big[\bigcup_{m=1}^{\infty} F_m\Big] \leq \sum_{m=1}^{\infty} \mathsf{P}\big[F_m\big] = \sum_{m=1}^{\infty} 0 = 0.$$

This result for the complement is what we wanted to show about E:

$$\mathsf{P}\big[E\big] = 1 - \mathsf{P}\big[E^{\mathbf{c}}\big] = 1 - 0 = 1.$$

□

Exercise V.8 (Independence assumption is needed in Lemma V.8)**.**
Find a sequence $E_1, E_2, \ldots \in \mathscr{F}$ of events (do not try independent) such that we have $\sum_{n=1}^{\infty} \mathsf{P}[E_n] = +\infty$, but $\mathsf{P}\big[\text{“}E_n \text{ occurs infinitely often”}\big] \neq 1$.

Example V.9 (Records)**.**
Consider an annual sports contest, in which one keeps track of the winner's scores for different years as well as the record score of all past years.

Suppose that the winner's score for year n is a real-valued random variable X_n, and suppose that $X_1, X_2, \ldots$ are independent and identically distributed — and moreover that the cumulative distribution function of X_n is continuous. (These assumptions are not completely unreasonable!)

Consider the event

$$E_n = \Big\{X_n > \max\{X_1, \ldots, X_{n-1}\}\Big\}$$

that a new record is made in the contest of year n. We leave it as an exercise to the reader to prove that $\mathsf{P}[E_n] = \frac{1}{n}$ and that $E_1, E_2, \ldots$ are independent.

Then, since the harmonic series diverges $\sum_n \mathsf{P}[E_n] = \sum_n \frac{1}{n} = +\infty$, the divergence part of Borel – Cantelli lemmas implies that $\mathsf{P}[E_n \text{ infinitely often}] = 1$. In other words: almost surely new records are made infinitely many times.

Consider then the event $F_n = E_n \cap E_{n+1}$ that records are broken in the consecutive years n and $n+1$. By what we know of $E_1, E_2, \ldots$, we get $\mathsf{P}[F_n] = \frac{1}{n(n+1)}$, and thus $\sum_n \mathsf{P}[F_n] = \sum_n \frac{1}{n^2+n} < +\infty$. Now the convergence part of Borel – Cantelli lemmas implies that $\mathsf{P}[F_n \text{ infinitely often}] = 0$. In other words: almost surely there are only finitely many times that new records are made in consecutive years.

Exercise V.9 (Decimal digits of a uniform random number)**.**
Let P be the uniform probability measure on the unit interval $[0,1]$ (cf. Example II.12). Consider the digits $D_k(\omega)$, $k \in \mathbb{N}$, of the decimal representation of a number $\omega \in [0,1]$, so that $\omega = \sum_{k=1}^{\infty} D_k(\omega)\, 10^{-k}$. For this exercise you can consider it known (*think about it anyway!*) that the digits $D_1, D_2, \ldots$ are simple random variables (on the sample space $\Omega = [0,1]$), and that they are independent.

Let Z_k be the run length of zeroes starting from the k:th digit:

- $Z_k = 0$ if the k:th digit is not zero, $D_k \neq 0$;
- $Z_k = m$ if $D_k = D_{k+1} = \cdots = D_{k+m-1} = 0$ and $D_{k+m} \neq 0$.

(a) Show that $\mathsf{P}\big[Z_k = m\big] = \frac{9}{10^{m+1}}$ for all $m \in \mathbb{Z}_{\geq 0}$.
(b) Fix $m \in \mathbb{Z}_{\geq 0}$. Show that $\mathsf{P}\big[Z_k = m \text{ for infinitely many } k\big] = 1$.
(c) Show that $\mathsf{P}\big[Z_k = k \text{ for infinitely many } k\big] = 0$.

Lecture VI

Events of the infinite horizon

Consider a sequence of random variables. The topic of this lecture is:

> What information about the sequence is not sensitive to the values of any finite number of individual members of the sequence?
>
> Which events can be decided and which random variables are determined by such information?

Although it might at first appear surprising, there are in fact many interesting events and random variables which are not affected by any finite number of individual values. We will give a number of examples.

We will moreover return to profound consequences of independence, and ask:

> Assuming moreover that the sequence of random variables is independent, what can be said about the probabilities of events that are not sensitive to any finite number of individual values?

Pertaining to the last question, we will prove Kolmogorov's 0-1 law which states that under the independence assumption, any event which is not sensitive to finitely many individual values has probability either zero or one, and any random variable which is not affected by finitely many individual values is almost surely constant. This probabilistic fact underlies some surprising phenomena, in particular related to phase transitions in physical systems.

VI.1. Definition of the tail σ-algebra

Throughout, let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space.

Let $X_1, X_2, X_3, \ldots$ be a sequence of random variables. Denote first of all by

$$\mathscr{X} := \sigma(X_1, X_2, \ldots) = \sigma\big((X_n)_{n\in\mathbb{N}}\big) \tag{VI.1}$$

the σ-algebra generated by the collection $(X_n)_{n\in\mathbb{N}}$ of random variables constituting the sequence (recall Definition IV.1). It should be interpreted as the information contained in the entire sequence.

We will also consider information contained in various subsets of the collection of random variables in the sequence. Specifically,

$$\mathscr{X}_k := \sigma(X_1, X_2, \ldots, X_k) = \sigma\big((X_n)_{n=1,\ldots,k}\big) \tag{VI.2}$$

describes the information contained in the in the first k members of the sequence, and

$$\mathscr{T}_k := \sigma(X_{k+1}, X_{k+2}, \ldots) = \sigma\big((X_n)_{n>k}\big) \tag{VI.3}$$

describes the information contained in the sequence without its first k members. In other words, $\mathscr{T}_k$ is the information in the sequence which is not affected by the first k values.

Our main interest lies in

$$\mathscr{T}_\infty := \bigcap_{k\in\mathbb{N}} \mathscr{T}_k. \tag{VI.4}$$

As an intersection of the σ-algebras $\mathscr{T}_k$, also $\mathscr{T}_\infty$ is a σ-algebra by Lemma I.8. It is called the *tail σ-algebra* of the sequence $X_1, X_2, \ldots$. Since $\mathscr{T}_\infty$ contains less information than any $\mathscr{T}_k$, $k \in \mathbb{N}$, it describes the information which is not affected by any finite number of changes in the sequence. The corresponding events $E \in \mathscr{T}_\infty$ are called *tail events*.

We start with examples of all of these different σ-algebras. For the examples, it is important to keep in mind that information in the first k members increases with k,

$$\mathscr{X}_1 \subset \mathscr{X}_2 \subset \mathscr{X}_3 \subset \cdots \subset \mathscr{X}_k \subset \mathscr{X}_{k+1} \subset \cdots \subset \mathscr{X},$$

whereas information without the first k members decreases with k

$$\mathscr{T}_\infty \subset \cdots \subset \mathscr{T}_{k+1} \subset \mathscr{T}_k \subset \cdots \subset \mathscr{T}_3 \subset \mathscr{T}_2 \subset \mathscr{T}_1 \subset \mathscr{X},$$

and of course all of these are contained in the information $\mathscr{X}$ about the entire sequence (which in turn is contained in complete information $\mathscr{F}$ on our probability space).

Example VI.1 (Examples of events and random variables related to a sequence)**.**
Suppose for concreteness that the sequence consists of real valued random variables, i.e.,

$$X_1, X_2, \ldots \in \mathrm{m}\mathscr{F}$$

are $\mathscr{F}$-measurable functions $\Omega \to \mathbb{R}$.

As a warm-up, note that by definition of $\mathscr{X}_n = \sigma(X_1, \ldots, X_n)$, for any $n \le \ell$ we have

$$X_n \in \mathrm{m}\mathscr{X}_n \subset \mathrm{m}\mathscr{X}_\ell \subset \mathrm{m}\mathscr{X}.$$

Because linear combinations of measurable functions are measurable by Proposition III.13, we get thus, for example, the following measurability of (finite) averages

$$\frac{X_1 + \cdots + X_\ell}{\ell} \in \mathrm{m}\mathscr{X}_\ell \subset \mathrm{m}\mathscr{X}. \tag{VI.5}$$

Also because upper and lower limits of measurable functions are measurable (Proposition III.14), we get the following

$$\liminf_n X_n \in \mathrm{m}\mathscr{X} \qquad \text{and} \qquad \limsup_n X_n \in \mathrm{m}\mathscr{X}. \tag{VI.6}$$

We claim that the random variables (VI.6) are, in fact, measurable even with respect to the tail σ-algebra $\mathscr{T}_\infty$. For this, observe first that we can write, for any $k \in \mathbb{N}$,

$$\liminf_n X_n = \sup_{m\in\mathbb{N}} \Big(\inf_{n\ge m} X_n \Big) = \sup_{m>k} \Big(\inf_{n\ge m} X_n \Big),$$

because $\inf_{n\ge m} X_n$ is increasing in m (infumum over smaller collections), so omitting finitely terms corresponding to $m = 1, 2, \ldots, k$ does not affect the supremum. Now $\inf_{n\ge m} X_n$ is the infimum of the countable collection of random variables $(X_n)_{n\ge m}$, which are by construction measurable with respect to $\mathscr{T}_k = \sigma(X_{k+1}, X_{k+2}, \ldots)$ when $m > k$. Proposition III.14 then first of all implies $\inf_{n\ge m} X_n \in \mathrm{m}\mathscr{T}_k$ for $m > k$, and the same proposition applied again to the supremum of these gives $\liminf_n X_n = \sup_{m>k} \big(\inf_{n\ge m} X_n \big) \in \mathrm{m}\mathscr{T}_k$. Since this holds for any k, and the tail σ-algebra is defined as the intersection $\mathscr{T}_\infty := \bigcap_k \mathscr{T}_k$, we get $\liminf_n X_n \in \mathrm{m}\mathscr{T}_\infty$. The limsup is handled similarly, and we conclude

$$\liminf_n X_n \in \mathrm{m}\mathscr{T}_\infty \qquad \text{and} \qquad \limsup_n X_n \in \mathrm{m}\mathscr{T}_\infty. \tag{VI.7}$$

Provided that the limit $\lim_{n\to\infty} X_n$ exists, it coincides with these, and is also $\mathscr{T}_\infty$-measurable.

We can use the above observations to check that the event of existence of the limit

$$E = \Big\{\omega \in \Omega \;\Big|\; \text{the limit } \lim_{n\to\infty} X_n(\omega) \text{ exists}\Big\}$$

is a tail event, $E \in \mathscr{T}_\infty$.[1] Indeed, the limit $\lim_{n\to\infty} X_n$ does not exist if and only if the upper and lower limits are different, which happens if and only if we can find some rational number $q \in \mathbb{Q}$ such that $\liminf_n X_n < q < \limsup_n X_n$. The event E of existence of limit is therefore

$$E = \bigg(\bigcup_{q\in\mathbb{Q}} \Big(\big\{ \liminf_n X_n < q \big\} \cap \big\{ \limsup_n X_n > q \big\} \Big) \bigg)^{\mathbf{c}}, \tag{VI.8}$$

where we are only using finite intersections, countable unions, and complements starting from the sets $\big\{ \liminf_n X_n < q \big\}$ and $\big\{ \limsup_n X_n > q \big\}$, which belong to the tail σ-algebra $\mathscr{T}_\infty$ by (VI.7). Thus we indeed have $E \in \mathscr{T}_\infty$.

A slightly less obvious example concerns the limit of averages (VI.5) as the number ℓ of terms tends to infinity,

$$\lim_{\ell\to\infty} \frac{X_1 + \cdots + X_\ell}{\ell}.$$

The limit may or may not exist, so let us instead consider, e.g., $\limsup_{\ell\to\infty} \frac{X_1+\cdots+X_\ell}{\ell}$. The key observation is that for fixed k, we have

$$\lim_{\ell\to\infty} \frac{X_1 + \cdots + X_k}{\ell} = 0$$

(the numerator remains constant as $\ell \to \infty$ while the denominator tends to infinity). Using this, we get

$$\begin{aligned} \limsup_{\ell\to\infty} \frac{X_1 + \cdots + X_\ell}{\ell} &= \limsup_{\ell\to\infty} \frac{(X_1 + \cdots + X_k) + (X_{k+1} + \cdots + X_\ell)}{\ell} \\ &= 0 + \limsup_{\ell\to\infty} \frac{X_{k+1} + \cdots + X_\ell}{\ell}. \end{aligned}$$

The random variables $(X_n)_{n>k}$ are by construction measurable with respect to $\mathscr{T}_k = \sigma(X_{k+1}, X_{k+2}, \ldots)$, so their linear combination $\frac{X_{k+1}+\cdots+X_\ell}{\ell}$ is also $\mathscr{T}_k$-measurable according to Proposition III.13. By Proposition III.14 we then deduce that that $\limsup_{\ell\to\infty} \frac{X_1+\cdots+X_\ell}{\ell} = \limsup_{\ell\to\infty} \frac{X_{k+1}+\cdots+X_\ell}{\ell}$ is $\mathscr{T}_k$-measurable. Since this holds for any k, and the tail σ-algebra is the intersection $\mathscr{T}_\infty := \bigcap_k \mathscr{T}_k$, we conclude

$$\limsup_{\ell} \frac{X_1 + \cdots + X_\ell}{\ell} \in \mathrm{m}\mathscr{T}_\infty.$$

Like before, it follows that the existence of the limit of averages is a tail event

$$\left\{\omega \in \Omega \;\middle|\; \text{the limit } \lim_{\ell\to\infty} \frac{X_1(\omega) + \cdots + X_\ell(\omega)}{\ell} \text{ exists}\right\} \in \mathscr{T}_\infty,$$

and provided the limit exists, it is measurable with respect to the tail σ-algebra $\mathscr{T}_\infty$.

In particular, the example shows that there are some interesting events in the tail σ-algebra $\mathscr{T}_\infty$. The following exercise contains a few more examples.

Exercise VI.1 (Tail events)**.**

Let $X_1, X_2, X_3, \ldots$ be a sequence of random numbers (i.e., $\mathbb{R}$-valued random variables)

[1] If the sequence $X_1, X_2, \ldots$ is bounded, then there is a slick way to check this. Form the difference $D = \limsup_n X_n - \liminf_n X_n$, which is $\mathscr{T}_\infty$-measurable as a linear combination (use Proposition III.13). The limit exists if and only if the difference is zero, $D = 0$. In other words, the event E is the preimage $E = D^{-1}(\{0\}) \in \mathscr{T}_\infty$ (since $\{0\} \in \mathscr{B}$ and $D \in \mathrm{m}\mathscr{T}_\infty$).

defined on a probability space $(\Omega, \mathscr{F}, \mathsf{P})$. Investigate which of the following events belong to the tail σ-algebra $\mathscr{T}_\infty$.

(a): $\Big\{\omega \in \Omega \;\Big|\; \text{the series } \sum_{n=1}^{\infty} X_n(\omega) \text{ converges}\Big\}$

(b): $\Big\{\omega \in \Omega \;\Big|\; \sum_{n=1}^{\infty} X_n(\omega) \le -42\Big\}$

(c): $\Big\{\omega \in \Omega \;\Big|\; \text{the sequence } X_1, X_2, \dots \text{ is bounded}\Big\}$

(d): $\Big\{\omega \in \Omega \;\Big|\; \forall\, \ell \in \mathbb{N}\ \exists n \in \mathbb{N} \text{ such that } X_n(\omega) = X_{n+1}(\omega) = \cdots = X_{n+\ell}(\omega)\Big\}$
$= \{\text{there exists arbitrarily long repetitions in the sequence } X_1, X_2, \dots\}$

VI.2. Kolmogorov's 0-1 law

If we assume that the sequence is independent, we get a remarkable result about the probabilities of tail events.

Theorem VI.2 (Kolmogorov's 0-1 law)**.**
Suppose that $X_1, X_2, \dots$ is a sequence of independent random variables. Then the following hold:

(a) *For any tail event $E \in \mathscr{T}_\infty$ we have either*
$$\mathsf{P}[E] = 1 \qquad \text{or} \qquad \mathsf{P}[E] = 0.$$

(b) *Any $\mathbb{R}$-valued random variable T which is measurable with respect to the tail σ-algebra $\mathscr{T}_\infty$ is almost surely constant, i.e., for some $c \in \mathbb{R}$ we have*
$$\mathsf{P}\big[T = c\big] = 1.$$

The proof strategy is almost as surprising as the statement: we will show that the information in the tail is independent of itself, $\mathscr{T}_\infty \perp\!\!\!\perp \mathscr{T}_\infty$.

Proof of Theorem VI.2. We will show that $\mathscr{T}_\infty \perp\!\!\!\perp \mathscr{T}_\infty$. The assertions (a) and (b) can then be concluded rather easily as follows.

For (a), suppose that $E \in \mathscr{T}_\infty$. Then by the independence property $\mathscr{T}_\infty \perp\!\!\!\perp \mathscr{T}_\infty$ we have
$$\mathsf{P}\big[E \cap E\big] = \mathsf{P}[E]\ \mathsf{P}[E].$$
But in view of the obvious fact $E \cap E = E$, this gives $\mathsf{P}[E] = \mathsf{P}[E]^2$. The only two solutions of this quadratic equation are 0 and 1, so we conclude $\mathsf{P}[E] \in \{0, 1\}$.

For (b), suppose that $T \in \mathrm{m}\mathscr{T}_\infty$. Then for any $x \in \mathbb{R}$ we have
$$\{T \le x\} = T^{-1}\big((-\infty, x]\big) \in \mathscr{T}_\infty,$$
so by part (a) we have
$$\mathsf{P}\big[T \le x\big] \in \{0, 1\}\,.$$
This shows that the cumulative distribution function of T never takes any other values except 0 and 1. If we define
$$c := \inf\big\{x \in \mathbb{R} \;\big|\; \mathsf{P}[T \le x] = 1\big\}\,,$$
then by right continuity of cumulative distribution functions (Proposition II.30) we have $\mathsf{P}[T \le c] = 1$ and $\mathsf{P}[T < c] = 0$. This gives
$$\mathsf{P}\big[T = c\big] = \mathsf{P}\big[T \le c\big] - \mathsf{P}\big[T < c\big] = 1,$$

i.e., $T = c$ almost surely as claimed in (b).

The proof will thus be complete once we have shown that $\mathscr{T}_\infty \perp\!\!\!\perp \mathscr{T}_\infty$. We will do this in a number of steps.

step 1, $\mathscr{X}_k \perp\!\!\!\perp \mathscr{T}_k$: We claim that the two σ-algebras $\mathscr{X}_k$ and $\mathscr{T}_k$ are independent. We will verify their independence using two π-systems which generate these σ-algebras.

Let $\mathscr{J}$ denote the collection of events of the form

$$A = \Big\{\omega \in \Omega \;\Big|\; X_1(\omega) \in B_1, X_2(\omega) \in B_2, \dots, X_k(\omega) \in B_k\Big\} \subset \Omega, \qquad \text{(VI.9)}$$
$$\text{where } B_1, B_2, \dots, B_k \in \mathscr{B}.$$

Then $\mathscr{J}$ is a π-system, which generates the first of our σ-algebras, $\sigma(\mathscr{J}) = \mathscr{X}_k$ (we leave it as an exercise[2] to the reader to verify this).

Similarly, let $\mathscr{J}'$ denote the collection of events of the form

$$A' = \Big\{\omega \in \Omega \;\Big|\; X_{k+1}(\omega) \in B_{k+1}, X_{k+2}(\omega) \in B_{k+2}, \dots, X_{k+r}(\omega) \in B_{k+r}\Big\} \subset \Omega, \qquad \text{(VI.10)}$$
$$\text{where } r \in \mathbb{N} \text{ and } B_{k+1}, B_{k+2}, \dots, B_{k+r} \in \mathscr{B}.$$

Then $\mathscr{J}'$ is a π-system, which generates the second of our σ-algebras, $\sigma(\mathscr{J}') = \mathscr{T}_k$ (we again leave it as an exercise to the reader to verify this).

Consider now $A \in \mathscr{J}$ and $A' \in \mathscr{J}'$ as in (VI.9) and (VI.10). Note that by independence of $X_1, X_2, \dots$ we have

$$\mathsf{P}\big[A\big] = \mathsf{P}\Big[X_1 \in B_1, \dots, X_k \in B_k\Big] = \prod_{j=1}^{k} \mathsf{P}\big[X_j \in B_j\big]$$

$$\mathsf{P}\big[A'\big] = \mathsf{P}\Big[X_{k+1} \in B_{k+1}, \dots, X_{k+r} \in B_{k+r}\Big] = \prod_{j=k+1}^{k+r} \mathsf{P}\big[X_j \in B_j\big].$$

The probability of the intersection $A \cap A'$ is similarly computed using independence,

$$\begin{aligned}\mathsf{P}\big[A \cap A'\big] &= \mathsf{P}\Big[X_1 \in B_1, \dots, X_k \in B_k, X_{k+1} \in B_{k+1}, \dots, X_{k+r} \in B_{k+r}\Big] \\ &= \prod_{j=1}^{k+r} \mathsf{P}\big[X_j \in B_j\big].\end{aligned}$$

These expressions show that $\mathsf{P}\big[A \cap A'\big] = \mathsf{P}\big[A\big]\ \mathsf{P}\big[A'\big]$. Since the π-systems $\mathscr{J}$ and $\mathscr{J}'$ generate the σ-algebras $\mathscr{X}_k$ and $\mathscr{T}_k$, it follows from Proposition V.5 that $\mathscr{X}_k$ and $\mathscr{T}_k$ are independent.

step 2, $\mathscr{X}_k \perp\!\!\!\perp \mathscr{T}_\infty$: In step 1 we showed $\mathscr{X}_k \perp\!\!\!\perp \mathscr{T}_k$. But since we have $\mathscr{T}_\infty \subset \mathscr{T}_k$, we a fortiori have also $\mathscr{X}_k \perp\!\!\!\perp \mathscr{T}_\infty$ (there are now fewer sets for which the condition of Definition V.1 needs to be verified!).

step 3, $\mathscr{X} \perp\!\!\!\perp \mathscr{T}_\infty$: We claim that the two σ-algebras $\mathscr{X}$ and $\mathscr{T}_\infty$ are independent. We will again verify their independence using π-systems.

Consider the collection $\mathscr{U} = \bigcup_{k\in\mathbb{N}} \mathscr{X}_k$. We claim that $\mathscr{U}$ is a π-system which generates $\mathscr{X}$ (note that $\mathscr{U}$ is the union of σ-algebras, but $\mathscr{U}$ is generally not a σ-algebra itself).

To check that $\mathscr{U}$ is a π-system, suppose that $A_1, A_2 \in \mathscr{U}$. Then since $\mathscr{U}$ is defined as the union of $\mathscr{X}_k$, $k \in \mathbb{N}$, we have $A_1 \in \mathscr{X}_{k_1}$ and $A_2 \in \mathscr{X}_{k_2}$ for some $k_1, k_2 \in \mathbb{N}$. Recall that the sequence of σ-algebras in the union is increasing, $\mathscr{X}_1 \subset \mathscr{X}_2 \subset \cdots$. Therefore by setting $k_0 = \max\{k_1, k_2\}$, we have $A_1, A_2 \in \mathscr{X}_{k_0}$. Since $\mathscr{X}_{k_0}$ is itself a σ-algebra, we thus have also $A_1 \cap A_2 \in \mathscr{X}_{k_0}$. Finally, since we have $\mathscr{X}_{k_0} \subset \bigcup_{k\in\mathbb{N}} \mathscr{X}_k = \mathscr{U}$, we get $A_1 \cap A_2 \in \mathscr{U}$, which shows that $\mathscr{U}$ is a π-system.

[2] As a hint for this exercise, it is worth noting that step 3 contains the details of a similar argument in a slightly more complicated situation.

Let us then check that $\mathscr{U}$ generates $\mathscr{X}$, i.e., $\sigma(\mathscr{U}) = \mathscr{X}$. We do this by showing inclusions in both directions. Note first that for any $n \in \mathbb{N}$ we have $X_n \in \mathrm{m}\mathscr{X}_n$ directly from the definition, and since $\mathscr{X}_n \subset \bigcup_{k\in\mathbb{N}} \mathscr{X}_k = \mathscr{U} \subset \sigma(\mathscr{U})$, we also have $X_n \in \mathrm{m}\sigma(\mathscr{U})$. Thus $\sigma(\mathscr{U})$ is a σ-algebra with respect to which each X_n is measurable, and by definition $\mathscr{X} = \sigma\big((X_k)_{k\in\mathbb{N}}\big)$ is the smallest such σ-algebra, so we get $\mathscr{X} \subset \sigma(\mathscr{U})$. On the other hand, for each $k \in \mathbb{N}$ we have $\mathscr{X}_k \subset \mathscr{X}$ so also $\mathscr{U} \subset \mathscr{X}$, and since $\sigma(\mathscr{U})$ is the smallest σ-algebra with this property, we get $\sigma(\mathscr{U}) \subset \mathscr{X}$. These two inclusions give the equality $\sigma(\mathscr{U}) = \mathscr{X}$.

Now we know that $\mathscr{U}$ is a π-system which generates $\mathscr{X}$. We use $\mathscr{T}_\infty$ itself as a π-system which generates $\mathscr{T}_\infty$, i.e., $\sigma(\mathscr{T}_\infty) = \mathscr{T}_\infty$ (recall that any σ-algebra is also a π-system). Let $A \in \mathscr{U}$ and $E \in \mathscr{T}_\infty$. We have $A \in \mathscr{X}_k$ for some $k \in \mathbb{N}$, because of the definition of $\mathscr{U}$ as a union. In step 2 we showed that $\mathscr{X}_k \perp\!\!\!\perp \mathscr{T}_\infty$, so we get $\mathsf{P}[A \cap E] = \mathsf{P}[A]\,\mathsf{P}[E]$. It therefore follows from Proposition V.5 that $\mathscr{X}$ and $\mathscr{T}_\infty$ are independent.

final step, $\mathscr{T}_\infty \perp\!\!\!\perp \mathscr{T}_\infty$: In step 3 we showed $\mathscr{X} \perp\!\!\!\perp \mathscr{T}_\infty$. But since we have $\mathscr{T}_\infty \subset \mathscr{X}$, we a fortiori have also $\mathscr{T}_\infty \perp\!\!\!\perp \mathscr{T}_\infty$ (there is less to verify!). This concludes the proof. □

Part (b) of Kolmogorov's 0-1 law is stated in Theorem above for real-valued random variables measurable with respect to the tail σ-algebra $\mathscr{T}_\infty$. This choice was made for the sake of concreteness — you may verify that the same conclusion holds much more generally.

Exercise VI.2. Let $\mathfrak{X}$ be a complete separable metric space. Assume that a random variable $T\colon \Omega \to \mathfrak{X}$ is $\mathscr{T}_\infty/\mathscr{B}(\mathfrak{X})$-measurable, where $\mathscr{T}_\infty$ is the tail σ-algebra of a sequence of independent random variables $X_1, X_2, \ldots$. Prove that T is almost surely constant, i.e., there exists a point $x \in \mathfrak{X}$ such that $\mathsf{P}[T = x] = 1$.

A few interesting examples

Various random series

Example VI.3 (Random series with independent terms)**.**
Suppose that $X_1, X_2, \ldots$ are independent $\mathbb{R}$-valued random variables. Consider the random series formed from the sequence,

$$\sum_{n=1}^{\infty} X_n.$$

Then the event that this series converges,

$$E = \left\{ \omega \in \Omega \;\middle|\; \sum_{n=1}^{\infty} X_n(\omega) \text{ converges} \right\},$$

belongs to the tail σ-algebra, $E \in \mathscr{T}_\infty$. Since the terms X_n were assumed independent, Kolmogorov's 0-1 law states that $\mathsf{P}[E] \in \{0, 1\}$, i.e., either the series $\sum_{n=1}^{\infty} X_n$ converges almost surely or it diverges almost surely. The theorem does not tell which of these two extremes occurs.

The following example is a simple but interesting special case of random series in Example VI.3. It in particular shows that both of the two extremes in the example are indeed possible.

Example VI.4 (Series with randomly assigned signs)**.**
Let $a_1, a_2, \ldots \geq 0$ be a sequence of non-negative real numbers and let $S_1, S_2, \ldots$ be a sequence of independent and identically distributed $\{\pm 1\}$-valued random variables with

$$\mathsf{P}\big[S_n = +1\big] = \frac{1}{2} \qquad \text{and} \qquad \mathsf{P}\big[S_n = -1\big] = \frac{1}{2}.$$

Consider the following series of a_n's with random signs ($S_n = \pm 1$),

$$\sum_{n=1}^{\infty} a_n S_n. \tag{VI.11}$$

This is a series of the type considered in Example VI.3, with terms $X_n = a_n S_n$. Therefore the series (VI.11) either converges almost surely or diverges almost surely. It depends on the given sequence $(a_n)_{n\in\mathbb{N}}$, which of these happens.

Note first that the series (VI.11) can only converge if its terms at least tend to zero. The absolute values of the terms are $|a_n S_n| = a_n$, so the terms tend to zero if and only if the given sequence $(a_n)_{n\in\mathbb{N}}$ satisfies $\lim_{n\to\infty} a_n = 0$. In particular, if we take $a_n = 1$ for all n (or anything else which does not tend to zero), then series (VI.11) diverges almost surely.

On the other hand, if the series $\sum_{n=1}^{\infty} a_n$ converges, then the series (VI.11) is always absolutely convergent, whatever the random signs. In particular, if we take $a_n = n^{-2}$ for $n \in \mathbb{N}$ (or anything else which constitutes a convergent series), then series (VI.11) converges almost surely.

The interesting borderline cases for random series of type (VI.11) are when $a_n \to 0$ as $n \to \infty$, but we have $\sum_{n=1}^{\infty} a_n = \infty$. An example of this is $a_n = n^{-1}$ for $n \in \mathbb{N}$. Then Kolmogorov's 0-1 law still says that we either have almost sure convergence or divergence, but which one is it now?

Another variant of random series is power series with random coefficients. These also have important applications, although we will only use them for the purpose of examples.

Example VI.5 (Random power series)**.**
Let $Y_0, Y_1, Y_2, \ldots$ be a sequence of independent random variables. The power series

$$F(z) = \sum_{n=0}^{\infty} Y_n z^n \tag{VI.12}$$

with these coefficients defines a random function F of a real (or complex) variable z. For any fixed value of $z \in \mathbb{R}$ (or $z \in \mathbb{C}$), the series (VI.12) is essentially a special case of Example VI.3, with terms $X_n = Y_n z^n$ (the only difference is that starting the indexing from $n = 0$ is more natural for power series).

We know from the theory of power series that the series (VI.12) has some radius of convergence R such that

(VI.12) is convergent for $|z| < R$ and (VI.12) is divergent for $|z| > R$.

Since the coefficients of the series are random, it seems that the radius of convergence R should be random, too. In fact, the Cauchy–Hadamard formula in the theory of power series tells us that R can be written in terms of the coefficients as

$$R = \frac{1}{\limsup_n \sqrt[n]{|Y_n|}}. \tag{VI.13}$$

Note that the function $y \mapsto \sqrt[n]{|y|}$ is continuous $\mathbb{R} \to \mathbb{R}$, and in particular Borel-measurable (by Corollary III.10), so $\sqrt[n]{|Y_n|}$ is a (measurable) random variable (by Proposition III.8). Then also the denominator, $\limsup_n \sqrt[n]{|Y_n|}$, is a random variable (by Proposition III.14). Finally, taking the reciprocal $s \mapsto \frac{1}{s}$ is continuous $[0, \infty] \to [0, \infty]$, so we see that the radius of convergence R of the random power series (VI.12) is indeed a random variable!

However, the radius of convergence R in (VI.13) is measurable with respect to the tail σ-algebra $\mathscr{T}_\infty$ (the limsup is insensitive to finite number of changes in the coefficients). Therefore, having assumed independence of the coefficients, we get from Kolmogorov's 0-1 law that R is almost surely equal to some constant c. In conclusion, the radius of convergence of this random power series is in fact essentially deterministic (non-random)!

Random walks

Example VI.6 (Escape probability of asymmetric simple random walk)**.**
Let $\theta \in [0,1]$ be a parameter, and let $X_1, X_2, \ldots$ be a sequence of independent and identically distributed $\{\pm 1\}$-valued random variables with

$$\mathsf{P}\big[X_n = +1\big] = \theta \qquad \text{and} \qquad \mathsf{P}\big[X_n = -1\big] = 1 - \theta.$$

The argument in Example VI.4 shows that the infinite series $\sum_{n=1}^{\infty} X_n$ diverges almost surely (we had $\theta = \frac{1}{2}$, but the same argument works for any θ). In this example we consider a different but related question.

We think of X_n as the n:th step of a random walker: $X_n = +1$ is interpreted as a step forward and $X_n = -1$ as a step backwards, and the parameter θ gives the probability of a forward step. If the walker starts from the origin, then her position after the first s steps is

$$W_s := \sum_{n=1}^{s} X_n.$$

Let us consider the question of whether the random walker eventually advances arbitrarily much: this is described by the event

$$\left\{ \omega \in \Omega \;\middle|\; \sup_s W_s(\omega) = +\infty \right\}$$

that the sequence $W_1, W_2, \ldots$ of walker's positions is not bounded from above. We leave it to the reader to check that this event $\{\sup_s W_s = +\infty\}$ belongs to the tail σ-algebra $\mathscr{T}_\infty$.[3] Having assumed the independence of the steps $X_1, X_2, \ldots$, it follows from Kolmogorov's 0-1 law that the probability of advancing arbitrarily much is either zero or one:

$$\mathsf{P}\Big[\sup_s W_s = +\infty\Big] \in \{0, 1\}.$$

In other words, either the walker almost surely advances arbitrarily much, or she almost surely does not. The theorem does not, however, tell which of these two options is true. The answer turns out to depend on the parameter θ: one can shown that we have

$$\mathsf{P}\big[\sup_s W_s = +\infty\big] = 1 \qquad \text{if and only if} \qquad \theta \geq \frac{1}{2}.$$

It is natural to refine the question of advancement of the walker slightly. By including also considerations of $\limsup_s W_s$ and $\liminf_s W_s$, the reader can show that also

$$\Big\{ \lim_{s\to\infty} W_s = +\infty \Big\} \qquad \text{and} \qquad \Big\{ \lim_{s\to\infty} W_s = -\infty \Big\}$$

are tail events. The former event describes the walker escaping towards $+\infty$, and the latter describes the the walker escaping towards $-\infty$. Concerning these, Kolmogorov's 0-1 law says that either the walker almost surely escapes towards $+\infty$ (resp. $-\infty$), or the probability of such escape is zero. Which is the case again depends on θ. One can shown that

$$\mathsf{P}\Big[\lim_{s\to\infty} W_s = +\infty \Big] = 1 \qquad \text{if and only if} \qquad \theta > \frac{1}{2}.$$

whereas symmetrically

$$\mathsf{P}\Big[\lim_{s\to\infty} W_s = -\infty \Big] = 1 \qquad \text{if and only if} \qquad \theta < \frac{1}{2}.$$

This seems to leave just one question: where on earth does that walker go if $\theta = \frac{1}{2}$?[4]

---

[3] It is important to note that we still talk about the tail σ-algebra of the sequence $X_1, X_2, \ldots$ of steps, not the tail σ-algebra of the sequence $W_1, W_2, \ldots$ of random walk positions.

[4] The answer is: all over the place! Namely, at $\theta = \frac{1}{2}$ we in fact have

$$\mathsf{P}\Big[\limsup_{s\to\infty} W_s = +\infty \;\text{ and }\; \liminf_{s\to\infty} W_s = -\infty \Big] = 1,$$

so the walker almost surely advances arbitrarily far in both directions and meanwhile returns to the origin infinitely often as well.

Lecture VII

Integration theory

Let $(S, \mathscr{S}, \mu)$ be a measure space. The goal of this lecture is to define, for all reasonable functions $f\colon S \to \mathbb{R}$, the integral

$$\int f \,\mathrm{d}\mu$$

of the function f with respect to the measure μ — denoted, when we want to emphasize the space S and the variable $s \in S$, also by

$$\int_S f(s) \,\mathrm{d}\mu(s).$$

This general theory of integration was originally conceived of by Henri Lebesgue. The theory is much more flexible than integration in the sense of Riemann, and it moreover leads to some remarkably powerful tools.

One of our main motivations for integration in probability theory is to get a precise mathematical definition of the expected value, which we will address in more detail in the next lecture (Lecture VIII). There are, however, also other equally important uses of the construction — special cases include:

Summation: Integration with respect to the counting measure (Example II.10) is just summation,

$$\int_S f(s) \,\mathrm{d}\mu_{\#}(s) = \sum_{s \in S} f(s).$$

From the general theory we develop for integrals, we will in particular get a general and precise definition of summation and many useful tools for calculating with sums.

Riemann integral and generalization: For real-valued functions f on the real line $\mathbb{R}$, integration with respect to the Lebesgue measure Λ (Example II.12) generalizes[1] the familiar Riemann integral. For this case, we therefore use also the very familiar notation

$$\int_{\mathbb{R}} f(x) \,\mathrm{d}\Lambda(x) = \int_{-\infty}^{\infty} f(x) \,\mathrm{d}x.$$

Expected value: The expected value $\mathsf{E}[X]$ of a real valued random variable X will be defined as the integral of the function $X\colon \Omega \to \mathbb{R}$ with respect to the probability measure P on $(\Omega, \mathscr{F})$,

$$\mathsf{E}\big[X\big] := \int_\Omega X(\omega) \,\mathrm{d}\mathsf{P}(\omega).$$

[1] To be precise, for any Riemann-integrable function $f\colon \mathbb{R} \to \mathbb{R}$ such that also $|f|$ is Riemann-integrable, the Riemann integral of f agrees with the integral of f with respect to the Lebesgue measure Λ. In principle both integrals are defined for somewhat more general functions f, but the Lebesgue integral is without a doubt the more fruitful generalization.

The construction of the integral will proceed by increasing the complexity of the allowed functions step by step:

(1) indicator functions
(2) non-negative simple functions
(3) non-negative measurable functions
(4) all integrable functions.

Many fundamental results about integration are likewise proved step by step — this proof strategy is affectionately referred to as the "*standard machine*".

For these steps, recall also the notations

$$\begin{aligned} \mathrm{m}\mathscr{S} &= \text{ the set of } \mathscr{S}\text{-measurable functions } S \to [-\infty, +\infty] \\ \mathrm{m}\mathscr{S}^+ &= \text{ the set of } \mathscr{S}\text{-measurable functions } S \to [0, +\infty] \\ \mathrm{s}\mathscr{S} &= \text{ the set of simple functions } S \to \mathbb{R} \\ \mathrm{s}\mathscr{S}^+ &= \text{ the set of simple functions } S \to [0, +\infty) \end{aligned}$$

In order to clearly distinguish between the (in principle) different definitions in each stage, during this lecture we use different notation for the integrals in the different stages as follows:

$\int^{\square}$ — the integral for non-negative simple functions (step 2)
$\int^{+}$ — the integral for non-negative measurable functions (step 3)
$\int$ — the integral for all integrable functions (step 4).

It will be shown in Lemmas VII.5 and VII.13, however, that the more general definitions coincide with the earlier ones, so in later lectures there will be no need to make these distinctions, and we can safely only use the notation $\int$.

Besides the construction, we prove some basic properties of the integral. In particular we will show *linearity*[2]

$$\begin{aligned} \int (cf)\,\mathrm{d}\mu &= c \int f\,\mathrm{d}\mu \\ \int (f_1 + f_2)\,\mathrm{d}\mu &= \int f_1\,\mathrm{d}\mu + \int f_2\,\mathrm{d}\mu \end{aligned}$$

and *monotonicity*

$$f \le g \qquad \Longrightarrow \qquad \int f\,\mathrm{d}\mu \le \int g\,\mathrm{d}\mu,$$

which in fact must be verified separately at each step of the construction, i.e., for the integrals $\int^{\square}$, $\int^{+}$, and $\int$.

Finally, in Section VII.4 we establish powerful general convergence theorems for integrals.

[2]In the end, the scalar c is allowed to be an arbitrary real number. During the steps which address integrals of non-negative functions, however, it will of course only be meaningful to allow non-negative scalars.

The approximation lemma

Especially when going from step (2) of the construction to step (3), i.e., from non-negative simple functions to all non-negative measurable functions, it will be important to keep in mind the following approximation result from Lecture III.

Lemma (Lemma III.18)**.** *For any non-negative measurable function*

$$f \in \mathrm{m}\mathscr{S}^+$$

there exists a sequence[3]

$$f_1, f_2, f_3, \ldots \in \mathrm{s}\mathscr{S}^+$$

of non-negative simple functions increasing pointwise to f, i.e.,

$$f_n(s) \uparrow f(s) \qquad \text{for all} \qquad s \in S.$$

VII.1. Integral for non-negative simple functions

Definition of integral of non-negative simple functions

For a subset $A \subset S$, the indicator function $\mathbb{I}_A$ takes the value 1 on A and 0 elsewhere,

$$\mathbb{I}_A(s) = \begin{cases} 1 & \text{if } s \in A \\ 0 & \text{if } s \notin A. \end{cases}$$

If $A \in \mathscr{S}$, then we want to define the integral of such a function simply as the measure of the set A,

$$\int^{\square} \mathbb{I}_A \, \mathrm{d}\mu := \mu[A]. \tag{VII.1}$$

This is step 1 in our definition of the integral. Note that already in this case it is possible for the integral to become infinite if $\mu[A] = +\infty$.

Any non-negative simple function $h \in \mathrm{s}\mathscr{S}^+$ can be written as

$$h = \sum_{j=1}^{n} a_j \, \mathbb{I}_{A_j},$$

where $a_1, \ldots, a_n$ are non-negative coefficients and $A_1, \ldots, A_n$ are measurable sets. To achieve linearity of the integral for simple functions, our only option is the following definition.

Definition VII.1 (Integral of a non-negative simple function)**.**
For $h = \sum_{j=1}^{n} a_j \, \mathbb{I}_{A_j}$ with $a_1, \ldots, a_n \in [0, +\infty)$ and $A_1, \ldots, A_n \in \mathscr{S}$, define

$$\int^{\square} h \, \mathrm{d}\mu := \sum_{j=1}^{n} a_j \, \mu[A_j]. \tag{VII.2}$$

[3] One concrete way of constructing the approximating sequence is to set $f_n = \varsigma_n \circ f$, where $\varsigma_n \colon [0, +\infty] \to [0, n]$ is the n:th staircase function with steps of size 2^{-n} and truncation at level n — see Equation (III.5) for definition and Figure III.2 for illustration.

A priori, this definition depends on the choices made in the expression of h as a linear combination of indicators, so we should verify well-definedness.

Exercise VII.1 (Well-definedness of the integral of non-negative simple functions)**.** Show that if $h \in \mathrm{s}\mathscr{S}^+$ can be written in two ways,

$$h = \sum_{j=1}^{n} a_j \, \mathbb{I}_{A_j} \qquad \text{and} \qquad h = \sum_{k=1}^{n'} a'_k \, \mathbb{I}_{A'_k},$$

as a non-negative linear combination of indicator functions, then we have

$$\sum_{j=1}^{n} a_j \, \mu[A_j] = \sum_{k=1}^{n'} a'_k \, \mu[A'_k].$$

Conclude that the integral of h is well-defined by (VII.2).

Properties of integral of non-negative simple functions

Let us then verify the basic properties of the integral defined in (VII.3). These will be used already in the next step.

Lemma VII.2 (Linearity of the integral for simple functions)**.**

(a) *If $h \in \mathrm{s}\mathscr{S}^+$ and $c \geq 0$, then $ch \in \mathrm{s}\mathscr{S}^+$ and*

$$\int^{\square} (ch) \, \mathrm{d}\mu = c \int^{\square} h \, \mathrm{d}\mu.$$

(b) *If $h, g \in \mathrm{s}\mathscr{S}^+$, then $h + g \in \mathrm{s}\mathscr{S}^+$ and*

$$\int^{\square} (h + g) \, \mathrm{d}\mu = \int^{\square} h \, \mathrm{d}\mu + \int^{\square} g \, \mathrm{d}\mu.$$

Proof. Both assertions are derived from Definition VII.1 relying on its well-definedness (Exercise VII.1), which permits us to use whichever decomposition to linear combination of indicators we find the most convenient.

proof of (a): Write $h \in \mathrm{s}\mathscr{S}^+$ as a linear combination of indicators with non-negative coefficients, $h = \sum_{j=1}^{n} a_j \, \mathbb{I}_{A_j}$, with $a_j \geq 0$ and $A_j \in \mathscr{S}$ for $j = 1, \ldots, n$. Then we have

$$ch = c \sum_{j=1}^{n} a_j \, \mathbb{I}_{A_j} = \sum_{j=1}^{n} c \, a_j \, \mathbb{I}_{A_j}.$$

Therefore, according to (VII.2), the integral is

$$\int^{\square} (ch) \, \mathrm{d}\mu = \sum_{j=1}^{n} c \, a_j \mu[A_j] = c \sum_{j=1}^{n} a_j \mu[A_j] = c \int^{\square} h \, \mathrm{d}\mu.$$

This proves the assertion (a).

proof of (b): Write $h \in \mathrm{s}\mathscr{S}^+$ and $g \in \mathrm{s}\mathscr{S}^+$ as linear combinations of indicators with non-negative coefficients, $h = \sum_{j=1}^{n} a_j \, \mathbb{I}_{A_j}$ and $g = \sum_{k=1}^{m} b_k \, \mathbb{I}_{B_k}$. Then we have

$$h + g = \sum_{j=1}^{n} a_j \, \mathbb{I}_{A_j} + \sum_{k=1}^{m} b_k \, \mathbb{I}_{B_k}.$$

Therefore, according to (VII.2), the integral is

$$\int^{\square} (h+g)\,\mathrm{d}\mu = \sum_{j=1}^{n} a_j\,\mu[A_j] + \sum_{k=1}^{m} b_k\,\mu[B_k] = \int^{\square} h\,\mathrm{d}\mu + \int^{\square} g\,\mathrm{d}\mu.$$

This proves the assertion (b). □

Lemma VII.3 (Monotonicity of the integral for simple functions)**.**
Suppose that $h, g \in \mathrm{s}\mathscr{S}^+$ and $h \le g$ pointwise. Then we have

$$\int^{\square} h\,\mathrm{d}\mu \le \int^{\square} g\,\mathrm{d}\mu.$$

Proof. We can first of all write

$$h = \sum_{j=1}^{n} a_j \mathbb{I}_{A_j}$$

with $a_1, \dots, a_n$ the finitely many different values of h and $A_j = h^{-1}(\{a_j\})$ the sets where this value is taken. Similarly we can write

$$g = \sum_{k=1}^{m} b_k \mathbb{I}_{B_k}$$

with $b_1, \dots, b_m$ the finitely many different values of g and $B_k = g^{-1}(\{b_k\})$. The intersections $A_j \cap B_k$ are the sets where simultaneously f takes the value a_j and g takes the value b_k — we can use them as refinements, and write the alternative expressions

$$h = \sum_{j=1}^{n}\sum_{k=1}^{m} a_j\,\mathbb{I}_{A_j\cap B_k} \qquad \text{and} \qquad g = \sum_{j=1}^{n}\sum_{k=1}^{m} b_k\,\mathbb{I}_{A_j\cap B_k},$$

which have the advantage that the same indicators appear in both. Observe that the assumption $h \le g$ implies that $a_j \le b_k$ whenever $A_j \cap B_k \neq \emptyset$. On the other hand, whenever $A_j \cap B_k = \emptyset$ we of course have $\mu[A_j \cap B_k] = 0$, so the desired conclusion

$$\int^{\square} h\,\mathrm{d}\mu = \sum_{j=1}^{n}\sum_{k=1}^{m} a_j\,\mu[A_j\cap B_k] \le \sum_{j=1}^{n}\sum_{k=1}^{m} b_k\,\mu[A_j\cap B_k] = \int^{\square} g\,\mathrm{d}\mu.$$

follows by comparing the non-zero terms in the sums. □

VII.2. Integral for non-negative measurable functions

Definition of integral of non-negative measurable functions

To preserve monotonicity, the integral of a nonnegative function f should be at least as large as the integral of any simple function $h \le f$ below it. The following definition is thus motivated by preserving monotonicity without unnecessarily giving up anything extra.

Definition VII.4 (Integral of a non-negative measurable function)**.**
For a non-negative measurable function $f \in \mathrm{m}\mathscr{S}^+$ we define

$$\int^{+} f\,\mathrm{d}\mu := \sup_{\substack{h\in \mathrm{s}\mathscr{S}^+ \\ 0\le h\le f}} \int^{\square} h\,\mathrm{d}\mu. \tag{VII.3}$$

Note that it is possible for the integral defined by the above supremum to become infinite — even for finite measures for which the integrals of all simple functions are finite.

We start by checking that the new more general definition (VII.3) of integral agrees with the earlier one (VII.2) whenever both are defined.

Lemma VII.5 (No conflict between the two first definitions of integral)**.** *For any non-negative simple function $f \in \mathrm{s}\mathscr{S}^+$ we have*

$$\int^{\square} f \, \mathrm{d}\mu = \int^{+} f \, \mathrm{d}\mu.$$

Proof. Assume that $f \in \mathrm{s}\mathscr{S}^+$.

For any $h \in \mathrm{s}\mathscr{S}^+$ such that $h \leq f$ we have by monotonicity, Lemma VII.3,

$$\int^{\square} h \, \mathrm{d}\mu \leq \int^{\square} f \, \mathrm{d}\mu.$$

By taking the supremum over such h as in the definition (VII.3), we get

$$\int^{+} f \, \mathrm{d}\mu = \sup_{\substack{h \in \mathrm{s}\mathscr{S}^+ \\ 0 \leq h \leq f}} \int^{\square} h \, \mathrm{d}\mu \leq \int^{\square} f \, \mathrm{d}\mu.$$

To prove the converse inequality, just consider $h = f$ in the supremum (VII.3), to get

$$\int^{+} f \, \mathrm{d}\mu = \sup_{\substack{h \in \mathrm{s}\mathscr{S}^+ \\ 0 \leq h \leq f}} \int^{\square} h \, \mathrm{d}\mu \geq \int^{\square} f \, \mathrm{d}\mu.$$

□

Properties of integral of non-negative measurable functions

The key properties of integrals continue to hold. Let us first verify monotonicity.

Proposition VII.6 (Monotonicity of the integral for non-negative functions)**.** *Suppose that $f, g \in \mathrm{m}\mathscr{S}^+$ and $f \leq g$ pointwise. Then we have*

$$\int^{+} f \, \mathrm{d}\mu \leq \int^{+} g \, \mathrm{d}\mu.$$

Proof. Whenever $h \in \mathrm{s}\mathscr{S}^+$ is such that $h \leq f$, we also have $h \leq g$, so the supremum in the definition of the integral of g is over a larger collection than the supremum in the definition of the integral of f. We get

$$\int^{+} f \, \mathrm{d}\mu = \sup_{\substack{h \in \mathrm{s}\mathscr{S}^+ \\ 0 \leq h \leq f}} \int^{\square} h \, \mathrm{d}\mu \;\leq\; \sup_{\substack{h \in \mathrm{s}\mathscr{S}^+ \\ 0 \leq h \leq g}} \int^{\square} h \, \mathrm{d}\mu = \int^{+} g \, \mathrm{d}\mu.$$

□

As a consequence, we get that the only way that a positive function can have vanishing integral is for the function to be equal to zero almost everywhere (i.e. except in a set of measure zero).

Proposition VII.7 (No nontrivial non-negative function has vanishing integral)**.**
If $f \in \mathrm{m}\mathscr{S}^+$ *and*

$$\int^+ f \,\mathrm{d}\mu = 0$$

then

$$\mu\Big[\,\big\{s \in S \;\big|\; f(s) > 0\big\}\,\Big] = 0.$$

Proof. For $n \in \mathbb{N}$, define

$$A_n = \Big\{ s \in S \;\Big|\; f(s) \geq \frac{1}{n} \Big\} = f^{-1}\Big([\frac{1}{n}, +\infty]\Big) \in \mathscr{S}.$$

We can then express the set of interest as

$$A = \Big\{ s \in S \;\Big|\; f(s) > 0 \Big\} = \bigcup_{n=1}^{\infty} A_n.$$

Consider the non-negative simple function $h = \frac{1}{n}\mathbb{I}_{A_n}$. By construction of A_n, we have $h \leq f$, and we can therefore estimate the supremum that defines the integral of f from below by

$$\int^+ f \,\mathrm{d}\mu \geq \int^{\square} \frac{1}{n}\mathbb{I}_{A_n} \,\mathrm{d}\mu = \frac{1}{n}\,\mu[A_n].$$

Using the assumption that the integral of f vanishes, we get

$$0 \leq \mu[A_n] \leq n \int^+ f \,\mathrm{d}\mu = 0,$$

so $\mu[A_n] = 0$. Finally, by the countable subadditivity (II.10) of Lemma II.19 (i.e., "the union bound") we get the conclusion

$$\mu[A] = \mu\Big[\bigcup_{n=1}^{\infty} A_n\Big] \leq \sum_{n=1}^{\infty} \mu[A_n] = \sum_{n=1}^{\infty} 0 = 0.$$

□

The converse to Proposition VII.7 also holds.

Exercise VII.2 (Almost everywhere vanishing functions have vanishing integrals)**.**
Check directly from Definition VII.4 that if $f \in \mathrm{m}\mathscr{S}^+$ and

$$\mu\Big[\,\big\{s \in S \;\big|\; f(s) > 0\big\}\,\Big] = 0,$$

then we have

$$\int^+ f \,\mathrm{d}\mu = 0.$$

At this stage, yet another fundamentally important property of integration arises, concerning the behavior of integrals under pointwise monotone approximation in the integrands (i.e., the functions to be integrated). The result, *Monotone convergence theorem (MCT)*, is at the same time

- "really all there is to integration theory"[4]
- a very practical tool for calculations.

[4]This is a quote from David Williams, [**Wil91**, Section 5.3]. One aspect of Williams' message is certainly to stress the importance of the MCT. Another aspect might be to point out that besides the MCT, all other steps in the construction of integral and derivation of its basic properties are rather intuitive and straightforward. For this latter reason we have chosen to postpone the somewhat lengthy proof of the Monotone convergence theorem to Appendix D, so that it does not interrupt the flow of the rest of the steps.

Theorem VII.8 (Monotone convergence theorem)**.**
If $f_1, f_2, \ldots \in \mathrm{m}\mathscr{S}^+$ and $f_n \uparrow f$ as $n \to \infty$, then we have

$$\int^+ f_n \,\mathrm{d}\mu \uparrow \int^+ f \,\mathrm{d}\mu \qquad \text{as } n \to \infty.$$

The proof is given in Appendix D.

Recall that by Lemma III.18, any non-negative measurable function $f \in \mathrm{m}\mathscr{S}^+$ can be approximated by simple functions $f_1, f_2, \ldots \in \mathrm{s}\mathscr{S}^+$ in a pointwise increasing way, $f_n \uparrow f$ as $n \to \infty$. Together with the observation of Lemma VII.5, Monotone convergence theorem thus gives the expression

$$\int^+ f \,\mathrm{d}\mu = \lim_{n\to\infty} \int^\square f_n \,\mathrm{d}\mu$$

for the integral of f as an increasing limit of integrals (VII.2) of simple functions. This should reassure the reader that integrals, despite their slightly abstract definition, are rather innocent constructions, after all.

It is now easy to verify that also linearity continues to hold.

Proposition VII.9 (Linearity of the integral for non-negative functions)**.**

(a) *If $f \in \mathrm{m}\mathscr{S}^+$ and $c \geq 0$, then $cf \in \mathrm{m}\mathscr{S}^+$ and*

$$\int^+ (cf) \,\mathrm{d}\mu = c \int^+ f \,\mathrm{d}\mu.$$

(b) *If $f, g \in \mathrm{m}\mathscr{S}^+$, then $f + g \in \mathrm{m}\mathscr{S}^+$ and*

$$\int^+ (f+g) \,\mathrm{d}\mu = \int^+ f \,\mathrm{d}\mu + \int^+ g \,\mathrm{d}\mu.$$

Proof. Both parts can be easily verified using the Monotone convergence theorem[5] and the corresponding properties of integrals of simple functions.

Let us only prove part (b) — the proof of part (a) is similar. So let $f, g \in \mathrm{m}\mathscr{S}^+$. As in Lemma III.18, pick sequences $f_1, f_2, \ldots \in \mathrm{s}\mathscr{S}^+$ and $g_1, g_2, \ldots \in \mathrm{s}\mathscr{S}^+$ of simple functions such that $f_n \uparrow f$ and $g_n \uparrow g$ as $n \to \infty$. Then by linearity of limits, we have also $f_n + g_n \uparrow f + g$ as $n \to \infty$. Now apply the Monotone convergence theorem (Theorem VII.8) to each of these three monotone approximations, and use also the linearity of integral for simple functions

[5]The careful reader wary of circular reasoning will now inspect that the proof of the Monotone convergence theorem in Appendix D does not rely on linearity of $\int^+$, but instead uses only properties of $\int^\square$ and the definition of $\int^+$.

(Lemma VII.2), to get

$$\begin{aligned}
&\int^{+} (f+g)\,\mathrm{d}\mu \\
&= \lim_{n\to\infty} \int^{\square} (f_n+g_n)\,\mathrm{d}\mu && \text{(by MCT and } f_n+g_n \uparrow f+g\text{)} \\
&= \lim_{n\to\infty} \left(\int^{\square} f_n\,\mathrm{d}\mu + \int^{\square} g_n\,\mathrm{d}\mu \right) && \text{(by Lemma VII.2)} \\
&= \lim_{n\to\infty} \int^{\square} f_n\,\mathrm{d}\mu + \lim_{n\to\infty} \int^{\square} g_n\,\mathrm{d}\mu && \text{(by linearity of limits)} \\
&= \int^{+} f\,\mathrm{d}\mu + \int^{+} g\,\mathrm{d}\mu && \text{(by MCT and } f_n \uparrow f \text{ and } g_n \uparrow g\text{).}
\end{aligned}$$

□

VII.3. Integral for integrable functions

For an arbitrary measurable function $f \in \mathrm{m}\mathscr{S}$, define two non-negative functions $f_+, f_- \colon S \to [0,+\infty]$ by

$$f_+(s) = \max\{f(s), 0\} \qquad \text{and} \qquad f_-(s) = \max\{-f(s), 0\}. \tag{VII.4}$$

Then we have $f_+, f_- \in \mathrm{m}\mathscr{S}^+$ by Proposition III.14, and we can write

$$f = f_+ - f_- \qquad \text{and} \qquad |f| = f_+ + f_-.$$

We call f_+ the *positive part* and f_- the *negative part* of f.

Definition VII.10 (Integrable function)**.**
We say that a function $f \in \mathrm{m}\mathscr{S}$ is *integrable* with respect to the measure μ and denote $f \in \mathcal{L}^1(\mu)$ if

$$\int^{+} |f|\,\mathrm{d}\mu < +\infty.$$

Remark VII.11 (Equivalent condition for integrability of a function)**.**
Since $|f| = f_+ + f_-$ and the integral of non-negative functions is additive, $\int^+ |f|\,\mathrm{d}\mu = \int^+ f_+\,\mathrm{d}\mu + \int^+ f_-\,\mathrm{d}\mu$, a measurable function is integrable if and only if

$$\int^{+} f_+\,\mathrm{d}\mu < +\infty \qquad \text{and} \qquad \int^{+} f_-\,\mathrm{d}\mu < +\infty.$$

Definition of integral of integrable functions

In view of the decomposition $f = f_+ - f_-$, our final definition of the integral is unsurprising.

Definition VII.12 (Integral of an integrable function)**.**
For $f \in \mathcal{L}^1(\mu)$ we define the integral as

$$\int f\,\mathrm{d}\mu = \int^{+} f_+\,\mathrm{d}\mu - \int^{+} f_-\,\mathrm{d}\mu, \tag{VII.5}$$

where f_+ and f_- are the positive and negative parts of the function f.

By Remark VII.11 above, the two terms in formula (VII.5) are finite, so the expression is well defined.

For non-negative functions $f \in \mathrm{m}\mathscr{S}^+$ the new definition of integral coincides with the earlier one.

Lemma VII.13 (No conflict between the two last definitions of integral)**.**
For any non-negative integrable function $f \in \mathrm{m}\mathscr{S}^+ \cap \mathcal{L}^1(\mu)$ we have

$$\int f \, \mathrm{d}\mu = \int^+ f \, \mathrm{d}\mu.$$

Proof. For a non-negative function $f \in \mathrm{m}\mathscr{S}^+$ we have $f_+ = f$ and $f_- = 0$. Therefore the agreement of the two integrals is obvious from the definition (VII.5). □

Properties of integral of integrable functions

Now that the functions involved may have both positive and negative signs, estimating integrals by the following triangle inequality becomes important.

Theorem VII.14 (Triangle inequality for integrals)**.**
For any integrable function $f \in \mathcal{L}^1(\mu)$ we have

$$\Big| \int f \, \mathrm{d}\mu \Big| \leq \int |f| \, \mathrm{d}\mu.$$

Proof. Observe first that the right hand side is

$$\int |f| \, \mathrm{d}\mu = \int^+ |f| \, \mathrm{d}\mu = \int^+ (f_+ + f_-) \, \mathrm{d}\mu = \int^+ f_+ \, \mathrm{d}\mu + \int^+ f_- \, \mathrm{d}\mu.$$

We will prove $\int f \, \mathrm{d}\mu \leq \int |f| \, \mathrm{d}\mu$ and $-\int f \, \mathrm{d}\mu \leq \int |f| \, \mathrm{d}\mu$, which together imply the assertion $\big| \int f \, \mathrm{d}\mu \big| \leq \int |f| \, \mathrm{d}\mu$.

By (VII.5) and the non-negativity of integral $\int f_- \, \mathrm{d}\mu$ of the non-negative function f_-, we have

$$\begin{aligned} \int f \, \mathrm{d}\mu &= \int^+ f_+ \, \mathrm{d}\mu - \int^+ f_- \, \mathrm{d}\mu \\ &\leq \int^+ f_+ \, \mathrm{d}\mu + \int^+ f_- \, \mathrm{d}\mu = \int |f| \, \mathrm{d}\mu. \end{aligned}$$

Similarly, we have

$$\begin{aligned} -\int f \, \mathrm{d}\mu &= -\int^+ f_+ \, \mathrm{d}\mu + \int^+ f_- \, \mathrm{d}\mu \\ &\leq \int^+ f_+ \, \mathrm{d}\mu + \int^+ f_- \, \mathrm{d}\mu = \int |f| \, \mathrm{d}\mu. \end{aligned}$$

□

In particular, in proving linearity, we use the triangle inequality to check that the sum of integrable functions is integrable in the first place.

Theorem VII.15 (Linearity of the integral)**.**

(a) *For any* $f \in \mathcal{L}^1(\mu)$ *and* $c \in \mathbb{R}$ *we have* $cf \in \mathcal{L}^1(\mu)$ *and*

$$\int cf \, \mathrm{d}\mu = c \int f \, \mathrm{d}\mu.$$

(b) *For any* $f, g \in \mathcal{L}^1(\mu)$ *we have* $f + g \in \mathcal{L}^1(\mu)$ *and*

$$\int (f + g) \, \mathrm{d}\mu = \int f \, \mathrm{d}\mu + \int g \, \mathrm{d}\mu.$$

Proof of part (a). If $c \geq 0$ then we have $(cf)_+ = cf_+$ and $(cf)_- = cf_-$. Integrability of cf follows easily by Remark VII.11, since

$$\int^+ (cf)_+ \, \mathrm{d}\mu = \int^+ cf_+ \, \mathrm{d}\mu = c \underbrace{\int^+ f_+ \, \mathrm{d}\mu}_{<+\infty} < +\infty$$

and similarly $\int^+ (cf)_- \, \mathrm{d}\mu < +\infty$. The asserted formula is also a direct consequence of the definition (VII.5) of $\int$, and linearity of $\int^+$ (Proposition VII.9),

$$\int^+ cf \, \mathrm{d}\mu = \int^+ cf_+ \, \mathrm{d}\mu - \int^+ cf_- \, \mathrm{d}\mu = c\Big(\int^+ f_+ \, \mathrm{d}\mu - \int^+ f_- \, \mathrm{d}\mu \Big) = c \int f \, \mathrm{d}\mu.$$

If $c < 0$ then we have $(cf)_+ = -cf_-$ and $(cf)_- = -cf_+$, and one can proceed similarly.

Proof of part (b). Note that $|f + g| \leq |f| + |g|$. If $f, g \in \mathcal{L}^1(\mu)$, we thus have

$$\int^+ |f + g| \, \mathrm{d}\mu \leq \int^+ \big(|f| + |g| \big) \, \mathrm{d}\mu = \int^+ |f| \, \mathrm{d}\mu + \int^+ |g| \, \mathrm{d}\mu < +\infty,$$

and therefore $f + g \in \mathcal{L}^1(\mu)$. The formula for the integral of $f + g$ can be proven by carefully considering which possibilities contribute to the positive and negative parts of $f + g$. □

Also monotonicity continues to hold.

Theorem VII.16 (Monotonicity of the integral)**.**
If $f, g \in \mathcal{L}^1(\mu)$ *and* $f \leq g$*, then we have*

$$\int f \, \mathrm{d}\mu \leq \int g \, \mathrm{d}\mu.$$

Proof. If we have the pointwise inequality $f \leq g$, then the positive and negative parts of these functions satisfy $f_+ \leq g_+$ and $f_- \geq g_-$. It therefore follows from the defining formula (VII.5) of $\int$, and monotonicity of $\int^+$ (Proposition VII.6), that

$$\begin{aligned} \int f \, \mathrm{d}\mu &= \int^+ f_+ \, \mathrm{d}\mu - \int^+ f_- \, \mathrm{d}\mu \\ &\leq \int^+ g_+ \, \mathrm{d}\mu - \int^+ g_- \, \mathrm{d}\mu = \int g \, \mathrm{d}\mu. \end{aligned}$$

□

VII.4. Convergence theorems for integrals

Finally, we establish some of the powerful standard tools of measure theory, which address when it is legitimate to interchange the order of a limit and an integration. Recalling that integration contains as special cases also expected values and

summation, we thus ask under what conditions we have

$$\begin{aligned}\lim_{n\to\infty}\int f_n(s)\,\mathrm{d}\mu(s) &\stackrel{?}{=} \int\Big(\lim_{n\to\infty} f_n(s)\Big)\,\mathrm{d}\mu(s)\\ \lim_{n\to\infty}\sum_{k=0}^{\infty} a_k^{(n)} &\stackrel{?}{=} \sum_{k=0}^{\infty}\big(\lim_{n\to\infty} a_k^{(n)}\big)\\ \lim_{n\to\infty}\mathsf{E}\big[X_n\big] &\stackrel{?}{=} \mathsf{E}\Big[\lim_{n\to\infty} X_n\Big].\end{aligned}$$

In particular, derivatives are limits of difference quotients, so the results can be applied to justifying the interchange of differentiation and integral, sum, or expected value, i.e.

$$\begin{aligned}\frac{\mathrm{d}}{\mathrm{d}\lambda}\int f_n(s;\lambda)\,\mathrm{d}\mu(s) &\stackrel{?}{=} \int\frac{\mathrm{d}}{\mathrm{d}\lambda} f_n(s;\lambda)\,\mathrm{d}\mu(s)\\ \frac{\mathrm{d}}{\mathrm{d}\lambda}\sum_{k=0}^{\infty} a_k(\lambda) &\stackrel{?}{=} \sum_{k=0}^{\infty}\Big(\frac{\mathrm{d}}{\mathrm{d}\lambda} a_k(\lambda)\Big)\\ \frac{\mathrm{d}}{\mathrm{d}\lambda}\mathsf{E}\big[f(X_n;\lambda)\big] &\stackrel{?}{=} \mathsf{E}\Big[\frac{\mathrm{d}}{\mathrm{d}\lambda} f(X_n;\lambda)\Big].\end{aligned}$$

To fully appreciate the positive results we will derive, it is advisable to first think about the ways the conclusion could fail.

Exercise VII.3 (Interchanging limit and integration is not always possible)**.**

(a) Consider the Lebesgue measure Λ on $\mathbb{R}$. For $n\in\mathbb{N}$, define $f_n\colon\mathbb{R}\to\mathbb{R}$ by

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

Calculate the following

$$\lim_{n\to\infty} f_n,\qquad \int_{\mathbb{R}} f_n\,\mathrm{d}\Lambda,\qquad \int_{\mathbb{R}}\big(\lim_{n\to\infty} f_n\big)\,\mathrm{d}\Lambda,\qquad \lim_{n\to\infty}\int_{\mathbb{R}} f_n\,\mathrm{d}\Lambda.$$

(b) Repeat the calculations of part (a) for the functions $\sqrt{f_n}$ and f_n^2, instead.

(c) Consider $\mathbb{Z}$ with the counting measure $\mu_\#$. For $n\in\mathbb{N}$, define $g_n\colon\mathbb{Z}\to\mathbb{R}$ by

$$g_n(k)=\begin{cases} 1 & \text{if } k=n\\ 0 & \text{if } k\ne n.\end{cases}$$

Calculate the following

$$\lim_{n\to\infty} g_n,\qquad \int_{\mathbb{Z}} g_n\,\mathrm{d}\mu_\#,\qquad \int_{\mathbb{Z}}\big(\lim_{n\to\infty} g_n\big)\,\mathrm{d}\mu_\#,\qquad \lim_{n\to\infty}\int_{\mathbb{Z}} g_n\,\mathrm{d}\mu_\#.$$

We will obtain several results about what conditions are sufficient to permit changing limits and integration. Besides

(MCT) Monotone convergence theorem (Theorem VII.8)

another such result of great importance is

(DCT) Dominated convergence theorem (Theorem VII.19)

For expected values, there is a very practical special case of the Dominated convergence theorem:

(BCT) Bounded convergence theorem (Corollary VII.21)

The Monotone convergence theorem was already introduced as a tool in step (3) of the construction. The proof of the Dominated convergence theorem relies on it, and another intermediate result that is occasionally useful on its own right as well:

Fatou's Lemma (Lemma VII.17)

Fatou's lemma

Fatou's lemma is a general convergence result for non-negative integrals, which does not even require the limits to exists! It instead addresses lower limits which always exist. A minor drawback is that the conclusion is merely an inequality in one direction.

Lemma VII.17 (Fatou's lemma)**.**
For any sequence $f_1, f_2, \ldots \in \mathrm{m}\mathscr{S}^+$ of non-negative measurable functions, we have

$$\int \left(\liminf_n f_n\right) \mathrm{d}\mu \leq \liminf_n \int f_n \,\mathrm{d}\mu. \tag{VII.6}$$

Proof. Define $g = \liminf_n f_n$. For $k \in \mathbb{N}$, define also

$$g_k := \inf_{n \geq k} f_n.$$

Then we have $g_k \in \mathrm{m}\mathscr{S}^+$ (by Proposition III.14), and $g_k \uparrow g$ as $k \to \infty$ by construction. By the Monotone convergence theorem, we therefore have

$$\int g_k \,\mathrm{d}\mu \uparrow \int g \,\mathrm{d}\mu \qquad \text{as } k \to \infty.$$

We have $f_n \geq g_k$ for any $n \geq k$, so monotonicity of integral gives $\int f_n \,\mathrm{d}\mu \geq \int g_k \,\mathrm{d}\mu$ for $n \geq k$, and thus also

$$\inf_{n \geq k} \int f_n \,\mathrm{d}\mu \;\geq\; \int g_k \,\mathrm{d}\mu.$$

In this inequality, take the limit as $k \to \infty$ to obtain

$$\liminf_n \int f_n \,\mathrm{d}\mu \;\geq\; \lim_{k\to\infty} \int g_k \,\mathrm{d}\mu \;=\; \int g \,\mathrm{d}\mu,$$

which is the asserted inequality. □

Exercise VII.4 (Strict inequality in Fatou's lemma)**.**
Find an example of a sequence of non-negative functions for which there is a strict inequality in Fatou's lemma, (VII.6).

Under suitable additional conditions, we get a result about upper limits of non-negative integrals as well.

Lemma VII.18 (Reverse Fatou's lemma)**.**
Suppose that $f_1, f_2, \ldots \in \mathrm{m}\mathscr{S}^+$ is a sequence of non-negative measurable functions such that there exists a non-negative measurable function $g \in \mathrm{m}\mathscr{S}^+$ which uniformly bounds the sequence, $f_n \leq g$ for all $n \in \mathbb{N}$, and which is itself integrable, $\int g \,\mathrm{d}\mu < +\infty$. Then we have

$$\limsup_n \int f_n \,\mathrm{d}\mu \leq \int \left(\limsup_n f_n\right) \mathrm{d}\mu. \tag{VII.7}$$

Proof. Apply Fatou's lemma to the sequence of functions $g - f_n$, $n \in \mathbb{N}$, and cancel the finite number $\int g \,\mathrm{d}\mu$ from both sides. □

Dominated convergence theorem

Arguably the most practical general convergence result is Lebesgue's dominated convergence theorem: it states that pointwise convergence and uniform boundedness of integrands by an integrable function are sufficient to permit taking the limit inside the integral.

Theorem VII.19 (Dominated convergence theorem)**.**
Suppose that $f_1, f_2, \ldots \in \mathrm{m}\mathscr{S}$ is a sequence of measurable functions such that there exists a non-negative measurable function $g \in \mathrm{m}\mathscr{S}^+$ which uniformly bounds the absolute values of the sequence, $|f_n| \leq g$ for all $n \in \mathbb{N}$, and which is itself integrable, $\int g \,\mathrm{d}\mu < +\infty$. Then if the pointwise limit $f = \lim_{n\to\infty} f_n$ exists, we have

$$\lim_{n\to\infty} \int |f_n - f| \,\mathrm{d}\mu = 0 \tag{VII.8}$$

and

$$\lim_{n\to\infty} \int f_n \,\mathrm{d}\mu = \int f \,\mathrm{d}\mu. \tag{VII.9}$$

Proof. Since $|f_n| \leq g$ for all $n \in \mathbb{N}$, and $f_n \to f$ as $n \to \infty$, we also have $|f| \leq g$. Therefore the triangle inequality gives

$$|f_n - f| \leq |f_n| + |f| \leq 2\,g.$$

We have $\int 2g < +\infty$ by integrability of g, so reverse Fatou's lemma can be applied to the sequence of functions $|f_n - f|$, $n \in \mathbb{N}$. Together with the assumed pointwise convergence $\lim_{n\to\infty} |f_n - f| = 0$, the reverse Fatou's lemma gives

$$\limsup_n \int |f_n - f| \,\mathrm{d}\mu \leq \int \left(\limsup_n |f_n - f|\right) \mathrm{d}\mu = \int 0 \,\mathrm{d}\mu = 0,$$

which proves the first assertion (VII.8).

Then use linearity and triangle inequality for integrals to get

$$\left| \int f_n \,\mathrm{d}\mu - \int f \,\mathrm{d}\mu \right| = \left| \int (f_n - f) \,\mathrm{d}\mu \right| \leq \int |f_n - f| \,\mathrm{d}\mu.$$

The right hand side tends to zero as $n \to \infty$ by the first part, (VII.8). This proves the second assertion (VII.9). □

Bounded convergence theorem

For finite measures (see Definition II.9) — and thus probability measures, in particular — there is a very easy and practical special case of the Dominated convergence theorem, known as the Bounded convergence theorem. The underlying reason for it, as well as many other simplifications in finite measure spaces, is the following.

Lemma VII.20 (Constant functions are integrable on finite measure spaces)**.**
Suppose that $(S, \mathscr{S}, \mu)$ is a finite measure space. Let $c \in \mathbb{R}$ be a constant, and let $g\colon S \to R$ be the constant function $g(s) = c$ for all $s \in S$. Then $g \in \mathcal{L}^1(\mu)$.

Proof. The constant function $g(s) = c$ is in particular a simple function, $g = c\,\mathbb{I}_S$, so we have

$$\int^{+} |g|\,\mathrm{d}\mu = \int^{\square} |c|\,\mathrm{d}\mu = |c|\ \underbrace{\mu[S]}_{<+\infty} < +\infty.$$

□

In particular, constant functions are legitimate choices for the dominating function g in the Dominated convergence theorem, Theorem VII.19. This immediately yields the following.

Corollary VII.21 (Bounded convergence theorem)**.**
Suppose that $(S, \mathscr{S}, \mu)$ is a finite measure space and $f_1, f_2, \ldots \in \mathrm{m}\mathscr{S}$ is a bounded sequence[6] *of measurable functions such that the pointwise limit function $f = \lim_{n\to\infty} f_n$ exists. Then we have*

$$\lim_{n\to\infty} \int f_n\,\mathrm{d}\mu = \int f\,\mathrm{d}\mu.$$

VII.5. Integrals over subsets and restriction of measures

Let $(S, \mathscr{S}, \mu)$ be a measure space.

We adopt the following notation for *integrals over subsets*: if $f\colon S \to \mathbb{R}$ is a function and if $R \subset S$ is a measurable subset, $R \in \mathscr{S}$, then the integral of f over R is defined as

$$\int_R f(s)\,\mathrm{d}\mu(s) := \int_S \mathbb{I}_R(s)\, f(s)\,\mathrm{d}\mu(s) \tag{VII.10}$$

whenever either $\mathbb{I}_R\, f \geq 0$ or $\mathbb{I}_R\, f \in \mathcal{L}^1(\mu)$.

A clear justification for the above convention is obtained in the following exercise.

Exercise VII.5 (Restriction of a measure to a subset)**.**
Let $(S, \mathscr{S}, \mu)$ is a measure space and $R \subset S$ a measurable subset, $R \in \mathscr{S}$.

(a) Let $\mathscr{S}'$ consist of those $A \subset R$ such that $A = B \cap R$ for some $B \in \mathscr{S}$, that is,

$$\mathscr{S}' := \left\{ B \cap R \;\middle|\; B \in \mathscr{S} \right\}.$$

Show that $\mathscr{S}'$ is a σ-algebra on R.

(b) When $A = B \cap R$ for some $B \in \mathscr{S}$, set

$$\mu'[A] := \mu[B \cap R].$$

Show that μ' is a measure on the measurable space $(R, \mathscr{S}')$.

(c) For any function $f\colon S \to \mathbb{R}$, denote by $f|_R\colon R \to \mathbb{R}$ the restriction of f to R. Prove that we have

$$\int_R f|_R\,\mathrm{d}\mu' = \int_S \mathbb{I}_R\, f\,\mathrm{d}\mu,$$

for (1) indicator functions f, (2) non-negative simple functions f, (3) non-negative measurable functions f, and finally for (4) all measurable functions f such that $f|_R \in \mathcal{L}^1(\mu')$.

[6] This means that there exists some $c \in [0, +\infty)$ such that $|f_n(s)| \leq c$ for all $s \in S$ and $n \in \mathbb{N}$.

VII.6. Riemann integral vs. Lebesgue integral

The reader can now verify that the Riemann integral (familiar from undergraduate calculus) and Lebesgue integral coincide in the following setups.

Exercise VII.6 (Riemann integral and Lebesgue integral on closed intervals)**.**
Consider a closed interval $[a,b] \subset \mathbb{R}$. Suppose that $f\colon [a,b] \to \mathbb{R}$ is a continuous function. Show that

$$\int_a^b f(x)\,\mathrm{d}x = \int_{[a,b]} f(x)\,\mathrm{d}\Lambda(x),$$

where the right hand side denotes the integral of f with respect to the Lebesgue measure Λ restricted to $[a,b] \subset \mathbb{R}$, and the left hand side denotes the Riemann integral of f.
Hint: *The Riemann integral is defined by the condition that the upper and lower Riemann sums associated to subdivisions of the interval $[a,b]$ tend to the same limit as the meshes of the subdivisions become finer. First show that these upper and lower Riemann sums associated to a subdivision are equal to the integrals of certain simple functions. Then use monotonicity to compare the Lebesgue integral of f to these.*

Exercise VII.7 (Improper Riemann integrals and Lebesgue integral)**.**

(a) Suppose that $f\colon [a,b) \to \mathbb{R}$ is continuous. Show that if $\int_{[a,b)} |f|\,\mathrm{d}\Lambda < +\infty$, then we have the following equality of the improper Riemann integral and the Lebesgue integral:

$$\int_a^b f(x)\,\mathrm{d}x := \lim_{\varepsilon \downarrow 0} \int_a^{b-\varepsilon} f(x)\,\mathrm{d}x = \int_{[a,b)} f(x)\,\mathrm{d}\Lambda(x).$$

(b) Suppose that $f\colon \mathbb{R} \to \mathbb{R}$ is continuous. Show that if $\int_{\mathbb{R}} |f|\,\mathrm{d}\Lambda < +\infty$, then we have the following equality of the improper Riemann integral and the Lebesgue integral:

$$\int_{-\infty}^{+\infty} f(x)\,\mathrm{d}x := \lim_{r \uparrow +\infty} \int_{-r}^{r} f(x)\,\mathrm{d}x = \int_{\mathbb{R}} f(x)\,\mathrm{d}\Lambda(x).$$

Hint: *Exercise VII.6 and Dominated convergence theorem.*

Remark VII.22 (On the relationship between Riemann-integral and Lebesgue integral)**.**

- There are plenty of functions for which the Riemann integral does not exist, while the Lebesgue integral is well-defined. A simple example is the indicator function $\mathbb{I}_{\mathbb{Q}\cap[a,b]}$ of the rational numbers on the interval. The Lebesgue integrals $\int \mathbb{I}_{\mathbb{Q}\cap[a,b]}\,\mathrm{d}\Lambda = 0$ and $\int \mathbb{I}_{[a,b]\setminus\mathbb{Q}}\,\mathrm{d}\Lambda = b-a$ reflect the fact that the countable set $\mathbb{Q}\cap[a,b]$ is negligible, in a measure-theoretic sense, compared to its complement $[a,b]\setminus\mathbb{Q}$ (which, in particular, is uncountable). The Riemann integral fails to account for this "evident" size difference.
- Pointwise limits of Riemann integrable functions need not be Riemann integrable, so one could not hope to have the general and powerful convergence theorems of Section VII.4 for the Riemann integral.[7]
- The coincidence of Riemann integral and Lebesgue integral for well-behaved functions allows us to use the familiar techniques such as the fundamental theorem of calculus, integration by parts, etc., without the need to prove them again.
- Given how common integration on the real axis is, it is convenient to have a concise notation for it. By virtue of the coincidence of the two notions of integration on $\mathbb{R}$ for all well-behaved functions, with almost no risk of confusion we can use the notation

$$\mathrm{d}x := \mathrm{d}\Lambda(x)$$

for integration with respect to Lebesgue measure Λ.

[7]Besides the generality (at once we constructed, e.g., integrals in all dimensions, expected values, infinite sums, ...), arguably the most significant advantage of the integration theory developed in this chapter is indeed the convergence theorems.

Lecture VIII

Expected values

In this lecture we begin examining the role and use of expected values in probability theory.

Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space. We will use the notation E for expected values. The *expected value* of a real-valued random variable $X\colon \Omega \to \mathbb{R}$ is just the integral with respect to the probability measure P,

$$\mathsf{E}\big[X\big] := \int_\Omega X(\omega)\, \mathrm{d}\mathsf{P}(\omega).$$

Let us summarize how it is constructed step by step for increasingly general random variables $X\colon \Omega \to \mathbb{R}$ as in the previous lecture (Lecture VII):

(1) If $X = \mathbb{I}_A$ for an event $A \in \mathscr{F}$, then we set $\mathsf{E}[X] = \mathsf{P}[A]$.
(2) If $X = \sum_{j=1}^n a_j \mathbb{I}_{A_j}$ is simple[1], then we set $\mathsf{E}[X] = \sum_{j=1}^n a_j \mathsf{P}[A_j]$.
(3) If X is non-negative, then we set

$$\mathsf{E}[X] = \sup_{\substack{H \in \mathrm{s}\mathscr{F}^+ \\ 0 \le H \le X}} \mathsf{E}[H].$$

(4) If X is integrable, i.e., $\mathsf{E}[X_+] < +\infty$ and $\mathsf{E}[X_-] < +\infty$, then we set

$$\mathsf{E}[X] = \mathsf{E}[X_+] - \mathsf{E}[X_-].$$

Since expected value is by definition a special case of integral, the basic properties of integrals lead to the corresponding basic properties of expected values:

Linearity: (Proposition VII.9 and Theorem VII.15)
For either $X, Y \in \mathrm{m}\mathscr{F}^+$ (two non-negative random variables), or for $X, Y \in \mathcal{L}^1(\mathsf{P})$ (two integrable random variables), we have

$$\mathsf{E}\big[X + Y\big] = \mathsf{E}\big[X\big] + \mathsf{E}\big[Y\big].$$

For either $X \in \mathrm{m}\mathscr{F}^+$ and $c \ge 0$, or for $X \in \mathcal{L}^1(\mathsf{P})$ and $c \in \mathbb{R}$, we have

$$\mathsf{E}\big[c\,X\big] = c\,\mathsf{E}\big[X\big].$$

Monotonicity: (Proposition VII.6 and Theorem VII.16)
For either $X, Y \in \mathrm{m}\mathscr{F}^+$ (two non-negative random variables) or for $X, Y \in \mathcal{L}^1(\mathsf{P})$ (two integrable random variables), we have

$$X \le Y \qquad \Longrightarrow \qquad \mathsf{E}\big[X\big] \le \mathsf{E}\big[Y\big]$$

[1] In Section VII.1 we were careful to only integrate simple functions which are non-negative, in order to avoid undefined expressions of the form $+\infty - \infty$. For expected values, because of finiteness $\mathsf{P}[\Omega] = 1 < +\infty$ of the total mass of a probability measure, this never causes problems, so formula (VII.2) can be directly taken as a definition of the expected value of any simple function (random variable).

Triangle inequality: (Theorem VII.14)
For $X \in \mathcal{L}^1(\mathsf{P})$, we have

$$\Big|\, \mathsf{E}\big[X\big] \,\Big| \;\leq\; \mathsf{E}\Big[|X|\Big].$$

Besides these basic properties, also the powerful convergence theorems of Section VII.4 continue hold for expected values.

VIII.1. Expected values in terms of laws

Recall from Definition III.6 that the law of a random variable $X\colon \Omega \to \mathbb{R}$ (or the distribution of X) is the probability measure P_X on $\mathbb{R}$ given by

$$P_X[B] = \mathsf{P}\big[X \in B\big] \qquad \text{for } B \in \mathscr{B}.$$

We next show how the expected value of $g(X)$ can be calculated as an integral with respect to the law P_X.

Theorem VIII.1 (Expected values in terms of laws)**.**
Let $X\colon \Omega \to \mathbb{R}$ be a random variable with law P_X, and let $g\colon \mathbb{R} \to \mathbb{R}$ be a Borel function. Then we have the equivalence of the following integrability properties

$$g(X) \in \mathcal{L}^1(\mathsf{P}) \qquad \Longleftrightarrow \qquad g \in \mathcal{L}^1(P_X).$$

If either (then both) of the above integrability properties holds, then we have

$$\mathsf{E}\big[g(X)\big] = \int_{\mathbb{R}} g(x)\, \mathrm{d}P_X(x).$$

Proof. The proof follows the "standard machine", i.e., we verify the statement for (1) indicator functions, (2) simple functions, (3) non-negative measurable functions, and finally (4) all measurable functions g.

step 1: Consider an indicator function $g = \mathbb{I}_B$ of a Borel set $B \in \mathscr{B}$. Observe, as in (IV.1) and (IV.2), that $\mathbb{I}_B(X(\omega)) = \mathbb{I}_{X^{-1}(B)}(\omega)$ for all $\omega \in \Omega$. Thus we have

$$\begin{aligned}
\mathsf{E}\big[\mathbb{I}_B(X)\big] &= \mathsf{E}\big[\mathbb{I}_{X^{-1}(B)}\big] && \text{(because } \mathbb{I}_B \circ X = \mathbb{I}_{X^{-1}(B)}\text{)}\\
&= \mathsf{P}\big[X^{-1}(B)\big] && \text{(by step 1 of def. of } \mathsf{E}\text{)}\\
&= P_X[B] && \text{(by def. of } P_X\text{)}\\
&= \int \mathbb{I}_B \,\mathrm{d}P_X. && \text{(by step 1 of def. of integral)}
\end{aligned}$$

This shows that $\mathsf{E}[g(X)] = \int g \,\mathrm{d}P_X$ for $g = \mathbb{I}_B$.

step 2: Consider a simple Borel function $g \in \mathrm{s}\mathscr{B}^+$. Write $g = \sum_{j=1}^n a_j \mathbb{I}_{B_j}$. We have

$$\begin{aligned}
\mathsf{E}\Big[\sum_{j=1}^n a_j\, \mathbb{I}_{B_j}(X)\Big] &= \sum_{j=1}^n a_j\, \mathsf{E}\big[\mathbb{I}_{B_j}(X)\big] && \text{(by linearity of } \mathsf{E}\text{)}\\
&= \sum_{j=1}^n a_j \int \mathbb{I}_{B_j}\, \mathrm{d}P_X && \text{(by step 1)}\\
&= \int \Big(\sum_{j=1}^n a_j\, \mathbb{I}_{B_j}\Big)\, \mathrm{d}P_X && \text{(by linearity of integral).}
\end{aligned}$$

This shows that $\mathsf{E}[g(X)] = \int g \,\mathrm{d}P_X$ for $g \in \mathrm{s}\mathscr{B}^+$.

step 3: Consider a non-negative Borel function $g \in \mathrm{m}\mathscr{B}^+$. Take a monotone increasing pointwise approximation $g_n \uparrow g$ as $n \to \infty$ by non-negative simple Borel functions $g_n \in \mathrm{s}\mathscr{B}^+$. We then also have $g_n(X) \uparrow g(X)$, and $g_n(X) \in \mathrm{s}\mathscr{F}^+$ are non-negative simple random variables. Using Monotone convergence theorem (Theorem VII.8) both for P and for P_X, we get

$$\begin{aligned} \mathsf{E}\big[g(X)\big] &= \lim_{n\to\infty} \mathsf{E}\big[g_n(X)\big] && \text{(by MCT for } \mathsf{P}) \\ &= \lim_{n\to\infty} \int g_n \,\mathrm{d}P_X && \text{(by step 2)} \\ &= \int g \,\mathrm{d}P_X. && \text{(by MCT for } P_X). \end{aligned}$$

step 4: Consider a Borel function $g \in \mathrm{m}\mathscr{B}$, and let $g_+, g_- \in \mathrm{m}\mathscr{B}^+$ be its positive and negative parts. The positive and negative parts of the random variable $g(X) = g \circ X$ are $(g(X))_+ = g_+ \circ X \in \mathrm{m}\mathscr{F}^+$ and $(g(X))_- = g_- \circ X \in \mathrm{m}\mathscr{F}^+$. By step 3 we thus have

$$\begin{aligned} \mathsf{E}\Big[\big(g(X)\big)_+\Big] &= \mathsf{E}\big[g_+(X)\big] = \int g_+ \,\mathrm{d}P_X \\ \text{and} \quad \mathsf{E}\Big[\big(g(X)\big)_-\Big] &= \mathsf{E}\big[g_-(X)\big] = \int g_- \,\mathrm{d}P_X. \end{aligned}$$

Therefore we first of all have the equivalent conditions

$$\begin{aligned} \mathsf{E}\Big[\big(g(X)\big)_+\Big] < +\infty \quad &\Longleftrightarrow \quad \int g_+ \,\mathrm{d}P_X < +\infty \\ \text{and} \quad \mathsf{E}\Big[\big(g(X)\big)_-\Big] < +\infty \quad &\Longleftrightarrow \quad \int g_- \,\mathrm{d}P_X < +\infty, \end{aligned}$$

which by Remark VII.11 shows the equivalence $g(X) \in \mathcal{L}^1(\mathsf{P}) \Leftrightarrow g \in \mathcal{L}^1(P_X)$. Moreover, when this integrability holds, step 4 of the definition of the integrals gives

$$\begin{aligned} \mathsf{E}\big[g(X)\big] &= \mathsf{E}\Big[\big(g(X)\big)_+\Big] - \mathsf{E}\Big[\big(g(X)\big)_-\Big] \\ \text{and} \quad \int g \,\mathrm{d}P_X &= \int g_+ \,\mathrm{d}P_X - \int g_- \,\mathrm{d}P_X. \end{aligned}$$

These coincide by the above equalities, which finishes the proof. □

Exercise VIII.1 (Discrete random numbers)**.**
A random variable is *discrete* if its range $A = X(\Omega)$ is finite or countably infinite. The probability mass function of a discrete random variable X is defined by $p_X(x) = \mathsf{P}\big[X = x\big]$. Prove that any discrete real-valued random variable satisfies:

(a) $\mathsf{E}\big[h(X)\big] = \sum_{x\in A} h(x)\, p_X(x)$ for all Borel functions $h : \mathbb{R} \to [0, \infty)$.
(b) $h(X) \in \mathcal{L}^1(\Omega, \mathscr{F}, \mathsf{P})$ if and only if $\sum_{x\in A} |h(x)|\, p_X(x) < \infty$.
(c) Explain why the formula in (a) is true for all $h \in \mathcal{L}^1(\mathbb{R}, \mathscr{B}, P_X)$.

Hint: *Recall the argument in the proof of Theorem VIII.1.*

Exercise VIII.2 (Almost sure equality)**.**
Let X and Y be real-valued random variables that are equal almost surely, that is,

$$\mathsf{P}\Big[X = Y\Big] = 1.$$

(a) In order to make sure that $\mathsf{P}[X = Y]$ is a meaningful probability, explain first why the set $\big\{\omega \in \Omega \;\big|\; X(\omega) = Y(\omega)\big\}$ is measurable.
(b) Prove that the laws of X and Y are the same, and conclude that we in particular have $\mathsf{E}[X] = \mathsf{E}[Y]$ (whenever the expected values exist).

Probability densities of continuous distributions

Definition VIII.2 (Probability density of a continuous distribution)**.**
Let $X \in \mathrm{m}\mathscr{F}$ be a real valued random variable. If there exists a Borel function

$$f_X : \mathbb{R} \to [0, +\infty)$$

such that[2]

$$P_X[B] := \mathsf{P}\big[X \in B\big] = \int_B f_X(x)\,\mathrm{d}x \tag{VIII.1}$$

for all $B \in \mathscr{B}$, then we say that X has a *continuous distribution* (or a *continuous law*), and we say that f_X is a *density function* of X.

Example VIII.3 (Gaussian distribution)**.**
Let $\mathfrak{m} \in \mathbb{R}$ and $\mathfrak{s} > 0$. A random variable X is said to have a *gaussian distribution* with mean $\mathfrak{m}$ and variance $\mathfrak{s}^2$ if its distribution is continuous and

$$f_X(x) = \frac{1}{\sqrt{2\pi\,\mathfrak{s}^2}}\,\exp\Big(-\frac{1}{2\,\mathfrak{s}^2}(x-\mathfrak{m})^2\Big)$$

is a density function of X.

Gaussian distributions are also called *normal distributions*, and the particular case of zero mean $\mathfrak{m} = 0$ and unit variance $\mathfrak{s}^2 = 1$ is called the *standard normal distribution*.

Example VIII.4 (Exponential distribution)**.**
Let $\lambda > 0$. A random variable X is said to have an *exponential distribution* with parameter λ if its distribution is continuous and

$$f_X(x) = \begin{cases} \lambda\, e^{-\lambda x} & \text{for } x \geq 0 \\ 0 & \text{for } x < 0 \end{cases}$$

is a density function of X.

Exercise VIII.3 (Random numbers with continuous distribution)**.**
Assume that X has a continuous law with a density function f_X.

(a) Show that X is integrable if and only if $\int_{\mathbb{R}} |x|\, f_X(x)\,\mathrm{d}x < \infty$.
(b) If X is integrable, show that expectation of X is given by

$$\mathsf{E}\big[X\big] = \int_{\mathbb{R}} x\, f_X(x)\,\mathrm{d}x.$$

(c) If X is integrable with expected value $\mathfrak{m} := \mathsf{E}[X]$, show that the variance of X can be computed as

$$\mathrm{Var}(X) := \mathsf{E}\big[(X - \mathsf{E}[X])^2\big] = \int_{\mathbb{R}} (x-\mathfrak{m})^2\, f_X(x)\,\mathrm{d}x.$$

(d) Can a random variable with continuous law have more than one density function?

Hint: *Remember the hint for Exercise VIII.1, and seek a unified argument leading to (a) and (b) — perhaps even (c).*

Exercise VIII.4 (Expected value and variance of an exponential random variable)**.**
Assume that a random variable X has an exponential distribution with parameter λ (see Example VIII.4). Use the previous exercise to calculate the expected value $\mathsf{E}[X]$ and the variance $\mathrm{Var}(X)$ of X.

[2]The right hand side of (VIII.1) is an integral over the subset $B \subset \mathbb{R}$ in the sense of (VII.10),

$$\int_B f_X(x)\,\mathrm{d}x \;:=\; \int_{\mathbb{R}} f_X(x)\,\mathbb{I}_B(x)\;\mathrm{d}\Lambda(x).$$

Exercise VIII.5 (The absolute value of a random number)**.**
Let $Y = |X|$ where X is a real-valued random variable.

(a) Prove that Y is a random variable.
(b) Assume that we know the cumulative distribution function $F_X(x) = \mathsf{P}[X \leq x]$ of X. What is the cumulative distribution function of Y?
(c) Assume that X has continuous distribution with a density function $f_X(x)$. Does Y also have a continuous distribution in this case? If yes, write down an expression for a density function f_Y of Y in terms of f_X. If not, explain why not.

VIII.2. Applications of convergence theorems for expected values

In this section we look at a few consequences of the convergence theorems in various aspects of probability theory.

Expected values of random series with non-negative terms

Let us observe that a non-negative random variable can only have finite expected value if the random variable is almost surely finite.[3]

Lemma VIII.5 (Finite expected value implies almost sure finiteness)**.**
If $X \in \mathrm{m}\mathscr{F}^+$ is a non-negative random variable such that $\mathsf{E}[X] < +\infty$, then we have $X < +\infty$ almost surely, i.e., $\mathsf{P}\big[X < +\infty\big] = 1$.

Proof. Let $A = \big\{\omega \in \Omega \;\big|\; X(\omega) = +\infty\big\}$ be the event that the random variable X takes an infinite value. Then for any $n \in \mathbb{N}$ we obviously have $X \geq n\,\mathbb{I}_A$, so by monotonicity of expected values we get

$$n\,\mathsf{P}[A] = \mathsf{E}\Big[n\,\mathbb{I}_A\Big] \leq \mathsf{E}\big[X\big].$$

Dividing this by n, we get $\mathsf{P}[A] \leq \frac{1}{n}\,\mathsf{E}[X]$. By assumption X has finite expected value, $\mathsf{E}\big[X\big] < +\infty$, so by letting $n \to \infty$ we get $\mathsf{P}[A] \leq 0$. This shows that $\mathsf{P}[A] = 0$, so passing to the complement we prove the claim

$$\mathsf{P}\big[X < +\infty\big] = \mathsf{P}\big[A^{\mathbf{c}}\big] = 1 - \mathsf{P}[A] = 1 - 0 = 1.$$

□

Next we note that in a random sum with non-negative terms, we are allowed to interchange the order of summation and expected value.

Lemma VIII.6 (Expected value of a series of non-negative random terms)**.**
Suppose that $X_1, X_2, \ldots \in \mathrm{m}\mathscr{F}^+$ is a sequence of non-negative random variables. Consider the random infinite series $\sum_{k=1}^{\infty} X_k$. Then we have

$$\mathsf{E}\Big[\sum_{k=1}^{\infty} X_k\Big] = \sum_{k=1}^{\infty} \mathsf{E}\big[X_k\big].$$

[3] A similar conclusion holds for integrals generally, not just expected values. We leave it to the reader to precisely formulate the statement and see that the proof carries through.

Proof. By definition, these infinite sums are the limits of their finite partial sums, in particular

$$\sum_{k=1}^{\infty} X_k(\omega) = \lim_{n\to\infty} \sum_{k=1}^{n} X_k(\omega) \qquad \text{for all } \omega \in \Omega.$$

By non-negativity of the terms, the sequence of partial sums is increasing, i.e.,

$$\sum_{k=1}^{n} X_k(\omega) \uparrow \sum_{k=1}^{\infty} X_k(\omega) \qquad \text{as } n \to \infty, \qquad \text{for all } \omega \in \Omega.$$

In this pointwise increasing approximation, we can apply the Monotone convergence theorem (MCT), and get

$$\begin{aligned}
\mathsf{E}\Big[\sum_{k=1}^{\infty} X_k\Big] &= \mathsf{E}\Big[\lim_{n\to\infty} \sum_{k=1}^{n} X_k\Big] && \text{(by definition of infinite sum)} \\
&= \lim_{n\to\infty} \mathsf{E}\Big[\sum_{k=1}^{n} X_k\Big] && \text{(by MCT)} \\
&= \lim_{n\to\infty} \sum_{k=1}^{n} \mathsf{E}\Big[X_k\Big] && \text{(by linearity of expected value)} \\
&= \sum_{k=1}^{\infty} \mathsf{E}\Big[X_k\Big]. && \text{(by definition of infinite sum)}
\end{aligned}$$

The assertion is thus proven. □

We can combine the above observations to obtain a sufficient condition for almost sure convergence of random series with non-negative terms.

Proposition VIII.7 (Guaranteeing almost sure convergence of a random series)**.** *Suppose that $X_1, X_2, \ldots \in \mathrm{m}\mathscr{F}^+$ is a sequence of non-negative random variables. Suppose moreover that we have*

$$\sum_{k=1}^{\infty} \mathsf{E}\big[X_k\big] < +\infty.$$

Then we have almost surely $\sum_{k=1}^{\infty} X_k < +\infty$ and almost surely $X_k \to 0$ as $k \to \infty$, i.e.,

$$\mathsf{P}\Big[\sum_{k=1}^{\infty} X_k < +\infty\Big] = 1 \qquad \textit{and} \qquad \mathsf{P}\Big[\lim_{k\to\infty} X_k = 0\Big] = 1.$$

Proof. Note that the convergence $\sum_{k=1}^{\infty} X_k < +\infty$ of the series implies that its terms tend to zero, $\lim_{k\to\infty} X_k = 0$, so for the corresponding events we have

$$\left\{\omega \in \Omega \;\middle|\; \sum_{k=1}^{\infty} X_k(\omega) < +\infty\right\} \subset \left\{\omega \in \Omega \;\middle|\; \lim_{k\to\infty} X_k(\omega) = 0\right\}$$

and for their probabilities therefore

$$\mathsf{P}\Big[\sum_{k=1}^{\infty} X_k < +\infty\Big] \leq \mathsf{P}\Big[\lim_{k\to\infty} X_k = 0\Big].$$

It therefore suffices to show that the probability on the left hand side is equal to one. By Lemma VIII.6 and the assumption of convergence of the series of expected values, we get

$$\mathsf{E}\Big[\sum_{k=1}^{\infty} X_k\Big] = \sum_{k=1}^{\infty} \mathsf{E}\Big[X_k\Big] < +\infty.$$

Lemma VIII.5 therefore implies what we claimed,

$$1 = \mathsf{P}\Big[\sum_{k=1}^{\infty} X_k < +\infty\Big] \leq \mathsf{P}\Big[\lim_{k\to\infty} X_k = 0\Big].$$

□

Indicators of events, in particular, are non-negative random variables. The first Borel – Cantelli lemma can thus be seen as consequence of the above observations.

Lemma (Borel–Cantelli lemma: convergence part, Lemma V.7)**.**
Suppose that $E_1, E_2, \ldots \in \mathscr{F}$ are such that $\sum_{k=1}^{\infty} \mathsf{P}[E_k] < +\infty$. Then we have

$$\mathsf{P}\big[\text{“}E_k \text{ occurs infinitely often”}\big] = 0.$$

Proof. Consider the indicators $\mathbb{I}_{E_k}$ of the events E_k, $k = 1, 2, \ldots$. Note that the random series

$$N := \sum_{k=1}^{\infty} \mathbb{I}_{E_k}$$

counts the (random) number N of events in the sequence $E_1, E_2, \ldots$, which occur. By assumption, we have

$$\sum_{k=1}^{\infty} \mathsf{E}\big[\mathbb{I}_{E_k}\big] = \sum_{k=1}^{\infty} \mathsf{P}[E_k] < +\infty.$$

Therefore by Proposition VIII.7 we in particular have that $\mathsf{P}[N < +\infty] = 1$, i.e., almost surely only finitely many of the events in the sequence $E_1, E_2, \ldots$ occur. The complementary event is that infinitely many of the events occur, so we have

$$\mathsf{P}\big[\text{“}E_k \text{ occurs infinitely often”}\big] = \mathsf{P}\big[N = +\infty\big] = 1 - \mathsf{P}\big[N < +\infty\big] = 1 - 1 = 0.$$

This finishes our alternative proof of the first Borel–Cantelli lemma. □

Differentiation inside expected values

As an application of the Dominated convergence theorem, the reader can verify that exchanging the order of expected values and differentiation is permitted for example in the following situation.

Exercise VIII.6 (Differentiation inside expectation)**.**
Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space and $X\colon \Omega \to \mathbb{R}$ a random variable. Suppose that $h\colon \mathbb{R} \times (a,b) \to \mathbb{R}$ is a continuous function, $(x,\lambda) \mapsto h(x,\lambda)$. Assume that the partial derivative with respect to the second variable, $\frac{\partial}{\partial\lambda}h\colon \mathbb{R} \times (a,b) \to \mathbb{R}$, is also continuous. Assume moreover, that for some integrable random variable $Y \in \mathcal{L}^1(\Omega, \mathscr{F}, \mathsf{P})$ we have

$$\text{for all } \lambda \in (a,b) \text{ and } \omega \in \Omega: \qquad \Big|\Big(\frac{\partial}{\partial\lambda}h\Big)(X(\omega),\lambda)\Big| \leq Y(\omega).$$

(a) Show that for any $\lambda, \lambda' \in (a,b)$ with $\lambda \neq \lambda'$ we have $|\frac{h(X,\lambda')-h(X,\lambda)}{\lambda'-\lambda}| \leq Y$.
Hint: *Write $h(x,\lambda_2) - h(x,\lambda_1) = \int_{\lambda_1}^{\lambda_2} (\frac{\partial}{\partial\lambda}h)(x,\lambda)\,\mathrm{d}\lambda$, and recall the triangle inequality for integrals.*

(b) Show that

$$\frac{\mathrm{d}}{\mathrm{d}\lambda}\mathsf{E}\Big[h(X,\lambda)\Big] = \mathsf{E}\Big[\big(\frac{\partial}{\partial\lambda}h\big)(X,\lambda)\Big].$$

Hint: *When $\lambda_n \to \lambda$, consider $\frac{1}{\lambda_n-\lambda}\Big(\mathsf{E}\big[h(X,\lambda_n)\big] - \mathsf{E}\big[h(X,\lambda)\big]\Big)$. Use part (a) and dominated convergence.*

Differentiation inside expected value is in particular often used for various generating functions.

Exercise VIII.7 (Differentiating the moment generating function)**.**
Define the moment generating function of $X\colon \Omega \to \mathbb{R}$ by $M_X(\lambda) := \mathsf{E}\big[e^{\lambda X}\big]$. Assume that $\mathsf{E}\big[e^{\varepsilon|X|}\big] < \infty$ for some $\varepsilon > 0$.

(a) Show that $M_X'(0) = \mathsf{E}[X]$.
Hint: *Find a way to apply Exercise VIII.6, perhaps by first showing that for $|\lambda| < \varepsilon$ one has $|xe^{\lambda x}| \leq Ce^{\varepsilon|x|}$.*

(d) Explain, without detailed calculations, why $M_X''(0) = \mathsf{E}[X^2]$. Find also a similar formula for $\mathsf{E}[X^n]$ for all $n \in \mathbb{N}$.

VIII.3. Space of p-integrable random variables

Let $X\colon \Omega \to \widehat{\mathbb{R}}$ be a random variable. Recall that we say that X is integrable and write $X \in \mathcal{L}^1(\mathsf{P})$ if

$$\mathsf{E}\big[|X|\big] < +\infty.$$

The following generalization is encountered very often.

Definition VIII.8 (p-integrability)**.**
Let $p > 0$. We say that a random variable X is *p-integrable* and denote $X \in \mathcal{L}^p(\mathsf{P})$ if

$$\mathsf{E}\big[|X|^p\big] < +\infty. \tag{VIII.2}$$

In probability spaces, p-integrability implies r-integrability for any $r \leq p$, as we will show in the next lemma. This is used in particular to conclude that if it is meaningful to talk about the *moment* $\mathsf{E}[X^p]$ of order p, it is also meaningful to talk about all lower order *moments* $\mathsf{E}[X^r]$. Let us emphasize that in very common measure spaces of infinite total mass, p-integrability does <u>not</u> imply integrability of lower order, so the lemma is specific to probability theory.

Lemma VIII.9 (Finiteness of lower order moments)**.**
Let $0 < r < p$. Suppose that $X \in \mathcal{L}^p(\mathsf{P})$. Then we have $X \in \mathcal{L}^r(\mathsf{P})$, and moreover

$$\mathsf{E}\big[|X|^r\big] \leq 1 + \mathsf{E}\big[|X|^p\big].$$

Proof. Since $0 < r < p$, we have that $|x|^r < |x|^p$ whenever $|x| > 1$, and $|x|^r \leq 1$ whenever $|x| \leq 1$. Define the event

$$A = \Big\{\omega \in \Omega \;\Big|\; |X(\omega)| > 1\Big\}.$$

Then we have the pointwise estimate

$$\begin{aligned} |X(\omega)|^r &\leq \mathbb{I}_{A^{\mathrm{c}}}(\omega) + \mathbb{I}_A(\omega)\,|X(\omega)|^p \\ &\leq 1 + |X(\omega)|^p. \end{aligned}$$

By monotonicity and linearity of the expected value, this gives

$$\mathsf{E}\Big[|X|^r\Big] \leq \mathsf{E}\Big[1 + |X|^p\Big] \leq 1 + \mathsf{E}\big[|X|^p\big].$$

Both asserted results follow from this. □

The space $\mathcal{L}^p(\mathsf{P})$ of all p-integrable random variables is a vector space.

Lemma VIII.10 (p-integrable random variables form a vector space)**.**
Let $p \geq 0$. Then we have the following.

(a) *If $X \in \mathcal{L}^p(\mathsf{P})$ and $a \in \mathbb{R}$, then $aX \in \mathcal{L}^p(\mathsf{P})$.*
(b) *If $X, Y \in \mathcal{L}^p(\mathsf{P})$, then $X + Y \in \mathcal{L}^p(\mathsf{P})$ and*

$$\mathsf{E}\Big[|X+Y|^p\Big] \leq 2^p \left(\mathsf{E}\big[|X|^p\big] + \mathsf{E}\big[|Y|^p\big]\right) \tag{VIII.3}$$

Proof. Part (a) is clear, since $\mathsf{E}\big[|aX|^p\big] = |a|^p\, \mathsf{E}\big[|X|^p\big]$. It therefore only remains to prove (b).

For any $x, y \in \mathbb{R}$ note that

$$|x+y| \leq |x| + |y| \leq 2\, \max\{|x|, |y|\}\,.$$

The mapping $t \mapsto t^p$ is increasing $[0,\infty) \to [0,\infty)$, so applying it to the above gives

$$|x+y|^p \leq 2^p\, \max\{|x|^p, |y|^p\}\,.$$

This of course implies also

$$|x+y|^p \leq 2^p \left(|x|^p + |y|^p\right).$$

Applying this inequality pointwise to the values of the random variables X and Y gives

$$|X(\omega)+Y(\omega)|^p \leq 2^p \left(|X(\omega)|^p + |Y(\omega)|^p\right) \qquad \text{for all } \omega \in \Omega.$$

Taking expected values and using monotonicity and linearity, we get

$$\mathsf{E}\Big[|X+Y|^p\Big] \leq 2^p \left(\mathsf{E}\big[|X|^p\big] + \mathsf{E}\big[|Y|^p\big]\right),$$

and the assertions of part (b) follow. □

Since a constant random variable c is p-integrable for any $p > 0$, the vector space property in particular has the following consequence.

Corollary VIII.11 (Adding a constant preserves p-integrability)**.**
If $X \in \mathcal{L}^p(\mathsf{P})$ and $c \in \mathbb{R}$ is a constant, then also $X + c \in \mathcal{L}^p(\mathsf{P})$.

The estimates in Lemma VIII.9 were rather crude. The following very often useful inequality is one way to improve them.

Exercise VIII.8 (Jensen's inequality)**.**
Suppose that $I \subset \mathbb{R}$ is an interval and $\phi\colon I \to \mathbb{R}$ is a convex function, i.e., a function such that for any $x, y \in I$ and $\lambda \in (0,1)$ we have

$$\phi\Big(\lambda\, x + (1-\lambda)\, y\Big) \;\leq\; \lambda\, \phi(x) + (1-\lambda)\, \phi(y).$$

Suppose moreover that $X\colon \Omega \to I$ is a random variable with values on the interval I.

(a) Show that for any $z \in I$ there exists a number $d \in \mathbb{R}$ such that $\phi(x) \geq \phi(z) + (x-z)\, d$.
Hint: *Consider the left and right derivatives of ϕ at z.*
(b) Show that if X is a random variable with values in $I \subset \mathbb{R}$, and both X and $\phi(X)$ are integrable, then we have

$$\phi\big(\mathsf{E}[X]\big) \leq \mathsf{E}\big[\phi(X)\big].$$

Hint: *Choose $z = \mathsf{E}[X]$ and apply part (a) together with monotonicity and linearlity of expected values.*

Exercise VIII.9 (The p-norm controls lower norms)**.**
Suppose that $X \in \mathcal{L}^p(\mathsf{P})$ for some $p > 0$. Let $0 < r < p$. Using Exercise VIII.8, show that

$$\Big(\mathsf{E}\Big[|X|^r\Big]\Big)^{1/r} \leq \Big(\mathsf{E}\Big[|X|^p\Big]\Big)^{1/p},$$

Lecture IX

Product spaces and Fubini's theorem

Imagine observing two random phenomena simultaneously. The set of possible outcomes then is a Cartesian product, it consists of pairs

$$\Omega_1 \times \Omega_2 = \big\{(\omega_1, \omega_2) \;\big|\; \omega_1 \in \Omega_1,\ \omega_2 \in \Omega_2\big\},$$

whose two components are outcomes of the two phenomena.

Example IX.1 (Two coin tosses)**.**
For one coin toss the sample space is $\{\mathrm{H}, \mathrm{T}\}$, representing the possible outcomes "heads" and "tails". For two coin tosses the sample space is

$$\{\mathrm{H}, \mathrm{T}\} \times \{\mathrm{H}, \mathrm{T}\} = \big\{(\mathrm{H}, \mathrm{H}),\ (\mathrm{H}, \mathrm{T}),\ (\mathrm{T}, \mathrm{H}),\ (\mathrm{T}, \mathrm{T})\big\}.$$

In this lecture we treat measures on such Cartesian product spaces. We construct in particular:

- the product σ-algebra on a Cartesian product of measurable spaces
- the product measure on a Cartesian product of (σ-finite) measure spaces.

One of the results of great practical importance is Fubini's Theorem (Theorem IX.9): for two measure spaces $(S_1, \mathscr{S}_1, \mu_1)$ and $(S_2, \mathscr{S}_2, \mu_2)$ and a function $f\colon S_1 \times S_2 \to \mathbb{R}$, under reasonable conditions the change of order of integration formula

$$\int_{S_1} \Big(\int_{S_2} f(s_1, s_2)\, \mathrm{d}\mu_2(s_2) \Big)\, \mathrm{d}\mu_1(s_1) = \int_{S_2} \Big(\int_{S_1} f(s_1, s_2)\, \mathrm{d}\mu_1(s_1) \Big)\, \mathrm{d}\mu_2(s_2)$$

holds. Recalling that expected values and summation are special cases of integration, this implies (under reasonable assumptions) formulas such as

$$\int \Big(\sum_{k=1}^{\infty} f_k(s) \Big)\, \mathrm{d}\mu(s) = \sum_{k=1}^{\infty} \int f_k(s)\, \mathrm{d}\mu(s),$$
$$\mathsf{E}\Big[\sum_{k=1}^{\infty} X_k \Big] = \sum_{k=1}^{\infty} \mathsf{E}\big[X_k\big],$$

etc.

Key tool: Monotone class theorem

In this lecture, we will repeatedly use the Monotone class theorem (Theorem C.2 from Appendix C). Let us recall its statement, and introduce convenient notation for it.

We denote as before

$$\begin{aligned} \mathrm{m}\mathscr{S} &\quad — \quad \text{the set of } \mathscr{S}\text{-measurable functions } S \to [-\infty, +\infty] \\ \mathrm{s}\mathscr{S} &\quad — \quad \text{the set of simple functions } S \to \mathbb{R} \end{aligned}$$

and in addition the set of bounded measurable functions will be denoted by

$$\mathrm{b}\mathscr{S} := \Big\{ f \in \mathrm{m}\mathscr{S} \;\Big|\; |f| \le C \text{ for some } C \in \mathbb{R} \Big\}.$$

We also continue to use the superscript + for the sets of non-negative functions of the corresponding types,

$$\begin{aligned} \mathrm{m}\mathscr{S}^+ &:= \Big\{ f \in \mathrm{m}\mathscr{S} \;\Big|\; f \ge 0 \Big\} \\ \mathrm{s}\mathscr{S}^+ &:= \Big\{ f \in \mathrm{s}\mathscr{S} \;\Big|\; f \ge 0 \Big\} \\ \mathrm{b}\mathscr{S}^+ &:= \Big\{ f \in \mathrm{b}\mathscr{S} \;\Big|\; f \ge 0 \Big\}. \end{aligned}$$

According to Definition C.1, a collection $\mathscr{H}$ of functions $S \to \mathbb{R}$ is said to be a monotone class if:

(MC-1) The constant function 1 belongs to $\mathscr{H}$.
(MC-$\mathbb{R}$) The class $\mathscr{H}$ is a vector space over $\mathbb{R}$.
(MC-$\uparrow$) If $f_1, f_2, \ldots \in \mathscr{H}$ is an increasing sequence of non-negative functions in $\mathscr{H}$ such that the pointwise limit $f_n \uparrow f$ is a bounded function f, then $f \in \mathscr{H}$.

The statement of the Monotone Class Theorem is the following.

Theorem (Monotone class theorem, Theorem C.2)**.**
Let $\mathscr{H}$ be a monotone class of bounded functions from S to $\mathbb{R}$ and let $\mathscr{J}$ be a π-system on S such that $\sigma(\mathscr{J}) = \mathscr{S}$. Suppose that

$$\mathbb{I}_A \in \mathscr{H} \qquad \text{for every } A \in \mathscr{J}.$$

Then we have

$$\mathrm{b}\mathscr{S} \subset \mathscr{H}.$$

IX.1. Product sigma algebra

Let $(S_1, \mathscr{S}_1)$ and $(S_2, \mathscr{S}_2)$ be two measurable spaces. As the first step, we have to equip the Cartesian product

$$S_1 \times S_2 = \big\{ (s_1, s_2) \;\big|\; s_1 \in S_1,\ s_2 \in S_2 \big\}$$

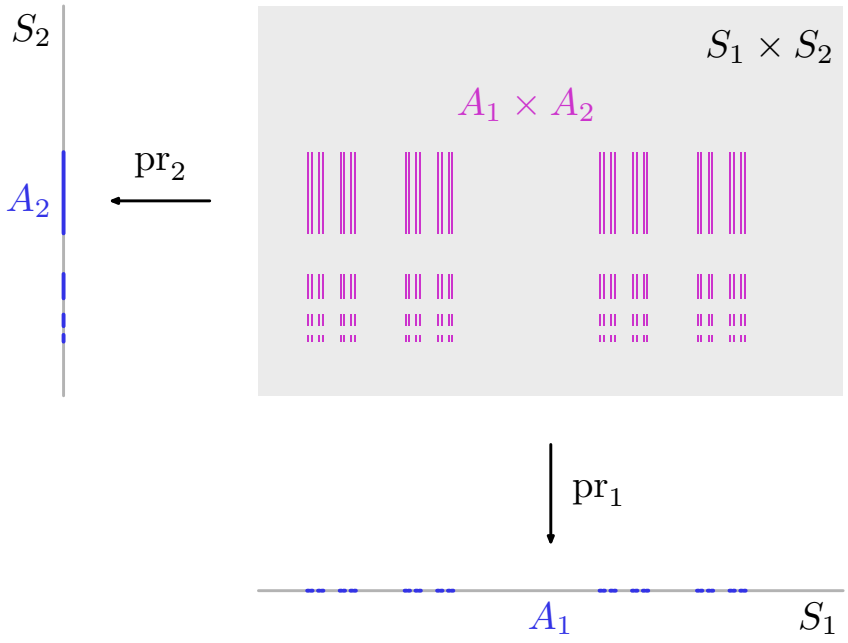
with a σ-algebra. We think of the coordinate projections

$$\begin{aligned} &\mathrm{pr}_1 \colon\ S_1 \times S_2 \to S_1 && \mathrm{pr}_1(s_1, s_2) = s_1 \\ &\mathrm{pr}_2 \colon\ S_1 \times S_2 \to S_2 && \mathrm{pr}_2(s_1, s_2) = s_2 \end{aligned}$$

as fundamental, and require that they must be measurable functions to $(S_1, \mathscr{S}_1)$ and $(S_2, \mathscr{S}_2)$, respectively.

Definition IX.2 (Product sigma algebra)**.**
The *product σ-algebra* $\mathscr{S}_1 \otimes \mathscr{S}_2$ on $S_1 \times S_2$ is the σ-algebra generated by the functions pr_1 and pr_2, i.e., the smallest σ-algebra on $S_1 \times S_2$ with respect to which $\mathrm{pr}_1 \colon S_1 \times S_2 \to S_1$ and $\mathrm{pr}_2 \colon S_1 \times S_2 \to S_2$ are measurable.

The following lemma says that Cartesian products of measurable sets are $\mathscr{S}_1 \otimes \mathscr{S}_2$-measurable, and that $\mathscr{S}_1 \otimes \mathscr{S}_2$ is the smallest σ-algebra with that property.



Lemma IX.3 (A pi system for the product sigma algebra)**.**
The collection

$$\mathscr{I} = \{A_1 \times A_2 \mid A_1 \in \mathscr{S}_1,\, A_2 \in \mathscr{S}_2\} \tag{IX.1}$$

is a π-system on $S_1 \times S_2$, and we have $\mathscr{S}_1 \otimes \mathscr{S}_2 = \sigma(\mathscr{I})$.

Exercise IX.1. Prove Lemma IX.3.

The Euclidean plane is just the Cartesian product $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ of two copies of the real axis. Does that mean that we have equipped the plane with two different σ-algebras: its Borel σ-algebra (generated by open subsets of the plane), and its product σ-algebra (generated by projections from the plane to the two axes)? In principle yes — but...

Exercise IX.2 (The two natural σ-algebras on the plane are the same)**.**
Let $\mathscr{B}$ be the Borel σ-algebra on the real line $\mathbb{R}$, and $\mathscr{B}(\mathbb{R}^2)$ the Borel σ-algebra on the plane $\mathbb{R}^2$. Show that

$$\mathscr{B} \otimes \mathscr{B} = \mathscr{B}(\mathbb{R}^2).$$

Hint: *A set $A \subset \mathbb{R}^n$ is open if and only if it can be written as a countable union of open boxes of the form $(a_1, b_1) \times \cdots \times (a_n, b_n)$ with a_i, b_i being rational numbers.*

Freezing one coordinate

Our first use of the Monotone Class Theorem is to show that with any function $f\colon S_1 \times S_2 \to \mathbb{R}$ which is measurable with respect to the product σ-algebra, keeping one of the two coordinates fixed and letting the other vary defines measurable functions $S_1 \to \mathbb{R}$ and $S_2 \to \mathbb{R}$.

Lemma IX.4 (Freezing a coordinate preserves measurability)**.**
Let $\mathscr{H}$ denote the class of functions $f\colon S_1 \times S_2 \to \mathbb{R}$ such that $f \in \mathrm{b}(\mathscr{S}_1 \otimes \mathscr{S}_2)$ and

$$\begin{aligned} &\forall s_1 \in S_1 \qquad s_2 \mapsto f(s_1, s_2) \quad \textit{is } \mathscr{S}_2\textit{-measurable } S_2 \to \mathbb{R} \\ &\forall s_2 \in S_2 \qquad s_1 \mapsto f(s_1, s_2) \quad \textit{is } \mathscr{S}_1\textit{-measurable } S_1 \to \mathbb{R}. \end{aligned}$$

Then we have $\mathscr{H} = \mathrm{b}(\mathscr{S}_1 \otimes \mathscr{S}_2)$.

Proof. By definition of $\mathscr{H}$ we have $\mathscr{H} \subset \mathrm{b}(\mathscr{S}_1 \otimes \mathscr{S}_2)$, so we must show $\mathscr{H} \supset \mathrm{b}(\mathscr{S}_1 \otimes \mathscr{S}_2)$.

Clearly the class $\mathscr{H}$ satisfies (MC-1), (MC-$\mathbb{R}$), and (MC-$\uparrow$). We will show that $\mathscr{H}$ contains the indicator functions of all sets in the π-system $\mathscr{I}$ of Lemma IX.3. Since $\sigma(\mathscr{I}) = \mathscr{S}_1 \otimes \mathscr{S}_2$, the Monotone Class Theorem then allows us to conclude that $\mathscr{H} \supset \mathrm{b}(\mathscr{S}_1 \otimes \mathscr{S}_2)$.

So suppose that $A_1 \in \mathscr{S}_1$ and $A_2 \in \mathscr{S}_2$, and consider the indicator function

$$\mathbb{I}_{A_1 \times A_2}\colon S_1 \times S_2 \to \mathbb{R}$$

of the set $A_1 \times A_2 \subset S_1 \times S_2$. We have

$$\mathbb{I}_{A_1 \times A_2}(s_1, s_2) = \mathbb{I}_{A_1}(s_1)\, \mathbb{I}_{A_2}(s_2),$$

so if $s_1 \in A_1$, the function $s_2 \mapsto \mathbb{I}_{A_1 \times A_2}(s_1, s_2)$ is the indicator $\mathbb{I}_{A_2}$ of the set $A_2 \subset S_2$, and if $s_1 \notin A_1$, it is zero. In either case, $s_2 \mapsto \mathbb{I}_{A_1 \times A_2}(s_1, s_2)$ is $\mathscr{S}_2$-measurable $S_2 \to \mathbb{R}$. Similarly one shows that $s_1 \mapsto \mathbb{I}_{A_1 \times A_2}(s_1, s_2)$ is $\mathscr{S}_1$-measurable $S_1 \to \mathbb{R}$, for any $s_2 \in S_2$. This shows that the indicator functions of all sets in the π-system $\mathscr{I}$ of Lemma IX.3 are in the class $\mathscr{H}$. This finishes the proof. □

IX.2. Product measure

Let $(S_1, \mathscr{S}_1, \mu_1)$ and $(S_2, \mathscr{S}_2, \mu_2)$ be two measure spaces. Our next goal is to construct a product measure, denoted by $\mu_1 \otimes \mu_2$, on the measurable space $(S_1 \times S_2, \mathscr{S}_1 \otimes \mathscr{S}_2)$.

First, in Section IX.2.1 we do that in the case when the measures are finite, and then in Section IX.2.2 we relax the assumption of finiteness to σ-finiteness.

IX.2.1. Product of two finite measures

We now consider the case that $(S_1, \mathscr{S}_1, \mu_1)$ and $(S_2, \mathscr{S}_2, \mu_2)$ are two finite measure spaces, i.e., we assume that the total masses of the two measures are finite

$$\mu_1[S_1] < +\infty \qquad \text{and} \qquad \mu_2[S_2] < +\infty.$$

This assumption will be used frequently below to conclude that a function which is bounded and measurable is necessarily integrable (recall Lemma VII.20).

Suppose that $f\colon S_1 \times S_2 \to \mathbb{R}$ is a bounded $\mathscr{S}_1 \otimes \mathscr{S}_2$-measurable function, i.e., $f \in \mathrm{b}(\mathscr{S}_1 \otimes \mathscr{S}_2)$. The functions $s_2 \mapsto f(s_1, s_2)$ and $s_1 \mapsto f(s_1, s_2)$ are clearly also

bounded, and they are measurable by Lemma IX.4. Thus they are integrable, and we can define the functions

$$\mathfrak{J}_1^f(s_1) := \int_{S_2} f(s_1, s_2)\, \mathrm{d}\mu_2(s_2) \tag{IX.2}$$

$$\mathfrak{J}_2^f(s_2) := \int_{S_1} f(s_1, s_2)\, \mathrm{d}\mu_1(s_1). \tag{IX.3}$$

The next lemma shows that these integrals over one of the variables define measurable functions. Moreover, it essentially establishes the conclusion of Fubini's theorem in the particular case of bounded measurable functions on finite measure spaces.

Lemma IX.5 (Fubini's theorem for bounded functions)**.**
Let $(S_1, \mathscr{S}_1, \mu_1)$ and $(S_2, \mathscr{S}_2, \mu_2)$ be two finite measure spaces, and let $f \in \mathrm{b}(\mathscr{S}_1 \otimes \mathscr{S}_2)$. We then have the following:

(i) *The function $s_1 \mapsto \mathfrak{J}_1^f(s_1)$ in* (IX.2) *is bounded and $\mathscr{S}_1$-measurable $S_1 \to \mathbb{R}$.*
(ii) *The function $s_2 \mapsto \mathfrak{J}_2^f(s_2)$ in* (IX.3) *is bounded and $\mathscr{S}_2$-measurable $S_2 \to \mathbb{R}$.*
(iii) *We have the equality*

$$\int_{S_1} \mathfrak{J}_1^f(s_1)\, \mathrm{d}\mu_1(s_1) = \int_{S_2} \mathfrak{J}_2^f(s_2)\, \mathrm{d}\mu_2(s_2).$$

Proof. Let $\mathscr{H}$ be the collection of functions $f \in \mathrm{b}(\mathscr{S}_1 \otimes \mathscr{S}_2)$ for which (i), (ii), and (iii) hold. We must show $\mathscr{H} = \mathrm{b}(\mathscr{S}_1 \otimes \mathscr{S}_2)$, and we will again do this using the Monotone Class Theorem.

We first claim that for any $A_1 \in \mathscr{S}_1$ and $A_2 \in \mathscr{S}_2$, the indicator $\mathbb{I}_{A_1 \times A_2} \colon S_1 \times S_2 \to \mathbb{R}$ is in the collection $\mathscr{H}$. Recall that $\mathbb{I}_{A_1 \times A_2}(s_1, s_2) = \mathbb{I}_{A_1}(s_1)\, \mathbb{I}_{A_2}(s_2)$, fix $s_1 \in S_1$, and integrate over the variable s_2 to get

$$\begin{aligned}
\mathfrak{J}_1^{\mathbb{I}_{A_1 \times A_2}}(s_1) &= \int_{S_2} \mathbb{I}_{A_1}(s_1)\, \mathbb{I}_{A_2}(s_2)\, \mathrm{d}\mu_2(s_2) \\
&= \mathbb{I}_{A_1}(s_1) \int_{S_2} \mathbb{I}_{A_2}(s_2)\, \mathrm{d}\mu_2(s_2) \\
&= \mu_2[A_2]\, \mathbb{I}_{A_1}(s_1).
\end{aligned}$$

As a constant multiple of an indicator of the measurable set A_1, the function $s_1 \mapsto \mathfrak{J}_1^{\mathbb{I}_{A_1 \times A_2}}(s_1)$ is $\mathscr{S}_1$-measurable and bounded, i.e., (i) holds for $\mathbb{I}_{A_1 \times A_2}$. Similarly one shows that

$$\mathfrak{J}_2^{\mathbb{I}_{A_1 \times A_2}}(s_2) = \mu_1[A_1]\, \mathbb{I}_{A_2}(s_2),$$

and concludes that $s_2 \mapsto \mathfrak{J}_2^{\mathbb{I}_{A_1 \times A_2}}(s_2)$ is $\mathscr{S}_2$-measurable and bounded, i.e., (ii) holds for $\mathbb{I}_{A_1 \times A_2}$. For (iii), calculate

$$\begin{aligned}
\int_{S_1} \mathfrak{J}_1^{\mathbb{I}_{A_1 \times A_2}}(s_1)\, \mathrm{d}\mu_1(s_1) &= \int_{S_1} \mu_2[A_2]\, \mathbb{I}_{A_1}(s_1)\, \mathrm{d}\mu_1(s_1) \\
&= \mu_2[A_2] \int_{S_1} \mathbb{I}_{A_1}(s_1)\, \mathrm{d}\mu_1(s_1) \\
&= \mu_2[A_2]\, \mu_1[A_1],
\end{aligned}$$

and similarly

$$\int_{S_2} \mathfrak{J}_2^{\mathbb{I}_{A_1 \times A_2}}(s_2)\, \mathrm{d}\mu_2(s_2) = \mu_1[A_1]\, \mu_2[A_2].$$

These two integrals coincide, so we get that also (iii) holds for $\mathbb{I}_{A_1 \times A_2}$, and thus $\mathbb{I}_{A_1 \times A_2} \in \mathscr{H}$.

It remains to prove that $\mathscr{H}$ is a monotone class. Properties (MC-1) and (MC-$\mathbb{R}$) are easy to verify. For property (MC-$\uparrow$), suppose that $f_1, f_2, \ldots \in \mathscr{H}$ are non-negative and $f_n(s_1, s_2) \uparrow f(s_1, s_2)$ for all $(s_1, s_2) \in S_1 \times S_2$, where $f \colon S_1 \times S_2 \to \mathbb{R}$ is bounded. Then for

any $s_1 \in S_1$, the functions $s_2 \mapsto f_n(s_1, s_2)$ increase pointwise to $s_2 \mapsto f(s_1, s_2)$ as $n \to \infty$, so by Monotone convergence theorem we get

$$\int_{S_2} f_n(s_1, s_2)\, \mathrm{d}\mu_2(s_2) \uparrow \int_{S_2} f(s_1, s_2)\, \mathrm{d}\mu_2(s_2),$$

i.e., $\mathfrak{J}_1^{f_n}(s_1) \uparrow \mathfrak{J}_1^f(s_1)$. Since $f_n \in \mathscr{H}$, the functions $s_1 \mapsto \mathfrak{J}_1^{f_n}(s_1)$ are $\mathscr{S}_1$-measurable, and as their pointwise limit also the function $s_1 \mapsto \mathfrak{J}_1^f(s_1)$ must therefore be. Moreover, since f is bounded and the measure μ_2 is finite, the function $s_1 \mapsto \mathfrak{J}_1^f(s_1)$ is bounded, so property (i) holds for f. Similarly one shows that $\mathfrak{J}_2^{f_n}(s_2) \uparrow \mathfrak{J}_2^f(s_2)$ for any $s_2 \in S_2$, and property (ii) for f follows. Finally, use the Monotone convergence theorem twice (once for μ_1 and once for μ_2) to calculate

$$\begin{aligned}
\int_{S_1} \mathfrak{J}_1^f(s_1)\, \mathrm{d}\mu_1(s_1) &= \lim_{n\to\infty} \int_{S_1} \mathfrak{J}_1^{f_n}(s_1)\, \mathrm{d}\mu_1(s_1) && (\mathfrak{J}_1^{f_n} \uparrow \mathfrak{J}_1^f \text{ and MCT for } \mu_1)\\
&= \lim_{n\to\infty} \int_{S_2} \mathfrak{J}_2^{f_n}(s_2)\, \mathrm{d}\mu_2(s_2) && (\text{property (iii) for } f_n \in \mathscr{H})\\
&= \int_{S_2} \mathfrak{J}_2^f(s_2)\, \mathrm{d}\mu_2(s_2). && (\mathfrak{J}_2^{f_n} \uparrow \mathfrak{J}_2^f \text{ and MCT for } \mu_2)
\end{aligned}$$

The equality above proves property (iii) for f, and shows that $f \in \mathscr{H}$, therefore establishing (MC-↑) and showing that $\mathscr{H}$ is indeed a monotone class. This finishes the proof. □

Property (iii) of the previous lemma shows that the two formulas below are equal, and the following definition is therefore unambiguous.

Definition IX.6 (Product measure)**.**
Let $(S_1, \mathscr{S}_1, \mu_1)$ and $(S_2, \mathscr{S}_2, \mu_2)$ be two finite measure spaces. The *product measure* $\mu_1 \otimes \mu_2$ on $S_1 \times S_2$ is defined by

$$\begin{aligned}
(\mu_1 \otimes \mu_2)[B] &= \int_{S_1} \Big(\int_{S_2} \mathbb{I}_B(s_1, s_2)\, \mathrm{d}\mu_2(s_2) \Big)\, \mathrm{d}\mu_1(s_1) \qquad \text{(IX.4)}\\
&= \int_{S_2} \Big(\int_{S_1} \mathbb{I}_B(s_1, s_2)\, \mathrm{d}\mu_1(s_1) \Big)\, \mathrm{d}\mu_2(s_2)
\end{aligned}$$

for any $B \in \mathscr{S}_1 \otimes \mathscr{S}_2$.

This formula indeed gives rise to a measure on the product space, as we verify next.

Lemma IX.7 (The product measure is a measure)**.**
Let $(S_1, \mathscr{S}_1, \mu_1)$ and $(S_2, \mathscr{S}_2, \mu_2)$ be two finite measure spaces, and define $\mu_1 \otimes \mu_2$ by formula (IX.4)*. Then $\mu_1 \otimes \mu_2$ is a measure on $(S_1 \times S_2, \mathscr{S}_1 \otimes \mathscr{S}_2)$.*

Proof. The defining formula (IX.4) only involves integrals of non-negative functions, so clearly $(\mu_1 \otimes \mu_2)[B] \in [0, +\infty]$ for any $B \in \mathscr{S}_1 \otimes \mathscr{S}_2$. We must check the two defining properties of a measure.

The indicator of the empty set $\emptyset \subset S_1 \times S_2$ is the zero function, $\mathbb{I}_\emptyset(s_1, s_2) = 0$. The integral of the zero function is zero, so from (IX.4) we directly get $(\mu_1 \otimes \mu_2)[\emptyset] = 0$.

Suppose that $B_1, B_2, \ldots \in \mathscr{S}_1 \otimes \mathscr{S}_2$ are disjoint, and let $B := \bigcup_{k\in\mathbb{N}} B_k$. Consider also the finite unions $U_n = B_1 \cup \cdots \cup B_n$ for $n \in \mathbb{N}$, to get an increasing sequence $U_1 \subset U_2 \subset \cdots$ of $\mathscr{S}_1 \otimes \mathscr{S}_2$-measurable sets increasing to $\bigcup_{n\in\mathbb{N}} U_n = B$. By disjointness we have $\mathbb{I}_{U_n} = \sum_{k=1}^n \mathbb{I}_{B_k}$. We first use this to calculate $(\mu_1 \otimes \mu_2)[U_n]$ from (IX.4), using linearity of

integration,

$$\begin{aligned}(\mu_1\otimes\mu_2)[U_n] &= \int_{S_1}\Big(\int_{S_2}\mathbb{I}_{U_n}(s_1,s_2)\,\mathrm{d}\mu_2(s_2)\Big)\,\mathrm{d}\mu_1(s_1)\\ &= \int_{S_1}\Big(\int_{S_2}\big(\sum_{k=1}^{n}\mathbb{I}_{B_k}(s_1,s_2)\big)\,\mathrm{d}\mu_2(s_2)\Big)\,\mathrm{d}\mu_1(s_1)\\ &= \sum_{k=1}^{n}\Big(\int_{S_1}\Big(\int_{S_2}\mathbb{I}_{B_k}(s_1,s_2)\,\mathrm{d}\mu_2(s_2)\Big)\,\mathrm{d}\mu_1(s_1)\Big) = \sum_{k=1}^{n}(\mu_1\otimes\mu_2)[B_k].\end{aligned}$$

On the other hand, the indicators of the increasing sequence U_n of sets increase pointwise to the indicator of the limit set B, i.e., $\mathbb{I}_{U_n}\uparrow\mathbb{I}_B$. We can use this to calculate $(\mu_1\otimes\mu_2)[B]$ from (IX.4), using Monotone convergence theorems (first for μ_2 and then for μ_1),

$$\begin{aligned}(\mu_1\otimes\mu_2)[B] &= \int_{S_1}\Big(\int_{S_2}\mathbb{I}_{B}(s_1,s_2)\,\mathrm{d}\mu_2(s_2)\Big)\,\mathrm{d}\mu_1(s_1)\\ &= \int_{S_1}\Big(\int_{S_2}\lim_{n\to\infty}\mathbb{I}_{U_n}(s_1,s_2)\,\mathrm{d}\mu_2(s_2)\Big)\,\mathrm{d}\mu_1(s_1)\\ &= \int_{S_1}\Big(\lim_{n\to\infty}\int_{S_2}\mathbb{I}_{U_n}(s_1,s_2)\,\mathrm{d}\mu_2(s_2)\Big)\,\mathrm{d}\mu_1(s_1) \qquad \text{(MCT for } \mu_2)\\ &= \lim_{n\to\infty}\int_{S_1}\Big(\int_{S_2}\mathbb{I}_{U_n}(s_1,s_2)\,\mathrm{d}\mu_2(s_2)\Big)\,\mathrm{d}\mu_1(s_1) \qquad \text{(MCT for } \mu_1)\\ &= \lim_{n\to\infty}(\mu_1\otimes\mu_2)[U_n].\end{aligned}$$

Combining these two calculations, we have shown that

$$\begin{aligned}(\mu_1\otimes\mu_2)\big[\bigcup_{k\in\mathbb{N}}B_k\big] = (\mu_1\otimes\mu_2)[B] &= \lim_{n\to\infty}(\mu_1\otimes\mu_2)[U_n]\\ &= \lim_{n\to\infty}\sum_{k=1}^{n}(\mu_1\otimes\mu_2)[B_k] = \sum_{k=1}^{\infty}(\mu_1\otimes\mu_2)[B_k],\end{aligned}$$

which is the countable additivity property for $\mu_1\otimes\mu_2$. □

It took some work to construct the product measure, but to characterize it is very easy:

Lemma IX.8 (A characterization of the product measure)**.**
Let $(S_1,\mathscr{S}_1,\mu_1)$ and $(S_2,\mathscr{S}_2,\mu_2)$ be two finite measure spaces. Then the product measure $\mu_1\otimes\mu_2$ is the unique measure ν on $(S_1\times S_2,\mathscr{S}_1\otimes\mathscr{S}_2)$ such that for all $A_1\in\mathscr{S}_1$ and $A_2\in\mathscr{S}_2$ we have

$$\nu[A_1\times A_2]=\mu_1[A_1]\,\mu_2[A_2].$$

Proof. An easy calculation (done already in the proof of Lemma IX.5) starting from the definition (IX.4) of the product measure gives

$$(\mu_1\otimes\mu_2)[A_1\times A_2]=\mu_1[A_1]\,\mu_2[A_2].$$

Thus $\mu_1\otimes\mu_2$ indeed satisfies the asserted formula.

Now if ν is another measure on $(S_1\times S_2,\mathscr{S}_1\otimes\mathscr{S}_2)$ for which the formula holds, then ν and $\mu_1\otimes\mu_2$ coincide on all sets of the π-system (IX.1), $\mathscr{I}=\big\{A_1\times A_2\;\big|\;A_1\in\mathscr{S}_1,\,A_2\in\mathscr{S}_2\big\}$. In particular the total masses are equal, $\nu[S_1\times S_1]=(\mu_1\otimes\mu_2)[S_1\times S_1]$. We can divide each by this finite total mass to get two probability measures, which agree on the π-system $\mathscr{I}$. Since $\sigma(\mathscr{I})=\mathscr{S}_1\otimes\mathscr{S}_2$, Dynkin's identification theorem then guarantees that these probability measures on $(S_1\times S_2,\mathscr{S}_1\otimes\mathscr{S}_2)$ are equal. Multiplying the total mass back, we conclude that $\nu=\mu_1\otimes\mu_2$. This shows the uniqueness of a measure satisfying the formula. □

Now we state and prove the main result about product measures.

Theorem IX.9 (Fubini's theorem)**.**
For a function $f\colon S_1 \times S_2 \to [-\infty, +\infty]$, consider the following three integrals:

$$\int_{S_1\times S_2} f \,\mathrm{d}(\mu_1 \otimes \mu_2) \tag{IX.5}$$

$$\int_{S_1} \Big(\int_{S_2} f(s_1, s_2)\,\mathrm{d}\mu_2(s_2)\Big)\,\mathrm{d}\mu_1(s_1) \tag{IX.6}$$

$$\int_{S_2} \Big(\int_{S_1} f(s_1, s_2)\,\mathrm{d}\mu_1(s_1)\Big)\,\mathrm{d}\mu_2(s_2). \tag{IX.7}$$

We have:

(a) *If f is non-negative and measurable, $f \in \mathrm{m}(\mathscr{S}_1 \otimes \mathscr{S}_2)^+$, then the integrals* (IX.5)*,* (IX.6)*, and* (IX.7) *are all in $[0, +\infty]$, and they are all equal.*
(b) *If f is integrable, $f \in \mathcal{L}^1(\mu_1 \otimes \mu_2)$, then the integrals* (IX.5)*,* (IX.6)*, and* (IX.7) *are all in $\mathbb{R}$, and they are all equal.*

Proof of part (a): We first claim that for all $f \in \mathrm{b}(\mathscr{S}_1 \otimes \mathscr{S}_2)$, then the integrals (IX.5), (IX.6), and (IX.7) are all equal as real numbers. The equality of the last two was in fact shown in part (iii) of Lemma IX.5, and here one proceeds similarly. Consider the collection $\mathscr{H}$ of functions $f \in \mathrm{b}(\mathscr{S}_1 \otimes \mathscr{S}_2)$ for which we have (i), (ii), and

(iii') The integrals (IX.5), (IX.6), and (IX.7) are all equal as real numbers.

With minor modifications to the proof of Lemma IX.5, one can prove that this collection $\mathscr{H}$ is a monotone class and contains indicators $\mathbb{I}_{A_1\times A_2}$ of sets $A_1 \times A_2$, and therefore $\mathscr{H} = \mathrm{b}(\mathscr{S}_1 \otimes \mathscr{S}_2)$.

It in particular follows, because simple functions are necessarily bounded, that all non-negative simple functions satisfy (iii'). For a given non-negative measurable $f \in \mathrm{m}(\mathscr{S}_1 \otimes \mathscr{S}_2)^+$ we choose a pointwise increasing approximation $f_n \uparrow f$ by non-negative simple functions $f_n \in \mathrm{s}(\mathscr{S}_1 \otimes \mathscr{S}_2)^+$. By (iii'), for these approximating functions we have the equalities

$$\begin{aligned}\int_{S_1\times S_2} f_n \,\mathrm{d}(\mu_1 \otimes \mu_2) &= \int_{S_1} \Big(\int_{S_2} f_n(s_1, s_2)\,\mathrm{d}\mu_2(s_2)\Big)\,\mathrm{d}\mu_1(s_1)\\ &= \int_{S_2} \Big(\int_{S_1} f_n(s_1, s_2)\,\mathrm{d}\mu_1(s_1)\Big)\,\mathrm{d}\mu_2(s_2).\end{aligned}$$

We show that these three increase to (IX.5), (IX.6), and (IX.7) for f, respectively, and the assertion (a) then follows. For the first, just use Monotone convergence theorem for the product measure $\mu_1 \otimes \mu_2$

$$\int_{S_1\times S_2} f_n \,\mathrm{d}(\mu_1 \otimes \mu_2) \;\uparrow\; \int_{S_1\times S_2} f \,\mathrm{d}(\mu_1 \otimes \mu_2).$$

Let us consider the second integral in the limit $n \to \infty$. Use first Monotone convergence theorem for μ_2 to get

$$\int_{S_2} f_n(s_1, s_2)\,\mathrm{d}\mu_2(s_2) \;\uparrow\; \int_{S_2} f(s_1, s_2)\,\mathrm{d}\mu_2(s_2),$$

which in particular shows that the inner integral in (IX.6) defines a non-negative measurable function of the variable s_1, as a pointwise limit of non-negative measurable functions. Use next the Monotone convergence theorem for μ_1 together with the previous result to get

$$\int_{S_1} \Big(\int_{S_2} f_n(s_1, s_2)\,\mathrm{d}\mu_2(s_2)\Big)\,\mathrm{d}\mu_1(s_1) \;\uparrow\; \int_{S_1} \Big(\int_{S_2} f(s_1, s_2)\,\mathrm{d}\mu_2(s_2)\Big)\,\mathrm{d}\mu_1(s_1).$$

The third one is shown similarly, with Monotone convergence theorem for μ_1 first and then for μ_2. This finishes the proof of (a).

Proof of part (b): Suppose that $f \in \mathcal{L}^1(\mu_1 \otimes \mu_2)$, i.e., that f is $\mathscr{S}_1 \otimes \mathscr{S}_2$-measurable and

$$\int_{S_1 \times S_2} |f| \, \mathrm{d}(\mu_1 \otimes \mu_2) < +\infty.$$

Write $f = f_+ - f_-$, where $f_+, f_- \in \mathrm{m}(\mathscr{S}_1 \otimes \mathscr{S}_2)^+$ are the positive and negative parts of f. From integrability of f, it follows that we have

$$\int_{S_1 \times S_2} f_+ \, \mathrm{d}(\mu_1 \otimes \mu_2) < +\infty \qquad \text{and} \qquad \int_{S_1 \times S_2} f_- \, \mathrm{d}(\mu_1 \otimes \mu_2) < +\infty.$$

We can then apply part (a) to f_+ to get that all of the following integrals finite and equal

$$\begin{aligned} \int_{S_1 \times S_2} f_+ \, \mathrm{d}(\mu_1 \otimes \mu_2) &= \int_{S_1} \Big(\underbrace{\int_{S_2} f_+(s_1, s_2) \, \mathrm{d}\mu_2(s_2)}_{=: \mathfrak{J}_1^{f_+}(s_1)} \Big) \mathrm{d}\mu_1(s_1) \\ &= \int_{S_2} \Big(\underbrace{\int_{S_1} f_+(s_1, s_2) \, \mathrm{d}\mu_1(s_1)}_{=: \mathfrak{J}_2^{f_+}(s_2)} \Big) \mathrm{d}\mu_2(s_2), \end{aligned}$$

and similarly for f_-. In particular this implies that for μ_1-almost every s_1 we have that $\mathfrak{J}_1^{f_+}(s_1) < +\infty$ and $\mathfrak{J}_1^{f_-}(s_1) < +\infty$. Noting that the positive and negative parts of $\mathfrak{J}_1^f$ are $(\mathfrak{J}_1^f)_+ = \mathfrak{J}_1^{f_+}$ and $(\mathfrak{J}_1^f)_- = \mathfrak{J}_1^{f_-}$, we furthermore see that $\mathfrak{J}_1^f \in \mathcal{L}^1(\mu_1)$. Similarly we get $\mathfrak{J}_2^f \in \mathcal{L}^1(\mu_2)$. Then comparing the definition of the integral (IX.5)

$$\int_{S_1 \times S_2} f \, \mathrm{d}(\mu_1 \otimes \mu_2) = \int_{S_1 \times S_2} f_+ \, \mathrm{d}(\mu_1 \otimes \mu_2) - \int_{S_1 \times S_2} f_- \, \mathrm{d}(\mu_1 \otimes \mu_2)$$

with the definitions of the double integrals (IX.6) and (IX.7) in terms of the positive and negative parts, the asserted equalities follow. □

Remark IX.10. In part (b) of Theorem IX.9 we assumed the integrability $f \in \mathcal{L}^1(\mu_1 \otimes \mu_2)$ with respect to the product measure $\mu_1 \otimes \mu_2$. By part (a), however, this integrability follows if f is measurable and either one of the double integrals

$$\int_{S_1} \Big(\int_{S_2} |f(s_1, s_2)| \, \mathrm{d}\mu_2(s_2) \Big) \mathrm{d}\mu_1(s_1) \quad \text{or} \quad \int_{S_2} \Big(\int_{S_1} |f(s_1, s_2)| \, \mathrm{d}\mu_1(s_1) \Big) \mathrm{d}\mu_2(s_2)$$

of the absolute value are finite.

IX.2.2. Product of two sigma finite measures

In Section IX.2.1 we constructed the product measure assuming that the two measures we started with were finite. Our next goal is to relax the assumption of finiteness. Let us, however, start by pointing out that it is necessary to make some assumptions in order to build a theory powerful enough to allow interchanging order of integrations. The following exercise contains a counterexample to the validity of such interchange property.

Exercise IX.3 (It is not always possible to change the order of integrations)**.** Consider the following two measure spaces $(S_1, \mathscr{S}_1, \mu_1)$ and $(S_2, \mathscr{S}_2, \mu_2)$. The underlying set for both is the unit interval,

$$S_1 = [0, 1] \qquad \text{and} \qquad S_2 = [0, 1],$$

the σ-algebra for both is the Borel σ-algebra

$$\mathscr{S}_1 = \mathscr{B}([0, 1]) \qquad \text{and} \qquad \mathscr{S}_2 = \mathscr{B}([0, 1])$$

of the interval. The first measure is taken to be the restriction of the Lebesgue measure

$$\mu_1 = \Lambda \qquad (\mu_1[(a, b)] = b - a)$$

to the interval, and the second measure is taken to be the counting measure

$$\mu_2 = \mu_{\#} \qquad (\mu_2[B] = \#B)$$

on the interval. Define the function $f\colon [0,1] \times [0,1] \to \mathbb{R}$ by

$$f(x,y) = \begin{cases} 1 & \text{if } x = y \\ 0 & \text{if } x \neq y. \end{cases}$$

(a) Verify that that f is a non-negative function, which is $\mathscr{S}_1 \otimes \mathscr{S}_2$-measurable.
(b) Calculate both

$$\int_{[0,1]} \Big(\int_{[0,1]} f(x,y)\, \mathrm{d}\mu_{\#}(y) \Big) \mathrm{d}\Lambda(x) \quad \text{and} \quad \int_{[0,1]} \Big(\int_{[0,1]} f(x,y)\, \mathrm{d}\Lambda(x) \Big) \mathrm{d}\mu_{\#}(y)$$

and compare the results.

Definition IX.11 (Sigma finite measures and measure spaces)**.**
Let $(S, \mathscr{S}, \mu)$ be a measure space. We say that the measure space is *σ-finite* or that the measure μ is *σ-finite* if there exists a sequence $A_1, A_2, \ldots \in \mathscr{S}$ of measurable sets which together cover the entire space, $S = \bigcup_{n=1}^{\infty} A_n$, and each has finite measure, $\mu[A_n] < +\infty$ for all $n \in \mathbb{N}$.

Example IX.12 (The Lebesgue measure is sigma finite)**.**
The real axis with the Lebesgue measure $(\mathbb{R}, \mathscr{B}, \Lambda)$ is a σ-finite measure space. An example of a sequence of finite measure sets which covers the real axis is $A_n = [-n, +n]$, for $n \in \mathbb{N}$.

Example IX.13 (The d-dimensional Lebesgue measure is sigma finite)**.**
The d-dimensional Euclidean space with the d-dimensional Lebesgue measure $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d), \Lambda^d)$ is a σ-finite measure space. An example of a sequence of finite measure sets which covers the real axis is balls $A_n = \big\{ x \in \mathbb{R}^d \;\big|\; \|x\| \leq n \big\}$ of radii n centered at the origin, for $n \in \mathbb{N}$.

The following (counter-)example underlies the problem encountered in Exercise IX.3.

Example IX.14 (The counting measure is sigma finite only on countable sets)**.**
The counting measure $\mu_{\#}$ on $(S, \mathscr{S})$ is σ-finite if and only if S is a countable set. (Recall that the countable union of finite sets is countable!)

Remark IX.15 (Sigma finiteness with disjoint pieces of finite measure)**.**
If $(S, \mathscr{S}, \mu)$ is a σ-finite measure space, then it is possible to choose disjoint measurable sets $A'_1, A'_2, \ldots \in \mathscr{S}$ which cover the entire space $\bigcup_{n=1}^{\infty} A'_n = S$ and each has finite measure $\mu[A'_n] < +\infty$ for all $n \in \mathbb{N}$. Indeed, if $A_1, A_2, \ldots$ form a sequence as in Definition IX.11, then it suffices to set $A'_n := A_n \setminus (A_1 \cup \cdots \cup A_{n-1})$ to achieve disjointness.

Suppose that $(S_1, \mathscr{S}_1, \mu_1)$ and $(S_2, \mathscr{S}_2, \mu_2)$ are two σ-finite measure spaces, and that $A_1^{(1)}, A_2^{(1)}, \ldots \in \mathscr{S}_1$ and $A_1^{(2)}, A_2^{(2)}, \ldots \in \mathscr{S}_2$ are their disjoint coverings by finite measure pieces,

$$S_1 = \bigcup_{n=1}^{\infty} A_n^{(1)}\,, \qquad S_2 = \bigcup_{n=1}^{\infty} A_n^{(2)}$$
$$\mu_1\big[A_n^{(1)}\big] < +\infty\,, \qquad \mu_2\big[A_n^{(2)}\big] < +\infty \qquad \forall n \in \mathbb{N}.$$

For any $n, m \in \mathbb{N}$, the truncation $\mu_1^{(n)}$ of μ_1 to $A_n^{(1)} \subset S_1$ and the truncation $\mu_2^{(m)}$ of μ_2 to $A_m^{(2)} \subset S_2$ are two finite measures (see Exercise II.1), so we know from

Section IX.2.1 how to construct on the product measure

$$\mu_1^{(n)} \otimes \mu_2^{(m)}.$$

The product of the two σ-finite measures can then be defined as the countable sum of these pieces,

$$(\mu_1 \otimes \mu_2)[B] := \sum_{n,m\in\mathbb{N}} \big(\mu_1^{(n)} \otimes \mu_2^{(m)}\big)[B].$$

We leave it as an exercise to the reader to check that this definition does not depend on the chosen pieces $A_n^{(i)}$.[1] Fubini's theorem continues to hold for products of σ-finite measures. The proof can be done by splitting to the countably many pieces $A_n^{(1)} \times A_m^{(2)} \subset S_1 \times S_2$ of finite measure.

[1]Hint: The formula $(\mu_1 \otimes \mu_2)[A_1 \times A_2] = \mu_1[A_1]\,\mu_2[A_2]$ holds for any sets $A_1 \in \mathscr{S}_1$, $A_2 \in \mathscr{S}_2$ of finite measure, and Dynkin's identification theorem can then be used to characterize the restrictions of the product measure to any pieces of finite measure.

Lecture X

Probability on product spaces

In this lecture we look at some of the uses of product sigma algebras, product measures, and Fubini's theorem.

We will in particular consider random vectors and joint distributions of random variables.

X.1. Joint laws

Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space. Recall from Definition III.6 that, in general, the law of a random variable $X\colon \Omega \to S$ is the probability measure on $(S, \mathscr{S})$ given by

$$P_X[B] = \mathsf{P}\big[X \in B\big].$$

As an example, for a real valued random variable $X\colon \Omega \to \mathbb{R}$, the law P_X is a probability measure on $(\mathbb{R}, \mathscr{B})$. To appreciate its role, recall from Theorem VIII.1 that for Borel functions $h\colon \mathbb{R} \to \mathbb{R}$ we have

$$\mathsf{E}\big[h(X)\big] = \int_{\mathbb{R}} h(x)\, \mathrm{d}P_X(x),$$

whenever $h(X) \in \mathcal{L}^1(\mathsf{P})$ or equivalently $h \in \mathcal{L}^1(P_X)$. Below we generalize the considerations to several real valued random variables.

Definition of the joint law of two real random variables

Consider then two real valued random variables X and Y defined on the same probability space $(\Omega, \mathscr{F}, \mathsf{P})$.

Definition X.1 (Joint law)**.**
For $X\colon \Omega \to \mathbb{R}$ and $Y\colon \Omega \to \mathbb{R}$, the *joint law* (or *joint distribution*) of X and Y is the probability measure on $(\mathbb{R}^2, \mathscr{B}(\mathbb{R}^2))$ given by

$$P_{X,Y}[A] = \mathsf{P}\big[(X, Y) \in A\big].$$

Remark X.2 (Joint law as the law of a random vector)**.**
The pair (X, Y) is a "random vector" in $\mathbb{R}^2$ — or more precisely

$$\begin{aligned} Z\colon\ & \Omega \to \mathbb{R}^2 \\ & \omega \mapsto Z(\omega) = \big(X(\omega), Y(\omega)\big) \end{aligned}$$

is. Indeed, we check that $Z\colon \Omega \to \mathbb{R}^2$ is $\mathscr{F}/\mathscr{B}(\mathbb{R}^2)$-measurable, and thus really a random variable with values in $\mathbb{R}^2$. To check this measurability, note first that for any $B_1, B_2 \in \mathscr{B}$,

we have

$$\begin{aligned} Z^{-1}(B_1 \times B_2) &= \Big\{\omega \in \Omega \;\Big|\; Z(\omega) \in B_1 \times B_2\Big\} \\ &= \Big\{\omega \in \Omega \;\Big|\; X(\omega) \in B_1,\ Y(\omega) \in B_2\Big\} \\ &= X^{-1}(B_1) \,\cap\, Y^{-1}(B_2) \end{aligned}$$

which is an event in $\mathscr{F}$ by the measurability of X and Y. The collection

$$\mathscr{J} = \big\{B_1 \times B_2 \;\big|\; B_1, B_2 \in \mathscr{B}\big\}$$

generates $\mathscr{B}\otimes\mathscr{B} = \mathscr{B}(\mathbb{R}^2)$ by Lemma IX.3 and Exercise IX.2, so this observation is sufficient for the measurability of $Z\colon \Omega \to \mathbb{R}^2$ according to Lemma III.9.

The joint law $P_{X,Y}$ of X and Y is thus nothing but the law P_Z of the random vector $Z = (X, Y)$.

Analogously to Theorem VIII.1, for a Borel function $h\colon \mathbb{R}^2 \to \mathbb{R}$ one can write the expected value of $h(X, Y)$ in terms of the law $P_{X,Y}$ as

$$\mathsf{E}\big[h(X,Y)\big] = \int_{\mathbb{R}^2} h(x,y)\,\mathrm{d}P_{X,Y}(x,y), \tag{X.1}$$

provided that $h(X, Y) \in \mathcal{L}^1(\mathsf{P})$ or equivalently $h \in \mathcal{L}^1(P_{X,Y})$. We leave the detailed verification of this to the reader.[1]

Lemma X.3 (Characterization of joint law)**.**
The law $P_{X,Y}$ is uniquely characterized by the property that

$$P_{X,Y}[B_1 \times B_2] = \mathsf{P}\big[(X,Y) \in B_1 \times B_2\big]$$

for all $B_1, B_2 \in \mathscr{B}$.

Proof. Recall that the collection $\mathscr{J}$ of sets of the form $B_1 \times B_2 \subset \mathbb{R}^2$ is a π-system which generates the product σ-algebra $\mathscr{B} \otimes \mathscr{B}$ (Lemma IX.3). By Dynkin's identification theorem (Theorem II.26), the probability measure $P_{X,Y}$ is uniquely characterized by its restriction to a generating π-system, so the assertion follows. □

The following example of joint laws is relevant to discrete time Markov processes on very general state spaces.

Exercise X.1 (Transition probability kernels)**.**
Let K be a transition probability kernel on $(S, \mathscr{S})$, i.e., a mapping $S\times\mathscr{S} \to [0, +\infty)$ denoted by $(s, A) \mapsto K_s[A]$ such that

- for any $A \in \mathscr{S}$, the mapping $s \mapsto K_s[A]$ is $\mathscr{S}$-measurable $S \to [0, +\infty)$
- for any $s \in S$, the mapping $A \mapsto K_s[A]$ is a probability measure on $(S, \mathscr{S})$.

Let μ be a probability measure on $(S, \mathscr{S})$.

(a) Define μK by

$$(\mu K)[A] = \int_S K_s[A]\,\mathrm{d}\mu(s), \qquad \text{for } A \in \mathscr{S}.$$

Show that μK is a probability measure on $(S, \mathscr{S})$.

[1] The idea is similar to Theorem VIII.1, but for this case, using the Monotone class theorem may be more convenient.

(b) Define, for $\mathscr{S} \otimes \mathscr{S}$-measurable subsets $B \subset S \times S$

$$\nu[B] = \int_S \left(\int_S \mathbb{I}_B(s_1, s_2) \, \mathrm{d}K_{s_1}(s_2) \right) \mathrm{d}\mu(s_1).$$

Show that ν is a probability measure on $(S \times S, \mathscr{S} \otimes \mathscr{S})$.

(c) Let $X = (X_1, X_2)$ be a random "vector" in $S \times S$ with distribution $P_X = \nu$ given by (b). Show that the distributions P_{X_1} and P_{X_2} of its components X_1 and X_2 are μ and μK, respectively.

Note: *(X_1, X_2) can be viewed as the first two values of a discrete time Markov process with initial distribution μ and transition kernel K (a measure valued generalization of transition matrix) on the state space S. The distribution of any number of first steps of the Markov process can be defined by generalizing the above construction. The existence of the whole Markov process can then be deduced from straightforward abstract extension theorems.*

The notion of joint law of n real valued random variables is a straightforward generalization — and it is nothing but the law of the n-dimensional random vector whose components are the real valued random variables.

Joint densities

Recall from (VIII.1) that a real valued random variable $X \colon \Omega \to \mathbb{R}$ is said to have a *continuous distribution*, if there is a Borel-measurable function $f_X \colon \mathbb{R} \to [0, +\infty]$ such that we have

$$P_X[B] = \mathsf{P}\big[X \in B\big] = \int_B f_X \, \mathrm{d}\Lambda \qquad \text{for all } B \in \mathscr{B}. \tag{X.2}$$

In the above formula and below we use the convention (VII.10) for integrals over subsets. The function f_X is called the *probability density* of (the law of) X.

The Lebesgue measure Λ^2 on $\mathbb{R}^2$ is the product measure $\Lambda^2 = \Lambda \otimes \Lambda$ of two Lebesgue measures on $\mathbb{R}$. It corresponds to the "area measure", as it is for example determined by the measures

$$\Lambda^2\big[(a_1, b_1) \times (a_2, b_2)\big] = (b_1 - a_1)\,(b_2 - a_2)$$

of rectangles with $a_1 \le b_1$, $a_2 \le b_2$.

Two random variables $X \colon \Omega \to \mathbb{R}$ and $Y \colon \Omega \to \mathbb{R}$ are said to have a *joint density* $f_{X,Y} \colon \mathbb{R}^2 \to [0, +\infty]$ if we have

$$P_{X,Y}[A] = \mathsf{P}\big[(X, Y) \in A\big] = \int_A f_{X,Y} \, \mathrm{d}\Lambda^2 \qquad \text{for all } A \in \mathscr{B}(\mathbb{R}^2). \tag{X.3}$$

The notion of joint density of more than two random variables is a straightforward generalization.

Proposition X.4 (Marginal densities from joint density)**.**
If X and Y have a joint density $f_{X,Y} \colon \mathbb{R}^2 \to [0, +\infty]$, then X has density $f_X \colon \mathbb{R} \to [0, +\infty]$ given by

$$f_X(x) = \int_{\mathbb{R}} f_{X,Y}(x, y) \, \mathrm{d}\Lambda(y)$$

and Y has density $f_Y\colon \mathbb{R} \to [0,+\infty]$ given by

$$f_Y(y) = \int_{\mathbb{R}} f_{X,Y}(x,y)\, \mathrm{d}\Lambda(x).$$

Proof. We prove this using Fubini's theorem, Theorem IX.9.

Define f_X by the formula in the statement. We have to verify that it is a density for X. Let $B \in \mathscr{B}$. Write

$$P_X[B] = \mathsf{P}\Big[X \in B\Big] = \mathsf{P}\Big[X \in B \text{ and } Y \in \mathbb{R}\Big] = \mathsf{P}\Big[(X,Y) \in B \times \mathbb{R}\Big]$$

and then use the fact that $f_{X,Y}$ is a joint density to write this as

$$P_X[B] = \int_{B\times\mathbb{R}} f_{X,Y}\, \mathrm{d}\Lambda^2 = \int_{\mathbb{R}^2} \mathbb{I}_{B\times\mathbb{R}}\, f_{X,Y}\, \mathrm{d}\Lambda^2.$$

Now observe that that $\mathbb{I}_{B\times\mathbb{R}}(x,y) = \mathbb{I}_B(x)$ and recall that $\Lambda^2 = \Lambda \otimes \Lambda$ is a product measure, and use Fubini's theorem to get

$$\begin{aligned} P_X[B] &= \int_{\mathbb{R}^2} \mathbb{I}_B(x)\, f_{X,Y}(x,y)\, \mathrm{d}(\Lambda\otimes\Lambda)(x,y) \\ &= \int_{\mathbb{R}} \Big(\int_{\mathbb{R}} \mathbb{I}_B(x)\, f_{X,Y}(x,y)\, \mathrm{d}\Lambda(y)\Big)\, \mathrm{d}\Lambda(x). \end{aligned}$$

The inner integral here is

$$\int_{\mathbb{R}} \mathbb{I}_B(x)\, f_{X,Y}(x,y)\, \mathrm{d}\Lambda(y) = \mathbb{I}_B(x) \int_{\mathbb{R}} f_{X,Y}(x,y)\, \mathrm{d}\Lambda(y) = \mathbb{I}_B(x)\, f_X(x).$$

Therefore, we have obtained

$$P_X[B] = \int_{\mathbb{R}} \mathbb{I}_B(x)\, f_X(x)\, \mathrm{d}\Lambda(x) = \int_B f_X(x)\, \mathrm{d}\Lambda(x),$$

which shows that f_X is a density for X.

The claim about the density of Y is proven similarly. □

In summary, Proposition X.4 says that the existence of a joint density for a random vector guarantees the existence of probability densities for its components (called *marginal densities*). The converse does not hold, in general.

Example X.5 (Marginal densities do not guarantee existence of joint density)**.**
Let X be a real valued random variable with continuous distribution and density $f_X\colon \mathbb{R} \to [0,+\infty]$. Define a two-dimensional random vector $Z = (X,X)$ whose two components are equal. Then both components have continuous distribution with density f_X. The joint law $P_{X,X}$, however, is supported on the line $\big\{(x,y)\,\big|\, x = y\big\} \subset \mathbb{R}^2$, which has measure zero (under the 2-dimensional Lebesgue measure Λ^2). From this, it is easy to see that there can not exist a joint density for $Z = (X,X)$.

Of course not all probability distributions have densities. The following exercise concerns a very simple probability distribution in the d-dimensional Euclidean space.

Exercise X.2 (Dirac measure on $\mathbb{R}^d$)**.**
The *Dirac measure* at a point $a \in \mathbb{R}^d$ is defined by

$$\delta_a[A] = \begin{cases} 1 & \text{if } a \in A, \\ 0 & \text{otherwise.} \end{cases}$$

(a) Show that δ_a is a probability measure on $(\mathbb{R}^d, \mathscr{B}(\mathbb{R}^d))$.
(b) Show that any Borel function $f : \mathbb{R}^d \to \mathbb{R}$ is δ_a-integrable (meaning that $\int_{\mathbb{R}^d} |f(x)|\, \mathrm{d}\delta_a(x) < +\infty$), and compute the integral $\int_{\mathbb{R}^d} f(x)\, \mathrm{d}\delta_a(x)$.

Consider now the case $d = 1$, and fix $a \in \mathbb{R}$.

(c) Does the measure δ_a have a probability density function, i.e., a Borel function $g\colon \mathbb{R}^n \to [0, +\infty]$ such that $\delta_a[A] = \int_{\mathbb{R}} \mathbb{I}_A(x)\, g(x)\, \mathrm{d}x$ for all $A \in \mathscr{B}$? If yes, find out an expression for it. If not, explain why.

X.2. Variances and covariances of square integrable random variables

Variances are important statistics of distributions of real valued random variables, and covariances are important statistics of the joint distributions of pairs of random variables. In order for these to be well-defined, we need the second moments of the random variables to exist.

The general notion of p-integrability was introduced in Definition VIII.8. For the present purposes, we are concerned with the particular case $p = 2$. A random variable $X \in \mathcal{L}^2(\mathsf{P})$ is said to be *square integrable* — we recall that this means the following finiteness of second moment

$$\mathsf{E}\big[X^2\big] < +\infty.$$

By Lemma VIII.10 we have the vector space property of square integrable random variables: if $X, Y \in \mathcal{L}^2(\mathsf{P})$ and $a, b \in \mathbb{R}$, then also $aX + bY \in \mathcal{L}^2(\mathsf{P})$. By Lemma VIII.9, square integrability implies in particular integrability, and thus $\mathcal{L}^2(\mathsf{P}) \subset \mathcal{L}^1(\mathsf{P})$ is a vector subspace in the space of integrable random variables.

The following is a fundamentally important inequality for square integrable random variables.

Theorem X.6 (Cauchy-Schwarz inequality)**.**
Suppose that $X, Y \in \mathcal{L}^2(\mathsf{P})$ are two square integrable random variables. Then the product XY is integrable, $XY \in \mathcal{L}^1(\mathsf{P})$, and we have

$$\Big|\, \mathsf{E}\big[XY\big] \,\Big| \leq \sqrt{\mathsf{E}\big[X^2\big]\ \mathsf{E}\big[Y^2\big]}. \tag{X.4}$$

Proof. For any two real numbers $x, y \in \mathbb{R}$ we have the following inequality

$$0 \leq (x - y)^2 = x^2 - 2xy + y^2.$$

By moving the cross-term to the other side and dividing by two we get $xy \leq \frac{1}{2}x^2 + \frac{1}{2}y^2$. By changing the sign of one of the numbers, we get also $-xy \leq \frac{1}{2}x^2 + \frac{1}{2}y^2$. Together these give

$$\big|xy\big| \leq \frac{1}{2}x^2 + \frac{1}{2}y^2.$$

Applying the above inequality to the values of the random variables X and Y, we get

$$\big|X(\omega)\, Y(\omega)\big| \leq \frac{1}{2}X(\omega)^2 + \frac{1}{2}Y(\omega)^2 \qquad \text{for all } \omega \in \Omega.$$

Taking the expected values and using monotonicity and linearity, as well as the assumption $X, Y \in \mathcal{L}^2(\mathsf{P})$, we deduce

$$\mathsf{E}\Big[\big|X\, Y\big|\Big] \leq \frac{1}{2}\,\mathsf{E}\big[X^2\big] + \frac{1}{2}\,\mathsf{E}\big[Y^2\big] < +\infty.$$

This shows that the product XY is indeed integrable, $XY \in \mathcal{L}^1(\mathsf{P})$.

Next observe that for any $t \in \mathbb{R}$ and any $\omega \in \Omega$ we have $0 \leq \big(tX(\omega) + Y(\omega)\big)^2$. Taking expected values and using linearity and monotonicity we get that

$$0 \leq \mathsf{E}\Big[\big(tX + Y\big)^2\Big] = \mathsf{E}\Big[t^2X^2 + 2tXY + Y^2\Big] = t^2\,\mathsf{E}\big[X^2\big] + 2t\,\mathsf{E}\big[XY\big] + \mathsf{E}\big[Y^2\big].$$

This is a quadratic polynomial in t which never becomes negative, so it can have at most one real root, and its discriminant is therefore either negative (if there are no roots) or zero (if there is one root), i.e.,

$$4\,\mathsf{E}\big[XY\big]^2 - 4\,\mathsf{E}\big[X^2\big]\,\mathsf{E}\big[Y^2\big] \leq 0.$$

Moving the second term to the other side and dividing by 4 yields

$$\mathsf{E}\big[XY\big]^2 \leq \mathsf{E}\big[X^2\big]\,\mathsf{E}\big[Y^2\big]. \tag{X.5}$$

Taking the square roots then gives the asserted inequality. □

Corollary X.7 (Square of expected value vs. expected value of square)**.**
If $X \in \mathcal{L}^2(\mathsf{P})$, then we have also $X \in \mathcal{L}^1(\mathsf{P})$ and

$$\mathsf{E}\big[X\big]^2 \leq \mathsf{E}\big[X^2\big]. \tag{X.6}$$

Proof. Note that the constant random variable 1 is square integrable, $1 \in \mathcal{L}^2(\mathsf{P})$. The asserted inequality is derived by applying the squared Cauchy-Schwarz inequality (X.5) to X and 1 as follows

$$\mathsf{E}\big[X\big]^2 = \mathsf{E}\big[X \cdot 1\big]^2 \overset{\text{(X.5)}}{\leq} \mathsf{E}\big[X^2\big]\,\underbrace{\mathsf{E}\big[1^2\big]}_{=1} = \mathsf{E}\big[X^2\big].$$

□

Suppose now that $X, Y \in \mathcal{L}^2(\mathsf{P})$ are square integrable random variables. Since we have $\mathcal{L}^2(\mathsf{P}) \subset \mathcal{L}^1(\mathsf{P})$ (Lemma VIII.9) the expected values $\mathsf{E}[X]$ and $\mathsf{E}[Y]$ are well-defined and finite. Moreover, since adding constants does not affect $\mathcal{L}^2(\mathsf{P})$ membership (Corollary VIII.11), we have also $X - \mathsf{E}[X] \in \mathcal{L}^2(\mathsf{P})$ and $Y - \mathsf{E}[Y] \in \mathcal{L}^2(\mathsf{P})$. By Theorem X.6 then, we have $(X - \mathsf{E}[X])(Y - \mathsf{E}[Y]) \in \mathcal{L}^1(\mathsf{P})$. Therefore the following definition makes sense.

Definition X.8 (Variance and covariance)**.**
The *variance* of a square integrable random variable $X \in \mathcal{L}^2(\mathsf{P})$ is

$$\mathsf{Var}(X) = \mathsf{E}\Big[\big(X - \mathsf{E}[X]\big)^2\Big].$$

The *covariance* of two square integrable random variables $X, Y \in \mathcal{L}^2(\mathsf{P})$ is

$$\mathsf{Cov}(X, Y) = \mathsf{E}\Big[\big(X - \mathsf{E}[X]\big)\,\big(Y - \mathsf{E}[Y]\big)\Big].$$

Proposition X.9 (Formulas for variance and covariance)**.**
For $X \in \mathcal{L}^2(\mathsf{P})$, we have

$$\mathsf{Var}(X) = \mathsf{E}\big[X^2\big] - \mathsf{E}\big[X\big]^2. \tag{X.7}$$

For $X, Y \in \mathcal{L}^2(\mathsf{P})$, we have

$$\mathsf{Cov}(X, Y) = \mathsf{E}\big[X\,Y\big] - \mathsf{E}\big[X\big]\,\mathsf{E}\big[Y\big]. \tag{X.8}$$

Proof. The first formula (X.7) is in fact a special case of the second (with $Y = X$), so it is enough to prove (X.8). Denote $\mathfrak{m}_X = \mathsf{E}[X]$ and $\mathfrak{m}_Y = \mathsf{E}[Y]$. Then calculate, by expanding and using linearity of expected values,

$$\begin{aligned}\mathsf{Cov}(X,Y) &= \mathsf{E}\Big[\big(X-\mathfrak{m}_X\big)\big(Y-\mathfrak{m}_Y\big)\Big] \\ &= \mathsf{E}\Big[XY - X\,\mathfrak{m}_Y - \mathfrak{m}_X\,Y + \mathfrak{m}_X\,\mathfrak{m}_Y\Big] \\ &= \mathsf{E}\big[XY\big] - \underbrace{\mathsf{E}[X]}_{=\mathfrak{m}_X}\mathfrak{m}_Y - \mathfrak{m}_X\,\underbrace{\mathsf{E}[Y]}_{=\mathfrak{m}_Y} + \mathfrak{m}_X\,\mathfrak{m}_Y \\ &= \mathsf{E}\big[XY\big] - \mathfrak{m}_X\,\mathfrak{m}_Y.\end{aligned}$$

This is the asserted formula. □

X.3. Independence and products

Equivalent conditions for independence of random numbers

Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space, and $X, Y\colon \Omega \to \mathbb{R}$ two real valued random variables on it. Denote the laws of X and Y by P_X and P_Y, respectively, and the joint law of X and Y by $P_{X,Y}$.

Theorem X.10 (Independence in terms of laws)**.**
For two real valued random variables $X, Y\colon \Omega \to \mathbb{R}$, the following conditions are equivalent:

(i) *X and Y are independent (i.e., $X \perp\!\!\!\perp Y$)*
(ii) *$P_{X,Y} = P_X \otimes P_Y$*
(iii) *for all $x, y \in \mathbb{R}$ we have $\mathsf{P}\big[X \le x,\, Y \le y\big] = \mathsf{P}\big[X \le x\big]\,\mathsf{P}\big[Y \le y\big]$.*

If moreover X and Y have a joint density $f_{X,Y}\colon \mathbb{R}^2 \to [0, +\infty]$, then (i), (ii), *and* (iii) *are also equivalent to*

(iv) *X and Y have densities f_X and f_Y, respectively, such that $f_{X,Y}(x,y) = f_X(x)\,f_Y(y)$ for Λ^2-almost all (x,y).*

In the proof, we repeatedly use the characterization of the product measure based on Lemma IX.8: if P and Q are two probability measures on $\mathbb{R}$, then $P \otimes Q$ is the unique measure on $\mathbb{R}^2$ such that

$$(P \otimes Q)[B_1 \times B_2] = P[B_1]\,Q[B_2]$$

for all $B_1, B_2 \in \mathscr{B}$.

Proof of Theorem X.10: The equivalence "(i) ⇔ (iii)" was already proven in Corollary V.6. We prove the implications "(i) ⇒ (ii)" and "(ii) ⇒ (iii)" to establish the first assertion about the equivalence of (i), (ii), and (iii). Then, under the additional hypothesis that X and Y have a joint density $f_{X,Y}$, we prove the implication "(ii) ⇒ (iv)" and leave it as an exercise to prove "(iv) ⇒ (iii)". The second assertion follows.

proof of "(i) ⇔ (iii)*":* Corollary V.6.

proof of "(i) ⇒ (ii)*":* Suppose that we have independence $X \perp\!\!\!\perp Y$. Let $B_1, B_2 \in \mathscr{B}$. Then we have $X^{-1}(B_1) \in \sigma(X)$ and $Y^{-1}(B_2) \in \sigma(Y)$, so independence can be used to calculate

$$\begin{aligned} P_{X,Y}[B_1 \times B_2] &= \mathsf{P}\Big[X^{-1}(B_1) \cap Y^{-1}(B_2)\Big] \\ &= \mathsf{P}\Big[X^{-1}(B_1)\Big]\ \mathsf{P}\Big[Y^{-1}(B_2)\Big] \qquad \text{(because } X \perp\!\!\!\perp Y\text{)} \\ &= P_X[B_1]\ P_Y[B_2]. \end{aligned}$$

By Lemma IX.8, the product measure $P_X \otimes P_Y$ is the unique measure for which this formula holds, so we conclude that indeed $P_{X,Y} = P_X \otimes P_Y$.

proof of "(ii) ⇒ (iii)*":* Suppose that we have $P_{X,Y} = P_X \otimes P_Y$. Let $x, y \in \mathbb{R}$, and consider $B_1 = (-\infty, x]$ and $B_2 = (-\infty, y]$. Then we have

$$\begin{aligned} \mathsf{P}\Big[X \le x,\ Y \le y\Big] &= P_{X,Y}\big[B_1 \times B_2\big] \\ &= P_X\big[B_1\big]\ P_Y\big[B_2\big] \qquad \text{(since } P_{X,Y} = P_X \otimes P_Y\text{)} \\ &= \mathsf{P}\big[X \le x\big]\ \mathsf{P}\big[Y \le y\big], \end{aligned}$$

which is what we needed to show.

proof of "(iv) ⇒ (iii)*" assuming joint density:* Exercise.

proof of "(ii) ⇒ (iv)*" assuming joint density:* Suppose that $P_{X,Y} = P_X \otimes P_Y$, and assume moreover that there exists a joint density $f_{X,Y} \colon \mathbb{R}^2 \to [0, +\infty)$ for X and Y. Recall from Proposition X.4 that X and Y then have densities

$$\begin{aligned} f_X(x) &= \int_{\mathbb{R}} f_{X,Y}(x,y)\ \mathrm{d}\Lambda(y) \\ f_Y(y) &= \int_{\mathbb{R}} f_{X,Y}(x,y)\ \mathrm{d}\Lambda(x). \end{aligned}$$

Define a new measure μ on $\mathbb{R}^2$ by the formula

$$\mu[A] := \int_A f_X(x)\, f_Y(y)\ \mathrm{d}\Lambda^2(x,y) \qquad \text{for all } A \in \mathscr{B}(\mathbb{R}^2) = \mathscr{B} \otimes \mathscr{B}.$$

By Fubini's theorem with the positive measurable function $(x,y) \mapsto f_X(x)\, f_Y(y)$ in this integral, for any $B_1, B_2 \in \mathscr{B}$ we get

$$\begin{aligned} \mu[B_1 \times B_2] &= \int_{B_1 \times B_2} f_X(x)\, f_Y(y)\ \mathrm{d}\Lambda^2(x,y) \\ &= \int_{B_2} f_Y(y) \Big(\underbrace{\int_{B_1} f_X(x)\ \mathrm{d}\Lambda(x)}_{= P_X[B_1]} \Big)\ \mathrm{d}\Lambda(y) \qquad \text{(Fubini)} \\ &= P_X[B_1] \int_{B_2} f_Y(y)\ \mathrm{d}\Lambda(y) \\ &= P_X[B_1]\ P_Y[B_2]. \end{aligned}$$

Again, by Lemma IX.8 the product measure $P_X \otimes P_Y$ is the only measure for which this formula holds, so we get $\mu = P_X \otimes P_Y$. But by assumption also $P_{X,Y} = P_X \otimes P_Y$. This shows that $P_{X,Y} = \mu$, which by the construction of μ shows that $(x,y) \mapsto f_X(x)\, f_Y(y)$ is a joint density of X and Y. Since also $f_{X,Y}$ is a joint density of X and Y, we must have $f_{X,Y}(x,y) = f_X(x)\, f_Y(y)$ for Λ^2-almost all points $(x,y) \in \mathbb{R}^2$. □

Exercise X.3 (Disintegration of independent random variables)**.**
Let X and Y be independent real valued random variables with laws P_X and P_Y. Let $h : \mathbb{R}^2 \to \mathbb{R}$ be Borel function.

(a) Prove that

$$\mathsf{E}\big[|h(X,Y)|\big] = \int_{\mathbb{R}} \mathsf{E}\big[|h(x,Y)|\big]\ \mathrm{d}P_X(x) = \int_{\mathbb{R}} \mathsf{E}\big[|h(X,y)|\big]\ \mathrm{d}P_Y(y).$$

(b) Prove that when $\mathsf{E}\big[|h(X,Y)|\big] < \infty$,

$$\mathsf{E}\big[h(X,Y)\big] = \int_{\mathbb{R}} \mathsf{E}\big[h(x,Y)\big]\,\mathrm{d}P_X(x) = \int_{\mathbb{R}} \mathsf{E}\big[h(X,y)\big]\,\mathrm{d}P_Y(y).$$

Hint: *You may start from* (X.1)*. Then keep in mind product measure and Fubini's theorem.*

Exercise X.4 (Product of two uniformly distributed numbers)**.**
Let U_1 and U_2 be independent uniformly distributed random variables on $[0,1]$, so that both have the function $\mathbb{I}_{[0,1]}$ as their probability density function. Define $X = U_1\,U_2$.

(a) Calculate the cumulative distribution function $F_X(x) = \mathsf{P}\big[X \le x\big]$.
(b) What is the distribution $P_X[B] = \mathsf{P}\big[X \in B\big]$ of X?
(c) Does X have a probability density function? If yes, find out an expression for it. If not, explain why.

Independence and expected value of product

Theorem X.11 (Expected value of product of independent random variables)**.**
Suppose that $X, Y \colon \Omega \to \mathbb{R}$ are two independent random variables which are integrable, $X, Y \in \mathcal{L}^1(\mathsf{P})$. Then also their product is integrable, $XY \in \mathcal{L}^1(\mathsf{P})$, and we have

$$\mathsf{E}\big[XY\big] = \mathsf{E}\big[X\big]\ \mathsf{E}\big[Y\big]. \tag{X.9}$$

Proof. We prove the statement using the "standard machine", i.e., successively for (1): indicators, (2): simple random variables, (3): non-negative random variables, and (4): all integrable random variables.

step 1 (indicator random variables): If $X = \mathbb{I}_A$ and $Y = \mathbb{I}_B$, then $XY = \mathbb{I}_{A\cap B}$. In this case we have

$$\mathsf{E}\big[XY\big] = \mathsf{E}\big[\mathbb{I}_{A\cap B}\big] = \mathsf{P}[A \cap B].$$

and

$$\mathsf{E}[X]\ \mathsf{E}[Y] = \mathsf{E}[\mathbb{I}_A]\ \mathsf{E}[\mathbb{I}_B] = \mathsf{P}[A]\ \mathsf{P}[B].$$

Of course $A = X^{-1}(\{1\})$ and $B = Y^{-1}(\{1\})$, so by the assumed independence of X and Y, we have $\mathsf{P}[A \cap B] = \mathsf{P}[A]\ \mathsf{P}[B]$. This shows that (X.9) holds for independent indicator random variables.

step 2 (simple random variables): Suppose that $X, Y \in \mathrm{s}\mathscr{F}$ are simple random variables. Write

$$X = \sum_{i=1}^{n} a_i\,\mathbb{I}_{A_i} \qquad \text{and} \qquad Y = \sum_{j=1}^{m} b_j\,\mathbb{I}_{B_j}$$

where $a_1, \dots, a_n$ are the distinct non-zero values of X, and $b_1, \dots, b_m$ are the distinct non-zero values of Y, and $A_i = X^{-1}(\{a_i\})$ and $B_j = Y^{-1}(\{b_j\})$. Independence of X and Y shows that $A_i \perp\!\!\!\perp B_j$, and thus also $\mathbb{I}_{A_i} \perp\!\!\!\perp \mathbb{I}_{B_j}$. Now calculate, using linearity and the result

of step 0,

$$\begin{aligned}\mathsf{E}\big[XY\big] &= \mathsf{E}\Big[\sum_{i=1}^{n}\sum_{j=1}^{m} a_i\, b_j\, \mathbb{I}_{A_i}\, \mathbb{I}_{B_j}\Big] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j\ \mathsf{E}\big[\mathbb{I}_{A_i}\mathbb{I}_{B_j}\big] \\ &= \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j\ \mathsf{E}[\mathbb{I}_{A_i}]\ \mathsf{E}[\mathbb{I}_{B_j}] \qquad \text{(by step 1)} \\ &= \mathsf{E}\Big[\sum_{i=1}^{n} a_i \mathbb{I}_{A_i}\Big]\ \mathsf{E}\Big[\sum_{j=1}^{m} b_j \mathbb{I}_{B_j}\Big] = \mathsf{E}[X]\ \mathsf{E}[Y].\end{aligned}$$

This shows that (X.9) holds for independent simple random variables.

step 3 (non-negative random variables): Assume that X and Y are non-negative random variables. Let $\varsigma_n\colon [0,+\infty] \to [0,n]$ be the n:th staircase function, and define the simple functions $X_n = \varsigma_n \circ X$ and $Y_n = \varsigma_n \circ Y$. Then independence of X and Y implies the independence of X_n and Y_n also, by Exercise V.2. Moreover, by construction we have the pointwise increasing approximations, $X_n \uparrow X$ and $Y_n \uparrow Y$, and also $X_n Y_n \uparrow XY$. Monotone convergence theorem implies that as $n \to \infty$, we have $\mathsf{E}[X_n] \uparrow \mathsf{E}[X]$, $\mathsf{E}[Y_n] \uparrow \mathsf{E}[Y]$, and $\mathsf{E}[X_n\, Y_n] \uparrow \mathsf{E}[XY]$. On the other hand, by step 2, we have

$$\mathsf{E}\big[X_n\, Y_n\big] = \mathsf{E}[X_n]\ \mathsf{E}[Y_n].$$

In the limit $n \to \infty$ we now get (X.9) for independent non-negative random variables.

step 4 (integrable random variables): Consider finally general integrable random variables $X, Y \in \mathcal{L}^1(\mathsf{P})$. Split these to positive and negative parts, $X = X_+ - X_-$ and $Y = Y_+ - Y_-$, with X_+, X_-, Y_+, Y_- non-negative. We have $X_+ = \max\{X, 0\}$ etc., so by Exercise V.2, independence of X and Y implies also the independences $X_+ \perp\!\!\!\perp Y_+$, $X_+ \perp\!\!\!\perp Y_-$, $X_- \perp\!\!\!\perp Y_+$, and $X_- \perp\!\!\!\perp Y_-$. Then we calculate

$$\begin{aligned}\mathsf{E}\big[X\,Y\big] &= \mathsf{E}\Big[\big(X_+ - X_-\big)\,\big(Y_+ - Y_-\big)\Big] \\ &= \mathsf{E}\big[X_+Y_+\big] - \mathsf{E}\big[X_+Y_-\big] - \mathsf{E}\big[X_-Y_+\big] + \mathsf{E}\big[X_-Y_-\big] \\ &= \mathsf{E}[X_+]\,\mathsf{E}[Y_+] - \mathsf{E}[X_+]\,\mathsf{E}[Y_-] - \mathsf{E}[X_-]\,\mathsf{E}[Y_+] + \mathsf{E}[X_-]\,\mathsf{E}[Y_-] \qquad \text{(by step 3)} \\ &= \Big(\mathsf{E}[X_+] - \mathsf{E}[X_-]\Big)\,\Big(\mathsf{E}[Y_+] - \mathsf{E}[Y_-]\Big) \\ &= \mathsf{E}[X]\ \mathsf{E}[Y].\end{aligned}$$

This finishes the proof. □

Proposition X.12 (Variance is additive for independent random variables)**.** *If $X, Y \in \mathcal{L}^2(\mathsf{P})$ are independent, then we have*

$$\mathsf{Cov}(X, Y) = 0 \tag{X.10}$$

and

$$\mathsf{Var}(X + Y) = \mathsf{Var}(X) + \mathsf{Var}(Y) \tag{X.11}$$

Proof. By independence and Theorem X.11, we have $\mathsf{E}\big[XY\big] = \mathsf{E}[X]\,\mathsf{E}[Y]$. Using formula (X.8) then immediately shows that the covariance vanishes, $\mathsf{Cov}(X, Y) = 0$.

Then calculate the variance of $X + Y$ using formula (X.7),

$$\begin{aligned}\mathsf{Var}(X+Y) &= \mathsf{E}\big[(X+Y)^2\big] - \mathsf{E}\big[X+Y\big]^2 \\ &= \mathsf{E}\big[X^2 + 2XY + Y^2\big] - \big(\mathsf{E}[X] + \mathsf{E}[Y]\big)^2 \\ &= \mathsf{E}[X^2] + 2\,\mathsf{E}[XY] + \mathsf{E}[Y^2] - \mathsf{E}[X]^2 - 2\,\mathsf{E}[X]\,\mathsf{E}[Y] - \mathsf{E}[Y]^2.\end{aligned}$$

Using again the consequence $\mathsf{E}\big[XY\big] = \mathsf{E}[X]\,\mathsf{E}[Y]$ of independence, we have a cancellation, and we get

$$\mathsf{Var}(X+Y) = \mathsf{E}[X^2] + \mathsf{E}[Y^2] - \mathsf{E}[X]^2 - \mathsf{E}[Y]^2.$$

This leads to (X.11), if we just recall formula (X.7) for the variances of X and Y. □

Exercise X.5 (Additivity of variance for independent random variables)**.**
Suppose that $X_1, \dots, X_n$ are independent. Show that then $\mathsf{Var}(\sum_{k=1}^n X_k) = \sum_{k=1}^n \mathsf{Var}(X_k)$. Show also, by finding a counterexample, that the formula does not hold in general without the assumption of independence.

X.4. A formula for the expected value

We finish this lecture by a straightforward, yet often practical use of the Fubini's theorem in a probabilistic setup. It gives a formula for the expected value of a non-negative random variable in terms of its cumulative distribution function.

Proposition X.13 (Expected value in terms of c.d.f.)**.**
Let $X\colon \Omega \to [0,+\infty)$ be a non-negative random number. Let $F_X\colon \mathbb{R} \to [0,1]$ be its cumulative distribution function

$$F_X(x) = \mathsf{P}\big[X \le x\big].$$

Then we have

$$\mathsf{E}\big[X\big] = \int_0^\infty \big(1 - F_X(x)\big)\,\mathrm{d}x.$$

Proof. Consider the product space

$$\Omega \times [0,+\infty),$$

equipped with the product σ-algebra $\mathscr{F} \otimes \mathscr{B}([0,+\infty))$ and with the product measure $\mathsf{P} \otimes \Lambda$ of the probability measure P on the sample space Ω and the Lebesgue measure Λ restricted to the non-negative real axis $[0,+\infty)$.

Define a subset

$$A := \Big\{(\omega, t) \;\Big|\; t < X(\omega)\Big\} \;\subset\; \Omega \times [0,+\infty).$$

This subset is measurable[2], $A \in \mathscr{F} \otimes \mathscr{B}([0,+\infty))$, so its indicator is a non-negative measurable function $\mathbb{I}_A \in \mathrm{m}\big(\mathscr{F} \otimes \mathscr{B}([0,+\infty))\big)^+$ concretely given by

$$\mathbb{I}_A(\omega,t) = \begin{cases} 1 & \text{if } t < X(\omega) \\ 0 & \text{if } t \geq X(\omega). \end{cases}$$

By Fubini's theorem, Theorem IX.9, it does not matter in which order we integrate the two variables of this function:

$$\int_0^\infty \Big(\int_\Omega \mathbb{I}_A(\omega,t) \,\mathrm{d}\mathsf{P}(\omega) \Big) \,\mathrm{d}t = \int_\Omega \Big(\int_0^\infty \mathbb{I}_A(\omega,t) \,\mathrm{d}t \Big) \,\mathrm{d}\mathsf{P}(\omega). \tag{X.12}$$

The rest of the proof consists of calculating both of the above expressions separately.

Let us start from the left-hand side of (X.12). The inner integal is calculated as

$$\int_\Omega \mathbb{I}_A(\omega,t) \,\mathrm{d}\mathsf{P}(\omega) = \int_\Omega \mathbb{I}_{\{t<X\}} \,\mathrm{d}\mathsf{P} \;=\; \mathsf{P}\big[t < X\big] \;=\; 1 - \mathsf{P}\big[t \geq X\big] \;=\; 1 - F(t).$$

The double integral on the left-hand side of (X.12) is therefore

$$\int_0^\infty \big(1 - F(t)\big) \,\mathrm{d}t. \tag{X.13}$$

Consider then the right-hand side of (X.12). The inner integal is calculated as

$$\int_0^\infty \mathbb{I}_A(\omega,t) \,\mathrm{d}t = \int_0^\infty \mathbb{I}_{[0,X(\omega))} \,\mathrm{d}t \;=\; \Lambda\Big[[0,X(\omega))\Big] \;=\; X(\omega).$$

The double integral on the right-hand side of (X.12) is therefore

$$\int_\Omega X(\omega) \,\mathrm{d}\mathsf{P}(\omega) = \mathsf{E}\big[X\big]. \tag{X.14}$$

By (X.12), the results (X.13) and (X.14) are equal, which proves the assertion. □

Variations of the idea above are very commonly used. As another example of the same idea, the following two exercises characterize the p-integrability of a non-negative random variable using its cumulative distribution function.

Exercise X.6 (Moments with cumulative distribution function)**.**
Let X be a non-negative random variable, and let $F(x) = \mathsf{P}[X \leq x]$ be its cumulative distribution function. Prove that for $p \geq 1$ we have

$$\mathsf{E}\big[X^p\big] = p \int_0^\infty t^{p-1} \big(1 - F(t)\big) \,\mathrm{d}t$$

Hint: *Try using as few of the following hints as possible:*

(1) *Find a function* $f\colon [0,\infty) \to [0,\infty)$ *such that for any* $x \geq 0$ *one has* $x^p = \int_0^x f(t) \,\mathrm{d}t$.
(2) *Then we have* $X(\omega)^p = \int_0^{X(\omega)} f(t) \,\mathrm{d}t$ *for any* $\omega \in \Omega$.

---

[2] Briefly, the measurability follows from writing A as the preimage $A = g^{-1}\big((0,+\infty)\big)$ of the Borel set $(0,+\infty)$ under the function $g\colon \Omega \times [0,+\infty) \to \mathbb{R}$ defined as $g = X \circ \mathrm{pr}_1 - \mathrm{pr}_2$, which is $\mathscr{F} \otimes \mathscr{B}$-measurable. Let us, however, spell out the details below.

Observe first that

$$(\omega,t) \mapsto \omega \qquad \text{and} \qquad (\omega,t) \mapsto t$$

are measurable with respect to the product σ-algebra $\mathscr{F} \otimes \mathscr{B}$ directly according to Definition IX.2 — they are exactly the projections pr_1 and pr_2 from $\Omega \times [0,+\infty)$ to the two factors. Further composing the former of these with the measurable function $X\colon \Omega \to [0,+\infty)$ gives that $(\omega,t) \mapsto X(\omega)$ is also $\mathscr{F} \otimes \mathscr{B}$-measurable. Then as a linear combination of these measurable functions, the function given by $g(\omega,t) = X(\omega) - t$ is $\mathscr{F} \otimes \mathscr{B}$-measurable. Now we can recognize

$$A = \Big\{ (\omega,t) \;\Big|\; X(\omega) > t \Big\} = \Big\{ (\omega,t) \;\Big|\; g(\omega,t) > 0 \Big\} = g^{-1}\big((0,+\infty)\big).$$

(3) *Now write the expected value of* X^p *as*

$$\mathsf{E}\big[X^p\big] = \mathsf{E}\bigg[\int_0^{X(\omega)} f(t)\,\mathrm{d}t\bigg],$$

and try to apply Fubini's theorem — but can you handle the random upper limit of the integral?

(4) *Note that you could have alternatively written* $x^p = \int_0^x f(t)\,\mathrm{d}t = \int_0^\infty \mathbb{I}_{[0,x)}(t)\,f(t)\,\mathrm{d}t$*, and the upper limit would not have posed any problems.*

Exercise X.7 (Asymptotics of c.d.f. and p-integrability)**.**

For a given $p \geq 1$, we say that a non-negative random variable X is p-integrable and write $X \in \mathcal{L}^p$ if and only if the random variable $|X|^p$ is integrable.

Assume that for some $\alpha > 0$ the cumulative distribution function F of X satisfies

$$\lim_{x\to\infty} \Big(x^\alpha \big(1 - F(x)\big)\Big) = c > 0.$$

Prove that $X \in \mathcal{L}^p$ if and only if $p < \alpha$.

Hint: *Use the previous exercise.*

Lecture XI

Probabilistic notions of convergence

Very often in stochastics, we want to assert that some sequence of random variables tends to a limit in a suitable probabilistic sense.

Some examples of such contexts could be:

- convergence of averages $\frac{X_1+\cdots+X_n}{n}$ as $n \to \infty$
- convergence of states X_t of a stochastic process as time t increases describes the long term behavior of the process
- limits of various random quantities as some parameter of the model tends to an idealized value, e.g.,
  - size of physical system $\to \infty$ in thermodynamics and statistical physics
  - size of input data $\to \infty$ in randomized algorithms
  - signal to noise ratio $\to \infty$ in communications.

In this lecture and the next we will treat two very famous and important convergence results concerning sums of independent identically distributed real valued random variables $X_1, X_2, \ldots$:

(LLN): "*Laws of large numbers*": Under what assumptions and in which sense do the averages tend to the expected value

$$\frac{1}{n}\sum_{j=1}^{n} X_j \longrightarrow \mathsf{E}[X_i] \ ?$$

(CLT): "*Central limit theorems*": Under what assumptions and in which sense do we have the normal approximation

$$\frac{1}{\sqrt{n}}\sum_{j=1}^{n} \Big(X_j - \mathsf{E}[X_j]\Big) \longrightarrow \mathrm{N}(0,\sigma^2) \ ?$$

XI.1. Notions of convergence in stochastics

For the rest of this section, assume that

$$X_1, X_2, X_3, \ldots \qquad \text{and} \qquad X \qquad \text{are real-valued random variables.}$$

Let us then introduce two important probabilistic notions of convergence.

Convergence almost surely

Under any possible outcome $\omega \in \Omega$, the realized values $X_1(\omega), X_2(\omega), \ldots$ of the random variables form a sequence of real numbers. Therefore for a fixed ω, the

meaning of the limit

$$\lim_{n\to\infty} X_n(\omega) = X(\omega) \tag{XI.1}$$

is familiar from undergraduate calculus. Now recall that we can form the event

$$E = \Big\{\omega \in \Omega \;\Big|\; \lim_{n\to\infty} X_n(\omega) = X(\omega)\Big\} \tag{XI.2}$$

consisting of those outcomes ω for which we have the limit (XI.1).

Pointwise convergence $X_n \to X$ means that (XI.1) holds for all outcomes, i.e., that $E = \Omega$. As pointed out already in Example O.2, however, this is often too much to ask for. Instead, it is meaningful to ask about the probability $\mathsf{P}[E]$ of the event (XI.2). If this event occurs almost surely, then we talk about almost sure limit.

Definition XI.1 (Convergence almost surely)**.**
We say that X_n *tends to* X *almost surely* as $n \to \infty$, if

$$\mathsf{P}\Big[\lim_{n\to\infty} X_n = X\Big] = 1.$$

In this case we denote $X_n \xrightarrow{\text{a.s.}} X$.

This probabilistic notion of a limit should be intuitively easy to understand — we are giving up pointwise convergence only on an exceptional event $E^{\mathbf{c}}$ which has probability zero. Although we have thus relaxed the extremely stringent requirement of pointwise convergence, this is still a very strong notion of convergence.

Convergence in probability

Occasionally the notion of almost sure convergence is still too much to hope for. The following notion of convergence is less restrictive.

Definition XI.2 (Convergence in probability)**.**
We say that X_n *tends to* X *in probability* as $n \to \infty$, if for all $\varepsilon > 0$ we have

$$\lim_{n\to\infty} \mathsf{P}\Big[|X_n - X| < \varepsilon\Big] = 1. \tag{XI.3}$$

In this case we denote $X_n \xrightarrow{\mathsf{P}} X$.

Remark XI.3. The limit (XI.3) is equivalent to the following limit of probabilities of complementary events

$$\lim_{n\to\infty} \mathsf{P}\Big[|X_n - X| \geq \varepsilon\Big] = 0. \tag{XI.4}$$

Because there exist many techniques to upper bound probabilities, this latter formulation is used more often in practice.

Comparison of the notions of convergence

To understand these notions and to appreciate their differences, it is instructive to unravel the definition of the limit (XI.1) and the limit in the above definition, and compare the notions:

pointwise convergence: $X_n \to X$ as $n \to \infty$ if

$$\forall \omega \in \Omega \quad \forall \varepsilon > 0 \quad \exists n_0(\omega, \varepsilon) \in \mathbb{N} \quad \forall n > n_0(\omega, \varepsilon) \ : \\ |X_n(\omega) - X(\omega)| < \varepsilon.$$

convergence almost surely: $X_n \xrightarrow{\text{a.s.}} X$ as $n \to \infty$ if

$$\mathsf{P}\Big[\Big\{\omega \in \Omega \;\Big|\; \forall \varepsilon > 0 \;\; \exists n_0(\omega, \varepsilon) \in \mathbb{N} \;\; \forall n > n_0(\omega, \varepsilon) : \\ |X_n(\omega) - X(\omega)| < \varepsilon\Big\}\Big] = 1.$$

convergence in probability: $X_n \xrightarrow{\mathsf{P}} X$ as $n \to \infty$ if

$$\forall \varepsilon > 0 \quad \forall \delta > 0 \quad \exists n_0(\varepsilon, \delta) \in \mathbb{N} \quad \forall n > n_0(\varepsilon, \delta) \ : \\ \mathsf{P}\Big[\Big\{\omega \in \Omega \;\Big|\; |X_n(\omega) - X(\omega)| < \varepsilon\Big\}\Big] > 1 - \delta.$$

We leave it as an exercise to the reader to verify that convergence almost surely is indeed stronger than convergence in probability.

Exercise XI.1 (Convergence almost surely implies convergence in probability)**.**
Let $X_1, X_2, \ldots$ be real valued random variables such that $X_n \xrightarrow{\text{a.s.}} X$. Show that $X_n \xrightarrow{\mathsf{P}} X$.

The converse is not true in general: convergence in probability does not imply almost sure convergence. However, it does imply almost sure convergence along some subsequence, as the following exercise shows.

Exercise XI.2 (Convergence in probability implies convergence a.s. along a subsequence)**.**
Assume that $X_1, X_2, \ldots$ are real valued random variables and $X_n \xrightarrow{\mathsf{P}} X$. Let $(a_k)_{k\in\mathbb{N}}$ and $(b_k)_{k\in\mathbb{N}}$ be two sequences of positive real numbers such that $a_k \downarrow 0$ and $\sum_{k=1}^{\infty} b_k < +\infty$ — for example $a_k = \frac{1}{k}$ and $b_k = 2^{-k}$.

(a) Show that there exists $(n_k)_{k\in\mathbb{N}}$ so that $n_k \in \mathbb{N}$ for all k and $n_1 < n_2 < \cdots$ and

$$\mathsf{P}\Big[|X_{n_k} - X| \geq a_k\Big] \leq b_k.$$

(b) With the sequence $(n_k)_{k\in\mathbb{N}}$ chosen as in part (a), show that

$$\mathsf{P}\Big[|X_{n_k} - X| \geq a_k \text{ for infinitely many } k\Big] = 0.$$

(c) With the sequence $(n_k)_{k\in\mathbb{N}}$ chosen as in part (a), show that $X_{n_k} \xrightarrow{\text{a.s.}} X$ as $k \to \infty$.

Hint: *Recall a suitable Borel-Cantelli lemma.*

Most notions of convergence behave well under continuous functions. The following exercise concerns the case of convergence in probability.

Exercise XI.3 (Continuous transformations and convergence in probability)**.**
Let $X, X_1, X_2, \ldots$ be real-valued random variables defined on a probability space $(\Omega, \mathscr{F}, \mathsf{P})$. Suppose that $X_n \xrightarrow{\mathsf{P}} X$.

(a) For any uniformly continuous[1] $h\colon \mathbb{R} \to \mathbb{R}$, show that we have $h(X_n) \xrightarrow{\mathsf{P}} h(X)$.

[1]The conclusion is actually valid for any continuous $h\colon \mathbb{R} \to \mathbb{R}$. Can you prove that, too?

(b) For any bounded[2] uniformly continuous $h\colon \mathbb{R} \to \mathbb{R}$, show that we have

$$\mathsf{E}\big[|h(X_n) - h(X)|\big] \to 0 \qquad \text{and} \qquad \mathsf{E}\big[h(X_n)\big] \to \mathsf{E}\big[h(X)\big].$$

XI.2. Weak and strong laws of large numbers

Having introduced the notions of convergence of random variables, we now return to laws of large numbers. In the rest of this lecture we will prove two versions of laws of large numbers, which differ primarily by the notion of convergence used.

Theorem XI.4 (Weak law of large numbers with bounded second moments)**.** *Let $X_1, X_2, \dots$ be independent real valued random variables. Assume that for some $\mathfrak{m} \in \mathbb{R}$ and $K_2 < +\infty$ we have*

$$\mathsf{E}\big[X_j\big] = \mathfrak{m} \quad \text{and} \quad \mathsf{E}\big[X_j^2\big] \le K_2 \qquad \text{for all } j \in \mathbb{N}.$$

Then we have

$$\frac{1}{n}\big(X_1 + \cdots + X_n\big) \xrightarrow{\ \mathsf{P}\ } \mathfrak{m} \qquad \text{as } n \to \infty.$$

Theorem XI.5 (Strong law of large numbers with bounded fourth moments)**.** *Let $X_1, X_2, \dots$ be independent real valued random variables. Assume that for some $\mathfrak{m} \in \mathbb{R}$ and $K_4 < +\infty$ we have*

$$\mathsf{E}\big[X_j\big] = \mathfrak{m} \quad \text{and} \quad \mathsf{E}\big[X_j^4\big] \le K_4 \qquad \text{for all } j \in \mathbb{N}.$$

Then we have

$$\frac{1}{n}\big(X_1 + \cdots + X_n\big) \xrightarrow{\ \text{a.s.}\ } \mathfrak{m} \qquad \text{as } n \to \infty.$$

Theorem XI.4 is called a *weak law of large numbers*, because it asserts the convergence of the averages in the weaker sense of convergence in probability. Correspondingly, Theorem XI.5 is called a *strong law of large numbers*, because it asserts the convergence of the averages in the stronger sense of convergence almost surely. In both statements, we have chosen to make assumptions on boundedness of suitable moments, in order to simplify their proofs. Versions of weak and strong laws of large numbers could be made with various different assumptions. In particular, in Section XI.5 we present the statement of one of the most general formulations of such results: Kolmogorov's strong law of large numbers. Such improvements are occasionally important, but the main objectives for the present lecture are simply to appreciate the differences in the notions of convergence, and to get a flavor of techniques that are used in proving laws of large numbers.

The following exercise requires some calculations, but it illustrates that weak and strong laws of large numbers are valid under different assumptions.

Exercise XI.4 (A subtle sequence of averages)**.** Let $X_3, X_4, \dots$ be independent random variables such that for $k = 3, 4, \dots$ we have

$$\mathsf{P}\big[X_k = 0\big] = 1 - \frac{1}{k\,\log(k)} \qquad \text{and} \qquad \mathsf{P}\big[X_k = +k\big] = \frac{1}{2k\,\log(k)} = \mathsf{P}\big[X_k = -k\big].$$

[2] The conclusions are not valid without the assumption of boundedness. Can you give a counterexample in that case?

(a) Calculate the expected value and variance of X_k.
(b) Show that we have

$$\sum_{j=3}^{\infty} \frac{1}{j\,\log(j)} = \infty \qquad \text{and} \qquad \frac{1}{n^2}\sum_{j=3}^{n} \frac{j}{\log(j)} \longrightarrow 0 \quad \text{as } n \to \infty.$$

Hint: *Recall that the integral function of* $x \mapsto \frac{1}{x\,\log(x)}$ *is* $x \mapsto \log(\log(x))$. *Consider breaking the second sum into two pieces,* $j \le b_n$ *and* $j > b_n$ *with a judiciously chosen* b_n*, so that you can easily estimate the pieces.*

For $n \ge 3$, define the average

$$A_n = \frac{1}{n-2}\sum_{k=3}^{n} X_k.$$

(c) Does the sequence $(A_n)_{n=3,4,\ldots}$ of averages converge almost surely?
(d) Does the sequence $(A_n)_{n=3,4,\ldots}$ of averages converge in probability?
Hint: *For this part you can use Chebyshev's inequality presented below, in Corollary XI.7.*

XI.3. Proof of the weak law of large numbers

The weak law of large numbers, Theorem XI.4, follows easily from inequalities of Markov and Chebyshev, which we present below.

Markov's inequality and Chebyshev's inequality

While almost obvious, the following observation is extremely useful.

Lemma XI.6 (Markov's inequality)**.**
If $X\colon \Omega \to \mathbb{R}$ *is a random variable, then for any* $a > 0$ *we have*

$$\mathsf{P}\Big[|X| \ge a\Big] \le \frac{1}{a}\,\mathsf{E}\Big[|X|\Big]. \tag{XI.5}$$

Proof. Let $a > 0$ and define the event

$$E = \Big\{\omega \in \Omega \;\Big|\; |X(\omega)| \ge a\Big\} = X^{-1}\big((-a,a)^{\mathbf{c}}\big).$$

For all $\omega \in \Omega$ we then have

$$|X(\omega)| \;\ge\; a\,\mathbb{I}_E(\omega),$$

so by monotonicity of expected value we get

$$\mathsf{E}\Big[|X|\Big] \ge \mathsf{E}\Big[a\,\mathbb{I}_E\Big] = a\,\mathsf{P}\big[E\big].$$

Dividing this by a gives the asserted inequality $\mathsf{P}\big[E\big] \le \frac{1}{a}\,\mathsf{E}\big[|X|\big]$. □

Markov's inequality leads to the following upper bound for the probability of fluctuations of a random variable X from its mean by more than a multiple of its standard deviation.

Corollary XI.7 (Chebyshev's inequality)**.**
Suppose that $X \in \mathcal{L}^2(\mathsf{P})$. *Denote* $\mathfrak{m} := \mathsf{E}[X]$ *and* $\mathfrak{s}^2 := \mathsf{Var}(X) = \mathsf{E}\big[(X-\mathfrak{m})^2\big]$.

Then for any $c > 0$ we have

$$\mathsf{P}\Big[|X - \mathfrak{m}| \geq c\Big] \leq \frac{\mathfrak{s}^2}{c^2}. \tag{XI.6}$$

Proof. Let $Y = (X - \mathfrak{m})^2$, so $\mathsf{E}\big[|Y|\big] = \mathfrak{s}^2$. Note that the event of interest can be written as

$$E = \Big\{\omega \in \Omega \;\Big|\; |X(\omega) - \mathfrak{m}| \geq c\Big\} = \Big\{\omega \in \Omega \;\Big|\; |Y(\omega)| \geq c^2\Big\}.$$

The assertion follows by applying Markov's inequality (XI.5) to Y with $a = c^2$:

$$\mathsf{P}\Big[|Y| \geq c^2\Big] \leq \frac{1}{c^2}\,\mathsf{E}\Big[|Y|\Big] = \frac{\mathfrak{s}^2}{c^2}.$$

□

Proving the weak law of large numbers with Chebyshev's inequality

With Chebyshev's inequality, it is easy to prove a weak law of large numbers.

Proof of Theorem XI.4. Let $S_n = \sum_{j=1}^n X_j$ and $Y_n = \frac{1}{n} S_n$. We want to show $Y_n \xrightarrow{\mathsf{P}} \mathfrak{m}$. Note that by linearity of expected value we have

$$\mathsf{E}\big[S_n\big] = \sum_{j=1}^n \underbrace{\mathsf{E}\big[X_j\big]}_{=\mathfrak{m}} = n\mathfrak{m} \quad \text{and} \quad \mathsf{E}\big[Y_n\big] = \frac{1}{n}\,\mathsf{E}\big[S_n\big] = \frac{1}{n} n\mathfrak{m} = \mathfrak{m}.$$

Note also that by assumption $\mathsf{E}\big[X_j^2\big] \leq K_2$ we get a bound on variance,

$$\mathsf{Var}(X_j) = \mathsf{E}\Big[X_j^2\Big] - \mathsf{E}\Big[X_j\Big]^2 \leq \mathsf{E}\Big[X_j^2\Big] \leq K_2.$$

By independence we have (see Proposition X.12 and Exercise X.5)

$$\mathsf{Var}(S_n) = \sum_{j=1}^n \underbrace{\mathsf{Var}(X_j)}_{\leq K_2} \leq nK_2 \quad \text{and} \quad \mathsf{Var}(Y_n) = \mathsf{Var}\big(\frac{1}{n} S_n\big) = \frac{1}{n^2}\,\mathsf{Var}(S_n) \leq \frac{K_2}{n}.$$

Therefore, applying Chebyshev's inequality to Y_n, we get for any $\varepsilon > 0$

$$\mathsf{P}\Big[|Y_n - \mathfrak{m}| \geq \varepsilon\Big] \leq \frac{\mathsf{Var}(Y_n)}{\varepsilon^2} \leq \frac{K_2}{n\,\varepsilon^2} \underset{n\to\infty}{\longrightarrow} 0,$$

which shows the convergence in probability $Y_n \xrightarrow{\mathsf{P}} \mathfrak{m}$. □

The ideas in the weak law of large numbers can be used to prove *Weierstrass' approximation theorem*, which states that polynomials are dense in the space of continuous functions on a compact interval, with respect to the uniform norm $\|\cdot\|_\infty$ (see Example B.10). The following exercise has the proof broken down to six easy steps.

Exercise XI.5 (Weierstrass' approximation theorem)**.**
Let $U_1, U_2, \ldots$ be independent random variables with uniform distribution on $[0,1]$. For $p \in [0,1]$, consider the events $E_1^{(p)}, E_2^{(p)}, \ldots$ defined by $E_j^{(p)} = \big\{U_j \leq p\big\}$.

(a) Define $S_n^{(p)} = \sum_{j=1}^n \mathbb{I}_{E_j^{(p)}}$. Calculate the expected value $\mathsf{E}\big[S_n^{(p)}\big]$, and show that the variance is $\mathsf{Var}(S_n^{(p)}) \leq \frac{n}{4}$.

(b) Show that for any $\delta > 0$ we have $\mathsf{P}\Big[\big|S_n^{(p)}/n - p\big| \geq \delta\Big] \leq \frac{1}{4n\delta^2}$.

Let $f\colon [0,1] \to \mathbb{R}$ be some function.

(c) Show that $B_n(p) := \mathsf{E}\Big[f(S_n^{(p)}/n)\Big]$ is a polynomial in p.

(d) Show that $\big|B_n(p) - f(p)\big| \le \mathsf{E}\Big[\,\big|f(S_n^{(p)}/n) - f(p)\big|\,\Big]$.

Suppose that $f\colon [0,1] \to \mathbb{R}$ is continuous. By compactness of $[0,1]$, we know the following. The function f is bounded: there exists a $K < +\infty$ such that $|f(p)| \le K$ for all $p \in [0,1]$. Also, f is uniformly continuous: for all $\varepsilon > 0$ there exists a $\delta > 0$ such that $|f(p) - f(q)| < \varepsilon$ whenever $|p - q| < \delta$.

(e) Let $\varepsilon > 0$ and choose $\delta > 0$ by uniform continuity of f as above. Consider the event $A_n^{(p)} = \Big\{\big|S_n^{(p)}/n - p\big| \ge \delta\Big\}$. Show that we have

$$\big|f(S_n^{(p)}/n) - f(p)\big| \le 2K\, \mathbb{I}_{A_n^{(p)}} + \varepsilon\, \mathbb{I}_{\Omega\setminus A_n^{(p)}}.$$

(f) Prove the Weierstrass' approximation theorem: for any $\varepsilon > 0$ there exists a polynomial B_n such that $\big|f(p) - B_n(p)\big| < 2\varepsilon$ for all $p \in [0,1]$.

XI.4. Proof of the strong law of large numbers

Proof of Theorem XI.5. We first prove the statement assuming $\mathfrak{m} = 0$, and then reduce the general case to this particular case by appropriately centering the random variables.

the case $\mathfrak{m} = 0$*:* Assume $\mathfrak{m} = 0$. Denote

$$S_n = \sum_{j=1}^{n} X_j.$$

The assertion of the strong law of large numbers is that the event

$$E := \left\{\omega \in \Omega \;\middle|\; \lim_{n\to\infty} \frac{S_n(\omega)}{n} = 0\right\}$$

occurs almost surely. Instead of considering E directly, however, we will consider another event

$$E' := \left\{\omega \in \Omega \;\middle|\; \sum_{n=1}^{\infty} \Big(\frac{S_n(\omega)}{n}\Big)^4 < +\infty\right\},$$

and show that it is almost sure, and that it implies the original event E.

To see that the occurrence of E' implies the occurrence of E, note that by definition of E' we have

$$\sum_{n=1}^{\infty} \Big(\frac{S_n(\omega)}{n}\Big)^4 < +\infty \qquad \text{for all } \omega \in E'.$$

In particular, the terms of this convergent sum must tend to zero,

$$\Big(\frac{S_n(\omega)}{n}\Big)^4 \underset{n\to\infty}{\longrightarrow} 0 \qquad \text{for all } \omega \in E'.$$

Applying furthermore the continuous function $t \mapsto t^{1/4}$ to this limit, we conclude that

$$\frac{S_n(\omega)}{n} \underset{n\to\infty}{\longrightarrow} 0 \qquad \text{for all } \omega \in E'.$$

This proves that $E' \subset E$, establishing the desired implication.

To prove that E' occurs almost surely, the main body of work consists of showing that

$$\sum_{n=1}^{\infty} \mathsf{E}\left[\Big(\frac{S_n}{n}\Big)^4\right] < +\infty. \tag{XI.7}$$

Once this is done, Lemmas VIII.6 and VIII.5 imply that we have $\sum_{n=1}^{\infty} \big(\frac{S_n}{n}\big)^4 < +\infty$ almost surely, i.e., that $\mathsf{P}[E'] = 1$. This will finish the proof, because we then have

$$1 = \mathsf{P}[E'] \le \mathsf{P}[E] \qquad \text{since } E' \subset E.$$

It remains to prove (XI.7). Note first of all that since $X_j \in \mathcal{L}^4(\mathsf{P})$ for all j by assumption, Lemma VIII.10 shows that also $S_n = X_1 + \cdots + X_n \in \mathcal{L}^4(\mathsf{P})$. We will compute the fourth moment of S_n by expanding the multinomial

$$\begin{aligned} S_n^4 &= \big(X_1 + \cdots + X_n\big)^4 \\ &= \sum_{1\le i\le n} X_i^4 + \sum_{\substack{1\le i,j\le n\\ i\neq j}} \frac{4!}{3!\,1!}\, X_i^3\, X_j + \frac{1}{2!} \sum_{\substack{1\le i,j\le n\\ i\neq j}} \frac{4!}{2!^2}\, X_i^2\, X_j^2 \\ &\qquad + \frac{1}{2!} \sum_{\substack{1\le i,j,k\le n\\ i,j,k \text{ different}}} \frac{4!}{2!\,1!^2}\, X_i^2\, X_j\, X_k + \frac{1}{4!} \sum_{\substack{1\le i,j,k,\ell\le n\\ i,j,k,\ell \text{ different}}} \frac{4!}{1!^4}\, X_i\, X_j\, X_k\, X_\ell. \end{aligned} \tag{XI.8}$$

We want to take expected values of this expansion, so let us make some observations regarding that. By assumption we have $X_i \in \mathcal{L}^4(\mathsf{P})$ for any $i \in \mathbb{N}$, so from Lemma VIII.9 we get finiteness of lower order moments $\mathsf{E}\big[|X_i|\big] < \infty$, $\mathsf{E}\big[|X_i|^2\big] < \infty$, and $\mathsf{E}\big[|X_i|^3\big] < \infty$ as well. For any $i \neq j$ the random variables X_i and X_j were assumed independent, so we have also the independence of $X_i^3 \in \mathcal{L}^1(\mathsf{P})$ and $X_j \in \mathcal{L}^1(\mathsf{P})$ (recall Exercise V.2). By Theorem X.11, we then get

$$\mathsf{E}\Big[X_i^3\, X_j\Big] = \mathsf{E}\big[X_i^3\big]\, \underbrace{\mathsf{E}\big[X_j\big]}_{=\mathfrak{m}=0} = 0.$$

Similarly one argues that

$$\mathsf{E}\Big[X_i^2\, X_j\, X_k\Big] = \mathsf{E}\big[X_i^2\big]\, \underbrace{\mathsf{E}\big[X_j\big]}_{=0}\, \underbrace{\mathsf{E}\big[X_k\big]}_{=0} = 0$$

and

$$\mathsf{E}\Big[X_i\, X_j\, X_k\, X_\ell\Big] = \underbrace{\mathsf{E}\big[X_i\big]}_{=0}\, \underbrace{\mathsf{E}\big[X_j\big]}_{=0}\, \underbrace{\mathsf{E}\big[X_k\big]}_{=0}\, \underbrace{\mathsf{E}\big[X_\ell\big]}_{=0} = 0.$$

Only two terms of the expansion (XI.8) of the fourth power thus contribute to the expected value,

$$\mathsf{E}\Big[S_n^4\Big] = \sum_{1\le i\le n} \mathsf{E}\Big[X_i^4\Big] + \frac{1}{2!} \sum_{\substack{1\le i,j\le n\\ i\neq j}} \frac{4!}{2!^2}\, \mathsf{E}\Big[X_i^2\, X_j^2\Big].$$

For the first contribution we have the bound $\mathsf{E}\big[X_i^4\big] \le K_4$ by assumption, and for the second we use Cauchy-Schwarz inequality

$$\mathsf{E}\Big[X_i^2\, X_j^2\Big] \le \sqrt{\mathsf{E}\big[X_i^4\big]\, \mathsf{E}\big[X_j^4\big]} \le \sqrt{K_4\, K_4} = K_4.$$

This gives

$$\begin{aligned} \mathsf{E}\Big[S_n^4\Big] &\le \sum_{1\le i\le n} K_4 + \frac{1}{2!} \sum_{\substack{1\le i,j\le n\\ i\neq j}} \frac{4!}{2!^2}\, K_4 = n\, K_4 + n(n-1)\, \frac{1}{2!}\, \frac{4!}{2!^2}\, K_4 \\ &= n\, K_4 + (3n^2 - 3n)\, K_4 \\ &\le 3\, n^2\, K_4. \end{aligned}$$

With this estimate we get

$$\sum_{n=1}^{\infty} \mathsf{E}\Big[\Big(\frac{S_n}{n}\Big)^4\Big] = \sum_{n=1}^{\infty} \frac{1}{n^4}\, \mathsf{E}\big[S_n^4\big] \le \sum_{n=1}^{\infty} \frac{3\, n^2\, K_4}{n^4} = 3K_4 \sum_{n=1}^{\infty} \frac{1}{n^2} < +\infty.$$

This is exactly (XI.7), and the proof is complete for the case $\mathfrak{m} = 0$.

the general case: Finally, let us reduce the general case, $\mathsf{E}[X_j] = \mathfrak{m} \in \mathbb{R}$ for all $j \in \mathbb{N}$, to the case above. We define $\widetilde{X}_j = X_j - \mathfrak{m}$. This is a "centered" version of the original term X_j, i.e., it has vanishing expected value

$$\mathsf{E}\big[\widetilde{X}_j\big] = 0.$$

If we can apply the previously proven centered case of the strong law of large numbers, we get

$$\frac{1}{n}\sum_{j=1}^{n}\widetilde{X}_j \xrightarrow{\text{a.s.}} 0.$$

But since $\widetilde{X}_j = X_j - \mathfrak{m}$, this translates to

$$\frac{1}{n}\sum_{j=1}^{n}X_j - \mathfrak{m} \xrightarrow{\text{a.s.}} 0,$$

which in turn is equivalent to the desired almost sure convergence

$$\frac{1}{n}\sum_{j=1}^{n}X_j \xrightarrow{\text{a.s.}} \mathfrak{m}.$$

The proof will thus be complete if we are permitted to apply the previous case to the centered random variables $\widetilde{X}_j$. The independence of $\widetilde{X}_1, \widetilde{X}_2, \dots$ follows from the assumed independence of $X_1, X_2, \dots$ (as in Exercise V.2(a)). To derive the boundedness of the fourth moments of of $\widetilde{X}_1, \widetilde{X}_2, \dots$ from the assumed boundedness of the fourth moments of $X_1, X_2, \dots$, we use (VIII.3):

$$\mathsf{E}\Big[|\widetilde{X}_j|^4\Big] = \mathsf{E}\Big[|X_j - \mathfrak{m}|^4\Big] \le 2^4\Big(\mathsf{E}\big[|X|^4\big] + \mathsf{E}\big[\mathfrak{m}^4\big]\Big) \le 16\,\big(K_4 + \mathfrak{m}^4\big).$$

We see that there exists a constant $\widetilde{K}_4 < +\infty$ (given by the last expression above) such that $\mathsf{E}\big[\widetilde{X}_j^4\big] \le \widetilde{K}_4$ for all $j \in \mathbb{N}$. This finishes the proof. □

XI.5. Kolmogorov's strong law of large numbers

Our laws of large numbers, Theorems XI.4 and XI.5, were formulated with assumptions of bounded moments of suitable order. The existence of expected values only requires that the random variables are integrable, $X_1, X_2, \dots \in \mathcal{L}^1(\mathsf{P})$. We finish by giving the statement of a result of Kolmogorov, which does not use any higher moment assumptions. The result is a strong law of large numbers, because it gives almost sure convergence, but it also guarantees yet another notion of convergence which we introduce before the statement.

Convergence in $\mathcal{L}^1$

The following notion of convergence already appeared implicitly in the Dominated convergence theorem (Theorem VII.19).

Definition XI.8 (Convergence in $\mathcal{L}^1$)**.**
Suppose that $X_1, X_2, \dots \in \mathcal{L}^1(\mathsf{P})$ and $X \in \mathcal{L}^1(\mathsf{P})$. We say that X_n *tends to* X *in* $\mathcal{L}^1$ as $n \to \infty$, if we have

$$\lim_{n\to\infty}\mathsf{E}\Big[|X_n - X|\Big] = 0.$$

In this case we denote $X_n \xrightarrow{\mathcal{L}^1} X$.

The notion is stronger than convergence in probability.

Exercise XI.6 (Convergence in $\mathcal{L}^1$ implies convergence in probability)**.**

Assume that $X_1, X_2, \ldots \in \mathcal{L}^1(\mathsf{P})$ are integrable random variables and $X_n \xrightarrow{\mathcal{L}^1} X$. Show that $X_n \xrightarrow{\mathsf{P}} X$.

Hint: *Apply Markov's inequality.*

Statement of Kolmogorov's strong law of large numbers

Theorem (Kolmogorov's strong law of large numbers)**.**

Let $X_1, X_2, \ldots \in \mathcal{L}^1(\mathsf{P})$ be independent and identically distributed integrable random variables with $\mathsf{E}[X_j] = \mathfrak{m}$ for all $j \in \mathbb{N}$. Then we have

$$\frac{1}{n}\big(X_1 + \cdots + X_n\big) \xrightarrow{\text{a.s.}} \mathfrak{m}$$

and

$$\frac{1}{n}\big(X_1 + \cdots + X_n\big) \xrightarrow{\mathcal{L}^1} \mathfrak{m}$$

as $n \to \infty$.

The proof can be found in, e.g., [**Wil91**, Chapters 12 and 14].

For the expected value to make sense, we had to at least assume $X_j \in \mathcal{L}^1(\mathsf{P})$ in the above statement. In Theorems XI.4 and XI.5 we assumed more — that $X_j \in \mathcal{L}^2(\mathsf{P})$ and $X_j \in \mathcal{L}^4(\mathsf{P})$, respectively. Note, however, that while the above result is thus relaxing the moment assumptions, it is not strictly speaking a generalization of the strong law of Theorem XI.5, because it assumes that the sequence consists of identically distributed random variables. Both formulations are useful.

Lecture XII

Central limit theorem and convergence in distribution

Consider a sequence $X_1, X_2, \ldots$ of independent and identically distributed random numbers, and form the sums

$$S_n = X_1 + \cdots + X_n$$

of the first n members of the sequence. We are interested in the behavior of the sums S_n with a large number n of terms.

If the random variables are integrable, $X_j \in \mathcal{L}^1(\mathsf{P})$, with expected values $\mathsf{E}[X_j] = \mathfrak{m}$, then we have $\mathsf{E}[S_n] = n\mathfrak{m}$. The law of large numbers[1] then says that the sum S_n concentrates around the value $n\mathfrak{m}$, when we look at it in a scale proportional to n, or more precisely

$$\frac{S_n - n\mathfrak{m}}{n} \longrightarrow 0 \qquad \text{(almost surely, in probability, and in } \mathcal{L}^1(\mathsf{P})).$$

If the random variables are square integrable, $X_j \in \mathcal{L}^2(\mathsf{P})$, with expected values $\mathsf{E}[X_j] = \mathfrak{m}$ and variances $\mathsf{Var}(X_j) = \mathfrak{s}^2$, then by the independence of the terms we have $\mathsf{Var}(S_n) = n\mathfrak{s}^2$. Chebyshev's inequality then says that the fluctuations of the sum S_n around the value $n\mathfrak{m}$ do not exceed a scale proportional to $\sqrt{n}$, or more precisely

$$\mathsf{P}\left[\frac{|S_n - n\mathfrak{m}|}{\sqrt{n}} \geq c\right] \leq \frac{n\mathfrak{s}^2}{(c\sqrt{n})^2} = \frac{\mathfrak{s}^2}{c^2} \qquad \text{(for any } c > 0).$$

To understand the behavior of the sums S_n for large n in detail, it is therefore meaningful to look at $S_n - n\mathfrak{m}$ on a scale proportional to $\sqrt{n}$. One can show that

$$\frac{S_n - n\mathfrak{m}}{\sqrt{n}}$$

does not converge anywhere almost surely or in probability, so in order to describe the limiting behavior, we need another notion of convergence. The appropriate notion is *convergence in distribution* (also known as *convergence in law* or *weak convergence*). Addressing exactly this, the Central limit theorem asserts that as n increases, the distribution of $\frac{S_n - n\mathfrak{m}}{\sqrt{n}}$ approaches a Gaussian distribution (Example VIII.3). An elementary interpretation of this is that the cumulative distribution functions have the following limit

$$\mathsf{P}\left[\frac{S_n - n\mathfrak{m}}{\mathfrak{s}\sqrt{n}} \leq x\right] \underset{n\to\infty}{\longrightarrow} \Phi(x) := \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-t^2/2}\, \mathrm{d}t \qquad \text{(for all } x \in \mathbb{R}).$$

The right hand side above is the cumulative distribution of the standard normal distribution (Gaussian distribution with mean zero and unit variance).

[1] In this formulation, we need Kolmogorov's strong law of large numbers.

XII.1. Characteristic functions

An important tool in the proof of the central limit theorem and in various other places in probability theory is characteristic functions. The characteristic function of a real valued random variable[2] is a single function, which encodes its distribution — it is essentially the Fourier transform of the law of the random variable.

Describing distributions in terms of their characteristic functions thus unifies the treatment of discrete distributions (which can be described by probability mass functions[3]) and continuous distributions (which can be described by probability densities[4]), as well as distributions, which are neither discrete nor continuous!

Complex valued random variables

The characteristic functions are complex valued, so we begin with a few remarks about complex valued random variables.

The set of complex numbers is denoted by $\mathbb{C}$, and the imaginary unit by $\mathbb{i}$ ($\mathbb{i} \in \mathbb{C}$ is a square root of -1, i.e., $\mathbb{i}^2 = -1$). A complex number $z \in \mathbb{C}$ can be written uniquely as $z = x + \mathbb{i}y$ with $x, y \in \mathbb{R}$. We call $x = \mathfrak{Re}(z)$ the real part and $y = \mathfrak{Im}(z)$ the imaginary part of z. The absolute value (or modulus) of z is $|z| := \sqrt{x^2 + y^2}$. We identify $\mathbb{C}$ with the Euclidean plane $\mathbb{R}^2$ (the set of complex numbers forms the "complex plane"), so that x and y are the two coordinates of $z = x + \mathbb{i}y$ in the plane. The absolute value $|z|$ is exactly the Euclidean norm, and we equip $\mathbb{C}$ with the topology of the plane, and correspondingly with the Borel σ-algebra $\mathscr{B}(\mathbb{C}) \cong \mathscr{B}(\mathbb{R}^2) = \mathscr{B} \otimes \mathscr{B}$.

A *complex valued random variable* $Z = X + \mathbb{i}\,Y$ is constructed out of a pair X, Y of real valued random variables[5] (the real and imaginary parts of Z). We say that Z is integrable if both its real and imaginary parts are, i.e., if $X, Y \in \mathcal{L}^1(\mathsf{P})$. We define the expected value of an integrable complex random variable Z to be the complex number

$$\mathsf{E}\big[Z\big] := \mathsf{E}\big[X\big] + \mathbb{i}\,\mathsf{E}\big[Y\big]$$

whose real and imaginary parts are the expected values of the real and imaginary parts of Z. Note that for $z = x + \mathbb{i}y$ we have $|x + \mathbb{i}\,y| \leq |x| + |y|$. From monotonicity, linearity, and triangle inequality we therefore get

$$\mathsf{E}\big[|Z|\big] \leq \mathsf{E}\Big[|X| + |Y|\Big] = \mathsf{E}\big[|X|\big] + \mathsf{E}\big[|Y|\big] < \infty,$$

if Z is integrable.

The exponential of a complex number $z \in \mathbb{C}$ is defined by the convergent power series

$$e^z = \sum_{n=0}^{\infty} \frac{1}{n!} z^n.$$

---

[2]More generally, characteristic functions of vector valued random variables could be defined, and have properties parallel with what we show in the setup of real valued random variables.

[3]See Definition II.15 and Exercise VIII.1, in particular.

[4]See Definition VIII.2 and Exercise VIII.3, in particular.

[5]As with random vectors, $Z \colon \Omega \to \mathbb{C}$ is $\mathscr{F}/\mathscr{B}(\mathbb{C})$-measurable if and only if $X, Y \colon \Omega \to \mathbb{R}$ are $\mathscr{F}/\mathscr{B}(\mathbb{R})$-measurable — see Remark X.2.

As a particular case, for $z = \mathbb{i}\phi$ with $\phi \in \mathbb{R}$ we have Euler's formula

$$e^{\mathbb{i}\,\phi} = \cos(\phi) + \mathbb{i}\,\sin(\phi).$$

Note that we have $\left|e^{\mathbb{i}\,\phi}\right| = \sqrt{\cos^2(\phi) + \sin^2(\phi)} = 1$. It is often convenient to write a complex number $z \in \mathbb{C}$ in polar coordinates as $z = r\,e^{\mathbb{i}\phi}$, where $r = |z|$ and $\phi \in [0, 2\pi)$.

Expected values of complex valued random variables have familiar properties.

Proposition XII.1 (Properties of expected values of complex random variables)**.**

Linearity: *If $c_1, c_2 \in \mathbb{C}$ are complex numbers and Z_1, Z_2 are integrable $\mathbb{C}$-valued random variables, then also $c_1Z_1 + c_2Z_2$ is an integrable $\mathbb{C}$-valued random variable and $\mathsf{E}\big[c_1Z_1 + c_2Z_2\big] = c_1\,\mathsf{E}[Z_1] + c_2\,\mathsf{E}[Z_2]$.*

Triangle inequality: *If Z is an integrable $\mathbb{C}$-valued random variable, then we have $\big|\mathsf{E}[Z]\big| \leq \mathsf{E}\big[|Z|\big]$.*

Dominated convergence: *Suppose that $Z_1, Z_2, \ldots$ are $\mathbb{C}$-valued random variables and $X \in \mathcal{L}^1(\mathsf{P})$ is an integrable random variable which dominates the absolute values, $|Z_n| \leq X$ for all $n \in \mathbb{N}$. Then if the pointwise limit $\lim_{n\to\infty} Z_n$ exists, we have $\mathsf{E}\big[\lim_{n\to\infty} Z_n\big] = \lim_{n\to\infty} \mathsf{E}[Z_n]$.*

Proof. Linearity is proved directly from the definition by splitting each of c_1, c_2, Z_1, Z_2 to real and imaginary parts. We leave the details as an exercise.

Triangle inequality can be proved as follows. The expected value is a complex number, so we can write it in polar coordinates as $\mathsf{E}[Z] = r\,e^{\mathbb{i}\phi}$, where $r = \big|\mathsf{E}[Z]\big|$ and $\phi \in [0, 2\pi)$. Then by linearity we have

$$r = e^{-\mathbb{i}\phi}\,\mathsf{E}[Z] = \mathsf{E}\big[e^{-\mathbb{i}\phi}Z\big] = \mathsf{E}\Big[\,\Re\mathfrak{e}\,\big(e^{-\mathbb{i}\phi}Z\big)\Big] + \mathbb{i}\,\mathsf{E}\Big[\,\Im\mathfrak{m}\,\big(e^{-\mathbb{i}\phi}Z\big)\Big].$$

Since $r \in \mathbb{R}$ by construction, the second term in fact has to vanish: $\mathsf{E}\big[\,\Im\mathfrak{m}\,\big(e^{-\mathbb{i}\phi}Z\big)\big] = 0$. If we furthermore use the fact that $\Re\mathfrak{e}(z) \leq |z|$ and monotonicity of real expected values, we thus get

$$r = \mathsf{E}\Big[\,\Re\mathfrak{e}\,\big(e^{-\mathbb{i}\phi}Z\big)\Big] \;\leq\; \mathsf{E}\Big[\big|e^{-\mathbb{i}\phi}Z\big|\Big] \;=\; \mathsf{E}\Big[|Z|\Big].$$

Remembering that $r = \big|\mathsf{E}[Z]\big|$, this gives the triangle inequality.

Dominated convergence follows by splitting Z_n to real and imaginary parts and applying dominated convergence theorem to these separately: the same integrable random variable X which dominates $|Z_n|$ also dominates $\Re\mathfrak{e}(Z_n)$ and $\Im\mathfrak{m}(Z_n)$. □

Exercise XII.1 (Independent complex random variables)**.**
Let $Z_1 = X_1 + \mathbb{i}\,Y_1$ and $Z_2 = X_2 + \mathbb{i}\,Y_2$ be two independent, integrable complex valued random variables. Show that the product Z_1Z_2 is integrable, and that we have

$$\mathsf{E}\big[Z_1\,Z_2\big] = \mathsf{E}\big[Z_1\big]\;\mathsf{E}\big[Z_2\big].$$

Definition and first properties of characteristic functions

Let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space and $X\colon \Omega \to \mathbb{R}$ a real valued random variable. Note that for any $\theta \in \mathbb{R}$ and any $\omega \in \Omega$, we have

$$e^{\mathbb{i}\,\theta X(\omega)} = \cos\big(\theta X(\omega)\big) + \mathbb{i}\,\sin\big(\theta X(\omega)\big). \tag{XII.1}$$

This shows that the real and imaginary parts of $e^{\mathbb{i}\,\theta X}$ are bounded random variables, and thus in particular integrable. Therefore the following definition makes sense.

Definition XII.2 (Characteristic function)**.**
The *characteristic function* of X is the function $\varphi_X\colon \mathbb{R} \to \mathbb{C}$ given by

$$\begin{aligned}\varphi_X(\theta) &= \mathsf{E}\Big[e^{\mathbb{i}\,\theta X}\Big] \\ &= \mathsf{E}\big[\cos(\theta X)\big] + \mathbb{i}\,\mathsf{E}\big[\sin(\theta X)\big].\end{aligned} \tag{XII.2}$$

Remark XII.3. The function $x \mapsto e^{\mathbb{i}\theta x}$ is continuous, and therefore a Borel function by Corollary III.10 (i.e. the real and imaginary parts $x \mapsto \cos(\theta x)$ and $x \mapsto \sin(\theta x)$ are). Therefore by Theorem VIII.1, the expected value in (XII.2) can be written using the distribution P_X of X,

$$\varphi_X(\theta) = \mathsf{E}\big[e^{\mathbb{i}\,\theta X}\big] = \int_{\mathbb{R}} e^{\mathbb{i}\theta x}\,\mathrm{d}P_X(x).$$

This shows that the characteristic function φ_X of X only depends on the distribution P_X of X. Soon we will show that φ_X in fact contains enough information to fully determine the distribution P_X.

Let us give a few examples of characteristic functions.

Example XII.4 (Characteristic function of exponential distribution)**.**
Suppose that $X \sim \mathrm{Exp}(\lambda)$ is exponentially distributed with parameter $\lambda > 0$ (see Example VIII.4), i.e., X has a probability density

$$f_X(x) \;=\; \lambda\, e^{-\lambda x}\, \mathbb{I}_{[0,+\infty)}(x).$$

Let us compute its characteristic function using the formula of Exercise VIII.3,

$$\begin{aligned}\varphi_X(\theta) = \mathsf{E}\big[e^{\mathbb{i}\,\theta X}\big] &= \int_{\mathbb{R}} e^{\mathbb{i}\,\theta x}\, f_X(x)\,\mathrm{d}x \\ &= \int_0^\infty e^{\mathbb{i}\,\theta x}\,\lambda\, e^{-\lambda x}\,\mathrm{d}x = \lambda \int_0^\infty e^{x(-\lambda+\mathbb{i}\,\theta)}\,\mathrm{d}x = \lambda\frac{-1}{-\lambda+\mathbb{i}\theta} = \frac{1}{1+\mathbb{i}\theta/\lambda}.\end{aligned}$$

Example XII.5 (Characteristic function of Poisson distribution)**.**
Suppose that $X \sim \mathrm{Poisson}(\lambda)$ is Poisson distributed with parameter $\lambda > 0$ (see Example II.16), i.e., X has a probability mass function

$$p_X(n) \;=\; \mathsf{P}\big[X = n\big] = e^{-\lambda x}\,\frac{\lambda^n}{n!} \qquad \text{for } n \in \mathbb{Z}_{\geq 0} = \{0, 1, 2, \ldots\}.$$

Let us compute its characteristic function using the formula of Exercise VIII.1,

$$\begin{aligned}\varphi_X(\theta) = \mathsf{E}\big[e^{\mathbb{i}\,\theta X}\big] &= \sum_{n=0}^{\infty} p_X(n)\, e^{\mathbb{i}\theta n} \\ &= \sum_{n=0}^{\infty} e^{-\lambda}\,\frac{\lambda^n}{n!}\, e^{\mathbb{i}\theta n} = e^{-\lambda}\sum_{n=0}^{\infty}\frac{1}{n!}\big(\lambda e^{\mathbb{i}\theta}\big)^n = e^{-\lambda}\, e^{\lambda e^{\mathbb{i}\theta}} = \exp\big(\lambda(e^{\mathbb{i}\theta}-1)\big).\end{aligned}$$

Exercise XII.2 (The characteristic function of a standard Gaussian random variable)**.**
Suppose that $X \sim \mathrm{N}(0,1)$ is a real valued random variable with standard normal distribution (see Example VIII.3), i.e., a continuous distribution with density function

$$f_X(x) = \frac{1}{\sqrt{2\pi}}\, e^{-\frac{1}{2}x^2} \qquad \text{for } x \in \mathbb{R}.$$

Hint: *You can consider it known that* $\int_{-\infty}^{+\infty} f_X(x)\,\mathrm{d}x = 1$*, and that the exponential of any complex number* $z \in \mathbb{C}$ *is given by the convergent series* $e^z = \sum_{n=0}^{\infty}\frac{1}{n!}z^n$*.*

(a) Let $t \in \mathbb{R}$. Show that $\mathsf{E}\big[e^{tX}\big] = e^{t^2/2}$.
Hint: *Express the expected value in terms of the density, and perform a suitable change of variables* $x' = x + c$*.*

(b) For $x, t \in \mathbb{R}$, show that $e^{|tx|} \leq e^{tx} + e^{-tx}$. Using this, prove that for any $t \in \mathbb{R}$ we have

$$\mathsf{E}\Big[\sum_{n=0}^{\infty} \frac{1}{n!} |tX|^n\Big] < +\infty.$$

(c) Prove that for any $t \in \mathbb{R}$ we have

$$\mathsf{E}\big[e^{tX}\big] = \sum_{n=0}^{\infty} \frac{1}{n!} t^n \, \mathsf{E}[X^n].$$

(d) By comparing (a) with (c), deduce that for $n \in \mathbb{N}$ we have

$$\mathsf{E}[X^n] = \begin{cases} \prod_{j=1}^{n/2} (2j-1) & \text{if } n \text{ is even} \\ 0 & \text{if } n \text{ is odd.} \end{cases}$$

(e) Prove that

$$\varphi_X(\theta) = e^{-\frac{1}{2}\theta^2} \qquad \text{for } \theta \in \mathbb{R}.$$

We now state properties of characteristic functions that hold in general. You may directly inspect that the characteristic functions in Examples XII.4, XII.5 and Exercise XII.2 indeed have the stated properties.

Proposition XII.6 (Basic properties of characteristic functions)**.**
Characteristic functions have the following properties:

(a) *We have* $\varphi_X(0) = 1$.
(b) *We have* $|\varphi_X(\theta)| \leq 1$ *for all* $\theta \in \mathbb{R}$.
(c) *The function* $\varphi_X \colon \mathbb{R} \to \mathbb{C}$ *is continuous.*
(d) *For any* $a, b \in \mathbb{R}$ *we have* $\varphi_{aX+b}(\theta) = e^{\mathrm{i}\, b\theta} \, \varphi_X(a\theta)$ *for all* $\theta \in \mathbb{R}$.
(e) *We have* $\varphi_{-X}(\theta) = \overline{\varphi_X(\theta)}$ *for all* $\theta \in \mathbb{R}$.
(f) *We have* $\varphi_X(-\theta) = \overline{\varphi_X(\theta)}$ *for all* $\theta \in \mathbb{R}$.

Proof. At $\theta = 0$ we of course have $\theta X(\omega) = 0$ for all $\omega \in \Omega$ and thus $e^{\mathrm{i}\,\theta X(\omega)} = 1$. We directly get $\varphi_X(0) = \mathsf{E}[1] = 1$, which proves (a).

Part (b) follows from triangle inequality: $|\varphi_X(\theta)| = \big|\mathsf{E}[e^{\mathrm{i}\theta X}]\big| \leq \mathsf{E}\big[|e^{\mathrm{i}\theta X}|\big] = \mathsf{E}[1] = 1$.

Continuity is proved as follows. We must show that for any sequence $\theta_1, \theta_2, \ldots \in \mathbb{R}$ such that $\theta_n \to \theta$, we have $\varphi_X(\theta_n) \to \varphi_X(\theta)$. Since $\theta_n \to \theta$, we get pointwise for all $\omega \in \Omega$, that $e^{\mathrm{i}\,\theta_n X(\omega)} \to e^{\mathrm{i}\,\theta X(\omega)}$, using the continuity of the exponential function. But the random variables $e^{\mathrm{i}\,\theta_n X}$ are also bounded, so we can use the Bounded convergence theorem (both real and imaginary parts are bounded real random variables which converge pointwise):

$$\varphi_X(\theta_n) = \mathsf{E}\Big[e^{\mathrm{i}\,\theta_n X}\Big] \longrightarrow \mathsf{E}\Big[e^{\mathrm{i}\,\theta X}\Big] = \varphi_X(\theta).$$

This proves part (c), the continuity of φ_X.

For part (d), observe that

$$e^{\mathrm{i}\,\theta\left(aX(\omega)+b\right)} = e^{\mathrm{i}\,\theta a X(\omega)} \, e^{\mathrm{i}\,\theta b}$$

and use linearity.

Parts (e) and (f) follow by noting that the complex conjugate of $e^{\mathrm{i}\theta X(\omega)}$ is $e^{-\mathrm{i}\theta X(\omega)}$. □

We can now for instance reduce the calculation of the characteristic function of a general Gaussian random variable to that of a standard Gaussian.

Exercise XII.3 (Characteristic function of a Gaussian random variable)**.**
Let $\mathfrak{m} \in \mathbb{R}$ and $\mathfrak{s} > 0$, and let $X \sim \mathrm{N}(\mathfrak{m}, \mathfrak{s}^2)$ (see Example VIII.3). Use Exercise XII.2 and Proposition XII.6(d) to show that

$$\varphi_X(\theta) = e^{\mathrm{i}\mathfrak{m}\theta - \frac{1}{2}\mathfrak{s}^2\theta^2} \qquad \text{for } \theta \in \mathbb{R}.$$

Another fundamental property of characteristic functions is that the characteristic function of a sum of independent terms is the pointwise product of the characteristic functions.

Exercise XII.4 (Characteristic function of a sum of independent terms)**.**
Suppose that X and Y are independent real valued random variables. Using Exercise XII.1, show that the characteristic function of their sum is

$$\varphi_{X+Y}(\theta) = \varphi_X(\theta)\,\varphi_Y(\theta) \qquad \text{for } \theta \in \mathbb{R}.$$

Lévy's inversion theorem

A fundamental property of the characteristic function of a random variable is that it contains all the information about the distribution of the random variable. This fact is made explicit by Lévy's inversion theorem, below.

Theorem XII.7 (Lévy's inversion theorem)**.**
Let $X \in \mathrm{m}\mathscr{F}$ be a real-valued random variable, P_X its distribution (a Borel probability measure on $\mathbb{R}$), and $\varphi_X \colon \mathbb{R} \to \mathbb{C}$ its characteristic function. Then for any $a, b \in \mathbb{R}$, $a < b$, we have

$$\begin{aligned} &\lim_{T \to +\infty} \frac{1}{2\pi} \int_{-T}^{+T} \frac{e^{-\mathrm{i}\theta a} - e^{-\mathrm{i}\theta b}}{\mathrm{i}\,\theta}\, \varphi_X(\theta)\, \mathrm{d}\theta \\ &= P_X\big[(a,b)\big] + \frac{1}{2} P_X\big[\,\{a\}\,\big] + \frac{1}{2} P_X\big[\,\{b\}\,\big]. \end{aligned}$$

In particular, φ_X uniquely determines P_X.

Moreover, if $\int_{\mathbb{R}} |\varphi_X(\theta)|\, \mathrm{d}\theta < +\infty$, then X has a continuous probability density function f_X given by

$$f_X(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-\mathrm{i}\theta x}\, \varphi_X(\theta)\, \mathrm{d}\theta.$$

The proof is given in Appendix F.

Exercise XII.5 (Sum of independent Gaussian random variables)**.**
Suppose that $X_1 \sim \mathrm{N}(\mathfrak{m}_1, \mathfrak{s}_1^2)$ and $X_2 \sim \mathrm{N}(\mathfrak{m}_2, \mathfrak{s}_2^2)$ are Gaussian random variables which are independent. Show that $X_1 + X_2 \sim \mathrm{N}(\mathfrak{m}_1 + \mathfrak{m}_2, \mathfrak{s}_1^2 + \mathfrak{s}_2^2)$.
Hint: *Use Lévy's inversion theorem together with Exercises XII.3 and XII.4.*

Exercise XII.6 (Sum of i.i.d. Bernoulli random variables)**.**

(a) Let $p \in [0,1]$. Calculate the characteristic function $\varphi_B(\theta) = \mathsf{E}[e^{\mathrm{i}\,\theta B}]$ of a random variable B such that $\mathsf{P}[B = 1] = p$ and $\mathsf{P}[B = 0] = 1 - p$ (we denote $B \sim \text{Bernoulli}(p)$).

(b) Let $p \in [0,1]$ and $n \in \mathbb{N}$. Calculate the characteristic function $\varphi_X(\theta) = \mathsf{E}[e^{\mathrm{i}\,\theta X}]$ of a random variable X such that $\mathsf{P}[X = k] = \binom{n}{k} p^k (1-p)^{n-k}$ for all $k \in \{0, 1, 2, \ldots, n\}$ (we denote $X \sim \mathrm{Bin}(n, p)$).

(c) Let $B_1, \dots, B_n$ be independent and identically distributed, with $\mathsf{P}[B_j = 1] = p$ and $\mathsf{P}[B_j = 0] = 1 - p$, for all j. Compute the characteristic function of $S = B_1 + \cdots + B_n$ using part (a) and Exercise XII.4. Compare with the result of part (b), and conclude that $S \sim \mathrm{Bin}(n, p)$.

Taylor expansion of a characteristic function

By Lévy's inversion theorem, the characteristic function φ_X of a random variable X contains all information about the distribution P_X of X. In particular, it should contain the information about the expected value, variance, etc. To see why this is at least formally true, write the power series expansion

$$e^{\mathbb{i}\,\theta X(\omega)} = \sum_{n=0}^{\infty} \frac{1}{n!}\big(\mathbb{i}\theta X(\omega)\big)^n = 1 + \mathbb{i}\theta X(\omega) - \frac{1}{2}\theta^2 X(\omega)^2 + \cdots \qquad \text{for all } \omega \in \Omega.$$

If the expected value could be taken term by term in this expansion, then we would get

$$\varphi_X(\theta) = \mathsf{E}\Big[e^{\mathbb{i}\,\theta X}\Big] \stackrel{?}{=} 1 + \mathbb{i}\theta\,\mathsf{E}[X] - \frac{1}{2}\theta^2\,\mathsf{E}[X^2] + \cdots$$

Formally, therefore, the expected value $\mathsf{E}[X]$ seems to be encoded in the first order term in the Taylor expansion of $\varphi_X(\theta)$ around the point $\theta = 0$, the variance $\mathsf{Var}(X) = \mathsf{E}[X^2] - \mathsf{E}[X]^2$ in the terms up to order two, and more generally moment $\mathsf{E}[X^n]$ of order n in the coefficient of θ^n. Of course, this can only be meaningful if the random variable has moments of the correct order, i.e., $X \in \mathcal{L}^p(\mathsf{P})$ for high enough $p \geq 1$.

The following lemma makes precise sense of the above formal observation for square integrable random variables.[6]

Proposition XII.8 (Taylor expansion of characteristic function)**.**
Let $X \in \mathcal{L}^2(\mathsf{P})$ be a square integrable random variable and let $\varphi_X \colon \mathbb{R} \to \mathbb{C}$ be its characteristic function. Then we have

$$\varphi_X(\theta) = 1 + \mathbb{i}\theta\,\mathsf{E}[X] - \frac{1}{2}\theta^2\,\mathsf{E}[X^2] + \epsilon(\theta), \tag{XII.3}$$

where the function $\epsilon \colon \mathbb{R} \to \mathbb{C}$ is an error term of smaller order than θ^2 in the sense that

$$\frac{|\epsilon(\theta)|}{|\theta|^2} \longrightarrow 0 \qquad \text{as } \theta \to 0.$$

Proof. The idea is to Taylor expand $e^{\mathbb{i}\theta X}$ up to order two, with a controlled error term.

Start by observing that we have $\frac{\mathrm{d}}{\mathrm{d}u} e^{\mathbb{i}\theta u} = \mathbb{i}\theta\, e^{\mathbb{i}\theta u}$, so for any $x \in \mathbb{R}$ we have

$$\int_0^x \mathbb{i}\theta\, e^{\mathbb{i}\theta u}\,\mathrm{d}u = e^{\mathbb{i}\theta x} - 1.$$

Let us solve for $e^{\mathbb{i}\theta x}$, and get

$$e^{\mathbb{i}\theta x} \;=\; 1 + \mathbb{i}\theta \int_0^x e^{\mathbb{i}\theta u}\,\mathrm{d}u.$$

[6] The reader should think about how to modify the assumptions, statement, and proof to see moments up to order n in the Taylor expansion of the characteristic function.

Apply the same observation again to the integrand $e^{\mathbb{i}\theta u}$, to get

$$\begin{aligned} e^{\mathbb{i}\theta x} &= 1 + \mathbb{i}\theta \int_0^x \Big(1 + \mathbb{i}\theta \int_0^u e^{\mathbb{i}\theta v}\, \mathrm{d}v\Big)\, \mathrm{d}u \\ &= 1 + \mathbb{i}\theta x - \theta^2 \int_0^x \Big(\int_0^u e^{\mathbb{i}\theta v}\, \mathrm{d}v\Big)\, \mathrm{d}u. \end{aligned}$$

In this expression, write still $e^{\mathbb{i}\theta v} = 1 + (e^{\mathbb{i}\theta v} - 1)$, and perform the integrations of the first term to get

$$e^{\mathbb{i}\theta x} = 1 + \mathbb{i}\theta x - \theta^2 \frac{x^2}{2} - \theta^2 \int_0^x \Big(\int_0^u (e^{\mathbb{i}\theta v} - 1)\, \mathrm{d}v\Big)\, \mathrm{d}u. \tag{XII.4}$$

The first three terms without integrations are the ones we care about, so let us introduce the following notation for the remainder that we want to get rid of,

$$R(\theta, x) := \int_0^x \Big(\int_0^v (e^{\mathbb{i}\theta v} - 1)\, \mathrm{d}v\Big)\, \mathrm{d}u$$

To estimate the magnitude of this remainder, note first that $|R(\theta, -x)| = |R(\theta, x)|$, so it is enough to consider $x \geq 0$. Then use the triangle inequality for integrals and the observation $e^{\mathbb{i}\theta v} - 1 = e^{\mathbb{i}\theta v/2}\big(e^{\mathbb{i}\theta v/2} - e^{-\mathbb{i}\theta v/2}\big) = 2\mathbb{i}\, e^{\mathbb{i}\theta v/2} \sin(\theta v/2)$ to get the upper bound

$$\big|R(\theta, x)\big| \leq \int_0^{|x|} \Big(\int_0^v \big|e^{\mathbb{i}\theta v} - 1\big|\, \mathrm{d}v\Big)\, \mathrm{d}u \leq 2 \int_0^{|x|} \Big(\int_0^v \big|\sin(\theta v/2)\big|\, \mathrm{d}v\Big)\, \mathrm{d}u$$

If we estimate the integrand in the last expression by $\big|\sin(\theta v/2)\big| \leq 1$, we find after the two integrations

$$\big|R(\theta, x)\big| \leq |x|^2, \tag{XII.5}$$

and if we estimate it by $\big|\sin(\theta v/2)\big| \leq \frac{1}{2}\,|\theta|\,|v|$, we find after integrations

$$\big|R(\theta, x)\big| \leq \frac{1}{6}\,|\theta|\,|x|^3,$$

which in particular shows that for any $x \in \mathbb{R}$ we have

$$\big|R(\theta, x)\big| \longrightarrow 0 \qquad \text{as } \theta \to 0. \tag{XII.6}$$

Let us now apply (XII.4) pointwise to the values of the random variable X, to get

$$e^{\mathbb{i}\theta X(\omega)} = 1 + \mathbb{i}\theta\, X(\omega) - \frac{1}{2}\theta^2\, X(\omega)^2 - \theta^2\, R\big(\theta, X(\omega)\big).$$

With this, we can write the error term $\epsilon(\theta)$ in the approximation (XII.3) in a manageable form. Namely, by linearity of expectation we have

$$\begin{aligned} \epsilon(\theta) &:= \varphi_X(\theta) - \Big(1 + \mathbb{i}\theta\, \mathsf{E}[X] - \frac{1}{2}\theta^2\, \mathsf{E}[X^2]\Big) \\ &= \mathsf{E}\Big[e^{\mathbb{i}\theta X} - 1 - \mathbb{i}\theta\, X + \frac{1}{2}\theta^2\, X^2\Big] \\ &= -\theta^2\, \mathsf{E}\big[R(\theta, X)\big]. \end{aligned}$$

Then use the triangle inequality for expected values to control the magnitude of the error term,

$$\big|\epsilon(\theta)\big| \leq |\theta|^2\, \mathsf{E}\Big[\big|R(\theta, X)\big|\Big].$$

The estimate (XII.5) shows that $\big|R(\theta, X)\big| \leq |X|^2$ for any θ, so by the assumption $X \in \mathcal{L}^2(\mathsf{P})$ we have an integrable upper bound and we can use the Dominated convergence theorem in

$$\lim_{\theta \to 0} \mathsf{E}\Big[\big|R(\theta, X)\big|\Big] = \mathsf{E}\Big[\lim_{\theta \to 0} \big|R(\theta, X)\big|\Big] \overset{\text{(XII.6)}}{=} \mathsf{E}[0] = 0.$$

We conclude that

$$\frac{|\epsilon(\theta)|}{|\theta|^2} \leq \mathsf{E}\Big[\big|R(\theta, X)\big|\Big] \longrightarrow 0$$

as $\theta \to 0$, and the proof is complete. □

XII.2. Convergence in distribution

The notion of convergence in distribution is different from notion of convergence in Lecture XI: it is not strictly speaking about convergence of random variables (functions on the sample space Ω), but rather the convergence of their distributions (measures on $\mathbb{R}$, in the case of real-valued random variables).[7]

Theorem XII.9 (Equivalent conditions for convergence in distribution)**.**
Let $X_1, X_2, \dots$ and X be real-valued random variables. Let also $F_1, F_2, \dots$ and F be their cumulative distribution functions, and let $\varphi_1, \varphi_2, \dots$ and φ be their characteristic functions, respectively. Then the following conditions are equivalent:

(i) *For all bounded continuous functions $f\colon \mathbb{R} \to \mathbb{R}$ we have $\mathsf{E}[f(X_n)] \to \mathsf{E}[f(X)]$ as $n \to \infty$.*
(ii) *We have $F_n(x) \to F(x)$ as $n \to \infty$ for all points $x \in \mathbb{R}$ such that F is continuous at x.*
(iii) *We have $\varphi_n(\theta) \to \varphi(\theta)$ as $n \to \infty$ for all $\theta \in \mathbb{R}$.*

The proof is given in Appendix F.

Definition XII.10 (Convergence in distribution)**.**
Let $X_1, X_2, \dots$ and X be real-valued random variables. We say that X_n *tend to X in distribution* (or *in law*) as $n \to \infty$ and denote $X_n \xrightarrow{\text{law}} X$, if any (then all) of the equivalent conditions of Theorem XII.9 hold.

Remark XII.11 (Convergence in distribution formulated in terms of distributions)**.**
Note that a bounded continuous function $f\colon \mathbb{R} \to \mathbb{R}$ is a Borel function (Corollary III.10), so the expected values $\mathsf{E}[f(X_n)]$ and $\mathsf{E}[f(X)]$ can be written using the distributions P_{X_n} and P_X (Theorem VIII.1). Therefore condition (i) of Theorem XII.9 can be written as

$$\int_{\mathbb{R}} f \, \mathrm{d}P_{X_n} \longrightarrow \int_{\mathbb{R}} f \, \mathrm{d}P_X \qquad \text{as } n \to \infty$$

for all such f. This formulation explains the terminology *convergence in distribution*.

XII.3. Central limit theorem

We are now ready to state and prove the Central limit theorem (CLT). It is one of the most central[8] theorems in probability and statistics. It is for instance the rigorous justification for various normal approximations that are commonly used.

[7] In fact, for convergence in distribution it is not even necessary that the random variables are defined on the same probability space: we could have $X\colon \Omega \to \mathbb{R}$ defined on $(\Omega, \mathscr{F}, \mathsf{P})$ but $X_n\colon \Omega_n \to \mathbb{R}$ each defined on its own probability space $(\Omega_n, \mathscr{F}_n, \mathsf{P}_n)$. In any case, the distributions P_{X_n} are probability measures on the real line $\mathbb{R}$, so the formulation of Remark XII.11 below still makes perfect sense. For simplicity of presentation, however, we choose not to keep explicitly mentioning and writing all the (possibly) different probability spaces during this lecture.

[8] pun intended / hence the name

Theorem XII.12 (Central limit theorem)**.**
Let $X_1, X_2, \ldots \in \mathcal{L}^2(\mathsf{P})$ be independent and identically distributed square integrable random variables. Denote

$$\mathfrak{m} := \mathsf{E}[X_j] \qquad \text{and} \qquad \mathfrak{s} := \sqrt{\mathsf{Var}(X_j)}.$$

Assume that $\mathfrak{s} > 0$.[9] *For all $n \in \mathbb{N}$, let $S_n = \sum_{j=1}^n X_j$. Then we have*

$$\frac{S_n - n\mathfrak{m}}{\mathfrak{s}\sqrt{n}} \xrightarrow{\text{law}} Z \qquad \text{as } n \to \infty, \tag{XII.7}$$

where Z is a random variable with standard normal distribution $\mathrm{N}(0,1)$.

Proof. By considering $\widetilde{X}_j = X_j - \mathfrak{m}$ if necessary, we may assume that $\mathfrak{m} = 0$. Likewise, by considering $\frac{X_j}{\mathfrak{s}}$ if necessary, we may assume that $\mathfrak{s} = 1$. The goal is then to show that $\frac{S_n}{\sqrt{n}} \xrightarrow{\text{law}} Z$. We will prove this by verifying condition (iii) of Theorem XII.9, i.e., the pointwise convergence of the characteristic functions of $\frac{S_n}{\sqrt{n}}$.

By the assumption of identical distributions, the characteristic functions of all X_j, $j \in \mathbb{N}$, are equal, so let us denote them by

$$\varphi(\theta) := \varphi_{X_j}(\theta) = \mathsf{E}\big[e^{\mathrm{i}\theta X_j}\big].$$

By Proposition XII.8 and assumptions $\mathfrak{m} = 0$ and $\mathfrak{s} = 1$, we have

$$\varphi(\theta) = 1 - \frac{1}{2}\theta^2 + \epsilon(\theta), \tag{XII.8}$$

where $\frac{\epsilon(\theta)}{|\theta|^2} \to 0$ as $\theta \to 0$.

Now calculate the characteristic function of the sum $S_n = \sum_{j=1}^n X_j$ using independence,

$$\varphi_{S_n}(\theta) = \mathsf{E}\Big[e^{\mathrm{i}\theta \sum_{j=1}^n X_j}\Big] \;=\; \mathsf{E}\Big[\prod_{j=1}^n e^{\mathrm{i}\theta X_j}\Big] \overset{(\perp\!\!\!\perp)}{=} \prod_{j=1}^n \mathsf{E}\big[e^{\mathrm{i}\theta X_j}\big] \;=\; \varphi(\theta)^n.$$

The characteristic function of $\frac{S_n}{\sqrt{n}}$ is then

$$\varphi_{S_n/\sqrt{n}}(\theta) = \mathsf{E}\Big[e^{\mathrm{i}\theta S_n/\sqrt{n}}\Big] \;=\; \varphi_{S_n}\Big(\frac{\theta}{\sqrt{n}}\Big) \;=\; \Big(\varphi\big(\frac{\theta}{\sqrt{n}}\big)\Big)^n.$$

By (XII.8), we have

$$\varphi\big(\frac{\theta}{\sqrt{n}}\big) \;=\; 1 - \frac{1}{2}\Big(\frac{\theta}{\sqrt{n}}\Big)^2 + \epsilon\big(\frac{\theta}{\sqrt{n}}\big) \;=\; 1 - \frac{\theta^2}{2n} + r_n,$$

where $\frac{r_n}{1/n} \to 0$ as $n \to \infty$. By substituting this in the expression for the characteristic function of $\frac{S_n}{\sqrt{n}}$, we get

$$\varphi_{S_n/\sqrt{n}}(\theta) \;=\; \Big(1 - \frac{\theta^2}{2n} + r_n\Big)^n.$$

The limit in Exercise XII.7 gives

$$\lim_{n\to\infty} \varphi_{S_n/\sqrt{n}}(\theta) \;=\; \lim_{n\to\infty} \Big(1 - \frac{\theta^2}{2n} + r_n\Big)^n \;=\; e^{-\frac{1}{2}\theta^2}.$$

Since this is the characteristic function $\varphi_Z(\theta)$ of a standard normal distributed random variable $Z \sim \mathrm{N}(0,1)$ according to Exercise XII.2, the proof is complete. □

Exercise XII.7 (Complex exponential as a limit)**.**
Let $z \in \mathbb{C}$ and suppose that $r_1, r_2, \ldots \in \mathbb{C}$ are such that $n\, r_n \to 0$ as $n \to \infty$. Show that

$$\lim_{n\to\infty} \Big(1 + \frac{z}{n} + r_n\Big)^n = e^z.$$

[9] This is always true unless X_j are almost surely constants. The case of almost surely constant random variables is not probabilistically interesting.

Appendix A

Set theory preliminaries

This appendix reviews necessary backgound on set theory, in particular the notion of countability, which is crucial in probability theory and measure theory.

A.1. Intersections and unions of sets

Let A, B be two sets. The *intersection* $A \cap B$ is defined as the set of those elements which belong to both A and B,

$$A \cap B = \left\{ x \;\middle|\; x \in A \text{ and } x \in B \right\}.$$

The *union* $A \cup B$ is defined as the set of those elements which belong to at least one of the sets A and B,

$$A \cup B = \left\{ x \;\middle|\; x \in A \text{ or } x \in B \right\}.$$

More generally, let $(A_i)_{i \in I}$ be a collection of sets A_i indexed by $i \in I$. The union of the collection is defined as

$$\bigcup_{i \in I} A_i = \left\{ x \;\middle|\; x \in A_i \text{ for some } i \in I \right\}.$$

The intersection of the collection is defined provided the collection is non-empty ($I \neq \emptyset$) as

$$\bigcap_{i \in I} A_i = \left\{ x \;\middle|\; x \in A_i \text{ for all } i \in I \right\}.$$

A collection $(A_i)_{i \in I}$ of sets is said to be *disjoint* if no two different members of the collection have common elements, i.e., for all $i, j \in I$, $i \neq j$, we have $A_i \cap A_j = \emptyset$. If the collection $(A_i)_{i \in I}$ of sets is disjoint, then we say that the union $\bigcup_{i \in I} A_i$ is a *disjoint union*. Disjoint unions enjoy additivity properties in probability theory and measure theory, according to the axiomatic properties in Definitions II.4 and II.5.

A.2. Set differences and complements

Let A, B be two sets. The *set difference* $A \setminus B$ is defined as the set of those elements which belong to A but do not belong to B,

$$A \setminus B = \left\{ x \;\middle|\; x \in A \text{ and } x \notin B \right\}.$$

When it is clear from the context that we are considering subsets of a particular reference set S (often the sample space $S = \Omega$), then the *complement* of a subset $A \subset S$, denoted by $A^{\mathbf{c}}$, is the set of those elements (of S) which do not belong to A

$$A^{\mathbf{c}} = S \setminus A = \left\{ x \in S \;\middle|\; x \notin A \right\}.$$

The following basic results of set theory tell how unions and intersections behave under complements. They are known as *De Morgan's laws*.

Proposition A.1 (De Morgan's laws)**.**
Let $(A_i)_{i\in I}$ be a non-empty indexed collection of subsets $A_i \subset S$ of a fixed set S. Then we have

$$\big(\bigcup_{i\in I} A_i\big)^{\mathbf{c}} = \bigcap_{i\in I} A_i^{\mathbf{c}} \qquad \text{and} \qquad \big(\bigcap_{i\in I} A_i\big)^{\mathbf{c}} = \bigcup_{i\in I} A_i^{\mathbf{c}}.$$

Exercise A.1. Prove the De Morgan's laws above.

A.3. Images and preimages of sets under functions

Let S and S' be two sets and

$$f\colon S \to S'$$

a function from S to S'. For $A \subset S$, the *image* of A under f is the subset $f(A) \subset S'$ consisting of all those elements $s' \in S'$ such that $s' = f(s)$ for some $s \in S$,

$$f(A) = \big\{ f(s) \;\big|\; s \in A \big\} \subset S'.$$

For $A' \subset S'$, the *preimage* of A' under f is the subset $f^{-1}(A') \subset S$ consisting of all those elements $s \in S$ whose image $f(s)$ belongs to the subset A',

$$f^{-1}(A') = \big\{ s \in S \;\big|\; f(s) \in A' \big\} \subset S.$$

Exercise A.2 (Properties of preimages)**.**
Show that

(a) $$f^{-1}\big(A'^{\mathbf{c}}\big) = \big(f^{-1}(A')\big)^{\mathbf{c}}$$

(b) $$f^{-1}\Big(\bigcup_i A'_i\Big) = \bigcup_i f^{-1}(A'_i)$$

(c) $$f^{-1}\Big(\bigcap_i A'_i\Big) = \bigcap_i f^{-1}(A'_i).$$

A.4. Cartesian products

Let A, B be two sets. The *Cartesian product* $A \times B$ is defined as the set of ordered pairs (a, b) whose first member a belongs to the set A and second member b belongs to the set B, i.e.,

$$A \times B = \big\{ (a, b) \;\big|\; a \in A,\ b \in B \big\}.$$

More generally, when $A_1, \ldots, A_n$ are sets, the n-fold Cartesian product $A_1 \times \cdots \times A_n$ is the set of ordered n-tuples $(a_1, \ldots, a_n)$ such that $a_k \in A_k$ for each $k = 1, \ldots, n$

$$A_1 \times \cdots \times A_n = \big\{ (a_1, \ldots, a_n) \;\big|\; a_1 \in A_1, \ldots, a_n \in A_n \big\}.$$

Even more generally, if $(A_j)_{j\in J}$ is a collection of sets indexed by $j \in J$, the Cartesian product $\prod_{j\in J} A_j$ is the set of indexed collections $(a_j)_{j\in J}$ such that $a_j \in A_j$ for

each $j \in J$,

$$\prod_{j\in J} A_j = \left\{(a_j)_{j\in J} \;\middle|\; \forall j \in J : a_j \in A_j\right\}.$$

The *axiom of choice* states that if all the factors are non-empty, $A_j \neq \emptyset$ for all $j \in J$, then the Cartesian product is also non-empty $\prod_{j\in J} A_j \neq \emptyset$.

As a particular case of Cartesian products, if each factor is the same set, $A_j = A$ for all $j \in J$, then the Cartesian product is alternatively denoted by $A^J := \prod_{j\in J} A$. In that case an element of A^J is an indexed collection $(a_j)_{j\in J}$ of elements of A, which can be naturally identified with the function $j \mapsto a_j$ from J to A. Therefore A^J is identified with the set of functions from J to A,

$$A^J = \{f : J \to A \text{ function}\}.$$

A.5. Power set

Given a set S, the set $\mathscr{P}(S)$ of all subsets $A \subset S$ of it is called the *power set* of S,

$$\mathscr{P}(S) = \left\{A \;\middle|\; A \subset S\right\}.$$

A subset $A \subset S$ can be specified by indicating for each element $s \in S$ whether it belongs to A or not, so it is natural to identify the the power set $\mathscr{P}(S)$ of S with the set $\{0,1\}^S$ of functions $S \to \{0,1\}$. In particular if S is a finite set with $\#S = n$ elements, then its power set is a finite set with $\#\mathscr{P}(S) = 2^n$ elements.

It is good to keep in mind that the power set readily provides an easy first example of many notions introduced in probability theory. For instance, in view of Definitions I.1, II.23, and C.3, the collection $\mathscr{P}(S)$ of all subsets of S is obviously a sigma algebra on S, a π-system on S, a d-system on S, etc.

A.6. Sequences of sets

A sequence $A_1, A_2, \ldots$ of sets is said to be *increasing* if

$$A_1 \subset A_2 \subset A_3 \subset \cdots.$$

In this case we denote

$$A_n \uparrow A,$$

where the *limit A of the increasing sequence of sets* is defined as the union

$$A = \bigcup_{n=1}^{\infty} A_n.$$

Likewise, a sequence $A_1, A_2, \ldots$ of sets is said to be *decreasing* if

$$A_1 \supset A_2 \supset A_3 \supset \cdots.$$

In this case we denote

$$A_n \downarrow A,$$

where the *limit A of the decreasing sequence of sets* is defined as the intersection

$$A = \bigcap_{n=1}^{\infty} A_n.$$

Sequences of sets which are either increasing or decreasing are said to be *monotone*.

Exercise A.3 (Characterization of limits of monotone sequences of sets)**.**
Show that the limits of monotone sequences of sets can be characterized as follows.

(a) Suppose that $A_n \uparrow A$. Show that

$$\begin{aligned} x \in A &\iff \exists m \in \mathbb{N} \text{ such that } \forall n \geq m : x \in A_n \\ x \notin A &\iff \forall n \in \mathbb{N} : x \notin A_n. \end{aligned}$$

(b) Suppose that $A_n \downarrow A$. Show that

$$\begin{aligned} x \in A &\iff \forall n \in \mathbb{N} : x \in A_n \\ x \notin A &\iff \exists m \in \mathbb{N} \text{ such that } \forall n \geq m : x \notin A_n. \end{aligned}$$

For a sequence $A_1, A_2, \ldots$ of sets, we define its *upper limit* as

$$\limsup_n A_n := \bigcap_{m \in \mathbb{N}} \bigcup_{n \geq m} A_n.$$

Note that if we define $C_m = \bigcup_{n \geq m} A_n$, then the sequence $C_1, C_2, \ldots$ of sets is decreasing, and its limit $\bigcap_{m \in \mathbb{N}} C_m$ is precisely $\limsup_n A_n$.

We also define the *lower limit* of the sequence of sets as

$$\liminf_n A_n := \bigcup_{m \in \mathbb{N}} \bigcap_{n \geq m} A_n.$$

Note that if we define $D_m = \bigcap_{n \geq m} A_n$, then the sequence $D_1, D_2, \ldots$ of sets is increasing, and $\liminf_n A_n$ is its limit.

Exercise A.4 (Characterization of upper and lower limits of sequences of sets)**.**
Show that the limsup and liminf of a sequence $A_1, A_2, \ldots$ of sets can be characterized as follows:

(a): $$\limsup_n A_n = \left\{ s \;\middle|\; \forall\, m \in \mathbb{N} : \exists\, n \geq m : s \in A_n \right\}$$

(b): $$\liminf_n A_n = \left\{ s \;\middle|\; \exists\, m \in \mathbb{N} : \forall\, n \geq m : s \in A_n \right\}.$$

A.7. Countable and uncountable sets

In probability theory we need to distinguish between sets of different sizes: finite sets, countably infinite sets, and uncountably infinite sets. In fact, if one had to summarize probability theory (and measure theory) in a single phrase, it might be:

> All countable operations in probability theory are defined to behave just as intuition dictates.

Comparison of cardinalities

Cardinality is the set-theoretic notion of the size or "number of elements" of a set. The idea is that when comparing the sizes of two sets A and B, we attempt to match the elements of A to the elements of B by functions $f\colon A \to B$. Under a surjective function, each element of B has at least some element(s) of A matched to it, and we then would interpret that A has at least as many elements as B. Under an injective function f, all elements of A are matched to some elements of B without any two different elements ever being matched to the same element, and then we would interpret that A has at most as many elements as B. Comparison of cardinalities is done by asking about the existence of such functions.

Definition A.2 (Comparison of cardinalities)**.**
Let A and B be sets. We say that the *cardinality* of A is less than or equal to the cardinality of B if there exists an injective function $f\colon A \to B$.

Remark A.3 (Comparison of cardinalities of finite sets)**.**
Suppose that A and B are two finite sets. Let $n = \#A$ be the number of elements in A and $m = \#B$ be the number of elements in B. Then it is easy to see that there exists an injective function $f\colon A \to B$ if and only if we have $n \leq m$. Thus for finite sets, the comparison of cardinalities amounts to just the comparison of the number of elements.

Example A.4 (Subsets can not have larger cardinality)**.**
If A is a subset of B, $A \subset B$, then the cardinality of A is less than or equal to the cardinality of B, because the inclusion mapping $\iota\colon A \to B$ defined by $\iota(x) = x$ for all $x \in A$ is injective.

Example A.5 (Transitivity of comparison of cardinalities)**.**
Let A, B, C be sets. Suppose that the cardinality of A is less than or equal to the cardinality of B and the cardinality of B is less than or equal to the cardinality of C. In that case there exists injective functions $f\colon A \to B$ and $\tilde{f}\colon B \to C$. The composition $\tilde{f} \circ f\colon A \to C$ is also injective, so we get that the cardinality of A is less than or equal to the cardinality of C. In other words, the comparison of cardinalities is transitive.

As suggested before Definition A.2, instead of requiring the existence of injective functions in the comparison of cardinalities, one can alternatively require the existence of surjective functions in the opposite direction. The following two exercises establish this alternative characterization. To solve these exercises, you are allowed to use the axiom of choice.

Exercise A.5 (Comparison of cardinalities with surjective functions: necessity)**.**
Show that if the cardinality of a non-empty set $A \neq \emptyset$ is less than or equal to the cardinality of a set B, then there exists a surjective function $g\colon B \to A$.

Exercise A.6 (Comparison of cardinalities with surjective functions: sufficiency)**.**
Show that if there exists a surjective function $g\colon B \to A$, then the cardinality of the set A is less than or equal to the cardinality of the set B.

Equal cardinalities

Definition A.6 (Equal cardinalities)**.**
We say that A and B have *equal cardinalities* if there exists an injective function $f\colon A \to B$ and an injective function $g\colon B \to A$.

Clearly if there exists a bijective function $f\colon A \to B$, then A and B have equal cardinalities (we may then take $g = f^{-1}$). The converse is also true, but it is not as obvious. The Schröder - Bernstein theorem states that if A and B have equal cardinalities, then there exists a bijective function $f\colon A \to B$ (you could try to prove this as an exercise).

Countable sets

For the purposes of probability theory and measure theory, countable cardinalities are the most crucial. We begin with the definition.

Definition A.7 (Countability)**.**
A set A is said to be *countable* if the cardinality of A is less than or equal to the cardinality of the set $\mathbb{N} = \{1, 2, 3, \ldots\}$ of natural numbers.

Example A.8 (Subsets of natural numbers are countable)**.**
From Example A.4 it follows that any subset $S \subset \mathbb{N}$, including the set $\mathbb{N}$ of natural numbers itself, is countable.

Since countable sets are so important, we unravel the definition once more, and provide an alternative characterization and two useful sufficient conditions.

Lemma A.9 (Criteria for countability)**.**

(a) *A set A is countable if and only if there exists an injective function $f\colon A \to \mathbb{N}$.*
(b) *A non-empty set $A \neq \emptyset$ is countable if and only if there exists a surjective function $g\colon \mathbb{N} \to A$.*
(c) *If B is a countable set and there exists an injective function $f\colon A \to B$, then also the set A is countable.*
(d) *If B is a countable set and there exists a surjective function $g\colon B \to A$, then also the set A is countable.*

Proof. Assertion (a) follows directly by combining Definitions A.2 and A.7.

Assertion (b) follows by combining Definition A.7 with the characterization of Exercises A.5 and A.6.

Assertions (c) and (d) are similarly obtained using the transitivity in Example A.5. □

Enumerations of countable sets

If A is countable and non-empty, then from Exercise A.5 it follows that there exists a surjective function $g\colon \mathbb{N} \to A$. We see that all elements of A are obtained in the following "list"

$$A = \{g(1), g(2), g(3), \ldots\}.$$

Note, however, that repetitions are allowed in the above "list", as g does not have to be injective. It is possible to remove repetitions and obtain an enumeration of the elements of A. To do this, one defines $a_k \in A$ as the k:th value in the list above omitting repetitions. If the set A is finite, however, then there are only finitely many different values and the enumeration terminates at some point. Thus, for a finite set A with n elements, we have an enumeration

$$A = \{a_1, a_2, \ldots, a_n\}.$$

An infinite set which is countable is said to be *countably infinite*, and for such a set A, we have an enumeration

$$A = \{a_1, a_2, a_3, \ldots\}.$$

Note also that if the elements of A can be enumerated as above without repetition, then the mapping $a_k \mapsto k$ is well defined and injective $A \to \mathbb{N}$. Therefore any set whose elements can be enumerated is countable.

Operations that preserve countability

In probability theory and measure theory, countable operations work well. It is therefore crucial to understand clearly which set theoretic operations preserve countability.

Suppose that A_1 and A_2 are countable sets. By definition, there exists injective functions $f_1 \colon A_1 \to \mathbb{N}$ and $f_2 \colon A_2 \to \mathbb{N}$. Consider the union $A_1 \cup A_2$, and note that it can be expressed as $A_1 \cup A_2 = A_1 \cup (A_2 \setminus A_1)$, where the latter is a disjoint union. The function $f \colon A_1 \cup A_2 \to \mathbb{N}$ defined "piecewise" by

$$f(x) = \begin{cases} 2f_1(x) & \text{if } x \in A_1 \\ 2f_2(x) + 1 & \text{if } x \in A_2 \setminus A_1 \end{cases}$$

is clearly injective: it maps elements of A_1 injectively to even natural numbers and the remaining elements injectively to odd natural numbers. From the existence of such an injective function we conclude that the union $A_1 \cup A_2$ is countable. Using this argument inductively, we get that finite unions of countable sets remain countable.

Lemma A.10 (Finite unions of countable sets are countable)**.**
Let $A_1, \ldots, A_n$ be countable sets. Then the union

$$A_1 \cup \cdots \cup A_n = \bigcup_{j=1}^{n} A_j$$

is also countable.

Example A.11 (The set of integers is countable)**.**
Consider the three sets:

$$A_1 = \{1, 2, 3, \ldots\}$$
$$A_2 = \{0\}$$
$$A_3 = \{-1, -2, -3, \ldots\}.$$

Each of them is countable: the set $A_1 = \mathbb{N}$ is countable by Example A.8, the set A_2 is countable because it is finite, and the set A_3 is countable because it is in bijection with $\mathbb{N}$ via $x \mapsto -x$. The set of all integers is the union of these three

$$\mathbb{Z} = \{\ldots, -2, -1, 0, 1, 2, \ldots\} = A_1 \cup A_2 \cup A_3$$

and as such $\mathbb{Z}$ is itself countable by Lemma A.10

Consider now the set $\mathbb{N} \times \mathbb{N}$, the Cartesian product of the set of natural numbers with itself. We claim that $\mathbb{N} \times \mathbb{N}$ is countable. To see this, note that the elements can be enumerated

$$\begin{aligned}\mathbb{N} \times \mathbb{N} = \big\{(n,m) \;\big|\; n,m \in \mathbb{N}\big\} = \Big\{&(1,1),\\ &(2,1),(1,2),\\ &(3,1),(2,2),(1,3),\\ &(4,1),(3,2),(2,3),(1,4),\\ &\qquad\qquad \dots \qquad\qquad \Big\}.\end{aligned}$$

The enumeration shows that $\mathbb{N} \times \mathbb{N}$ is also countable, as it gives rise to an injective function $h\colon \mathbb{N} \times \mathbb{N} \to \mathbb{N}$.

Suppose now that A_1 and A_2 are countable sets. Then there exists injective functions $f_1\colon A_1 \to \mathbb{N}$ and $f_2\colon A_2 \to \mathbb{N}$. Now define $f\colon A_1 \times A_2 \to \mathbb{N}$ by

$$f(x_1,x_2) = h\big(f_1(x_1), f_2(x_2)\big) \qquad \text{for } x_1 \in A_1,\ x_2 \in A_2,$$

where $h\colon \mathbb{N} \times \mathbb{N} \to \mathbb{N}$ is the injective function given by the above enumeration of $\mathbb{N} \times \mathbb{N}$. The function $f\colon A_1 \times A_2 \to \mathbb{N}$ is injective, and we thus see that the Cartesian product $A_1 \times A_2$ of countable sets A_1 and A_2 is again countable. Using this observation inductively, we get that Cartesian products of finitely many countable sets remain countable.

Proposition A.12 (Finite Cartesian produts of countable sets are countable)**.**
Let $A_1, \dots, A_n$ be countable sets. Then the Cartesian product

$$A_1 \times \cdots \times A_n$$

is also countable.

Example A.13 (The d-dimensional integer lattice $\mathbb{Z}^d$ is countable)**.**
Let $d \in \mathbb{N}$. The set

$$\mathbb{Z}^d = \underbrace{\mathbb{Z} \times \cdots \times \mathbb{Z}}_{d \text{ times}} = \Big\{(x_1,\dots,x_d) \in \mathbb{R}^d \;\Big|\; x_1,\dots,x_d \in \mathbb{Z}\Big\}$$

of points with integer coordinates in the d-dimensional Euclidean space $\mathbb{R}^d$ is countable, since it is the Cartesian product of d copies of the countable set $\mathbb{Z}$.

Example A.14 (The set of rational numbers is countable)**.**
Consider the set $\mathbb{Q} \subset \mathbb{R}$ of rational numbers. The mapping $g\colon \mathbb{Z} \times \mathbb{N} \to \mathbb{Q}$ defined by $g(n,m) = \frac{n}{m}$ is surjective onto $\mathbb{Q}$. As a Cartesian product of the countable sets $\mathbb{Z}$ and $\mathbb{N}$, the set $\mathbb{Z} \times \mathbb{N}$ is countable by Proposition A.12. From Lemma A.9(d) and the existence of the surjective function g we get that also the set $\mathbb{Q}$ of rational numbers is countable.

Using the above proposition about Cartesian products of countable sets we can strengthen our earlier observation about unions of countable sets: it turns out that countable unions of countable sets remain countable.

Proposition A.15 (Countable unions of countable sets are countable)**.**
Let $A_1, A_2, A_3, \ldots$ be countable sets. Then the union

$$\bigcup_{j=1}^{\infty} A_j = A_1 \cup A_2 \cup A_3 \cup \cdots$$

is also countable.

Proof. We may assume that the sets A_j are non-empty, $A_j \neq \emptyset$, for all $j \in \mathbb{N}$ (empty terms can be omitted from the union). Then, for each $j \in \mathbb{N}$ there exists a surjective function

$$g_j \colon \mathbb{N} \to A_j.$$

We now define a function g on $\mathbb{N} \times \mathbb{N}$ by

$$g(j,k) = g_j(k) \qquad \text{for } j,k \in \mathbb{N},$$

and observe that this function

$$g \colon \mathbb{N} \times \mathbb{N} \to \bigcup_{j=1}^{\infty} A_j$$

is surjective onto the union $\bigcup_{j=1}^{\infty} A_j$. This shows that the cardinality of $\bigcup_{j=1}^{\infty} A_j$ is less than or equal to the cardinality of $\mathbb{N} \times \mathbb{N}$. Since $\mathbb{N} \times \mathbb{N}$ is countable, this shows that the union $\bigcup_{j=1}^{\infty} A_j$ is also countable. □

Uncountable sets

A set which is not countable is said to be *uncountable*. Since all finite sets are countable, an uncountable set is necessarily infinite.

In the previous section we saw that some rather nontrivial set theoretic operations preserve countability. We now provide examples of uncountable sets by a useful standard argument known as Cantor's diagonal extraction. The argument shows that countable Cartesian products of countable sets (or even of finite sets) are generally not countable.

Example A.16 (The set of binary sequences is uncountable)**.**
For each $j \in \mathbb{N}$, let $A_j = \{0,1\}$. Consider the Cartesian product

$$B = \prod_{j=1}^{\infty} A_j = \{0,1\}^{\mathbb{N}} = \Big\{ (b_1, b_2, \ldots) \;\Big|\; b_1, b_2, \ldots \in \{0,1\} \Big\}$$

of the sets $A_1, A_2, \ldots$, which is most concretely interpreted as the set of infinite binary sequences $b = (b_1, b_2, \ldots)$ of zeroes and ones. The set B is a countable Cartesian product of finite sets. We claim that B itself is uncountable.

The diagonal argument proceeds by supposing, on the contrary, that B is countable. If this were the case, then we could find an enumeration

$$B = \left\{ b^{(1)}, b^{(2)}, b^{(3)}, \ldots \right\},$$

where the m:th element $b^{(m)}$ is a binary sequence

$$b^{(m)} = (b_1^{(m)}, b_2^{(m)}, b_3^{(m)}, \ldots).$$

Now define a binary sequence $b' = (b'_1, b'_2, b'_3, \ldots)$ by choosing for each $j \in \mathbb{N}$ the j:th "digit" $b'_j \in \{0,1\}$ to be different from $b_j^{(j)}$, the j:th "digit" of the j:th element $b^{(j)}$ in the enumeration. This construction of $b' \in B$ is known as diagonal extraction. Now for any $m \in \mathbb{N}$, the binary sequence b' differs from $b^{(m)}$ at least in the m:th digit, so $b' \neq b^{(m)}$. But the element $b' \in B$ should appear in the enumeration of B, so we have derived a contradiction. We conclude that B can not be enumerated. Therefore B is in fact uncountable.

Example A.17 (The set of real numbers is uncountable)**.**

Let $B = \{0,1\}^{\mathbb{N}}$ be the set of binary sequences as in the previous example. Consider also the subset $B' \subset B$ of those binary sequences which contain infinitely many zeroes,

$$B' = \left\{ (b_1, b_2, \ldots) \in \{0,1\}^{\mathbb{N}} \;\middle|\; \forall m \in \mathbb{N} : \exists n \geq m : b_n = 0 \right\}.$$

We first claim that B' is also uncountable.

The complement $B \setminus B'$ is the set of binary sequences which end with repeated ones,

$$B \setminus B' = \left\{ (b_1, b_2, \ldots) \in \{0,1\}^{\mathbb{N}} \;\middle|\; \exists m \in \mathbb{N} : \forall n \geq m : b_n = 1 \right\}.$$

Let

$$R_m = \left\{ (b_1, b_2, \ldots) \in \{0,1\}^{\mathbb{N}} \;\middle|\; \forall n \geq m : b_n = 1 \right\}.$$

denote the set of sequences where a repetition of ones has started by the m:th "digit". Note that R_m is a finite set, $\#R_m = 2^{m-1}$, since we are only free to choose the values of the first $m-1$ "digits". As a countable union of these finite sets, the complement

$$B \setminus B' = \bigcup_{m=1}^{\infty} R_m.$$

is countable by Proposition A.15. Now if B' would be countable, then the union $B = B' \cup (B \setminus B')$ would also be countable, which is a contradiction with the conclusion of Example A.16. We thus conclude that B' is uncountable.

To prove that the set of real numbers is uncountable, we note that any real number has a binary expansion. More specifically, any number $x \in [0,1)$ has a binary expansion with its "digit sequence" in B'. Indeed, define a function f on $[0,1)$ by

$$f(x) = (b_1, b_2, \ldots) \qquad \text{where } b_j = \lfloor 2^j x \rfloor.$$

The sequence $f(x) = (b_1, b_2, \ldots)$ is a binary expansion of x,

$$x = \sum_{j=1}^{\infty} b_j \, 2^{-j}.$$

It is easy to see that $f(x) \in B'$ and that $f\colon [0,1) \to B'$ is surjective onto B' (for $b \in B'$ and $x = \sum_{j=1}^{\infty} b_j \, 2^{-j}$ we indeed have $f(x) = b$). Therefore we conclude that the cardinality of B' is less than or equal to the cardinality of $[0,1)$. But since B' is uncountable, also the set $[0,1)$ is uncountable.

Since $[0,1) \subset \mathbb{R}$ is a subset and $[0,1)$ is uncountable, also the set $\mathbb{R}$ of real numbers is uncountable.

Appendix B

Topological preliminaries

This appendix reviews necessary backgound on topological notions, in particular properties of real numbers.

B.1. Topological properties of the real line

The set $\mathbb{R}$ of real number inherits its topology from the natural notion of distance of points on the line: the distance $\varrho(x, y)$ of two numbers $x, y \in \mathbb{R}$ is the absolute value of their difference

$$\varrho(x, y) := |x - y|.$$

The distance function $\varrho\colon \mathbb{R} \times \mathbb{R} \to [0, +\infty)$ is called the metric on $\mathbb{R}$, see Section B.2 for a summary of metric space topology more generally. Here we first remind the reader of some fundamental properties of the topology of the real line specifically.

Extended real line

Frequently during the present course it is convenient to extend the real line $\mathbb{R}$ by two symbols, $-\infty$ and $+\infty$, and consider the *extended real line*

$$\widehat{\mathbb{R}} = \mathbb{R} \cup \{-\infty\} \cup \{+\infty\}\,. \tag{B.1}$$

We also alternatively denote the extended real line by $\widehat{\mathbb{R}} = [-\infty, +\infty]$, because it has the topology of a closed interval with $-\infty$ and $+\infty$ as its endpoints, as will be detailed later in Example B.19. Likewise, we denote

$$\begin{aligned}
[-\infty, +\infty) &= \mathbb{R} \cup \{-\infty\} \\
(-\infty, +\infty] &= \mathbb{R} \cup \{+\infty\} \\
(-\infty, +\infty) &= \mathbb{R}.
\end{aligned}$$

Supremum and infimum

One of the key defining properties of real numbers is the completeness property that non-empty bounded subsets have least upper bounds and greatest lower bounds. In the present text, supremum and infimum are used for these notions generalized to the setup of the extended real line, and to a setup where we do not even require non-emptiness and boundedness of the subset.

Definition B.1 (Supremum and infimum)**.**

Let $A \subset \widehat{\mathbb{R}}$ be a subset of the extended real line.

The *supremum*, $\sup(A)$, is the least upper bound of A, i.e., the smallest $M \in [-\infty, \infty]$ such that $x \leq M$ for all $x \in A$.

The *infimum*, $\inf(A)$, is the greatest lower bound of A, i.e., the greatest $M \in [-\infty, \infty]$ such that $x \geq M$ for all $x \in A$.

For non-empty subsets $\emptyset \neq A \subset \mathbb{R}$ of the real line, we have $\sup(A) \in (-\infty, +\infty]$, whereas for the empty set $\emptyset$ we have $\sup(\emptyset) = -\infty$. Likewise, for non-empty subsets $\emptyset \neq A \subset \mathbb{R}$ of the real line, we have $\inf(A) \in [-\infty, +\infty)$, whereas for the empty set $\emptyset$ we have $\inf(\emptyset) = +\infty$.

For indexed collections $(x_j)_{j \in J}$, we denote the supremum and infimum also by

$$\sup_{j \in J} x_j := \sup\Big(\big\{ x_j \;\big|\; j \in J \big\} \Big) \qquad \text{and} \qquad \inf_{j \in J} x_j := \inf\Big(\big\{ x_j \;\big|\; j \in J \big\} \Big).$$

Sequences of numbers

In probability theory, we frequently encounter sequences of real numbers. Crualial notions about sequences include in particular convergence (limits), monotonicity, and upper and lower limits (limsup and liminf).

Convergence of sequences of numbers

The usual notion of limit of a sequence of real numbers is the following.

Definition B.2 (Convergence of real number sequences)**.**

Let $x_1, x_2, x_3, \ldots$ be a sequence of real numbers.

The sequence is said to *converge to* (or *tend to*) a *limit* $x \in \mathbb{R}$ if for all $\varepsilon > 0$ there exists $n_0 = n_0(\varepsilon) \in \mathbb{N}$ such that for all $n \geq n_0$ we have $|x_n - x| < \varepsilon$. We then denote

$$\lim_{n \to \infty} x_n = x \qquad \text{or} \qquad x_n \underset{n \to \infty}{\longrightarrow} x.$$

In addition to the above usual notion of limit of the sequence inside $\mathbb{R}$, we consider also limits $+\infty$ and $-\infty$.

Definition B.3 (Convergence of real number sequences towards infinities)**.**

Let $x_1, x_2, x_3, \ldots$ be a sequence of real numbers.

The sequence is said to *converge to* (or *tend to*) $+\infty$ if for all $M > 0$ there exists $n_0 = n_0(M) \in \mathbb{N}$ such that for all $n \geq n_0$ we have $x_n > M$. We then denote $\lim_{n \to \infty} x_n = +\infty$ or $x_n \to +\infty$.

The sequence is said to *converge to* (or *tend to*) $-\infty$ if for all $M < 0$ there exists $n_0 = n_0(M) \in \mathbb{N}$ such that for all $n \geq n_0$ we have $x_n < M$. We then denote $\lim_{n \to \infty} x_n = -\infty$ or $x_n \to -\infty$.

Monotone sequences of numbers

A sequence $x_1, x_2, x_3, \ldots$ of real numbers is said to be *increasing* if

$$x_1 \leq x_2 \leq x_3 \leq \cdots .$$

If the sequence $x_1, x_2, x_3, \ldots$ is increasing, then it has a limit $x \in (-\infty, +\infty]$, and we denote

$$x_n \uparrow x.$$

It is easy to see that the limit of an increasing sequence is its supremum

$$x = \sup_{n \in \mathbb{N}} x_n.$$

Likewise, a sequence $x_1, x_2, x_3, \ldots$ of real numbers is said to be *decreasing* if

$$x_1 \geq x_2 \geq x_3 \geq \cdots .$$

If the sequence $x_1, x_2, x_3, \ldots$ is decreasing, then it has a limit $x \in [-\infty, +\infty)$, and we denote

$$x_n \downarrow x.$$

It is easy to see that the limit of a decreasing sequence is its infimum

$$x = \inf_{n \in \mathbb{N}} x_n.$$

Sequences of numbers which are either increasing or decreasing are said to be *monotone*.

Upper and lower limits of sequences

For a sequence $x_1, x_2, x_3, \ldots$ of real numbers, we define the *upper limit*

$$\limsup_n x_n := \inf_{m \in \mathbb{N}} \left(\sup_{n \geq m} x_n \right)$$

Note that if we define $c_m = \sup_{n \geq m} x_n$, then the sequence $c_1, c_2, \ldots$ of numbers in $(-\infty, +\infty]$ is decreasing, and its limit is $\lim_{m \to \infty} c_m = \limsup_n x_n$.

We also define the *lower limit*

$$\liminf_n x_n := \sup_{m \in \mathbb{N}} \left(\inf_{n \geq m} x_n \right)$$

Note that if we define $d_m = \inf_{n \geq m} x_n$, then the sequence $d_1, d_2, \ldots$ of numbers in $[-\infty, +\infty)$ is increasing, and its limit is $\lim_{m \to \infty} d_m = \liminf_n x_n$.

Proposition B.4 (Limit in terms of upper and lower limits)**.**
For any sequence $x_1, x_2, x_3, \ldots$ of real numbers we have

$$\liminf_n x_n \leq \limsup_n x_n.$$

The equality above holds if and only if the sequence is convergent, and it this case $\liminf_n x_n = \limsup_n x_n = \lim_{n \to \infty} x_n$.

Exercise B.1. Prove Proposition B.4.

Countability properties on the real line

We make use of the following facts on a few occasions.

Exercise B.2 (Open subsets of the real line are countable unions of open intervals)**.**
Show that any open set $V \subset \mathbb{R}$ is the union of at most countably many open intervals.

Hint: *Show that every point $x \in V$ is contained in some interval $(a, b) \subset V$ with rational endpoints $a, b \in \mathbb{Q}$.*

Proposition B.5 (Open sets as countable unions of disjoint open intervals)**.**
Any open set $V \subset \mathbb{R}$ is the union of countably many disjoint open intervals.

Exercise B.3. Give a proof of the above proposition, using the following steps.

(a) Show that any open set $V \subset \mathbb{R}$ is the union of countably many open intervals.
Hint: *Show that every point $x \in V$ is contained in some interval $(a, b) \subset V$ with rational endpoints $a, b \in \mathbb{Q}$.*

(b) Show that any open set $V \subset \mathbb{R}$ is the union of countably many disjoint open intervals.
Hint: *Show that every point $x \in V$ is contained in a unique maximal interval (a, b) within the set V. Use part (a) to show that there are at most countably many different such maximal intervals.*

Proposition B.6 (Monotone functions have countably many discontinuities)**.**
A monotone function $f\colon \mathbb{R} \to \mathbb{R}$ has countably many points of discontinuity.

Exercise B.4. Prove the above proposition.
Hint: *Consider f restricted to an interval $[k, k+1]$. For a given $m \in \mathbb{N}$, how many jumps of size at least $\frac{1}{m}$ can f have on such an interval?*

B.2. Metric space topology

Basic concepts of metric space topology

Recall that a metric space is a set $\mathfrak{X}$ equipped with a metric, i.e., a function $\varrho\colon \mathfrak{X} \times \mathfrak{X} \to [0, \infty)$ such that

$$\begin{aligned} &\varrho(x,y) = 0 \quad \Leftrightarrow \quad x = y && && (\varrho\text{-Sep})\\ &\varrho(x,y) = \varrho(y,x) && \forall x, y \in \mathfrak{X} && (\varrho\text{-Sym})\\ &\varrho(x,y) \le \varrho(x,z) + \varrho(z,y) && \forall x, y, z \in \mathfrak{X}. && (\varrho\text{-Tri}) \end{aligned}$$

Example B.7 (Real line as a metric space)**.**
The set of real numbers $\mathbb{R}$ equipped with the usual metric $\varrho(x, y) = |x - y|$ is a metric space.

Example B.8 (Euclidean space as a metric space)**.**
Consider the d-dimensional real vector space $\mathbb{R}^d$. The *Euclidean norm* of a vector $x = (x_1, \dots, x_d) \in \mathbb{R}^d$ is

$$\|x\| = \sqrt{x_1^2 + \cdots + x_d^2}.$$

The space $\mathbb{R}^d$ equipped with the metric $\varrho(x, y) = \|x - y\|$ is a metric space.

Example B.9 (Discrete metric)**.**
Let S be any set. If we think of the points of S as discrete, then we may define the *discrete metric* on S by

$$\varrho_{\mathrm{dis}}(x,y) = \begin{cases} 1 & \text{if } x \neq y \\ 0 & \text{if } x = y. \end{cases}$$

The set S equipped with the metric ϱ_{dis} is a metric space.

Example B.10 (Uniform norm)**.**
Consider the space of continuous real-valued functions on a closed interval $[a,b] \subset \mathbb{R}$

$$\mathcal{C}([a,b]) := \{f \colon [a,b] \to \mathbb{R} \text{ continuous}\}.$$

The *supremum norm* (or *uniform norm*) of a function $f \in \mathcal{C}([a,b])$ is

$$\|f\|_\infty = \sup_{x \in [a,b]} |f(x)|.$$

The space $\mathcal{C}([a,b])$ equipped with the metric $\varrho(f,g) = \|f-g\|_\infty$ is a metric space.

Exercise B.5. Verify in each of the examples above that the given metric indeed satisfies the axioms (ϱ-Sep), (ϱ-Sym), and (ϱ-Tri).

We will use the following topological notions in metric spaces.

Definition B.11 (Open balls)**.**
Let $x \in \mathfrak{X}$ and $r > 0$. The (open) *ball* of radius r centered at x is the subset

$$\mathcal{B}_r(x) = \left\{ y \in \mathfrak{X} \,\middle|\, \varrho(x,y) < r \right\}.$$

Definition B.12 (Open sets)**.**
A subset $A \subset \mathfrak{X}$ is *open*, if for all its points some ball centered at that point is contained in the set A (i.e., $\forall x \in A\ \exists r > 0 : \mathcal{B}_r(x) \subset A$).

Exercise B.6 (Open balls are open sets)**.**
Prove that any open ball $\mathcal{B}_r(x) \subset \mathfrak{X}$ is an open set.

Definition B.13 (Closed sets)**.**
A subset $A \subset \mathfrak{X}$ is *closed*, if its complement $\mathfrak{X} \setminus A$ is open.

Note that for example the empty set $\emptyset \subset \mathfrak{X}$ and the whole space $\mathfrak{X}$ are both open and closed. There are also sets which are neither open nor closed.

Definition B.14 (Limit of a sequence)**.**
A sequence $(x_n)_{n\in\mathbb{N}}$ of points $x_n \in \mathfrak{X}$ *converges* to $x \in \mathfrak{X}$ if $\varrho(x_n, x) \to 0$ as $n \to \infty$. We then call x the *limit* of the sequence and denote

$$\lim_{n\to\infty} x_n = x \qquad \text{or} \qquad x_n \underset{n\to\infty}{\longrightarrow} x.$$

Exercise B.7 (The limit of a real number sequence in terms of metric)**.**
Verify that the usual notion of a limit of a sequence of real numbers given in Definition B.2 coincides with Definition B.14 in the special case when the metric space is $\mathbb{R}$ (Example B.7).

Definition B.15 (Continuous functions between metric spaces)**.**
If $(\mathfrak{X}^{(1)}, \varrho^{(1)})$ and $(\mathfrak{X}^{(2)}, \varrho^{(2)})$ are two metric spaces, then a function $f\colon \mathfrak{X}^{(1)} \to \mathfrak{X}^{(2)}$ is *continuous* if for any convergent sequence $(x_n)_{n\in\mathbb{N}}$ of points $x_n \in \mathfrak{X}^{(1)}$ the sequence $(f(x_n))_{n\in\mathbb{N}}$ converges in $\mathfrak{X}^{(2)}$ and

$$\lim_{n\to\infty} f(x_n) = f(\lim_{n\to\infty} x_n).$$

The notions of convergence of sequences and continuity of functions can be formulated in purely topological terms, without reference to metric, only using the notion of open sets.

Proposition B.16 (Topological characterization of continuity)**.**
A function $f\colon \mathfrak{X}^{(1)} \to \mathfrak{X}^{(2)}$ between two metric spaces $(\mathfrak{X}^{(1)}, \varrho^{(1)})$ and $(\mathfrak{X}^{(2)}, \varrho^{(2)})$ is continuous if and only if for every open set V in $\mathfrak{X}^{(2)}$, the preimage $f^{-1}(V) = \big\{x \in \mathfrak{X}^{(1)} \;\big|\; f(x) \in V\big\}$ is open in $\mathfrak{X}^{(1)}$.

Proposition B.17 (Topological characterization of limits)**.**
A sequence $(x_n)_{n\in\mathbb{N}}$ of points in a metric space $(\mathfrak{X}, \varrho)$ converges to $x \in \mathfrak{X}$ if and only if for every open set $U \subset \mathfrak{X}$ containing the point x, there exists $n_0 = n_0(U)$ such that for all $n \geq n_0$ we have $x_n \in U$.

Exercise B.8. Verify that the convergence of a sequence and continuity of a function can be equivalently defined in terms of open sets as stated in Propositions B.17 and B.16 above.

Definition B.18 (Homeomorphism)**.**
If $(\mathfrak{X}^{(1)}, \varrho^{(1)})$ and $(\mathfrak{X}^{(2)}, \varrho^{(2)})$ are two metric spaces, then a function $f\colon \mathfrak{X}^{(1)} \to \mathfrak{X}^{(2)}$ is a *homeomorphism* if f is bijective and both $f\colon \mathfrak{X}^{(1)} \to \mathfrak{X}^{(2)}$ and its inverse $f^{-1}\colon \mathfrak{X}^{(2)} \to \mathfrak{X}^{(1)}$ are continuous. Two spaces are *homeomorphic* if there exists a homeomorphism between them.

According to Proposition B.16, for a bijective function f to be a homeomorphism, a characterizing property is that a subset $U \subset \mathfrak{X}^{(1)}$ is open if and only if its image $f(U) \subset \mathfrak{X}^{(2)}$ is open. Since all topological properties can be formulated using the notion of open sets, homeomorphisms are exactly the mappings which preserve all topological properties.

Example B.19 (The topology of the extended real line)**.**
Consider the function $h\colon [-\frac{\pi}{2}, +\frac{\pi}{2}] \to [-\infty, +\infty]$ defined by

$$h(s) = \begin{cases} -\infty & \text{if } s = -\frac{\pi}{2} \\ \tan(s) & \text{if } -\frac{\pi}{2} < s < +\frac{\pi}{2} \\ +\infty & \text{if } s = +\frac{\pi}{2}. \end{cases}$$

It is easy to see that h is bijective.

The restriction of the function h to the open interval $(-\frac{\pi}{2}, +\frac{\pi}{2})$ is continuous (it is the trigonometric function $\tan\colon (-\frac{\pi}{2}, +\frac{\pi}{2}) \to \mathbb{R}$) and has continuous inverse (the inverse is $\arctan\colon \mathbb{R} \to (-\frac{\pi}{2}, +\frac{\pi}{2})$). It is therefore provides a homeomorphism from the open interval $(-\frac{\pi}{2}, +\frac{\pi}{2})$ to the real line $\mathbb{R}$.

We can define a topology on the extended real line $\widehat{\mathbb{R}} = [-\infty, +\infty]$ by requiring that $h\colon [-\frac{\pi}{2}, +\frac{\pi}{2}] \to \widehat{\mathbb{R}}$ is a homeomorphism. With this definition, in particular, a subset $A \subset \widehat{\mathbb{R}}$

is open if and only if $h^{-1}(A) \subset [-\frac{\pi}{2}, +\frac{\pi}{2}]$ is open. The topology of $\widehat{\mathbb{R}}$ can be obtained for example using the metric $\varrho(x, y) = |h^{-1}(x) - h^{-1}(y)|$, but since this choice of metric is not canonical and in particular does not agree with the usual metric on the subset $\mathbb{R} \subset \widehat{\mathbb{R}}$, we usually prefer not to use an explicitly chosen metric on $\widehat{\mathbb{R}}$.

Exercise B.9 (Convergence on the extended real line)**.**
Verify that our definition of convergence of a real sequence $(x_n)_{n\in\mathbb{N}}$ to $+\infty$ (respectively to $-\infty$) in Definition B.3 is equivalent to the convergence of that sequence in the topological space $\widehat{\mathbb{R}}$ to the point $+\infty \in \widehat{\mathbb{R}}$ (resp. to $-\infty \in \widehat{\mathbb{R}}$).

Appendix C

Dynkin's identification and monotone class theorem

In this appendix we give the proof of Dynkin's identification theorem (Theorem II.26) and we state and prove a related result, the Monotone class theorem (Theorem C.2 below), which was used in Lectures IV and IX.

It is possible to study Sections C.2 and C.3 immediately after Lecture II, where Dynkin's identification theorem was stated.

Another option is to study this entire appendix after Lecture IV, where the Monotone class theorem is first used. This latter approach may be more convenient, since the techniques of proofs of both results are very closely related.

C.1. Monotone class theorem

Definition C.1 (Monotone class)**.**
A collection $\mathscr{H}$ of bounded functions from S to $\mathbb{R}$ is said to be a *monotone class* if it satisfies the following conditions:

(MC-1) The constant function 1 belongs to $\mathscr{H}$.
(MC-$\mathbb{R}$) The class $\mathscr{H}$ is a vector space[1] over $\mathbb{R}$.
(MC-$\uparrow$) If $f_1, f_2, \ldots \in \mathscr{H}$ is an increasing sequence of non-negative functions in $\mathscr{H}$ such that the pointwise limit $f = \lim_{n\to\infty} f_n$ is a bounded function, then $f \in \mathscr{H}$.

Theorem C.2 (Monotone class theorem)**.**
Suppose that $\mathscr{H}$ is a monotone class of bounded functions from S to $\mathbb{R}$. Let $\mathscr{J}$ be a π-system on S. Then if $\mathscr{H}$ contains the indicator function $\mathbb{I}_A$ of every set $A \in \mathscr{J}$ in the π-system, then $\mathscr{H}$ contains all bounded $\sigma(\mathscr{J})/\mathscr{B}$-measurable functions.

C.2. Auxiliary results

In the proof of both Dynkin's identification theorem and Monotone class theorem, we use the following definitions and auxiliary observations.

Definition C.3 (D-system)**.**
A collection $\mathscr{D}$ of subsets of S is said to be a *d-system* on S if it satisfies the following conditions:

[1] That $\mathscr{H}$ is a vector space means that it is stable under taking linear combinations of functions: if we have $f_1, f_2 \in \mathscr{H}$ and $c_1, c_2 \in \mathbb{R}$, then we also have $c_1 f_1 + c_2 f_2 \in \mathscr{H}$.

(D-1) $S \in \mathscr{D}$
(D-d) if $A, B \in \mathscr{D}$ and $A \subset B$, then $B \setminus A \in \mathscr{D}$
(D-↑) if $A_1, A_2, \ldots \in \mathscr{D}$ is an increasing sequence of subsets and $A = \bigcup_{n\in\mathbb{N}} A_n$ is its limit, then $A \in \mathscr{D}$.

Proposition C.4 (A characterization of sigma algebras)**.**
A collection $\mathscr{C}$ of subsets of S is a σ-algebra if and only if it is both a d-system and a π-system.

Proof. The "if" direction is obvious: a σ-algebra clearly satisfies both the conditions of a d-system and of a π-system. It thus suffices to prove the "only if" part.

Suppose that $\mathscr{C}$ is both a π-system and a d-system. We must show that $\mathscr{C}$ is a σ-algebra.

By property (D-1) we have $S \in \mathscr{C}$, so $\mathscr{C}$ satisfies condition (Σ-1) of σ-algebras. If $A \in \mathscr{C}$, then we have $A^{\mathbf{c}} = S \setminus A \in \mathscr{C}$ by properties (D-1) and (D-d), so $\mathscr{C}$ satisfies condition (Σ-**c**) of σ-algebras. It remains to show condition (Σ-$\cup$), i.e., that $\mathscr{C}$ is stable under countable unions.

Consider first the union of just two sets $A_1, A_2 \in \mathscr{C}$ in the collection. We have $A_1^{\mathbf{c}}, A_2^{\mathbf{c}} \in \mathscr{C}$ by d-system properties and then $A_1^{\mathbf{c}} \cap A_2^{\mathbf{c}} \in \mathscr{C}$ by π-system properties. Using De Morgan's law we then observe that

$$A_1 \cup A_2 = S \setminus (A_1^{\mathbf{c}} \cap A_2^{\mathbf{c}}) \in \mathscr{C},$$

again by d-system properties. From this, by induction one gets that $\mathscr{C}$ is stable under finite unions.

Now consider a countable sequence of sets $A_1, A_2, \ldots \in \mathscr{C}$. Denote $G_n = A_1 \cup \cdots \cup A_n$. The induction above allows to conclude that $G_n \in \mathscr{C}$. Then $G_1 \subset G_2 \subset \cdots$ is an increasing sequence of subsets belonging to the collection $\mathscr{C}$, so by d-system property (D-↑) we get that also the limit $G = \bigcup_{n\in\mathbb{N}} G_n$ belongs to the collection $\mathscr{C}$. But by construction $\bigcup_{n\in\mathbb{N}} G_n = \bigcup_{n\in\mathbb{N}} A_n$, so we conclude that $\bigcup_{n\in\mathbb{N}} A_n \in \mathscr{C}$. This shows that $\mathscr{C}$ also satisfies condition (Σ-$\cup$) of σ-algebras, and finishes the proof. □

Definition C.5 (D-system generated by a collection of subsets)**.**
The d-system generated by a collection $\mathscr{I}$ of subsets of S is the smallest d-system $\mathscr{D}$ which contains the collection $\mathscr{I}$. We denote the d-system generated by the collection $\mathscr{I}$ by $d(\mathscr{I})$.

Remark C.6 (Well-definedness of generated d-systems)**.**
The definition makes sense again essentially because the intersection of d-systems is a d-system. The smallest d-system with the property that they contain the collection $\mathscr{I}$ is the intersection of all such d-systems (the intersection is over a non-empty collection since at the very least the power set $\mathscr{P}(S)$ is such a d-system).

Lemma C.7 (Dynkin's lemma)**.**
Suppose that $\mathscr{J}$ is a π-system. Then the d-system $d(\mathscr{J})$ generated by $\mathscr{J}$ and the σ-algebra $\sigma(\mathscr{J})$ generated by $\mathscr{J}$ coincide, $d(\mathscr{J}) = \sigma(\mathscr{J})$.

Proof. Since any sigma-algebra is a d-system, we clearly have $d(\mathscr{J}) \subset \sigma(\mathscr{J})$. By Proposition C.4 it thus suffices to show that $d(\mathscr{J})$ is also a π-system. We show this in two steps.

In the first step our goal is to show that whenever $B \in d(\mathscr{J})$ and $C \in \mathscr{J}$, we have $B \cap C \in d(\mathscr{J})$. Define therefore the collection

$$\mathscr{D}_1 = \Big\{ B \in d(\mathscr{J}) \;\Big|\; B \cap C \in d(\mathscr{J}) \text{ for all } C \in \mathscr{J} \Big\}$$

of sets B with this property. Rephrasing our goal, we wish to show that this collection is simply $\mathscr{D}_1 = d(\mathscr{J})$. By construction we have $\mathscr{D}_1 \subset d(\mathscr{J})$. Moreover, since $\mathscr{J}$ is a π-system, we have $\mathscr{J} \subset \mathscr{D}_1$. Since $d(\mathscr{J})$ is the smallest d-system containing $\mathscr{J}$, showing that $\mathscr{D}_1$ is a d-system will thus achieve our goal: $\mathscr{D}_1 = d(\mathscr{J})$.

For any $C \in \mathscr{J}$ we have $S \cap C = C \in \mathscr{J} \subset d(\mathscr{J})$, so we get that $S \in \mathscr{D}_1$, i.e., property (D-1) holds for $\mathscr{D}_1$. If $A, B \in \mathscr{D}_1$ and $A \subset B$ and $C \in \mathscr{J}$, then we have

$$\big(B \setminus A\big) \cap C = \big(\underbrace{B \cap C}_{\in d(\mathscr{J})}\big) \setminus \big(\underbrace{A \cap C}_{\in d(\mathscr{J})}\big) \in d(\mathscr{J})$$

by property (D-d) of d-system $d(\mathscr{J})$. We get $B \setminus A \in \mathscr{D}_1$, and we thus see that property (D-d) holds for $\mathscr{D}_1$. If $A_1, A_2, \ldots \in \mathscr{D}_1$ is an increasing sequence of subsets and $C \in \mathscr{J}$, then we have

$$\Big(\bigcup_{n\in\mathbb{N}} A_n\Big) \cap C = \bigcup_{n\in\mathbb{N}} \big(\underbrace{A_n \cap C}_{\in d(\mathscr{J})}\big) \in d(\mathscr{J})$$

by property (D-↑) of d-system $d(\mathscr{J})$. We get $\bigcup_{n\in\mathbb{N}} A_n \in \mathscr{D}_1$, and we thus see that property (D-↑) holds for $\mathscr{D}_1$. We conclude that $\mathscr{D}_1$ is a d-system, and therefore that $\mathscr{D}_1 = d(\mathscr{J})$, which was our first goal.

In the second step our goal is to show that whenever $A, B \in d(\mathscr{J})$, we have $A \cap B \in d(\mathscr{J})$. Define therefore the collection

$$\mathscr{D}_2 = \Big\{ A \in d(\mathscr{J}) \;\Big|\; A \cap B \in d(\mathscr{J}) \text{ for all } B \in d(\mathscr{J}) \Big\}$$

of sets A with this property. Rephrasing our second goal, we wish to show that this collection is simply $\mathscr{D}_2 = d(\mathscr{J})$. From the first step we got that $\mathscr{J} \subset \mathscr{D}_2$ and by construction we have $\mathscr{D}_2 \subset d(\mathscr{J})$. Since $d(\mathscr{J})$ is the smallest d-system containing $\mathscr{J}$, showing that $\mathscr{D}_2$ is a d-system will thus achieve our second goal: $\mathscr{D}_2 = d(\mathscr{J})$. The proof that $\mathscr{D}_2$ is a d-system is exactly parallel to the similar argument in the first step.

The conclusion of the second step precisely says that $d(\mathscr{J})$ is a π-system, and the proof is thus complete. □

C.3. Proof of Dynkin's identification theorem

Recall the statement of the nontrivial direction of Dynkin's identification theorem (Theorem II.26): we assume that

- P_1 and P_2 are two probability measures on a measurable space $(\Omega, \mathscr{F})$
- $\mathscr{J}$ is a π-system on Ω such that $\sigma(\mathscr{J}) = \mathscr{F}$
- for all $E \in \mathscr{J}$ we have $\mathsf{P}_1[E] = \mathsf{P}_2[E]$.

Then the claim is that the two probability measures are in fact identical,

$$\mathsf{P}_1 = \mathsf{P}_2,$$

i.e., that the equalities $\mathsf{P}_1[E] = \mathsf{P}_2[E]$ hold for all events $E \in \mathscr{F}$.

Proof of Dynkin's identification theorem (Theorem II.26). Let

$$\mathscr{D} = \Big\{ E \in \mathscr{F} \;\Big|\; \mathsf{P}_1[E] = \mathsf{P}_2[E] \Big\}$$

be the collection of those E for which the desired equality $\mathsf{P}_1[E] = \mathsf{P}_2[E]$ holds. To show that $\mathsf{P}_1 = \mathsf{P}_2$ we must show that this collection contains all events, $\mathscr{D} = \mathscr{F}$.

We claim that $\mathscr{D}$ is a d-system on Ω. The defining properties are checked as follows:

- We have $\mathsf{P}_1[\Omega] = 1 = \mathsf{P}_2[\Omega]$ by definition of probability measures. Therefore we see that $\Omega \in \mathscr{D}$, which shows property (D-1) for $\mathscr{D}$.

- Suppose that $A, B \in \mathscr{D}$ and $A \subset B$. Write $B = A \cup (B \setminus A)$, which is a disjoint union, so by finite additivity of probabilities for disjoint sets (Lemma II.19) we get
$$\mathsf{P}_1[B] = \mathsf{P}_1[A] + \mathsf{P}_1[B \setminus A]$$
and similarly for P_2. We solve this for the probability of $B \setminus A$ and get
$$\begin{aligned} \mathsf{P}_1[B \setminus A] &= \mathsf{P}_1[B] - \mathsf{P}_1[A] \\ &= \mathsf{P}_2[B] - \mathsf{P}_2[A] \qquad \text{(because } A, B \in \mathscr{D}\text{)} \\ &= \mathsf{P}_2[B \setminus A]. \end{aligned}$$
This shows that $B \setminus A \in \mathscr{D}$, and establishes property (D-d) for $\mathscr{D}$.
- Suppose that $A_1, A_2, \ldots \in \mathscr{D}$ and $A_n \uparrow A$. By monotone convergence of probabilities (Theorem II.22) we get $\mathsf{P}_1[A_n] \uparrow \mathsf{P}_1[A]$ and similarly for P_2. Therefore we have
$$\begin{aligned} \mathsf{P}_1[A] &= \lim_{n\to\infty} \mathsf{P}_1[A_n] \\ &= \lim_{n\to\infty} \mathsf{P}_2[A_n] \qquad \text{(because } A_n \in \mathscr{D}\text{)} \\ &= \mathsf{P}_2[A]. \end{aligned}$$
This shows that $A \in \mathscr{D}$, and establishes property (D-↑) for $\mathscr{D}$.

We have shown that $\mathscr{D}$ is a d-system. By assumption, $\mathscr{J}$ is contained in it, $\mathscr{J} \subset \mathscr{D}$, so also the d-system generated by $\mathscr{J}$ must be contained in it, $d(\mathscr{J}) \subset \mathscr{D}$. Since $\mathscr{J}$ is a π-system, we get from Dynkin's lemma (Lemma C.7) that $d(\mathscr{J}) = \sigma(\mathscr{J})$, and by assumption we have $\sigma(\mathscr{J}) = \mathscr{F}$. Therefore we conclude $\mathscr{F} \subset \mathscr{D}$. By definition of $\mathscr{D}$ this means that $\mathsf{P}_1[E] = \mathsf{P}_2[E]$ hold for all events $E \in \mathscr{F}$. □

C.4. Proof of Monotone class theorem

With the preparations in Section C.2 we can prove also the Monotone class theorem. It is worth observing that the proof steps are exactly those of the "standard machine" of integration theory (see Lecture VII).

Proof of Theorem C.2 (Monotone class theorem). Let $\mathscr{H}$ be a monotone class of bounded functions from S to $\mathbb{R}$. Define $\mathscr{D}$ as the collection of subsets $A \subset S$ whose indicator belongs to the monotone class, $\mathbb{I}_A \in \mathscr{H}$. Properties (MC-1), (MC-$\mathbb{R}$), and (MC-↑) of $\mathscr{H}$ imply that $\mathscr{D}$ has properties (D-1), (D-d), and (D-↑), i.e., $\mathscr{D}$ is a d-system.

Now assume, as in the statement, that $\mathscr{J}$ is a π-system such that $\mathscr{H}$ contains the indicator function of each member of $\mathscr{J}$. Then we have $\mathscr{J} \subset \mathscr{D}$. Since $\mathscr{D}$ is a d-system, also $d(\mathscr{J}) \subset \mathscr{D}$. Furthermore $d(\mathscr{J}) = \sigma(\mathscr{J})$ by Dynkin's lemma, so we actually have $\sigma(\mathscr{J}) \subset \mathscr{D}$. In other words, all indicator functions of sets in $\sigma(\mathscr{J})$ are in the monotone class $\mathscr{H}$.

Any bounded simple $\sigma(\mathscr{J})/\mathscr{B}$-measurable function is a finite linear combination of indicator functions of sets in $\sigma(\mathscr{J})$, so by the above observation $\sigma(\mathscr{J}) \subset \mathscr{D}$ and the vector space property (MC-$\mathbb{R}$) we have that the monotone class $\mathscr{H}$ contains such bounded simple functions.

If $f\colon S \to \mathbb{R}$ is a non-negative bounded $\sigma(\mathscr{J})/\mathscr{B}$-measurable function, by the approximation lemma (Lemma III.18) we may find a sequence $f_1, f_2, \ldots$ of non-negative simple $\sigma(\mathscr{J})/\mathscr{B}$-measurable functions such that $f_n \uparrow f$. We observed that the simple functions are in the monotone class, $f_n \in \mathscr{H}$. Therefore, by property (MC-↑), their limit is as well, $f \in \mathscr{H}$.

We have shown that $\mathscr{H}$ contains all non-negative bounded $\sigma(\mathscr{J})/\mathscr{B}$-measurable functions. For a general bounded $\sigma(\mathscr{J})/\mathscr{B}$-measurable function f, write $f = f_+ - f_-$ where $f_+ = \max(f, 0)$ and $f_- = \max(-f, 0)$ are non-negative, bounded and $\sigma(\mathscr{J})/\mathscr{B}$-measurable. As this linear combination, f itself belongs to the monotone class, $f \in \mathscr{H}$. This finishes the proof. □

Appendix D

Monotone convergence theorem

This appendix is devoted to the proof of the Monotone convergence theorem from Lecture VII. Let us recall its statement.

Theorem (Monotone convergence theorem, Theorem VII.8)**.**
If $f_1, f_2, \ldots \in \mathrm{m}\mathscr{S}^+$ and $f_n \uparrow f$ as $n \to \infty$, then we have

$$\int^+ f_n \,\mathrm{d}\mu \uparrow \int^+ f \,\mathrm{d}\mu \qquad \text{as } n \to \infty.$$

We will prove this in a number of steps, gradually increasing the generality of the non-negative limit function f as well as the non-negative approximating functions f_n.

D.1. Monotone convergence theorem for simple functions

First, notice that the Monotone convergence theorem for integrals is certainly closely related to the following monotone convergence of measures, which we proved in Lecture II.

Proposition (Part (II.9) of Lemma II.19)**.**
If $A_1, A_2, \ldots \in \mathscr{S}$ and $A_n \uparrow A$ as $n \to \infty$, then we have $\mu[A_n] \uparrow \mu[A]$.

With this initial observation, our first step is the following monotone convergence result for simple functions approximating an indicator function.

Lemma D.1 (Monotone convergence for simple approximations of an indicator)**.**
If we have $A \in \mathscr{S}$, and if $h_1, h_2, \ldots \in \mathrm{s}\mathscr{S}^+$ are such that $h_n \uparrow \mathbb{I}_A$ as $n \to \infty$, then we have

$$\int^\square h_n \,\mathrm{d}\mu \uparrow \mu[A] \qquad \text{as } n \to \infty.$$

Proof. Since we have $h_n \leq \mathbb{I}_A$, the monotonicity of the integral $\int^\square$ (Lemma VII.3) yields

$$\int^\square h_n \,\mathrm{d}\mu \leq \int^\square \mathbb{I}_A \,\mathrm{d}\mu = \mu[A],$$

so it suffices to prove that

$$\liminf_n \int^\square h_n \,\mathrm{d}\mu \geq \mu[A].$$

Let $0 < \varepsilon < 1$. Define $A_n = \big\{s \in S \;\big|\; h_n(s) > 1 - \varepsilon\big\}$. Then by the assumption $h_n \uparrow \mathbb{I}_A$, we have $A_n \uparrow A$. Monotone convergence of measures (Lemma II.19) thus gives $\mu[A_n] \uparrow \mu[A]$.

Note also that $h_n \geq (1-\varepsilon)\,\mathbb{I}_{A_n}$ by construction of A_n, so the monotonicity of the integral $\int^{\square}$ gives

$$\int^{\square} h_n \,\mathrm{d}\mu \geq (1-\varepsilon)\,\mu[A_n].$$

Taking the lower limit as $n \to \infty$ of this inequality and recalling $\mu[A_n] \uparrow \mu[A]$, we get

$$\liminf_n \int^{\square} h_n \,\mathrm{d}\mu \geq (1-\varepsilon)\,\mu[A].$$

Since $\varepsilon > 0$ can be taken arbitrarily small, we conclude that

$$\liminf_n \int^{\square} h_n \,\mathrm{d}\mu \geq \mu[A],$$

which finishes the proof. □

Next we prove the monotone convergence theorem for simple functions.

Lemma D.2 (Monotone convergence theorem for simple functions)**.** *If $h_1, h_2, \ldots \in \mathrm{s}\mathscr{S}^+$ and $h_n \uparrow h \in \mathrm{s}\mathscr{S}^+$ as $n \to \infty$, then we have*

$$\int^{\square} h_n \,\mathrm{d}\mu \uparrow \int^{\square} h \,\mathrm{d}\mu \qquad \text{as } n \to \infty.$$

Proof. Write $h = \sum_{k=1}^m a_k \,\mathbb{I}_{A_k}$, with $a_1, \ldots, a_m > 0$ and $A_1, \ldots, A_m \in \mathscr{S}$ disjoint. Then we have that $\frac{1}{a_k}\,\mathbb{I}_{A_k} h_n \uparrow \mathbb{I}_{A_k}$ as $n \to \infty$ by assumption $h_n \uparrow h$. Now the assertion follows from Lemma D.1 and linearity of the integral. □

D.2. Monotone convergence theorem for general non-negative functions

Above we established the Monotone convergence theorem for simple functions. The next steps first relax the assumption that the limit function is simple, and then relax the assumption that the approximating functions are simple.

Let us start by verifying that for any non-negative measurable function there exists at least some sequence of simple functions for which the conclusion of the Monotone convergence theorem holds.

Lemma D.3 (Monotone convergence for some simple approximating sequence)**.** *For any $f \in \mathrm{m}\mathscr{S}^+$ there exists a sequence $g_1, g_2, \ldots \in \mathrm{s}\mathscr{S}^+$ such that as $n \to \infty$, we have*

$$g_n \uparrow f \qquad \text{and} \qquad \int^{\square} g_n \,\mathrm{d}\mu \uparrow \int^{+} f \,\mathrm{d}\mu.$$

Proof. By Definition VII.4 we have

$$\int^{+} f \,\mathrm{d}\mu := \sup_{\substack{h \in \mathrm{s}\mathscr{S}^+ \\ 0 \leq h \leq f}} \int^{\square} h \,\mathrm{d}\mu,$$

so there exists some sequence $h_1, h_2, \ldots \in \mathrm{s}\mathscr{S}^+$ such that the integral of f is approximated as

$$\int^{\square} h_n \,\mathrm{d}\mu \uparrow \int^{+} f \,\mathrm{d}\mu \qquad \text{as } n \to \infty. \tag{D.1}$$

In addition, by the approximation lemma (Lemma III.18), there exists a sequence $f_1, f_2, \ldots \in \mathrm{s}\mathscr{S}^+$ such that the function f is approximated as

$$f_n \uparrow f \qquad \text{as } n \to \infty. \tag{D.2}$$

Define a new sequence $g_1, g_2, \ldots \in \mathrm{s}\mathscr{S}^+$ in terms of the two sequences above as

$$g_n := \max\{f_n, h_1, h_2, \ldots, h_n\}. \tag{D.3}$$

By construction this sequence is pointwise increasing, $g_1 \le g_2 \le \cdots$. The constructed functions are also simple, $g_n \in \mathrm{s}\mathscr{S}^+$ (the possible values of the maximum of finitely many simple functions are the finitely many values of these simple functions together).

Moreover, we clearly have $f_n \le g_n \le f$, so from (D.2) we also get

$$g_n \uparrow f \qquad \text{as } n \to \infty.$$

Likewise, we clearly have $h_n \le g_n \le f$, so by monotonicity of the integral $\int^\square$ and definition of the integral $\int^+$ we get the inequalities

$$\int^\square h_n \,\mathrm{d}\mu \le \int^\square g_n \,\mathrm{d}\mu \le \int^+ f \,\mathrm{d}\mu.$$

These inequalities, together with (D.1), give

$$\int^\square g_n \,\mathrm{d}\mu \uparrow \int^+ f \,\mathrm{d}\mu \qquad \text{as } n \to \infty.$$

This finishes the proof. □

In the remaining steps of the proof of the Monotone convergence theorem, we use the following auxiliary result about monotone increasing arrays twice.

Lemma D.4 (Monotone arrays)**.**
Let $\left(t_n^{(r)}\right)_{n\in\mathbb{N}, r\in\mathbb{N}}$ be an array of numbers $t_n^{(r)} \in [0, +\infty]$, which is increasing in both indices:

$$t_1^{(r)} \le t_2^{(r)} \le t_3^{(r)} \le \cdots \qquad \text{for all } r \in \mathbb{N} \tag{D.4}$$

$$t_n^{(1)} \le t_n^{(2)} \le t_n^{(3)} \le \cdots \qquad \text{for all } n \in \mathbb{N}. \tag{D.5}$$

For any $r \in \mathbb{N}$, denote the limit of the increasing sequence (D.4) *by*

$$t^{(r)} := \lim_{n\to\infty} t_n^{(r)},$$

and for any $n \in \mathbb{N}$, denote the limit of the increasing sequence (D.5) *by*

$$t_n := \lim_{r\to\infty} t_n^{(r)}.$$

Then the sequences $t^{(1)}, t^{(2)}, \ldots$ and $t_1, t_2, \ldots$ are both increasing, and their limits

$$t^{(\infty)} := \lim_{r\to\infty} t^{(r)} \qquad \text{and} \qquad t_\infty := \lim_{n\to\infty} t_n$$

coincide,

$$t^{(\infty)} = t_\infty.$$

Proof. We may assume that the array of numbers is uniformly bounded in the sense that for some $M < +\infty$ we have $t_n^{(r)} \le M$ for all $n \in \mathbb{N}$, $r \in \mathbb{N}$ (consider for example $\arctan(t_n^{(r)})$ if the original table is not uniformly bounded).

Recall the definition $t_n = \lim_{r\to\infty} t_n^{(r)}$, for all $n \in \mathbb{N}$. Since for any n and r we have

$$t_n^{(r)} \le t_{n+1}^{(r)} \le M,$$

taking the limits as $r \to \infty$ we get

$$t_n \leq t_{n+1} \leq M.$$

This shows that the sequence $t_1, t_2, \ldots$ is increasing and bounded by M. It therefore has a limit, which we denote by $t_\infty = \lim_{n\to\infty} t_n$.

Similarly one shows that the sequence $t^{(1)}, t^{(2)}, \ldots$ is increasing and bounded. Denote its limit by $t^{(\infty)} = \lim_{r\to\infty} t^{(r)}$.

Let $\varepsilon > 0$. Since $t_n \uparrow t_\infty$ as $n \to \infty$, we can choose some n_0 such that $t_{n_0} > t_\infty - \frac{1}{2}\varepsilon$. Then, since $t^{(r)}_{n_0} \uparrow t_{n_0}$ as $r \to \infty$, we can choose some r_0 such that $t^{(r_0)}_{n_0} > t_{n_0} - \frac{1}{2}\varepsilon$. We now see, because the sequences are increasing, that

$$t^{(\infty)} \geq t^{(r_0)} \geq t^{(r_0)}_{n_0} > t_{n_0} - \frac{1}{2}\varepsilon > t_\infty - \varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, this shows $t^{(\infty)} \geq t_\infty$. Completely symmetrically one obtains the opposite inequality $t_\infty \geq t^{(\infty)}$. We conclude the equality $t_\infty = t^{(\infty)}$, and the lemma is proven. □

We now improve the result of Lemma D.3, and show that the specific choice of the approximating sequence of simple functions did not matter.

Lemma D.5 (Monotone convergence for simple approximating sequences)**.** *For any $f \in \mathrm{m}\mathscr{S}^+$ and any sequence $h_1, h_2, \ldots \in \mathrm{s}\mathscr{S}^+$ such that $h_n \uparrow f$ as $n \to \infty$, we have*

$$\int^\square h_n \,\mathrm{d}\mu \uparrow \int^+ f \,\mathrm{d}\mu.$$

Proof. Given $f \in \mathrm{m}\mathscr{S}^+$, use Lemma D.3 to get a sequence $g^{(1)}, g^{(2)}, \ldots \in \mathrm{s}\mathscr{S}^+$ such that

$$g^{(r)} \uparrow f \qquad \text{and} \qquad \int^\square g^{(r)} \,\mathrm{d}\mu \uparrow \int^+ f \,\mathrm{d}\mu \qquad \text{as } r \to \infty.$$

Suppose now that $h_1, h_2, \ldots \in \mathrm{s}\mathscr{S}^+$ is another sequence such that $h_n \uparrow f$ as $n \to \infty$. Define, for all $n \in \mathbb{N}$ and $r \in \mathbb{N}$ the function

$$f^{(r)}_n := \min\left\{h_n, g^{(r)}\right\},$$

which is simple, $f^{(r)}_n \in \mathrm{s}\mathscr{S}^+$. Observe that in this situation

$$f^{(r)}_n \uparrow h_n \qquad \text{as } r \to \infty$$
$$\text{and} \qquad f^{(r)}_n \uparrow g^{(r)} \qquad \text{as } n \to \infty.$$

Also denote the integrals of these functions by

$$t^{(r)}_n := \int^\square f^{(r)}_n \,\mathrm{d}\mu.$$

Consider first a fixed $r \in \mathbb{N}$. Since $f^{(r)}_n \uparrow g^{(r)}$ as $n \to \infty$, we can apply the already proven Monotone convergence theorem for simple functions, Lemma D.2, to get

$$t^{(r)}_n = \int^\square f^{(r)}_n \,\mathrm{d}\mu \uparrow \int^\square g^{(r)} \,\mathrm{d}\mu =: t^{(r)} \qquad \text{as } n \to \infty.$$

Now letting $r \to \infty$, we have by the choice of the sequence $g^{(1)}, g^{(2)}, \ldots$ that

$$t^{(r)} = \int^\square g^{(r)} \,\mathrm{d}\mu \uparrow \int^+ f \,\mathrm{d}\mu =: t^{(\infty)} \qquad \text{as } r \to \infty.$$

Consider then a fixed $n \in \mathbb{N}$. Since $f_n^{(r)} \uparrow h_n$ as $r \to \infty$, we can again apply the already proven Monotone convergence theorem for simple functions, Lemma D.2, to get

$$t_n^{(r)} = \int^{\square} f_n^{(r)} \,\mathrm{d}\mu \uparrow \int^{\square} h_n \,\mathrm{d}\mu =: t_n \qquad \text{as } r \to \infty.$$

The integrals t_n of the functions h_n form an increasing sequence (the sequence of functions is increasing and the integral is monotone), so they have a limit, which we denote by t_∞:

$$t_n = \int^{\square} h_n \,\mathrm{d}\mu \uparrow t_\infty \qquad \text{as } n \to \infty.$$

It now follows from Lemma D.4 on monotone increasing arrays that $t_\infty = t^{(\infty)}$. Recalling the definition of $t^{(\infty)}$, we have obtained that as $n \to \infty$,

$$\int^{\square} h_n \,\mathrm{d}\mu = t_n \uparrow t_\infty = t^{(\infty)} = \int^{+} f \,\mathrm{d}\mu$$

and the proof is complete. □

Now, finally, we are ready to prove the Monotone convergence theorem in full generality.

Proof of the Monotone convergence theorem (Theorem VII.8). Let $f_1, f_2, \ldots \in \mathrm{m}\mathscr{S}^+$ be a sequence of non-negative measurable functions, which is pointwise increasing:

$$0 \le f_1(s) \le f_2(s) \le \cdots \qquad \text{for all } s \in S.$$

Let $f\colon S \to [0, +\infty]$ be the pointwise limit of this increasing sequence,

$$f(s) := \lim_{n\to\infty} f_n(s) \qquad \text{for } s \in S,$$

so that we have $f_n \uparrow f$. By Proposition III.14 this limit function is measurable, $f \in \mathrm{m}\mathscr{S}^+$.

For each $n \in \mathbb{N}$, use the staircase function construction of Lemma III.18 to get an increasing approximation of f_n by simple functions: define $f_n^{(r)} := \varsigma_r \circ f_n$ so that

$$f_n^{(r)} \uparrow f_n \qquad \text{as } r \to \infty.$$

Note also that for a fixed $r \in \mathbb{N}$, the r:th staircase function $\varsigma_r\colon [0, +\infty] \to [0, r]$ is monotone, so when applied to the increasing sequence $f_1 \le f_2 \le \cdots$ it produces an increasing sequence $f_1^{(r)} \le f_2^{(r)} \le \cdots$. By left-continuity of ς_r, the limit of this increasing sequence is $f^{(r)} := \varsigma_r \circ f$ (see Remark III.20), i.e.,

$$f_n^{(r)} \uparrow f^{(r)} \qquad \text{as } n \to \infty.$$

In particular, we thus see that the array of simple functions $(f_n^{(r)})_{n\in\mathbb{N}, r\in\mathbb{N}}$ is pointwise increasing in both of its indices, n and r.

Now for $n \in \mathbb{N}$ and $r \in \mathbb{N}$, denote

$$t_n^{(r)} := \int^{\square} f_n^{(r)} \,\mathrm{d}\mu.$$

Since the array of functions is increasing in both indices, by the monotonicity of the integral $\int^{\square}$, also the array $(t_n^{(r)})_{n\in\mathbb{N}, r\in\mathbb{N}}$ of their integrals is increasing in both indices.

Consider first a fixed $r \in \mathbb{N}$. Then we have $f_n^{(r)} \uparrow f^{(r)}$ as $n \to \infty$, where $f^{(r)} = \varsigma_r \circ f$. The functions here are all simple, so we can apply the already proven Monotone convergence theorem for simple functions, Lemma D.2, to get that

$$t_n^{(r)} := \int^{\square} f_n^{(r)} \,\mathrm{d}\mu \uparrow \int^{\square} f^{(r)} \,\mathrm{d}\mu =: t^{(r)} \qquad \text{as } n \to \infty.$$

The functions $f^{(r)} = \varsigma_r \circ f$ themselves constitute an increasing approximation of f by simple functions,

$$f^{(r)} \uparrow f \qquad \text{as } r \to \infty.$$

Therefore we can use the already proven Monotone convergence theorem with simple approximating functions, Lemma D.5, to get

$$t^{(r)} := \int^{\square} f^{(r)} \, \mathrm{d}\mu \uparrow \int^{+} f \, \mathrm{d}\mu =: t^{(\infty)} \qquad \text{as } r \to \infty.$$

Consider then a fixed $n \in \mathbb{N}$. Then we have by construction $f_n^{(r)} \uparrow f_n$ as $r \to \infty$. The approximating functions here are all simple, so we can apply the already proven Monotone convergence theorem with simple approximating functions, Lemma D.5, to get that

$$t_n^{(r)} := \int^{\square} f_n^{(r)} \, \mathrm{d}\mu \uparrow \int^{+} f_n \, \mathrm{d}\mu =: t_n \qquad \text{as } r \to \infty.$$

At this stage we apply again Lemma D.4: it says that as $n \to \infty$ we have $t_n \uparrow t_\infty = t^{(\infty)}$. Recalling what t_n and $t^{(\infty)}$ are, we have obtained

$$\int^{+} f_n \, \mathrm{d}\mu = t_n \uparrow t_\infty = t^{(\infty)} = \int^{+} f \, \mathrm{d}\mu.$$

This is exactly the assertion of the Monotone convergence theorem. □

Appendix E

Orthogonal projections and conditional expected values

This appendix concerns orthogonal projections in the space $\mathcal{L}^2(\mathsf{P})$ of square integrable random variables, and conditional expected values with respect to σ-algebras.

E.1. Geometry of the space of square integrable random variables

The Cauchy-Schwarz inequality, Equation (X.4) in Theorem X.6,

$$\Big|\,\mathsf{E}\big[XY\big]\,\Big| \leq \sqrt{\mathsf{E}\big[X^2\big]\;\mathsf{E}\big[Y^2\big]} \qquad \text{for } X, Y \in \mathcal{L}^2(\mathsf{P})$$

underlies a lot of familiar geometry in the space $\mathcal{L}^2(\mathsf{P})$ of square integrable random variables. We begin by defining inner products, norm, and distances, and establishing results of familiar geometric flavor about them.

Inner product, norm, and distance

Definition E.1 (Inner product and norm for square integrable random variables)**.** For $X, Y \in \mathcal{L}^2(\mathsf{P})$, we denote

$$\langle X, Y\rangle := \mathsf{E}\big[XY\big] \tag{E.1}$$

We call this the *inner product* of the random variables X and Y. If the inner product vanishes, $\langle X, Y\rangle = 0$, then we say that X and Y are *orthogonal* and we denote $X \perp Y$.

For $X \in \mathcal{L}^2(\mathsf{P})$, we denote

$$\|X\| := \sqrt{\langle X, X\rangle} = \sqrt{\mathsf{E}\big[X^2\big]} \tag{E.2}$$

We call this the *norm* (or more specifically 2-*norm*) of the random variable X.

In this notation, the Cauchy-Schwarz inequality reads

$$\big|\langle X, Y\rangle\big| \;\leq\; \|X\|\,\|Y\|.$$

Corollary X.7, in turn, amounts to the following bound

$$\mathsf{E}\big[|X|\big] \leq \|X\|$$

for expected values in terms of the norm.

The norm leads to a notion of distance in $\mathcal{L}^2(\mathsf{P})$: for two square integrable random variables $X, Y \in \mathcal{L}^2(\mathsf{P})$, we interpret the norm of their difference

$$\|X - Y\|$$

as the distance between the two.[1]

Note that the distance satisfies the usual triangle inequality.

Lemma E.2 (Triangle inequality)**.**
For any $X, Y \in \mathcal{L}^2(\mathsf{P})$, we have

$$\|X+Y\| \le \|X\| + \|Y\|. \tag{E.3}$$

Proof. We have, by bilinearity of the inner product and Cauchy-Schwarz inequality

$$\begin{aligned}\|X+Y\|^2 \;=\; \langle X+Y, X+Y\rangle \;&=\; \langle X,X\rangle + \underbrace{2\,\langle X,Y\rangle}_{\le 2\,\|X\|\,\|Y\|} + \langle Y,Y\rangle \\ &\le\; \|X\|^2 + 2\,\|X\|\,\|Y\| + \|Y\|^2 \;=\; \Big(\|X\| + \|Y\|\Big)^2.\end{aligned}$$

The assertion follows by taking square roots. □

Regarding orthogonality, we have the following familiar formula.

Lemma E.3 (Pythagoras' theorem)**.**
If $X, Y \in \mathcal{L}^2$ are orthogonal, $X \perp Y$, then we have

$$\|X+Y\|^2 = \|X\|^2 + \|Y\|^2.$$

Proof. By direct calculation using bilinearity and symmetry of the inner product, we get

$$\begin{aligned}\|X+Y\|^2 &= \langle X+Y, X+Y\rangle \\ &= \langle X,X\rangle + 2\,\underbrace{\langle X,Y\rangle}_{=0} + \langle Y,Y\rangle \;=\; \|X\|^2 + \|Y\|^2.\end{aligned}$$

□

A similar calculation yields another useful formula.

Lemma E.4 (Parallelogram law)**.**
If $X, Y \in \mathcal{L}^2$, then we have

$$\|X+Y\|^2 + \|X-Y\|^2 = 2\,\|X\|^2 + 2\,\|Y\|^2.$$

Proof. By direct calculation using bilinearity and symmetry of the inner product, we get

$$\begin{aligned}\|X+Y\|^2 + \|X-Y\|^2 &= \langle X+Y, X+Y\rangle + \langle X-Y, X-Y\rangle \\ &= 2\,\langle X,X\rangle + 0\,\langle X,Y\rangle + 2\,\langle Y,Y\rangle \;=\; 2\,\|X\|^2 + 2\,\|Y\|^2.\end{aligned}$$

□

[1] It is common to quotient the space $\mathcal{L}^p(\mathsf{P})$ by the equivalence relation

$$X \overset{\text{a.s.}}{=} Y \qquad\Longleftrightarrow\qquad \mathsf{P}\Big[\Big\{\omega\in\Omega \;\Big|\; X(\omega) = Y(\omega)\Big\}\Big] = 1$$

of almost sure equality. This is quite natural, because we have $\mathsf{E}\big[|X|^p\big] = \mathsf{E}\big[|Y|^p\big]$ whenever $X \overset{\text{a.s.}}{=} Y$, and moreover we have $\mathsf{E}\big[|X|^p\big] = 0$ if and only if $X \overset{\text{a.s.}}{=} 0$. The quotient space $L^p(\mathsf{P}) = \mathcal{L}^p(\mathsf{P})/\overset{\text{a.s.}}{=}$ is a vector space, and the formula $\|X\|_p = \big(\mathsf{E}\big[|X|^p\big]\big)^{1/p}$ defines a norm in it. The statements of Proposition E.6 and Exercise E.1 then assert that $L^2(\mathsf{P})$ and $L^p(\mathsf{P})$ are Banach spaces, i.e., a complete normed vector spaces.

There would be certain advantages in identifying random variables which are almost surely equal. We, however, choose not to use this quotient space — agreeing with [**Wil91**], we find it preferable that random variables are functions on Ω instead of equivalence classes of such functions.

The notion of distance, in turn, leads to a notion of convergence of sequences. Compare this notion of convergence in the space $\mathcal{L}^2(\mathsf{P})$ of square integrable random variables with Defition XI.8 on convergence in the space $\mathcal{L}^1(\mathsf{P})$ of integrable random variables.[2]

Definition E.5 (Convergence in $\mathcal{L}^2$)**.**
Suppose that $X_1, X_2, \ldots \in \mathcal{L}^2(\mathsf{P})$ and $X \in \mathcal{L}^2(\mathsf{P})$. We say that X_n *tends to* X *in* $\mathcal{L}^2$ as $n \to \infty$ and denote $X_n \xrightarrow{\mathcal{L}^2} X$, if we have $\|X_n - X\| \to 0$ or equivalently $\mathsf{E}\big[(X_n - X)^2\big] \to 0$ as $n \to \infty$.

Completeness of square integrable random variables

The space of square integrable random variables has the following completeness property.

Proposition E.6 (The space of square integrable random variables is complete)**.**
Suppose that $X_1, X_2, \ldots \in \mathcal{L}^2(\mathsf{P})$ *is a sequence of square integrable random variables which is* Cauchy *in the sense that*

$$\lim_{m\to\infty} \sup_{n,n'\geq m} \|X_n - X_{n'}\| = 0. \tag{E.4}$$

Then there exists a square integrable random variable $X \in \mathcal{L}^2(\mathsf{P})$ *such that we have* $X_n \xrightarrow{\mathcal{L}^2} X$.

Proof. Assuming (E.4), we can choose $m_1 < m_2 < \cdots$ such that

$$\|X_n - X_{n'}\| \leq 2^{-k} \qquad \text{whenever } n, n' \geq m_k. \tag{E.5}$$

By Corollary X.7, we then have also

$$\mathsf{E}\Big[\big|X_n - X_{n'}\big|\Big] \leq \sqrt{\mathsf{E}\big[\big(X_n - X_{n'}\big)^2\big]} = \|X_n - X_{n'}\| \leq 2^{-k} \quad \text{ for } n, n' \geq m_k.$$

In particular we get $\mathsf{E}\big[\big|X_{m_{k+1}} - X_{m_k}\big|\big] \leq 2^{-k}$ and thus

$$\sum_{k=1}^{\infty} \mathsf{E}\Big[\big|X_{m_{k+1}} - X_{m_k}\big|\Big] < +\infty.$$

According to Lemma VIII.6, it follows that almost surely the series

$$\sum_{k=1}^{\infty} \big(X_{m_{k+1}} - X_{m_k}\big)$$

converges absolutely. Let Y denote the sum of this almost surely convergent series,

$$\begin{aligned} Y = \sum_{k=1}^{\infty} \big(X_{m_{k+1}} - X_{m_k}\big) &= \lim_{\ell\to\infty} \sum_{k=1}^{\ell} \big(X_{m_{k+1}} - X_{m_k}\big) \\ &= \lim_{\ell\to\infty} \Big(X_{m_{\ell+1}} - X_{m_\ell} + X_{m_\ell} - X_{m_{\ell-1}} + \cdots + X_{m_3} - X_{m_2} + X_{m_2} - X_{m_1}\Big) \\ &= \lim_{\ell\to\infty} \Big(X_{m_{\ell+1}} - X_{m_1}\Big). \end{aligned}$$

[2]The limit of a sequence converging in $\mathcal{L}^2(\mathsf{P})$ (or in $\mathcal{L}^1(\mathsf{P})$, for that matter) is not strictly speaking uniquely defined. Rather, if $X, X' \in \mathcal{L}^2(\mathsf{P})$ are $\mathcal{L}^2$-limits of the same sequence, then one can only conclude that they are almost surely equal, $X \overset{\text{a.s.}}{=} X'$. We have chosen to embrace this slight non-uniqueness of limits. The alternative approach would again be to quotient by the equivalence relation of almost sure equality, $\overset{\text{a.s.}}{=}$.

By setting $X = Y + X_{m_1}$, we conclude that almost surely

$$\lim_{\ell\to\infty} X_{m_\ell} = X.$$

For any $n \geq m_k$, as a consequence of (E.5) and Fatou's lemma, we obtain

$$\begin{aligned}\mathsf{E}\Big[\big(X_n - X\big)^2\Big] &= \mathsf{E}\Big[\lim_{\ell\to\infty}\big(X_n - X_{m_\ell}\big)^2\Big] \\ &\leq \liminf_{\ell}\ \mathsf{E}\Big[\big(X_n - X_{m_\ell}\big)^2\Big] \\ &= \liminf_{\ell}\ \|X_n - X_{m_\ell}\|^2 \ \leq\ 4^{-k}.\end{aligned}$$

This shows first of all that $X_n - X \in \mathcal{L}^2(\mathsf{P})$, so by the vector space property we get that $X = X_n - (X_n - X)$ is square integrable as well. Moreover, it shows that

$$\lim_{n\to\infty}\ \mathsf{E}\Big[\big(X_n - X\big)^2\Big] = 0,$$

which completes the proof. □

Exercise E.1 (Completeness of $\mathcal{L}^p$)**.**
Prove an analoguous statement for $\mathcal{L}^p(\mathsf{P})$, using Exercise VIII.9 instead of Corollary X.7.

Orthogonal projections to closed subspaces

Definition E.7 (Closed subspaces of square integrable random variables)**.**
A vector subspace $\mathcal{V} \subset \mathcal{L}^2(\mathsf{P})$ in the space of square integrable random variables is said to be *closed*, if for any sequence $X_1, X_2, \ldots \in \mathcal{V}$ which converges in $\mathcal{L}^2(\mathsf{P})$, a limit can be found also within the subspace, i.e., $X_n \xrightarrow{\mathcal{L}^2} X \in \mathcal{V}$.

We say that $Y \in \mathcal{L}^2(\mathsf{P})$ is *orthogonal* to the subspace and denote $Y \perp \mathcal{V}$, if for all $X \in \mathcal{V}$ we have $Y \perp X$.

Given a random variable and a subspace, the natural notion of projecting the random variable to the subspace is obtained by finding the nearest point within the subspace. The existence and almost uniqueness of such a nearest point is guaranteed by the following proposition, which also explains why this is an orthogonal projection.

Proposition E.8 (Orthogonal projection)**.**
Let $\mathcal{V} \subset \mathcal{L}^2(\mathsf{P})$ be a closed subspace of square integrable random variables and let $Y \in \mathcal{L}^2(\mathsf{P})$ be a square integrable random variable. Define the distance of Y to the subspace $\mathcal{V}$ by

$$\Delta\ :=\ \inf_{X\in\mathcal{V}} \|Y - X\|. \tag{E.6}$$

Then for a random variable $Z \in \mathcal{V}$ the following conditions are equivalent:

$$\textit{(i):}\ \|Y - Z\| = \Delta \qquad\qquad \textit{(ii):}\ Y - Z\ \perp\ \mathcal{V}.$$

Furthermore, there exists a random variable $Z \in \mathcal{V}$ with these properties, and if $\widetilde{Z} \in \mathcal{V}$ is another such random variable, then we have $Z = \widetilde{Z}$ almost surely.

Proof. We first show the equivalence of the conditions (i) and (ii), then prove the existence of Z satisfying (ii), and finally prove the almost uniqueness.

proof of (ii) ⇒ (i): Suppose that $Z \in \mathcal{V}$ satisfies (ii). Let $Z' \in \mathcal{V}$ be any other point in the subspace. Because we have $Z - Z' \in \mathcal{V}$, property (ii) implies $Y - Z \perp Z - Z'$. Therefore Pythagoras' theorem gives

$$\begin{aligned}\|Y - Z'\|^2 &= \|Y - Z + Z - Z'\|^2 \\ &= \|Y - Z\|^2 + \|Z - Z'\|^2 \;\geq\; \|Y - Z\|^2.\end{aligned}$$

Therefore Z minimizes the distance to Y within the subspace, $\|Y - Z\| = \Delta$, i.e., (i) holds.

proof of (i) ⇒ (ii): Suppose that $Z \in \mathcal{V}$ satisfies (i), i.e., $\|Y - Z\| = \Delta$. Let $V \in \mathcal{V}$, and for $t \in \mathbb{R}$ consider the vector $Z + tV \in \mathcal{V}$. Then by definition (E.6) we have

$$\begin{aligned}0 &\leq \|Y - Z - tV\|^2 - \Delta^2 \\ &= \underbrace{\|Y - Z\|^2}_{=\Delta^2} - 2t\,\langle Y - Z, V\rangle + t^2\,\|V\|^2 - \Delta^2 \\ &= -2t\,\langle Y - Z, V\rangle + t^2\,\|V\|^2.\end{aligned}$$

If $\langle Y - Z, V\rangle \neq 0$, then this polynomial would obtain negative values for small positive or small negative t, which is a contradiction. Therefore we must have $\langle Y - Z, V\rangle = 0$, i.e., (ii) holds.

proof of existence of minimizer: By definition (E.6), we can find a sequence $Z_1, Z_2, \ldots \in \mathcal{V}$ such that

$$\|Y - Z_n\|^2 \;\leq\; \Delta^2 + \frac{1}{n}.$$

By the parallelogram law, Lemma E.4, for any n, n' we have

$$2\,\|Y - Z_n\|^2 + 2\,\|Y - Z_{n'}\|^2 \;=\; \|2Y - Z_n - Z_{n'}\|^2 + \|Z_n - Z_{n'}\|^2.$$

Note here, that we have $\|2Y - Z_n - Z_{n'}\|^2 = 4\,\|Y - \frac{Z_n + Z_{n'}}{2}\|^2 \geq 4\Delta^2$, since also $\frac{Z_n + Z_{n'}}{2} \in \mathcal{V}$. We thus conclude that

$$\begin{aligned}\|Z_n - Z_{n'}\|^2 \;&=\; 2\,\|Y - Z_n\|^2 + 2\,\|Y - Z_{n'}\|^2 - \|2Y - Z_n - Z_{n'}\|^2 \\ &\leq\; 2\big(\Delta^2 + \frac{1}{n}\big) + 2\big(\Delta^2 + \frac{1}{n'}\big) - 4\Delta^2 \\ &\leq\; \frac{2}{n} + \frac{2}{n'}.\end{aligned}$$

This shows that the sequence $Z_1, Z_2, \ldots$ is Cauchy. Therefore by Proposition E.6, the sequence $Z_1, Z_2, \ldots$ converges in $\mathcal{L}^2(\mathsf{P})$. Since $\mathcal{V}$ is a closed subspace, a limit remains in the subspace, $Z_n \xrightarrow{\mathcal{L}^2} Z \in \mathcal{V}$.

By triangle inequality, we have

$$\Delta \leq \|Y - Z\| = \|Y - Z_n + Z_n - Z\| \leq \underbrace{\|Y - Z_n\|}_{\to\Delta} + \underbrace{\|Z_n - Z\|}_{\to 0} \underset{n\to\infty}{\longrightarrow} \Delta,$$

so we conclude that $\|Y - Z\| = \Delta$. This shows the existence of a random variable $Z \in \mathcal{V}$ satisfying (i), and therefore also (ii).

proof of almost uniqueness: Suppose that $Z, \widetilde{Z} \in \mathcal{V}$ are two random variables satisfying (i) and (ii). Then $Z - \widetilde{Z} \in \mathcal{V}$ and thus by (ii) we have $Z - \widetilde{Z} \perp Y - Z$. Therefore property (i) and Pythagoras theorem again lead to

$$\begin{aligned}\Delta^2 = \|Y - \widetilde{Z}\|^2 &= \|Y - Z + Z - \widetilde{Z}\|^2 \\ &= \|Y - Z\|^2 + \|Z - \widetilde{Z}\|^2 = \Delta^2 + \|Z - \widetilde{Z}\|^2.\end{aligned}$$

This shows that $\|Z - \widetilde{Z}\|^2 = 0$, which implies $Z = \widetilde{Z}$ almost surely. □

Given $Y \in \mathcal{L}^2(\mathsf{P})$ and a closed subspace $\mathcal{V} \subset \mathcal{L}^2(\mathsf{P})$, a random variable Z which satisfies (i) and (ii) is called (a version of) the orthogonal projection of Y to $\mathcal{V}$.

E.2. Conditional expected values

Throughout, let $(\Omega, \mathscr{F}, \mathsf{P})$ be a probability space. Given an integrable random variable Y and a sub-σ-algebra $\mathscr{G} \subset \mathscr{F}$, we will define and study the *conditional expected value* $\mathsf{E}[Y|\mathscr{G}]$ of Y given $\mathscr{G}$. This should be interpreted as the best estimate of Y that can be made based on the information $\mathscr{G}$.[3] We start by discussing the case when Y is square-integrable.

Orthogonal projections as best estimates

Observe that if $\mathscr{G} \subset \mathscr{F}$ is a sub-σ-algebra of $\mathscr{F}$, then the space

$$\mathcal{L}^2(\mathsf{P}) \cap \mathrm{m}\mathscr{G}$$

of square integrable $\mathscr{G}$-measurable random variables is a closed subspace of $\mathcal{L}^2(\mathsf{P})$: it is clearly a vector subspace, and satisfies the completeness property of Proposition E.6 — therefore any sequence in it which converges has a limit in it (convergent sequences are necessarily Cauchy).

Suppose now that the σ-algebra $\mathscr{G}$ represents information available to us, and our task is to form an estimate $\widehat{Y}$ of a random quantity $Y \in \mathcal{L}^2(\mathsf{P})$ based on that information. To formalize the property that the estimate is made based on information $\mathscr{G}$, we require that $\widehat{Y}$ is a $\mathscr{G}$-measurable random variable.[4] The difference $Y - \widehat{Y}$ between the actual value and our estimate is the error we commit, and thus the norm $\|Y - \widehat{Y}\|$ tells us about the magnitude of the error.[5] In this sense, the best possible estimate is obtained by choosing $\widehat{Y} \in \mathcal{L}^2(\mathsf{P}) \cap \mathrm{m}\mathscr{G}$ which minimizes $\|Y - \widehat{Y}\|$. According to Proposition E.8, this is achieved by letting $\widehat{Y}$ be the orthogonal projection of Y onto the subspace $\mathcal{V} = \mathcal{L}^2(\mathsf{P}) \cap \mathrm{m}\mathscr{G}$.

We then observe a key property that the best estimate $\widehat{Y}$ satisfies.

Lemma E.9 (Orthogonal projection to $\mathscr{G}$-measurable random variables)**.** *For $\mathscr{G} \subset \mathscr{F}$ a sub-σ-algebra and $Y \in \mathcal{L}^2$, let $\widehat{Y}$ be (a version of) the orthogonal projection of Y to the closed subspace $\mathcal{L}^2(\mathsf{P}) \cap \mathrm{m}\mathscr{G}$ of $\mathscr{G}$-measurable square integrable random variables. Then for any $G \in \mathscr{G}$ we have*

$$\mathsf{E}\big[\mathbb{I}_G \widehat{Y}\big] \;=\; \mathsf{E}\big[\mathbb{I}_G Y\big].$$

Proof. Obviously $\mathbb{I}_G$ is $\mathscr{G}$-measurable and square integrable, so $\mathbb{I}_G \in \mathcal{V} := \mathcal{L}^2(\mathsf{P}) \cap \mathrm{m}\mathscr{G}$. By definition, $\widehat{Y}$ satisfies property (ii) of Proposition E.8: $Y - \widehat{Y} \perp \mathcal{V}$. Thus we in particular have $Y - \widehat{Y} \perp \mathbb{I}_G$. This can be explicitly written as

$$0 = \langle Y - \widehat{Y},\, \mathbb{I}_G \rangle = \langle Y,\, \mathbb{I}_G \rangle - \langle \widehat{Y},\, \mathbb{I}_G \rangle = \mathsf{E}\big[Y\, \mathbb{I}_G\big] - \mathsf{E}\big[\widehat{Y}\, \mathbb{I}_G\big].$$

The asserted equality follows. □

This observation is the motivating idea, which underlies the abstract definition of conditional expected values.

---

[3] Recall from Lecture IV the interpretation of σ-algebras as information.

[4] Morally, the value $\widehat{Y}$ of the estimate should be a deterministic function of the known information (see Doob's representation theorem, Theorem IV.5).

[5] In particular, $\widehat{Y}$ must be square integrable for the magnitude of error in this sense to be finite.

Definition of conditional expected value

Let now $Y \in \mathcal{L}^1(\mathsf{P})$ and let $\mathscr{G} \subset \mathscr{F}$ be a sub-σ-algebra.

Definition E.10 (Conditional expected value)**.**
A random variable $\widehat{Y} \in \mathcal{L}^1(\mathsf{P})$ is said to be (a version of) the *conditional expected value* (denoted $\mathsf{E}[Y|\mathscr{G}]$) of Y given $\mathscr{G}$, if $\widehat{Y} \in \mathrm{m}\mathscr{G}$ and

$$\mathsf{E}\big[\mathbb{I}_G\, \widehat{Y}\big] = \mathsf{E}\big[\mathbb{I}_G\, Y\big] \qquad \text{for all } G \in \mathscr{G}. \tag{E.7}$$

In particular, if $Y \in \mathcal{L}^2(\mathsf{P})$, then by Lemma E.9 the orthogonal projection of Y to $\mathcal{L}^2(\mathsf{P}) \cap \mathrm{m}\mathscr{G}$ is (a version of) this conditional expected value. In the general case of $Y \in \mathcal{L}^1(\mathsf{P})$, we still have to show that such conditional expected values exist. We start, however, by first addressing their uniqueness (up to the usual amount of ambiguity).

Lemma E.11 (Almost uniqueness of conditional expected values)**.**
Suppose that both $\widehat{Y}$ and $\widehat{Y}'$ are conditional expected values of Y given $\mathscr{G}$. Then the two are almost surely equal, $\widehat{Y} \overset{\text{a.s.}}{=} \widehat{Y}'$.

Proof. For $n \in \mathbb{N}$, let

$$G_n := \left\{ \omega \in \Omega \;\middle|\; \widehat{Y}(\omega) - \widehat{Y}'(\omega) \geq \frac{1}{n} \right\}.$$

Then we have $G_n \in \mathscr{G}$, since $\widehat{Y}$ and $\widehat{Y}'$ are $\mathscr{G}$-measurable. By Markov's inequality, we get

$$\begin{aligned} \frac{1}{n}\,\mathsf{P}[G_n] \leq \mathsf{E}\Big[\mathbb{I}_{G_n}\,(\widehat{Y} - \widehat{Y}')\Big] &= \mathsf{E}\big[\mathbb{I}_{G_n}\,\widehat{Y}\big] - \mathsf{E}\big[\mathbb{I}_{G_n}\,\widehat{Y}'\big] \\ &= \mathsf{E}\big[\mathbb{I}_{G_n}\,Y\big] - \mathsf{E}\big[\mathbb{I}_{G_n}\,Y\big] = 0, \end{aligned}$$

where we used the definition of conditional expected values for both $\widehat{Y}$ and $\widehat{Y}'$. We conclude that $\mathsf{P}[G_n] = 0$, and then using the union bound also

$$\mathsf{P}\big[\widehat{Y} > \widehat{Y}'\big] = \mathsf{P}\Big[\bigcup_{n\in\mathbb{N}} G_n\Big] \leq \sum_{n\in\mathbb{N}} \mathsf{P}[G_n] = \sum_{n\in\mathbb{N}} 0 = 0.$$

By changing the roles of $\widehat{Y}$ and $\widehat{Y}'$, one similarly derives $\mathsf{P}\big[\widehat{Y} < \widehat{Y}'\big] = 0$, and therefore

$$\mathsf{P}\big[\widehat{Y} \neq \widehat{Y}'\big] = 0.$$

By passing to the complementary events, this concludes the proof of $\widehat{Y} \overset{\text{a.s.}}{=} \widehat{Y}'$. □

We will denote the conditional expected values of Y given $\mathscr{G}$ by

$$\mathsf{E}[Y|\mathscr{G}] \;\in\; \mathcal{L}^1(\mathsf{P}) \cap \mathrm{m}\mathscr{G},$$

and not worry too much about the possibility that different choices for it could be made, since any two choices are anyway almost surely equal.

Admitting that conditional expected values exist, the reader can now for example verify the following.

Exercise E.2 (Conditional expected value preserves non-negativity)**.**
Show that if $Y \geq 0$ (almost surely) then also $\mathsf{E}[Y|\mathscr{G}] \geq 0$ (almost surely).

The remaining task is to show that the conditional expected value $\mathsf{E}[Y|\mathscr{G}]$ exists not only when $Y \in \mathcal{L}^2(\mathsf{P})$, but generally for any $Y \in \mathcal{L}^1(\mathsf{P})$. We first do this by

assuming non-negativity of Y, and then it is routine to deal with the general case by splitting to positive and negative parts.

Conditional expected value for integrable random variables

Let Y be a non-negative integrable random variable. Denote by $Y \wedge n$ this random variable truncated at level n:

$$\big(Y \wedge n\big)(\omega) = \min\{Y(\omega), n\} \qquad \text{for } \omega \in \Omega.$$

Then $Y \wedge n$ is bounded and in particular square integrable, $Y \wedge n \in \mathcal{L}^2(\mathsf{P})$. Moreover, we have $Y \wedge n \uparrow Y$ as $n \to \infty$.

We use this approximation by square integrable random variables to construct the conditional expected value of non-negative integrable Y.

Lemma E.12 (Truncation approximation to conditional expected values)**.**
Let Y be a non-negative integrable random variable, and let Z_n be the orthogonal projection of $Y \wedge n$ to the subspace $\mathcal{L}^2(\mathsf{P}) \cap \mathrm{m}\mathscr{G}$. Then there exists an integrable $\mathrm{m}\mathscr{G}$-measurable random variable Z such that $Z_n \uparrow Z$ as $n \to \infty$ (almost surely), and we have

$$\mathsf{E}\big[\mathbb{I}_G\, Z\big] = \mathsf{E}\big[\mathbb{I}_G\, Y\big] \qquad \textit{for any } G \in \mathscr{G}.$$

Proof. Since $Y \wedge (n+1) \geq Y \wedge n$ for any $n \in \mathbb{N}$, it follows from linearity of projection and Exercise E.2 that $Z_{n+1} \geq Z_n$ (almost surely). Therefore the sequence $Z_1, Z_2, \ldots$ of (almost surely) non-negative random variables is (almost surely) increasing, and thus has a limit Z. By the $\mathrm{m}\mathscr{G}$-measurability of each Z_n, the limit Z is also $\mathrm{m}\mathscr{G}$-measurable. Now let $G \in \mathscr{G}$. Then using Lemma E.9 once and the Monotone convergence theorem twice, we calculate

$$\begin{aligned}\mathsf{E}\big[\mathbb{I}_G\, Z\big] = \mathsf{E}\Big[\mathbb{I}_G \big(\lim_{n\to\infty} Z_n\big)\Big] &= \lim_{n\to\infty} \mathsf{E}\big[\mathbb{I}_G\, Z_n\big] \\ &= \lim_{n\to\infty} \mathsf{E}\big[\mathbb{I}_G\, (Y \wedge n)\big] \\ &= \mathsf{E}\Big[\mathbb{I}_G \big(\lim_{n\to\infty} (Y \wedge n)\big)\Big] = \mathsf{E}\big[\mathbb{I}_G\, Y\big].\end{aligned}$$

This is the asserted property of the limit Z, and taking $G = \Omega$ in particular shows that indeed $Z \in \mathcal{L}^1(\mathsf{P})$. □

Proposition E.13 (Existence of conditional expected values)**.**
For any $Y \in \mathcal{L}^1(\mathsf{P})$, (a version of) the conditional expected value $\mathsf{E}[Y|\mathscr{G}]$ exists.

Proof. Decompose Y to its positive and negative parts, $Y = Y_+ - Y_-$. Lemma E.12 can be used to construct conditional expected values $\widehat{Y}_+$ and $\widehat{Y}_-$ of the non-negative integrable random variables Y_+ and Y_- given $\mathscr{G}$. Then $\mathsf{E}[Y|\mathscr{G}] := \widehat{Y}_+ - \widehat{Y}_-$ clearly satisfies the defining property of conditional expected value of Y given $\mathscr{G}$. □

Properties of conditional expected value

In the following we summarize a number of important properties of conditional expected values.

Theorem E.14 (Properties of conditional expected values)**.**
Conditional expected values satisfy the following properties (interpreted in the almost sure sense), when Y and $Y_1, Y_2, \ldots$ are integrable random variables.

(i) *If $Y \in \mathrm{m}\mathscr{G}$, then we have $\mathsf{E}[Y|\mathscr{G}] = Y$.*
(ii) *We have $\mathsf{E}\big[\mathsf{E}[Y|\mathscr{G}]\big] = \mathsf{E}[Y]$.*
(iii) *For $c_1, c_2 \in \mathbb{R}$, we have $\mathsf{E}\big[c_1 Y_1 + c_2 Y_2 \,\big|\, \mathscr{G}\big] = c_1\,\mathsf{E}[Y|\mathscr{G}] + c_2\,\mathsf{E}[Y_2|\mathscr{G}]$.*
(iv) *If $\mathscr{H} \subset \mathscr{G} \subset \mathscr{F}$ are σ-algebras, then we have $\mathsf{E}\big[\,\mathsf{E}[Y|\mathscr{G}] \,\big|\, \mathscr{H}\big] = \mathsf{E}[Y|\mathscr{H}]$.*
(v) *If $Z \in \mathrm{m}\mathscr{G}$ and $ZY \in \mathcal{L}^1(\mathsf{P})$, then we have $\mathsf{E}[ZY|\mathscr{G}] = Z\,\mathsf{E}[Y|\mathscr{G}]$.*
(vi) *If $\mathscr{G} \perp\!\!\!\perp \sigma(Y)$, then we have $\mathsf{E}[Y|\mathscr{G}] = \mathsf{E}[Y]$.*

Remark E.15 (Interpretations of the properties of conditional expected values)**.**
The properties in the theorem above have rather intuitive interpretations.

(i) The best estimate of a known quantity $Y \in \mathrm{m}\mathscr{G}$ is the quantity Y itself.
(ii) The best estimate of a quantity Y is unbiased, in the sense that it has the same expected value as the quantity Y itself.
(iii) The best estimate of a linear combination of quantities is the corresponding linear combination of the best estimates.
(iv) Suppose that a person H possesses less information than a person G. If H tries to form an estimate about the best estimate that G makes about some quantity Y, then the best she can do is to use her own best estimate of the quantity Y.
(v) Known quantities can be treated like constants when forming best estimates.
(vi) Any information that is independent of Y can not be used to estimate Y any better than the expected value $\mathsf{E}[Y]$ of Y.

Proof of Theorem E.14. Property (i) is immediate from the defining equation (E.7) and the (almost) uniqueness of conditional expected value (Lemma E.11).

Property (ii) follows by taking $G = \Omega \in \mathscr{G}$ in the defining equation (E.7) of conditional expected value.

Property (iii) is a consequence of the linearity of the defining equation (E.7) and the (almost) uniqueness of conditional expected value (Lemma E.11).

To prove property (iv), note first that both $\mathsf{E}[Y|\mathscr{H}]$ and $\mathsf{E}\big[\,\mathsf{E}[Y|\mathscr{G}] \,\big|\, \mathscr{H}\big]$ are by construction $\mathscr{H}$-measurable and integrable, so it only remains to verify the defining equation (E.7). For $H \in \mathscr{H} \subset \mathscr{G} \subset \mathscr{F}$, using the defining property of conditional expected values, we get

$$\mathsf{E}\Big[\mathbb{I}_H\,\mathsf{E}\big[\,\mathsf{E}[Y|\mathscr{G}] \,\big|\, \mathscr{H}\big]\Big] = \mathsf{E}\Big[\mathbb{I}_H\,\mathsf{E}[Y|\mathscr{G}]\Big] \;=\; \mathsf{E}\big[\mathbb{I}_H\,Y\big].$$

Since conditional expected values are (almost) uniquely defined, this proves that

$$\mathsf{E}\big[\,\mathsf{E}[Y|\mathscr{G}] \,\big|\, \mathscr{H}\big] = \mathsf{E}[Y|\mathscr{H}].$$

We leave it as an exercise to the reader to prove property (v) by the "standard machine", i.e, by verifying the claim successively when the random variable Z is a $\mathscr{G}$-measurable indicator, a simple random variable, a non-negative random variable, and finally in the full generality of the assertion.

To prove property (vi), note that the assumed independence $\mathscr{G} \perp\!\!\!\perp \sigma(Y)$ implies that $\mathbb{I}_G \perp\!\!\!\perp Y$ for any $G \in \mathscr{G}$. Therefore we get

$$\mathsf{E}\big[\mathbb{I}_G\,Y\big] = \mathsf{E}[\mathbb{I}_G]\,\mathsf{E}[Y] \;=\; \mathsf{E}\Big[\,\mathbb{I}_G\,\mathsf{E}[Y]\,\Big],$$

which by (almost) uniqueness of conditional expected value shows that $\mathsf{E}[Y|\mathscr{G}] = \mathsf{E}[Y]$. □

Also for example Monotone convergence theorem, Dominated convergence theorem, Fatou's lemma, Jensen's inequality, etc. hold in the appropriate form for conditional expected values, and their proofs are straightforward modifications of the corresponding ones for usual expected values.

Appendix F

Characteristic functions

This appendix is devoted to the proof of two results from Lecture XII: Lévy's inversion theorem (Theorem XII.7), by which the distribution of a random variable is recovered from its characteristic function, and Theorem XII.9 about equivalent conditions for convergence in distribution formulated in terms of cumulative distribution functions or characteristic functions.

F.1. Lévy's inversion theorem

Let

$$X\colon \Omega \to \mathbb{R}$$

be a real valued random variable. In this section, we use the following notational conventions regarding its distribution:

- $\nu = P_X$, the distribution of X, a probability measure on $\mathbb{R}$ such that
$$\nu[B] = \mathsf{P}\big[X \in B\big] \qquad \text{for Borel sets } B \in \mathscr{B}.$$
- $\varphi = \varphi_X$, the characteristic function of X, a function $\mathbb{R} \to \mathbb{C}$ such that
$$\varphi(\theta) = \mathsf{E}\big[e^{\mathbb{i}\,\theta X}\big] = \int_{\mathbb{R}} e^{\mathbb{i}\theta x}\,\mathrm{d}\nu(x) \qquad \text{for } \theta \in \mathbb{R}.$$
- $F = F_X$, the cumulative distribution function of X, a function $\mathbb{R} \to [0,1]$ such that
$$F(x) = \mathsf{P}\big[X \le x\big] = \nu\big[(-\infty, x]\big] \qquad \text{for } x \in \mathbb{R}.$$

The statement of Lévy's inversion theorem is the following.

Theorem (Lévy's inversion theorem, Theorem XII.7)**.**
For any $a, b \in \mathbb{R}$, $a < b$, we have

$$\lim_{T\to+\infty} \frac{1}{2\pi} \int_{-T}^{+T} \frac{e^{-\mathbb{i}\theta a} - e^{-\mathbb{i}\theta b}}{\mathbb{i}\,\theta}\,\varphi(\theta)\,\mathrm{d}\theta \tag{F.1}$$

$$= \nu\big[(a,b)\big] + \frac{1}{2}\,\nu\big[\,\{a\}\,\big] + \frac{1}{2}\,\nu\big[\,\{b\}\,\big] \tag{F.2}$$

Moreover, if $\int_{\mathbb{R}} |\varphi(\theta)|\,\mathrm{d}\theta < +\infty$, then X has a continuous probability density function f_X given by

$$f_X(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-\mathbb{i}\theta x}\,\varphi(\theta)\,\mathrm{d}\theta. \tag{F.3}$$

Remark F.1. Formula (F.2) is occasionally written in terms of the cumulative distribution function as well. Recall from Proposition II.30 that the cumulative distribution function F is

increasing and right-continuous: if $x_n \downarrow x$ then $F(x_n) \downarrow F(x)$. Since it is increasing and bounded, the left limits also exist: they are defined by

$$F(x^-) := \lim_{x' \uparrow x} F(x').$$

By monotone convergence of probability measures (Theorem II.22), the left limits can be expressed in terms of the distribution ν as the following measures of open semi-infinite intervals

$$\begin{aligned} F(x^-) &= \lim_{n\to\infty} F\big(x - \frac{1}{n}\big) \\ &= \lim_{n\to\infty} \nu\Big[\big(-\infty, x - \frac{1}{n}\big]\Big] \\ &= \nu\Big[\bigcup_{n=1}^{\infty} \big(-\infty, x - \frac{1}{n}\big]\Big] \;=\; \nu\big[(-\infty, x)\big]. \end{aligned}$$

In particular, if there is a discontinuity in the cumulative distribution function at a point x, then the size of the jump $F(x) - F(x^-)$ is the probability mass located at the single point x,

$$F(x) - F(x^-) = \nu\big[(-\infty, x]\big] - \nu\big[(-\infty, x)\big] = \nu\big[\,\{x\}\,\big]$$

(at continuity points x of F there is no probability mass, $\nu[\{x\}] = 0$).

For any $a < b$, the increment from a to b of the cumulative distribution function is

$$F(b) - F(a) = \nu\big[(-\infty, b]\big] - \nu\big[(-\infty, a]\big] = \nu\big[(a, b]\big].$$

If we replace $F(x)$ by the average $\frac{1}{2}\big(F(x) + F(x^-)\big)$ of the left and right limits, then the corresponding increment becomes

$$\frac{1}{2}\Big(F(b) + F(b^-)\Big) - \frac{1}{2}\Big(F(a) + F(a^-)\Big) = \nu\big[(a, b)\big] + \frac{1}{2}\nu\big[\,\{a\}\,\big] + \frac{1}{2}\nu\big[\,\{b\}\,\big].$$

This allows to alternatively write (F.2) in terms of the cumulative distribution function F.

In the proof we use the following auxiliary calculation.

Lemma F.2 (An auxiliary integral)**.**
For $r \in \mathbb{R}$ define

$$S(r) := \int_0^r \frac{\sin(\theta)}{\theta}\, \mathrm{d}\theta.$$

Then the limits as $r \uparrow +\infty$ and as $r \downarrow -\infty$ of this integral are

$$\lim_{r\uparrow+\infty} S(r) = \frac{\pi}{2} \qquad \text{and} \qquad \lim_{r\downarrow-\infty} S(r) = -\frac{\pi}{2}.$$

Moreover, for any $c \in \mathbb{R}$, we have

$$\int_0^r \frac{\sin(c\,\theta)}{\theta}\, \mathrm{d}\theta \;=\; S(c\,r).$$

Proof. Let us start from the last part: performing the change of variables $\theta' = c\,\theta$ we obtain the asserted formula

$$\int_0^r \frac{\sin(c\,\theta)}{\theta}\, \mathrm{d}\theta \;=\; \int_0^{cr} \frac{\sin(\theta')}{\theta'}\, \mathrm{d}\theta' \;=\; S(c\,r).$$

In particular taking $c = -1$ we see that $S(-r) = S(r)$, so it suffices to consider $r > 0$.

The improper Riemann integral $\lim_{r\uparrow+\infty} S(r)$ is one which is routine to evaluate with complex analysis and residue calculus. Indeed, the principal value integral of $\frac{e^{\mathrm{i}z}}{z}$ over the real line evaluates to $\mathrm{i}\,\pi$ times the residue at $z = 0$,

$$\text{P.V.} \int_{-\infty}^{\infty} \frac{e^{\mathrm{i}z}}{z}\, \mathrm{d}z = \mathrm{i}\pi\, \mathrm{Res}_{z=0}\Big(\frac{e^{\mathrm{i}z}}{z}\Big) = \mathrm{i}\pi.$$

The result follows from this by taking the imaginary part and using parity. □

Proof of Theorem XII.7. We first prove the equality of (F.1) and (F.2). Then we prove the assertion about the probability density assuming integrability of the characteristic function.

proof of (F.1)=(F.2)*:* For any $u, v \in \mathbb{R}$, we have $|e^{\mathbb{i}u} - e^{\mathbb{i}v}| \leq |u - v|$. We deduce that

$$\int_{\mathbb{R}} \left(\int_{-T}^{+T} \underbrace{\left| \frac{e^{\mathbb{i}\theta(x-a)} - e^{\mathbb{i}\theta(x-b)}}{\mathbb{i}\,\theta} \right|}_{\leq \frac{|\mathbb{i}\theta b - \mathbb{i}\theta a|}{|\mathbb{i}\theta|} \leq |b-a|} \mathrm{d}\theta \right) \mathrm{d}\nu(x) \;\leq\; 2T\,|b-a| \;<\; +\infty.$$

This finiteness of the double integral of absolute value justifies the use of Fubini's theorem to exchange the order of integrations in expression (F.1)

$$\begin{aligned} \int_{-T}^{+T} \frac{e^{-\mathbb{i}\theta a} - e^{-\mathbb{i}\theta b}}{\mathbb{i}\,\theta}\, \varphi(\theta)\, \mathrm{d}\theta &= \int_{-T}^{+T} \frac{e^{-\mathbb{i}\theta a} - e^{-\mathbb{i}\theta b}}{\mathbb{i}\,\theta} \left(\int_{\mathbb{R}} e^{\mathbb{i}\theta x}\, \mathrm{d}\nu(x) \right) \mathrm{d}\theta \\ &= \int_{\mathbb{R}} \left(\int_{-T}^{+T} \frac{e^{\mathbb{i}\theta(x-a)} - e^{\mathbb{i}\theta(x-b)}}{\mathbb{i}\,\theta}\, \mathrm{d}\theta \right) \mathrm{d}\nu(x). \end{aligned}$$

In the last expression, the inner integral over θ is such that the imaginary part of the integrand is an odd function and the real part of the integrand is an even function, so only the real part survives integration over the symmetric interval $[-T, T]$. We are able to express the result in terms of the integrals $S(r)$ of Lemma F.2:

$$\begin{aligned} \int_{-T}^{+T} \frac{e^{\mathbb{i}\theta(x-a)} - e^{\mathbb{i}\theta(x-b)}}{\mathbb{i}\,\theta}\, \mathrm{d}\theta &= \int_{-T}^{+T} \frac{\sin\big(\theta(x-a)\big) - \sin\big(\theta(x-b)\big)}{\theta}\, \mathrm{d}\theta \\ &= 2\int_{0}^{+T} \frac{\sin\big(\theta(x-a)\big)}{\theta}\, \mathrm{d}\theta - 2\int_{0}^{+T} \frac{\sin\big(\theta(x-b)\big)}{\theta}\, \mathrm{d}\theta \\ &= 2\,S\big((x-a)T\big) - 2\,S\big((x-b)T\big). \end{aligned}$$

As $T \to +\infty$, it follows from Lemma F.2 that we have

$$\int_{-T}^{+T} \frac{e^{\mathbb{i}\theta(x-a)} - e^{\mathbb{i}\theta(x-b)}}{\mathbb{i}\,\theta}\, \mathrm{d}\theta \longrightarrow \begin{cases} 2\pi & \text{if } a < x < b \\ \pi & \text{if } x = a \text{ or } x = b \\ 0 & \text{if } x < a \text{ or } x > b, \end{cases}$$

and the left hand side expression is uniformly bounded in T (as an upper bound for its absolute value we may use for example $4 \sup_{r>0} |S(r)| < +\infty$). The bounded convergence theorem therefore says that the limit $T \to +\infty$ can be interchanged with the integration over the probability measure ν:

$$\int_{\mathbb{R}} \left(\int_{-T}^{+T} \frac{e^{\mathbb{i}\theta(x-a)} - e^{\mathbb{i}\theta(x-b)}}{\mathbb{i}\,\theta}\, \mathrm{d}\theta \right) \mathrm{d}\nu(x) \underset{T\to\infty}{\longrightarrow} 2\pi\,\nu\big[(a,b)\big] + \pi\,\nu\big[\,\{a\}\,\big] + \pi\,\nu\big[\,\{b\}\,\big].$$

This concludes the proof of the first assertion, the equality of (F.1) and (F.2).

proof of probability density part: Suppose now that $\int_{\mathbb{R}} |\varphi(x)|\, \mathrm{d}x < +\infty$. Note that the function (F.3)

$$f_X(x) = \frac{1}{2\pi} \int_{\mathbb{R}} e^{-\mathbb{i}\theta x}\, \varphi(\theta)\, \mathrm{d}\theta.$$

is continuous: if $x_n \to x$ then we have the pointwise limit $e^{-\mathbb{i}\theta x_n}\, \varphi(\theta) \to e^{-\mathbb{i}\theta x}\, \varphi(\theta)$ by continuity of the exponential function, and domination $|e^{-\mathbb{i}\theta x_n}\, \varphi(\theta)| \leq |\varphi(\theta)|$ by an integrable function, so it follows from dominated convergence theorem that $f_X(x_n) \to f_X(x)$. It therefore remains to show that this f_X is a probability density for the measure ν.

Next note that

$$\left| \frac{e^{-\mathbb{i}\theta a} - e^{-\mathbb{i}\theta b}}{\mathbb{i}\theta}\, \varphi(x) \right| \leq |a-b|\, |\varphi(x)|.$$

Therefore by the Dominated convergence theorem (domination by constant multiple of the integrable function $|\varphi(x)|$), the left hand side (F.1) can be written simply as

$$\frac{1}{2\pi}\int_{\mathbb{R}} \frac{e^{-\mathrm{i}\theta a} - e^{-\mathrm{i}\theta b}}{\mathrm{i}\,\theta}\,\varphi(\theta)\,\mathrm{d}\theta.$$

Also by dominated convergence (domination again by multiple of $|\varphi(x)|$), if $a_n \to a$ and $b_n \to b$, we get

$$\frac{1}{2\pi}\int_{\mathbb{R}} \frac{e^{-\mathrm{i}\theta a_n} - e^{-\mathrm{i}\theta b_n}}{\mathrm{i}\,\theta}\,\varphi(\theta)\,\mathrm{d}\theta \;\longrightarrow\; \frac{1}{2\pi}\int_{\mathbb{R}} \frac{e^{-\mathrm{i}\theta a} - e^{-\mathrm{i}\theta b}}{\mathrm{i}\,\theta}\,\varphi(\theta)\,\mathrm{d}\theta.$$

Using the result of the first part and rewriting (F.2) in terms of the cumulative distribution function F, we get

$$\frac{F(b_n)+F(b_n^-)}{2} - \frac{F(a_n)+F(a_n^-)}{2} \longrightarrow \frac{F(b)+F(b^-)}{2} - \frac{F(a)+F(a^-)}{2}.$$

This shows that the cumulative distribution function F is continuous.

Again, since we have

$$\Big|\frac{e^{-\mathrm{i}\theta a} - e^{-\mathrm{i}\theta b}}{\mathrm{i}\theta(b-a)}\,\varphi(x)\Big| \le |\varphi(x)|,$$

we may calculate the derivative of the cumulative distribution function F by dominated convergence as follows:

$$\begin{aligned} F'(a) = \lim_{b\to a}\frac{F(b)-F(a)}{b-a} &= \lim_{b\to a}\frac{1}{2\pi}\int_{\mathbb{R}}\frac{e^{-\mathrm{i}\theta a} - e^{-\mathrm{i}\theta b}}{\mathrm{i}\,\theta\,(b-a)}\,\varphi(\theta)\,\mathrm{d}\theta \\ &= \frac{1}{2\pi}\int_{\mathbb{R}} e^{\mathrm{i}\theta a}\,\varphi(\theta)\,\mathrm{d}\theta. \end{aligned}$$

Thus the derivative is given by the expression in (F.3), $F'(x) = f_X(x)$. It follows that f_X is a probability density, because for any interval $(a,b) \subset \mathbb{R}$ we have

$$\nu\big[(a,b)\big] = F(b) - F(a) = \int_a^b F'(x)\,\mathrm{d}x = \int_a^b f_X(x)\,\mathrm{d}x$$

and intervals form a π-system which uniquely determine the probability measure ν. □

Let us at this point also prove a lemma which allows us to control the amount of probability mass that is outside a large interval $[-r,+r]$ in terms of the behavior of the characteristic function $\varphi(\theta)$ near $\theta = 0$.

Lemma F.3 (A bound on the probability mass outside an interval)**.** *Let ν be a probability measure on $(\mathbb{R},\mathscr{B})$, and let $\varphi(\theta) = \int_{\mathbb{R}} e^{\mathrm{i}\theta x}\,\mathrm{d}\nu(s)$ be its characteristic function. Then for any $r > 0$ we have*

$$\nu\Big[\mathbb{R}\setminus[-r,+r]\Big] \le \frac{r}{2}\int_{-2/r}^{2/r}\big(1-\varphi(\theta)\big)\,\mathrm{d}\theta.$$

Proof. Let us denote $u = 2/r$ for convenience. Start with the following calculation of an integral

$$\int_{-u}^{u}(1-e^{\mathrm{i}\theta x})\,\mathrm{d}\theta = 2u - \frac{1}{\mathrm{i}x}\big(e^{\mathrm{i}ux} - e^{-\mathrm{i}ux}\big) = 2u - \frac{2\sin(ux)}{x} = 2u\Big(1-\frac{\sin(ux)}{ux}\Big).$$

Let us then integrate both sides of this equation over the variable x with respect to the measure ν. On the left hand side we get, using Fubini's theorem,

$$\begin{aligned} \int_{\mathbb{R}}\Big(\int_{-u}^{u}(1-e^{\mathrm{i}\theta x})\,\mathrm{d}\theta\Big)\,\mathrm{d}\nu(x) &= \int_{-u}^{u}\Big(\int_{\mathbb{R}}(1-e^{\mathrm{i}\theta x})\,\mathrm{d}\nu(x)\Big)\,\mathrm{d}\theta \\ &= \int_{-u}^{u}\big(1-\varphi(\theta)\big)\,\mathrm{d}\theta. \end{aligned}$$

Dividing by u and equating with what we get on the right hand side gives

$$\frac{1}{u}\int_{-u}^{u}\big(1-\varphi(\theta)\big)\,\mathrm{d}\theta \;=\; 2\int_{\mathbb{R}}\Big(1-\frac{\sin(ux)}{ux}\Big)\,\mathrm{d}\nu(x).$$

Because $\frac{\sin(\xi)}{\xi}\le 1$ for any $\xi\in\mathbb{R}$, we see that the integrand on the right hand side here is non-negative, $1-\frac{\sin(ux)}{ux}\ge 0$. Therefore omitting from the integral over $\mathbb{R}$ the part over $[-r,r]\subset\mathbb{R}$ yields a lower bound

$$\frac{1}{u}\int_{-u}^{u}\big(1-\varphi(\theta)\big)\,\mathrm{d}\theta \;\ge\; 2\int_{\mathbb{R}\setminus[-r,r]}\Big(1-\frac{\sin(ux)}{ux}\Big)\,\mathrm{d}\nu(x).$$

Then observe that for $|x|>r=2/u$ we have $|\frac{\sin(ux)}{ux}|\le\frac{1}{u|x|}\le\frac{1}{2}$, so the integrand on the remaining part satisfies $1-\frac{\sin(ux)}{ux}\ge\frac{1}{2}$, which yields

$$\frac{1}{u}\int_{-u}^{u}\big(1-\varphi(\theta)\big)\,\mathrm{d}\theta \;\ge\; 2\int_{\mathbb{R}\setminus[-r,r]}\frac{1}{2}\,\mathrm{d}\nu(x) \;=\; \nu\Big[\mathbb{R}\setminus[-r,+r]\Big].$$

Recalling that $u=2/r$, this is exactly the asserted inequality. □

F.2. Equivalent conditions for convergence in distribution

Theorem (Theorem XII.9)**.**
Let $X_1, X_2, \dots$ and X be real-valued random variables. Let also $\nu_1, \nu_2, \dots$ and ν be their laws, and let $F_1, F_2, \dots$ and F be their cumulative distribution functions, and let $\varphi_1, \varphi_2, \dots$ and φ be their characteristic functions, respectively. Then the following conditions are equivalent:

(i) *For all bounded continuous functions $f\colon\mathbb{R}\to\mathbb{R}$ we have*

$$\int_{\mathbb{R}} f(x)\,\mathrm{d}\nu_n(x)\longrightarrow\int_{\mathbb{R}} f(x)\,\mathrm{d}\nu(x)\qquad\text{as } n\to\infty.$$

(ii) *We have $F_n(x)\to F(x)$ as $n\to\infty$ for all points $x\in\mathbb{R}$ such that F is continuous at x.*

(iii) *We have $\varphi_n(\theta)\to\varphi(\theta)$ as $n\to\infty$ for all $\theta\in\mathbb{R}$.*

Proof. We first show the equivalence of (i) and (ii) by proving both implications "(i) ⇒ (ii)" and "(ii) ⇒ (i)".

Then we show that "(i) ⇒ (iii)". Finally, we show that "(iii) ⇒ (ii)", making use of the previously established implications. The equivalence of all three conditions then follows.

proof of "(i) ⇒ (ii)": Assume (i), i.e., that for all bounded continuous functions $f\colon\mathbb{R}\to\mathbb{R}$ we have

$$\int_{\mathbb{R}} f(x)\,\mathrm{d}\nu_n(x)\underset{n\to\infty}{\longrightarrow}\int_{\mathbb{R}} f(x)\,\mathrm{d}\nu(x).$$

Let $x_0\in\mathbb{R}$ be a continuity point of F. Fix $\varepsilon>0$. Then by continuity of F at x_0, for some $\delta>0$ we have

$$F(x_0-\delta)>F(x_0)-\varepsilon\qquad\text{and}\qquad F(x_0+\delta)<F(x_0)+\varepsilon. \tag{F.4}$$

Now define two piecewise linear functions $f^-, f^+\colon\mathbb{R}\to\mathbb{R}$ by

$$f^-(x)=\begin{cases}1 & \text{if } x\le x_0-\delta\\ \frac{x_0-x}{\delta} & \text{if } x_0-\delta<x<x_0\\ 0 & \text{if } x_0\le x\end{cases},\qquad f^+(x)=\begin{cases}1 & \text{if } x\le x_0\\ \frac{x_0+\delta-x}{\delta} & \text{if } x_0<x<x_0+\delta\\ 0 & \text{if } x_0+\delta\le x\end{cases}.$$

These functions are chosen so that we have the pointwise inequalities

$$\mathbb{I}_{(-\infty,x_0-\delta]}(x) \,\le\, f^-(x) \,\le\, \mathbb{I}_{(-\infty,x_0]}(x) \,\le\, f^+(x) \,\le\, \mathbb{I}_{(-\infty,x_0+\delta]}(x). \tag{F.5}$$

If we integrate (F.5) over x with respect to the measure ν, the indicators integrate to values of the cumulative distribution function F, and by monotonicity of integral, we get

$$F(x_0-\delta) \,\le\, \int f^-\,\mathrm{d}\nu \,\le\, F(x_0) \,\le\, \int f^+\,\mathrm{d}\nu \,\le\, F(x_0+\delta).$$

Combining with inequalities (F.4) and rearranging terms, we derive the estimates

$$\int f^+\,\mathrm{d}\nu - \varepsilon < F(x_0) < \int f^-\,\mathrm{d}\nu + \varepsilon.$$

Similarly we can integrate (F.5) with respect to ν_n for any $n \in \mathbb{N}$ to get

$$\int f^-\,\mathrm{d}\nu_n \,\le\, F_n(x_0) \,\le\, \int f^+\,\mathrm{d}\nu_n.$$

To estimate $F_n(x_0) - F(x_0)$, we can combine these with the previous inequalities and rearrange to the form

$$\int f^-\,\mathrm{d}\nu_n - \int f^-\,\mathrm{d}\nu - \varepsilon \;<\; F_n(x_0) - F(x_0) \;<\; \int f^+\,\mathrm{d}\nu_n - \int f^+\,\mathrm{d}\nu + \varepsilon.$$

Now since the functions f^- and f^+ are bounded and continuous, by our assumption (i) there exists some N such that for $n \ge N$ we have

$$\Big|\int f^-\,\mathrm{d}\nu_n - \int f^-\,\mathrm{d}\nu\Big| < \varepsilon \quad \text{and} \quad \Big|\int f^+\,\mathrm{d}\nu_n - \int f^+\,\mathrm{d}\nu\Big| < \varepsilon.$$

For $n \ge N$, our estimate on $F_n(x_0) - F(x_0)$ therefore yields

$$-2\varepsilon \;<\; F_n(x_0) - F(x_0) \;<\; +2\varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, we can conclude $\lim_{n\to\infty} F_n(x_0) = F(x_0)$, which establishes (ii).

proof of "(ii) $\Rightarrow$ (i)*":* Assume (ii), i.e., that $\lim_{n\to\infty} F_n(x) = F(x)$ for all $x \in D$, where $D \subset \mathbb{R}$ is the set of continuity points of F. The increasing function F can only have countably many points of discontinuity, so the complement $\mathbb{R} \setminus D$ is countable, and thus $D \subset \mathbb{R}$ is dense.[1]

Let $\varepsilon > 0$. Since $F(x) \downarrow 0$ as $x \downarrow -\infty$ and $F(x) \uparrow 1$ as $x \uparrow +\infty$ (Proposition II.30), and since $D \subset \mathbb{R}$ is dense, we can choose $a, b \in D$, $a < b$, such that $F(b) - F(a) > 1 - \varepsilon$. Moreover, since $\lim_{n\to\infty} F_n(a) = F(a)$ and $\lim_{n\to\infty} F_n(b) = F(b)$, there exists some N_1 such that we have

$$F_n(b) - F_n(a) > 1 - 2\varepsilon \qquad \text{for all } n \ge N_1.$$

Let $f\colon \mathbb{R} \to \mathbb{R}$ be bounded and continuous. On the compact interval $[a,b] \subset \mathbb{R}$, the function f is uniformly continuous, so for some $\delta > 0$ we have $|f(x) - f(y)| < \varepsilon$ whenever $|x - y| < \delta$ and $x, y \in [a,b]$. Now choose points $a = c_0 < c_1 < \cdots < c_{k-1} < c_k = b$ such that $c_j - c_{j-1} < \delta$ and $c_j \in D$ for all $j = 1, \ldots, k$. Then for any $j = 1, \ldots, k$, we get

$$\big|f(x) - f(c_j)\big| < \varepsilon \qquad \text{for} \qquad x \in [c_{j-1}, c_j].$$

Define the simple function $h\colon \mathbb{R} \to \mathbb{R}$ by

$$h(x) = \sum_{j=1}^{k} f(c_j)\, \mathbb{I}_{(c_{j-1},c_j]}(x)$$

The above estimate shows that $|f(x) - h(x)| < \varepsilon$ for all $x \in [a,b]$. By boundedness of f, there exists a constant $K > 0$ such that $|f(x)| \le K$ for all $x \in \mathbb{R}$. Since h vanishes outside $(a,b]$, the triangle inequality for integral with respect to ν_n gives

$$\Big|\int_{\mathbb{R}} f\,\mathrm{d}\nu_n - \int_{\mathbb{R}} h\,\mathrm{d}\nu_n\Big| \,\le\, \underbrace{\int_{(a,b]} |f-h|\,\mathrm{d}\nu_n}_{\le\varepsilon} + \underbrace{\int_{\mathbb{R}\setminus(a,b]} |f|\,\mathrm{d}\nu_n}_{\le K\,\nu_n\left[\mathbb{R}\setminus(a,b]\right]}.$$

[1] In fact, the proof of this implication only relies on having pointwise convergence of the cumulative distribution functions in some dense set.

When $n \geq N_1$, we have $\nu_n\big[\mathbb{R} \setminus (a,b]\big] = 1 - \nu_n\big[(a,b]\big] = 1 - (F_n(b) - F_n(a)) < 2\varepsilon$, and thus the triangle inequality implies

$$\Big| \int_{\mathbb{R}} f \,\mathrm{d}\nu_n - \int_{\mathbb{R}} h \,\mathrm{d}\nu_n \Big| \leq \varepsilon + K\, 2\varepsilon = (1 + 2K)\,\varepsilon.$$

Similarly, integrating now with respect to ν instead, one shows that

$$\Big| \int_{\mathbb{R}} f \,\mathrm{d}\nu - \int_{\mathbb{R}} h \,\mathrm{d}\nu \Big| \leq (1 + K)\,\varepsilon.$$

It remains to consider the integrals of the function h with respect to both ν_n and ν. These integrals are expressible in terms of the cumulative distribution functions,

$$\begin{aligned} \int_{\mathbb{R}} h \,\mathrm{d}\nu_n &= \sum_{j=1}^{k} f(c_j)\, \nu_n\big[(c_{j-1}, c_j]\big] \\ &= \sum_{j=1}^{k} f(c_j)\, \big(F_n(c_j) - F_n(c_{j-1})\big) \end{aligned}$$

and similarly

$$\int_{\mathbb{R}} h \,\mathrm{d}\nu = \sum_{j=1}^{k} f(c_j)\, \big(F(c_j) - F(c_{j-1})\big).$$

The difference of the integrals of h with respect to these two can therefore be estimated as

$$\begin{aligned} \Big| \int_{\mathbb{R}} h \,\mathrm{d}\nu - \int_{\mathbb{R}} h \,\mathrm{d}\nu_n \Big| &= \Big| \sum_{j=1}^{k} f(c_j)\, \big(F(c_j) - F_n(c_j) - F(c_{j-1}) + F_n(c_{j-1})\big) \Big| \\ &\leq \sum_{j=1}^{k} |f(c_j)|\, \Big(\big|F(c_j) - F_n(c_j)\big| + \big|F(c_{j-1}) + F_n(c_{j-1})\big| \Big) \\ &\leq 2kK \max_{j=1,\dots,k} \big|F(c_j) - F_n(c_j)\big|. \end{aligned}$$

By our assumption (ii), we have $\lim_{n\to\infty} F_n(c_j) = F(c_j)$ for each $j = 1, \dots, k$, so there exists N_2 such that for $n \geq N_2$ we have $\max_{j=1,\dots,k} |F(c_j) - F_n(c_j)| < \frac{\varepsilon}{k}$, and thus

$$\Big| \int_{\mathbb{R}} h \,\mathrm{d}\nu - \int_{\mathbb{R}} h \,\mathrm{d}\nu_n \Big| \leq 2K\varepsilon.$$

Combining the estimates we have obtained, for $n \geq \max(N_1, N_2)$, we have

$$\begin{aligned} &\Big| \int_{\mathbb{R}} f \,\mathrm{d}\nu - \int_{\mathbb{R}} f \,\mathrm{d}\nu_n \Big| \\ &\leq \underbrace{\Big| \int_{\mathbb{R}} f \,\mathrm{d}\nu - \int_{\mathbb{R}} h \,\mathrm{d}\nu \Big|}_{\leq (1+K)\varepsilon} + \underbrace{\Big| \int_{\mathbb{R}} h \,\mathrm{d}\nu - \int_{\mathbb{R}} h \,\mathrm{d}\nu_n \Big|}_{\leq 2K\varepsilon} + \underbrace{\Big| \int_{\mathbb{R}} h \,\mathrm{d}\nu_n - \int_{\mathbb{R}} f \,\mathrm{d}\nu_n \Big|}_{\leq (1+2K)\varepsilon} \\ &\leq (2 + 5K)\varepsilon. \end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, this shows that $\int f \,\mathrm{d}\nu_n \to \int f \,\mathrm{d}\nu$ as $n \to \infty$, so we have established (i).

proof of "(i) ⇒ (iii)*":* Assume that we have $\mathsf{E}[f(X_n)] \to \mathsf{E}[f(X)]$ as $n \to \infty$ for all bounded continuous functions $f\colon \mathbb{R} \to \mathbb{R}$. In particular, for any fixed $\theta \in \mathbb{R}$ the functions $f(x) = \cos(\theta x)$ and $f(x) = \sin(\theta x)$ are bounded and continuous, so we get

$$\begin{aligned} \varphi_n(\theta) = \mathsf{E}\Big[\exp(\mathrm{i}\theta X_n) \Big] &= \mathsf{E}\big[\cos(\theta X_n) \big] + \mathrm{i}\, \mathsf{E}\big[\sin(\theta X_n) \big] \\ &\xrightarrow[n\to\infty]{} \mathsf{E}\big[\cos(\theta X) \big] + \mathrm{i}\, \mathsf{E}\big[\sin(\theta X) \big] = \varphi(\theta). \end{aligned}$$

proof of "(iii) ⇒ (ii)*":* Assume (iii), i.e., for all $\theta \in \mathbb{R}$ we have $\varphi_n(\theta) \to \varphi(\theta)$ as $n \to \infty$. Let us denote by $D \subset \mathbb{R}$ the set of continuity points of F. Our proof of (ii) consists of proving two parts:

(1) *Existence of subsequential limits:* Given any subsequence indexed by $(n_k)_{k\in\mathbb{N}}$, we can find a further subsequence indexed by $(n_{k_\ell})_{\ell\in\mathbb{N}}$ and a cumulative distribution function $\widetilde{F}$ such that $F_{n_{k_\ell}}(x) \to \widetilde{F}(x)$ as $\ell \to \infty$ at all continuity points x of $\widetilde{F}$.

(2) *Uniqueness of subsequential limits:* If a cumulative distribution function $\widetilde{F}$ is the limit of F_{n_k} along some subsequence, pointwise in the set of continuity points of $\widetilde{F}$, then we must have $\widetilde{F} = F$.

From these two together we derive that $F_n(x) \to F(x)$ for all $x \in D$. Indeed, suppose by contradiction that there exists $x \in D$ such that $F_n(x)$ does not tend to $F(x)$. Then there exists some $\varepsilon > 0$ and a subsequence $(n_k)_{k\in\mathbb{N}}$ such that $|F_{n_k}(x) - F(x)| \geq \varepsilon$ for all $k \in \mathbb{N}$. Applying (1) to this subsequence, we find a further subsequence indexed by $(n_{k_\ell})_{\ell\in\mathbb{N}}$ and $\widetilde{F}$ such that $F_{n_{k_\ell}} \to \widetilde{F}$, pointwise in the set of continuity points of $\widetilde{F}$. Then by (2) we should have $\widetilde{F} = F$, but this is a contradiction, since $|F_{n_{k_\ell}}(x) - F(x)| \geq \varepsilon$ holds in the subsequence. This shows that (1) and (2) indeed imply (ii).

Proving uniqueness of subsequential limits (2) is straightforward using the already established implications "(ii) ⇒ (i)" and "(i) ⇒ (iii)". Indeed, suppose that F_{n_k} converges to $\widetilde{F}$, pointwise in the set of continuity points of $\widetilde{F}$. By the already established "(ii) ⇒ (i)", this implies the convergence $\int f \,\mathrm{d}\nu_{n_k} \to \int f \,\mathrm{d}\widetilde{\nu}$ for all bounded continuous f, where the probability measure $\widetilde{\nu}$ has $\widetilde{F}$ as its cumulative distribution function. This, in turn, by the already established "(ii) ⇒ (i)", implies that $\varphi_{n_k}(\theta) \to \widetilde{\varphi}(\theta)$ for all $\theta \in \mathbb{R}$, where $\widetilde{\varphi}$ is the characteristic function of $\widetilde{\nu}$. But since the entire sequence $\varphi_1, \varphi_2, \ldots$ of characteristic functions has limit φ, the limit of the subsequence must be the same, $\widetilde{\varphi} = \varphi$. Therefore we also get $\widetilde{\nu} = \nu$ (by Lévy's inversion theorem) and $\widetilde{F} = F$, and the uniqueness part (2) follows.

It remains to prove the existence of subsequential limits (1). The basic idea is to use Cantor's diagonal extraction. In order to keep the notation simpler, let us show how to extract a convergent subsequence from the entire sequence $(F_n)_{n\in\mathbb{N}}$ — the same argument works for extracting a convergent subsequence from any given subsequence. The set $\mathbb{Q}$ of rational numbers is countable, so let $\mathbb{Q} = \{q^{(1)}, q^{(2)}, \ldots\}$ be an enumeration of it. Observe that for any rational point $q^{(j)} \in \mathbb{Q}$, the sequence $\big(F_n(q^{(j)})\big)_{n\in\mathbb{N}}$ of values at this point is a bounded sequence: $0 \leq F_n(q^{(j)}) \leq 1$ for all $n \in \mathbb{N}$. In particular, considering $q^{(1)} \in \mathbb{Q}$ first, we can find a subsequence indexed by $(n_k^{(1)})_{k\in\mathbb{N}}$ such that

$$\lim_{k\to\infty} F_{n_k^{(1)}}(q^{(1)})$$

exists. Then considering $q^{(2)} \in \mathbb{Q}$, we can find a subsequence of this already chosen subsequence, indexed now by $(n_k^{(2)})_{k\in\mathbb{N}}$ such that

$$\lim_{k\to\infty} F_{n_k^{(2)}}(q^{(2)})$$

exists, and since this is a subsequence of the previous one, also

$$\lim_{k\to\infty} F_{n_k^{(2)}}(q^{(1)}) = \lim_{k\to\infty} F_{n_k^{(1)}}(q^{(1)}).$$

Continuing inductively, we find subsequences $(n_k^{(\ell)})_{k\in\mathbb{N}}$, $\ell = 1, 2, 3, \ldots$, such that

$$\text{the limit} \qquad \lim_{k\to\infty} F_{n_k^{(\ell)}}(q^{(j)}) \qquad \text{exists for all } j \leq \ell.$$

Then if we consider the subsequence of diagonally selected indices $n_\ell := n_\ell^{(\ell)}$, we get that

$$\text{the limit} \qquad \lim_{\ell\to\infty} F_{n_\ell}(q^{(j)}) \qquad \text{exists for all } j \in \mathbb{N},$$

since apart from finitely many first members, the index sequence $(n_\ell)_{\ell\in\mathbb{N}}$ is a subsequence of $(n_k^{(j)})_{k\in\mathbb{N}}$. The limits

$$G(q^{(j)}) := \lim_{\ell\to\infty} F_{n_\ell}(q^{(j)})$$

define a function on the set of rational numbers

$$G\colon \mathbb{Q} \to [0, 1].$$

As the limit of the increasing functions F_{n_ℓ} (restricted to $\mathbb{Q}$), this function is also increasing:

$$G(q) \le G(q') \qquad \text{when } q, q' \in \mathbb{Q} \text{ and } q < q'.$$

We now claim that $G(q) \downarrow 0$ as $q \downarrow -\infty$ and $G(q) \uparrow 1$ as $q \uparrow -\infty$ — this is the place where the assumption (iii) is used. Let $\varepsilon > 0$. Then, since $\varphi(0) = 1$ and $\theta \mapsto \varphi(\theta)$ is continuous at $\theta = 0$ (by Proposition XII.6), there exists $\delta > 0$ such that

$$\big|1 - \varphi(\theta)\big| < \varepsilon \qquad \text{when} \qquad |\theta| < \delta.$$

In particular we have

$$\frac{1}{\delta} \int_{-\delta}^{\delta} \big(1 - \varphi(\theta)\big)\, \mathrm{d}\theta \;<\; \varepsilon.$$

By assumption (iii) we have $\varphi_n(\theta) \to \varphi(\theta)$ for all θ. Since we have $\big|1-\varphi_n(\theta)\big| \le 1 + |\varphi_n(\theta)| \le 2$, the Bounded convergence theorem (Corollary VII.21) implies

$$\frac{1}{\delta} \int_{-\delta}^{\delta} \big(1 - \varphi_n(\theta)\big)\, \mathrm{d}\theta \underset{n\to\infty}{\longrightarrow} \frac{1}{\delta} \int_{-\delta}^{\delta} \big(1 - \varphi(\theta)\big)\, \mathrm{d}\theta,$$

and we can conclude that there exists an N_ε such that for $n \ge N_\varepsilon$ we have

$$\frac{1}{\delta} \int_{-\delta}^{\delta} \big(1 - \varphi_n(\theta)\big)\, \mathrm{d}\theta \;<\; 2\varepsilon.$$

In view of Lemma F.3 this gives for $n \ge N_\varepsilon$

$$\nu_n\big[\mathbb{R} \setminus [-R, R]\big] < 2\varepsilon,$$

where $R = 2/\delta$. For the corresponding cumulative distribution functions this implies

$$\begin{aligned} F_n(x) &< 2\varepsilon \qquad && \text{for } x < -R \\ F_n(x) &> 1 - 2\varepsilon \qquad && \text{for } x > -R. \end{aligned}$$

The inequalities inherited by the subsequential limit $G(q) = \lim_{\ell\to\infty} F_{n_\ell}(q)$ are then

$$\begin{aligned} G(q) &\le 2\varepsilon \qquad && \text{for } q < -R \\ G(q) &\ge 1 - 2\varepsilon \qquad && \text{for } q > -R. \end{aligned}$$

Since $\varepsilon > 0$ was arbitrary, this shows that $G(q) \downarrow 0$ as $q \downarrow -\infty$ and $G(q) \uparrow 1$ as $q \uparrow -\infty$.

The function G is not yet the cumulative distribution function we are looking for: it is only defined on the rational numbers, and it is not necessarily right-continuous. But we can use it to define

$$\widetilde{F}(x) = \inf_{\substack{q > x \\ q \in \mathbb{Q}}} G(q)$$

(this in fact defines the smallest right continuous function above G). From the definition it is clear that if $q', q'' \in \mathbb{Q}$ and $q' < x < q''$, then

$$G(q') \le \widetilde{F}(x) \le G(q'').$$

We leave it to the reader to check that this $\widetilde{F}$ is right continuous, increasing, and has limits 0 and 1 at $-\infty$ and $+\infty$, respectively. Therefore $\widetilde{F}$ is a cumulative distribution function of some probability measure $\widetilde{\nu}$ on $\mathbb{R}$.

The only remaining claim is that the subsequence $(F_{n_\ell})_{\ell\in\mathbb{N}}$ converges to $\widetilde{F}$ pointwise in the set $\widetilde{D} \subset \mathbb{R}$ of continuity points of $\widetilde{F}$. Consider $x \in \widetilde{D}$ and let $\varepsilon > 0$. Then by continuity of $\widetilde{F}$ at x we have for some $\delta > 0$

$$\widetilde{F}(x - \delta) > \widetilde{F}(x) - \varepsilon \qquad \text{and} \qquad \widetilde{F}(x + \delta) < \widetilde{F}(x) + \varepsilon.$$

Choose $q' \in \mathbb{Q} \cap (x - \delta, x)$. Then by the definition $G(q') = \lim_{\ell\to\infty} F_{n_\ell}(q')$ there exists some L_ε such that for all $\ell \ge L_\varepsilon$ we have

$$F_{n_\ell}(q') > G(q') - \varepsilon,$$

which yields

$$F_{n_\ell}(x) \ge F_{n_\ell}(q') > G(q') - \varepsilon \ge \widetilde{F}(x - \delta) - \varepsilon > \widetilde{F}(x) - 2\varepsilon.$$

Since $\varepsilon > 0$ was arbitrary, we get

$$\liminf_{\ell} F_{n_\ell}(x) \geq \widetilde{F}(x).$$

By choosing $q'' \in \mathbb{Q} \cap (x, x+\delta)$, one can similarly argue that

$$\limsup_{\ell} F_{n_\ell}(x) \leq \widetilde{F}(x),$$

and these two combined imply

$$\lim_{\ell \to \infty} F_{n_\ell}(x) = \widetilde{F}(x).$$

This establishes the existence of subsequential limits (1), and finishes the proof. □

Index

References

Bibliography

[Dur10] Richard Durrett. *Probability: Theory and Examples*. Cambridge University Press, 4th edition, 2010.

[JP04] Jean Jacod and Philip Protter. *Probability Essentials*. Cambridge Univ. Press, 2nd edition, 2004.

[Kin16] Juha Kinnunen. *Measure and integral*. Aalto University, 2016. `https://math.aalto.fi/~jkkinnun/files/measure_and_integral.pdf`.

[Wil91] David Williams. *Probability with Martingales*. Cambridge Univ. Press, 1991.