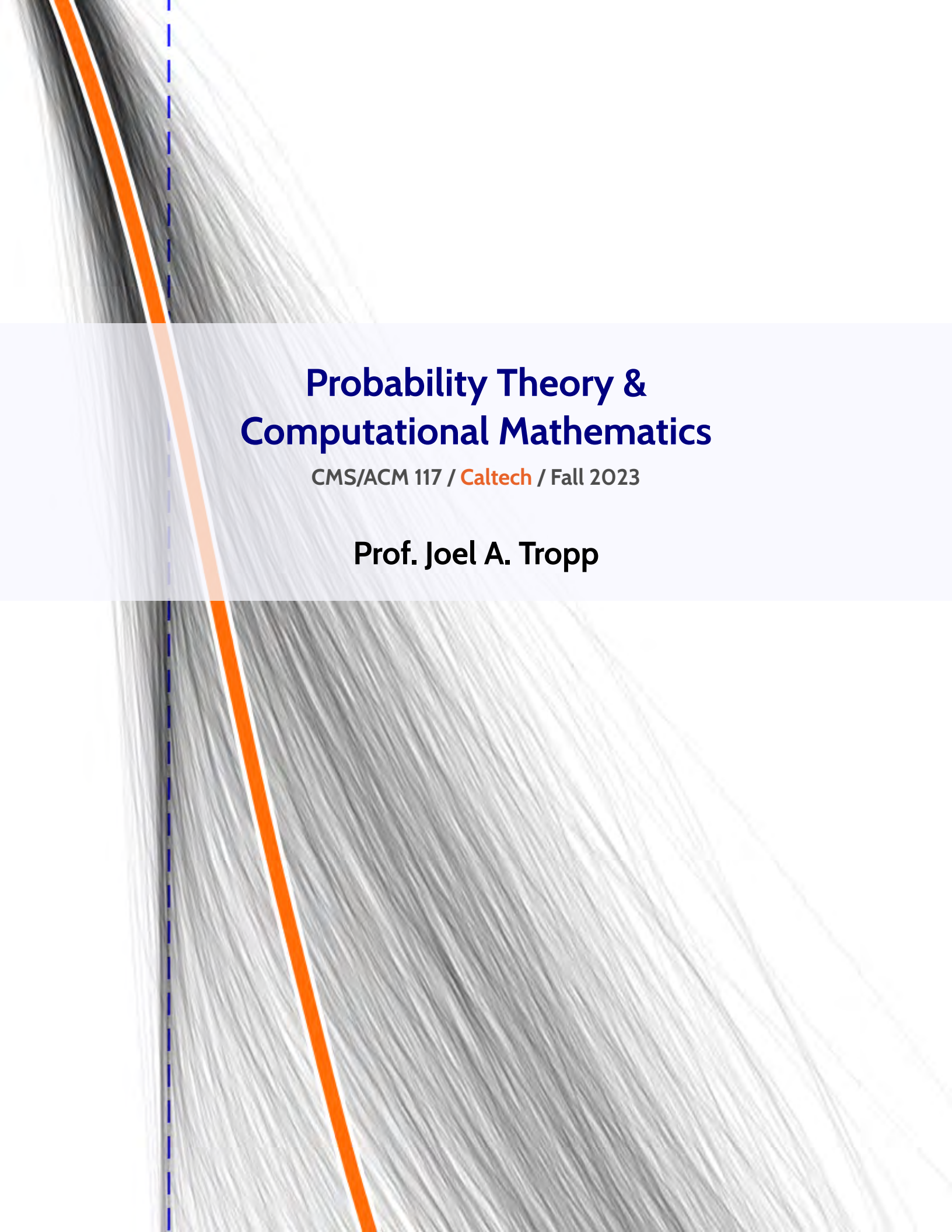


Probability Theory & Computational Mathematics

CMS/ACM 117 / Caltech / Fall 2023

Prof. Joel A. Tropp

Typeset on December 14, 2023

Cite as:
Joel A. Tropp, *CMS/ACM 117: Probability Theory & Computational Mathematics*, Caltech CMS Lecture Notes 2023-01, Pasadena, December 2023.

Available from https://doi.org/10.7907/q75sz-e1e79.

Cover image: Sample paths of a randomized block Krylov method for estimating the largest eigenvalue of a symmetric matrix.

Contents

Preface

"*Alea iacta est.* The die has been cast."

—Julius Caesar, after crossing the Rubicon, 49 BCE

"The reader of any book is entitled to ask why it had to be written at all and, if the book absolutely had to exist, why it couldn't have been shorter."

—Walter Russell Mead

CMS/ACM 117 is a first-year graduate course on probability theory and stochastic processes for students in computing and mathematical sciences. It is not intended to be a first course in probability, and our focus will be on developing theoretical foundations, rather than providing a toolkit for calculation. Nevertheless, we will touch on a few substantive applications of the theory to demonstrate its implications for practical problems.

Course overview

Modern probability is expressed in the language of measure theory. Although measure theory has a bad reputation, it can be engaging and accessible if we do not venture too deep within the labyrinth of details. The probabilistic gloss also breathes life into the subject.

The course notes begin with a rigorous, but unfussy, overview of measure and integration theory. Our goal is to develop good geometric intuitions for the concepts of measure and integral. When studying probability, one must learn some abstract measure theory, so we introduce these ideas early on so that the student gains the confidence that only comes with practice. From the appendices, a keen reader can also learn the foundational results on existence and uniqueness of measures, including the construction of the Lebesgue measure.

In Fall 2023, the measure theory background was treated in an optional boot camp outside of class. In the future, it will return to the main sequence of lectures.

This development sets the stage for a quick treatment of Kolmogorov's axiomatic definition of probability. We introduce new probabilistic language that adds a vivid interpretation to the measure-theoretic constructs. The key new idea in probability theory is that algebras of events encode our knowledge about the world; the concept of independence is best understood through this lens.

Next, we introduce our first stochastic process: a sum of independent random variables. We develop the tools for understanding the finite-time and asymptotic behavior of independent sums. These ideas include concentration inequalities, large-deviation principles, laws of large numbers, the central limit theorem, and more. These results require us to explore what it means for two probability distributions to be similar to each other.

Afterward, we turn our attention to the fundamental concept of conditional expectation. How does our current knowledge about the world affect our predictions for the future? We present these ideas in an intuitive way using the concept of best

approximation of a random variable. Conditioning also allows us to talk about the relative density of one distribution with respect to another, a major ingredient in Bayesian statistics and other fields.

To gain some experience with conditioning, we study discrete-time martingales. These are random processes where the future depends only on the past. We develop the basic theory of maximal inequalities and martingale convergence. These tools have a wide swath of applications, including techniques for prediction, filtering, adaptive testing, online learning, and stochastic optimization.

Along the way, we will explore applications of probability theory in computational statistics, computational mathematics, computer science, electrical engineering, and control theory.

Previous iterations of this course included the development of other sequential random processes, namely Markov chains. This term, we have removed this material to make the course more manageable.

These notes

The Fall 2023 edition of CMS/ACM 117 is the fifth instantiation of this class. Hopefully, this version of the course is approaching asymptotic stability. Future upgrades to the notes may supersede the current version.

These lecture notes diverge somewhat from CMS/ACM 117 as it was taught in Fall 2023, because they contain additional material that was not covered in the classroom. At present, they are intended as a reference for the students who have taken the class. The notes have been prepared with some care. Nevertheless, they are not fully polished, and they may still contain repetitions, omissions, errors, and inconsistencies. In particular, they lack full scholarly citations. *Caveat lector*!

Activities

The lecture notes are full of exercises and problems. Exercises contain material that is essential for your understanding, and they are usually quite easy. Problems are intended to give you more practice with the concepts or to expand your understanding; they may be more difficult or lengthier than exercises. Applications are designed to show how tools from probability theory are used in computational mathematics; they may involve coding and simulation.

More challenging activities may be marked with stars; the number of stars gives a rough indication of the difficulty.

Starred sections and asides

This course is designed for students arriving from a wide range of academic experiences. Students with more applied backgrounds may not have seen much of the material in this course, while students from more theoretical backgrounds may have seen the majority. Students who are just entering this subject may want to focus on the unstarred material, while students with more exposure should take the time the understand the starred sections and appendices. The hope is that everyone will learn something that is new and interesting for them.

For those with more experience, consider this: "Never underestimate the joy people derive from hearing something they already know."
—Enrico Fermi

Similarly, asides are reserved for technical comments. These sidebars address questions that may occur to you as you read or they may point toward more advanced parts of probability theory. This material falls outside the scope of the class, so you may skip it with impunity.

Prerequisites

The prerequisites for this course are differential and integral calculus (e.g., Caltech Math 1ac), intermediate linear algebra (e.g., Math 1b and ACM 104), and applied

probability (e.g., Math 3 and ACM 116). Exposure to linear analysis (e.g., ACM 107a) and functional analysis (e.g., ACM 107b) is valuable but not necessary.

This course demands experience with basic facts about set theory, real numbers, functions, point–set topology, sequences, series, convergence, continuity, derivatives, Riemann integrals, metric spaces, and linear spaces. At Caltech, this material is covered in the undergraduate class Math 108a. You will also benefit from some exposure to measure theory, which is covered in Math 108b. Some good textbooks include

- [Fol99] Folland, *Real Analysis*, 2nd ed., Wiley, 1999.
- [Roy88] Royden, *Real Analysis*, 3rd ed., Macmillan, 1988.
- [Rud76] Rudin, *Principles of mathematical analysis*, 3rd ed., Wiley, 1976.
- [Tao16] Tao, *Analysis I*, 3rd ed., Springer, 2014.

It may be possible to brush up on this background as the course proceeds.

Supplemental textbooks

There is no required textbook for the course. Some relatively recent books that cover related material include

- [GS01] Grimmett & Stirzaker, *Probability and random processes*, Oxford, 2001.
- [Wil91] Williams, *Probability with martingales*, Cambridge, 1991.
- [Dur19] Durrett, *Probability theory and examples*, 5th ed., Cambridge, 2019.
- [Pol02] Pollard, *A user's guide to measure theoretic probability*, Cambridge, 2002.
- [Bil12] Billingsley, *Probability and measure*, 3rd ed., Wiley, 2012.
- [Dud02] Dudley, *Real analysis and probability*, 2nd ed., Cambridge, 2002.
- [Kal02] Kallenberg, *Foundations of modern probability*, 2nd ed., Springer, 2002.

These books are arranged in rough order of difficulty. Grimmett & Stirzaker is a rigorous book on applied probability, focusing on probability models and problem solving. Williams is a charming, short book on martingales, but it is also idiosyncratic and telegraphic. Durrett's book seems to be the current standard in graduate mathematics programs. Pollard's book contains a personal perspective on probability, with loads of intuition. Billingsley's book is more technical. Dudley's book gives a thorough treatment of the foundations of set theory, real analysis, and probability. Last, Kallenberg contains a comprehensive, rigorous overview of topics in modern probability.

Acknowledgements

These notes are adapted from the lecture notes for CMS 117, Fall 2019, as transcribed in the classroom. The 2019 notes were prepared by Jeremy Bernstein with contributions from Richard Kueng. Many students have helped to find mistakes in the text. All remaining errors are the fault of the instructor.

Joel A. Tropp
Steele Family Professor of Applied & Computational Mathematics
California Institute of Technology

jtropp@caltech.edu
https://tropp.caltech.edu/

Pasadena, California
December 2023

Notation and Definitions

> "You sound like a physicist," she said.
> "There's no reason to be insulting."
>
> —*Dr. No*, Percival Everett, 2022

The notation in this course is standard in probability theory and related fields. This section contains the main definitions and conventions. Other notation will be introduced as needed.

Set theory

The Pascal notation, := or =:, generates a definition. Sets without any particular internal structure are denoted with sans serif capitals: $\mathsf{A}, \mathsf{B}, \mathsf{E}$. Collections of sets are written in a calligraphic font: $\mathscr{A}, \mathscr{B}, \mathscr{F}$. The power set (that is, the collection containing all subsets) of a set E is written as $\mathscr{P}(\mathsf{E})$.

The symbol $\emptyset$ is reserved for the empty set. We use braces to denote a set. The character $\in$ (or, rarely, $\ni$) is the member-of relation. The set-builder notation

$$\{x \in \mathsf{A} : P(x)\}$$

carves out the (unique) set of elements that belong to a set A and that satisfy the predicate P. The operator # returns the cardinality of a set. Basic set operations include union ($\cup$), intersection ($\cap$), symmetric difference ($\triangle$), set difference ($\setminus$), and the complement (^{c}) with respect to a fixed set. We often write $\dot\cup$ for the union of *disjoint* sets.

The natural numbers $\mathbb{N} := \{1, 2, 3, \dots\}$. Ordered tuples and sequences are written with parentheses, e.g.,

$$(a_1, a_2, a_3, \dots, a_n) \quad \text{or} \quad (a_1, a_2, a_3, \dots)$$

Alternative notations include things like $(a_i : i \in \mathbb{N})$ or $(a_i)_{i\in\mathbb{N}}$ or simply (a_i).

The relations $\subseteq$ and $\supseteq$ indicate set containment. For a sequence of increasing (resp., decreasing) sets, we may use arrows to denote the union (resp., intersection):

$$\mathsf{A}_i \uparrow \bigcup_{i=1}^{\infty} \mathsf{A}_i \qquad \text{when } \mathsf{A}_1 \subseteq \mathsf{A}_2 \subseteq \mathsf{A}_3 \subseteq \cdots;$$
$$\mathsf{A}_i \downarrow \bigcap_{i=1}^{\infty} \mathsf{A}_i \qquad \text{when } \mathsf{A}_1 \supseteq \mathsf{A}_2 \supseteq \mathsf{A}_3 \supseteq \cdots.$$

We typically use italic lowercase letters like f, g, h for functions. To introduce a function f with domain A and codomain B, we write

$$f : \mathsf{A} \to \mathsf{B} \quad \text{where} \quad f : a \mapsto f(a).$$

For a subset $\mathsf{E} \subseteq \mathsf{A}$, the set $f(\mathsf{E})$ is the image of E. For a subset $\mathsf{F} \subseteq \mathsf{B}$, the set $f^{-1}(\mathsf{F})$ is the preimage of F. The circle $\circ$ composes functions.

The set $\mathsf{A} \times \mathsf{B}$ is the Cartesian product of sets A and B. The symbol A^n denotes the n-fold product of A with itself: $\mathsf{A}^n = \mathsf{A} \times \cdots \times \mathsf{A}$, with A repeated n times. More generally, A^{I} is a repeated product of A indexed by a set I. For each $i \in \mathsf{I}$, the function $\pi_i : \mathsf{A}^{\mathsf{I}} \to \mathsf{A}$ refers to the ith coordinate projection: $\pi_i((a_j : j \in \mathsf{I})) = a_i$.

Real analysis

We mainly work in the field $\mathbb{R}$ of real numbers, equipped with the absolute value $|\cdot|$. The extended real numbers $\overline{\mathbb{R}} := \mathbb{R} \cup \{\pm\infty\}$ are defined with the usual rules of arithmetic and order. In particular, we instate the conventions that $0/0 = 0$ and $0 \cdot \pm\infty = 0$. Expressions involving competing infinities ($\infty - \infty$) are undefined, and we do not allow division by infinity. We use the standard (American) notation for open and closed intervals; e.g.,

$$(a, b) := \{x \in \overline{\mathbb{R}} : a < x < b\} \quad \text{and} \quad [a, b] := \{x \in \overline{\mathbb{R}} : a \le x \le b\}.$$

Occasionally, we will visit the rational field $\mathbb{Q}$ or the complex field $\mathbb{C}$. The imaginary unit, i, is written in an upright font. Euler's constant is denoted as e; by default every logarithm has base e.

We use modern conventions for words describing order; these may be slightly different from what you are used to. In this course, we enforce the definition that *positive* means ≥ 0 and *negative* means ≤ 0. For example, the positive integers compose the set $\mathbb{Z}_+ := \{0, 1, 2, 3, \dots\}$ and the positive reals compose the set $\mathbb{R}_+ := \{x \in \mathbb{R} : x \ge 0\}$. When required, we may deploy the phrase *strictly positive* to mean > 0 and *strictly negative* to mean < 0. For instance, the set $\mathbb{R}_{++} := \{x \in \mathbb{R} : x > 0\}$ contains the strictly positive real numbers. Similarly, *increasing* means "never going down" and *decreasing* means "never going up."

Warning: Positive means ≥ 0!

We often write maximum ($\vee$) and minimum ($\wedge$) of two numbers using infix notation. Given a nonempty set $\mathsf{A} \subseteq \overline{\mathbb{R}}$ of extended real numbers, define

$$\sup(\mathsf{A}) := \text{least upper bound on } \mathsf{A} \text{ in } \overline{\mathbb{R}};$$
$$\inf(\mathsf{A}) := \text{greatest lower bound on } \mathsf{A} \text{ in } \overline{\mathbb{R}}.$$

Warning: The infix maximum and minimum have the opposite appearance to what you might expect!

The supremum and infimum always exist, but they may be infinite.

Each of the following expressions means that the sequence $(x_i : i \in \mathbb{N})$ has limiting value x:

$$\lim_{i\to\infty} x_i = x \quad \text{or} \qquad x_i \to x \text{ as } i \to \infty \quad \text{or} \quad x_i \to x.$$

We may use vertical arrows to indicate that a real-valued sequence increases or decreases to its limiting value: $x_i \uparrow x$ or $x_i \downarrow x$. Recall that a monotone sequence always has a limit in the extended reals. The limit superior and limit inferior are defined as

A monotone sequence is either increasing or decreasing.

$$\limsup_{i\to\infty} x_i := \lim_{i\to\infty} \sup_{n\ge i} x_n = \inf_{i\in\mathbb{N}} \sup_{n\ge i} x_n;$$
$$\liminf_{i\to\infty} x_i := \lim_{i\to\infty} \inf_{n\ge i} x_n = \sup_{i\in\mathbb{N}} \inf_{n\ge i} x_n.$$

These limits always exist, but they may be infinite.

For a pair of functions with a common domain, we understand relations and other operations in the pointwise sense. For example, $f = g$ means $f(x) = g(x)$ for all x in their domain. When the functions are real-valued, we may write $f \wedge g$ for the function $x \mapsto (f(x) \wedge g(x))$ with the same domain or $f\,g$ for the function $x \mapsto (f(x)g(x))$ with the same domain. Similarly, $f \le g$ means that $f(x) \le g(x)$ for all x in the domain. We often use a compact set-builder notation, such as $\{f \le g\} := \{x : f(x) \le g(x)\}$, for brevity.

Much the same way, for a sequence $(f_j : j \in \mathbb{N})$ of functions with a common domain, the expression $f_j \to f$ means pointwise convergence of the f_j to a limiting function f. For real-valued functions, we may use vertical arrows to denote pointwise monotone convergence. For example,

$$f_j \uparrow f \quad \text{if and only if} \quad f_{j+1} \ge f_j \quad \text{and} \quad f_j \to f.$$

The increasing limit always exists pointwise, but it may take (some) infinite values. Likewise, $f_j \downarrow f$ refers to decreasing monotone convergence.

Measure and integral

Given a collection $\mathscr{C}$ of subsets of X, the symbol $\sigma(\mathscr{C})$ denotes the smallest σ-algebra on X that contains all the sets in $\mathscr{C}$.

For a topological space X, the family $\mathscr{B}(\mathsf{X})$ contains the Borel sets in X. In particular, $\mathscr{B}(\mathbb{R}^n)$ contains the Borel sets in $\mathbb{R}^n$.

The family of Borel sets is the sigma-algebra generate by all open sets.

To denote measures, we typically use Greek letters in the middle of the alphabet (μ, ν). The letter λ is reserved for the Lebesgue measure on $\mathbb{R}$, and λ^n is the Lebesgue measure on $\mathbb{R}^n$. We write δ_x for the Dirac mass of intensity one, concentrated at a point x.

Given a subset $\mathsf{A} \subseteq \mathsf{X}$ of a domain, we define the real-valued 0–1 indicator function

$$\mathbb{1}_{\mathsf{A}} : \mathsf{X} \to \mathbb{R} \quad \text{where} \quad \mathbb{1}_{\mathsf{A}}(x) := \begin{cases} 1, & x \in \mathsf{A}; \\ 0, & x \notin \mathsf{A}. \end{cases}$$

Indicator functions and sets are in one-to-one correspondence, so we can switch between them at will.

We have many, many equivalent notations for integrals. For a measure μ (on a σ-algebra) on X and a function $f : \mathsf{X} \to \overline{\mathbb{R}}$, we may write

$$\mu(f) := \int_{\mathsf{X}} f(x)\, \mu(\mathrm{d}x) =: \int_{\mathsf{X}} f(x)\, \mathrm{d}\mu(x) =: \int_{\mathsf{X}} f\, \mathrm{d}\mu$$

It is common that we omit the domain X from these notations. To integrate over a subset $\mathsf{A} \subseteq \mathsf{X}$, we may write

$$\mu(f; \mathsf{A}) := \int_{\mathsf{A}} f(x)\, \mu(\mathrm{d}x) := \int_{\mathsf{X}} \mathbb{1}_{\mathsf{A}}(x)\, f(x)\, \mu(\mathrm{d}x).$$

When integrating over an interval of the real line, we sometimes use the classic notation where the limits of the interval are placed in the integral sign:

$$\int_a^b f\, \mathrm{d}\mu := \int_{(a,b]} f\, \mathrm{d}\mu \quad \text{where } a < b \text{ and } a, b \in \mathbb{R}.$$

Note, however, that the former notation requires care because a general measure μ can have an atom at one of the endpoints. When the measure is the Lebesgue measure λ, the differential is often abbreviated as well: $\mathrm{d}x := \lambda(\mathrm{d}x)$.

We also use the arrow notation to refer to modes of convergence that require measure or integral. In these cases, we will give an explicit qualification to emphasize the type of convergence. For example, we may write $f_j \to f$ λ-almost everywhere or $h_j \to h$ in $\mathsf{L}_1(\lambda)$.

For weak convergence of measures, we use the seismic arrow: $\mu_n \rightsquigarrow \mu$. In parallel, we use the seismic arrow $\rightsquigarrow$ to refer to converge in distribution for random variables $(X_n \rightsquigarrow X)$ and distribution functions $(F_n \rightsquigarrow F)$.

Linear algebra

We usually denote scalars with lowercase Greek letters (α, β). Lowercase boldface italics $(\boldsymbol{u}, \boldsymbol{v})$ refer to vectors. Uppercase boldface italics $(\boldsymbol{A}, \boldsymbol{B})$ are associated with matrices or linear maps. The symbol * denotes the (conjugate) transpose of a vector or matrix. The operator tr returns the trace of a square matrix.

Norms and pseudonorms are denoted with double bars: $\|\cdot\|$. We typically add a subscript to refer to a specific norm, such as the Euclidean norm $\|\cdot\|_{\ell_2}$.

Probability

We write $(\Omega, \mathscr{F}, \mathbb{P})$ for the probability space with sample space Ω, with master σ-algebra $\mathscr{F}$, and with probability measure $\mathbb{P}$ defined on $\mathscr{F}$. The map $\mathbb{P}(\cdot)$ returns the probability of an event in $\mathscr{F}$. The operator $\mathbb{E}[\cdot]$ returns the expectation of a random variable (taking values in a linear space). We only include the brackets when it is necessary for clarity, and we impose the convention that nonlinear functions bind before the expectation. At rare times, we may also use $\mathbb{P}$ to denote expectation with respect to a probability measure $\mathbb{P}$.

Uppercase italic letters (near the end of the Roman alphabet) usually refer to real random variables: S, T, W, X, Y, Z. We often write μ_X for the law of a random variable X, while F_X denotes its (cumulative) distribution function.

We use small capitals for named distributions. For example, UNIFORM or NORMAL. The symbol $\sim$ means "has the distribution."

The sigma-algebra generated by a real random variable X is defined as $\sigma(X) := \{X^{-1}(\mathsf{B}) : \mathsf{B} \in \mathscr{B}(\mathbb{R})\}$. Similar notations are in force for random variables taking values in other measures spaces.

For each real number $p > 0$, the space $\mathsf{L}_p := \mathsf{L}_p(\Omega, \mathscr{F}, \mathbb{P})$ contains all real random variables X whose pth absolute moment is finite:

$$\mathsf{L}_p := \{X : \mathbb{E}\,|X|^p < +\infty\}.$$

The operator $\mathrm{Var}[\cdot]$ returns the variance of a random variable in L_2, while $\mathrm{Cov}(\cdot, \cdot)$ computes the covariance of a pair of random variables in L_2.

The symbol $\perp$ means that two random variables in L_2 are orthogonal: $X \perp Y$ if and only if $\mathbb{E}[XY] = 0$. In contrast, the notation $X \perp\!\!\!\perp Y$ means that X and Y are *independent* random variables.

As usual, we write $\mathbb{E}[X \mid \mathscr{G}]$ for the conditional expectation of the random variable X with respect to the σ-algebra $\mathscr{G}$. Related notations include the conditional expectation with respect to a family of events or a family of random variables. For example, $\mathbb{E}[X \mid \mathsf{A}, \mathsf{B}] := \mathbb{E}[X \mid \sigma(\{\mathsf{A}, \mathsf{B}\})]$ and $\mathbb{E}[X \mid Y, Z] := \mathbb{E}[X \mid \sigma(Y, Z)]$.

0. Probability + CMS

Agenda:

1. What is probability?
2. Probability models
3. Randomized trace estimators
4. Stochastic gradient
5. Markov chains
6. Probability and measure

"Jedenfalls bin ich überzeugt, daß der nicht würfelt."
"I, at any rate, am convinced that [God] does not throw dice."

—Albert Einstein, 1926

"The gods may throw a dice,
Their minds as cold as ice."

—*The Winner Takes it All*, ABBA, 1980

Probability theory is the study of regular patterns that arise from random phenomena. The field of statistics exploits this fact to make inferences about the state of the world. These regularities can also be used to develop efficient algorithms for solving a wide range of computational problems.

In this introductory section, we give a simple example of the patterns that can arise from a simple probability experiment. Then we discuss how probability models can arise and some of the fields where they are in use. Afterward, we present several applications of probability theory in contemporary computational mathematics. Last, we summarize the concepts that are required to make sense of the application examples. This motivation helps us appreciate why we need to understand measure theory to work with probability models. In the first lecture, we enter into our treatment of measure theory.

0.1 What is probability theory?

Even my 6-year-old thought that this would be a surprising outcome.

Say that you flip a fair penny 100 times, and you observe that heads turns up 97 times. Are you surprised? In a word, yes. While the outcome of this experiment is random, you do not expect it to be totally irregular. In fact, you will probably take the result as evidence that the coin is not fair after all.

When we say that the coin is fair, we mean that each sequence HTHHHTHHTH... or HHHHHHHHHH... is equally likely to occur. So no particular sequence is very common. Nevertheless, if we ask summary questions like "How many heads?", then the result is almost deterministic. Indeed, we strongly suspect that, in the long-term, the proportion of heads is close to one-half. We anticipate this pattern based on our experience living in the world. When we observe a pattern that violates our intuition, we may draw inferences about how that outcome arose.

Moving beyond coins, suppose that we perform a probabilistic experiment and record the outcome. From the word *probabilistic*, we understand that repeating the experiment gives an *unpredictable* result each time. Nevertheless, when we look at a large number of experiments, we encounter *predictable* phenomena.

Probability is the study of the predictable patterns that arise from random phenomena. For instance, it is predictable that a fair coin turns up heads about half the time, and we will learn to quantify this statement precisely.

Inversely, *statistics* uses probability theory to infer the state of the world from observed outcomes of probabilistic experiments. For example, if a coin turns up heads 97 times out of 100, we can be confident (but not certain) that the coin is unfair.

Our challenge is to develop the mathematical foundations for probability. Statistics is a primary beneficiary of this effort, but probability also enriches the mathematical sciences, computing, and engineering.

Activity 0.1 (Probability experiments). What are situations where probability arises in your field? It is likely that your research program involves more than flipping coins. ▪

0.2 Probability models

The real world is complex and messy. To help us understand the world and make predictions, we frame simplifying or reductive models. *Probability models* describe phenomena that are not fully predictable. They encode our uncertainty using distributions, rather than worst-case considerations.

"All models are wrong, but some are useful."
—George E. P. Box, 1970s

0.2.1 Sources of randomness

So, how does randomness arise? Why might we want to use a probability model for applications in science and engineering?

Sources of randomness:

1. Uncertain initial conditions
2. Sensitivity to initial conditions
3. Measurement errors
4. Statistical mechanics
5. Quantum mechanics
6. Uncertain mechanisms
7. Sampling from a population
8. Random number generators
9. The whim of Tyche

Since the laws of classical physics are deterministic, one may imagine that a classical phenomenon becomes completely predictable if we know the state of the world. In practice, however, we rarely have complete information about the initial conditions (e.g., the starting position, velocity, and angular momentum of a tossed penny). We can model this uncertainty using probability. Many phenomena (e.g., the weather) exhibit a sensitive dependency on the initial conditions, so it may only be reasonable for us to talk about a distribution of outcomes, some more likely than others. In a similar vein, a system may be so complicated (e.g., the positions of all molecules in a cup of coffee) that we must use probabilistic models to summarize its behavior; this is the insight behind the field of statistical mechanics. Beyond that, measurements are inherently inaccurate, and measurement errors are commonly modeled using probability.

To the best of our understanding, quantum mechanics gives a precise and accurate description of the nanoscale. In the quantum world, probability is an irreducible fact of life: Born's rule tells us that every measurement of a quantum system yields a random outcome. Nevertheless, at a macroscopic scale, the aggregation of many random outcomes can lead to behavior that does not appear random at all. (We have already seen a similar effect in long sequences of coin flips or the bulk behavior of the molecules in a gas.) Indeed, the law of large numbers helps resolve the tension between the randomness of the quantum world and the apparent determinism of the classical world.

Physical laws often admit simple mathematical expressions that have been validated by extensive experiments. In other human endeavors, we may not have a complete understanding of mechanisms or even a collection of empirical laws that govern the phenomena under investigation. This kind of challenge emerges in social sciences (economics, sociology, psychology) or in hard sciences that involve complex systems (cell biology, neuroscience). We can try to model our uncertainty using probability.

Another important source of randomness is sampling from a population. Suppose that we isolate a number of individuals from a larger (perhaps infinite) group. If these

individuals are not chosen in any particular fashion, then it may be reasonable to model them as a random sample. This allows us to regard the individuals as representative of the full population, so we can infer things about the population from the sample. The fields of statistics and statistical learning theory are based on this idea. In fact, rigorous statistical studies reinforce this model by randomly assigning participating subjects to different treatment conditions.

Randomness also plays a central role in modern computational science. We can design algorithms that exploit probability theory by making random choices during their execution. A classic example is Monte Carlo integration, which approximates an integral by averaging the values of the integrand at some randomly chosen points. Another example is the stochastic gradient algorithm (Section 0.4), which minimizes a function by taking small random steps that, on average, point in the direction of the negative gradient of the objective. There are many more examples. We will leave aside the question about how a computer can get hold of random numbers in the first place.

Activity 0.2 (Sources of randomness). Can you think of other ways that probability and randomness arise in your research area? ▪

Activity 0.3 (Validation). Models do not need to be perfect to be valuable, but they do need to describe salient aspects of the phenomenon under study. When working with a model, you must verify the assumptions and confirm the predictions by reference to reality. In your field, what steps can you take to validate probability models? ▪

0.2.2 Sequential probability models and applications

The full version of this course focuses on three particular types of probability models:

- **A sum of independent random variables:** This model can be used to study a sequence of independent experiments (like our coin flips).
- **A martingale sequence:** This model describes a repeated game of chance where a player's strategy may depend on the past (like the stock market).
- **A Markov chain:** This model describes random sequences where the next element depends only on the current element (like a random walk).

Each of these stochastic processes has its own distinctive theory, as well as a wide range of applications.

As mathematical scientists, applying for grants in the twenty-first century, we may be particularly interested in using these probability models for

- statistics,
- signal and image processing,
- information theory,
- learning and decision-making,
- control theory,
- numerical analysis,
- uncertainty quantification,
- computer algorithms, or
- quantum information science.

We will continue our discussion with some contemporary applications of probability theory in computational science. By the end of the course, we will understand our probability models and a few of their applications well.

0.3 Randomized trace estimators

In this section, our goal is to introduce a randomized algorithm for approximating the trace of a positive-semidefinite matrix. This algorithm is a particular, but untraditional, example of a Monte Carlo method. The method depends on properties of a sum of independent random variables.

Let $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ be a positive-semidefinite matrix. If we can easily access individual entries of the matrix, then there is no impediment to evaluating the trace directly:

A positive-semidefinite matrix is a (conjugate) symmetric matrix whose eigenvalues are all positive.

$$\operatorname{tr} \boldsymbol{A} = \sum\nolimits_{i=1}^{n} a_{ii}.$$

In other words, we compute and sum the diagonal entries of the matrix. Done.

In certain applications, we cannot easily access entries of the matrix. Nevertheless, we may be able to compute the matrix–vector product $\boldsymbol{u} \mapsto \boldsymbol{A}\boldsymbol{u}$ for an arbitrary vector $\boldsymbol{u} \in \mathbb{R}^n$. Can we design an algorithm that approximates the trace using a small number of applications of this primitive?

0.3.1 A Monte Carlo method

There is a beautiful and simple randomized method for trace estimation that operates in this setting. Construct a random vector $\boldsymbol{Z} = (Z_1, \dots, Z_n) \in \mathbb{R}^n$ with *independent* entries where

We use *highlighting* to emphasize important words that you should not overlook.

$$Z_i \sim \textsc{uniform}\{\pm 1\} \quad \text{for each } i = 1, \dots, n.$$

Note that the expectation $\mathbb{E}\, Z_i = 0$ for each i. Using the matrix–vector multiplication primitive, we can form the random quantity

$$X := \boldsymbol{Z}^* (\boldsymbol{A}\boldsymbol{Z}).$$

Expanding the quadratic form, we have a detailed representation:

$$X = \sum\nolimits_{i,j=1}^{n} Z_i Z_j a_{ij} = \sum\nolimits_{i=1}^{n} a_{ii} + \sum\nolimits_{i \neq j} Z_i Z_j a_{ij}.$$

We quickly obtain the expectation of this random variable:

$$\mathbb{E}\, X = \sum\nolimits_{i=1}^{n} a_{ii} + \sum\nolimits_{i \neq j} (\mathbb{E}\, Z_i)(\mathbb{E}\, Z_j) \cdot a_{ij} = \sum\nolimits_{i=1}^{n} a_{ii} = \operatorname{tr} \boldsymbol{A}.$$

Indeed, the expectation operator is linear, and the expectation of a product of independent random variables equals the product of expectations. By a similar calculation, we may also compute the variance:

If you learn only one thing in this class, then you should learn that expectation is linear.

$$\operatorname{Var}[X] := \mathbb{E}[X^2] - (\mathbb{E}\, X)^2 = \sum\nolimits_{i \neq j} a_{ij}^2 < (\operatorname{tr} \boldsymbol{A})^2.$$

This argument takes a little more work, and we leave it as an exercise. The final inequality depends on the fact that $\boldsymbol{A}$ is positive semidefinite.

Although the random variable X is an unbiased estimator for the trace, it is not an adequate estimate because its standard deviation may be on the same scale as the trace, the thing we are trying to compute. We soon realize that we can enhance the estimator by averaging independent copies:

The standard deviation is the square-root of the variance.

$$\bar{X}_k := \frac{1}{k} \sum\nolimits_{i=1}^{k} X_i \quad \text{where } X_1, \dots, X_k \text{ are independent copies of } X. \tag{0.1}$$

The estimator $\bar{X}_k$ has the properties that

$$\mathbb{E}[\bar{X}_k] = \operatorname{tr} \boldsymbol{A} \quad \text{and} \quad \operatorname{Var}[\bar{X}_k] \leq k^{-1} (\operatorname{tr} \boldsymbol{A})^2.$$

Indeed, expectation is linear, and the variance of an independent sum is the sum of the variances. By this device, we can reduce the variability of the estimator. With only a constant (say, $k = 10$) summands, we can achieve a result that is often within a factor of two of the correct answer. The trace estimator (0.1) is a simple example of a *Monte Carlo* method.

This discussion is based on the most elementary facts about expectation and independence. Unfortunately, it does not tell us much about the *probability* that the estimator achieves a desired level of precision at a given number k of samples. To address this task, we need to develop *concentration inequalities* that describe how sharply a random variable peaks around its expectation. As the number k of samples grows without bound, the consistency of this estimator follows from *the law of large numbers*, while the *central limit theorem* precisely describes the fluctuations of the error.

0.3.2 *Application: Smoothed least squares

It may not be obvious why it would be valuable to estimate a trace by means of the matrix–vector multiplication primitive. Let us give an example from computational statistics. In fact, this is the original application of the randomized trace estimator.

Consider paired real-valued data $\{(x_i, y_i) : i = 1, \dots, n\} \subseteq \mathbb{R} \times \mathbb{R}$, where the covariates x_i are arranged in (strictly) increasing order. Suppose that we want to fit a slowly varying function to the data. We can accomplish this task by discretizing the function at the covariates. Then we formulate a least-squares problem with a smoothness penalty:

$$\text{minimize}_{\boldsymbol{f} \in \mathbb{R}^n} \quad \|\boldsymbol{y} - \boldsymbol{f}\|_{\ell_2}^2 + \zeta \cdot \|\boldsymbol{D}\boldsymbol{f}\|_{\ell_2}^2. \tag{0.2}$$

The first term enforces fidelity between the approximation $\boldsymbol{f} := (f_1, \dots, f_n) \in \mathbb{R}^n$ and the observed data $\boldsymbol{y} := (y_1, \dots, y_n) \in \mathbb{R}^n$. In the second term, the matrix $\boldsymbol{D} \in \mathbb{R}^{(n-1)\times n}$ is a (bidiagonal) first-difference operator, say

$$(\boldsymbol{D}\boldsymbol{f})_i := \frac{f_{i+1} - f_i}{x_{i+1} - x_i} \quad \text{for } i = 1, \dots, n-1.$$

As a consequence, the second term in (0.2) is small if and only if the fitted model $\boldsymbol{f}$ varies slowly. The regularization parameter $\zeta > 0$ negotiates a tradeoff between the fidelity and the smoothness.

Using calculus, we quickly determine that the solution to the smoothed least-squares problem (0.2) takes the form

$$\boldsymbol{f} = \boldsymbol{A}_\zeta \boldsymbol{y} \quad \text{where} \quad \boldsymbol{A}_\zeta = (\mathbf{I} + \zeta \boldsymbol{D}^* \boldsymbol{D})^{-1}.$$

The matrix $\mathbf{I} + \zeta \boldsymbol{D}^* \boldsymbol{D}$ is tridiagonal and positive definite, so we can apply its inverse $\boldsymbol{A}_\zeta$ with $O(n)$ arithmetic operations (for example, by Cholesky factorization and triangular solves). On the other hand, we do not have direct access to entries of $\boldsymbol{A}_\zeta$.

Why are we concerned about the trace of the matrix $\boldsymbol{A}_\zeta$? This question arises when we try to select the best value of the regularization parameter ζ. A standard approach to this task is to minimize the generalized cross-validation functional:

$$\text{gcv}(\zeta) := n \cdot \left[\frac{\|(\mathbf{I} - \boldsymbol{A}_\zeta)\boldsymbol{y}\|_{\ell_2}}{\text{tr}(\mathbf{I} - \boldsymbol{A}_\zeta)} \right]^2.$$

This quantity reflects the "effective number of degrees of freedom" in the model, per unit of approximation error, at the regularization level ζ. The effective degrees of

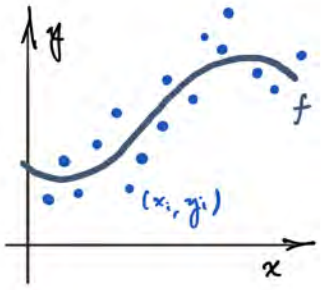


Figure 0.1 Smoothed least squares

freedom measures the smoothness of the model. We prefer smooth models but allow rougher models if warranted by a significant improvement in approximation quality.

In most cases, to minimize the gcv functional, we need to evaluate it numerically for regularization parameters ζ drawn from a grid of values. To do so, we need to estimate the trace of $\mathbf{I} - \boldsymbol{A}_\zeta$ for many choices of ζ. But we can only access $\boldsymbol{A}_\zeta$ by applying it to vectors. Therefore, randomized trace estimation is a natural tool for accelerating generalized cross-validation.

0.4 Stochastic gradient algorithms

In many learning systems, the goal is to find a reductive model that minimizes the error in explaining some observed data. In this section, we consider a basic least-squares problem that arises in statistical learning theory. We show how to convert this problem into a stochastic optimization problem. Then we introduce an algorithm, called *stochastic gradient iteration*, for solving this problem. Stochastic gradient and its variants are among the most widely used computational tools in modern machine learning. One approach to studying these algorithms involves tools from martingale theory.

0.4.1 From least squares to stochastic least squares

Suppose that we have acquired paired data $\{(\boldsymbol{a}_i, b_i) : i = 1, \dots, n\} \subseteq \mathbb{R}^d \times \mathbb{R}$. We might like to fit a model that approximates the responses b_i as a (pure) linear function of the covariates $\boldsymbol{a}_i$. The most basic approach is the ordinary least squares formulation:

$$\text{minimize}_{\boldsymbol{x} \in \mathbb{R}^d} \quad \frac{1}{2n} \sum\nolimits_{i=1}^{n} \big(\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle - b_i\big)^2.$$

As you know, this is among the simplest problems in statistical machine learning.

The normalized sum can be interpreted as an average. This observation suggests a probabilistic interpretation. Introduce the empirical measure μ of the data set:

$$\mu = \textsc{uniform}\{(\boldsymbol{a}_i, b_i) : i = 1, \dots, n\}.$$

Using the empirical measure, we can rewrite the least-squares problem as

$$\text{minimize}_{\boldsymbol{x} \in \mathbb{R}^d} \quad \mathbb{E}_{(\boldsymbol{a},b)\sim\mu} \left[\tfrac{1}{2}\big(\langle \boldsymbol{a}, \boldsymbol{x} \rangle - b\big)^2\right]. \tag{0.3}$$

The symbol $\sim$ means "has the distribution."

Defining the random vector $\boldsymbol{\xi} = (\boldsymbol{a}, b) \sim \mu$ and the bivariate function $f(\boldsymbol{x}; \boldsymbol{\xi}) = \frac{1}{2}\big(\langle \boldsymbol{a}, \boldsymbol{x} \rangle - b\big)^2$, we arrive at the compact formulation

$$\text{minimize}_{\boldsymbol{x} \in \mathbb{R}^d} \quad \mathbb{E}_{\boldsymbol{\xi}\sim\mu} f(\boldsymbol{x}; \boldsymbol{\xi}). \tag{0.4}$$

The expression (0.4) is a general way of writing a *stochastic optimization problem*. Our example is just one instance from a broad class.

At this stage, one may wonder what kind of object is the empirical measure μ. To what extent can we pass between sums and expectations?

0.4.2 Stochastic gradients

How can we solve the stochastic optimization problem (0.3)? For an unconstrained optimization problem with a continuous objective function defined on a Euclidean space, the simplest solution concept is *gradient descent*: we repeatedly take small steps

in the direction of the negative gradient of the objective function. To that end, let us compute the gradient of the objective in (0.3)–(0.4):

$$\begin{aligned}\nabla_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{\xi}} f(\boldsymbol{x}; \boldsymbol{\xi}) &= \mathbb{E}_{\boldsymbol{\xi}} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}; \boldsymbol{\xi}) \\ &= \mathbb{E}_{\mu} \nabla_{\boldsymbol{x}} \left[\tfrac{1}{2} \big(\langle \boldsymbol{a}, \boldsymbol{x} \rangle - b \big)^2 \right] = \mathbb{E}_{\mu} \left[\big(\langle \boldsymbol{a}, \boldsymbol{x} \rangle - b \big) \boldsymbol{a} \right].\end{aligned}$$

For now, we blithely pass the gradient through the expectation, even though this step requires justification. You may also realize that it is expensive to compute this expectation. Even in our discrete setting, it involves iteration over the entire set of n data pairs. In a more general case, where the support of the distribution μ has infinite cardinality, the expectation might not be tractable at all.

The idea behind the stochastic gradient algorithm is simple. We replace the gradient by an easily computable random variable $\boldsymbol{g} \in \mathbb{R}^n$ that gives an unbiased estimator for the gradient:

$$\mathbb{E}\, \boldsymbol{g} = \nabla_{\boldsymbol{x}} \mathbb{E}_{\boldsymbol{\xi}} f(\boldsymbol{x}; \boldsymbol{\xi}) = \mathbb{E}_{\mu} \left[\big(\langle \boldsymbol{a}, \boldsymbol{x} \rangle - b \big) \boldsymbol{a} \right].$$

For example, we can just take a random draw $\boldsymbol{\xi} = (\boldsymbol{a}, b)$ from the distribution μ and form

$$\boldsymbol{g} = \big(\langle \boldsymbol{a}, \boldsymbol{x} \rangle - b \big) \boldsymbol{a} \quad \text{where } (\boldsymbol{a}, b) \sim \mu.$$

In our discrete setting, this amounts to choosing one data point at random and constructing the vector

$$\boldsymbol{g} = \big(\langle \boldsymbol{a}_I, \boldsymbol{x} \rangle - b_I \big) \boldsymbol{a}_I \quad \text{where } I \sim \textsc{uniform}\{1, \dots, n\}.$$

Our randomized gradient approximation no longer requires us to compute an expectation. But the stochastic gradient it is correct on average, and that turns out to be enough.

0.4.3 The stochastic gradient iteration

To solve the optimization problem (0.3), we proceed as follows. Make an initial guess, say $\boldsymbol{x}_0 \in \mathbb{R}^d$, for the solution. At each iteration $j = 1, 2, 3, \dots$, construct a randomized gradient estimate:

$$\boldsymbol{g}_j = \big(\langle \boldsymbol{a}, \boldsymbol{x}_{j-1} \rangle - b \big) \boldsymbol{a} \quad \text{where } (\boldsymbol{a}, b) \sim \mu.$$

Update the current iterate:

$$\boldsymbol{x}_j = \boldsymbol{x}_{j-1} - \eta_j \boldsymbol{g}_j \quad \text{for a step size parameter } \eta_j > 0.$$

This process repeats indefinitely.

In contrast to a classic optimization algorithm, the sequence $(\boldsymbol{x}_0, \boldsymbol{x}_1, \boldsymbol{x}_2, \dots)$ of iterates is composed of random elements. Under what conditions does this sequence converge to a minimizer of the stochastic least-squares problem (0.3)? In what sense? Does the initialization $\boldsymbol{x}_0$ matter? How should we select the step sizes? What is the role of the distribution of the data, encapsulated in the measure μ?

These questions seem daunting. Indeed, at each iteration, the random iterate depends on the entire history of the algorithm. Yet we can easily prove finite-time guarantees using *maximal inequalities for martingales*, while asymptotic guarantees follow from the *martingale convergence theorem*.

0.5 Markov chains

In computational mathematics, one of the basic challenges is to draw samples from a complicated probability distribution. Although there are reliable routines for producing uniform random variables or normal random variables, there may be no direct approach by which we can sample from a more complicated model. A remarkable and powerful idea is to construct a *sequence of distributions* that ultimately converges to the desired distribution. This approach is called *Markov chain Monte Carlo* (MCMC). In this section, we give a short description of an MCMC algorithm for drawing samples from a basic model in statistical physics.

0.5.1 The Ising model on a graph

The Ising model was designed as a stylized description of ferromagnetism. We construct a graph that reflects which atoms in a material, say a crystal lattice, are adjacent. We assume that each atom can have a positive or negative spin. When most of the spins are oriented the same way, the net magnetic moment is large and the material will generate a significant magnetic field. On the other hand, when the spins are incoherent, there is little magnetic effect. We may ask to sample a typical configuration of spins.

Formally, let $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ be an undirected combinatorial graph on a finite set V of vertices with edge set E. To each vertex in the graph, we assign a positive spin $(+1)$ or a negative spin (-1). We can encode the family of assignments in a configuration function $\boldsymbol{\sigma} : \mathsf{V} \to \{\pm 1\}$. The nearest-neighbor Ising model is a probability distribution over these configurations:

$$\mu(\boldsymbol{\sigma}) := \frac{1}{Z_\beta} \exp\left(\beta \sum\nolimits_{\{u,v\} \in \mathsf{E}} \sigma(u)\sigma(v)\right) \quad \text{for } \boldsymbol{\sigma} \in \{\pm 1\}^{\mathsf{V}}. \tag{0.5}$$

In this expression, $\beta > 0$ is a fixed parameter called the *inverse temperature*, and Z_β is a normalizing constant known as the *partition function*. This type of distribution μ is usually called a *Gibbs measure*.

The intuition behind the formula (0.5) is that the probability of a configuration is largest when neighboring vertices tend to share the same spin (either positive or negative). If the parameter β is large (i.e., the temperature is low), the model places most of the probability mass on configurations where spins are aligned. When the parameter β is small (i.e., the temperature is high), we are more likely to see heterogeneous configurations where spins at neighboring vertices differ.

For a more picayune example, imagine a social network $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ where the edge set E reflects which members of the community V are friends. We may imagine that each individual has a preference for Coke $(+1)$ or for Pepsi (-1). While friends tend to have similar preferences, their affinity depends on the interaction strength β. The probability $\mu(\boldsymbol{\sigma})$ expresses the likelihood that the network exhibits a particular configuration $\boldsymbol{\sigma}$ of tastes in soft drinks. Sampling this distribution is a pressing issue in viral marketing.

0.5.2 The Metropolis–Hastings chain

For a general graph G, it is not obvious how we can draw a random sample from the Gibbs measure μ of the Ising model (0.5). Indeed, the configuration space $\{\pm 1\}^{\mathsf{V}}$ may be quite large. Individual spins have a complicated dependency because of the graph structure. Moreover, it is challenging just to compute the partition function Z_β.

Instead, suppose that we start with an initial configuration $\boldsymbol{\sigma}_0$. Is there some way to make simple random updates to the configuration to obtain a sequence $(\boldsymbol{\sigma}_n : n \in \mathbb{N})$

of random configurations that tends toward a sample from the Gibbs distribution μ? In fact, we can accomplish this task via an elegant method called the *Metropolis–Hastings algorithm*.

At each iteration, the Metropolis–Hastings algorithm tells us to update the current configuration $\boldsymbol{\sigma} \in \{\pm 1\}^{\mathsf{V}}$ according to the following procedure.

1. **Proposal:** Choose a random vertex $u \sim \textsc{uniform}(\mathsf{V})$. Form a candidate configuration $\boldsymbol{\sigma}^?$ by flipping the spin at vertex u:

$$\sigma^?(v) := \begin{cases} -\sigma(v), & v = u; \\ \sigma(v), & v \neq u \end{cases} \quad \text{for } v \in \mathsf{V}.$$

2. **Acceptance ratio:** Compute the ratio α of the Gibbs measure μ at the candidate configuration $\boldsymbol{\sigma}^?$ and at the current configuration $\boldsymbol{\sigma}$:

$$\alpha := \frac{\mu(\boldsymbol{\sigma}^?)}{\mu(\boldsymbol{\sigma})} = \exp\left(\beta \sum\nolimits_{v:\{u,v\}\in\mathsf{E}} (-2\sigma(u))\sigma(v)\right).$$

   This quantity reflects how much the probability increases or decreases by flipping the spin at the site u.

3. **Update:** Draw a random variable $X \sim \textsc{uniform}[0,1]$. Update the current configuration from $\boldsymbol{\sigma}$ to $\boldsymbol{\sigma}'$ according to the rule

$$\boldsymbol{\sigma}' := \begin{cases} \boldsymbol{\sigma}^?, & X \leq \alpha; \\ \boldsymbol{\sigma}, & X > \alpha. \end{cases}$$

   In other words, we move to the candidate configuration $\boldsymbol{\sigma}^?$ with probability $1 \wedge \alpha$. Otherwise, we remain at the original configuration $\boldsymbol{\sigma}$.

Let us point out a few key features of this algorithm. First, the candidate configuration $\boldsymbol{\sigma}^?$ is random, but it is easy to construct. Second, we do not need to know the partition function Z_β to compute the acceptance ratio α, and we only need to sum over vertices adjacent to the selected vertex u. Third, we always move to the candidate configuration $\boldsymbol{\sigma}^?$ if it has larger probability than the current configuration $\boldsymbol{\sigma}$. We may elect to move to a less probable configuration, but we inject randomness into this decision.

0.5.3 Markov chains

To recap, the Metropolis–Hastings algorithm generates a sequence of random configurations taking values in the space $\{\pm 1\}^{\mathsf{V}}$. At each step in the sequence, the next configuration depends on the current configuration, but it is independent from the trajectory of the algorithm before the current time. This kind of random sequence is called a *Markov chain*.

In the theory of Markov chains, some of the basic questions center on their long-term behavior. Does there exist a distribution that is stationary under the dynamics of the Markov chain? Is this stationary distribution unique? Does the Markov chain converge to a stationary distribution from any initialization? How long does it take before the distribution of the chain is close to the stationary distribution? Can we compute moments of the stationary distribution by averaging along the trajectory of the chain?

The Metropolis–Hastings algorithm emerged as general recipe for designing a Markov chain with a given stationary distribution.

Problem 0.4 (Metropolis–Hastings for Ising model: Gibbs measure is stationary). Suppose that we draw $\boldsymbol{\sigma}$ from the distribution μ defined in (0.5). Show that the output $\boldsymbol{\sigma}'$ of the Metropolis–Hastings algorithm in Section 0.5.2 also follows the distribution μ.

In other words, the Gibbs measure μ is stationary under the Metropolis–Hastings dynamics. This fact hints that the sequence of random configurations might converge to a sample from the Gibbs measure. Indeed, it can be shown that Metropolis–Hastings allows us to simulate the distribution of the Ising model on a graph.

Because of the fundamental importance of Markov chains, we used to introduce some basic definitions and applications at the end of the course. Unfortunately, we would need an entire term to fully address the questions that we posed in the earlier paragraphs. ACM 216 takes up this study for Markov chains with discrete configuration spaces, and ACM sometimes offer courses on Markov chains with continuous state spaces.

0.6 Probability and measure

We have now encountered several different computational applications of probability. These applications already raise a large number of questions about the foundations of probability theory. To develop appropriate answers to these questions, we need to learn about measure theory.

0.6.1 Probability concepts

Here are some of the questions that we must confront:

- What is a probability distribution? What events have probabilities?
- What is a random variable? Is it the same as a distribution?
- How do we define the expectation of a random variable? What properties does expectation have?
- How do we bound the probability that a random variable takes values far from its expectation?
- What does it mean for a pair of random variables to be similar? When does a sequence of random variables converge?
- What does it mean for two probability distributions to be similar? When does a sequence of distributions converge?
- How do we define independence? How do we condition on prior knowledge?
- How can we make sense of an infinite sequence of independent random variables?
- What behavior should we expect an independent sum to exhibit? What about a martingale? A Markov chain?

0.6.2 The role of measure theory

To address these questions in a systematic way, we need to learn measure theory. There are many reasons that measure theory is the appropriate language for talking about probability.

1. **What is an event?** An event is a collection of outcomes of a probability experiment to which we can ascribe a probability. In simple discrete settings (coins, dice, etc.), all sets of outcomes are events. But as soon as we move to the continuous setting (height, lifetimes, etc.), we encounter a problem. There is no way to consistently assign probabilities to all subsets of the real line. We need a way to delineate which events are legitimate and which are not.

2. **Discrete versus continuous:** In elementary probability courses, we usually hone our intuition with discrete probability models, and we develop a complete set of definitions and formulas for this case. Afterward, we proceed to continuous probability models, and we develop another complete set of definitions and formulas for this case. Many students will perceive an analogy between these situations. Measure theory allows us to treat all random variables (discrete, continuous, mixed) on an equal footing, with one set of definitions that is valid in all cases.
3. **Univariate versus multivariate:** Similarly, elementary probability courses begin with treatments of individual random variables before proceeding to pairs of random variables, random vectors, and so on. Each case is burdened with its own terminology and formulas. Measure theory allows us to regard a family of random variables as determining a distribution on a space of higher dimension. There is no fundamental distinction between univariate and multivariate models.
4. **Independence:** What does it mean for two probabilistic events to be independent from each other? In elementary courses, we typically define a pair of random variables to be independent when the joint cumulative distribution function factors. This definition is unintuitive and hard to work with. Measure theory allows us to describe independence naturally through the construction of product measures.
5. **Conditioning:** How does knowledge about the world inform our predictions? Introductory probability courses also present a long catalog of formulas for conditioning of discrete random variable and another catalog for conditioning of continuous random variables. It is in no way clear, however, that we can mix these ideas. Measure theory provides a natural set of tools for handling cases that cannot be addressed by manipulation of probability masses or densities.
6. **Approximation of distributions:** In its most basic form, the central limit theorem states that a standardized sum of independent uniform ± 1 random variables converges to a continuous normal distribution. But it must be perplexing to contemplate discrete distributions that have a continuous limit. By treating all probability distributions as measures, this conceptual difficulty evaporates.

At this stage, it may not be clear exactly what we mean by this encomium to measure theory. Once we have finished the course, and you look back on this section, you will appreciate how measure-theoretic probability resolves all of these issues.

In the first lecture, we will begin our study of measure theory. These concepts are easier to understand without the added burden of a probabilistic interpretation, so we will first introduce the major concepts in the comfortable setting of the integers and then we will upgrade the basic ideas to the real line. Once the foundations are solid, we will recast measure theory as a language for discussing probability.

Notes

Monte Carlo methods were invented to perform challenging integrations that arise in computational physics. In practice, Monte Carlo methods are often enhanced with more sophisticated sampling techniques (e.g., importance sampling or control variates). For many problems, where it is intractable to produce unbiased samples, we can implement a Markov chain that drives an initial distribution toward the desired target. The resulting technique, called Markov Chain Monte Carlo (MCMC), has fundamental importance in contemporary machine learning and computational science.

MCMC was invented in the 1940s by the physicists Nick Metropolis and Stanislaw Ulam, who were on the staff of the Manhattan Project. In the 1970s, MCMC was revitalized by Hastings [Has70]. Image processing applications emerged in the 1980s with the work of Geman & Geman [GG84]. MCMC methods entered the mainstream of computational statistics in the 1990s; for example, see [Tie94]. They are now central tools for Bayesian solution of inverse problems [Stu10]. For an accessible introduction to Markov chains on discrete state spaces, see the book of Levin & Peres [LP17].

Randomized trace estimators were proposed by Girard [Gir89] for generalized cross-validation of smoothed least-squares problems. The specific estimator that we described was proposed by Hutchinson [Hut90]. See [MT20] for a more in-depth discussion and more recent extensions.

The stochastic gradient algorithm was first proposed by Robbins and Monro [RM51] for rootfinding problems that arise in statistics (quantile estimation, least-squares computation). These methods have been rediscovered many times in other research communities. Stochastic gradient achieved a renewed prominence in the machine learning community after an influential paper [Bot98] of Bottou. At present, variants of stochastic gradient are widely used in statistical machine learning and for the training of artificial neural networks.

Our discussion of the importance of measure theory for probability theory owes a great debt to Pollard's book [Pol02].

We have presented the tired quotation of George Box about the fact that models are, well, models. The discrepancy between reality, models of reality, and our perception of reality is an ancient theme in philosophy. For example,

> "*Dao ke dao, fei chang dao. / Ming ke ming, fei chang ming.*"
>
> "The truth can be known, but it is not the truth known to you. / Things can be named, but the names are not the things."
>
> —Laozi, *Dao Deching*, circa 400 BCE

You may also recall Plato's "Allegory of the Cave," where the prisoners imagine that the shadows of objects are the reality of the objects. In his meditation on a ball of wax, Descartes considered the falseness of our senses, and he argued that our knowledge of the world is uncertain. He concluded that, on some level, models for reality can capture truths that we cannot directly perceive.

Lecture bibliography

[Bot98] L. Bottou. "Online Algorithms and Stochastic Approximations". In: *Online Learning and Neural Networks*. Revised, Oct. 2012. Cambridge University Press, 1998. URL: http://leon.bottou.org/papers/bottou-98x.

[GG84] S. Geman and D. Geman. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6.6 (1984), pages 721–741. DOI: 10.1109/TPAMI.1984.4767596.

[Gir89] D. A. Girard. "A fast "Monte Carlo cross-validation" procedure for large least squares problems with noisy data". In: *Numer. Math.* 56.1 (1989), pages 1–23. DOI: 10.1007/BF01395775.

[Has70] W. K. Hastings. "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57.1 (1970), pages 97–109. DOI: 10.1093/biomet/57.1.97.

[Hut90] M. F. Hutchinson. "A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines". In: *Comm. Statist. Simulation Comput.* 19.2 (1990), pages 433–450. DOI: 10.1080/03610919008812864.

[LP17] D. A. Levin and Y. Peres. *Markov chains and mixing times*. Second edition of [MR2466937], With contributions by Elizabeth L. Wilmer, With a chapter on "Coupling from the past" by James G. Propp and David B. Wilson. American Mathematical Society, Providence, RI, 2017. DOI: 10.1090/mbk/107.

[MT20] P.-G. Martinsson and J. A. Tropp. "Randomized numerical linear algebra: Foundations and algorithms". In: *Acta Numerica* (2020).

[Pol02] D. Pollard. *A user's guide to measure theoretic probability*. Cambridge University Press, 2002.

[RM51] H. Robbins and S. Monro. "A stochastic approximation method". In: *Ann. Math. Statistics* 22 (1951), pages 400–407. DOI: 10.1214/aoms/1177729586.

[Stu10] A. M. Stuart. "Inverse problems: a Bayesian perspective". In: *Acta Numer.* 19 (2010), pages 451–559. DOI: 10.1017/S0962492910000061.

[Tie94] L. Tierney. "Markov chains for exploring posterior distributions". In: *Ann. Statist.* 22.4 (1994). With discussion and a rejoinder by the author, pages 1701–1762. DOI: 10.1214/aos/1176325750.

I. measure theory

1. Measures on the Integers

Agenda:

1. Distributions of mass
2. Motivation for definition
3. Measures on the integers
4. Basic properties
5. Examples
6. Specifying a measure

"Truth is truth
To th'end of reck'ning."

—William Shakespeare, *Measure for Measure*, Act V, Scene 1

Measure theory is a branch of mathematics that provides tools for describing a "distribution of mass" over a domain. The mathematical abstraction is the same, regardless of whether we are talking about distributions of physical mass or distributions of probability or whatnot.

Modern probability is written in the language of measure theory, and we cannot develop a mature understanding of the subject without it. Although there are more elementary approaches, we really need the technical apparatus of measure theory to properly build foundations (e.g., conditional expectations) and to pursue modern applications (e.g., optimal transportation).

The first task of measure theory is to define a measure, which is the basic object that describes a distribution of mass over a domain. Later, we will use measures to introduce integrals, which allow us to add up the values of a function on the domain, weighted by the distribution of mass.

We will begin our study in a drastically simplified setting, the case of distributions over the integers. In this environment, we can develop intuitions without worrying about technical matters. In the next lecture, we will expand on the basic ideas to give the abstract definition of a measure. The abstraction is important for understanding measures on the real line. As always, our goal is to be correct but never fussy.

1.1 Distributions on the integers

So, what does measure theory allow us to measure? All kinds of things. For instance,

- Combinatorial content, such as counting points.
- Geometric content, such as length, area, or volume.
- Physical content, such as mass or charge.
- Probability or likelihood, which is our main interest.

In each case, we can think about some kind of substance that is distributed over a domain. We will use the word "mass" generically to refer to the substance that is being distributed.

The first step in our program is to study mass distributions, or measures, on the integers. The distribution simply places some amount of mass on each integer. This model is not rich enough to support all of the examples that might interest us, but it already allows us to identify some of the core properties of a measure. In the next few lectures, we will gradually introduce the additional technical ideas that are required to define measures on the real line.

Example 1.1 (Counting). Let $\mathsf{A} \subseteq \mathbb{Z}$ be a subset of the integers. We may be interested in counting the number of points in the set A. To that end, define $\#\mathsf{A}$ to be the number of points in the set A. We will call $\#$ the *counting measure* (on the integers).

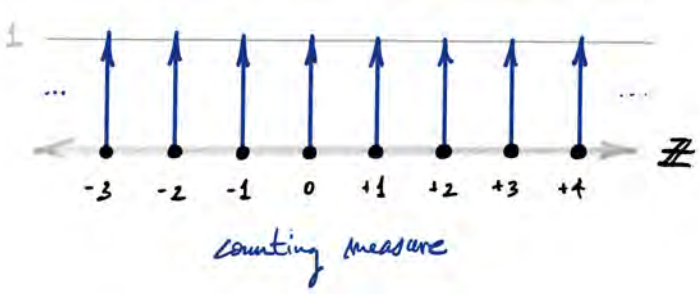
As a concrete example, consider an interval $\mathsf{A} = \{a+1, a+2, \ldots, a+k\}$ where $a \in \mathbb{Z}$. Then $\#\mathsf{A} = k$ is just the length of the interval. On the other hand, for a set A that contains an infinite number of points, we set $\#\mathsf{A} = +\infty$.

At first glance, counting may not seem to have much to do with distributions. In fact, it is just an elementary example. Counting is associated with the distribution that places one unit of mass on each integer. Thus, for a set A with finite cardinality,

$$\#\mathsf{A} = \sum\nolimits_{i \in \mathsf{A}} 1.$$

If A has infinite cardinality, we can interpret the sum as the limit of an increasing sequence of finite partial sums, which equals $+\infty$.

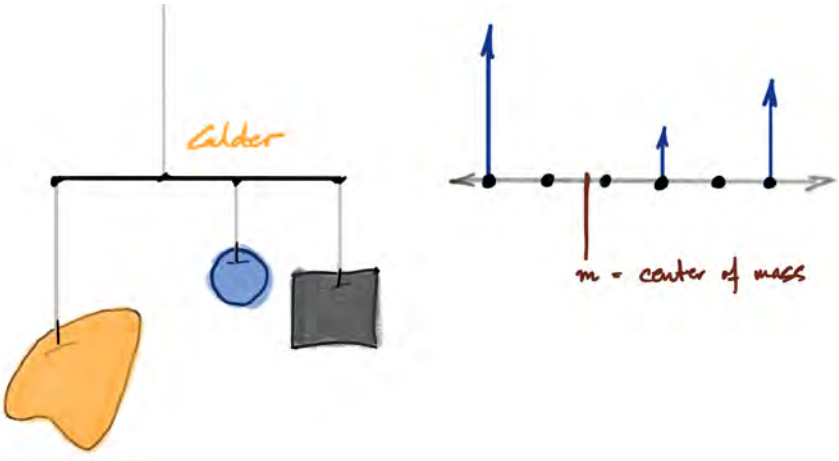
It will be valuable to develop schematics that can give us intuition for distributions. Here is a picture of the counting measure:



Each spike has height one, and it represents one unit of mass, concentrated at an integer point. ∎

Example 1.2 (Physical mass). Alexander Stirling Calder was a mid-20th century American sculptor, well-known for his whimsical constructions of measures. Here is an illustration of a Calder mobile, along with an associated schematic that describes how the mass is distributed over the integer points:

Calder grew up in Pasadena. His father, also named Alexander Calder, was a prominent sculptor, who designed the frieze over the original Pasadena Hall in 1910. These sculptures are now located on the bridge between Church Hall and Crellin Hall.

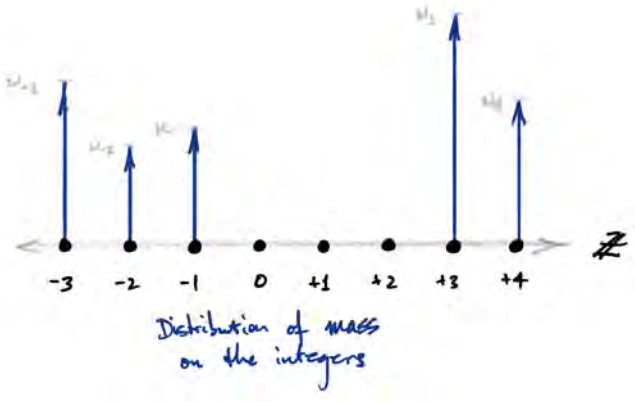


We imagine that the horizontal rod is massless and the strings have no width.

It is natural to describe a general distribution of physical mass over the integers using a sequence $(w_i : i \in \mathbb{Z})$ of positive real numbers. The number w_i specifies how

Warning: Recall that positive always means ≥ 0! ∎

much mass is placed at the integer i. The total mass carried by a set $\mathsf{A} \subseteq \mathbb{Z}$ is simply $\sum_{i \in \mathsf{A}} w_i$. Here is a picture of what the distribution might look like.



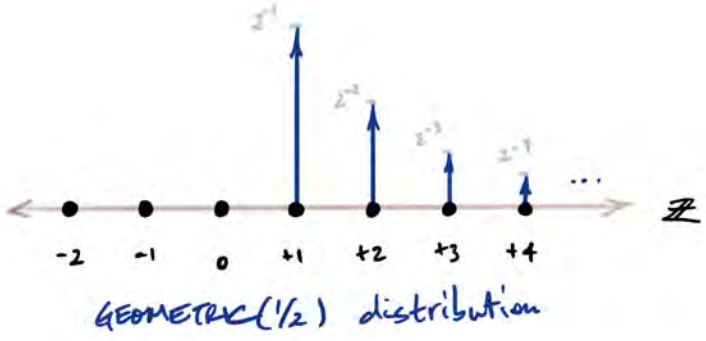
As in Example 1.1, the spikes represent point masses, which have no spatial extent.

The center of mass m of the system is the location at which we must support the number line to make it balance. In other words, the torque about the point m is zero:

$$\sum\nolimits_{i \in \mathbb{Z}} (i - m)\, w_i = 0. \tag{1.1}$$

We have added up the values of a function $i \mapsto i - m$, weighted by the distribution $(w_i : i \in \mathbb{Z})$ of mass. This is an explicit example of integrating a function against a distribution, a task we will learn to accomplish in some generality. ▪

Example 1.3 (Probability). A probabilistic experiment has a distribution of possible outcomes, some more likely than others. For a truly boring instance, consider the process of flipping a fair penny until we encounter the first heads. By elementary reasoning, for each natural number $i \in \mathbb{N}$, the probability that the first heads appears on the ith flip is 2^{-i}. Here is an illustration of this distribution of probability:



This is a particular example of the geometric distribution, which is a family of probability distributions that are supported on the natural numbers.

We can describe a general probability distribution over the integers via a sequence $(p_i : i \in \mathbb{Z})$ of positive numbers that sums to one: $\sum_{i \in \mathbb{Z}} p_i = 1$. The normalization reflects the fact that the total probability of all the outcomes must equal one, or 100%. The probability that the outcome lies in a set $\mathsf{A} \subseteq \mathbb{Z}$ is just $\sum_{i \in \mathsf{A}} p_i$, the total probability mass carried by the set.

The expectation m of the probability distribution is given by the series

$$m = \sum\nolimits_{i \in \mathbb{Z}} i\, p_i. \tag{1.2}$$

You may notice the strong analogy between (1.1) and (1.2), which suggests a mechanical interpretation of the expectation.

In other words, we average the values of the function $i \mapsto i$, weighted by the distribution $(p_i : i \in \mathbb{Z})$ of probability. This is another explicit example of an integral, and we begin to see why probability theory demands a robust theory of integration. ▪

> **Aside: (Signed distributions).** We can consider a more general class of distributions on the integers that may place either a positive or negative mass at each integer. This situation arises in physics, where it models a distribution of electric charge on the integers. Mathematically, we describe this kind of distribution using an object called a *signed measure*. For the moment, however, we will only consider distributions that are positive.

1.2 What properties should a measure have?

In the last section, we discussed several situations where we may encounter a distribution of mass over the integers. A measure is simply a mathematical object that describes a distribution of mass. We will need this formalism to give a unified treatment of distributions over domains that are more general than the integers.

Fortunately, we can use distributions over the integers to get acquainted with the basic ingredients in the definition of a measure. The archetype of a measure on the integers is the counting measure, which reports the number of points in a set (see Example 1.1). From this example, we can identify some intuitive properties that all measures should share.

1.2.1 Measures are defined on sets

What kind of function is the counting measure? After a moment of thought, we realize that the counting measure reports the number of points *in a set of integers*. That is, we have defined $\#\mathsf{A}$ for a set $\mathsf{A} \subseteq \mathbb{Z}$. In contrast, it does not make any sense to talk about the number of points in an individual integer $a \in \mathbb{Z}$, even though we can talk about the number of points in the singleton set $\{a\}$. In other words, the domain of the counting measure $\#$ on the integers is the power set $\mathcal{P}(\mathbb{Z})$, which contains all subsets of integers.

Evidently, $\#\{a\} = 1$.

This innocuous observation points to a central idea in measure theory:

> A measure is a function, defined on sets, that reports the amount of mass carried by each set.

Different measures describe different ways to assign mass to sets. In the subsequent lectures, we will refine this idea in an essential way.

> **Aside:** Although a measure is defined on subsets of the domain, it may not be defined on every subset. This issue is truly fundamental, but we can ignore it until the next lecture.

1.2.2 Measures take positive values, which may be infinite

What is the range of the counting measure? It is very easy to see that the cardinality of a set of integers may take any value in $\mathbb{Z}_+ \cup \{+\infty\}$, the set of positive integers together with $+\infty$. By consideration of our other examples (physical mass, probability), we quickly realize that it is too restrictive to require a measure to take integer values.

From these observations, we extract several conclusions. First, a measure should assign a positive amount of mass to a set. Second, it is eminently reasonable for the measure to assign an infinite amount of mass to some sets. In summary, the range of a

measure is contained in $\overline{\mathbb{R}}_+ := \mathbb{R}_+ \cup \{+\infty\} = [0, +\infty]$, the set of positive real numbers, together with $+\infty$.

1.2.3 Measures are finitely additive

At this stage, we have decided that a measure is a function that maps a set to a positive number. But what kind of function? What is the core property of the counting measure that other measures must share?

A pair (A, B) of sets is *disjoint* when they have a trivial intersection: $\mathsf{A} \cap \mathsf{B} = \emptyset$. In other words, disjoint sets do not overlap.

Observe that the total number of points in a union of two *disjoint* sets equals the sum of the number of points in each of the sets:

$$\#(\mathsf{A} \dot\cup \mathsf{B}) = (\#\mathsf{A}) + (\#\mathsf{B}) \quad \text{for disjoint } \mathsf{A}, \mathsf{B} \subseteq \mathbb{Z}. \tag{1.3}$$

The symbol $\dot\cup$ denotes the union of disjoint sets.

The formula (1.3) expresses the *finite additivity* of the counting measure. This is a very natural requirement for any distribution of mass: the total mass carried by two disjoint sets must equal the sum of the masses of the two sets.

Observe that the finite additivity property (1.3) has several formal implications. First, it allows us to compute the number of points in a set difference:

$$\#(\mathsf{A} \setminus \mathsf{E}) = (\#\mathsf{A}) - (\#\mathsf{E}) \quad \text{when } \mathsf{E} \subseteq \mathsf{A} \subseteq \mathbb{Z}.$$

Second, we can extend the additivity rule to a disjoint family containing a *finite* number of sets:

$$\#\big(\dot\bigcup_{i=1}^{n} \mathsf{A}_i\big) = \sum_{i=1}^{n} (\#\mathsf{A}_i) \quad \text{for disjoint sets } \mathsf{A}_i \subseteq \mathbb{Z}. \tag{1.4}$$

The words *family* and *collection* are alternative terms for "set." For euphony, they are often used to refer to a set whose elements are sets.

A family of sets is *disjoint* when each pair of sets in the family is disjoint.

This identity follows by iteration of (1.3).

1.2.4 Measures are countably additive

Measure theory demands that we upgrade the finite additivity rule (1.4) to a stronger property called *countable additivity*. For the counting measure, countable additivity is the trivial statement that

$$\#\big(\dot\bigcup_{i=1}^{\infty} \mathsf{A}_i\big) = \sum_{i=1}^{\infty} (\#\mathsf{A}_i) \quad \text{for disjoint sets } \mathsf{A}_i \subseteq \mathbb{Z}. \tag{1.5}$$

Here, we may consider any countable family of disjoint sets.

A set is *countable* if it can be placed in one-to-one correspondence with a subset of the natural numbers. For example, finite sets are countable; the integers are countable; and the rational numbers are countable. The real numbers are *not* countable.

We will require a similar property to hold for every measure. To see what this looks like, consider a general mass distribution $(w_i : i \in \mathbb{Z})$ as in Example 1.2. For any subset $\mathsf{A} \subseteq \mathbb{Z}$, we can unambiguously define the total mass in the set via $\mu(\mathsf{A}) := \sum_{i \in \mathsf{A}} w_i$. It is not too hard to check that

$$\mu\big(\dot\bigcup_{i=1}^{\infty} \mathsf{A}_i\big) = \sum_{i=1}^{\infty} \mu(\mathsf{A}_i) \quad \text{for disjoint sets } \mathsf{A}_i \subseteq \mathbb{Z}. \tag{1.6}$$

This identity expresses the countable additivity of μ. Since the integers $\mathbb{Z}$ are countable, we will often invoke this relation to translate information about the mass on singleton sets to the mass on general sets of integers.

In a wider context, countable additivity allows us to effortlessly handle sequences of sets, to perform limiting operations, and to construct integrals. In general, if we wish to enjoy the fruits of the countable additivity property, we must enforce it explicitly because it does *not* follow from finite additivity (!). The crucial role of *countable* additivity is a brilliant technical insight that powers the entire field of measure theory.

Exercise 1.4 (Countable additivity). Verify that (1.6) holds for the function $\mu(\mathsf{A}) = \sum_{i \in \mathsf{A}} w_i$ with $w_i \geq 0$. **Hint:** A (countable) sum of positive numbers has an unambiguous value, no matter the order of summation.

1.3 Measures on the integers

In the last section, we explored the properties of the counting measure, and we argued that any distribution of mass on the integers should share similar properties. In this section, we will give a rigorous definition of a measure, which crystallizes these ideas.

1.3.1 Formal definition

Let us reiterate the key properties we have uncovered. First, a measure is defined on sets of integers, and it reports the total mass carried by a set. Second, measures are countably additive: the measure of a countable union of disjoint sets must equal the sum of the measures of the sets. The definition of a measure just collects these principles.

Definition 1.5 (Measure on the integers). A *measure on the integers* is a positive function defined on sets of integers:

$$\mu : \mathcal{P}(\mathbb{Z}) \to [0, +\infty].$$

A measure has two distinguished properties:

1. **Empty set:** $\mu(\emptyset) = 0$.
2. **Countable additivity:** If $(\mathsf{A}_i : i \in \mathbb{N})$ is a countable sequence of *disjoint* sets in $\mathcal{P}(\mathbb{Z})$, then the measure of the union is the sum of the measures:

$$\mu\big(\dot{\bigcup}_{i=1}^{\infty} \mathsf{A}_i\big) = \sum_{i=1}^{\infty} \mu(\mathsf{A}_i). \tag{1.7}$$

Exercise 1.6 (Counting measure). Let us verify that Definition 1.5 is not vacuous. Show that the counting measure $\# : \mathcal{P}(\mathbb{Z}) \to [0, +\infty]$ is a measure on the integers.

Exercise 1.7 (Set differences). Let μ be a measure on the integers. Confirm that $\mu(\mathsf{A} \setminus \mathsf{E}) = \mu(\mathsf{A}) - \mu(\mathsf{E})$ when $\mathsf{E} \subseteq \mathsf{A} \subseteq \mathbb{Z}$.

Warning 1.8 (Details, details). There are a number of pitfalls that often trip students up when they first encounter the definition of a measure:

1. Measures assign mass to *sets*, not to points.
2. Measures can return the value $+\infty$.
3. Measures are assumed to be *countably* additive.
4. The definition of countable additivity requires *disjointess* of the sets.

Be careful! ∎

Problem 1.9 (*Finite additivity). A function $\mu_0 : \mathcal{P}(\mathbb{Z}) \to [0, +\infty]$ is *finitely additive* if it satisfies

$$\mu_0\big(\dot{\bigcup}_{i=1}^{n} \mathsf{A}_i\big) = \sum_{i=1}^{n} \mu_0(\mathsf{A}_i) \quad \text{for all } \textit{disjoint } \mathsf{A}_i \subseteq \mathbb{Z} \text{ and } n \in \mathbb{N}. \tag{1.8}$$

Confirm that a measure μ on the integers is always finitely additive. Exhibit a finitely additive function μ_0 with $\mu_0(\emptyset) = 0$ that is not a measure.

Conclude that it is necessary to assume that a function on sets is countably additive if we wish to use this property. The proof of Proposition 1.20 gives a first hint about why countable additivity is so valuable.

> **Aside: (Signed measures).** More precisely, we have defined a *positive measure* on all sets of the integers. We omit the qualification that the measure is positive because this is the most common case by far. There is a related object, called a *signed measure*, which can take negative values. In that case, we always use the word "signed" to maintain the distinction.

1.3.2 Basic properties

Measures are required to satisfy the countable additivity property (1.7), which involves disjoint sets. Even for sets that are not disjoint, the measure still satisfies some elegant rules. As a first example, let us show that a measure is monotone.

Example 1.10 (Measure: Monotonicity). Let μ be a measure on the integers. Consider *nested* sets $\mathsf{A} \subseteq \mathsf{B} \subseteq \mathbb{Z}$. We have the disjoint decomposition

$$\mathsf{B} = \mathsf{A} \dot{\cup} (\mathsf{B} \setminus \mathsf{A}).$$

Using (countable) additivity and positivity of the measure, we find that

$$\mu(\mathsf{B}) = \mu(\mathsf{A}) + \mu(\mathsf{B} \setminus \mathsf{A}) \geq \mu(\mathsf{A}).$$

In summary, the measure is monotone with respect to set inclusion: $\mathsf{A} \subseteq \mathsf{B}$ implies that $\mu(\mathsf{A}) \leq \mu(\mathsf{B})$. This fact gives additional support to the heuristic that measures model distributions of mass. If you enlarge a set, the amount of mass that it carries can only increase. ∎

Exercise 1.11 (Measure: Properties). Let μ be a measure on the integers. Prove the following claims.

1. **Inclusion–exclusion:** For all sets $\mathsf{A}, \mathsf{B} \subseteq \mathbb{Z}$,

   $$\mu(\mathsf{A} \cup \mathsf{B}) + \mu(\mathsf{A} \cap \mathsf{B}) = \mu(\mathsf{A}) + \mu(\mathsf{B}).$$

   What happens when $\mathsf{B} \subseteq \mathsf{A}$?
2. **Countable subadditivity:** Consider a countable sequence $(\mathsf{A}_i : i \in \mathbb{N})$ of sets of integers, not necessarily disjoint. Show that

   $$\mu\big(\textstyle\bigcup_{i=1}^{\infty} \mathsf{A}_i\big) \leq \sum_{i=1}^{\infty} \mu(\mathsf{A}_i). \tag{1.9}$$

   Hint: Reason about the increasing sets $\mathsf{B}_n := \bigcup_{i=1}^{n} \mathsf{A}_i$ for $n \in \mathbb{N}$.

You may find that Venn diagrams are helpful here. In each case, the key idea is to rewrite a set as the disjoint union of two or more subsets.

Exercise 1.12 (Measure: Monotone limits). Let μ be a measure on the integers. Verify that the measure interacts well with monotone limits.

1. **Increasing limits:** For an increasing sequence $\mathsf{A}_1 \subseteq \mathsf{A}_2 \subseteq \mathsf{A}_3 \subseteq \cdots \subseteq \mathbb{Z}$, show that

   $$\mu\big(\textstyle\bigcup_{i=1}^{\infty} \mathsf{A}_i\big) = \lim_{i \to \infty} \mu(\mathsf{A}_i).$$

2. **Decreasing limits:** For a decreasing sequence $\mathbb{Z} \supseteq \mathsf{A}_1 \supseteq \mathsf{A}_2 \supseteq \mathsf{A}_3 \supseteq \cdots$, show that

   $$\mu\big(\textstyle\bigcap_{i=1}^{\infty} \mathsf{A}_i\big) = \lim_{i \to \infty} \mu(\mathsf{A}_i),$$

   provided that $\mu(\mathsf{A}_i) < +\infty$ for some index i. **(*)** Give an example to show that the statement may fail without the qualification.

1.3.3 Finite measures

It is useful to make a distinction among measures that distribute a finite amount of mass and measures that distribute an infinite amount of mass.

Definition 1.13 (Finite measure). Let μ be a measure on the integers. We say that μ is a *finite measure* if the total mass $\mu(\mathbb{Z}) < +\infty$. Otherwise, μ is not finite.

We will encounter examples of finite and non-finite measures in the next section.

1.4 Livestock

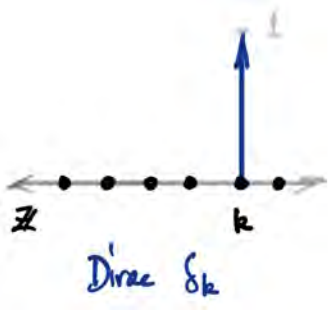
You cannot run a ranch without any cattle. Likewise, you need to become acquainted with examples of measures if you want to develop an operational understanding of measure theory. Furthermore, you need to become familiar with the methods for transforming measures to obtain new measures.

1.4.1 Basic examples

Here are some of the basic classes of measures on the integers.

Example 1.14 (Dirac measure). Let us exhibit a family of measures that are simple but very important. For a point $k \in \mathbb{Z}$, the *Dirac measure* δ_k is defined for each set $\mathsf{A} \subseteq \mathbb{Z}$ via the rule

$$\delta_k(\mathsf{A}) := \mathbb{1}_{\mathsf{A}}(k) := \begin{cases} 1, & k \in \mathsf{A}; \\ 0, & k \notin \mathsf{A}. \end{cases} \tag{1.10}$$



The measure δ_k is often called the *point mass* at k. We illustrate δ_k as a "spike" at the point k with height 1. The Dirac measure is obviously a finite measure. ∎

Exercise 1.15 (Weights and measures). Consider a sequence $(w_i : i \in \mathbb{Z})$ of positive numbers that may be infinite; that is, $w_i \in [0, +\infty]$ for each i. Define

$$\mu(\mathsf{A}) := \sum\nolimits_{i \in \mathsf{A}} w_i \quad \text{for } \mathsf{A} \subseteq \mathbb{Z}.$$

Verify that μ is a measure on the integers. Deduce that μ is the unique measure on the integers with $\mu(\{i\}) = w_i$ for each i. Under what conditions is μ a finite measure?

1.4.2 Measures from measures

Next, let us explore some of the transformation rules that produce new measures from existing measures.

"Haste still pays haste, and leisure answers leisure, / Like doth quit like, and measure still for measure."
—William Shakespeare

Exercise 1.16 (Restriction of measures). Let μ be a measure on the integers, and fix a set $\mathsf{E} \subseteq \mathbb{Z}$. We can define the *restriction* of the measure μ to the set E via

$$\nu(\mathsf{A}) := \mu(\mathsf{A} \cap \mathsf{E}) \quad \text{for } \mathsf{A} \subseteq \mathbb{Z}.$$

Show that ν is a measure. In what circumstances is ν a finite measure?

Example 1.17 (Positive linear combinations). A measure on the integers is a particular type of function that takes extended real values. As a consequence, we can scale a measure by a positive number, and we can add measures. For instance, if μ, ν are measures on the integers and $\alpha, \beta \in \mathbb{R}_+$ are positive numbers, then we can define the measure $\alpha\mu + \beta\nu$ via the rule

$$(\alpha\mu + \beta\nu)(\mathsf{A}) := \alpha\mu(\mathsf{A}) + \beta\nu(\mathsf{A}) \quad \text{for all } \mathsf{A} \subseteq \mathbb{Z}.$$

For example, $\# + \delta_0$ is a measure (what is it?). More generally, we can form a positive linear combination of (a finite number of) measures. ▪

Example 1.18 (*Mapping). Another way to construct new measures is via mapping. Let μ be a measure on the integers, and let $f : \mathbb{Z} \to \mathbb{Z}$ be a function. Then we can define a measure ν, called the *push-forward* of the measure μ by the function f:

$$\nu(\mathsf{A}) := \mu(f^{-1}(\mathsf{A})) \quad \text{for each } \mathsf{A} \subseteq \mathbb{Z}. \tag{1.11}$$

We have written $f^{-1}(\mathsf{A}) := \{k \in \mathbb{Z} : f(k) \in \mathsf{A}\}$ for the *preimage* of the function. It is common to denote the push-forward measure as $\nu = f_*\mu$. ▪

Exercise 1.19 (*Preimages and mapping). Let $f : \mathbb{Z} \to \mathbb{Z}$ be a function. Show that the preimage of a union is the union of the preimages:

$$f^{-1}\big(\bigcup_{i\in\mathsf{I}} \mathsf{A}_i\big) = \bigcup_{i\in\mathsf{I}} f^{-1}(\mathsf{A}_i) \quad \text{where } \mathsf{A}_i \subseteq \mathbb{Z} \text{ for each } i \in \mathsf{I}.$$

This statement is true for every index set I. In fact, it is valid for all functions f, regardless of the domain and codomain.

Deduce that (1.11) defines a measure. Compute $f_*\delta_0$ and $f_*\#$. Are there situations where you can simplify the formulas?

1.5 Specifying a measure on the integers

What information do we need to completely describe a measure on the integers? *A priori*, to specify a measure, we need to provide a rule that delivers the value of the measure on every subset of integers. In some cases, it is natural to define the measure directly on subsets (e.g., the counting measure or the Dirac measure). In other cases, it may be more productive to pursue other representations.

We have already encountered a fully general construction of a measure on the integers in Exercise 1.15. Indeed, we have the following converse result.

Proposition 1.20 (Integer measures: Characterization). Every measure μ on the integers can be written uniquely in the form $\mu(\mathsf{A}) = \sum_{i\in\mathsf{A}} w_i$ where the weights $w_i \in [0, +\infty]$.

Proof. Define $w_i := \mu(\{i\})$ for each $i \in \mathbb{Z}$. For each subset $\mathsf{A} \subseteq \mathbb{Z}$, we can evidently write $\mathsf{A} = \dot{\bigcup}_{i\in\mathsf{A}} \{i\}$. By countable additivity of the measure μ, we have the identity $\mu(\mathsf{A}) = \sum_{i\in\mathsf{A}} \mu(\{i\}) = \sum_{i\in\mathsf{A}} w_i$. ∎

When we study measures on more general domains, however, we may lack an analog of Proposition 1.20. For instance, when the domain is uncountable, we cannot decompose a set into a countable union of singleton sets as we did in the proof. In these situations, we need to find more flexible classes of sets that we can use as building blocks.

Let us see how we might design other kinds of representations for measures in the present context. Suppose that we knew the measure of every interval $(j, k] \cap \mathbb{Z}$ where $j, k \in \mathbb{Z}$. It seems as if this data should be adequate to determine the measure of every set of integers. This intuition is correct.

There is a separate concern, however, that the measures of intervals need to be self-consistent. For instance, we certainly cannot have $\mu(\{0\}) = 2$ and $\mu(\{0, 1\}) = 1$. For *finite* measure on the integers, there is an easy way to guarantee consistency by working with the *distribution function* of the measure.

Definition 1.21 (Distribution function: Integer measure). Let μ be a *finite* measure on the integers. The *(cumulative) distribution function* (abbreviated *cdf* or *df*) of the measure is defined as

$$F(k) := \mu((-\infty, k] \cap \mathbb{Z}) \quad \text{for } k \in \mathbb{Z}.$$

Distribution functions are useful tools for probability theory. We will spend more time with them later on. For now, let us present some results, which state that a distribution function characterizes a measure.

Proposition 1.22 (Distribution function on integers: Properties). Let $F : \mathbb{Z} \to \mathbb{R}_+$ be the distribution function of a *finite* measure μ on the integers. Define $M := \mu(\mathbb{Z}) < +\infty$. Then F enjoys two properties:

1. **Increasing:** If $j \le k$, then $F(j) \le F(k)$.
2. **Asymptotic limits:** We have $\lim_{k \downarrow -\infty} F(k) = 0$ and $\lim_{k \uparrow +\infty} F(k) = M$.

Exercise 1.23 (Distribution function on integers). Prove Proposition 1.22.

Remarkably, the converse is also true. Any function on the integers with these two properties is the distribution function of a unique finite measure.

Theorem 1.24 (Distribution function on integers = finite integer measure). Let $F : \mathbb{Z} \to \mathbb{R}_+$ be a function that satisfies Proposition 1.22(1)–(2). Then there is a unique finite measure μ on the integers with

$$\mu((j, k] \cap \mathbb{Z}) = F(k) - F(j) \quad \text{for all } j, k \in \mathbb{Z} \text{ with } j < k.$$

Proof. Define $w_i := F(i) - F(i-1)$ for each $i \in \mathbb{Z}$. Since F is increasing and finite, the weights w_i are positive. Construct the function

$$\mu(\mathsf{A}) := \sum\nolimits_{i \in \mathsf{A}} w_i \quad \text{for all } \mathsf{A} \subseteq \mathbb{Z}.$$

By a telescope, this function has the advertised property:

$$\mu((j, k] \cap \mathbb{Z}) = \sum\nolimits_{i=j+1}^{k} w_i = F(k) - F(j) \quad \text{for } j, k \in \mathbb{Z} \text{ with } j < k.$$

Taking limits as $j \to -\infty$ and $k \to +\infty$, we quickly determine that the total mass $\mu(\mathbb{Z}) = \sum_{i \in \mathbb{Z}} w_i = M < +\infty$. According to Exercise 1.15, the function μ is indeed a finite measure, which is uniquely determined by F. ∎

Problem 1.25 (*Discrete distribution functions). A measure μ on the integers is *locally finite* if $\mu(\{i\}) < +\infty$ for each $i \in \mathbb{Z}$. Show that locally finite measures on the integers are in one-to-one correspondence with the class of increasing functions $F : \mathbb{Z} \to \mathbb{R}$ that satisfy $F(0) = 0$.

Problems

Exercise 1.26 (Preimage: Set operations). Let $f : \mathsf{X} \to \mathsf{Y}$ be a function. The preimage of a subset of the codomain is defined as

$$f^{-1}(\mathsf{B}) := \{x \in \mathsf{X} : f(x) \in \mathsf{B}\} \subseteq \mathsf{X} \quad \text{for all } \mathsf{B} \subseteq \mathsf{Y}.$$

1. Show that the preimage of a complement is the complement of the preimage:

$$f^{-1}(\mathsf{B}^{\mathsf{c}}) = (f^{-1}(\mathsf{B}))^{\mathsf{c}}.$$

   The complement of a subset of a domain is defined with respect to the domain. For example, if $\mathsf{B} \subseteq \mathsf{Y}$, then $\mathsf{B}^{\mathsf{c}} := \mathsf{Y} \setminus \mathsf{B}$.
2. Verify that the preimage f^{-1} distributes over unions, intersections, and set differences. For example,

$$f^{-1}(\mathsf{B} \cup \mathsf{C}) = f^{-1}(\mathsf{B}) \cup f^{-1}(\mathsf{C}) \quad \text{for all } \mathsf{B}, \mathsf{C} \subseteq \mathsf{Y}.$$

3. For functions f and g with compatible domains, confirm that $(f \circ g)^{-1} = g^{-1} \circ f^{-1}$.

Notes

The material in this lecture is standard, but it is not very common to emphasize measures on the integers because there are more elementary ways to present these ideas. We have chosen this introduction to help build intuition for the concept of a measure in a concrete setting where there are no technicalities.

2. Abstract Measure Spaces

> "[Abstract art is] a product of the untalented, sold by the unprincipled to the utterly bewildered."
>
> —Al Capp

Agenda:
1. Measurable sets
2. Sigma-algebras
3. Measurable spaces
4. Measure spaces
5. Examples
6. How to specify a measure

We introduced the concept of a measure in the friendly territory of the integers. This concrete setting helps us visualize a distribution of mass and intuit some of the properties that a distribution of mass should have.

But the ambitious reader may wish to distribute mass over domains that are more general than the integers ($\mathbb{Z}$). The most important case is the real line ($\mathbb{R}$). Other related examples include the plane ($\mathbb{R}^2$) and ordinary space ($\mathbb{R}^3$). Later, when we enter into probability theory, we will want to assign likelihoods to the outcomes of general probability experiments (e.g., coins, dice, card games, the Kentucky Derby...).

Although mathematicians are sometime accused of an excessive love of abstraction, these examples intimate that we really do need to consider distributions of mass over general domains. Abstraction allows us to isolate what is truly essential. We will see the power of this approach in the next lecture when we discuss measures on the real line. Later, it will provide a clean setting for developing probability theory.

The main goal of this lecture is to introduce a foundational concept from measure theory, the notion of a σ-algebra of sets. After exploring this important idea, we will give a general definition of a measure and present some examples. We dedicate the next lecture to the study of measures on the real line.

2.1 Measurable sets

One of the core ideas behind measure theory is that we assign mass to subsets of the domain, rather than to points. This raises the question: *Which subsets?*

2.1.1 Why is this an issue?

We ducked the question when we introduced measures on the integers $\mathbb{Z}$. Indeed, we defined these measures on the power set $\mathcal{P}(\mathbb{Z})$, which contains *all subsets* of the integers. This approach is successful because the integers compose a countable set, and so every subset of $\mathbb{Z}$ is also countable. Therefore, we can use countable additivity to define a measure on singleton sets and to extend it to all subsets. (See Proposition 1.20.)

Recall that a countable set is in one-to-one correspondence with a subset of the natural numbers. An *uncountable* set is not countable.

This program crashes when we try to define measures on uncountable domains, such as the real line $\mathbb{R}$. Roughly speaking, there is no way to break down an arbitrary subset of the real line as a countable union of "nice" subsets. As a consequence, we cannot hope to define a measure on "nice" subsets and extend it to all subsets via countable additivity.

The difficulty is both serious and inexorable. The essential example of a distribution of mass on the real line is a uniform distribution. Under the uniform distribution, the mass of each interval is proportional to its length. In the late 19th century, mathematicians sought a way to define the "length" of an arbitrary subset of the real line. After many failed attempts, they eventually realized that there is no consistent definition of "length" that is valid for every subset of the real line. Shockingly, we cannot break down a general set into pieces that have well-defined lengths. The field of measure theory was invented at the outset of the 20th century as a way to resolve the concept of length. We will discuss the history and the details in the next lecture.

For now, suppose that we want to define a distribution of mass over a domain. We have argued that a measure is a countably additive function that reports the amount of mass carried by subsets of the domain.

> The key idea is that we should only try to define a measure on a "nice" class of subsets of the domain, called *measurable sets*.

Only by ratcheting down our expectations can we design a successful theory.

"If at first you don't succeed, lower your standards."
—*Tommy Boy*, 1995

2.1.2 Measurable sets

Let X be an abstract set, called the *domain*, whose elements are called *points*. The domain is often a familiar environment (such as $\mathbb{Z}$ or $\mathbb{R}$ or $\mathbb{R}^n$), but measure theory does not require the domain to have any extra structure. A measure will describe a distribution of mass over the domain X.

To that end, we equip the domain with a collection $\mathcal{F} \subseteq \mathcal{P}(\mathsf{X})$ of subsets, called *measurable sets*. Measures only assign mass to measurable sets; other subsets of the domain are out of bounds. The family $\mathcal{F}$ of measurable sets cannot be totally arbitrary, or else we cannot hope to define measures on the family $\mathcal{F}$. Rather, the family $\mathcal{F}$ must be stable under certain set operations. These set operations ensure that it makes sense to define a countably additive function on $\mathcal{F}$.

In this context, some authors use the word *closed* instead of the word *stable*.

2.2 Measurable spaces

In this section, we rigorously define the concept of a *measurable space*, the arena where measure theory takes place. A measurable space involves a domain and a collection of measurable sets. Our first task is to develop the mathematical framework for describing the measurable sets.

2.2.1 Sigma-algebras

Let us introduce the key definition underlying the construction of measurable sets. This object is called a σ-algebra. It provides the scaffolding for abstract measure theory, and we cannot overstate its importance for this class.

The prefix σ means "countable."

Definition 2.1 (Sigma-algebra of sets). Let X be a domain. A family $\mathcal{F} \subseteq \mathcal{P}(\mathsf{X})$ of subsets of X is called a *σ-algebra* on the domain X if it satisfies three properties.

1. **Nothing and everything:** The empty set $\emptyset$ and the domain X belong to $\mathcal{F}$.
2. **Complements:** If a set $\mathsf{A} \in \mathcal{F}$, then its complement $\mathsf{A}^c := \mathsf{X} \setminus \mathsf{A}$ belongs to $\mathcal{F}$.
3. **Countable unions and intersections:** The family is stable under *countable* unions

Some authors use the term *σ-field* instead of σ-algebra. The latter is more better; see Problem 2.49.

and intersections:

$$A_i \in \mathcal{F} \text{ for } i \in \mathbb{N} \quad \text{implies that} \quad \bigcup_{i=1}^{\infty} A_i \in \mathcal{F} \quad \text{and} \quad \bigcap_{i=1}^{\infty} A_i \in \mathcal{F}.$$

Exercise 2.2 (Sigma-algebra: Minimal definition). Some of the requirements in Definition 2.1 are redundant. Which ones can be removed?

Here are some particular examples of σ-algebras.

Example 2.3 (Trivial σ-algebra). The family $\mathcal{F} = \{\emptyset, X\}$ is a σ-algebra on X. ▪

Example 2.4 (Almost trivial σ-algebra). Fix a set $A \subseteq X$. Then $\mathcal{F} = \{\emptyset, A, A^c, X\}$ is a σ-algebra on X. ▪

Example 2.5 (Complete σ-algebra). The power set $\mathcal{P}(X)$ is a σ-algebra on X. ▪

As we will see, the complete σ-algebra appears in the construction of measures on the integers. We will also encounter the Borel σ-algebra, a more complicated object that is essential for constructing measures on the real line.

> **Aside:** Why not allow *uncountable* unions? That is a bridge too far. If we required the measurable sets to be stable under uncountable unions, then we might end up with so many measurable sets that we could not define measures consistently. Sigma-algebras are designed to cooperate with the countable additivity of measures.

2.2.2 Algebraic properties of sigma-algebras

By construction, a σ-algebra is stable under complements, countable unions, and countable intersections. We can deduce some additional stability properties directly from the definition.

Exercise 2.6 (Sigma-algebra: Differences). Show that a σ-algebra is stable under set differences. For sets A, B in the σ-algebra, the sets $A \setminus B$ and $B \setminus A$ and $A \triangle B$ also belong to the σ-algebra.

The intersection of two σ-algebras is always a σ-algebra. This fact and its generalizations play an important role in constructions of σ-algebras.

Exercise 2.7 (Sigma-algebra: Intersections). Let $\mathcal{F}$ and $\mathcal{G}$ be σ-algebras on X. Show that $\mathcal{F} \cap \mathcal{G}$ is also a σ-algebra on X. Argue that the intersection of an arbitrary family of σ-algebras on X remains a σ-algebra.

The intersection of two collections of sets contains exactly those sets that appear in both collections.

Exercise 2.8 (Sigma-algebra: Restriction). Let $\mathcal{F}$ be a σ-algebra on a domain X, and let $E \in \mathcal{F}$ be an element of the σ-algebra. Define the *restriction* of the σ-algebra to E:

$$\mathcal{F}|_E := \{E \cap F : F \in \mathcal{F}\}.$$

Show that $\mathcal{F}|_E$ is a σ-algebra on E.

2.2.3 Generation of sigma-algebras

A collection of subsets of the domain may or may not be a σ-algebra. Nevertheless, we can always construct a minimal σ-algebra that contains the collection.

Conceptually, it seems that we might want to repeatedly add subsets that are missing and stop as soon as arrive at a σ-algebra. This approach, however, is hard to make rigorous. Instead, we will begin with the complete σ-algebra and remove as many subsets as possible, keeping the initial family of subsets and retaining the σ-algebra property.

Definition 2.9 (σ-algebra: Generation). Let $\mathcal{S} \subseteq \mathcal{P}(\mathsf{X})$ be a collection of subsets of X. The family $\mathcal{S}$ generates a unique minimal σ-algebra:

$$\sigma(\mathcal{S}; \mathsf{X}) := \big\{\mathsf{A} \subseteq \mathsf{X} : \mathsf{A} \text{ belongs to every } \sigma\text{-algebra } \mathcal{F} \text{ on } \mathsf{X} \text{ with } \mathcal{S} \subseteq \mathcal{F}\big\}.$$

We often call $\sigma(\mathcal{S}; \mathsf{X})$ the *smallest* σ-algebra on X that contains $\mathcal{S}$. The domain X is omitted from the notation when it is clear from context: $\sigma(\mathcal{S}) := \sigma(\mathcal{S}; \mathsf{X})$.

Similar constructions arise throughout mathematics. For example, the span of a set of vectors is the smallest linear subspace that contains those vectors. Likewise, the convex hull is the smallest convex set that contains a given set.

It remains to verify that Definition 2.9 actually produces a minimal σ-algebra.

Proposition 2.10 (Generated σ-algebras). Let $\mathcal{S} \subseteq \mathcal{P}(\mathsf{X})$ be a collection of subsets of X. Then $\sigma(\mathcal{S}; \mathsf{X})$ is a σ-algebra on X. Moreover, if $\mathcal{S} \subseteq \mathcal{F}$ for another σ-algebra $\mathcal{F}$ on X, then $\sigma(\mathcal{S}; \mathsf{X}) \subseteq \mathcal{F}$.

Proof. We can interpret the definition of the generated σ-algebra $\sigma(\mathcal{S}; \mathsf{X})$ as the intersection of all σ-algebras on X that contain the distinguished collection $\mathcal{S}$ of sets. The intersection is nonempty because the power set $\mathcal{P}(\mathsf{X})$ is a σ-algebra that contains $\mathcal{S}$. By Exercise 2.7, the intersection of σ-algebras remains a σ-algebra.

Finally, since $\sigma(\mathcal{S})$ is the intersection of all σ-algebras that contain $\mathcal{S}$, it must be the case that $\sigma(\mathcal{S})$ is a subset of any particular σ-algebra that contains $\mathcal{S}$. ∎

Here are some basic examples of generated σ-algebras.

Example 2.11 (Almost trivial σ-algebra). Let A be a subset of the domain X. Then $\sigma(\{\mathsf{A}\}) = \{\emptyset, \mathsf{A}, \mathsf{A}^{\mathsf{c}}, \mathsf{X}\}$.

To see why, observe that this is the minimal list of sets that must appear in $\sigma(\{\mathsf{A}\})$. Indeed, every σ-algebra contains nothing ($\emptyset$) and everything (X). The generated σ-algebra contains the set A, so it also contains the complement A^{c}. These four sets already compose a σ-algebra, so they form the smallest σ-algebra generated by A. ▪

Exercise 2.12 (Small σ-algebras). Let $\mathsf{A}, \mathsf{B} \subseteq \mathsf{X}$. What is $\sigma(\{\mathsf{A}, \mathsf{B}\})$?

Exercise 2.13 (Countable domains). Let X be a *countable* domain. Show that the singleton sets generate the complete σ-algebra. That is, $\sigma(\{\{x\} : x \in \mathsf{X}\}) = \mathcal{P}(\mathsf{X})$.

Problem 2.14 (Uncountable domains).** Let X be an *uncountable* domain. Show that the singleton sets do not generate the complete σ-algebra. That is, $\sigma(\{\{x\} : x \in \mathsf{X}\}) \subsetneq \mathcal{P}(\mathsf{X})$.

> **Aside:** The minimality property of a generated σ-algebra also provides a versatile theoretical tool. Later on, we will use this fact to verify that particular set collections are σ-algebras. This argument arises in the development of the integral and again in the construction of product measures.

2.2.4 Borel sigma-algebras

Generated σ-algebras are a powerful mechanism for building classes of measurable sets that contain specific families of "elementary sets." An important example is the Borel σ-algebra on the real line, which will play a central role in the definition of measures on the real line.

Example 2.15 (Borel σ-algebra on $\mathbb{R}$). On the real line $\mathbb{R}$, we may regard open intervals (a, b) as a class of elementary sets. We define the Borel σ-algebra on $\mathbb{R}$ as

$$\mathcal{B}(\mathbb{R}) := \sigma(\{(a, b) : a < b \text{ and } a, b \in \mathbb{R}\}).$$

As it happens, the Borel σ-algebra also contains all open subsets of the real line because every open set in the line is a countable union of open intervals. ▪

We can extend this example to a much wider setting. Here are two cases of particular importance in probability theory.

Example 2.16 (Borel σ-algebra on $\mathbb{R}^n$). In the Euclidean space $\mathbb{R}^n$, we may regard open Euclidean balls as a class of elementary sets. For reference, the open Euclidean balls are the sets

$$\mathsf{D}(\boldsymbol{x}; r) := \{\boldsymbol{y} \in \mathbb{R}^n : \|\boldsymbol{y} - \boldsymbol{x}\|_2 < r\}.$$

Recall that $\|\cdot\|_2$ is the ordinary Euclidean norm on $\mathbb{R}^n$.

We define the Borel σ-algebra on $\mathbb{R}^n$ as

$$\mathscr{B}(\mathbb{R}^n) := \sigma(\{\mathsf{D}(\boldsymbol{x}; r) : \boldsymbol{x} \in \mathbb{R}^n \text{ and } r > 0\}),$$

As before, the Borel σ-algebra on $\mathbb{R}^n$ contains all the open subsets of $\mathbb{R}^n$. ▪

Example 2.17 (*Borel σ-algebra on a metric space). Let $(\mathsf{X}, \mathrm{dist})$ be a *separable* metric space. The Borel σ-algebra on X is defined as

A separable metric space contains a countable subset that is dense in the whole space.

$$\mathscr{B}(\mathsf{X}) := \sigma(\{\mathsf{D}(\boldsymbol{x}; r) : \boldsymbol{x} \in \mathsf{X} \text{ and } r > 0\}),$$

where the open ball $\mathsf{D}(\boldsymbol{x}; r) := \{\boldsymbol{y} \in \mathsf{X} : \mathrm{dist}(\boldsymbol{y}, \boldsymbol{x}) < r\}$. As in the previous examples, the Borel σ-algebra on X contains all the open subsets of X. This claim requires the separability assumption. ▪

2.2.5 Measurable spaces

To summarize, we have introduced the concept of a σ-algebra, which is a collection of sets that includes the empty set and that is stable under complements, countable unions, and countable intersections. We may now describe the stage where measure theory plays out.

Definition 2.18 (Measurable space). Let X be a domain equipped with a σ-algebra $\mathscr{F}$. The pair $(\mathsf{X}, \mathscr{F})$ is called a *measurable space*. In this context, the elements of $\mathscr{F}$ are called *measurable sets* or *$\mathscr{F}$-measurable sets*.

Here are some simple examples of measurable spaces that frequently arise.

Example 2.19 (The trivial measurable space). The pair $(\mathsf{X}, \{\emptyset, \mathsf{X}\})$ is a measurable space. The only measurable sets are the empty set and the whole domain. As we will see, this trivial example plays a role in probability theory. ▪

Example 2.20 (Finite measurable spaces). Let X be a finite set. Then $(\mathsf{X}, \mathscr{P}(\mathsf{X}))$ is a measurable space. ▪

Example 2.21 (A countable measurable space). The space $(\mathbb{Z}, \mathscr{P}(\mathbb{Z}))$ is a measurable space. The measurable sets consist of all subsets of the integers. This is the measurable space where we define measures on the integers, as in Lecture 1. ▪

Example 2.22 (A multivariate measurable space). The space $(\mathbb{N}^2, \mathscr{P}(\mathbb{N}^2))$ is a measurable space. The measurable sets consist of all subsets of pairs of natural numbers. ▪

Example 2.23 (The complete measurable space). In general, the space $(\mathsf{X}, \mathscr{P}(\mathsf{X}))$ is a measurable space where every subset of X is measurable. As we have seen, this construction is useful when X is countable. On the other hand, when X is uncountable, the power set contains too many measurable sets for us to build a successful theory. ▪

Example 2.24 (The real line with its Borel sets). The pair $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ is a measurable space. This construction will allow us to rigorously define distributions of mass over the real line, including the uniform distribution of mass. We will wait until the next lecture to elaborate on this example. ▪

Warning 2.25 ("Measurable"). The data of a measurable space does *not* include a measure. In a moment, we will define a measure to be a type of function whose domain consists of measurable sets. In other words, the term "measurable" refers to a potentiality. ▪

2.3 Abstract measures

We may now introduce the notion of an abstract measure, which is a function that reports the mass of a measurable set. This section presents the formal definition, recounts the basic properties and examples, and introduces some additional terminology.

2.3.1 Measures

To reiterate, a measure is a function that reports the amount of mass carried by each measurable set. As in the case of measures on the integers, we require that the measure is countably additive.

Definition 2.26 (Measure). Let $(\mathsf{X}, \mathcal{F})$ be a measurable space. A *(countably additive) measure* is a function $\mu : \mathcal{F} \to [0, +\infty]$ that satisfies two properties:

1. **Empty set:** $\mu(\emptyset) = 0$.
2. **Countable additivity:** Let $(\mathsf{A}_i \in \mathcal{F} : i \in \mathbb{N})$ be a *disjoint* sequence of *measurable* sets. Then
$$\mu\big(\dot{\bigcup}_{i=1}^{\infty} \mathsf{A}_i\big) = \sum_{i=1}^{\infty} \mu(\mathsf{A}_i). \tag{2.1}$$

Warning: See Warning 1.8! ▪

Definition 2.27 (Measure space). A *measure space* is a triple $(\mathsf{X}, \mathcal{F}, \mu)$ where X is a domain, $\mathcal{F}$ is a σ-algebra on X, and μ is a measure on $\mathcal{F}$.

Warning: In contrast to a measurable space, the data in a measure space includes a measure. ▪

We have already seen the importance of countable additivity in the construction of measures on the integers (Proposition 1.20). Countable additivity will be critical for taking limits and for designing a theory of integration.

Problem 2.28 (*Finite additivity). Let $(\mathsf{X}, \mathcal{F})$ be a measurable space. A function $\mu_0 : \mathcal{F} \to [0, +\infty]$ is *finitely additive* if it satisfies

$$\mu_0\big(\dot{\bigcup}_{i=1}^{n} \mathsf{A}_i\big) = \sum_{i=1}^{n} \mu_0(\mathsf{A}_i) \quad \text{for all } \textit{disjoint } \mathsf{A}_i \in \mathcal{F} \text{ and } n \in \mathbb{N}. \tag{2.2}$$

Confirm that a measure μ on $(\mathsf{X}, \mathcal{F})$ is always finitely additive.

Show that the finite additivity property (2.2) follows from the simpler condition

$$\mu_0(\mathsf{A} \dot{\cup} \mathsf{B}) = \mu_0(\mathsf{A}) + \mu_0(\mathsf{B}) \quad \text{for } \textit{disjoint } \mathsf{A}, \mathsf{B} \in \mathcal{F}.$$

Why is this assumption inadequate to establish countable addivity (2.1)?

Assume that X is infinite. Exhibit a finitely additive function μ_0 with $\mu_0(\emptyset) = 0$ that is not a measure. Thus, we must enforce countable additivity if we want to use it.

> **Aside:** Why not *uncountable* additivity? You cannot add up an uncountable number of positive quantities unless there are only countably many nonzero terms, so there is no sensible notion of uncountable additivity.

2.3.2 Basic properties

Abstract measures satisfy the same properties that we established for integer measures. The proofs are exactly the same. Let us set these statements down for reference.

Proposition 2.29 (Measure: Properties). Let $(X, \mathcal{F}, \mu)$ be a measure space.

1. **Monotonicity:** For nested $\mathcal{F}$-measurable sets $A \subseteq B$, we have $\mu(A) \leq \mu(B)$.
2. **Inclusion–exclusion:** For all measurable sets $A, B \in \mathcal{F}$,
$$\mu(A \cup B) + \mu(A \cap B) = \mu(A) + \mu(B).$$
3. **Countable subadditivity:** Consider a countable sequence $(A_i : i \in \mathbb{N})$ of measurable sets, not necessarily disjoint. Then
$$\mu\big(\bigcup_{i=1}^{\infty} A_i\big) \leq \sum_{i=1}^{\infty} \mu(A_i). \tag{2.3}$$

Proposition 2.30 (Measure: Monotone limits). Let $(X, \mathcal{F}, \mu)$ be a measure space.

1. **Increasing limits:** For an increasing sequence $A_1 \subseteq A_2 \subseteq A_3 \subseteq \cdots$ of measurable sets,
$$\mu\big(\bigcup_{i=1}^{\infty} A_i\big) = \lim_{i\to\infty} \mu(A_i).$$
2. **Decreasing limits:** For a decreasing sequence $A_1 \supseteq A_2 \supseteq A_3 \supseteq \cdots$ of measurable sets,
$$\mu\big(\bigcap_{i=1}^{\infty} A_i\big) = \lim_{i\to\infty} \mu(A_i),$$
provided that $\mu(A_i) < +\infty$ for some index i.

Problem 2.31 (*Measure: Continuity at zero). Let $(X, \mathcal{F})$ be a measurable space, and let $\mu_0 : \mathcal{F} \to [0, +\infty]$ be a finitely additive function, as in (2.2). Prove that μ_0 is a measure if and only if μ_0 satisfies the condition
$$\mu_0(A_i) \downarrow 0 \quad \text{when} \quad A_i \downarrow \emptyset \quad \text{and} \quad \mu_0(A_i) < +\infty \text{ for some } i.$$

That is to say, a measure is continuous at zero. This property may feel more intuitive than countable additivity.

2.3.3 Finite measures

Measures that we encounter in everyday life have either a finite amount of mass, or they have an infinite amount of mass that is nicely distributed. We will give some examples in the next subsection.

Definition 2.32 (Finite; σ-finite). Let $(X, \mathcal{F}, \mu)$ be a measure space.

- **Finite measure:** If the total mass $\mu(X) < +\infty$, then we say that μ is a *finite measure*.
- **Sigma-finite measure:** We say that μ is a *σ-finite measure* if we can cover X by countably many measurable sets A_i, each with finite measure. That is,
$$\bigcup_{i=1}^{\infty} A_i = X \quad \text{and} \quad \mu(A_i) < +\infty \quad \text{for each } i \in \mathbb{N}.$$

There is a special case of particular importance for us.

Definition 2.33 (Probability measure). A finite measure μ with total mass $\mu(\mathsf{X}) = 1$ is called a *probability measure*. It describes a distribution of probability mass over the domain X.

Aside: Measures that are not σ-finite can exhibit counterintuitive behavior. We will often exclude them from consideration.

2.3.4 Examples of measures

We may now explore some of the basic examples of measures.

Example 2.34 (Dirac measure). Let $(\mathsf{X}, \mathcal{F})$ be a measurable space. Let $t \in \mathsf{X}$ be a fixed point. The *Dirac measure* concentrated at t is given by

$$\delta_t(\mathsf{A}) := \mathbb{1}_{\mathsf{A}}(t) := \begin{cases} 1, & t \in \mathsf{A}; \\ 0, & t \notin \mathsf{A} \end{cases} \quad \text{for } \mathsf{A} \in \mathcal{F}.$$

This measure is also called the *point mass* at t. Clearly, δ_t is a probability measure. ▪

Example 2.35 (Counting measure). Let $(\mathsf{X}, \mathcal{F})$ be a measurable space. The *counting measure* $\#$ is defined as

$$\#\mathsf{A} := \begin{cases} \operatorname{card}(\mathsf{A}), & \mathsf{A} \text{ is finite}; \\ +\infty, & \text{otherwise} \end{cases} \quad \text{for } \mathsf{A} \in \mathcal{F}.$$

This measure reports the number of points in a measurable set. If X is finite, then $\#$ is a finite measure. If X is countable, then $\#$ is a σ-finite measure. On the other hand, if X is uncountable, then $\#$ is *not* σ-finite. ▪

Example 2.36 (Uniform measure on a finite set). Let X be a *finite* set. Then $(\mathsf{X}, \mathcal{P}(\mathsf{X}))$ is a measurable space. We can define a measure μ where

$$\mu(\mathsf{A}) := \frac{\#\mathsf{A}}{\#\mathsf{X}} \quad \text{for each } \mathsf{A} \subseteq \mathsf{X}.$$

This is called the *uniform measure* on X. It is clearly a probability measure. ▪

Example 2.37 (Measures on the integers). Consider the measurable space $(\mathbb{Z}, \mathcal{P}(\mathbb{Z}))$. Fix a sequence $(w_i : i \in \mathbb{N})$ of positive numbers that may be infinite. We can define a measure

$$\mu(\mathsf{A}) := \sum\nolimits_{i \in \mathsf{A}} w_i \quad \text{for each } \mathsf{A} \subseteq \mathbb{Z}.$$

In other words, we add up the masses w_i for the indices i that appear in the set A. The measure μ is finite if and only if $\sum_{i \in \mathbb{Z}} w_i < +\infty$. The measure μ is σ-finite if and only if $w_i < +\infty$ for all i. In what circumstances is μ a probability measure? ▪

Exercise 2.38 (Measures on pairs of integers). Consider the measurable space $(\mathbb{Z}^2, \mathcal{P}(\mathbb{Z}^2))$. Show that measures are in one-to-one correspondence with functions $\boldsymbol{w} : \mathbb{Z}^2 \to [0, +\infty]$.

We can also define measures on domains with smaller σ-algebras. Let us give an explicit example to show that it is not necessary for a measure to be defined on all subsets of a domain.

Example 2.39 (Measures on the trivial σ-algebra). Consider a domain X equipped with the trivial σ-algebra $\mathcal{F} = \{\emptyset, \mathsf{X}\}$. There is a one-parameter family of measures defined on this σ-algebra. For a positive $\alpha \in [0, +\infty]$, these measures take the form

$$\mu(\emptyset) = 0 \quad \text{and} \quad \mu(\mathsf{X}) = \alpha.$$

You can easily check that μ is countably additive on $\mathcal{F}$. It is not defined on any nontrivial subset of the domain. ▪

Exercise 2.40 (Measures on small σ-algebras). Consider the domain X equipped with the almost trivial σ-algebra $\mathcal{F} := \{\emptyset, \mathsf{A}, \mathsf{A}^c, \mathsf{X}\}$ for a set $\mathsf{A} \subseteq \mathsf{X}$. Describe all of the measures defined on the measurable space $(\mathsf{X}, \mathcal{F})$.

Now, consider the σ-algebra $\mathcal{F} := \sigma(\{\mathsf{A}, \mathsf{B}\})$ generated by two sets $\mathsf{A}, \mathsf{B} \subseteq \mathsf{X}$. Describe all of the measures defined on $(\mathsf{X}, \mathcal{F})$. For simplicity, you may want to consider the case of finite measures.

2.3.5 Measures from measures

As before, we can obtain new measures from old measures via the following transformations.

Exercise 2.41 (Restriction). Let $(\mathsf{X}, \mathcal{F}, \mu)$ be a measure space, and let E be a measurable set. Define the restriction of μ to E:

$$\nu(\mathsf{A}) := \mu(\mathsf{A} \cap \mathsf{E}) \quad \text{for measurable } \mathsf{A} \in \mathcal{F}.$$

Confirm that the restriction is a measure on $(\mathsf{X}, \mathcal{F})$.

Exercise 2.42 (Positive combinations). Let $(\mathsf{X}, \mathcal{F})$ be a measurable space equipped with two measures μ and ν. For positive scalars $\alpha, \beta \geq 0$, define the function

$$(\alpha\mu + \beta\nu)(\mathsf{A}) := \alpha\mu(\mathsf{A}) + \beta\nu(\mathsf{B}) \quad \text{for measurable } \mathsf{A} \in \mathcal{F}.$$

Check that the positive combination is a measure on $(\mathsf{X}, \mathcal{F})$.

> **Aside:** You may notice that the push-forward $f_*\mu$ has disappeared from this list. The reason is that the construction requires further assumptions on the function f. We will turn back to this matter when we develop a theory of integration.

2.3.6 Negligible sets and almost-everywhere sets

We continue with a few more general definitions about measures. Let us introduce some important terminology for sets that carry no mass or whose complement carries no mass.

Definition 2.43 (Negligible; almost everywhere). Let $(\mathsf{X}, \mathcal{F}, \mu)$ be a measure space.

- **Negligible sets:** A measurable set A is called a *negligible set* for the measure μ when $\mu(\mathsf{A}) = 0$.
- **Almost everywhere sets:** We say that a measurable set A is *μ-almost everywhere* when its *complement* is a μ-negligible set: $\mu(\mathsf{A}^c) = 0$. You will often see the abbreviations *μ-a.e.* or just *a.e.*

The term *null set* is more common, but less informative, than the term negligible set.

> **Warning:** These concepts depend on the measure! ▪

Let us mention one of the major use cases for this definition. Consider a measure μ and two functions $f, g : \mathsf{X} \to \mathsf{Y}$. We say that f and g are *equal μ-almost everywhere* if

$$\mu(\{x \in \mathsf{X} : f(x) \neq g(x)\}) = 0.$$

As we will see, functions that are equal almost everywhere often behave as if they were the same function. The careful reader will realize that we must make further demands on the functions f, g to be sure that the points where the functions differ compose a measurable set. We will return to this issue when we develop a theory of integration.

Exercise 2.44 (Dirac almost everywhere). Consider the measurable space $(\mathbb{Z}, \mathcal{P}(\mathbb{Z}))$. Let δ_0 be a Dirac measure at zero. Describe all negligible sets for δ_0. Describe all the almost everywhere sets for δ_0.

Exercise 2.45 (Negligible sets: Countable union). For each $i \in \mathbb{N}$, let A_i be a negligible set for a measure μ. Show that $\bigcup_{i=1}^{\infty} \mathsf{A}_i$ is a negligible set for μ.

2.3.7 Atoms

As we have seen, measures may concentrate mass on a point. There is special terminology for describing this situation.

Definition 2.46 (Atom). Let $(\mathsf{X}, \mathcal{F}, \mu)$ be a measure space. We say that μ has an *atom* at the point $x \in \mathsf{X}$ when $\mu(\{x\}) > 0$.

When we illustrate a measure that has an atom at a point, we use a spike to indicate the location and strength of the atom. All of the measures we have considered so far have atoms.

In the next lecture, we will introduce measures on the real line. For example, the Lebesgue measure models a uniform distribution of mass, with no mass concentrated at any point. In other words, the Lebesgue measure has no atoms, in contrast to the examples we have studied so far.

2.4 How do we construct a measure?

What information do we need to completely describe a measure? As before, we may give a rule that specifies the value of the measure on every measurable set. This is straightforward for examples like the Dirac measure or the counting measure. Beyond that, to define a measure on a countable domain, we can enumerate the values that the measure assigns to each singleton set and invoke countable additivity to extend the measure to all subsets. (See Proposition 1.20.)

In more general measurable spaces, however, life is hard. The measurable sets may be very complicated, and they may not have any explicit description. What do we do in these cases?

In many situations, we can begin with a small family of *elementary sets* that are easy to describe (e.g., open intervals of the real line). The measurable sets are obtained as the smallest σ-algebra that contains all of the elementary sets. We can try to construct a measure by specifying its value on elementary sets (e.g., the lengths of the intervals) and then extending the partial definition to the entire family of measurable sets. To execute this program, we must also ensure that the partial data is consistent with a unique measure.

This approach is called *measure extension*. Appendix A presents the statement and proof of the Hahn–Kolmogorov theorem, a foundational result on measure extension. This theorem is the main tool we use to verify that there exist measures that meet various desiderata. Measure extension theorems are rarely needed for workaday applications of measure theory. Theoretically minded readers will want to understand how measures are constructed, but most users will not need to explore the guts of this machinery.

Problems

Exercise 2.47 (Indicators). Subsets of a domain are in one-to-one correspondence with indicator functions. Each $A \subseteq X$ is associated with the 0–1 indicator function $\mathbb{1}_A : X \to \mathbb{R}$ defined by $\mathbb{1}_A(x) = 1$ when $x \in A$ and $\mathbb{1}_A(x) = 0$ when $x \notin A$. Set operations have algebraic analogs for indicator functions. This observation is useful because it is often easier to reason about indicators than about sets.

1. Show that $\mathbb{1}_{A\cap B} = \min\{\mathbb{1}_A, \mathbb{1}_B\} = \mathbb{1}_A\mathbb{1}_B$. Find similar formulas that express the union, the complement with respect to X, the set difference, and the symmetric difference in terms of indicators. How can you represent subset and superset relations with indicators?
2. **(Inclusion–exclusion).** For arbitrary sets $A, B \subseteq X$, use indicator calculus to show that
$$\mathbb{1}_{A\cup B} = \mathbb{1}_A + \mathbb{1}_B - \mathbb{1}_{A\cap B}.$$
3. **(Deep inclusion–exclusion).** For arbitrary sets $A_1, \dots, A_n \subseteq X$, show that
$$\mathbb{1}_{\bigcup_{i=1}^n A_i} = \sum_{k=1}^n (-1)^{k+1} \sum_{i_1<i_2<\cdots<i_k} \mathbb{1}_{A_{i_1}\cap\cdots\cap A_{i_k}}.$$
In the sum, each index i_j takes values $1, \dots, n$, and the indices $i_1, i_2, \dots, i_k$ are strictly increasing. **Hint:** Take the complement of the union.
4. **(Set limits).** For $i \in \mathbb{N}$, let $A_i \subseteq \mathbb{R}$ be sets. Define the set limit superior and limit inferior as
$$\overline{A} := \limsup_{i\to\infty} A_i \qquad \text{via} \qquad \mathbb{1}_{\overline{A}} = \limsup_{i\to\infty} \mathbb{1}_{A_i};$$
$$\underline{A} := \liminf_{i\to\infty} A_i \qquad \text{via} \qquad \mathbb{1}_{\underline{A}} = \liminf_{i\to\infty} \mathbb{1}_{A_i}.$$
Using indicator calculus, express the set limits using only set operations. One of these sets is interpreted as "points that appear in an infinite number of the sets A_i" and one of these sets is interpreted as "points that eventually appear in a set A_i." Which is which?
5. Use indicator calculus to give short proofs of the following set identities. For all sets A, B, E, F,
$$(A \cup B)\triangle(E \cup F) \subseteq (A\triangle E) \cup (B\triangle F);$$
$$(A \cap B)\triangle(E \cap F) \subseteq (A\triangle E) \cup (B\triangle F).$$
These identities play a role in the proof of the Hahn–Kolmogorov theorem (Theorem A.12).

Exercise 2.48 (More inclusion–exclusion). The simple inclusion–exclusion rule for two sets extends to a more general result. This fact has many applications in set theory and combinatorics.

1. Let $(X, \mathcal{F}, \mu)$ be a measure space. For measurable sets $A_1, \dots, A_n \in \mathcal{F}$, establish the general inclusion–exclusion principle:
$$\mu\big(\textstyle\bigcup_{i=1}^n A_i\big) = \sum_{i=1}^n \mu(A_i) - \sum_{i_1<i_2} \mu(A_{i_1} \cap A_{i_2}) + \sum_{i_1<i_2<i_3} \mu(A_{i_1} \cap \cdots \cap A_{i_3}) - \cdots + (-1)^{n+1}\mu(A_1 \cap \cdots \cap A_n).$$
Each of the sums ranges over indices $i_j = 1, \dots, n$ where the indices $i_1, i_2, \dots, i_k$ are strictly increasing. **Hint:** You can prove this by induction on n.

2. **(**Valuation).** Alternatively, define the function $\mu(\mathbb{1}_A) := \mu(A)$ for measurable sets $A \in \mathcal{F}$. Argue that μ extends to a linear functional on the linear space $\operatorname{lin}\{\mathbb{1}_A : A \in \mathcal{F}\}$, which is a subset of the real-valued functions on X. This is not easy; see Section 5.5.
3. **(Keys).** At a party, each of n guests gives their car keys to the host (don't drink and drive!). How many ways can the keys be returned so that no guest receives the correct set of keys? Can you make an approximation to simplify the result? **Hint:** Choose the sets A_i carefully, and use inclusion–exclusion for the counting measure.

Exercise 2.49 (Set algebras). In this problem, we explore another important set-theoretic notion. Let X be a domain. A family $\mathcal{A} \subseteq \mathcal{P}(X)$ of subsets of X is called a *set algebra* on the domain X if it satisfies three properties:

1. **Nothing and everything:** The empty set $\emptyset$ and the domain X belong to $\mathcal{A}$.
2. **Complements:** If a set $A \in \mathcal{A}$, then its complement $A^c := X \setminus A$ belongs to $\mathcal{A}$.
3. **Unions and intersections:** If $A, B \in \mathcal{A}$, then the union and intersection belong to $\mathcal{A}$:

$$A \cup B \in \mathcal{A} \quad \text{and} \quad A \cap B \in \mathcal{A}.$$

A σ-algebra upgrades the demands on an algebra by requiring stability under *countable* unions and intersections.

1. Some of the requirements in the definition of an algebra are redundant. Which ones can be removed?
2. Show that an algebra is stable under *finite* unions and intersections.
3. Show that a set algebra is stable under set differences.
4. Show that the intersection of an arbitrary family of set algebras on X remains a set algebra on X.
5. **(Generated algebras).** For a family $\mathcal{S} \subseteq \mathcal{P}(X)$, we can construct the smallest algebra on X containing $\mathcal{S}$:

$$\operatorname{algebra}(\mathcal{S}; X) := \{A \subseteq X : A \text{ belongs to every algebra } \mathcal{A} \text{ on } X \text{ with } \mathcal{S} \subseteq \mathcal{A}\}.$$

   Show that $\operatorname{algebra}(\mathcal{S}; X)$ is a well-defined set algebra.
6. **(*Set algebras are algebras).** As in Exercise 2.47, we can pass from the set algebra $\mathcal{A}$ to a collection of indicator functions: $\{\mathbb{1}_A : A \in \mathcal{A}\}$. Equip this collection with the multiplication operation $\mathbb{1}_A \odot \mathbb{1}_B := \mathbb{1}_{A \cap B}$ and the addition operation $\mathbb{1}_A \oplus \mathbb{1}_B := \mathbb{1}_{A \triangle B}$. With these definitions, show that the indicators of a set algebra compose an algebra of functions over the field $\mathbb{F}_2$.
7. **(*Algebras are not always σ-algebras).** Let X be an infinite set. Find an example of an algebra on X that is not a σ-algebra. **Hint:** One easy example is called the *co-finite algebra*.

Notes

You can find the material in this lecture in any book on probability theory or measure theory. For example, see Dudley [Dud02] or Folland [Fol99] for a treatment as part of a real analysis course. For a presentation with a more probabilistic flavor, see Billingsley [Bil12], Pollard [Pol02], or Williams [Wil91].

3. Measures on the Real Line

Agenda:
1. Distributions on the real line
2. Borel sets
3. Borel measures
4. Lebesgue measure
5. Support
6. Specifying a measure

"It is not length of life, but depth of life."

—Ralph Waldo Emerson

"This report, by its very length, defends itself against the risk of being read."

—Winston Churchill

As we have learned, a measure describes a distribution of mass over a domain. It reports the mass that is carried by subsets of the domain. We started with the example of a measure on the integers, which assigns mass to every subset of integers. In more general settings, however, measures may not be defined on all subsets of the domain. Rather, measures are only defined on a family of measurable sets. We introduced the concept of a σ-algebra to formalize the properties of a family of measurable sets. This definition dovetails with the countable additivity property of a measure.

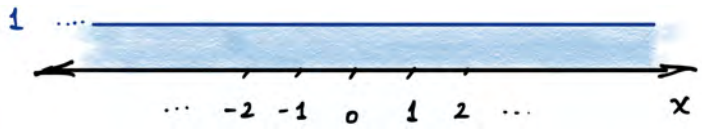
The reason for all of this abstraction will now become clear. In this lecture, we will turn to the problem of defining a distribution of mass over the real line. The archetype of a measure on the real line is the uniform measure, which assigns a "length" to certain subsets of the real line. As it happens, we cannot give a consistent meaning to the concept of "length" for all subsets of the real line. Therefore, we must identify a suitable class of measurable subsets of the real line for which length is meaningful.

To begin our investigation, we will discuss some simple examples of distributions of mass over the real line, and we will present schematics that allow us to visualize these distributions. Afterward, we introduce the Borel class of measurable subsets of the real line, along with the related notion of a Borel measure defined on these sets. This discussion culminates in the construction of the Lebesgue measure, which reports the length of every Borel-measurable set. Afterward, we present more examples of measures on the real line and talk about how we can specify a measure using a distribution function.

3.1 Distributions on the real line

Measure theory allows us to model distributions of mass over very general domains. In particular, we would very much like to describe distributions of mass over the real line. Let us give some examples of how these distributions arise naturally in geometry, mechanics, and probability.

Example 3.1 (Uniform distribution on the real line). The uniform distribution on the real line places mass on the entire line with a constant density of mass per unit length. We can visualize this distribution using a schematic:



The shading indicates that the distribution has a density, and the height of the shaded region reflects the local density of mass at each point. Since the density of mass is constant, the top boundary of the region is a constant function.

In contrast, we used a spike to illustrate mass that is concentrated at a point.

We will construct a measure λ on certain subsets of the real line that models the uniform distribution with a particular normalization. This measure has the distinctive property that

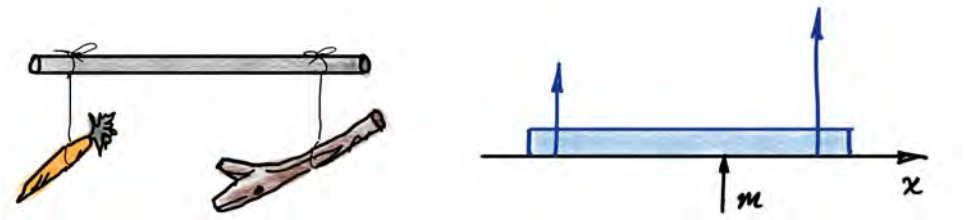
$$\lambda((a,b)) = |b-a| \quad \text{for } a, b \in \mathbb{R} \text{ with } a < b.$$

In other words, the measure λ reports the length of each open interval. More generally, the measure λ assigns a well-determined length to every one of the measurable sets.

The characteristic property of the uniform distribution on the real line is that it is invariant under translation. If we shift the distribution to the left or right, it does not change. Correspondingly, the length of an interval does not change if we shift the interval. ▪

Mathematically, the term *uniform* refers to any distribution that is invariant under an appropriate class of transformations.

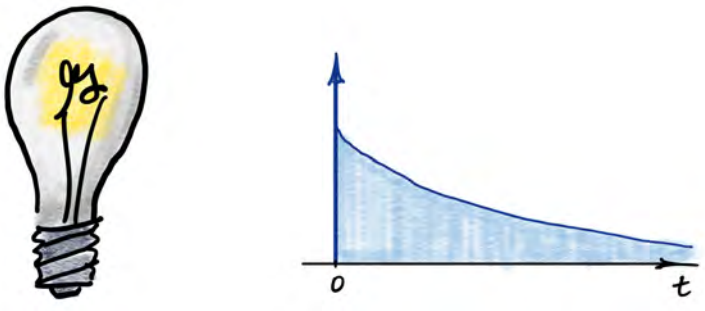
Example 3.2 (Mass). Consider a homogeneous metal rod with weights hanging from the ends. We can model the rod itself as a uniform distribution of mass. (The scaling reflects the density of the rod, which depends on whether it is made of iron, aluminum, etc.) The weights place a positive amount of mass at specific points.



This distribution has both a continuous component (the shaded region) and a discrete component (the spikes). A measure on the real line can handle both of these features with grace.

We have marked the center of mass m of the system. This is the point m about which the total torque is zero. Heuristically, the torque at the position x is the length $(x - m)$ of the lever arm times the "local mass" at the point x. To define the center of mass of a general distribution, we must make sense of these concepts. To answer this challenge, we will develop a theory of integration for general measures. ▪

Example 3.3 (Probability). A probabilistic experiment has a distribution of possible outcomes, some more likely than others. Here is an anodyne example. When we install a new light bulb, we do not know how long it will last before it burns out. There is a strictly positive possibility that it will burn out immediately. If not, the lifetime can be modeled by an exponential distribution. The overall distribution is neither discrete nor continuous.



We can illustrate this probability distribution using a spike to represent the atom at time zero and a shaded region to represent the density for positive times. Our theory of integration will allow us to evaluate the expected lifetime of the lightbulb. ▪

3.2 Borel sets and Borel measures

In this section, we will initiate our construction of measures on the real line. The main challenge is to identify a suitable class of measurable sets. Fortunately, we can make short work of this task using the abstraction of a measurable space.

3.2.1 Intervals

At a minimum, our general definition of measures on the real line must support the construction of a uniform measure. As we have discussed, the uniform distribution on the real line is associated with the concept of length. We may exploit this connection to identify a family of elementary sets that had better be measurable.

Suppose that we want to construct a measure that generalizes the length to a wide class of sets of real numbers. Where should we start? If all is right in the world, the measure should be defined for all open intervals. Indeed,

$$\text{length}(a, b) = |b - a| \quad \text{for } a, b \in \mathbb{R} \text{ with } a < b.$$

In view of this trivial observation, we may regard the open intervals as a class of elementary sets that we absolutely cannot do without. Unfortunately, the open intervals compose a rather small collection of sets. There are many other sets that ought to have a well-defined length. For example, we can obviously assign a length to a finite union of disjoint open intervals.

So what other sets should we include? This seems like a tricky matter. But we can dispose of it efficiently by using the machinery of measurable spaces. We simply form the minimal σ-algebra that contains all of the open intervals. This is a natural setting for defining a uniform measure on the real line.

> **Aside:** Why focus on open intervals? We could just as well start with closed intervals or half-open intervals. But the choice of open intervals lends itself better to generalization, and it is consistent with definitions that we will encounter later.

3.2.2 The collection of Borel sets

We are now prepared to reintroduce a fundamental collection of subsets in the real line. This class will serve as the family of measurable sets that we use to construct measures on the real line.

Definition 3.4 (Borel sets). We define $\mathcal{B}(\mathbb{R})$ to be the *smallest* σ-algebra of subsets of $\mathbb{R}$ that contains all the open intervals. That is,

$$\mathcal{B}(\mathbb{R}) := \sigma(\{(a,b) \subset \mathbb{R} : a, b \in \mathbb{R} \text{ and } a < b\}).$$

The sets in $\mathcal{B}(\mathbb{R})$ are called *Borel-measurable sets* or *Borel sets* in the real line.

The definition of Borel sets is not very explicit, so it is worth taking a moment to investigate what kinds of sets are Borel.

Exercise 3.5 (Borel sets). Verify that the following sets are Borel.

- **The empty set:** The set $\emptyset$ is Borel.
- **The real line:** The set $\mathbb{R}$ is Borel.
- **Set differences:** If A and B are Borel, then $\mathsf{A} \setminus \mathsf{B}$ and $\mathsf{A} \triangle \mathsf{B}$ are Borel.
- **Singletons:** For each $a \in \mathbb{R}$, the set $\{a\}$ is a Borel set. **Hint:** Represent the singleton as a countable intersection of decreasing open intervals.
- **Countable sets:** Show that every countable subset of $\mathbb{R}$ is Borel. In particular, the integers $\mathbb{Z}$ and the rationals $\mathbb{Q}$ are Borel sets.
- **Intervals:** For $a, b \in \mathbb{R}$, the half-open interval $(a, b]$, the closed interval $[a, b]$, and the semi-infinite intervals $(-\infty, a]$ and $(b, +\infty)$ are all Borel.
- **Open sets:** If G is an open subset of $\mathbb{R}$, then G is Borel. **Hint:** Every open set in $\mathbb{R}$ is a countable union of open intervals.
- **Closed sets:** If F is a closed subset of $\mathbb{R}$, then F is Borel. **Hint:** The complement of a closed set is an open set.

Exercise 3.6 (Borel sets: Other generators). Definition 3.4 states that the Borel sets in the real line are generated by the collection of finite open intervals. Show that we can define the Borel sets as the σ-algebra generated by any one of the following classes.

- **Half-open intervals:** $(a, b]$ for $a, b \in \mathbb{R}$.
- **Closed intervals:** $[a, b]$ for $a, b \in \mathbb{R}$.
- **Semi-infinite open intervals:** $(-\infty, a)$ for $a \in \mathbb{R}$.
- **Semi-infinite closed intervals:** $(-\infty, a]$ for $a \in \mathbb{R}$.

Hint: Exercise 3.5 already implies that the Borel sets contains all of these types of intervals. For the reverse direction, you must argue that you can represent open intervals using countable combinations from each of these classes.

In the course of human events, practically every set of real numbers that you encounter will be a Borel set. Nevertheless, you should keep in mind that there are (many!) subsets of the real line that are not Borel.

Warning: There are subsets of $\mathbb{R}$ that are not Borel. ■

Aside: The Borel sets are in one-to-one correspondence with the real numbers $\mathbb{R}$. The power set $\mathcal{P}(\mathbb{R})$, which contains all subsets of real numbers, has strictly larger cardinality than $\mathcal{B}(\mathbb{R})$. Surprisingly, there is a concrete construction of a non-Borel set, due to Lusin. There are easier, but less explicit, constructions of non-Borel sets that require the axiom of choice; see Appendix B.

3.2.3 Extended Borel sets

It is often necessary to work with functions taking extended values, and this fact of life requires us to define an appropriate class of Borel sets.

Definition 3.7 (Extended Borel sets). We define $\mathscr{B}(\overline{\mathbb{R}})$ to be the *smallest* σ-algebra of subsets of $\overline{\mathbb{R}}$ that contains all the open intervals: That is,

$$\mathscr{B}(\overline{\mathbb{R}}) := \sigma(\{(a,b) : a,b \in \overline{\mathbb{R}}\}; \overline{\mathbb{R}}).$$

The sets in $\mathscr{B}(\overline{\mathbb{R}})$ are called *Borel sets* in the extended real line.

Exercise 3.8 (Extended Borel sets). Confirm that $\{-\infty\}$ and $\{+\infty\}$ and $\overline{\mathbb{R}}$ are extended Borel sets. Deduce that every Borel set in $\mathscr{B}(\mathbb{R})$ also belongs to $\mathscr{B}(\overline{\mathbb{R}})$.

3.2.4 Borel measures

Now that we have constructed a σ-algebra of measurable sets in $\mathbb{R}$, we may introduce a class of measures on $\mathbb{R}$.

Definition 3.9 (Borel measure on the real line). Consider the measurable space $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. A measure $\mu : \mathscr{B}(\mathbb{R}) \to [0, +\infty]$ is called a *Borel measure* on $\mathbb{R}$.

Unless otherwise noted, we always equip $\mathbb{R}$ with the class $\mathscr{B}(\mathbb{R})$ of Borel measurable sets, and all measures on $\mathbb{R}$ are understood to be Borel measures. We often omit the qualification "Borel", and we simply refer to measurable subsets of the real line and measures on the real line.

Activity 3.10 (Borel measures). Borel measures on the real line compose a particular class of abstract measures that has special importance. As such, Borel measures enjoy all of the same properties as an abstract measure. This is a good time to review the general definitions and results presented in Section 2.3. Think concretely about what these statements mean for measures on the real line. Draw some pictures! ▪

3.2.5 Discrete Borel measures

Let us present a few simple examples of discrete Borel measures that are similar in spirit to examples we have considered before.

Example 3.11 (Dirac measure on the real line). For a point $t \in \mathbb{R}$, we can define the Dirac measure δ_t. For each Borel set $\mathsf{B} \in \mathscr{B}(\mathbb{R})$,

$$\delta_t(\mathsf{B}) := \mathbb{1}_{\mathsf{B}}(t) := \begin{cases} 1, & t \in \mathsf{B}; \\ 0, & t \notin \mathsf{B}. \end{cases}$$

By general considerations, δ_t is a measure on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. It has an atom at t. ▪

Example 3.12 (Measure supported on the integers). Let $(w_i \in [0, +\infty] : i \in \mathbb{Z})$ be a sequence of positive weights. We can define a Borel measure

$$\mu(\mathsf{B}) := \sum\nolimits_{i \in \mathsf{B} \cap \mathbb{Z}} w_i \quad \text{for each Borel set } \mathsf{B} \subseteq \mathbb{R}.$$

This measure has an atom at each integer i where $w_i > 0$. When restricted to subsets of the integers, the Borel measure μ coincides with a measure on the integers, as defined in Lecture 1. Nevertheless, μ is defined on the entire class of Borel subsets of the real line, which is larger than the collection $\mathscr{P}(\mathbb{Z})$ of subsets of the integers. ▪

Example 3.13 (Discrete measure). We can extend Example 3.12 by considering a sequence $(w_i \in [0, +\infty] : i \in \mathsf{I})$ of positive weights indexed by a *countable* set I. Define a Borel measure

$$\mu(\mathsf{B}) := \sum\nolimits_{i \in \mathsf{B} \cap \mathsf{I}} w_i \quad \text{for each Borel set } \mathsf{B} \subseteq \mathbb{R}. \tag{3.1}$$

This measure has an atom at each point $i \in \mathsf{I}$ where $w_i > 0$. If a Borel measure can be expressed in the form (3.1), we say that the measure is *discrete*. ▪

3.3 The Lebesgue measure

In the late 19th century, mathematicians began a serious attack on the following question: "How do we define the length of a subset of the real line?" Our geometric intuitions lead to some sensible definitions, but it turns out that these definitions are fraught with peril. The field of measure theory was initially developed to resolve the confusion about the meaning of the word "length." The critical steps of this project were completed in Henri Lebesgue's 1902 doctoral dissertation.

We have already laid the groundwork for defining the length on Borel sets. In this section, we will complete the construction.

3.3.1 The length of elementary sets

In this section, we outline some of the properties that the "length" should have. *This discussion is not rigorous because length is never defined.* The formal construction of a measure that models the length appears in the next subsection.

So, how might we assign a length to a subset of the real line? Let us start small. Surely, the half-open interval $(a, b]$ must have length $|b - a|$ for all real numbers $a < b$. What about more complicated sets? Consider a finite union of *disjoint* half-open intervals; for example,

$$\mathsf{A} = (a_1, b_1] \cup (a_2, b_2] \cup \cdots \cup (a_n, b_n] \quad \text{with } a_i < b_i \le a_{i+1} \text{ for each } i.$$

In this construction, it is more convenient to work with half-open intervals rather than with open intervals. Half-open intervals link together more neatly, and the complement of a half-open interval is a half-open interval. As we saw in Exercise 3.6, the half-open intervals also generate the Borel sets.

We want the length to be finitely additive, so the length of this disjoint union should equal the sum of the lengths of its components:

$$\operatorname{length}(\mathsf{A}) = \textstyle\sum_{i=1}^{n} |b_i - a_i|. \tag{3.2}$$

This formula requires some thought because we have to affirm that it gives the same result, no matter how we break up the set into subintervals.

If the length is to become a measure, then it must also be countably additive. Thus, it also seems reasonable that we should be able to assign a length to a *countable, disjoint* union of intervals, presumably by adding up the lengths of the intervals:

$$\mathsf{A} = \dot{\bigcup}_{i=1}^{\infty} (a_i, b_i] \quad \text{implies} \quad \operatorname{length}(\mathsf{A}) = \textstyle\sum_{i=1}^{\infty} |b_i - a_i|. \tag{3.3}$$

This expression takes a partial step toward countable additivity. As before, we need to argue that the formula (3.3) gives a well-defined result. This is one of the core technical challenges in constructing a measure that agrees with the length.

Exercise 3.14 (*Length: Finite additivity). Show that (3.2) does not depend on how we decompose the set A as a finite union of disjoint half-open intervals.

3.3.2 Lebesgue measure

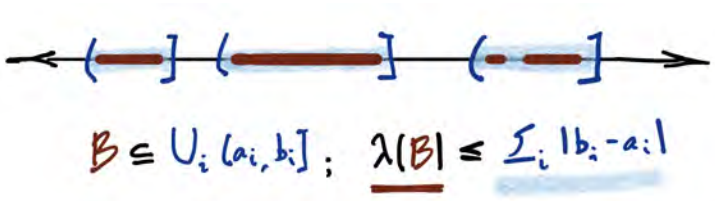
What about sets that are more complicated still? We may not be able to break up a Borel set $\mathsf{B} \in \mathscr{B}(\mathbb{R})$ into a countable number of half-open intervals. Instead, we will *cover* the set B by a countable union of (disjoint) half-open intervals. This union has an unambiguous length, and it serves as an upper bound for the length of the set B. Among all such covers, we search for the shortest one.

This approach results in the definition of the Lebesgue measure, which assigns a well-determined length to each Borel set.

Definition 3.15 (Lebesgue measure). The *Lebesgue measure* is the Borel measure $\lambda : \mathscr{B}(\mathbb{R}) \to [0, +\infty]$ given by the rule

$$\lambda(\mathsf{B}) := \inf \Big\{ \textstyle\sum_{i=1}^{\infty} |b_i - a_i| : \mathsf{B} \subseteq \dot{\bigcup}_{i=1}^{\infty} (a_i, b_i] \Big\}. \tag{3.4}$$

The function that appears in (3.4) is called the *exterior length* of the Borel set. Let us emphasize that the definition is sensible, even without proving that (3.3) is valid. The construction is akin to shrink-wrapping a package:



You can easily confirm that the exterior length coincides with the elementary length on a finite union of half-open intervals.

It is not evident that Definition 3.15 yields a reasonable construction of the length for all Borel sets. Nor is it clear that the definition results in a Borel measure. Let us state a major theorem that speaks to these concerns.

Theorem 3.16 (Lebesgue measure). The Lebesgue measure $\lambda : \mathscr{B}(\mathbb{R}) \to [0, +\infty]$, defined in (3.4), has the following properties.

1. **Measure:** The Lebesgue measure λ is a Borel measure. In particular, it is countably additive.
2. **Intervals:** Each interval has Lebesgue measure $\lambda((a, b]) = |b - a|$ for all real numbers $a < b$. In particular, $\lambda((0, 1]) = 1$.
3. **Translation invariance:** The Lebesgue measure is invariant under translation: $\lambda(\mathsf{B} + t) = \lambda(\mathsf{B})$ for all Borel B and all $t \in \mathbb{R}$.
4. **Uniqueness:** The Lebesgue measure is the only Borel measure that satisfies requirements (1)–(3).

Proof. The proof of Theorem 3.16 appears in Appendix A.3. You are invited to check the length property and the translation invariance yourself. The remaining assertions (countable additivity, uniqueness) are very difficult. ∎

In summary, the Lebesgue measure λ reports the length of every Borel set. The distinctive property of the Lebesgue measure is the translation invariance. This result ensures that the Lebesgue measure models a uniform distribution of mass on the real line. The normalization just establishes a particular scaling: the measure places one unit of mass per unit of length.

Exercise 3.17 (Lebesgue measure: Singletons). For a point $x \in \mathbb{R}$, use the definition of the Lebesgue measure to compute the Lebesgue measure $\lambda(\{x\})$ of the singleton set containing x.

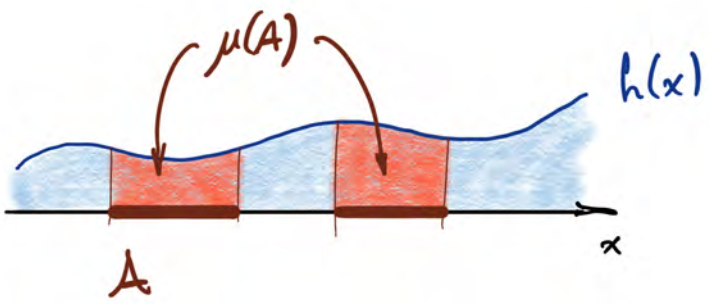
Exercise 3.18 (Lebesgue almost everywhere). Let λ be the Lebesgue measure. Use the definition to confirm that the empty set is negligible for λ. Deduce that $\mathbb{R}$ is an almost everywhere set for λ. Find two more examples of λ-negligible sets and two more examples of λ-almost everywhere sets. Be creative!

Exercise 3.19 (Lebesgue measure: σ-finiteness). Explain why the Lebesgue measure λ is σ-finite.

> **Aside:** At this stage, we can obtain a more concrete understanding of what Borel sets look like. For each $\varepsilon > 0$, every Borel set with finite Lebesgue measure can be expressed as a finite union of disjoint half-open intervals, united with a Borel set that has Lebesgue measure less than ε. Proving this fact is a challenging problem that requires the machinery from Appendix A. See Problem A.19.

3.3.3 Looking ahead: Measures with density

A natural way to describe a distribution of mass on the real line is to start with the local density, the mass per unit length.



Let $h : \mathbb{R} \to \mathbb{R}_+$ be a positive function (say, bounded and continuous) that models the local density. Formally, we can define a measure μ on each Borel set A via the expression

$$\text{“} \quad \mu(\mathsf{A}) := \int_{\mathsf{A}} h(x)\,\mathrm{d}x. \quad \text{”}$$

That is, the measure μ adds up the local density on the Borel set A. When A is a finite union of intervals, we can simply use a Riemann integral here. Ahead of us, we have a big job of understanding what the integral means for all Borel sets A. The Lebesgue measure will play a central role in this construction.

> **Aside:** It is not the case that all Borel measures are continuous (with a local density), or discrete (point masses), or positive linear combinations thereof. There also exist *singular measures*, such as the Cantor distribution, that have neither a density nor a discrete distribution. Although these examples may seem exotic, singular measures arise naturally in the study of continuous stochastic processes, such as Brownian motion. Singular measures will not play a role in this course.

3.3.4 *Unmeasurable sets

The perspicacious reader will realize that the definition (3.4) of the exterior length could be applied to any subset of the real line. Why have we restricted our attention to the Borel sets?

Suppose that we use (3.4) to define the exterior length $\lambda^* : \mathscr{P}(\mathbb{R}) \to [0, +\infty]$ for an *arbitrary* subset of the real line. Unfortunately, λ^* does not behave itself. Indeed, there are (many!) disjoint pairs of sets for which the exterior length of the union is strictly smaller than the sum of the exterior lengths of the sets:

$$\lambda^*(\mathsf{A} \dot\cup \mathsf{B}) < \lambda^*(\mathsf{A}) + \lambda^*(\mathsf{B}) \quad \text{for certain disjoint subsets } \mathsf{A}, \mathsf{B} \text{ of } \mathbb{R}.$$

This fact violates all our intuitions about the length: the parts are greater than the whole. For Borel sets, however, the exterior length behaves in accordance with our expectations about length.

We have averred that some subsets of the real line cannot be assigned a length in a satisfactory way. This is the fundamental reason that we have to work with a smaller collection of subsets, such as the Borel sets. For the assiduous student, Appendix B contains a more in-depth discussion about sets that cannot be assigned a length.

3.4 Support

A Borel measure may only place its mass on a part of the real line. It is helpful to have terminology and a rigorous definition to describe the locations where there is mass.

Definition 3.20 (Borel measure: Support). Let μ be a Borel measure on the real line. The *support* of the measure μ is defined as the set of points where every open neighborhood has strictly positive measure:

$$\operatorname{supp}(\mu) := \{x \in \mathbb{R} : \mu(x - \varepsilon, x + \varepsilon) > 0 \text{ for all } \varepsilon > 0\}.$$

The support is always a closed subset of $\mathbb{R}$, hence a Borel set.

Example 3.21 (Borel measure: Support). For the Lebesgue measure, $\operatorname{supp}(\lambda) = \mathbb{R}$. Indeed, every open interval satisfies $\lambda(x - \varepsilon, x + \varepsilon) = 2\varepsilon > 0$, so every point $x \in \mathbb{R}$ belongs to the support.

For the Dirac measure at zero, $\operatorname{supp}(\delta_0) = \{0\}$. Indeed, every open interval about zero satisfies $\delta_0(-\varepsilon, +\varepsilon) = 1$. On the other hand, for each point $a \neq 0$, we can find an open interval about a with zero measure: $\delta_0(a - |a|/2, a + |a|/2) = 0$. ∎

3.5 Specifying a Borel measure

Let us pose the same type of question that we asked at the end of each of the previous two lectures: How can we specify a Borel measure on the real line?

In Lecture 1, we gave some specific answers that were tailored to the elementary case of a measure on the integers. In Lecture 2, we described an abstract approach that extends a partial definition from a class of elementary sets to the full class of measurable sets. Here, we will unite these two perspectives to arrive at a powerful approach for representing measures on the real line.

Definition 3.22 (Distribution function). Let μ be a *finite* Borel measure on the real line. The *(cumulative) distribution function* (abbreviated *cdf* or *df*) of the measure is defined as

$$F_\mu(a) := \mu((-\infty, a]) \quad \text{for } a \in \mathbb{R}.$$

Exercise 3.23 (Some distribution functions). What are the distribution functions of some basic Borel measures?

1. Let $\mu = \delta_t$ be the Dirac measure at the point $t \in \mathbb{R}$. Compute the distribution function.
2. Define $\mu(\mathsf{B}) := \lambda(\mathsf{B} \cap (0, 1])$ on the Borel sets $\mathsf{B} \in \mathscr{B}(\mathbb{R})$. This is the uniform measure restricted to the interval $(0, 1]$. Compute its distribution function.

At a basic level, we can appreciate how the distribution function might serve to represent a Borel measure. Indeed, Exercise 3.6 shows that Borel sets are generated by the class of semi-infinite intervals $(-\infty, a]$ for $a \in \mathbb{R}$. As a consequence, it is easy to imagine that knowledge of the measure on these elementary sets is necessary and sufficient to determine its values on all of the Borel sets. This intuition is correct, but the formal deduction is quite intricate. See Appendix A.

Distribution functions play an important role in probability theory. Anticipating these developments, let us outline some useful results, which state that a distribution function characterizes a finite Borel measure.

Proposition 3.24 (Distribution function: Properties). Let $F : \mathbb{R} \to \mathbb{R}_+$ be the distribution function of a *finite* Borel measure μ on the real line. Set $M := \mu(\mathbb{R}) < +\infty$. Then F enjoys three properties:

1. **Increasing:** If $a \leq b$, then $F(a) \leq F(b)$.
2. **Asymptotic limits:** We have $\lim_{a \downarrow -\infty} F(a) = 0$ and $\lim_{a \uparrow +\infty} F(a) = M$.
3. **Right continuous:** For each $x \in \mathbb{R}$, we have $\lim_{a \downarrow x} F(a) = F(x)$.

Problem 3.25 (Distribution function). Prove Proposition 3.24.

Remarkably, the converse is also true. Any function with these distinguished properties is the distribution function of a unique finite measure.

Theorem 3.26 (Distribution function = measure). Let $F : \mathbb{R} \to \mathbb{R}_+$ be a function that satisfies the properties listed in Proposition 3.24(1)–(3). Then there is a unique finite measure μ for which

$$\mu((a, b]) = F(b) - F(a) \quad \text{for all } a, b \in \mathbb{R} \text{ with } a \leq b.$$

Proof. See Problem A.17 in Appendix A. ∎

Quiz

Respond to the following questions with one of the alternatives: Always True (T) / Always False (F).

1. The set $\{\pi\}$ is a Borel set.
2. Every subset of the real line is a Borel set.
3. If A, B are Borel sets, then $\mathsf{A} \cap \mathsf{B}$ is a Borel set.
4. The Lebesgue measure of $(1, 2) \cup [3, 4]$ is two.
5. A Borel measure μ is finitely additive:

$$\mathsf{A} = \dot{\bigcup}_{i=1}^{n} \mathsf{A}_i \quad \text{implies} \quad \mu(\mathsf{A}) = \sum_{i=1}^{n} \mu(\mathsf{A}_i) \quad \text{for Borel sets } \mathsf{A}_i.$$

6. If the measure of a set is zero, then the set is empty.

Problems

Exercise 3.27 (Lebesgue measure: Rationals). Explain why the set $\mathbb{Q}$ of rational numbers is countable. Show that $\mathbb{Q}$ is a Borel set. Confirm that the Lebesgue measure $\lambda(\mathbb{Q}) = 0$.

Problem 3.28 (*Lebesgue measure: Cantor set). Recall the construction of Cantor's ternary set. Begin with the unit interval. Remove the middle third. Remove the middle third of

each of the two remaining subintervals. Repeat. This yields a sequence

$$\begin{aligned}
\mathsf{T}_0 &:= [0,1];\\
\mathsf{T}_1 &:= [0,1/3] \cup [2/3,1];\\
\mathsf{T}_2 &:= [0,1/9] \cup [2/9 \cup 1/3] \cup [2/3,7/9] \cup [8/9,1]; \ldots
\end{aligned}$$

Cantor's ternary set T is obtained as the (decreasing) set limit of this process. Equivalently, T is the set of numbers in the interval $[0,1]$ whose base-three expansion does not contain the numeral 1.

Prove that Cantor's ternary set T is a Borel set. Explain why T is uncountable. Confirm that the Lebesgue measure $\lambda(\mathsf{T}) = 0$.

Notes

You can find the material in this lecture in any book on probability theory or measure theory. For example, see Billingsley [Bil12] or Folland [Fol99]. Probability texts focus their attention on Borel sets and Borel measures, in part because probability involves many different types of distributions. Real analysis and measure theory books place more emphasis on Lebesgue sets, and the Lebesgue measure is often the main object of study.

For an explicit example of non-Borel set, see the book [Kec95] by Caltech logician, Alexander Kechris.

Lecture bibliography

[Bil12] P. Billingsley. *Probability and measure*. Anniversary ed. John Wiley & Sons Inc., 2012.

[Fol99] G. B. Folland. *Real analysis*. Second. Modern techniques and their applications, A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1999.

[Kec95] A. S. Kechris. *Classical descriptive set theory*. Springer-Verlag, New York, 1995. DOI: 10.1007/978-1-4612-4190-4.

4. Integration on the Real Line

> "There is nothing new under the sun, but there are new suns."
>
> —Octavia Butler

Agenda:

1. Sums weighted by mass
2. Measurable functions
3. Integrating positive functions
4. Integrating signed functions
5. Examples
6. Riemann vs. Lebesgue

We have now completed our introduction to the concept of a measure, which models a distribution of mass over a domain. Our next step is to develop a general theory of integration, which allows us to add up the values of a function, weighted by the local mass. We have seen hints that a flexible method for integration is valuable for problems in geometry, mechanics, and probability.

This lecture begins with a geometric description of our new approach to integration. Once we perceive the ideas, we can introduce the class of functions that we are allowed to integrate. Afterward, we give a formal definition of the integral that comports with our geometric picture. We discuss some of the basic properties of the new integral, and we compare it with the familiar Riemann integral.

We will develop the basic concepts of integration in the concrete setting of Borel measures on the real line. The next lecture will recapitulate these ideas in an abstract setting. The abstract definitions are really no different, but the simpler presentation here may help to build intuition.

4.1 Sums weighted by mass

To reiterate: an integral sums the values of a function, weighted by the local mass, over a domain. Applications include

- **Geometry:** For a uniform distribution of mass on the real line (modeled by the Lebesgue measure), the integral of a function computes the signed area enclosed between a function and the horizontal axis.
- **Mechanics:** For a distribution of mass in a one-dimensional mechanical system, we can define an integral that returns the center of mass of the distribution.
- **Probability:** For a distribution of probability, the integral can be used to find the expected value of the distribution.

In this lecture, our goal is to develop a general method for integrating a real-valued function against a distribution of mass on the real line, given by a Borel measure. In particular, the Lebesgue measure λ describes a uniform distribution of mass over the real line, where the mass of an interval equals its length. In this situation, the integral computes the signed area between the function and the horizontal axis. To fix ideas, you may wish to visualize this special case for the remainder of this lecture.

You can see that there is a potential mismatch in the definitions of functions and measures. Indeed, functions take values at points, whereas the Lebesgue measure is defined on sets. It may also be puzzling that the Lebesgue measure of a singleton set is zero, which suggests that the uniform distribution does not put any mass anywhere.

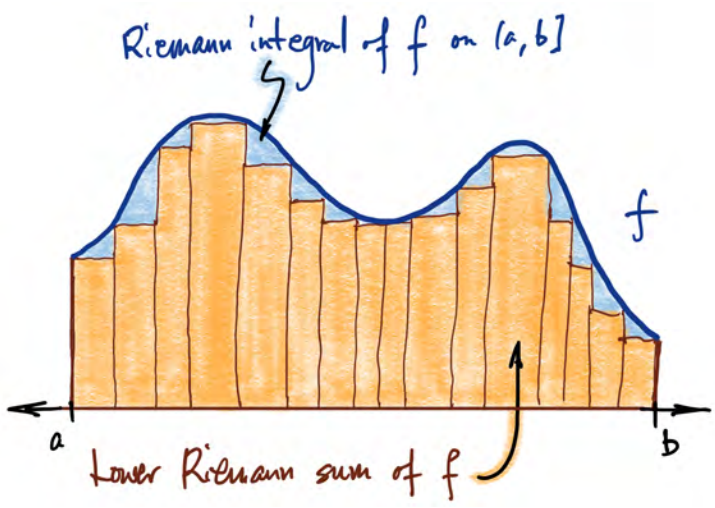


Figure 4.1 (The Riemann integral). The region under the curve is sliced into narrow vertical bands.

To resolve this conflict, we first review the geometric construction of the Riemann integral. This treatment reveals some shortcomings that we want to address. Then we outline another approach to integration, called the *Lebesgue integral*, that will lead to a more satisfactory theory. Lebesgue integrals provide the foundation for much of modern analysis and for probability theory.

4.1.1 Riemann integrals

In the simplest case, the integral computes the area enclosed between a *positive* function and the horizontal axis. You have learned to denote this quantity using the symbols

$$\int_a^b f(x)\,\mathrm{d}x \quad \text{where } f : [a, b] \to \mathbb{R}_+.$$

To define this object, we approximate the region under the curve by vertical rectangles that do not overlap. The rough area under the curve is obtained by adding up the areas of the rectangles (given by the length of the base times the height). See Figure 4.1. By making the rectangles narrower, we can improve the quality of the approximation. This geometric approach to integration dates back to the ancient Greeks (Eudoxus, Archimedes) and the ancient Chinese (Liu Hui, Zu Chongzhi).

Bernhard Riemann, in the 19th century, was the first to give a rigorous treatment of the integral. He formalized the notion of subdividing the area into rectangles of increasingly small width. If the total area of the rectangles tends to a well-defined limit as the width tends to zero, then the integral is declared to equal this limit. Most real analysis books present a variant of Riemann's approach, called a *Darboux integral*. See Appendix C for an overview of this construction.

The Riemann integral weights function values by a uniform distribution of mass over the real line. We can extend the Riemann integral to more general distributions of mass. The heuristic approach is to multiply the height of each of the vertical panels by the *mass* carried on its base, rather than the length of the base. This construction can be made rigorous, and it leads to an object called the *Riemann–Stieltjes integral*.

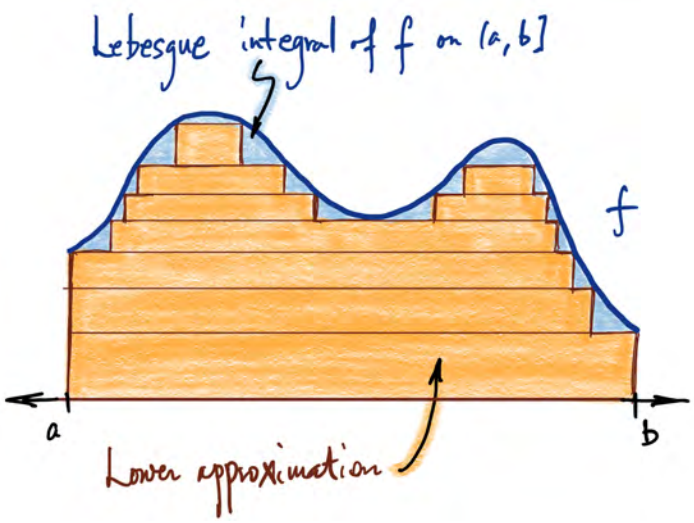


Figure 4.2 (The Lebesgue integral). The region under the curve is sliced into narrow horizontal bands.

These intuitive ideas work well for some types of functions, such as continuous functions on a compact interval. Nevertheless, we can easily escape from the class of Riemann-integrable functions. For example, neither unbounded functions nor functions on unbounded intervals are considered to be Riemann integrable, although we can sometimes achieve satisfactory results for these examples by taking limits (called *improper Riemann integrals*).

Unfortunately, when we start taking limits of Riemann integrals, we quickly run into problems. There is no satisfactory theory that describes when we can interchange a limit with a Riemann integral. This leads to vexing outcomes. For instance, we can construct a pointwise-convergent sequence of Riemann-integrable functions whose limit is not Riemann integrable. This difficulty remains even when we restrict our attention to nice functions (e.g., bounded, continuous) or to gentle kinds of convergence (e.g, monotone increase). Since limits are among the basic operations in analysis (and probability!), this shortcoming of the Riemann integral is serious.

The construction of the Riemann integral also depends on being able to partition the domain of the function into increasingly tiny, nonoverlapping pieces. As a consequence, we cannot extend the Riemann integral naturally to functions defined on more general domains. As we will see (Lecture 7), this issue makes Riemann integrals unsuitable for serious probability.

4.1.2 Lebesgue integrals

Let us return to our motivating problem. How can we compute the area between a positive function and the horizontal axis?

The simple, but astonishing, idea of Henri Lebesgue was to cut up the area under the function *horizontally* instead of vertically. The content of each horizontal rectangle is given by its height times the length of its base. By taking the limit as the *height* of each rectangle tends to zero, we can compute the area under the curve. See Figure 4.2 for an illustration of this idea.

Lebesgue invented the concept of a measure, constructed the Lebesgue measure, and designed the new integral in his 1902 doctoral thesis, *Intégrale, longeur, aire.*

We can also realize this geometric idea using functions. Suppose that $f : \mathbb{R} \to \mathbb{R}_+$

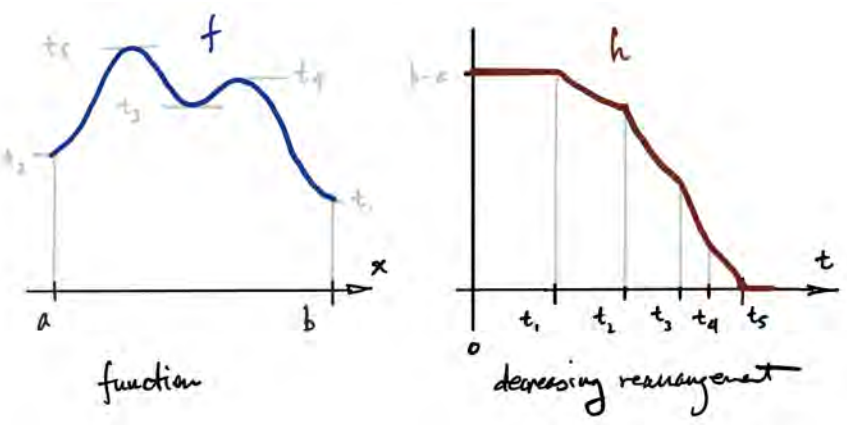


Figure 4.3 (Decreasing rearrangement). For a positive-valued function $f : \mathbb{R} \to \mathbb{R}_+$, the decreasing rearrangement $h : \mathbb{R}_+ \to \mathbb{R}_+$, defined in (4.1), returns the total length of the super-level set $\{x \in \mathbb{R} : f(x) > t\}$ for each $t \geq 0$. The rearrangement is positive and decreasing. The area under the function f is the same as the area under the rearrangement h.

is a "nice" *positive-valued* function that we wish to integrate. Then we can construct another function

$$h(t) := \lambda\{x \in \mathbb{R} : f(x) > t\} \quad \text{for } t \geq 0. \tag{4.1}$$

The function h is sometimes called the *decreasing rearrangement* of f.

Evidently, h is positive and decreasing (why?). You can interpret the value $h(t)$ as the total length of the (broken) horizontal line, positioned at the location t along the vertical axis and lying beneath the function f. See Figure 4.3 for a schematic.

The strong inequality in (4.1) could also be replaced by a weak inequality. The strong inequality simplifies a few technical arguments.

Here is a second way to visualize the function h. Suppose that we turn the graph of the function f clockwise by 90°. At each point t on the horizontal axis, we consolidate the (broken) vertical line above t inside the rotated curve by sliding the pieces downward to form an interval sitting on the horizontal axis. This construction indicates that the area under h equals the area under the function f.

We want to construct an integral that returns the area under the function f induced by a subdivision into horizontal panels. Therefore, we anticipate that the right approach is to define this integral via the relation

$$\int_{\mathbb{R}} f(x)\, \lambda(\mathrm{d}x) := \int_0^\infty \lambda\{x \in \mathbb{R} : f(x) > t\}\, \mathrm{d}t = \int_0^\infty h(t)\, \mathrm{d}t. \tag{4.2}$$

You should interpret the right-hand side of the definition as an improper Riemann integral.

The notation for the integral is intended to suggest that each function value $f(x)$ is weighted by the local mass $\lambda(\mathrm{d}x)$ carried by an "infinitesimal" set at the point x. As we will discover, the formulation (4.2) leads to a well-defined integral with many remarkable properties.

The definition (4.2) focuses our attention on the super-level sets of the integrand: $\{x \in \mathbb{R} : f(x) > t\}$. The Lebesgue measure λ simply computes the total length of each super-level set. It should now be obvious that we can weight function values by a different distribution of mass by replacing λ with another Borel measure. Furthermore, we perceive that there is an opportunity to construct the integral of a real-valued function defined on any domain that carries a measure (see Lecture 5).

4.2 Borel measurable functions

Before we can define the integral properly, we need to take a step back again and ask what kinds of functions we may try to integrate. Observe that the preliminary definition (4.2) of the integral does not even make sense unless we can apply the measure λ to the super-level sets. In other words, each super-level set of the integrand must be a Borel set. Equivalently,

$$f^{-1}(t, +\infty) := \{x \in \mathbb{R} : f(x) > t\} \text{ is Borel for each } t \in \mathbb{R}.$$

Now is a good time to review the definition of a preimage and its properties (Exercise 1.26).

To attend to this the matter, we must introduce the concept of a (Borel) measurable function.

4.2.1 Measurability

If we want to compute the measure of the super-level sets of the integrand, these sets must be measurable. This important insight leads us to our next definition.

Definition 4.1 (Measurable function). We say that a function $f : \mathbb{R} \to \mathbb{R}$ is *measurable* if the preimage of each semi-infinite interval $(t, +\infty)$ is a Borel set. That is,

$$f^{-1}(t, +\infty) := \{x \in \mathbb{R} : f(x) > t\} \in \mathscr{B}(\mathbb{R}) \quad \text{for all } t \in \mathbb{R}. \tag{4.3}$$

For emphasis, we may refer to f as a *Borel-measurable* function.

Warning: Measurability of a function does not involve a measure!

The condition (4.3) on super-level sets implies that a measurable function satisfies an (apparently) much stronger property.

Proposition 4.2 (Borel measurability). A function $f : \mathbb{R} \to \mathbb{R}$ is Borel measurable if and only if

$$f^{-1}(\mathsf{B}) := \{x \in \mathbb{R} : f(x) \in \mathsf{B}\} \in \mathscr{B}(\mathbb{R}) \quad \text{for all Borel } \mathsf{B} \in \mathscr{B}(\mathbb{R}). \tag{4.4}$$

For comparison, recall that a function is continuous if the preimage of every open set is an open set.

That is, the preimage of every Borel set is a Borel set.

**Proof.* This argument exhibits a fundamental technique for working with Borel sets. It hinges on the fact that $\mathscr{B}(\mathbb{R})$ is the *smallest* σ-algebra on $\mathbb{R}$ that contains all semi-infinite intervals.

It is clear that the condition (4.4) for Borel sets implies the condition (4.3) for super-level sets because the semi-infinite interval $(t, +\infty)$ is a Borel set for each $t \in \mathbb{R}$. See Exercise 3.5.

To prove the converse, suppose that the function $f : \mathbb{R} \to \mathbb{R}$ satisfies the condition (4.3) on super-level sets. Introduce the collection $\mathscr{C}$ of sets whose preimage under f is a Borel set:

$$\mathscr{C} := \{\mathsf{C} \subseteq \mathbb{R} : f^{-1}(\mathsf{C}) \in \mathscr{B}(\mathbb{R})\}.$$

We *claim* that $\mathscr{C}$ is a σ-algebra. Given this claim, we may complete the argument. The condition (4.3) ensures that $\mathscr{C}$ contains every semi-infinite interval $(t, +\infty)$. According to Exercise 3.6, the Borel class $\mathscr{B}(\mathbb{R})$ is the *smallest* σ-algebra that contains the semi-infinite intervals. Therefore, $\mathscr{B}(\mathbb{R}) \subseteq \mathscr{C}$. In particular, $f^{-1}(\mathsf{B})$ is a Borel set for every Borel set B.

To establish the claim, we first prove that the family $\mathscr{C}$ is stable under complements. That is, if $\mathsf{C} \in \mathscr{C}$, then $\mathsf{C}^{\mathsf{c}} \in \mathscr{C}$. Recall that $\mathsf{C} \in \mathscr{C}$ means that $f^{-1}(\mathsf{C}) \in \mathscr{B}(\mathbb{R})$. Since the Borel sets are stable under complements,

$$f^{-1}(\mathsf{C}^{\mathsf{c}}) = (f^{-1}(\mathsf{C}))^{\mathsf{c}} \in \mathscr{B}(\mathbb{R}).$$

We have used the fact (Exercise 1.26) that the preimage commutes with the complement. Therefore, $\mathsf{C}^c \in \mathscr{C}$. A very similar argument establishes that $\mathscr{C}$ is stable under countable unions. These two properties are enough to conclude that $\mathscr{C}$ is a σ-algebra. ∎

Using Proposition 4.2, we can easily confirm that points where a measurable function takes a specific value compose a measurable set. This is an important application.

Exercise 4.3 (Measurable function: Zero set). Let $f : \mathbb{R} \to \mathbb{R}$ be a measurable function. Then $\{x \in \mathbb{R} : f(x) = 0\}$ and $\{x \in \mathbb{R} : f(x) \neq 0\}$ are both measurable sets.

The argument behind Proposition 4.2 can be used to establish many related principles. For example, we can check the measurability of a function by examining the preimages of finite half-open intervals.

Exercise 4.4 (*Measurability: Intervals). Let $f : \mathbb{R} \to \mathbb{R}$ be a function. Suppose that the preimage $f^{-1}(a, b]$ is measurable for all real numbers $a < b$. Deduce that f is measurable.

4.2.2 Extended values

Functions with extended real values can easily arise from limiting processes, so we need an appropriate definition for measurability. (Sorry!)

Definition 4.5 (Measurable function: Extended values). We say that a function $f : \mathbb{R} \to \overline{\mathbb{R}}$ with extended real values is *measurable* when

$$f^{-1}(t, +\infty] := \{x \in \mathbb{R} : f(x) > t\} \in \mathscr{B}(\mathbb{R}) \quad \text{for all } t \in \mathbb{R}.$$

For emphasis, we may also say that f is *Borel measurable*.

Equivalently, $f : \mathbb{R} \to \overline{\mathbb{R}}$ is measurable when each *extended* Borel set $\mathsf{B} \in \mathscr{B}(\overline{\mathbb{R}})$ has a measurable preimage: $f^{-1}(\mathsf{B}) \in \mathscr{B}(\mathbb{R})$. There is no conceptual difference from the case of real-valued functions.

See Definition 3.7 of the extended Borel sets.

We have chosen to allow extended-valued functions because they lead to a simpler presentation of integration theory with fewer special cases and qualifications. It does take some practice to get used to working with functions that take infinite values. Keep in mind the conventions that $0/0 = 0$ and $0 \cdot (\pm\infty) = 0$. We never allow division by $\pm\infty$. We must also avoid *competing infinities*: expressions of the form $\infty - \infty$ are undefined.

> There is a broad principle that we can deal with infinite numbers or signed numbers, but not both at the same time. Thus, we will allow positive functions to take the value $+\infty$, but we will require signed functions to remain finite.

4.2.3 Examples

Our prospective definition (4.2) of the integral leads inexorably to the concept of a measurable function. Let us give some important examples of measurable functions. These results are a straightforward consequence of the definition.

Example 4.6 (Measurability: Indicators). Let $\mathsf{B} \in \mathscr{B}(\mathbb{R})$ be a Borel set. Then the indicator $\mathbb{1}_{\mathsf{B}} : \mathbb{R} \to \mathbb{R}$ is a measurable function. Conversely, if $\mathsf{B} \subseteq \mathbb{R}$ is *not* Borel, then the indicator $\mathbb{1}_{\mathsf{B}}$ is *not* measurable.

We just need to examine the semi-infinite intervals $(t, +\infty)$ for each $t \in \mathbb{R}$. The

preimage is one of three alternatives:

$$(\mathbb{1}_{\mathsf{B}})^{-1}(t,+\infty) = \begin{cases} \mathbb{R}, & t < 0; \\ \mathsf{B}, & t \in [0,1); \\ \emptyset, & t \geq 1. \end{cases}$$

Obviously, $\emptyset$ and $\mathbb{R}$ are Borel sets. We see that the indicator $\mathbb{1}_{\mathsf{B}}$ is measurable if and only if the set B is Borel. ▪

Exercise 4.7 (Measurability: Constant functions). Show that each constant function $f : \mathbb{R} \to \mathbb{R}$ is measurable. Recall that a constant function satisfies $f(x) = c$ for all $x \in \mathbb{R}$, where $c \in \mathbb{R}$.

The next three examples show that many familiar classes of real-valued functions are measurable.

Example 4.8 (Measurability: Continuous functions). Let $f : \mathbb{R} \to \mathbb{R}$ be a continuous function. Then f is Borel measurable.

Indeed, for each $t \in \mathbb{R}$, the preimage $f^{-1}(t,+\infty)$ of the open interval $(t,+\infty)$ is an open set because f is continuous. Finally, recall that every open set is a Borel set. ▪

Exercise 4.9 (Measurability: Monotone functions). Let $f : \mathbb{R} \to \overline{\mathbb{R}}$ be a monotone function. Prove that f is Borel measurable.

Exercise 4.10 (Measurability: Convex functions). Let $f : \mathbb{R} \to \overline{\mathbb{R}}$ be a lower-semicontinuous (lsc) convex function. Prove that f is Borel measurable. **Hint:** Look up the definition of lsc. Is convexity required?

4.2.4 Stability properties

Next, we will argue that most set-theoretic, algebraic, and analytic combinations of measurable functions remain measurable. This is basically all you need to know to work with measurable functions. The proofs are not of great importance for this course.

Example 4.11 (Measurability: Composition). Let $f, g : \mathbb{R} \to \mathbb{R}$ be measurable functions. Then the composition $f \circ g$ is measurable. See Problem 5.36 for an extension.

To prove this claim, recall that $(f \circ g)(x) := f(g(x))$ for $x \in \mathbb{R}$. By Exercise 1.26, the preimage of a set under composition satisfies

$$(f \circ g)^{-1}(\mathsf{A}) = g^{-1}(f^{-1}(\mathsf{A})) \quad \text{for each } \mathsf{A} \subseteq \mathbb{R}.$$

If A is a Borel set, the preimage $f^{-1}(\mathsf{A})$ under the measurable function f is a Borel set, and consequently the preimage $g^{-1}(f^{-1}(\mathsf{A}))$ under the measurable function g is also Borel. ▪

Example 4.12 (Measurability: Positive and negative part). Let $f : \mathbb{R} \to \overline{\mathbb{R}}$ be a function that may take extended real values. Define the positive and negative parts:

$$\begin{aligned} f_+(x) &:= \max\{+f(x), 0\}; \\ f_-(x) &:= \max\{-f(x), 0\} \end{aligned} \quad \text{for } x \in \mathbb{R}.$$

Warning: Both the positive part f_+ and the negative part f_- are positive-valued functions! ▪

As an exercise, confirm that $f = f_+ - f_-$ and that $|f| = f_+ + f_-$.

We assert that both the positive and negative parts of a measurable function are measurable. For example, let us consider the positive part. Observe that

$$(f_+)^{-1}(t,+\infty] = \begin{cases} f^{-1}(t,+\infty], & t \geq 0; \\ \mathbb{R}, & t < 0. \end{cases}$$

Since f is measurable, so is the positive part f_+. ∎

Using the ideas from the last example, you can establish some related results.

Exercise 4.13 (Measurability: Abs, min, max). Let $f, g : \mathbb{R} \to \overline{\mathbb{R}}$ be measurable functions. Confirm that the following functions are measurable.

- **Absolute value:** By direct argument, prove that $|f|$ is measurable.
- **Minimum:** The pointwise minimum $f \wedge g := \min\{f, g\}$ is measurable.
- **Maximum:** The pointwise maximum $f \vee g := \max\{f, g\}$ is measurable.

Next, we show sums of measurable functions are measurable.

Example 4.14 (Measurability: Positive sums). Let $f, g : \mathbb{R} \to \overline{\mathbb{R}}_+$ be positive, measurable functions that may take the value $+\infty$. Then $f + g$ is measurable.

Warning: We require *positive* values to avoid competing infinities $(\infty - \infty)$. ∎

To check this statement, it suffices to show that the super-level set $\{x \in \mathbb{R} : f(x) + g(x) > t\}$ is measurable for each $t \in \mathbb{R}$. Fix the level t. Observe that

$$f(x) + g(x) > t \quad \text{if and only if} \quad f(x) > q > t - g(x) \quad \text{for some } q \in \mathbb{Q}.$$

It is crucial that q is a rational number!

Using this observation, we can write the super-level set as a countable union over the rationals:

$$\{x \in \mathbb{R} : f(x) + g(x) > t\} = \bigcup\nolimits_{q \in \mathbb{Q}} \left[f^{-1}(q, +\infty] \cap g^{-1}(t - q, +\infty] \right].$$

We can start to see why it is so important that Borel sets are stable over countable unions.

By measurability of f, g, each element of the union is the intersection of two Borel sets, so is a Borel set. Finally, a countable union of Borel sets is Borel. ∎

Similar arguments allow us to show that algebraic combinations of signed functions remain measurable.

Exercise 4.15 (Measurability: Algebraic combinations). Let $f, g : \mathbb{R} \to \mathbb{R}$ be *finite-valued* measurable functions. Prove that the following functions are measurable.

Warning: We require *finite* values to avoid competing infinities $(\infty - \infty)$. ∎

- **Sums:** The sum $f + g$ is measurable.
- **Products:** The product fg is measurable.
- **Linearity:** Deduce that the set of finite-valued Borel measurable functions is a linear space (in fact, an algebra).

4.2.5 Countable combinations and limits

Operations involving a countable number of measurable functions produce measurable functions. These results allow us to form limits. By permitting functions to take extended values, we can obtain clean statements without any extraneous conditions. The fact that the Borel sets form a σ-algebra is also a crucial ingredient here.

Example 4.16 (Measurability: Countable infimum and supremum). For each $j \in \mathbb{N}$, let $f_j : \mathbb{R} \to \overline{\mathbb{R}}$ be a measurable function that may take extended real values. Then the pointwise infimum, $\inf_{j \in \mathbb{N}} f_j$, and the pointwise supremum, $\sup_{j \in \mathbb{N}} f_j$, are measurable functions.

This is the place where it is really critical that measurable sets are stable under countable unions.

Let us establish the result for the supremum. To do so, we write its super-level set at $t \in \mathbb{R}$ as a countable union of Borel sets:

$$\{x \in \mathbb{R} : \sup\nolimits_{j \in \mathbb{N}} f_j(x) > t\} = \bigcup\nolimits_{j=1}^{\infty} \{x \in \mathbb{R} : f_j(x) > t\}.$$

Indeed, $\sup_j f_j(x) > t$ if and only if $f_j(x) > t$ for at least one index j. ∎

The same approach applies to many related examples.

Exercise 4.17 (Measurability: Limits). For each $j \in \mathbb{N}$, let $f_j : \mathbb{R} \to \overline{\mathbb{R}}$ be a measurable function that may take extended real values.

- **Superior limit:** The superior limit, $\limsup_{j\to\infty} f_j$, is a measurable function. **Hint:** $\limsup_{j\to\infty} f_j = \inf_{j\geq 1} \sup_{k\geq j} f_k$.
- **Inferior limit:** The inferior limit, $\liminf_{j\to\infty} f_j$, is measurable.
- **Limits:** If the limit, $\lim_{j\to\infty} f_j$, exists pointwise in $\overline{\mathbb{R}}$, then it is measurable.
- **Set of convergence:** Deduce that the set $\{x \in \mathbb{R} : \lim_{j\to\infty} f_j(x) \text{ exists}\}$ is Borel.

Each of these results requires the stability of Borel sets under countable unions!

Exercise 4.18 (Measurability: Series). For each $j \in \mathbb{N}$, let $f_j : \mathbb{R} \to \overline{\mathbb{R}}_+$ be a measurable function that takes *positive* values. Verify that the series $\sum_{j=1}^{\infty} f_j$ determines a measurable function. The situation is more complicated for signed functions. Under what conditions is an infinite sum of signed functions measurable?

Warning 4.19 (Non-measurability: Uncountable combinations). It is problematic to perform uncountably indexed operations with measurable functions. For example, the supremum of an uncountable family of measurable functions may not be measurable. This kind of function does arise in statistics and optimization. Take care! ▪

Aside: If you have studied measure theory, you may have encountered Lebesgue-measurable functions. These are functions $f : \mathbb{R} \to \mathbb{R}$ for which the preimage $f^{-1}(\mathsf{B})$ of each Borel set B is a Lebesgue set. A rather unpleasant feature of this definition is that it does not interact neatly with composition. In particular, the composition of a Lebesgue measurable function followed by a continuous function need not be Lebesgue measurable.

4.2.6 The takeaway

Measurability is a fundamental concept that is crucial for defining integrals. At the same time, the following principle suggests that we usually do not need to worry too much about whether a real-valued function is Borel measurable.

If you encounter a function $f : \mathbb{R} \to \overline{\mathbb{R}}$ in applications, that function is quite likely to be Borel measurable. Warning 4.19 outlines the main exceptions to this principle.

As a consequence, we will not place a lot of emphasis on measurability at this point. It will become more important in probability theory when we study conditioning.

4.3 The Lebesgue integral on the real line

We are now prepared to give a rigorous definition of the integral of a function, weighted by a distribution of mass on the real line. This construction was pioneered by Henri Lebesgue in his doctoral thesis, so it is now called the *Lebesgue integral* in his honor.

The development of the Lebesgue integral proceeds in two stages. First, we define the integral for positive functions. Second, we use the primitive definition to extend the integral to signed functions.

4.3.1 The integral of a positive function

We begin with the simplest case: the integral of a positive function on the real line with respect to a Borel measure. The definition matches the geometric picture we gave

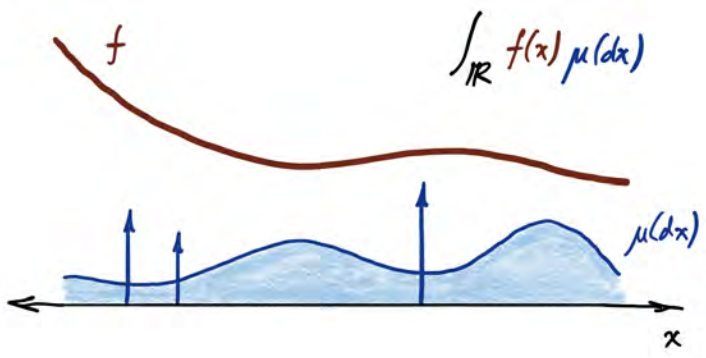


Figure 4.4 (Integral of a positive function). The integral weights the values $f(x)$ of a positive function by the local mass $\mu(\mathrm{d}x)$ and sums over the real line $\mathbb{R}$.

in Figure 4.2. Just add up the measures of the horizontal bands.

Definition 4.20 (Lebesgue integral: Positive function). Fix a Borel measure μ on the real line. Let $f : \mathbb{R} \to \overline{\mathbb{R}}_+$ be a *positive* measurable function that may take the value $+\infty$. Define the *Lebesgue integral* of f with respect to μ as

$$\int_{\mathbb{R}} f(x)\,\mu(\mathrm{d}x) := \int_0^{+\infty} \mu\{x \in \mathbb{R} : f(x) > t\}\,\mathrm{d}t. \tag{4.5}$$

The right-hand side of (4.5) is a well-defined improper Riemann integral (see Appendix C), which may take the value $+\infty$.

Warning: The Lebesgue integral need not involve the Lebesgue measure! ■

Figure 4.4 illustrates what the Lebesgue integral is designed to do in the context of the real line. Heuristically, we are weighting each value $f(x)$ of the function by the mass $\mu(\mathrm{d}x)$ located near x, and adding up these quantities. The *notation* $\mu(\mathrm{d}x)$ is designed to suggest the measure of an "infinitesimal set" containing the point x, which is something like the local density of mass at x. In Lecture 22, we will see that there is (sometimes) a deeper connection with densities.

Warning: The $\mathrm{d}x$ is a piece of notation, not a defined object. ■

To make complete sense of Definition 4.20, we need a few more observations. First of all, our assumption that f is measurable ensures that each super-level set $\{x \in \mathbb{R} : f(x) > t\}$ is a Borel set. Therefore, the measure μ of the super-level set is defined. Second, consider the function

$$h_\mu(t) := \mu\{x \in \mathbb{R} : f(x) > t\} \quad \text{for } t \geq 0.$$

The function h_μ is positive and decreasing. This type of function always has a well-defined (improper) Riemann integral, although the value can equal $+\infty$. See the discussion in Appendix C.

In other words, what the integral is actually doing is computing the measure μ of each super-level set $\{x \in \mathbb{R} : f(x) > t\}$ and summing over the levels t. This is the content of the rigorous definition.

4.3.2 Properties of the integral for positive functions

The Lebesgue integral for positive functions has a number of important properties that are easy to verify. We simply need to refer back to the definition and the invoke properties of the Riemann integral.

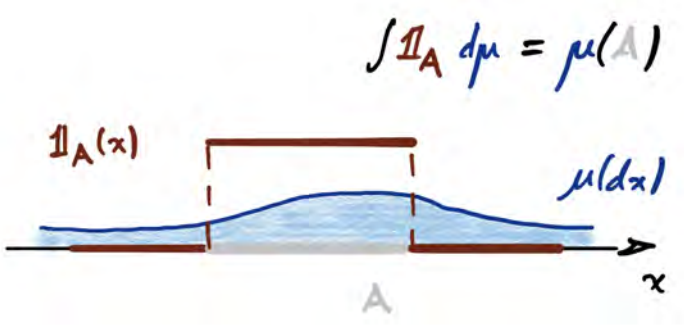


Figure 4.5 (Integral of an indicator). The integral of an indicator function of a Borel set equals the measure of the set.

Example 4.21 (Lebesgue integral: Indicators). Fix a Borel measure μ on the real line. Let $\mathsf{B} \in \mathscr{B}(\mathbb{R})$ be a Borel set. We can easily calculate the integral of the indicator $\mathbb{1}_{\mathsf{B}}$ of the set. Referring back to Example 4.6, we find that

$$\begin{aligned}\int_{\mathbb{R}} \mathbb{1}_{\mathsf{B}}(x)\, \mu(\mathrm{d}x) &= \int_0^\infty \mu\{x \in \mathbb{R} : \mathbb{1}_{\mathsf{B}}(x) > t\}\, \mathrm{d}t \\ &= \int_0^1 \mu(\mathsf{B})\, \mathrm{d}t + \int_1^\infty \mu(\emptyset)\, \mathrm{d}t = \mu(\mathsf{B}).\end{aligned}$$

We will often use basic properties of the Riemann integral without comment (here, domain decomposition). Thus, the integral of the indicator function of a set equals the measure of the set. See Figure 4.5 for an illustration. In particular, the integral of the zero function is zero. ∎

Example 4.22 (Lebesgue integral: Dirac measure). For $a \in \mathbb{R}$, consider the Dirac measure δ_a on the real line. It is illuminating to compute the integral of a positive, measurable function $f : \mathbb{R} \to \mathbb{R}_+$ with respect to the Dirac measure. Indeed,

$$\begin{aligned}\int_{\mathbb{R}} f(x)\, \delta_a(\mathrm{d}x) &= \int_0^\infty \delta_a\{x \in \mathbb{R} : f(x) > t\}\, \mathrm{d}t \\ &= \int_0^\infty \mathbb{1}_{[0,f(a))}(t)\, \mathrm{d}t = f(a).\end{aligned}$$

Indeed, the Dirac measure δ_a of the super-level set $\{x \in \mathbb{R} : f(x) > t\}$ equals one if and only if $0 \le t < f(a)$. This identity motivates the use of a "spike" to illustrate the Dirac measure. ∎

Exercise 4.23 (Lebesgue integral: Monotonicity for positive functions). Fix a Borel measure μ on the real line. For positive measurable functions $f, g : \mathbb{R} \to \overline{\mathbb{R}}_+$, show that the integral is monotone:

$$0 \le f \le g \text{ pointwise} \quad \text{implies} \quad \int_{\mathbb{R}} f(x)\, \mu(\mathrm{d}x) \le \int_{\mathbb{R}} g(x)\, \mu(\mathrm{d}x).$$

In particular, the integral of a positive function is positive. **Hint:** At each fixed level $t \ge 0$, compare the super-level set of f and g.

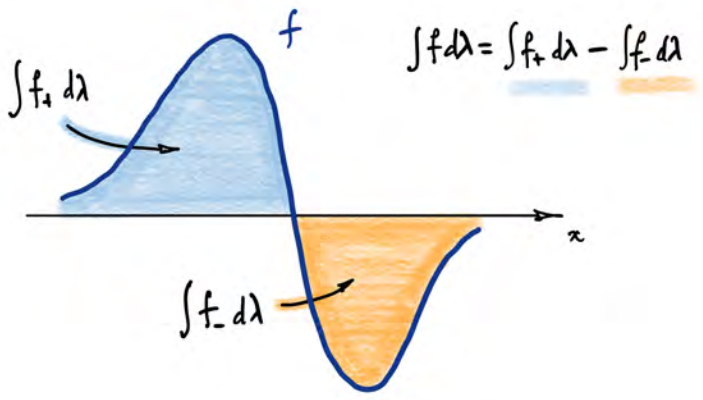


Figure 4.6 (Integral of a signed function). We define the integral of a signed function by integrating its positive and negative parts and computing the difference. This figure illustrates the case of an integral with respect to the Lebesgue measure λ.

Exercise 4.24 (Lebesgue integral: Positive homogeneity). Fix a Borel measure μ on the real line. Let $f : \mathbb{R} \to \overline{\mathbb{R}}_+$ be a positive, measurable function. Prove that

$$\int_{\mathbb{R}} \alpha f(x)\, \mu(\mathrm{d}x) = \alpha \int_{\mathbb{R}} f(x)\, \mu(\mathrm{d}x) \quad \text{for } \textit{positive } \alpha \in \mathbb{R}_+.$$

In particular, the integral of a positive constant c equals $c \cdot \mu(\mathbb{R})$. **Hint:** Make a linear change of variables in the definition of the integral.

4.3.3 The integral of a measurable function

To integrate a signed function, we just integrate the positive and negative parts separately and then combine the results; see Figure 4.6. To avoid competing infinities, we will require that the absolute value of the function has a finite integral.

Definition 4.25 (Integrable function). We say that a *finite-valued* measurable function $f : \mathbb{R} \to \mathbb{R}$ is *integrable* with respect to the Borel measure μ when

$$\int_{\mathbb{R}} |f(x)|\, \mu(\mathrm{d}x) < +\infty.$$

For brevity, we may also say that f is *μ-integrable*. The class of μ-integrable real-valued functions is often denoted as $\mathsf{L}_1(\mu)$.

Exercise 4.26 (Integrable function: Positive and negative parts). Suppose that $f : \mathbb{R} \to \mathbb{R}$ is integrable with respect to the Borel measure μ. Deduce that the positive and negative parts are integrable with respect to μ:

$$\int_{\mathbb{R}} f_+(x)\, \mu(\mathrm{d}x) < +\infty \quad \text{and} \quad \int_{\mathbb{R}} f_-(x)\, \mu(\mathrm{d}x) < +\infty.$$

Hint: Invoke Exercise 4.12 and Exercise 4.23.

With these preparations complete, we are prepared to define the Lebesgue integral of a signed function.

Definition 4.27 (Lebesgue integral). Let $f : \mathbb{R} \to \mathbb{R}$ be a function that is *integrable* with respect to the Borel measure μ. Then we may define the *Lebesgue integral* of f with respect to μ to be

$$\int_{\mathbb{R}} f(x)\, \mu(\mathrm{d}x) := \int_{\mathbb{R}} f_+(x)\, \mu(\mathrm{d}x) - \int_{\mathbb{R}} f_-(x)\, \mu(\mathrm{d}x).$$

Otherwise, the function f does not admit a Lebesgue integral.

Definition 4.27 presents no new complications. Exercise 4.26 guarantees that both the positive and negative parts of a measurable function have finite integrals, so there is no possibility of competing infinities ($\infty - \infty$) when we subtract their values.

Exercise 4.28 (Lebesgue integral: Consistency). Let $f : \mathbb{R} \to \mathbb{R}_+$ be a *positive*, measurable function that is integrable with respect to the Borel measure μ. Show that Definition 4.20 and Definition 4.27 give the same value.

Exercise 4.29 (Lebesgue integral: Absolute value). Assuming that $f : \mathbb{R} \to \mathbb{R}$ is integrable with respect to the Borel measure μ, check that

$$\left| \int_{\mathbb{R}} f(x)\, \mu(\mathrm{d}x) \right| \leq \int_{\mathbb{R}} |f(x)|\, \mu(\mathrm{d}x).$$

Hint: This argument only requires the definitions and monotonicity of the integral for positive functions. In particular, it is not necessary to check that the integral is linear.

Exercise 4.30 (Lebesgue integral: Basic properties). Without restriction to positive functions, show that the Lebesgue integral is monotone. Show that the integral is homogeneous: we can pass any finite scalar through the integral.

4.3.4 Integration over a set

It is often the case that we want to integrate a function over a subset of the domain. With Lebesgue integrals, this task can be accomplished in a straightforward way.

Definition 4.31 (Lebesgue integral: Subset). Fix a Borel measure μ on the real line and a measurable function $f : \mathbb{R} \to \overline{\mathbb{R}}$ whose Lebesgue integral with respect to μ is defined. For each Borel set $\mathsf{B} \in \mathscr{B}(\mathbb{R})$, we define the integral over the set via the expression

$$\int_{\mathsf{B}} f(x)\, \mu(\mathrm{d}x) := \int_{\mathbb{R}} \mathbb{1}_{\mathsf{B}}(x) f(x)\, \mu(\mathrm{d}x).$$

4.3.5 Linearity

The critical fact about the Lebesgue integral is that it is linear: the integral of a sum is the sum of the integrals.

Theorem 4.32 (Lebesgue integral: Linearity). Fix a Borel measure μ on the real line.

Let $f, g : \mathbb{R} \to \mathbb{R}$ be μ-integrable functions. Then

$$\int_{\mathbb{R}} (\alpha f + \beta g)(x)\, \mu(\mathrm{d}x) = \alpha \int_{\mathbb{R}} f(x)\, \mu(\mathrm{d}x) + \beta \int_{\mathbb{R}} g(x)\, \mu(\mathrm{d}x) \quad \text{for all } \alpha, \beta \in \mathbb{R}.$$

In Lecture 5, we will prove Theorem 4.32 in a more general setting. The argument is somewhat involved.

4.3.6 Negligible sets

Another basic fact about Lebesgue integrals is that they are insensitive to the values of a function on a negligible set. The following result encapsulates the basic fact.

Proposition 4.33 (Lebesgue integral: Negligible sets). Let $f, g : \mathbb{R} \to \overline{\mathbb{R}}$ be measurable functions. If f and g are equal μ-almost everywhere, then their integrals are equal:

$$\mu\{x \in \mathbb{R} : f(x) \neq g(x)\} = 0 \quad \text{implies} \quad \int_{\mathbb{R}} f(x)\, \mu(\mathrm{d}x) = \int_{\mathbb{R}} g(x)\, \mu(\mathrm{d}x).$$

In Lecture 5, we establish Proposition 4.33 in a more general setting.

Exercise 4.34 (Lebesgue integral: Rationals). For a positive, measurable function $f : \mathbb{R} \to \mathbb{R}_+$, compute the integral $\int_{\mathbb{R}} \mathbb{1}_{\mathbb{Q}}(x) f(x)\, \lambda(\mathrm{d}x)$ by pure thought.

4.3.7 Examples

As we have insinuated, the integral has a wide range of applications. Let us elaborate on some of the examples we have already mentioned.

Example 4.35 (Center of mass). Suppose that μ is a finite Borel measure on the real line that describes a distribution of physical mass (see Example 3.2). The center of mass m of the distribution is the point about which the total torque is zero:

$$\int_{\mathbb{R}} (x - m)\, \mu(\mathrm{d}x) = 0.$$

Indeed, the torque at m due to the mass at a point $x \in \mathbb{R}$ is the length $(x - m)$ of the lever arm times that local mass $\mu(\mathrm{d}x)$ at the point x. By the linearity property of the integral (Theorem 4.32),

$$m = \frac{1}{\mu(\mathbb{R})} \int_{\mathbb{R}} x\, \mu(\mathrm{d}x).$$

This computation requires the assumption that the identity function $x \mapsto x$ is integrable with respect to μ. Heuristically, the system cannot place too much mass at very long distances from the origin.

The construction of the integral allows us to give a unified way to compute the center of mass for any one-dimensional mechanical system, even if it involves both extended mass (like a rod) and point masses (like hanging weights). ▪

Example 4.36 (Expectation). Suppose that μ is a Borel probability measure on the real line, so $\mu(\mathbb{R}) = 1$. See Example 3.3 for an illustration. The expected value m of the distribution is the quantity

$$m = \int_{\mathbb{R}} x\, \mu(\mathrm{d}x).$$

In other words, we weight each outcome $x \in \mathbb{R}$ by the local probability mass $\mu(\mathrm{d}x)$ and sum. Once again, the computation requires the assumption that $x \mapsto x$ is integrable with respect to μ. Heuristically, the probability of very large values must not be too big. Note the analogy with the center of mass of a mechanical system.

The construction of the integral provides a unified way to compute the expectation of any distribution of probability, even if the distribution has continuous parts and point masses (like the lifetime of a lightbulb). ▪

4.4 Riemann versus Lebesgue

Students often ask: What is the relationship between the Riemann integral and the Lebesgue integral with respect to the Lebesgue measure? When we see an integral sign, do we interpret it as a Riemann integral or a Lebesgue integral? How do we calculate Lebesgue integrals in practice? This section speaks to these concerns.

The best way to think about the Lebesgue integral is to regard it as an *upgrade* of the Riemann integral. We use similar notation because the approaches are designed to accomplish the same goal. As with Riemann integrals, you can use calculus to evaluate Lebesgue integrals. With Lebesgue integrals, we gain some additional tools: a clear definition of the class of integrable functions and a suite of limit theorems (Lecture 5). Moreover, we can define Lebesgue integrals in a wider setting.

4.4.1 Riemann implies Lebesgue

For both Riemann and Lebesgue integrals, the geometric purpose is similar: they are designed to sum up function values. As a consequence, the two approaches usually give the same answer when they are both valid. In particular, every function that is (properly) Riemann integrable is also Lebesgue integrable.

Proposition 4.37 (Riemann implies Lebesgue). Suppose that $f : [a, b] \to \mathbb{R}$ is a (bounded) Riemann integrable function. Then f is also Lebesgue integrable with respect to the Lebesgue measure λ on the interval $[a, b]$, and the integrals coincide:

$$\int_{[a,b]} f(x)\, \lambda(\mathrm{d}x) = \int_a^b f(x)\, \mathrm{d}x.$$

Proof. See Appendix C.10. ■

Proposition 4.37 gives us immediate access to familiar tools for working with the Riemann integral, such as elementary antiderivatives, change of variables formulas, and so forth. We will take these calculus rules for granted, but see the Problems section for a taste. Similar results are valid for Riemann–Stieltjes integrals.

Warning 4.38 (Improper integrals). There are functions that are not Lebesgue integrable but still have *improper* Riemann integrals. A classic example is

$$\int_{-\infty}^{+\infty} \frac{\sin(\pi x)}{\pi x}\, \mathrm{d}x = 1.$$

The Lebesgue integral of the integrand f does not exist because f_+ and f_- both have infinite integrals with respect to Lebesgue measure. ▪

> **Aside:** There is a more general construction, called the Denjoy–Perron–Henstock–Kurzweil integral, that allows us to integrate a larger class of functions that includes all (improperly) Riemann-integrable functions and all Lebesgue-integrable functions on the real line.

4.4.2 All our integrals are Lebesgue integrals

In fact, once we are comfortable with Lebesgue integrals and their properties, we can forget about Riemann integrals entirely. The next result recasts the formula in Definition 5.11 purely in terms of Lebesgue integrals.

Proposition 4.39 (Integration by parts). Let μ be a Borel measure, and let $f : \mathbb{R} \to \overline{\mathbb{R}}_+$ be a *positive*, measurable function. Then

$$\begin{aligned}\int_{\mathbb{R}} f(x)\, \mu(\mathrm{d}x) &:= \int_0^\infty \mu\{x \in \mathbb{R} : f(x) > t\}\, \mathrm{d}t \\ &= \int_{\mathbb{R}_+} \mu\{x \in \mathbb{R} : f(x) > t\}\, \lambda(\mathrm{d}t).\end{aligned}$$

In this expression, we interpret the right-hand side as a *Lebesgue* integral over the positive real line!

Proof. This boils down to an application of Proposition 4.37. See Appendix C.11 for the details. ∎

As a consequence of this statement, you can regard every integral you see in this class as a Lebesgue integral, unless explicitly stated otherwise. To perform concrete computations, you may still rely on familiar calculus rules. Concerning the terminology, Proposition 4.39 is the simplest case of the integration by parts formula; see Problem 6.26 for a generalization.

4.4.3 Outlook

So what do we gain by using the Lebesgue integral? We obtain flexibility and tools that are simply not available if we stick with Riemann integrals. In the next lecture, we will give an abstract treatment of the Lebesgue integral that provides all these benefits.

First, we use the same procedure to define the Lebesgue integral on domains that are more general than the real line. Then we will be able to integrate real-valued functions defined on an arbitrary measure space. This extension is critical when we develop an axiomatic model of probability theory (Lecture 7).

Second, Lebesgue integrals are equipped with a robust convergence theory, which allows us to compute the integral of a sequence of functions. These results play a crucial role in analysis, and they will also support the development of limit theorems in probability.

Problems

Exercise 4.40 (Elementary antiderivatives). In integral calculus, we learn to compute definite Riemann integrals using antiderivatives. Similar results hold for Lebesgue integrals with respect to the Lebesgue measure. In this problem, we assume that $a \le b$ and $a, b \in \mathbb{R}$. In each case, first argue that the integrand is a measurable function.

1. **Powers:** For real $p \neq 1$, use a direct argument to confirm that the Lebesgue integral of the power function satisfies
$$\int_{[a,b]} x^p \, \lambda(\mathrm{d}x) = \frac{1}{p+1}(b^{p+1} - a^{p+1}) \quad \text{assuming } 0 < a \leq b.$$
Hint: Use Definition 4.20 to refer the matter back to a familiar Riemann integral.
2. **Reciprocals:** By direct argument, show that the Lebesgue integral of the reciprocal function satisfies
$$\int_{[a,b]} x^{-1} \, \lambda(\mathrm{d}x) = \log(b) - \log(a) \quad \text{provided that } 0 < a \leq b.$$
3. **Exponentials** By direct argument, show that the Lebesgue integral of the exponential function satisfies
$$\int_{[a,b]} \mathrm{e}^{\theta x} \, \lambda(\mathrm{d}x) = \theta^{-1}(\mathrm{e}^{\theta b} - \mathrm{e}^{\theta a}) \quad \text{for } \theta \neq 0.$$
4. ***Cosines:** By direct argument, show that the Lebesgue integral of the cosine function satisfies
$$\int_{[a,b]} \cos(\theta x) \, \lambda(\mathrm{d}x) = \theta^{-1}(\sin(\theta b) - \sin(\theta a)) \quad \text{for } \theta \neq 0.$$
Hint: Decompose the domain $[a, b]$ of integration into regions where cos is monotone.
5. **FTC:** From these examples, you can see that the direct approach to computing Lebesgue integrals can be inefficient or difficult. Fortunately, there is a general result. Suppose that $f : [a, b] \to \mathbb{R}$ is a continuous function. Let $F : [a, b] \to \mathbb{R}$ be an antiderivative; that is, $F' = f$ on (a, b). Use Proposition 4.37 to check that the Lebesgue integral satisfies the Fundamental Theorem of Calculus (FTC):
$$\int_{[a,b]} f(x) \, \lambda(\mathrm{d}x) = F(b) - F(a).$$
Explain how this result yields all of the previous statements.
6. **Change of variables:** Let $u : [A, B] \to [a, b]$ be a strictly increasing, continuously differentiable function, and suppose that $f : [a, b] \to \mathbb{R}$ is continuously differentiable. Use the FTC to confirm that
$$\int_{[A,B]} f'(u(x))u'(x) \, \lambda(\mathrm{d}x) = \int_{[a,b]} f'(x) \, \lambda(\mathrm{d}x).$$

Exercise 4.41 (Powers: Integrability). Let p be a real number. Consider the functions $f_p, g_p : \mathbb{R} \to \mathbb{R}$ defined by
$$f_p(x) = \begin{cases} x^p, & x \geq 1 \\ 0, & \text{otherwise;} \end{cases} \qquad g_p(x) = \begin{cases} x^p, & 0 < x < 1 \\ 0, & \text{otherwise.} \end{cases}$$
Convince yourself that f_p and g_p are measurable. For which values of p are the functions f_p integrable with respect to λ? What about g_p?

Exercise 4.42 (Layer-cake representation). Let $f : \mathbb{R} \to \mathbb{R}_+$ be a positive, measurable function. Show that
$$f(a) = \int_0^\infty \mathbb{1}_{\{x \in \mathbb{R} : f(x) > t\}}(a) \, \lambda(\mathrm{d}t).$$
Hint: Apply Proposition 4.39 with the Dirac measure δ_a.

Notes

The quotation at the head of this chapter is from the wonderful speculative fiction writer, Octavia Butler. She was born and raised here in Pasadena and Altadena. Her work received multiple Hugo and Nebula prizes, and she was the first speculative fiction author to be awarded a MacArthur Fellowship. Her archive is held by the Huntington Library. I particularly recommend *Kindred* [But03].

Our presentation of the Lebesgue integral, using super-level sets, is adapted from the analysis textbook [LL01] of Lieb & Loss. This approach has several benefits. It gives a clear geometric picture of what the Lebesgue integral is doing, and it motivates the definition of the class of measurable functions. On the other hand, this construction requires the use of the improper Riemann integral, and we will need to rely on theoretical facts about the Riemann integral.

Most books on probability theory and measure theory define the Lebesgue integral using approximation by simple functions (Section 5.6). This approach is more self-contained because it does not rely on properties of the Riemann integral, and it also extends more readily to functions that take values in a linear space. On the other hand, it also has some deficiencies. The construction based on simple functions makes it hard to appreciate where measurability comes from and why it is essential for the integral. (Roughly, the positive measurable functions are the increasing limits of positive simple functions.) Furthermore, it also requires a nontrivial argument to prove that the integral of a positive simple function is well-defined.

Altogether, the approach via super-level sets seems more intuitive. The mathematically oriented reader should understand both perspectives, in part because simple functions play an important role in proving facts about integrals.

There are at least two more approaches to defining the Lebesgue integral, using ideas from functional analysis. This perspective originally emerged from Bourbaki's program. It has been championed by Peter Lax [Lax02] and Barry Simon [Sim15]; David Pollard [Pol02] also expresses admiration.

Suppose that we want to construct the Lebesgue integral with respect to the Lebesgue measure on the compact interval $[0, 1]$. The first approach begins with the linear space of continuous functions on $[0, 1]$, equipped with the $\mathsf{L}_1(\lambda)$ norm. In this special case, the L_1 norm can be defined using an ordinary Riemann integral. This normed linear space is completed to obtain the Banach space $\mathsf{L}_1(\lambda)$ of equivalence classes of λ-integrable functions. We can extend this idea to other measures by using the Riemann–Stieltjes integral.

The second approach begins with the convex cone $\mathsf{M}_+(\mathsf{X})$ of positive measurable functions on a measurable space $(\mathsf{X}, \mathcal{F})$. An integral is defined to be a positive functional on $\mathsf{M}_+(\mathsf{X})$ that is positive homogeneous, additive, and satisfies a monotone convergence rule. By the Riesz–Kakutani representation theorem, these positive functions are in one-to-one correspondence with positive measures. This approach places the abstract properties of an integral front and center, although the construction is not particularly concrete. See [Pol02, Sec. 2.3] and Problem 5.45.

Although these ideas are elegant, they demand comfort with functional analysis. For this course, we prefer to start with a more elementary construction of the integral. This approach helps us gain intuition for functional analysis in a relatively simple setting.

Lecture bibliography

[But03] O. E. Butler. *Kindred*. Beacon, 2003.

[Lax02] P. D. Lax. *Functional analysis*. Wiley-Interscience, 2002.

[LL01] E. H. Lieb and M. Loss. *Analysis*. 2nd ed. American Mathematical Society, 2001. DOI: 10.1090/gsm/014.

[Pol02] D. Pollard. *A user's guide to measure theoretic probability*. Cambridge University Press, 2002.

[Sim15] B. Simon. *Real analysis*. With a 68 page companion booklet. American Mathematical Society, 2015. DOI: 10.1090/simon/001.

5. Abstract Integration

> "**Agent Dale Cooper:** Wait a minute, wait a minute. You know, this is—excuse me—a *damn* fine cup of coffee. I've had I can't tell you how many cups of coffee in my life, and this... this is one of the best."
>
> —*Twin Peaks*, 2001

Agenda:

1. Measurable functions
2. The Lebesgue integral
3. Convergence theorems
4. Almost everywhere convergence
5. Proof of integral properties

In the last lecture, we introduced the Lebesgue integral with respect to a Borel measure on the real line. This integral adds up the values of a function, weighted by a distribution of mass. From the construction, we can perceive an opportunity to define the integral of a real-valued function on a measure space. In this lecture, we will give a complete treatment of integration on an abstract measure space, along with the major convergence theorems for integrals. These foundations are essential for an axiomatic treatment of probability theory.

For most readers, the key concepts are the properties of the integral, packaged in Theorem 5.14, and the three major convergence theorems (monotone convergence, Fatou's lemma, and dominated convergence). The proofs in this lecture are not particularly hard, but they are somewhat involved. The technical details are not really necessary for a working understanding of the subject, so most of the material is starred. More mathematically oriented reader will want to understand the arguments, which is why they are included here.

In this lecture, we fix a general measure space $(\mathsf{X}, \mathscr{F}, \mu)$ with domain X, sigma-algebra $\mathscr{F}$, and measure μ. This is all of the structure that we need to develop a theory of integration.

5.1 Compact notation for set-builder

We will often be working with functions. The standard set-builder notation for sets of function values quickly becomes cumbersome. This seems like a good time to introduce a more efficient script.

Our new notation delineates the set of points where a function satisfies some condition. For example, suppose that $f, g : \mathsf{X} \to \mathbb{R}$. Then we may write things like

$$\{f > g\} := \{x \in \mathsf{X} : f(x) > g(x)\};$$
$$\{f = g\} := \{x \in \mathsf{X} : f(x) = g(x)\}.$$

We will use many similar expressions without further comment. It takes a little practice to get used to this convention. But it saves a lot of letters, which ultimately makes things easier to understand.

5.2 The space of measurable functions

As before, we can only integrate measurable functions. In this section, we quickly introduce the appropriate definitions and summarize the stability properties of measurable functions. Last, we describe how to approximate positive measurable functions by means of *positive simple functions*.

5.2.1 Measurable functions

To begin, we need to introduce the appropriate concept of a measurable function on the domain. See Problem 5.36 for more context and further extensions.

Definition 5.1 (Measurable function). Let $(\mathsf{X}, \mathcal{F})$ be a measurable space. A function $f : \mathsf{X} \to \overline{\mathbb{R}}$ taking extended real values is *measurable* when

$$f^{-1}(t, +\infty] \in \mathcal{F} \quad \text{for all } t \in \overline{\mathbb{R}}.$$

In other words, the preimage of each semi-infinite interval must be a measurable set. For emphasis, we may say that f is an $\mathcal{F}$-measurable function.

Warning: The definition of a measurable function does not involve a measure. ∎

Exercise 5.2 (*Measurability). Prove that a function $f : \mathsf{X} \to \overline{\mathbb{R}}$ is $\mathcal{F}$-measurable if and only if

$$f^{-1}(\mathsf{B}) \in \mathcal{F} \quad \text{for all extended Borel sets } \mathsf{B} \in \mathcal{B}(\overline{\mathbb{R}}).$$

Hint: See the proof of Proposition 4.2.

Exercise 5.3 (*Measurability: Continuous functions). Suppose that the measurable space $(\mathsf{X}, \mathcal{F})$ is also a topological space where every open set belongs to $\mathcal{F}$. Verify that each continuous function $f : \mathsf{X} \to \mathbb{R}$ is measurable.

Although the definition of an $\mathcal{F}$-measurable function is similar to the definition of an (extended) Borel measurable function (Definition 4.5), we would like to inject a note of caution in this general setting.

Warning 5.4 (Measurability: Role of measurable sets). The definition of a measurable function depends on the class $\mathcal{F}$ of measurable sets. When $\mathcal{F}$ is a small σ-algebra, the stock of measurable functions may be limited. This point is not central right now, but it will play a role in probability theory when we discuss conditioning. ∎

Exercise 5.5 (*Measurability: Trivial σ-algebra). Consider a domain X equipped with the trivial σ-algebra $\mathcal{F} = \{\emptyset, \mathsf{X}\}$. Give a complete description of the class of measurable functions $f : \mathsf{X} \to \overline{\mathbb{R}}$.

Repeat this exercise for the almost trivial σ-algebra $\mathcal{F} = \{\emptyset, \mathsf{A}, \mathsf{A}^c, \mathsf{X}\}$ where $\mathsf{A} \subseteq \mathsf{X}$ is a subset of the domain.

5.2.2 Stability properties and limits

As with Borel measurable functions, the measurable functions on a measurable space are stable under a wide range of operations. The proofs parallel those in Sections 4.2.4 and 4.2.5, so we simply collect the results.

Proposition 5.6 (Measurable functions: Algebraic operations). Fix a measurable space $(\mathsf{X}, \mathcal{F})$. Let $f, g : \mathsf{X} \to \overline{\mathbb{R}}$ be measurable functions.

1. **Constants:** Constant functions are measurable.

2. **Indicators:** For a measurable set $\mathsf{A} \in \mathscr{F}$, the indicator function $\mathbb{1}_{\mathsf{A}}$ is a measurable function.
3. **Sign parts:** The positive part f_+, the negative part f_-, and the absolute value $|f|$ are measurable functions.
4. **Min and max:** The minimum $f \wedge g$ and the maximum $f \vee g$ are measurable.
5. **Sum:** The sum $f + g$ is measurable, *provided* that f, g are both positive or both finite-valued.
6. **Product:** The product fg is measurable, *provided* that f, g are both positive or both finite-valued.
7. **Linear space:** In particular, the finite-valued measurable functions compose a linear space (in fact, an algebra).

The notation for this linear space is not standardized.

Proposition 5.7 (Measurable functions: Countable operations). Fix a measurable space $(\mathsf{X}, \mathscr{F})$. For each $j \in \mathbb{N}$, let $f_j : \mathsf{X} \to \overline{\mathbb{R}}$ be a measurable function.

1. **Infimum and supremum:** The infimum, $\inf_{j\in\mathbb{N}} f_j$, and the supremum, $\sup_{j\in\mathbb{N}} f_j$, are measurable.
2. **Inferior and superior limits:** The inferior limit, $\liminf_{j\to\infty} f_j$, and the superior limit, $\limsup_{j\to\infty} f_j$, are measurable.
3. **Limits:** If $\lim_{j\to\infty} f_j$ exists pointwise, then it is measurable.
4. **Set of convergence:** The set $\{x \in \mathsf{X} : \lim_{j\to\infty} f_j(x) \text{ exists}\}$ is measurable.

Exercise 5.8 (Measurable functions). Prove Proposition 5.6 and Proposition 5.7.

5.2.3 *Positive simple functions

To establish properties of the integral, it is helpful to start with functions where we can compute the value of the integral with our bare hands. To that end, we introduce the class of *positive* linear combinations of indicator functions:

$$\mathsf{SF}_+ := \mathsf{SF}_+(\mathscr{F}) := \left\{ \sum\nolimits_{i=1}^{n} \alpha_i \mathbb{1}_{\mathsf{A}_i} : \alpha_i \in \mathbb{R}_+ \text{ and } \mathsf{A}_i \in \mathscr{F} \text{ and } n \in \mathbb{N} \right\}. \tag{5.1}$$

The elements of the class SF_+ are called *positive simple functions*. The key property of a simple function is that it takes only a finite number of values. We anticipate that the integral of a positive simple function satisfies the linearity relation

$$\text{“} \qquad \int_{\mathsf{X}} \left(\sum\nolimits_{i=1}^{n} \alpha_i \mathbb{1}_{\mathsf{A}_i} \right) \mu(\mathrm{d}x) = \sum\nolimits_{i=1}^{n} \alpha_i \mu(\mathsf{A}_i). \qquad \text{”}$$

Our construction of the integral leads to this result, but it is not a triviality.

Exercise 5.9 (Simple functions: Limit superior). Confirm that each positive simple function is measurable. For a sequence $(s_j : j \in \mathbb{N})$ of positive simple functions, explain why $\limsup_{j\to\infty} s_j$ is measurable.

Simple functions are important because we can approximate every positive measurable function by a simple function that is pointwise smaller; see Figure 5.1. To establish this fact, we introduce the *staircase maps* $Q_j : \overline{\mathbb{R}}_+ \to \mathbb{R}_+$ for each $j \in \mathbb{N}$:

$$Q_j(x) := \begin{cases} j, & x > j; \\ (i-1)2^{-j}, & (i-1)2^{-j} < x \le i2^{-j} \le j \quad \text{for } i \in \mathbb{N}; \\ 0, & x = 0. \end{cases} \tag{5.2}$$

The function Q_j quantizes and thresholds positive, extended real numbers.

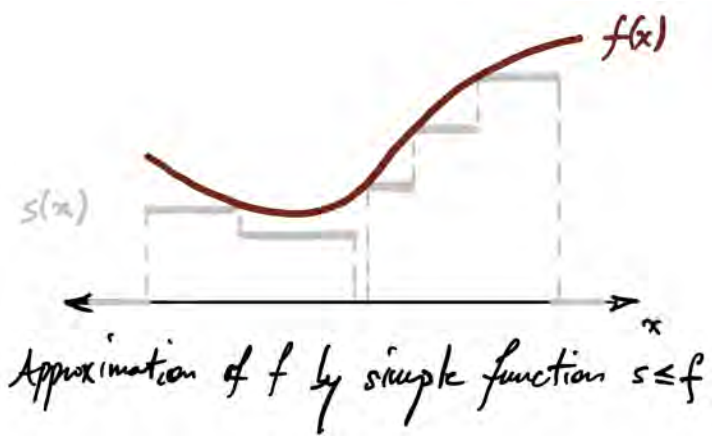


Figure 5.1 (Staircase approximation). Every positive measurable function can be approximated by a positive simple function.

Exercise 5.10 (Staircase approximation). Let $f : \mathsf{X} \to \overline{\mathbb{R}}_+$ be a positive, measurable function.

1. Prove that $Q_j \circ f$ is a positive simple function for each $j \in \mathbb{N}$.
2. Check that $Q_j \circ f \leq f$ pointwise for each $j \in \mathbb{N}$.
3. Show that $(Q_j \circ f) \uparrow f$ pointwise as $j \to \infty$.

Later, we will use the staircase approximation in conjunction with a fundamental limit theorem to extend results for the integral on positive simple functions to all positive measurable functions.

5.3 The Lebesgue integral

In this section, we begin the development of the Lebesgue integral on the measure space $(\mathsf{X}, \mathcal{F}, \mu)$. We first consider the integral of a positive function. Then we extend the integral to signed functions by passing to the positive and negative parts. Afterward, we outline the main properties of the integral.

5.3.1 Positive functions

To integrate a positive-valued function over a measure space, we simply compute the total measure of the super-level sets of the function.

Definition 5.11 (Lebesgue integral: Positive functions). Consider a *positive*, measurable function $f : \mathsf{X} \to \overline{\mathbb{R}}$. The Lebesgue integral of f with respect to μ is defined as

$$\int_{\mathsf{X}} f(x)\, \mu(\mathrm{d}x) := \int_0^{+\infty} \mu\{x \in \mathsf{X} : f(x) > t\}\, \mathrm{d}t.$$

The right-hand side is an improper Riemann integral (Appendix C), and it is always well-defined.

As before, the right-hand side gives sense to the Lebesgue integral. It allows us to apply familiar tools for Riemann integrals, such as elementary antiderivatives, change of variables, and so forth.

5.3.2 Signed functions

To integrate a signed function over a measure space, we first define the class of integrable functions.

Definition 5.12 (Integrable function). We say that a *finite-valued* measurable function $f : \mathsf{X} \to \mathbb{R}$ is *integrable* with respect to the measure μ when

$$\int_{\mathsf{X}} |f(x)| \, \mu(\mathrm{d}x) < +\infty.$$

In this case, we may also say that f is *μ-integrable*.

The positive and negative parts of an integrable function are also integrable. Therefore, we may define the integral of a signed function in terms by integrating the positive and negative parts separately.

Definition 5.13 (Lebesgue integral). Let $f : \mathsf{X} \to \mathbb{R}$ be a function that is integrable with respect to the measure μ. Then we may define the *Lebesgue integral* to be

$$\int_{\mathsf{X}} f(x) \, \mu(\mathrm{d}x) := \int_{\mathsf{X}} f_+(x) \, \mu(\mathrm{d}x) - \int_{\mathsf{X}} f_-(x) \, \mu(\mathrm{d}x).$$

In this case, we may also say that f is *μ-integrable*.

As before, the integrability assumption ensures that the integral is well-defined. For positive functions, Definition 5.13 is consistent with Definition 5.11.

5.3.3 Notation for integrals

There are many common notations for Lebesgue integrals, and you should be familiar with them so that you can fluently read the mathematical literature. First, there are several alternative expressions for the differential:

$$\int_{\mathsf{X}} f(x) \, \mu(\mathrm{d}x) =: \int_{\mathsf{X}} f(x) \, \mathrm{d}\mu(x) =: \int_{\mathsf{X}} f \, \mathrm{d}\mu =: \int f \, \mathrm{d}\mu.$$

These expressions all mean the same thing. As in the last term, we may omit the domain of integration, in which case the integration takes place over the full domain of the integrand f. We introduce parallel notation for the integral over a subdomain:

$$\int_{\mathsf{A}} f(x) \, \mu(\mathrm{d}x) := \int_{\mathsf{X}} \mathbb{1}_{\mathsf{A}}(x) \, f(x) \, \mu(\mathrm{d}x) \quad \text{for measurable } \mathsf{A} \in \mathcal{F}.$$

The left-hand integral may be abbreviated further, as in the penultimate display. It is also very convenient to use functional notation:

$$\mu(f) := \int_{\mathsf{X}} f(x) \, \mu(\mathrm{d}x) \quad \text{or} \quad \mu(f; \mathsf{A}) := \int_{\mathsf{A}} f(x) \, \mu(\mathrm{d}x).$$

Throughout these notes, we vary how we write the integral, depending on what part of the formula requires your attention.

For integrals with respect to the Lebesgue measure λ on the measurable space $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$, the differential $\lambda(\mathrm{d}x)$ is often written more compactly as $\mathrm{d}x$. Thus,

$$\int_{\mathsf{A}} f(x) \, \lambda(\mathrm{d}x) = \int_{\mathsf{A}} f(x) \, \mathrm{d}\lambda(x) = \int_{\mathsf{A}} f \, \mathrm{d}\lambda = \int_{\mathsf{A}} f(x) \, \mathrm{d}x.$$

These notations are consistent with the familiar notation for Riemann integrals.

5.3.4 Structural properties

We have developed the Lebesgue integral in two stages. First, we defined the integral for all positive functions. Second, we defined the integral for all measurable functions, with some restrictions to avoid problems with infinities.

In parallel, we will generally present facts about integrals in two parts, one for positive functions (that may have integral $+\infty$) and one for signed functions (requiring the integral to be finite).

We may now state an omnibus theorem that describes the major properties of the Lebesgue integral. For succinctness, we include results for both positive and integrable functions together, but we emphasize that these cases are slightly different in spirit.

Theorem 5.14 (Lebesgue integral: Properties). Let $(\mathsf{X}, \mathcal{F}, \mu)$ be a measure space. Let $f, g : \mathsf{X} \to \overline{\mathbb{R}}$ be measurable functions whose integrals $\mu(f)$ and $\mu(g)$ are defined.

Refer back to Section 5.3.3 for the functional notation for integrals.

1. **Zero:** The integral of the zero function is zero: $\mu(0) = 0$.
2. **Indicators:** For a measurable set $\mathsf{A} \in \mathcal{F}$, the integral $\mu(\mathbb{1}_{\mathsf{A}}) = \mu(\mathsf{A})$.
3. **Positivity:** For a *positive* function $f \geq 0$, the integral is positive: $\mu(f) \geq 0$.
4. **Monotonicity:** If $f \leq g$, then $\mu(f) \leq \mu(g)$.
5. **Positive linearity:** For *positive* functions f, g and *positive* scalars $\alpha, \beta \geq 0$,
$$\mu(\alpha f + \beta g) = \alpha\mu(f) + \beta\mu(g).$$
6. **Linearity:** For *μ-integrable* functions f, g and *all* scalars $\alpha, \beta \in \mathbb{R}$,
$$\mu(\alpha f + \beta g) = \alpha\mu(f) + \beta\mu(g).$$
7. **Almost definite:** For a *positive* function $f \geq 0$, a zero integral $\mu(f) = 0$ implies that $f = 0$ μ-almost everywhere.
8. **Almost equal:** If $f = g$ μ-almost everywhere, then $\mu(f) = \mu(g)$.

Proof. See Section 5.5 for a complete proof of Theorem 5.14. ∎

Let us take a moment to discuss the contents of Theorem 5.14. First, we remark that the results are not independent from each other; some of them can easily be deduced from others. Second, keep in mind that positive functions and their integrals are allowed to take the value $+\infty$, while signed functions and their integrals must remain finite.

Properties (1) and (2) allow us to evaluate specific elementary Lebesgue integrals. The key fact here is that the integral of the indicator of a measurable set equals the measure of the set. This always remains true—even for very complicated sets.

Properties (3) and (4) concern the interaction between the integral and the pointwise order on functions. Point (3) states that the integral is a positive operator: it maps positive functions to positive numbers. Point (4) shows that the integral respects the pointwise partial order relation on functions.

Properties (5) and (6) are both related to the homogeneity and additivity of the integral. We can pass scalars through the integral, and the integral of a sum is the sum of the integrals. Note that there is a subtle distinction between the statements that reflects the differences between positive and signed functions. *Linearity is the single most important property of the Lebesgue integral.*

Last, Properties (7) and (8) state that the integral is insensitive to the value of a function on a negligible set. In particular, for a function f that is zero μ-almost

everywhere, the integral $\mu(f) = 0$. For positive functions only, the converse of the latter statement is also true.

Most of the results in Theorem 5.14 are easy to derive from Definitions 5.11 and 5.13 of the Lebesgue integral. The two results on linearity make for a striking exception. They require a significant number of intermediate steps, and they hinge on the monotone convergence theorem (Theorem 5.18).

Exercise 5.15 (Integral properties). Try to prove items (1)–(4) and (7)–(8) in Theorem 5.14. **Hint:** See Section 4.3.2.

5.3.5 The linear space of integrable functions

It is valuable to develop notation for the class of signed functions whose Lebesgue integral is defined.

Definition 5.16 (Space of integrable functions). Let $(\mathsf{X}, \mathscr{F}, \mu)$ be a measure space. Define the functional

$$\|f\|_{\mathsf{L}_1(\mu)} := \int_{\mathsf{X}} |f(x)| \, \mu(\mathrm{d}x) \quad \text{for measurable } f : \mathsf{X} \to \mathbb{R}.$$

Introduce the class of finite-valued functions with a finite integral:

$$\mathsf{L}_1(\mu) := \left\{ f : \mathsf{X} \to \mathbb{R} \text{ measurable} : \|f\|_{\mathsf{L}_1(\mu)} < +\infty \right\}.$$

The class $\mathsf{L}_1(\mu)$ is called the space of μ-integrable functions.

Exercise 5.17 (The space of integrable functions). Explain in detail why $\mathsf{L}_1(\mu)$ is a linear space. Show that $\|\cdot\|_{\mathsf{L}_1(\mu)}$ is a pseudonorm on $\mathsf{L}_1(\mu)$.

A *pseudonorm* is a positive function that is positive homogeneous and satisfies the triangle inequality. The pseudonorm of a nonzero function may be equal to zero.

The linear space $\mathsf{L}_1(\mu)$ has central importance in functional analysis and probability theory. We postpone a detailed discussion to Lecture 11.

5.4 Convergence theorems

A core technical problem in analysis is to understand when we can interchange two limiting processes. In particular, we may ask when we can pass a limit through an integral. A major deficiency of the Riemann integral is the lack of a crisp answer to this question. In contrast, the Lebesgue integral is designed so that limits behave predictably. It is *not true* that we can always swap a Lebesgue integral with a limit, but we can perform this operation under simple and easily verified conditions.

5.4.1 Monotone convergence

Levi's monotone convergence theorem is the central fact in the theory of Lebesgue integration. For an *increasing* sequence of positive functions, the limit of the integrals equals the integral of the limit. We can use this result to establish more flexible convergence theorems, and it also plays a critical role in developing other properties of the integral (including linearity!). See the Problems section for some applications.

An *increasing* sequence of functions satisfies

$$f_{j+1}(x) \geq f_j(x)$$

for all $x \in \mathsf{X}$ and $j \in \mathbb{N}$. An increasing sequence always has a pointwise limit $f : \mathsf{X} \to \overline{\mathbb{R}}$ that can take the value $+\infty$.

Theorem 5.18 (Lebesgue integral: Monotone convergence). Let $(\mathsf{X}, \mathscr{F}, \mu)$ be a measure space. Consider a pointwise *increasing* sequence $(f_j : \mathsf{X} \to \overline{\mathbb{R}}_+)_{j \in \mathbb{N}}$ of *positive*,

measurable functions. That is,

$$f_j(x) \uparrow f(x) \quad \text{for each } x \in \mathsf{X}.$$

Then the sequence of Lebesgue integrals increases, converging to its limiting value:

$$\mu(f_j) \uparrow \mu(f).$$

Let us emphasize that the proof of Theorem 5.18 only depends on Definition 5.11. It uses none of the properties listed in Theorem 5.14.

Proof. Since the functions are increasing, the limiting function $f = \sup_j f_j$ is positive and measurable (Example 4.16). Therefore, its integral $\mu(f)$ is defined.

The super-level sets of the functions f_j compose an increasing sequence of sets:

$$\{f_j > t\} \uparrow \{f > t\} \quad \text{as } j \to \infty \text{ for each } t \geq 0.$$

Indeed, the increase follows from the fact that $f_j(x) > t$ implies that $f_{j+1}(x) > t$. We obtain the limit from the observation that $f(x) > t$ if and only if $\sup_j f_j(x) > t$.

On the interval $t \geq 0$, define the positive, decreasing functions

$$h_j(t) := \mu\{f_j > t\} \quad \text{for } j \in \mathbb{N};$$
$$h(t) := \mu\{f > t\}.$$

By the increasing limit property of a measure (Proposition 2.30), we see that $h_j(t) \uparrow h(t)$ for each $t \geq 0$. We may conclude that

$$\mu(f_j) = \int_0^\infty h_j(t)\,\mathrm{d}t \uparrow \int_0^\infty h(t)\,\mathrm{d}t = \mu(f),$$

via doubly monotone convergence of Riemann integrals (Theorem C.8). ∎

Exercise 5.19 (Monotone convergence: Integrable functions). Extend Theorem 5.18 to the setting where $(f_j : j \in \mathbb{N})$ is a pointwise increasing sequence of *μ-integrable* functions. **Hint:** Use linearity, which itself is a consequence of Theorem 5.18.

5.4.2 Fatou's lemma

We continue with another convergence result, which gives a lower bound on the smallest values attained by a sequence of integrals.

Theorem 5.20 (Lebesgue integral: Fatou's lemma). Let $(\mathsf{X}, \mathcal{F}, \mu)$ be a measure space. Consider a sequence $(f_j : \mathsf{X} \to \overline{\mathbb{R}}_+)_{j\in\mathbb{N}}$ of *positive*, measurable functions. Then the inferior limit of integrals is bounded below:

$$\liminf_{j\to\infty} \mu(f_j) \geq \mu\big(\liminf_{j\to\infty} f_j\big).$$

Warning: This result is false without the positivity assumption! ∎

This argument uses some of the simpler properties of the integral from Theorem 5.14.

Proof. The function $f := \liminf_{j\to\infty} f_j$ is positive and measurable (Exercise 4.17), so its integral $\mu(f)$ is defined.

Recall that the inferior limit can be expressed as

$$f = \liminf_{j\to\infty} f_j = \lim_{j\to\infty} \inf_{k\geq j} f_k =: \lim_{j\to\infty} g_j.$$

For each $j \in \mathbb{N}$, we have introduced the positive, measurable function $g_j := \inf_{k \geq j} f_k$. By construction, the sequence $(g_j : j \in \mathbb{N})$ is pointwise increasing. Furthermore, $\mu(g_j) \leq \mu(f_k)$ for each $k \geq j$ by the monotonicity of the integral (Theorem 5.14).

An application of monotone convergence (Theorem 5.18) delivers

$$\begin{aligned}\mu(f) = \mu(\lim_{j\to\infty} g_j) &= \lim_{j\to\infty} \mu(g_j) \\ &\leq \lim_{j\to\infty} \inf_{k\geq j} \mu(f_k) = \liminf_{j\to\infty} \mu(f_j).\end{aligned}$$

This is the required result. ∎

A convenient feature of Theorem 5.20 is that it requires neither the sequence of functions nor their integrals to have a limit. As such, we can apply it impulsively, without stopping to check that the limits exist. The cost for this flexibility is that the theorem only yields a lower bound.

Exercise 5.21 (Fatou gap). Find a sequence $(f_j : \mathbb{R} \to \mathbb{R})_{j\in\mathbb{N}}$ of positive functions on the real line where $\liminf_{j\to\infty} \lambda(f_j) > \lambda(\liminf_{j\to\infty} f_j)$. **Hint:** The mass can "leak out" at $\pm\infty$.

> **Aside:** Fatou's lemma states that the Lebesgue integral is lower semicontinuous on the class of positive, measurable functions.

5.4.3 Dominated convergence

The dominated convergence theorem is our main workhorse when we need to take limits of Lebesgue integrals. It gives a simple sufficient condition under which we can exchange the integral with a limit. See the Problems section for some applications.

Theorem 5.22 (Lebesgue integral: Dominated convergence). Let $(\mathsf{X}, \mathcal{F}, \mu)$ be a measure space. Consider a pointwise *convergent* sequence $(f_j : \mathsf{X} \to \mathbb{R})_{j\in\mathbb{N}}$ of measurable functions: $f_j \to f$. Suppose that each function in the sequence is dominated by a *fixed, integrable* function:

$$|f_j| \leq |g| \text{ for each } j \in \mathbb{N} \quad \text{where} \quad g \in \mathsf{L}_1(\mu).$$

Then we can exchange the limit with the integral:

$$\lim_{j\to\infty} \mu(f_j) = \mu\big(\lim_{j\to\infty} f_j\big).$$

That is, $\mu(f_j) \to \mu(f)$.

> **Warning:** The dominating function g cannot depend on the index j. ∎

Unlike monotone convergence, the proof of dominated convergence depends on the integral properties obtained in Theorem 5.14. In particular, it relies on linearity.

Proof. First, assume that $f_j \geq 0$ for each index j. By monotonicity of the integral, the functions f_j and the limit function f are integrable because the dominating function $|g|$ is integrable. Recall that integrable functions only take finite values.

By Fatou's lemma (Theorem 5.20),

$$\liminf_{j\to\infty} \mu(f_j) \geq \mu(\liminf_{j\to\infty} f_j) = \mu(f).$$

For each index j, we have $|g| - f_j \geq 0$. Another application of Fatou's lemma yields

$$\liminf_{j\to\infty} \mu(|g| - f_j) \geq \mu(|g| - \limsup_{j\to\infty} f_j) = \mu(|g| - f).$$

Using the linearity of the integral, we can simplify this inequality to read

$$\limsup_{j\to\infty} \mu(f_j) \le \mu(f).$$

Combining these bounds, we find that

$$\liminf_{j\to\infty} \mu(f_j) \ge \mu(f) \ge \limsup_{j\to\infty} \mu(f_j).$$

The inferior and superior limits coincide, so the limit exists. We deduce that the integrals satisfy $\mu(f_j) \to \mu(f)$.

Now, suppose that the functions f_j are merely integrable. Both the sequence of positive parts $((f_j)_+ : j \in \mathbb{N})$ and the sequence of negative parts $((f_j)_- : j \in \mathbb{N})$ are dominated by $|g|$. Using Definition 5.13 and applying the result for positive functions twice, we reach the conclusion that $\mu(f_j) \to \mu(f)$. ∎

Exercise 5.23 (Lebesgue integral: Continuity fails). Find a sequence $(f_j : \mathbb{R} \to \mathbb{R})_{j\in\mathbb{N}}$ where $f_j \to f$ pointwise but $\lambda(f_j) \not\to \lambda(f)$. **Hint:** The mass can "leak out" at $\pm\infty$.

5.4.4 Convergence pointwise and convergence almost everywhere

As a general principle, Lebesgue integrals have no interest in what a function does on a negligible set. We can apply this intuition to extend the convergence theorems to the case where the sequences converge almost everywhere.

First, we take a moment to elaborate on the difference between pointwise convergence and almost-everywhere convergence. Consider a measure space $(\mathsf{X}, \mathcal{F}, \mu)$. Let $(f_j : \mathsf{X} \to \mathbb{R})_{j\in\mathbb{N}}$ be a sequence of measurable functions, and let $f : \mathsf{X} \to \mathbb{R}$ be another measurable function. We compare two convergence concepts:

Pointwise convergence:	$f_j(x) \to f(x)$	for every $x \in \mathsf{X}$.
Almost-everywhere convergence:	$f_j(x) \to f(x)$	for μ-almost every $x \in \mathsf{X}$.

More precisely, almost-everywhere convergence means that

$$\mu\{x \in \mathsf{X} : f_j(x) \nrightarrow f(x)\} = 0.$$

Warning: The practical import of almost-everywhere convergence depends on the measure. ∎

That is, the set of points where the sequence fails to converge is a negligible set for μ. It is clear that pointwise convergence implies almost-everywhere convergence, but the converse is not true.

It may helpful to examine some simple examples. Let λ be the Lebesgue measure on the real line. Consider the functions

$$f_j(x) = \begin{cases} x^j, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases} \quad \text{for } j \in \mathbb{N}.$$

The function $f = \mathbb{1}_{\{1\}}$ is a pointwise limit and a λ-almost-everywhere limit. The zero function $f = 0$ is another λ-almost-everywhere limit but not a pointwise limit. Of course, the two λ-almost everywhere limits agree λ-almost everywhere.

It is not hard to construct sequences that converge almost everywhere but fail to converge pointwise. A very simple example is

$$f_j(x) = \begin{cases} (-1)^j x^j, & 0 \le x \le 1 \\ 0, & \text{otherwise} \end{cases} \quad \text{for } j \in \mathbb{N}.$$

Indeed, since $f_j(1)$ oscillates between ± 1, the sequence fails to converge at the point $x = 1$. But it still converges λ-almost everywhere to the zero function.

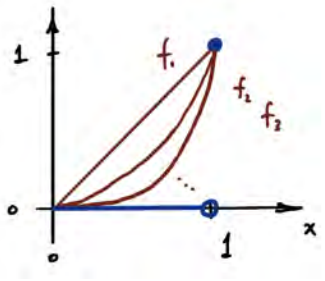


Figure 5.2 (Pw versus ae). On the interval $[0, 1]$, the functions $f_j(x) = x^j$ converge pointwise to a nonzero function, but they converge λ-almost everywhere to zero.

Exercise 5.24 (Almost everywhere for δ_0). Consider the measurable space $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Let δ_0 be the Dirac measure at zero. Let $f_j : \mathbb{R} \to \mathbb{R}$ be measurable functions. Find a simple characterization of what if means for $(f_j : j \in \mathbb{N})$ to converge almost everywhere for δ_0.

Exercise 5.25 (Monotone convergence, almost everywhere). Prove that Theorem 5.18 remains valid under the weaker condition that the functions are increasing almost everywhere. That is, we assume that $\{x \in \mathsf{X} : f_j(x) \text{ is increasing}\}$ is a μ-almost-everywhere set.

Exercise 5.26 (Dominated convergence, almost everywhere). Prove that Theorem 5.22 remains valid under the weaker condition that the functions are convergent almost everywhere. That is, we assume that $\{x \in \mathsf{X} : \lim_{j\to\infty} f_j(x) \text{ exists}\}$ is a μ-almost-everywhere set.

5.5 *Properties of the integral: Proofs

In this section, we give a complete proof of Theorem 5.14. We rely on monotone convergence (Theorem 5.18), whose proof is already finished. Throughout this section, $(\mathsf{X}, \mathscr{F}, \mu)$ is a fixed measure space.

5.5.1 Indicators

We begin with the easy fact that the μ-integral of the indicator function of a set coincides with the measure of the set. Let $\mathsf{A} \in \mathscr{F}$ be a measurable set with indicator $\mathbb{1}_{\mathsf{A}}$. The super-level sets of this function satisfy

$$\{\mathbb{1}_{\mathsf{A}} > t\} = \begin{cases} \mathsf{A}, & 0 \le t < 1; \\ \emptyset, & 1 \le t. \end{cases}$$

By Definition 5.11 of the integral of a positive function,

$$\mu(\mathbb{1}_{\mathsf{A}}) = \int_0^\infty \mu\{\mathbb{1}_{\mathsf{A}} > t\}\, \mathrm{d}t = \int_0^1 \mu(\mathsf{A})\, \mathrm{d}t = \mu(\mathsf{A}).$$

In particular, taking $\mathsf{A} = \emptyset$, we confirm that the integral of the zero function is zero.

5.5.2 Monotonicity properties

The fact that a positive function has a positive integral is obvious. Just glance at Definition 5.11 (the Lebesgue integral of a positive function), and recall that the improper Riemann integral of a positive, decreasing function is a positive number (Theorem C.7).

Next, we develop monotonicity properties for positive, measurable functions $f, g : \mathsf{X} \to \overline{\mathbb{R}}_+$. Suppose that $0 \le f \le g$. For each $t \ge 0$,

$$\mu\{f > t\} \le \mu\{g > t\}.$$

Indeed, the super-level sets satisfy the containment $\{f > t\} \subseteq \{g > t\}$, and measures are monotone (Proposition 2.29). The result follows from Definition 5.11 and the monotonicity of the improper Riemann integral (Theorem C.7).

Last, we consider integrable functions $f, g : \mathsf{X} \to \mathbb{R}$. Suppose that $f \le g$. The positive and negative parts satisfy

$$f_+ \le g_+ \quad \text{and} \quad f_- \ge g_-.$$

Use Definition 5.13 (the Lebesgue integral of a signed function) and the monotonicity of the Lebesgue integral for positive functions (obtained in the previous paragraph).

5.5.3 Linearity properties

The proof that the integral is linear is surprisingly involved. We begin with the case where the functions are positive; we remove the restriction afterward. It is straightforward to check that the integral is (positively) homogeneous; the major effort arises in the proof of additivity.

Positive homogeneity

First, we check that the integral is positively homogeneous. Fix a positive, measurable function $f : \mathsf{X} \to \overline{\mathbb{R}}_+$ and a positive scalar $\alpha \in \mathbb{R}_+$. Calculate

$$\mu(\alpha f) = \int_0^\infty \mu\{\alpha f > t\} \, \mathrm{d}t = \alpha \int_0^\infty \mu\{f > t\} \, \mathrm{d}t = \alpha \mu(f). \tag{5.3}$$

We have used the linear change of variables $t \mapsto \alpha t$ in the Riemann integral. (This fact can be established by direct examination of the lower and upper sums. Otherwise, take limits in Corollary C.6.)

Additivity for positive simple functions

Next, we establish that the integral is additive for positive, measurable functions that take a finite number of values. This is the class $\mathsf{SF}_+ := \mathsf{SF}_+(\mathcal{F})$ we encountered in (5.1). This result takes several steps, which we parcel into propositions.

Proposition 5.27 (Positive simple function: Standard form). A positive simple function $f \in \mathsf{SF}_+$ can always be written in *standard form*:

$$f = \sum\nolimits_{i=0}^{n} \alpha_i \mathbb{1}_{\mathsf{A}_i} \quad \text{where } \alpha_i \geq 0 \text{ and the } \mathsf{A}_i \in \mathcal{F} \text{ are disjoint sets.}$$

We can also upgrade the representation to *canonical form* where the coefficients are distinct and the sets cover the domain: $0 = \alpha_0 < \alpha_1 < \cdots < \alpha_n$ and $\dot{\bigcup}_{i=0}^{n} \mathsf{A}_i = \mathsf{X}$.

Proof. By definition, the range of a positive simple function f contains a finite number of distinct values. We may construct the canonical representation as follows.

$$f = \sum\nolimits_{\alpha \in \operatorname{range} f} \alpha \, \mathbb{1}_{\{f = \alpha\}}.$$

Since f is measurable, the level set $\{f = \alpha\}$ is measurable for each α. For distinct values of α, the level sets $\{f = \alpha\}$ are disjoint. Finally, the level sets cover the whole domain. ∎

Proposition 5.28 (Lebesgue integral: Positive simple function). Let $f \in \mathsf{SF}_+$ be a positive simple function, presented in *standard* form (Proposition 5.27). Then

$$f = \sum\nolimits_{i=0}^{n} \alpha_i \mathbb{1}_{\mathsf{A}_i} \quad \text{implies} \quad \mu(f) = \sum\nolimits_{i=0}^{n} \alpha_i \mu(\mathsf{A}_i).$$

In particular, the value of the integral does not depend on the choice of the standard form representation of the simple function f.

Proof. Let $f = \sum_{i=0}^{n} \alpha_i \mathbb{1}_{\mathsf{A}_i}$, where the A_i are disjoint. Suppose that f takes on the *distinct* values $0 \leq t_1 < \cdots < t_r$. Write $t_0 = 0$.

Introduce the decreasing mass rearrangement $h(t) := \mu\{f > t\}$ for $t \geq 0$. From its definition, we see that the function h is constant on each interval $[t_{j-1}, t_j)$ for $j = 1, \ldots, r$, and h is zero on the interval $[t_r, +\infty)$. Furthermore, since the sets A_i are disjoint, f exceeds t on the set A_i if and only if $\alpha_i > t$. Thus,

$$h(t) = \mu\big(\dot{\bigcup}_{i:\alpha_i > t} \mathsf{A}_i\big) = \sum\nolimits_{i:\alpha_i > t} \mu(\mathsf{A}_i).$$

We have used (finite) additivity of the measure.

With this information, we can now compute the integral of f using our bare hands:

$$\begin{aligned}\mu(f) = \int_0^\infty h(t)\,\mathrm{d}t &= \sum\nolimits_{j=1}^{r} (t_j - t_{j-1}) \cdot h(t_{j-1}) \\ &= \sum\nolimits_{j=1}^{r} (t_j - t_{j-1}) \sum\nolimits_{i:\alpha_i > t_{j-1}} \mu(\mathsf{A}_i) \\ &= \sum\nolimits_{j=1}^{r} t_j \sum\nolimits_{i:\alpha_i = t_j} \mu(\mathsf{A}_i) = \sum\nolimits_{i=0}^{n} \alpha_i \mu(\mathsf{A}_i).\end{aligned}$$

The second relation holds because h is constant on each interval $[t_{j-1}, t_j)$. We have used summation by parts to pass from the second line to the third. ∎

Proposition 5.29 (Lebesgue integral: Additivity for positive simple functions). Let $f, g \in \mathsf{SF}_+$ be positive simple functions. Then $\mu(f + g) = \mu(f) + \mu(g)$.

Proof. Let f and g be presented in *canonical* form:

$$f = \sum\nolimits_{i=0}^{m} \alpha_i \mathbb{1}_{\mathsf{A}_i} \quad \text{and} \quad g = \sum\nolimits_{j=0}^{n} \beta_j \mathbb{1}_{\mathsf{B}_j}.$$

As usual, the coefficients $\alpha_i, \beta_j \geq 0$. The argument hinges on the disjoint cover property:

$$\dot{\bigcup}_{i=0}^{m} \mathsf{A}_i = \mathsf{X} \quad \text{and} \quad \dot{\bigcup}_{j=0}^{n} \mathsf{B}_j = \mathsf{X}.$$

This representation makes it easy to write the sum $f + g$ in standard form:

$$f + g = \sum\nolimits_{i=0}^{m} \sum\nolimits_{j=0}^{n} (\alpha_i + \beta_j)\, \mathbb{1}_{\mathsf{A}_i \cap \mathsf{B}_j}.$$

Indeed, the family $\{\mathsf{A}_i \cap \mathsf{B}_j \text{ for all } i, j\}$ is a disjoint cover of X. Therefore, every point $x \in \mathsf{X}$ belongs to exactly one of these sets, and $(f + g)(x) = \alpha_i + \beta_j$ when $x \in \mathsf{A}_i \cap \mathsf{B}_j$.

Proposition 5.28 now delivers the integral of the sum:

$$\begin{aligned}\mu(f + g) &= \sum\nolimits_{i=0}^{m} \sum\nolimits_{j=0}^{n} (\alpha_i + \beta_j)\, \mu(\mathsf{A}_i \cap \mathsf{B}_j) \\ &= \sum\nolimits_{i=0}^{m} \alpha_i \left[\sum\nolimits_{j=0}^{n} \mu(\mathsf{A}_i \cap \mathsf{B}_j) \right] + \sum\nolimits_{j=0}^{n} \beta_j \left[\sum\nolimits_{i=0}^{m} \mu(\mathsf{A}_i \cap \mathsf{B}_j) \right] \\ &= \sum\nolimits_{i=0}^{m} \alpha_i \mu(\mathsf{A}_i) + \sum\nolimits_{j=0}^{n} \beta_j \mu(\mathsf{B}_j) = \mu(f) + \mu(g).\end{aligned}$$

To compute the first large bracket, we used the fact that $\{\mathsf{A}_i \cap \mathsf{B}_j \text{ for all } j\}$ is a disjoint family with union A_i, and we used the additivity of the measure. The second bracket is computed in an analogous fashion. Finally, we applied Proposition 5.28 again to recognize the integrals $\mu(f)$ and $\mu(g)$. ∎

Additivity for positive functions

To derive additivity for positive functions, we use approximation by simple functions. This argument relies on monotone convergence (Theorem 5.18)!

Proposition 5.30 (Lebesgue integral: Additivity for positive functions). Let $f, g : \mathsf{X} \to \overline{\mathbb{R}}_+$ be positive, measurable functions. Then $\mu(f + g) = \mu(f) + \mu(g)$.

Proof. In (5.2), we introduced the sequence $(Q_j : j \in \mathbb{N})$ of staircase maps. According to Exercise 5.10, these maps have two key properties. First, $Q_j \circ f$ is a positive simple function for each $j \in \mathbb{N}$. Second, the sequence $Q_j \circ f \uparrow f$ pointwise. An evident consequence is that $(Q_j \circ f) + (Q_j \circ g) \uparrow f + g$ pointwise.

By Proposition 5.29,

$$\mu((Q_j \circ f) + (Q_j \circ g)) = \mu(Q_j \circ f) + \mu(Q_j \circ g).$$

To conclude, apply monotone convergence (Theorem 5.18) three times. ∎

Positive linearity

At this point, we stop to collect our results in an important theorem. This result is significant because it holds for all positive functions, regardless of whether they have finite integrals.

Theorem 5.31 (Lebegue integral: Positive linearity). Let $f, g : X \to \overline{\mathbb{R}}_+$ be *positive*, measurable functions, and let $\alpha, \beta \geq 0$ be *positive* scalars. Then

$$\mu(\alpha f + \beta g) = \alpha\mu(f) + \beta\mu(g).$$

Proof. Combine the additivity of the integral for positive functions, Proposition 5.30, with the positive homogeneity property (5.3). ∎

Linearity for integrable functions

Finally, we arrive at the last step in our proof that the Lebesgue integral is linear. Let us emphasize again that linearity is the most important property of an integral.

Theorem 5.32 (Lebesgue integral: Linearity). Let $f, g : X \to \mathbb{R}$ be *μ-integrable* functions, and let $\alpha, \beta \in \mathbb{R}$ be scalars. Then

$$\mu(\alpha f + \beta g) = \alpha\mu(f) + \beta\mu(g).$$

Proof. First, we check homogeneity. Note that scaling preserves the integrability of the function f. Indeed, by the positive homogeneity property (5.3),

$$\mu(|\alpha f|) = \mu(|\alpha| \cdot |f|) = |\alpha| \cdot \mu(|f|) < +\infty \quad \text{for } \alpha \in \mathbb{R}.$$

Restrict attention to the case $\alpha \geq 0$. Using the Definition 5.13 of the Lebesgue integral and (5.3), we see that

$$\mu(\alpha f) = \mu(\alpha f_+ - \alpha f_-) = \mu(\alpha f_+) - \mu(\alpha f_-) = \alpha\mu(f_+) - \alpha\mu(f_-) = \alpha\mu(f).$$

The case $\alpha < 0$ is similar.

Next, we prove that the integral is additive. To that end, note that a sum of integral functions remains integrable:

$$\mu(|f + g|) \leq \mu(|f| + |g|) \leq \mu(|f|) + \mu(|g|) < +\infty.$$

The first relation follows from the triangle inequality for $|\cdot|$ and the monotonicity of the integral (Section 5.5.2). By Definition 5.13 and Theorem 5.31 on positive linearity,

$$\begin{aligned} \mu(f + g) &= \mu((f_+ + g_+) - (f_- + g_-)) \\ &= \mu(f_+ + g_+) - \mu(f_- + g_-) \\ &= [\mu(f_+) + \mu(g_+)] - [\mu(f_-) + \mu(g_-)] = \mu(f) + \mu(g). \end{aligned}$$

We have used the fact that the positive and negative parts of an integrable function are integrable, as are sums of integrable functions.

Together, additivity and the homogeneity readily imply linearity. ∎

5.5.4 Negligible sets

In this section, we establish that negligible sets do not play a role in determining the value of the integral.

First, assume that $f : \mathsf{X} \to \overline{\mathbb{R}}_+$ is a positive function. We will verify that $\mu(f) = 0$ if and only if $f = 0$ μ-almost everywhere. For the reverse direction, assume that $f = 0$ μ-almost everywhere. Equivalently, $\mathsf{E} := \{f > 0\}$ has measure $\mu(\mathsf{E}) = 0$. By monotonicity, the set $\{f > t\} \subseteq \mathsf{E}$ has measure zero for each $t > 0$. By Definition 5.11 of the integral,

$$\mu(f) = \int_0^\infty \mu\{f > t\} \, \mathrm{d}t = 0.$$

For the forward direction, suppose that $\mu\{f > 0\} > 0$. It follows that $\mu\{f > \varepsilon\} > \varepsilon$ for some $\varepsilon > 0$. (Why?) Therefore,

$$\mu(f) = \int_0^\infty \mu\{f > t\} \, \mathrm{d}t \geq \int_0^\varepsilon \mu\{f > t\} \, \mathrm{d}t \geq \int_0^\varepsilon \mu\{f > \varepsilon\} \, \mathrm{d}t \geq \varepsilon^2.$$

In particular, the integral $\mu(f) \neq 0$.

Next, suppose that $f, g : \mathsf{X} \to \overline{\mathbb{R}}_+$ are positive functions that are equal μ-almost everywhere. The reader may verify that the pointwise minimum $f \wedge g$ coincides with both f and g μ-almost everywhere. In particular, $f - (f \wedge g) = 0$ μ-almost everywhere. By the first part,

$$\mu(f) = \mu\big(f - (f \wedge g) + (f \wedge g)\big) = \mu\big(f - (f \wedge g)\big) + \mu(f \wedge g) = \mu(f \wedge g).$$

We have used the fact (Theorem 5.31) that the integral is additive for positive functions. By the same argument, $\mu(g) = \mu(f \wedge g)$. We conclude that $\mu(f) = \mu(g)$.

Finally, suppose that $f, g : \mathbb{R} \to \mathbb{R}$ are μ-integrable functions that are equal μ-almost everywhere. The reader may verify that $f_+ = g_+$ and $f_- = g_-$ almost everywhere for μ. Thus,

$$\mu(f) = \mu(f_+) - \mu(f_-) = \mu(g_+) - \mu(g_-) = \mu(g).$$

We have used Definition 5.13 of the Lebesgue integral and the result from the last paragraph on positive functions that are equal almost everywhere.

Exercise 5.33 (Negligible sets: Without additivity). It is possible to prove these results without using the fact that the integral is additive for positive functions. Give it a try.

5.6 *The Lebesgue integral via simple functions

The most important property of an integral is linearity. Although our construction of the Lebesgue integral via super-level sets is geometrically intuitive, the linearity property does not follow easily (see Section 5.5.3). Moreover, our treatment of the Lebesgue integral relies heavily on properties of the Riemann integral.

In this section, we explore another approach to defining the Lebesgue integral that makes the linearity property more-or-less obvious. We will confirm that the two definitions are equivalent. With extra work (not included here), the approach in this section can also be used to develop the Lebesgue integral without any reference to the Riemann integral.

The approach in this section is the standard way to introduce Lebesgue integrals. Mathematically inclined readers should become familiar with these ideas.

5.6.1 From indicators to positive simple functions

Let $(\mathsf{X}, \mathscr{F}, \mu)$ be a measure space. For now, let us *forget* that we have already defined the Lebesgue integral and start from scratch. If all is right in the world, the μ-integral of the indicator $\mathbb{1}_{\mathsf{A}}$ of a measurable set A should equal the measure of the set. Thus, we begin by *defining* the integral for the class of indicator functions:

$$\int_{\mathsf{X}} \mathbb{1}_{\mathsf{A}} \, \mathrm{d}\mu := \mu(\mathsf{A}) \quad \text{for all } \mathsf{A} \in \mathscr{F}.$$

Note that we allow the integral to take the value $+\infty$ when the set A has infinite measure.

Next, we wish to extend the integral to the class SF_+ of positive simple functions (5.1). We do so by forcing the integral to be linear:

$$\int_{\mathsf{X}} \left(\sum_{i=1}^{n} \alpha_i \mathbb{1}_{\mathsf{A}_i} \right) \mathrm{d}\mu := \sum_{i=1}^{n} \alpha_i \int_{\mathsf{X}} \mathbb{1}_{\mathsf{A}_i} \, \mathrm{d}\mu = \sum_{i=1}^{n} \alpha_i \mu(\mathsf{A}_i). \tag{5.4}$$

Since we have insisted that the coefficients α_i are positive, the integral may take the value $+\infty$, but the definition cannot produce any competing infinities ($\infty - \infty$). The major difficulty is to confirm that (5.4) is a legal definition.

Problem 5.34 (*Lebesgue integral: Well-definition for simple functions). Prove that (5.4) gives a well-defined result. That is, the value of the integral does not depend on how we write the simple function as a positive linear combination of indicators. **Hint:** We may require the α_i to be distinct and the A_i to be disjoint sets that cover $\mathbb{R}$; see Proposition 5.28.

5.6.2 The integral of a positive function

Our next goal is to extend the integral to all positive, measurable functions. To do so, we approximate measurable functions by simple functions that are pointwise smaller; see Figure 5.1. This illustration suggests how to construct the integral. For a positive measurable function $f : \mathsf{X} \to \overline{\mathbb{R}}_+$, define

$$\int_{\mathsf{X}} f \, \mathrm{d}\mu := \sup \left\{ \int_{\mathsf{X}} s \, \mathrm{d}\mu : s \in \mathsf{SF}_+ \text{ and } s \le f \text{ pointwise} \right\}. \tag{5.5}$$

This formula obviously produces a well-defined result in the range $[0, +\infty]$. To work with this definition, you may recall that Exercise 5.10 provides a concrete mechanism for approximating a positive, measurable function below by a positive simple function.

The integral defined in (5.5) inherits positive linearity and other properties from the definition of the integral for simple functions. A direct proof of this claim requires some effort. In our setting, an alternative approach to establishing integral properties is to verify that the new definition of the integral agrees with the old definition.

Proposition 5.35 (Lebesgue integral: Equivalence of definitions). For all positive, measurable functions, the definition (5.5) of the Lebesgue integral via simple functions agrees with Definition 5.11 via super-level sets.

Proof. Let $(\mathsf{X}, \mathscr{F}, \mu)$ be a measure space, and let $f : \mathsf{X} \to \overline{\mathbb{R}}_+$ be a positive, measurable function. We need notation to distinguish the two integrals from each other. To that end, let $\mu(f)$ denote the integral obtained from Definition 5.11, for which we have established monotone convergence (Theorem 5.18) plus monotonicity and linearity (Theorem 5.14). In contrast, let $\mu_0(f)$ denote the integral (5.5) obtained by approximation with simple functions.

The goal is to prove that $\mu(f) = \mu_0(f)$. We already know that $\mu(s) = \mu_0(s)$ for every positive simple function $s \in \mathsf{SF}_+$. Indeed, the value of $\mu_0(s)$ in definition (5.4) agrees with the value of $\mu(s)$ computed in Proposition 5.28.

Recall the definition (5.2) of the staircase map Q_j. According to Exercise 5.10, the positive simple functions $Q_j \circ f$ increase pointwise to f. Thus,

$$\mu_0(f) \geq \sup_j \mu_0(Q_j \circ f) = \sup_j \mu(Q_j \circ f) = \mu(f).$$

The last relation is monotone convergence for μ.

On the other hand, let $(s_j \in \mathsf{SF}_+ : j \in \mathbb{N})$ be a maximizing sequence for the supremum in (5.5). In particular, $s_j \leq f$ for all $j \in \mathbb{N}$. Then

In this context, a *maximizing sequence* has the property that $\mu_0(s_j) \uparrow \mu(f)$.

$$\begin{aligned}\mu_0(f) = \sup_j \mu_0(s_j) = \sup_j \mu(s_j) \\ \leq \sup_j \mu(\sup_{k \leq j} s_k) = \mu(\sup_j \sup_{k \leq j} s_k) \leq \mu(f).\end{aligned}$$

The two inequalities rely on the fact that μ is monotone; the last equality holds by monotone convergence for μ. ∎

5.6.3 The integral of a measurable function

Finally, to integrate a measurable function $f : \mathsf{X} \to \mathbb{R}$ that may take positive and negative values, we split it into positive and negative parts, as before. Assume that $\int_{\mathsf{X}} |f| \, \mathrm{d}\mu < +\infty$. Then we set

$$\int_{\mathsf{X}} f \, \mathrm{d}\mu := \int_{\mathsf{X}} f_+ \, \mathrm{d}\mu - \int f_- \, \mathrm{d}\mu. \tag{5.6}$$

This formula produces a well-defined, finite value. It is not hard to see that the integral (5.6) inherits linearity from the integral (5.5) for positive functions.

Problems

Problem 5.36 (Measurability). Let $(\mathsf{X}, \mathscr{F})$ and $(\mathsf{Y}, \mathscr{G})$ be measurable spaces. We say that a function $f : \mathsf{X} \to \mathsf{Y}$ is *measurable* if

$$f^{-1}(\mathsf{G}) \in \mathscr{F} \quad \text{for all } \mathsf{G} \in \mathscr{G}. \tag{5.7}$$

1. Consider the special case where $(\mathsf{Y}, \mathscr{G}) = (\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Confirm that the definition (5.7) is consistent with Definition 5.1 of a measurable function.
2. Suppose that $\mathscr{G} = \sigma(\mathcal{S}; \mathsf{Y})$ is the σ-algebra generated by a family $\mathcal{S} \subseteq \mathscr{P}(\mathsf{Y})$. Prove that a function $f : \mathsf{X} \to \mathsf{Y}$ is measurable if and only if

   $$f^{-1}(\mathsf{S}) \in \mathscr{F} \quad \text{for each } \mathsf{S} \in \mathcal{S}.$$

   In other words, we only need to check the sets that generate the σ-algebra for the codomain of f. **Hint:** See the proof of Proposition 4.2.
3. Deduce Proposition 4.2 directly from (2).
4. Let $f : \mathsf{X} \to \mathsf{Y}$ and $g : \mathsf{Y} \to \mathsf{Z}$ be measurable functions defined on suitable measurable spaces. Show that the composition $g \circ f$ is measurable.

> **Aside:** In categorical terms, measurable functions are the morphisms between measurable spaces. For an analogy, recall that continuous functions are the morphisms between topological spaces.

Exercise 5.37 (Yet more inclusion–exclusion). Let $(\mathsf{X}, \mathscr{F}, \mu)$ be a measure space. Let $\mathsf{A}_1, \dots, \mathsf{A}_n$ be measurable sets. In Exercise 2.47, we derived the inclusion–exclusion identity

$$\mathbb{1}_{\bigcup_{i=1}^n \mathsf{A}_i} = \sum_{k=1}^n (-1)^{k+1} \sum_{i_1 < i_2 < \cdots < i_k} \mathbb{1}_{\mathsf{A}_{i_1} \cap \cdots \cap \mathsf{A}_{i_k}}.$$

Deduce that

$$\mu\big(\textstyle\bigcup_{i=1}^n \mathsf{A}_i\big) = \sum_{k=1}^n (-1)^{k+1} \sum_{i_1 < i_2 < \cdots < i_k} \mu(\mathsf{A}_{i_1} \cap \cdots \cap \mathsf{A}_{i_k}).$$

Hint: The integral is a linear functional.

Exercise 5.38 (Lebesgue integral: Downward monotone convergence). Let $(\mathsf{X}, \mathscr{F}, \mu)$ be a measure space, and assume that μ is a *finite* measure. Consider a *decreasing* sequence $(f_j : \mathsf{X} \to \overline{\mathbb{R}}_+)$ of *positive*, measurable functions: $f_j \downarrow f$. Prove that $\mu(f_j) \downarrow \mu(f)$.

Exercise 5.39 (Lebegue integral: Tonelli's theorem for sums). Let $(\mathsf{X}, \mathscr{F}, \mu)$ be a measure space. Let $(f_j : \mathsf{X} \to \overline{\mathbb{R}}_+)_{j \in \mathbb{N}}$ be a sequence of *positive*, measurable functions. Show that, without further qualification,

$$\int_{\mathsf{X}} \Big(\sum_{j=1}^\infty f_j \Big) \, \mathrm{d}\mu = \sum_{j=1}^\infty \int_{\mathsf{X}} f_j \, \mathrm{d}\mu.$$

Hint: Use monotone convergence!

Exercise 5.40 (Lebesgue integral: Domain decomposition). Let $(\mathsf{X}, \mathscr{F}, \mu)$ be a measure space. Suppose that $\mathsf{X} = \dot{\bigcup}_{j=1}^\infty \mathsf{E}_j$ for measurable sets E_j. If the measurable function $f : \mathsf{X} \to \mathbb{R}$ is either positive or integrable, then

$$\int_{\mathsf{X}} f \, \mathrm{d}\mu = \sum_{j=1}^\infty \int_{\mathsf{E}_j} f \, \mathrm{d}\mu.$$

Problem 5.41 (Borel–Cantelli I). Let $(\mathsf{X}, \mathscr{F}, \mu)$ be a measure space. For a sequence $(f_j : \mathsf{X} \to \overline{\mathbb{R}}_+)_{j \in \mathbb{N}}$ of positive, measurable functions, prove that

$$\sum_{j=1}^\infty \mu(f_j) < +\infty \quad \text{implies} \quad \mu\big(\limsup_{j \to \infty} f_j \big) = 0.$$

Specialize this result to indicator functions, and then write the statement in terms of sets. What is the interpretation? **Hint:** Write the limit superior as an inf–sup. Note that the integral of the supremum tends to zero.

Problem 5.42 (Lebesgue integral: Differentiation under the integral). We often encounter integrals that are parameterized by a real variable. Under modest conditions, we can differentiate the integral with respect to the parameter by passing the derivative through the integral.

Let μ be a Borel measure on $\mathbb{R}$, and let $f : \mathbb{R} \times (a, b) \to \mathbb{R}$ be a bivariate function, where $a, b \in \mathbb{R}$. Fix a point $y_0 \in (a, b)$. Assume that $f(\cdot, y)$ is μ-integrable for each $y \in (a, b)$. Define

$$F(y) := \int_{\mathbb{R}} f(x, y) \, \mu(\mathrm{d}x) \quad \text{for } y \in (a, b).$$

1. Assume that $\lim_{y\to y_0} f(x,y) = f(x,y_0)$ for every $x \in \mathbb{R}$. Suppose that $|f(x,y)| \le |g(x)|$ for all $y \in (a,b)$, where g is μ-integrable. Use dominated convergence to conclude that
$$\lim_{y\to y_0} F(y) = F(y_0).$$
In particular, if $f(x,\cdot)$ is continuous for each x, then F is continuous at y_0. **Hint:** You can interpret the limit as the limit of a sequence.
2. Suppose the partial derivative $\partial_y f : \mathbb{R}\times(a,b) \to \mathbb{R}$ exists. Assume that $|(\partial_y f)(x,y)| \le |g(x)|$ for each $y \in (a,b)$, where g is a fixed μ-integrable function. Deduce that F is differentiable at y_0, and
$$F'(y_0) = \int_{\mathbb{R}} (\partial_y f)(x,y_0)\,\mu(\mathrm{d}x).$$
Hint: Use the mean-value theorem to argue that g also dominates the difference quotients of $f(x,\cdot)$.
3. For $a > 0$, consider the parameterized integral
$$F(a) := \int_{\mathbb{R}_+} \mathrm{e}^{-ax}\,\lambda(\mathrm{d}x) = \frac{1}{a}.$$
Compute two alternative expressions for the nth derivative $F^{(n)}(a)$ by differentiating this relation repeatedly with respect to a. Specialize this formula to $a = 1$, and discuss the result.
4. **(*)** Here is another example that can be treated by the same approach:
$$F(a) := \int_{\mathbb{R}_+} \mathrm{e}^{-ax}\,\frac{\sin(x)}{x}\,\lambda(\mathrm{d}x) = \frac{\pi}{2} - \arctan(a) \quad \text{for } a > 0.$$
Verify the identity by differentiating under the integral sign and using standard tools from calculus.

Problem 5.43 (Densities). Let $(\mathsf{X},\mathcal{F},\mu)$ be a measure space, and let $f : \mathsf{X} \to \overline{\mathbb{R}}_+$ be a positive, measurable function. We can define a function $\nu : \mathcal{F} \to [0,+\infty]$ on measurable sets via
$$\nu(\mathsf{A}) := \int_{\mathsf{A}} f\,\mathrm{d}\mu \quad \text{for all } \mathsf{A} \in \mathcal{F}.$$
The function f is called the *density* of ν with respect to μ. It is often written $f = \mathrm{d}\nu/\mathrm{d}\mu$. See Exercise 6.27 for an explanation of this notation.

1. Prove that ν is a measure with $\nu(\mathsf{X}) = \mu(f)$.
2. From the definition of the integral, show that two positive functions f and g are both densities of μ if and only if $f = g$ λ-almost everywhere. **Hint:** Prove the forward direction by contraposition. For the reverse direction, first establish the case $f = 0$. Then consider the measurable function $f - f \wedge g$.
3. Explain why $\mu(\mathsf{A}) = 0$ implies that $\nu(\mathsf{A}) = 0$ for every measurable set A. We write this condition as $\nu \ll \mu$, and we say that ν is *absolutely continuous* with respect to μ.
4. Consider the measure space $(\mathbb{R},\mathcal{B}(\mathbb{R}),\lambda)$. Define
$$\gamma(\mathsf{B}) := \frac{1}{\sqrt{2\pi}} \int_{\mathsf{B}} \mathrm{e}^{-x^2/2}\,\lambda(\mathrm{d}x) \quad \text{for Borel } \mathsf{B} \in \mathcal{B}(\mathbb{R}).$$
Show that γ is a measure, which is called the *standard Gaussian measure*. What is its density with respect to the Lebesgue measure λ? **(*)** What is the total mass of the measure?

The converse of (3) is called the Radon–Nikodým theorem: $\nu \ll \mu$ implies that ν has a density with respect to μ. See Lecture 22.

5. Consider the measure space $(\mathbb{N}, \mathcal{P}(\mathbb{N}), \#)$. Define

$$\mu(\mathsf{A}) := \frac{1}{\mathrm{e}} \int_{\mathsf{A}} \frac{1}{k!} \, \#(\mathrm{d}k) = \sum\nolimits_{k \in \mathsf{A}} \frac{1}{\mathrm{e}\, k!} \quad \text{for } \mathsf{A} \subseteq \mathbb{N}.$$

   Show that μ is a probability measure on $\mathbb{N}$, called the *standard Poisson measure*. What is its density with respect to the counting measure #?

Problem 5.44 (Push-forward of a measure). Let $(\mathsf{X}, \mathcal{F}, \mu)$ be a measure space, and let $f : \mathsf{X} \to \mathbb{R}$ be a measurable function on the domain X. Define a function $\nu : \mathcal{B}(\mathbb{R}) \to [0, +\infty]$ via

$$\nu(\mathsf{B}) := \mu(f^{-1}(\mathsf{B})) \quad \text{for all } \mathsf{B} \in \mathcal{B}(\mathbb{R}).$$

The measure ν is called the *push-forward* of the measure μ by the function f. It is commonly denoted $\nu = f_* \mu$.

1. Prove that ν is a Borel measure on $\mathbb{R}$. **Hint:** Use the definition of measurability and the properties of the preimage.
2. Establish the change of variables formula:

$$\nu(g) = \mu(g \circ f). \tag{5.8}$$

   Hint: Start with the case where g is positive and use Definition 5.11 of the Lebesgue integral. This problem becomes much harder if you define the integral as a limit of simple functions.

Problem 5.45 (*Measures from positive operators). Let $(\mathsf{X}, \mathcal{F})$ be a measurable space. Introduce the set of positive, measurable functions: $\mathsf{L}_+ := \{f : \mathsf{X} \to \overline{\mathbb{R}}_+ \text{ measurable}\}$. Let $\boldsymbol{T} : \mathsf{L}_+ \to [0, +\infty]$ be an operator that satisfies

1. **Monotonicity:** $f \leq g$ implies that $\boldsymbol{T}(f) \leq \boldsymbol{T}(g)$ for all $f, g \in \mathsf{L}_+$.
2. **Positive linearity:** $\boldsymbol{T}(\alpha f + \alpha g) = \alpha \boldsymbol{T}(f) + \beta \boldsymbol{T}(g)$ for all $f, g \in \mathsf{L}_+$ and $\alpha, \beta \geq 0$.
3. **Monotone convergence:** For every increasing sequence $(f_j : j \in \mathbb{N})$ of functions in L_+ with pointwise limit f, we have $\boldsymbol{T}(f_j) \uparrow \boldsymbol{T}(f)$.

Define a function $\mu : \mathcal{F} \to [0, +\infty]$ on measurable sets via

$$\mu(\mathsf{A}) := \boldsymbol{T}(\mathbb{1}_{\mathsf{A}}) \quad \text{for all } \mathsf{A} \in \mathcal{F}.$$

Prove that μ is a measure. Deduce that measures are in one-to-one correspondence with these positive operators.

Notes

All of this material is standard, and some version of these results may be found in any book on real analysis. Nevertheless, the construction of the integral using super-level sets is an unusual choice; it is motivated by the presentation in Lieb & Loss [LL01]. The problem on differentiation under the integral sign is drawn from Folland's book [Fol99], while the examples of this method are extracted from Conrad's note [Con]. The problem on constructing measures from positive operators is adapted from Pollard's book [Pol02].

Lecture bibliography

[Con] K. Conrad. “Differentiation under the integral sign”. Available online. URL: https://kconrad.math.uconn.edu/blurbs/analysis/diffunderint.pdf.

[Fol99] G. B. Folland. *Real analysis*. Second. Modern techniques and their applications, A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1999.

[LL01] E. H. Lieb and M. Loss. *Analysis*. 2nd ed. American Mathematical Society, 2001. DOI: 10.1090/gsm/014.

[Pol02] D. Pollard. *A user’s guide to measure theoretic probability*. Cambridge University Press, 2002.

6. Product Measures

Agenda:

1. Products of measurable spaces
2. Product measure
3. Fubini–Tonelli
4. Integration by parts

> "Don't find customers for your products, find products for your customers."
>
> —Seth Godin

> "Tout dans la nature se modèle sur la sphère, le cône et le cylindre, il faut apprendre à peindre sur ces figures simples, on pourra ensuite faire tout ce qu'on voudra."
>
> "Everything in nature is modeled on the sphere, the cone, and the cylinder. You must learn to paint these simple figures. You will then be able to paint anything that you want."
>
> —Paul Cézanne, 1904

The theory of measure and integrals was initially developed to clarify the notion of "length". We outlined these ideas in Lecture 3, where we encountered the construction of the Borel sets and the definition of the Lebesgue measure.

In this lecture, we turn to another classic problem: How can we assign a consistent notion of "area" to subsets of the real plane? We know that the area of a rectangle should equal the product of its width and its height, and our goal is to extend this elementary idea to a wider class of sets. This labor requires the machinery of abstract measure theory (Lecture 2) and abstract integration (Lecture 5).

Related questions arise in probability theory. Given two "independent" probabilistic experiments, the probability that the pair of outcomes is in a set $A \times B$ equals the product of the probability that the first outcome belongs to A and the probability that the second outcome belongs to B. The problem is to determine the probability that the pair of outcomes belongs to a set that is not "rectangular". We will spend some energy on the probabilistic interpretation later.

After developing a way to assign areas in the plane, we turn to the problem of integrating a function on the plane. This investigation is tied to the question about when we can interchange two integrals, which is addressed by the fundamental theorem of Fubini and Tonelli.

6.1 Products of measurable space

A geometric rectangle is the product of a horizontal line segment and a vertical line segment. The area of a rectangle is the product of the length of the horizontal segment and the vertical segment. It stands to reason that we must begin with Cartesian products of sets if we want to understand area and its generalizations.

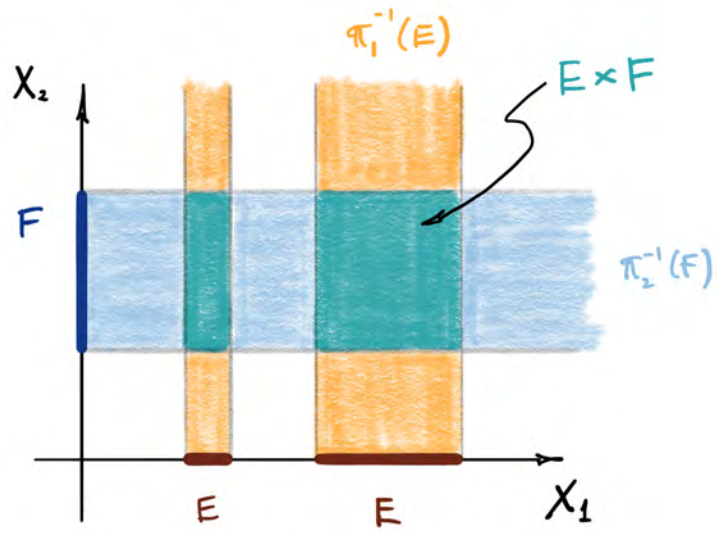


Figure 6.1 (Cylinders and rectangles). The preimage $\pi_1^{-1}(\mathsf{E})$ of a set $\mathsf{E} \subseteq \mathcal{X}_1$ under the coordinate projection π_1 onto X_1 is called a Cartesian cylinder. Similarly, the preimage of a measurable set in X_2 under the coordinate projection π_2 is a Cartesian cylinder. The intersection of two Cartesian cylinders forms a Cartesian rectangle.

6.1.1 Cylinders and rectangles

Consider two domains X_1 and X_2. Recall that the (Cartesian) product of the domains is

$$\mathsf{X} := \mathsf{X}_1 \times \mathsf{X}_2 = \{(x_1, x_2) : x_1 \in \mathsf{X}_1 \text{ and } x_2 \in \mathsf{X}_2\}.$$

For arbitrary subsets $\mathsf{E}, \mathsf{F} \subseteq \mathsf{X}$, we can construct the (Cartesian) *rectangle* $\mathsf{E} \times \mathsf{F} \subseteq \mathsf{X}$.

Associated with the product space are the two coordinate projection maps:

$$\begin{aligned} \pi_1 &: \mathsf{X} \to \mathsf{X}_1 \quad \text{where} \quad \pi_1 : (x_1, x_2) \mapsto x_1; \\ \pi_2 &: \mathsf{X} \to \mathsf{X}_2 \quad \text{where} \quad \pi_2 : (x_1, x_2) \mapsto x_2. \end{aligned}$$

Using the coordinate projection maps, we can lift arbitrary sets in the factor spaces to obtain (Cartesian) *cylinders*.

$$\begin{aligned} \pi_1^{-1}(\mathsf{E}) &= \{(x_1, x_2) \in \mathsf{X} : x_1 \in \mathsf{E}\} \quad \text{for } \mathsf{E} \subseteq \mathsf{X}_1; \\ \pi_2^{-1}(\mathsf{F}) &= \{(x_1, x_2) \in \mathsf{X} : x_2 \in \mathsf{F}\} \quad \text{for } \mathsf{F} \subseteq \mathsf{X}_2. \end{aligned}$$

In the two-dimensional product X, the intersection of a horizontal and vertical cylinder yields a Cartesian rectangle. See Figure 6.1 for an illustration.

6.1.2 Measurable cylinders and product-measurable sets

Consider two measurable spaces $(\mathsf{X}_1, \mathcal{F}_1)$ and $(\mathsf{X}_2, \mathcal{F}_2)$. We would like to construct a natural product of these measurable spaces. To do so, we must decide which sets will be measurable in the product domain $\mathsf{X} = \mathsf{X}_1 \times \mathsf{X}_2$.

Right now, we only have a notion of what sets are measurable in X_1 and in X_2. As we have seen, we can lift the measurable sets in each domain to obtain cylinders:

$$\begin{aligned} &\pi_1^{-1}(\mathsf{E}) \quad \text{for measurable } \mathsf{E} \in \mathcal{F}_1; \\ &\pi_2^{-1}(\mathsf{F}) \quad \text{for measurable } \mathsf{F} \in \mathcal{F}_2. \end{aligned}$$

A cylinder obtained from a measurable set is called a *measurable cylinder*. It is reasonable to insist that every measurable cylinder should be a measurable set in the product space. To achieve this goal, we simply *define* the measurable sets in the product to be the elements of the σ-algebra generated by all measurable cylinders.

Let us formalize this construction.

Definition 6.1 (Product of measurable spaces). Let $(\mathsf{X}_i, \mathscr{F}_i)$ be measurable spaces for $i = 1, 2$. The *product* of the measurable spaces

$$(\mathsf{X}, \mathscr{F}) := (\mathsf{X}_1, \mathscr{F}_1) \times (\mathsf{X}_2, \mathscr{F}_2) := \prod\nolimits_{i=1,2} (\mathsf{X}_i, \mathscr{F}_i)$$

is the measurable space with the domain $\mathsf{X} = \mathsf{X}_1 \times \mathsf{X}_2$. We equip it with the *product σ-algebra* $\mathscr{F}$, which is generated by all measurable cylinders:

$$\mathscr{F} := \mathscr{F}_1 \times_\sigma \mathscr{F}_2 := \sigma\big(\{\pi_1^{-1}(\mathsf{E}) : \mathsf{E} \in \mathscr{F}_1\} \cup \{\pi_2^{-1}(\mathsf{F}) : \mathsf{F} \in \mathscr{F}_2\}; \mathsf{X}\big).$$

We abbreviate $(\mathsf{X}_1, \mathscr{F}_1)^2$ for the product of a measurable space with itself.

By construction, the product σ-algebra is the smallest σ-algebra on the product space $\mathsf{X}_1 \times \mathsf{X}_2$ in which the coordinate projections π_1, π_2 are measurable functions (see Problem 5.36). This is the key reason that we construct the measurable sets using cylinders.

This construction is analogous to the product topology, which is the smallest topology on the product where each coordinate projection is a continuous function.

Exercise 6.2 (Products of measurable spaces). In some instances, we can compute the product of measurable spaces easily. Let $(\mathsf{X}_i, \mathscr{F}_i)$ be a measurable spaces for $i = 1, 2$.

- **Trivial σ-algebras:** Suppose that $\mathscr{F}_i = \{\emptyset, \mathsf{X}_i\}$ for $i = 1, 2$. What is the product σ-algebra?
- **Finite σ-algebras:** Suppose that $\mathscr{F}_i$ has finite cardinality for $i = 1, 2$. Show that the product σ-algebra has finite cardinality.
- **Complete σ-algebras:** Suppose that X_i is countable, and let $\mathscr{F}_i = \mathscr{P}(\mathsf{X}_i)$ for $i = 1, 2$. Show that the product σ-algebra is the complete σ-algebra on $\mathsf{X}_1 \times \mathsf{X}_2$.

Exercise 6.3 (*Products: Associativity). Let $(\mathsf{X}_i, \mathscr{F}_i)$ be measurable spaces for $i = 1, 2, 3$. Check that the product is associative:

$$((\mathsf{X}_1, \mathscr{F}_1) \times (\mathsf{X}_2, \mathscr{F}_2)) \times (\mathsf{X}_3, \mathscr{F}_3) = (\mathsf{X}_1, \mathscr{F}_1) \times ((\mathsf{X}_2, \mathscr{F}_2) \times (\mathsf{X}_3, \mathscr{F}_3)).$$

As a consequence, we can write the product measurable space without parentheses:

$$(\mathsf{X}_1, \mathscr{F}_1) \times (\mathsf{X}_2, \mathscr{F}_2) \times (\mathsf{X}_3, \mathscr{F}_3).$$

Hint: Show that the product σ-algebra is

$$\sigma(\{\pi_i^{-1}(\mathsf{E}_i) : \mathsf{E}_i \in \mathscr{F}_i \text{ and } i = 1, 2, 3\}).$$

By repeating this construction, we can define n-fold products.

6.1.3 Measurable rectangles

Consider a Cartesian rectangle $\mathsf{E} \times \mathsf{F}$ obtained from two measurable sets $\mathsf{E} \in \mathscr{F}_1$ and $\mathsf{F} \in \mathscr{F}_2$. This rectangle is the intersection of two measurable cylinders:

$$\mathsf{E} \times \mathsf{F} = \pi_1^{-1}(\mathsf{E}) \cap \pi_2^{-1}(\mathsf{F}).$$

Since the measurable cylinders generate the σ-algebra of measurable sets, the rectangle $\mathsf{E} \times \mathsf{F}$ must also be measurable. Indeed, σ-algebras are stable under (countable) intersection. See Section 6.1.6 for further discussion.

Exercise 6.4 (Rectangles generate the product-measurable sets in two dimensions). Let $(\mathsf{X}_i, \mathscr{F}_i)$ be measurable spaces for $i = 1, 2$. Show that the rectangles generate the product σ-algebra:

$$\mathscr{F} = \mathscr{F}_1 \times_\sigma \mathscr{F}_2 = \sigma\{\mathsf{E} \times \mathsf{F} : \mathsf{E} \in \mathscr{F}_1 \text{ and } \mathsf{F} \in \mathscr{F}_2\}.$$

Let us note that this example only involves the product of two measurable spaces.

6.1.4 Example: Borel sets in Euclidean spaces

On the real line, the length is associated with the uniform distribution of mass, which places one unit of mass per unit of length on the entire line. Similarly, the concept of area is associated with a uniform distribution of mass over the Euclidean plane ($\mathbb{R}^2$). This distribution should place one unit of mass per unit of area on the entire plane. Before we turn to the construction, we need to introduce an appropriate measurable space. Let us introduce and reconcile two possible approaches.

We reasoned that we should be able to define the length of any open interval in the real line $\mathbb{R}$. This led to the definition of the Borel sets $\mathscr{B}(\mathbb{R})$ as the smallest σ-algebra generated by the open intervals of $\mathbb{R}$. Similarly, we should be able to define the area of any open Euclidean ball in the plane $\mathbb{R}^2$. This idea leads to the definition of the Borel measurable sets in the plane.

Definition 6.5 (Borel sets: Euclidean plane). The class $\mathscr{B}(\mathbb{R}^2)$ of Borel sets in the Euclidean plane $\mathbb{R}^2$ is the smallest σ-algebra generated by open Euclidean balls:

$$\mathscr{B}(\mathbb{R}^2) := \sigma\{\mathsf{D}(\boldsymbol{x}; r) : \boldsymbol{x} \in \mathbb{R}^2 \text{ and } r > 0\}.$$

We have written $\mathsf{D}(\boldsymbol{x}; r) := \{\boldsymbol{y} \in \mathbb{R}^2 : \|\boldsymbol{y} - \boldsymbol{x}\|_2 < r\}$ for the open ball centered at $\boldsymbol{x} \in \mathbb{R}^2$ with radius $r > 0$.

As usual, $\|\cdot\|_2$ denotes the Euclidean norm.

Exercise 6.6 (*Borel sets: Euclidean plane). Show that $\mathscr{B}(\mathbb{R}^2)$ is the σ-algebra generated by open subsets of the Euclidean plane $\mathbb{R}^2$. **Hint:** Every open set in the plane is a countable union of open Euclidean balls. Look up "second-countable space."

Just now, we have introduced the notion of a product of two measurable spaces. In particular, we can consider the product $(\mathbb{R}, \mathscr{B}(\mathbb{R}))^2$ of the real line with itself. The basic idea here is that every (measurable) cylinder in the plane is product-measurable, and this gives rise to another class $\mathscr{B}(\mathbb{R})^2$ of measurable sets. How can we reconcile these two constructions?

Proposition 6.7 (Product of Borel sets: Euclidean plane). Consider the measurable space $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$ given by the real line equipped with its Borel sets. Then

$$(\mathbb{R}, \mathscr{B}(\mathbb{R}))^2 = (\mathbb{R}^2, \mathscr{B}(\mathbb{R}^2)).$$

In other words, if we "square" the real line with its Borel sets, we obtain the real plane equipped with its Borel sets.

**Proof.* First, observe that the coordinate projections $\pi_i : \mathbb{R}^2 \to \mathbb{R}$ are continuous with respect to the Euclidean topology (generated by open balls in $\mathbb{R}^2$). Therefore, Exercise 5.3 implies that π_i is a $\mathscr{B}(\mathbb{R}^2)$-measurable function because $\mathscr{B}(\mathbb{R}^2)$ is also generated by open balls in $\mathbb{R}^2$. In detail,

$$\pi_i^{-1}(\mathsf{B}) \in \mathscr{B}(\mathbb{R}^2) \quad \text{for each } \mathsf{B} \in \mathscr{B}(\mathbb{R}) \text{ and } i = 1, 2.$$

As a consequence,

$$\mathscr{B}(\mathbb{R})^2 = \sigma\{\pi_i^{-1}(\mathsf{B}) : \mathsf{B} \in \mathscr{B}(\mathbb{R}) \text{ and } i = 1, 2\} \subseteq \mathscr{B}(\mathbb{R}^2).$$

For the converse, recall that the Borel sets $\mathscr{B}(\mathbb{R}^2)$ in the plane are generated by open sets (Exercise 6.6). Every open set G in the plane is a countable union of open rectangles of the form $(a, b) \times (c, d)$. Now, each of these open rectangles is the intersection of two cylinders, namely $(a, b) \times \mathbb{R}$ and $\mathbb{R} \times (c, d)$. As a consequence, G is a countable combination of cylinders, so $\mathsf{G} \in \mathscr{B}(\mathbb{R})^2$. We deduce that $\mathscr{B}(\mathbb{R}^2) \subseteq \mathscr{B}(\mathbb{R})^2$. ■

Activity 6.8 (Borel sets: Euclidean plane). For $\mathsf{E}, \mathsf{F} \in \mathscr{B}(\mathbb{R})$, note that the rectangle $\mathsf{E} \times \mathsf{F}$ is a Borel set in the plane. Deduce that each singleton $\{\boldsymbol{x}\} \subseteq \mathbb{R}^2$ is Borel. Check that geometric rectangles, like $(a, b) \times (c, d)$ and $(a, b] \times (c, d]$, are Borel. Note that semi-infinite rectangles $(-\infty, b] \times (-\infty, d]$ are Borel. Explain why every open set and every closed set in the plane is Borel. Check that each half-space $\{\boldsymbol{x} \in \mathbb{R}^2 : \boldsymbol{a}^\intercal \boldsymbol{x} \leq b\}$ is Borel. What about convex sets? Can you think of more examples? ▪

We can extend these notions to higher-dimensional Euclidean spaces.

Definition 6.9 (Borel sets: Euclidean space). For $n \in \mathbb{N}$, the class $\mathscr{B}(\mathbb{R}^n)$ of Borel sets in the Euclidean space $\mathbb{R}^n$ is the smallest σ-algebra generated by open Euclidean balls:

$$\mathscr{B}(\mathbb{R}^n) := \sigma\{\mathsf{D}(\boldsymbol{x}; r) : \boldsymbol{x} \in \mathbb{R}^n \text{ and } r > 0\}.$$

We have written $\mathsf{D}(\boldsymbol{x}; r) := \{\boldsymbol{y} \in \mathbb{R}^n : \|\boldsymbol{y} - \boldsymbol{x}\|_2 < r\}$ for the open ball centered at $\boldsymbol{x} \in \mathbb{R}^n$ with radius $r > 0$.

Exercise 6.10 (*Product of Borel sets: Euclidean space). For $n \in \mathbb{N}$, show that the n-fold product of the real line satisfies $(\mathbb{R}, \mathscr{B}(\mathbb{R}))^n = (\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$.

Warning 6.11 (*Borel sets: Infinite products). The statement analogous to Exercise 6.10 fails for an infinite product. Indeed, the open sets in $\mathscr{B}(\mathbb{R}^{\mathbb{N}})$ may only contain the intersections of a *finite* number of cylinders induced by open sets in the factor spaces. ▪

6.1.5 Measurable functions on product spaces

Suppose that $(\mathsf{X}, \mathscr{F})$ is a product of measurable spaces $(\mathsf{X}_i, \mathscr{F}_i)$ for $i = 1, 2$. We can specialize the Definition 5.1 of a measurable function to this setting. In detail, let $f : \mathsf{X}_1 \times \mathsf{X}_2 \to \overline{\mathbb{R}}$ be a bivariate function. Then f is *(product) measurable* when

$$f^{-1}(t, +\infty] = \{(x, y) \in \mathsf{X}_1 \times \mathsf{X}_2 : f(x, y) > t\} \in \mathscr{F} \quad \text{for each } t \in \mathbb{R}.$$

What kind of functions are product measurable? As usual, the indicator function $\mathbb{1}_{\mathsf{A}}$ of a product-measurable set $\mathsf{A} \in \mathscr{F}$ is measurable. All linear combinations and products of measurable functions are measurable. All countable combinations of measurable functions are measurable. See Section 5.2.1 for general principles.

Exercise 6.12 (Product functions). Suppose that $(\mathsf{X}, \mathscr{F})$ is a product of measurable spaces $(\mathsf{X}_i, \mathscr{F}_i)$ for $i = 1, 2$. Let $f : \mathsf{X}_1 \to \mathbb{R}$ and $g : \mathsf{X}_2 \to \mathbb{R}$ be measurable functions. Show that the function $(x, y) \mapsto f(x)g(y)$ is product measurable.

Exercise 6.13 (*Measurable functions on $\mathbb{R}^n$). Consider the product space $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$. Show that every continuous function $f : \mathbb{R}^n \to \mathbb{R}$ is product measurable. Show that every lower-semicontinuous convex function $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ is product measurable.

In the particular case of $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$, we have the same principle about measurable functions as we did in $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Indeed, most functions $f : \mathbb{R}^n \to \overline{\mathbb{R}}$ that you encounter in practice are product measurable.

6.1.6 *General products of measurable spaces

For an arbitrary index set I, we can construct the product of an indexed family $\{(\mathsf{X}_i, \mathscr{F}_i) : i \in \mathsf{I}\}$ of measurable spaces by adapting Definition 6.1. First, form the Cartesian product $\mathsf{X} = \prod_{i\in\mathsf{I}} \mathsf{X}_i$ of the domains. Equip the product domain X with the σ-algebra

$$\mathscr{F} := \sigma\big(\pi_i^{-1}(\mathsf{E}_i) : \mathsf{E}_i \in \mathscr{F}_i \text{ for some } i \in \mathsf{I}\big).$$

In other words, $\mathscr{F}$ is generated by all measurable cylinders $\pi_i^{-1}(\mathsf{E}_i)$ induced by all measurable sets E_i in the factor spaces X_i.

Now, consider a Cartesian rectangle $\mathsf{R} := \prod_{i\in\mathsf{I}} \mathsf{E}_i$ formed as a product of *measurable* sets $\mathsf{E}_i \in \mathscr{F}_i$. When the index set I is *countable*, every such rectangle R is a measurable set in the product.

In contrast, when I is *uncountable*, this rectangle R need not be measurable because the σ-algebra is only guaranteed to contain *countable* intersections of cylinders! Although this technicality is sometimes inconvenient, the construction we have given (starting with cylinders, not rectangles) is the mathematically natural one. Indeed, it is the smallest family of measurable sets for which the coordinate projections are all measurable functions.

This issue may seem esoteric, but it can arise in the study of continuous stochastic processes. These processes contain an uncountable number of random variables. We cannot simultaneously constrain the value of each random variable and be confident that these outcomes compose a measurable set (called an event in this context). More advanced probability courses address this matter, but we will not discuss it further in this class.

Aside: For a comparison, consider a family of topological spaces equipped with the product topology. In this setting, an *open cylinder* is defined as the preimage under the coordinate projection of an open set in one of the factor spaces. The product topology is the (smallest) topology generated by the open cylinders. Meanwhile, consider a Cartesian rectangle obtained as the product of one open set from each factor space. When we form the product of an *infinite* number of topological spaces, the product topology may not contain all such rectangles because the topology only contains *finite* intersections of open cylinders. Be careful!

6.2 Product measures

Consider two measure spaces $(\mathsf{X}_1, \mathscr{F}_1, \mu_1)$ and $(\mathsf{X}_2, \mathscr{F}_2, \mu_2)$. Let $(\mathsf{X}, \mathscr{F})$ be the product of the associated measurable spaces. Our next job is to equip the product measurable space with a canonical product measure $\mu = \mu_1 \times \mu_2$.

6.2.1 Existence and uniqueness of product measure

For measurable rectangles, the value of the product measure should certainly equal the product of the measures of the sides:

$$\mu(\mathsf{E} \times \mathsf{F}) := (\mu_1 \times \mu_2)(\mathsf{E} \times \mathsf{F}) := \mu_1(\mathsf{E}) \cdot \mu_2(\mathsf{F}) \quad \text{for } \mathsf{E} \in \mathscr{F}_1 \text{ and } \mathsf{F} \in \mathscr{F}_2. \tag{6.1}$$

This definition agrees with our elementary concept of area. Indeed, when $\mu_1 = \mu_2 = \lambda$ is the Lebesgue measure, then the product measure $\mu = \lambda \times \lambda$ of a Borel rectangle is

the length subtended in its first coordinate times the length subtended in the second coordinate.

Of course, the problem remains that measurable sets in the product space are more complicated than simple rectangles. Indeed, the product σ-algebra contains countable unions and intersections of rectangles, which can be very intricate. The next result guarantees that there is a unique measure on the product space that satisfies (6.1).

Theorem 6.14 (Product measure: Existence and uniqueness). Let $(\mathsf{X}_i, \mathcal{F}_i, \mu_i)$ be *σ-finite* measure spaces for $i = 1, 2$. The product $(\mathsf{X}, \mathcal{F}) := (\mathsf{X}_1, \mathcal{F}_1) \times (\mathsf{X}_2, \mathcal{F}_2)$ carries a unique measure $\mu := \mu_1 \times \mu_2$, called the *product measure*, that satisfies

$$\mu(\mathsf{E} \times \mathsf{F}) = \mu_1(\mathsf{E}) \cdot \mu_2(\mathsf{F}) \quad \text{for all } \mathsf{E} \in \mathcal{F}_1 \text{ and } \mathsf{F} \in \mathcal{F}_2. \tag{6.2}$$

The triple $(\mathsf{X}, \mathcal{F}, \mu)$ is called the *product* of the measure spaces.

Warning: The construction of product measures fails without σ-finiteness! ▪

The proof of Theorem 6.14 appears in Appendix D. The argument appeals to the Hahn–Kolmogorov theorem and some tools from integration theory. Exercise 6.19 shows that we can also construct the product of a finite number of measures in the natural way.

Example 6.15 (Product of Lebesgue measures). Consider the measure space $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \lambda)$, the real line equipped with Borel sets and the Lebesgue measure. According to Proposition 6.7, the product of this space with itself is $(\mathbb{R}^2, \mathscr{B}(\mathbb{R}^2), \lambda \times \lambda)$. By construction, the product measure satisfies

$$(\lambda \times \lambda)(\mathsf{A} \times \mathsf{B}) = \lambda(\mathsf{A}) \cdot \lambda(\mathsf{B}) \quad \text{for Borel } \mathsf{A}, \mathsf{B} \subseteq \mathbb{R}.$$

For example, when the sets are half-open intervals,

$$(\lambda \times \lambda)((a, b] \times (c, d]) = \lambda((a, b]) \cdot \lambda((c, d]) = |b - a| \cdot |d - c|.$$

This formula is valid for all real numbers that satisfy $a < b$ and $c < d$. Thus, the area of a rectangle is the product of side lengths.

For an arbitrary Borel set $\mathsf{B} \in \mathscr{B}(\mathbb{R}^2)$, we interpret $(\lambda \times \lambda)(\mathsf{B})$ as the *area* of the set B. Since the Borel σ-algebra contains all open sets and all closed sets, we may assign an area to every subset of the plane that is either open or closed, and many more besides.

We usually write $\lambda^2 := \lambda \times \lambda$ for the Lebesgue product measure on $(\mathbb{R}^2, \mathscr{B}(\mathbb{R}^2))$. The formation of products may be repeated. For $n \in \mathbb{N}$, we will write $\lambda^n := \lambda \times \cdots \times \lambda$ for the n-fold product of the Lebesgue measure, defined on the measurable space $(\mathbb{R}^n, \mathscr{B}(\mathbb{R}^n))$. For a Borel set $\mathsf{B} \in \mathscr{B}(\mathbb{R}^n)$, we interpret $\lambda^n(\mathsf{B})$ as the n-dimensional *volume* of B. ▪

Exercise 6.16 (Product measure: Dirac measures). Consider measure spaces $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \delta_x)$ and $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \delta_y)$ for points $x, y \in \mathbb{R}$. What is the product space? Compute the product measure of a measurable rectangle.

Exercise 6.17 (Product measure: Line measure). Consider measure spaces $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \delta_x)$ and $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \lambda)$ for a point $x \in \mathbb{R}$. What is the product space? Compute the product measure of a measurable rectangle.

Exercise 6.18 (Product measure: Counting measures). Consider the discrete measure space $(\mathbb{N}, \mathscr{P}(\mathbb{N}), \#)$. What is the product of this measure space with itself? Compute the product measure of a measurable rectangle. Can you compute the product measure of a general product-measurable set? **Hint:** Reduce to singleton sets by using countable additivity.

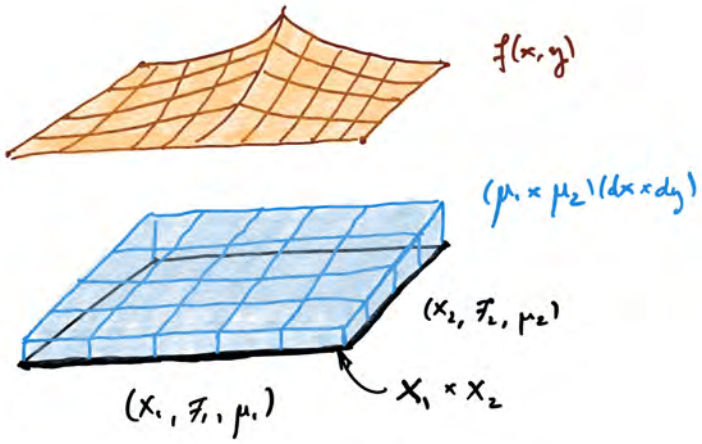


Figure 6.2 (Integration over a product space). On a product space $\mathsf{X}_1 \times \mathsf{X}_2$, the integral sums the values of a bivariate function $f : \mathsf{X}_1 \times \mathsf{X}_2 \to \overline{\mathbb{R}}$ weighted by the local product measure $(\mu_1 \times \mu_2)(\mathrm{d}x \times \mathrm{d}y) = \mu_1(\mathrm{d}x)\,\mu_2(\mathrm{d}y)$. The function f does not need to have a product structure.

Exercise 6.19 (Product measure: Properties). Consider measure spaces $(\mathsf{X}_i, \mathscr{F}_i, \mu_i)$ for $i = 1, 2, 3$. Show that the formation of product measures is associative:

$$(\mu_1 \times \mu_2) \times \mu_3 = \mu_1 \times (\mu_2 \times \mu_3).$$

In particular, the product $\mu_1 \times \mu_2 \times \mu_3$ is well defined, and it has the property that

$$(\mu_1 \times \mu_2 \times \mu_3)(\mathsf{E}_1 \times \mathsf{E}_2 \times \mathsf{E}_3) = \mu_1(\mathsf{E}_1) \cdot \mu_2(\mathsf{E}_2) \cdot \mu_3(\mathsf{E}_3)$$

for all measurable sets $\mathsf{E}_i \in \mathscr{F}_i$ for $i = 1, 2, 3$.

Define the operator $\boldsymbol{R} : (x, y) \mapsto (y, x)$. Show that product measures are commutative, in the sense that

$$(\mu_1 \times \mu_2)(\mathsf{A}) = (\mu_2 \times \mu_1)(\boldsymbol{R}\mathsf{A}) \quad \text{for measurable } \mathsf{A} \in \mathscr{F}_1 \times \mathscr{F}_2.$$

The action of an operator on a set is understood to mean the set obtained by applying the operator to each element.

> **Aside:** We cannot necessarily construct the product of an infinite number of measures without further assumptions. The Kolmogorov extension theorem (Appendix D) describes one important situation where the infinite product of measures is meaningful.

6.2.2 Integration over product measure spaces

A product of measure spaces is just another measure space, so we can compute integrals with respect to the product measure. As usual, the integral sums the values of a measurable function, weighted by the local mass given by the product measure, over the product domain. See Figure 6.2.

For integrals over product spaces, the notation is a little different from the univariate case. Form the product $(\mathsf{X}, \mathscr{F}, \mu)$ of two measure spaces $(\mathsf{X}_1, \mathscr{F}_1, \mu_1)$ and $(\mathsf{X}_2, \mathscr{F}_2, \mu_2)$.

For a measurable function $f : \mathsf{X} \to \mathbb{R}$ whose integral is defined, we use the notation

$$\int_{\mathsf{X}} f(\boldsymbol{x})\, \mu(\mathrm{d}\boldsymbol{x}).$$

The variable of integration $\boldsymbol{x} = (x_1, x_2)$ is written in boldface to emphasize that it is a pair of coordinates rather than a single coordinate. We think about the differential $\mathrm{d}\boldsymbol{x}$ as an infinitesimal rectangle at the point $\boldsymbol{x}$.

We may also write out the product measure $\mu = \mu_1 \times \mu_2$ in full, in which case it is common to use other notations for the differential. For a function $f : \mathsf{X}_1 \times \mathsf{X}_2 \to \mathbb{R}$ where the integral is defined,

$$\begin{aligned}\int_{\mathsf{X}} f(\boldsymbol{x})\, \mu(\mathrm{d}\boldsymbol{x}) &=: \int_{\mathsf{X}_1 \times \mathsf{X}_2} f(x, y)\, (\mu_1 \times \mu_2)(\mathrm{d}x \times \mathrm{d}y) \\ &=: \int_{\mathsf{X}_1 \times \mathsf{X}_2} f(x, y)\, (\mu_1 \times \mu_2)(\mathrm{d}x\, \mathrm{d}y).\end{aligned}$$

These expressions suggest that the two variables may change independently. We can think about the differential as representing an infinitesimal box at (x, y) that has infinitesimal width $\mathrm{d}x$ and infinitesimal height $\mathrm{d}y$. The second notation has the same interpretation. For the Lebesgue measure λ^2, it is quite common to drop the measure from the notation, so $\mathrm{d}\boldsymbol{x} := \lambda^2(\mathrm{d}\boldsymbol{x})$.

Example 6.20 (Product of Lebesgue measures). Let $\mathsf{B} \in \mathscr{B}(\mathbb{R}^2)$ be a Borel set in the plane. As always, the integral of the indicator function of a set is the measure of the set. For the product of Lebesgue measures,

$$\int_{\mathbb{R}^2} \mathbb{1}_{\mathsf{B}}(\boldsymbol{x})\, \lambda^2(\mathrm{d}\boldsymbol{x}) = \lambda^2(\mathsf{B}).$$

We interpret the right-hand side as the area of the Borel set B. ∎

6.3 Interchange of integrals

As you know, we can also compute the area of a plane region by slicing it into thin vertical strips and summing the areas of the strips along the horizontal direction. Likewise, we can compute the area by slicing the region into thin horizontal strips and summing the areas of the strips along the vertical direction. Similarly, if a function describes the density of mass in a plane region, we can find the total mass by adding up the mass of vertical strips or by adding up the mass of horizontal strips.

These geometric principles are intuitive. They suggest that we can compute the integral of a bivariate function with respect to the Lebesgue product measure $\lambda^2 = \lambda \times \lambda$ by integrating with respect to one coordinate and then the other, in either order. More generally, we would like to understand when we can compute integrals with respect to a product measure $\mu_1 \times \mu_2$ by integrating along one coordinate and then the other.

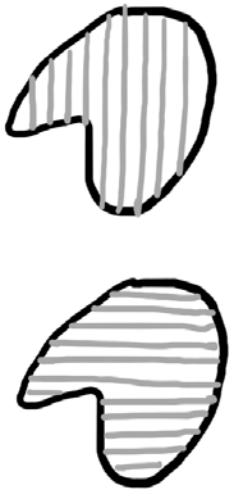


Figure 6.3 (Slicing area). We can compute the area of a plane region by summing the areas of vertical slices or the areas of horizontal slices.

6.3.1 *Measurability of sections

We can only integrate functions that are measurable. Therefore, the first step in our investigation requires us to understand some properties of measurable bivariate functions defined on a product space. In particular, we must verify that the univariate sections of these functions remain measurable. The key fact is a related section property for sets; see Figure 6.4.

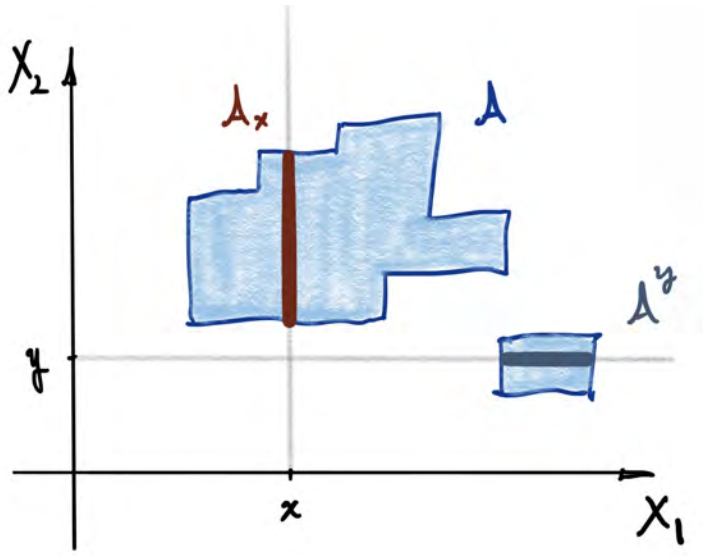


Figure 6.4 (Sections). Given a measurable set A in the product space $\mathsf{X}_1 \times \mathsf{X}_2$, each section A_x and A^y is a measurable set in the factor space.

Proposition 6.21 (Product space: Section property for sets). Let $(\mathsf{X}_i, \mathcal{F}_i)$ be measurable spaces for $i = 1, 2$. For a set $\mathsf{A} \subseteq \mathsf{X}_1 \times \mathsf{X}_2$, define the sections

$$\mathsf{A}_x := \{y \in \mathsf{X}_2 : (x, y) \in \mathsf{A}\} \quad \text{for } x \in \mathsf{X}_1;$$
$$\mathsf{A}^y := \{x \in \mathsf{X}_1 : (x, y) \in \mathsf{A}\} \quad \text{for } y \in \mathsf{X}_2.$$

If the set $\mathsf{A} \in \mathcal{F}_1 \times \mathcal{F}_2$ belongs to the product σ-algebra, then each section A_x is a measurable set in $\mathcal{F}_2$ and each section A^y is a measurable set in $\mathcal{F}_1$.

**Proof.* This argument hinges on the fact that the product $\mathcal{F} := \mathcal{F}_1 \times \mathcal{F}_2$ is the *smallest* σ-algebra that contains all measurable cylinders. Introduce the family $\mathcal{F}_{\text{sect}}$ that contains each product-measurable set whose sections are all measurable:

$$\mathcal{F}_{\text{sect}} := \{\mathsf{A} \in \mathcal{F} : \text{all sections } \mathsf{A}_x \text{ and } \mathsf{A}^y \text{ are measurable}\} \subseteq \mathcal{F}.$$

Clearly, the empty set $\emptyset$ and the product domain $\mathsf{X}_1 \times \mathsf{X}_2$ belong to $\mathcal{F}_{\text{sect}}$. In addition, every measurable cylinder belongs to $\mathcal{F}_{\text{sect}}$. The family $\mathcal{F}_{\text{sect}}$ is stable under complements because

$$(\mathsf{A}^{\mathsf{c}})_x = (\mathsf{A}_x)^{\mathsf{c}} \quad \text{and} \quad (\mathsf{A}^{\mathsf{c}})^y = (\mathsf{A}^y)^{\mathsf{c}}.$$

Likewise, the family $\mathcal{F}_{\text{sect}}$ is stable under (countable) unions because

$$\Big(\bigcup_{i=1}^{\infty} \mathsf{A}_i\Big)_x = \bigcup_{i=1}^{\infty} (\mathsf{A}_i)_x \quad \text{and} \quad \Big(\bigcup_{i=1}^{\infty} \mathsf{A}_i\Big)^y = \bigcup_{i=1}^{\infty} (\mathsf{A}_i)^y.$$

In summary, $\mathcal{F}_{\text{sect}}$ is a σ-algebra that contains all measurable cylinders. Therefore, $\mathcal{F}_{\text{sect}}$ must contain the product σ-algebra $\mathcal{F}$. We conclude that $\mathcal{F}_{\text{sect}} = \mathcal{F}$. ∎

With this result at hand, we can easily derive an analogous section property for functions.

Exercise 6.22 (Product space: Section property for functions). Let $(\mathsf{X}_i, \mathcal{F}_i)$ be measurable spaces for $i = 1, 2$. For a function $f : \mathsf{X}_1 \times \mathsf{X}_2 \to \overline{\mathbb{R}}$ on the product, define the sections

$$f_x : y \mapsto f(x, y) \quad \text{for each } x \in \mathsf{X}_1;$$
$$f^y : x \mapsto f(x, y) \quad \text{for each } y \in \mathsf{X}_2.$$

If f is measurable with respect to the product σ-algebra $\mathcal{F}_1 \times \mathcal{F}_2$, then the section f_x is $\mathcal{F}_2$-measurable for each x and the section f^y is $\mathcal{F}_1$-measurable for each y.

6.3.2 The Fubini–Tonelli theorem

The notorious Fubini–Tonelli theorem states that we can integrate a bivariate measurable function with respect to a product measure by arranging the integrals in any order we like. This result is true under, essentially, minimal conditions.

Theorem 6.23 (Fubini–Tonelli). Consider *σ-finite* measure spaces $(\mathsf{X}_i, \mathcal{F}_i, \mu_i)$ for $i = 1, 2$. Let $f : \mathsf{X}_1 \times \mathsf{X}_2 \to \overline{\mathbb{R}}$ be a Borel measurable function.

1. **Positive case:** For a *positive* function $f \geq 0$,

$$\begin{aligned} \int_{\mathsf{X}_1 \times \mathsf{X}_2} f(x, y)\, (\mu_1 \times \mu_2)(\mathrm{d}x \times \mathrm{d}y) & \\ = \int_{\mathsf{X}_1} \left(\int_{\mathsf{X}_2} f(x, y)\, \mu_2(\mathrm{d}y) \right) \mu_1(\mathrm{d}x) & \qquad (6.3) \\ = \int_{\mathsf{X}_2} \left(\int_{\mathsf{X}_1} f(x, y)\, \mu_1(\mathrm{d}x) \right) \mu_2(\mathrm{d}y). & \end{aligned}$$

2. **Integrable case:** The identities (6.3) also hold for a function f that is *integrable* with respect to $\mu_1 \times \mu_2$. In detail, integrability means that

$$\int_{\mathsf{X}} |f(x, y)|\, (\mu_1 \times \mu_2)(\mathrm{d}x \times \mathrm{d}y) < +\infty.$$

Warning: This result can fail for measures that are not σ-finite! ■

Warning: Confirm that the function f is either positive or integrable before applying this result! ■

The proof of Theorem 6.23 involves some new set theoretic tools. We postpone the argument to Appendix D.4.

Let us emphasize that the statement of Theorem 6.23 only makes sense because of the section property (Exercise 6.22). Indeed, the univariate integrals are understood to mean

$$\int_{\mathsf{X}_1} f(x, y)\, \mu_1(\mathrm{d}x) := \int_{\mathsf{X}_1} f^y(x)\, \mu_1(\mathrm{d}x);$$
$$\int_{\mathsf{X}_2} f(x, y)\, \mu_2(\mathrm{d}y) := \int_{\mathsf{X}_2} f_x(y)\, \mu_2(\mathrm{d}y).$$

These definitions require that all sections are measurable.

In the special case of the product $\lambda \times \lambda$ of Lebesgue measure on $\mathbb{R}^2$, the Fubini–Tonelli theorem implies that the area of a measurable set in the plane equals both the integral of the lengths of the horizontal slices and the integral of the lengths of the vertical slices. We can obtain this statement by applying the theorem to the indicator function $\mathbb{1}_{\mathsf{B}}$ of a Borel set $\mathsf{B} \in \mathcal{B}(\mathbb{R}^2)$.

Exercise 6.24 (Product measure: Dirac measures). Consider measure spaces $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \delta_x)$ and $(\mathbb{R}, \mathcal{B}(\mathbb{R}), \delta_y)$ for points $x, y \in \mathbb{R}$. Compute the product measure $(\delta_x \times \delta_y)(\mathsf{B})$ of

an arbitrary Borel set $\mathsf{B} \in \mathscr{B}(\mathbb{R}^2)$. **Hint:** Write the measure of the set as an integral, and invoke Fubini–Tonelli.

Exercise 6.25 (Product measure: Line measure). Consider measure spaces $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \delta_x)$ and $(\mathbb{R}, \mathscr{B}(\mathbb{R}), \lambda)$ for a point $x \in \mathbb{R}$. Compute the product measure $(\delta_x \times \lambda)(\mathsf{B})$ of an arbitrary Borel set $\mathsf{B} \in \mathscr{B}(\mathbb{R}^2)$.

6.4 Integration by parts

The Fubini–Tonelli theorem is an essential tool. One of the most important applications of this result is a generalized integration by parts formula.

Problem 6.26 (Integration by parts). Let $(\mathsf{X}, \mathscr{F}, \mu)$ be a σ-finite measure space. Let $f : \mathsf{X} \to \mathbb{R}_+$ be a positive measurable function. Consider an increasing, continuously differentiable function $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ with $\varphi(0) = 0$. Establish the identity

$$\int_{\mathsf{X}} (\varphi \circ f)(x)\, \mu(\mathrm{d}x) = \int_0^\infty \mu\{x \in \mathbb{R} : f(x) > t\}\, \varphi'(t)\, \lambda(\mathrm{d}t).$$

Instantiate the cases $\varphi(t) = t$ and $\varphi(t) = t^p$ for $p > 0$ and $\varphi(t) = \mathrm{e}^t - 1$. **Hint:** On the right-hand side, write the measure of the super-level set as the integral of an indicator function, and invoke Fubini–Tonelli. You will also need the fundamental theorem of calculus.

(*) What happens if $\varphi(0) \neq 0$? Can you modify this result to handle the case where φ is not necessarily increasing? What about the situation where φ is increasing but may not be differentiable?

Problems

Exercise 6.27 (Densities). Let μ be a σ-finite measure, and let $f : \mathbb{R} \to \overline{\mathbb{R}}_+$ be a positive, measurable function. Recall that there is a measure $\nu : \mathscr{B}(\mathbb{R}) \to [0, +\infty]$ on Borel sets, determined by

$$\nu(\mathsf{B}) := \int_{\mathsf{B}} f \,\mathrm{d}\mu \quad \text{for all } \mathsf{B} \in \mathscr{B}(\mathbb{R}).$$

For a ν-integrable function $g : \mathbb{R} \to \mathbb{R}$, show that

$$\int_{\mathbb{R}} g \,\mathrm{d}\nu = \int_{\mathbb{R}} gf \,\mathrm{d}\mu.$$

This formulation explains the notation $f = \mathrm{d}\nu/\mathrm{d}\mu$ for the density f. **Hint:** Start with the case where g is positive. Write out the definition of the integral on the left-hand side. Express the measure of the super-level set as the integral of an indicator so that we have access to Fubini–Tonelli.

There is a direct proof of this formula using approximation of g by simple functions; see Proposition 9.5. This approach may be less natural, given our definition of the integral in terms of super-level sets.

Problem 6.28 (*Linear transformations). Let $\boldsymbol{T} : \mathbb{R}^2 \to \mathbb{R}^2$ be an invertible linear map with (nonzero) determinant, $\det(\boldsymbol{T})$. For each λ^2-integrable function $f : \mathbb{R}^2 \to \mathbb{R}$, we will argue that

$$\int_{\mathbb{R}^2} f(\boldsymbol{x})\, \lambda^2(\mathrm{d}\boldsymbol{x}) = \int_{\mathbb{R}^2} f(\boldsymbol{T}\boldsymbol{x}) \cdot |\det(\boldsymbol{T})|\, \lambda^2(\mathrm{d}\boldsymbol{x}). \tag{6.4}$$

This formula implements the linear change of variables $\boldsymbol{x} \mapsto \boldsymbol{T}\boldsymbol{x}$.

1. **(*)** Show that every invertible linear map $\boldsymbol{T} : \mathbb{R}^2 \to \mathbb{R}^2$ can be written as the product of finitely many row-reduction operations of the form
$$\begin{aligned} \boldsymbol{T}_0 &: (x_1, x_2) \mapsto (x_2, x_1); \\ \boldsymbol{T}_1 &: (x_1, x_2) \mapsto (\alpha x_1, x_2) && \text{for } \alpha \neq 0; \\ \boldsymbol{T}_2 &: (x_1, x_2) \mapsto (x_1 + \beta x_2, x_2) && \text{for } \beta \in \mathbb{R}. \end{aligned}$$
2. Show that the result (6.4) holds when $\boldsymbol{T}$ is any one of $\boldsymbol{T}_0, \boldsymbol{T}_1, \boldsymbol{T}_2$. **Hint:** Use Fubini–Tonelli, one-dimensional change of variables formulas (see Exercise 4.40), and the translation invariance (Theorem 3.16) of the Lebesgue measure λ.
3. Deduce that the result (6.4) holds for every invertible linear map $\boldsymbol{T}$. **Hint:** The determinant of a product is the product of determinants.
4. Conclude that the Lebesgue measure λ^2 is invariant under orthogonal linear transformations. **Hint:** The determinant of an orthogonal matrix equals one.
5. *A fortiori*, show that the Lebesgue measure λ^2 is invariant under all rigid motions (i.e., an orthogonal linear transformation followed by a translation).
6. Consider a Borel set $\mathsf{B} \in \mathscr{B}(\mathbb{R})$ in the real line. We can form another Borel set $\mathsf{C} = \mathsf{B} \times \{0\} \subseteq \mathbb{R}^2$ that lies on the horizontal axis in the plane. Explain why $\lambda^2(\mathsf{C}) = 0$. Deduce that $\lambda^2(\boldsymbol{T}\mathsf{C}) = 0$ for each rigid motion $\boldsymbol{T}$. Conclude that "one-dimensional" Borel sets in the plane are negligible with respect to λ^2.

Warning: A subset of a line in $\mathbb{R}^2$ need not be a Borel set in $\mathbb{R}^2$. ∎

7. **(*Multivariate case).** Formulate and prove an extension of (6.4) for invertible linear transformations on $\mathbb{R}^n$.

Problem 6.29 **(**Bivariate change of variables).** Let $\boldsymbol{G} : \Omega \to \mathbb{R}^2$ be a diffeomorphism on an open subset $\Omega \subseteq \mathbb{R}^2$. That is, $\boldsymbol{G}$ is an injective function with a continuous derivative $\mathrm{D}\boldsymbol{G} : \Omega \to \mathbb{R}^{2\times 2}$ that is everywhere invertible: $|\det(\mathrm{D}\boldsymbol{G}(\boldsymbol{x}))| > 0$ for all $\boldsymbol{x} \in \mathbb{R}^2$. For each λ^2-integrable function $f : \Omega \to \mathbb{R}$, we will prove that

$$\int_{\boldsymbol{G}(\Omega)} f(\boldsymbol{x})\, \lambda^2(\mathrm{d}\boldsymbol{x}) = \int_{\Omega} f(\boldsymbol{G}(\boldsymbol{x})) \cdot |\det(\mathrm{D}\boldsymbol{G}(\boldsymbol{x}))|\, \lambda^2(\mathrm{d}\boldsymbol{x}). \tag{6.5}$$

The *image* of the set Ω under the function $\boldsymbol{G}$ is defined as
$$\boldsymbol{G}(\Omega) := \{\boldsymbol{G}\boldsymbol{y} : \boldsymbol{y} \in \Omega\}.$$

This formula implements the nonlinear change of variables $\boldsymbol{x} \mapsto \boldsymbol{G}(\boldsymbol{x})$. The (absolute value of the) determinant of the derivative is often called the *Jacobian* of the transformation. We begin with a sequence of standard reductions. This pattern of argument is common when proving facts about Lebesgue integrals.

1. If the identity (6.5) holds for each positive, measurable function $f : \Omega \to \mathbb{R}_+$, deduce that it holds for all λ^2-integrable functions.
2. If the identity (6.5) holds for each positive, simple function $f : \Omega \to \mathbb{R}_+$, use monotone convergence (Theorem 5.18) to argue that it holds for each positive, measurable function.
3. If the identity (6.5) holds when $f = \mathbb{1}_{\mathsf{B}}$, the indicator of a Borel set $\mathsf{B} \subseteq \Omega$, confirm that it holds for all positive, simple functions.
4. It is helpful to assume that $\mathrm{D}\boldsymbol{G}$ is a bounded function. To do so, we restrict attention to a *compact* set $\Omega' \subset \Omega$. Verify that $\boldsymbol{G}(\Omega')$ is compact, and check that the restriction $\mathrm{D}\boldsymbol{G} : \Omega' \to \mathbb{R}^{2\times 2}$ is bounded.
5. **(*)** Show that there is an increasing sequence $\Omega_1 \subseteq \Omega_2 \subseteq \cdots \subset \Omega$ of compact sets Ω_i for which $\bigcup_{i=1}^{\infty} \Omega_i = \Omega$.

This method is called *exhaustion by compact sets*.

6. Fix an arbitrary compact subset $\Omega' \subset \Omega$. Suppose that (6.5) holds when $f = \mathbb{1}_{\mathsf{B}'}$, the indicator of an arbitrary Borel set $\mathsf{B}' \subseteq \boldsymbol{G}(\Omega')$. Using exhaustion by compact sets, invoke monotone convergence to show that the identity (6.5) remains valid for the indicator $\mathbb{1}_{\mathsf{B}}$ of each Borel set $\mathsf{B} \subseteq \Omega$.

7. **(**)** A *half-open geometric rectangle* is a set of the form $\mathsf{R} = (a, b] \times (c, d]$ where $a < b$ and $c < d$. Fix a parameter $\varepsilon > 0$. Using the machinery from Appendix A, show that each *bounded* Borel set $\mathsf{B} \in \mathscr{B}(\mathbb{R}^2)$ can be written as a finite union of half-open geometric rectangles plus a set $\mathsf{E} \subseteq \mathbb{R}^2$ with Lebesgue measure $\lambda^2(\mathsf{E}) < \varepsilon$:
$$\mathsf{B} = \mathsf{R}_1 \cup \mathsf{R}_2 \cup \cdots \cup \mathsf{R}_m \cup \mathsf{E}.$$
Hint: Show that $\mathscr{B}(\mathbb{R}^2)$ is the "completion" of the algebra generated by the family $\{(a, b] \times (c, d] : a < b, c < d\}$ of half-open rectangles.
8. Using the last two parts, show that it is enough to prove that
$$\int_{\mathbb{R}^2} \mathbb{1}_{\mathsf{R}}(\boldsymbol{x})\, \lambda^2(\mathrm{d}x) = \int_{\mathbb{R}^2} \mathbb{1}_{\mathsf{R}}(\boldsymbol{G}(\boldsymbol{x})) \cdot |\det(\mathrm{D}\boldsymbol{G}(\boldsymbol{x}))|\, \lambda^2(\mathrm{d}\boldsymbol{x}) \tag{6.6}$$
where $\mathsf{R} \subseteq \Omega$ is an arbitrary half-open geometric rectangle.
9. Let $\mathsf{R} \subseteq \Omega$ be any half-open geometric rectangle. Consider the function
$$J(\mathsf{R}) := \frac{\int_{\boldsymbol{G}^{-1}(\mathsf{R})} |\det(\mathrm{D}\boldsymbol{G}(\boldsymbol{x}))|\, \lambda^2(\mathrm{d}\boldsymbol{x})}{\lambda^2(\mathsf{R})}.$$
We can interpret $J(\mathsf{R})$ as the average value of the Jacobian over the set R. By subdividing the rectangle into a large number of tiny, congruent rectangles, argue that $J(\mathsf{R}) = 1$. Deduce that (6.6) is valid. **Hint:** The derivative $\mathrm{D}\boldsymbol{G}$ is bounded and continuous, so it is essentially a *linear* map on a small rectangle. Invoke (6.4) to handle this case.
10. **(*)** Formulate and prove an extension of (6.5) for integrals with respect to the Lebesgue measure λ^n on the n-dimensional Euclidean space $\mathbb{R}^n$. **Hint:** There is no conceptual difference from the two-dimensional case.

Exercise 6.30 (Polar coordinates). In two dimensions, one of the most valuable transformations replaces Cartesian coordinates (x, y) with polar coordinates $(r \cos\theta, r \sin\theta)$ where $r > 0$ and $\theta \in (0, 2\pi)$. Use the formula (6.5) and Fubini–Tonelli to show that
$$\int_{\mathbb{R}\times\mathbb{R}} f(x, y)\, \lambda(\mathrm{d}x)\, \lambda(\mathrm{d}y) = \int_{\theta\in[0,2\pi]} \int_{r\in\mathbb{R}_+} f(r\cos\theta, r\sin\theta)\, r\, \lambda(\mathrm{d}r)\, \lambda(\mathrm{d}\theta).$$
This expression is valid for any measurable function $f : \mathbb{R}^2 \to \mathbb{R}$ that is either positive of λ^2-integrable.

Exercise 6.31 (Gaussian integrals). This exercise outlines the classic calculation of a Gaussian integral. Define
$$I := \int_{\mathbb{R}} \mathrm{e}^{-x^2}\, \lambda(\mathrm{d}x).$$
1. By applying Fubini–Tonelli (with justification), confirm that
$$I^2 = \int_{\mathbb{R}\times\mathbb{R}} \mathrm{e}^{-(x^2+y^2)} (\lambda \times \lambda)(\mathrm{d}x \times \mathrm{d}y).$$
2. Use Exercise 6.30 to change to polar coordinates.
3. Compute the resulting integral with Fubini–Tonelli. What is the value of I?
4. Recall that the standard Gaussian measure on the real line takes the form
$$\gamma(\mathsf{B}) := \frac{1}{\sqrt{2\pi}} \int_{\mathsf{B}} \mathrm{e}^{-x^2/2}\, \lambda(\mathrm{d}x) \quad \text{for all Borel } \mathsf{B} \in \mathscr{B}(\mathbb{R}).$$
Confirm that γ is a Borel probability measure on the real line.

5. The standard Gaussian measure on $\mathbb{R}^n$ takes the form

$$\gamma^n(\mathsf{B}) := \frac{1}{(2\pi)^{n/2}} \int_{\mathsf{B}} \mathrm{e}^{-\|\boldsymbol{x}\|_2^2/2} \, \lambda^n(\mathrm{d}\boldsymbol{x}) \quad \text{for all Borel } \mathsf{B} \in \mathscr{B}(\mathbb{R}^n).$$

Confirm that γ^n is a Borel probability measure on $\mathbb{R}^n$.

Problem 6.32 (*Sinc). For a parameter $M > 0$, observe that

$$\int_{[0,M]} \frac{\sin(x)}{x} \, \lambda(\mathrm{d}x) = \int_{x \in [0,M]} \sin(x) \left(\int_{a \in \mathbb{R}_+} \mathrm{e}^{-ax} \lambda(\mathrm{d}a) \right) \lambda(\mathrm{d}x).$$

1. Compute the limiting value of this integral as $M \to \infty$. **Hint:** In sequence, use Fubini–Tonelli, the fundamental theorem of calculus (FTC), dominated convergence, and then the FTC again.
2. Prove that $x \mapsto \sin(x)/x$ is not Lebesgue integrable with respect to λ on $\mathbb{R}_+$.
3. For a parameter $a > 0$, confirm the following identity:

$$\int_{\mathbb{R}_+} \frac{\sin(x)}{x} \mathrm{e}^{-ax} \, \lambda(\mathrm{d}x) = \frac{\pi}{2} - \arctan(a).$$

Use Fubini–Tonelli here, rather than differentiating under the integral.

Notes

This lecture is based on [LL01, Chap. 1] and [Tao11, Sec. 1.7]. The discussion of Borel sets is adapted from [Wil91]. The change of variables theorems are adapted from Folland's book [Fol99], but the proof here is somewhat different in detail. For nonlinear transformations, the key step in the argument is extracted from [Sch15]. Some of the applications of Fubini–Tonelli are drawn from Folland's book [Fol99] and Driver's notes [Dri12].

Lecture bibliography

[Dri12] B. K. Driver. "Analysis tools with examples". Available online. 2012. URL: https://mathweb.ucsd.edu/~bdriver/240A-C-2016-17/Lecture_Notes/2012%20Notes/240Lecture_Notes_Ver8.pdf.

[Fol99] G. B. Folland. *Real analysis*. Second. Modern techniques and their applications, A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1999.

[LL01] E. H. Lieb and M. Loss. *Analysis*. 2nd ed. American Mathematical Society, 2001. DOI: 10.1090/gsm/014.

[Sch15] R. Schwartz. "The change of variables formula". Available online. 2015. URL: https://www.math.brown.edu/reschwar/M114/notes8.pdf.

[Tao11] T. Tao. *An introduction to measure theory*. American Mathematical Society, Providence, RI, 2011. DOI: 10.1090/gsm/126.

[Wil91] D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

II. probability foundations

7. Probability Spaces

> "The purpose of this monograph is to give an axiomatic foundation for the theory of probability. The author set himself the task of putting in their natural place, among the general notions of modern mathematics, the basic concepts of probability theory—concepts which until recently were considered to be quite peculiar.
>
> "This task would have been a rather hopeless one before the introduction of Lebesgue's theories of measure and integration. However, after Lebesgue's publication of his investigations, the analogies between measure of a set and probability of an event, and between integral of a function and mathematical expectation of a random variable, became apparent. These analogies allowed of further extensions; thus, for example, various properties of independent random variables were seen to be in complete analogy with the corresponding properties of orthogonal functions. But if probability theory was to be based on the above analogies, it still was necessary to make the theories of measure and integration independent of the geometric elements which were in the foreground with Lebesgue. This has been done by Fréchet.
>
> "While a conception of probability theory based on the above general viewpoints has been current for some time among certain mathematicians, there was lacking a complete exposition of the whole system, free of extraneous complications..."
>
> —A. N. Kolmogorov (1933), transl. Morrison (1950)

Agenda:

1. Kolmogorov's model
2. The sample space
3. The σ-algebra of events
4. Probability measures

In today's lecture, we will introduce Kolmogorov's axiomatic model for probability theory, laid out in his 1933 monograph, *Grundbegriffe der Wahrscheinlichkeitsrechnung* or *Foundations of Probability Calculus*.

Kolmogorov reportedly wrote this book to fund repairs on the roof of his *dacha*.

So, what was going on *before* 1933? Was the world deterministic? No! In fact, Palamedes was said to be rolling dice during the Siege of Troy. The point of Kolmogorov's formulation was to ground probability firmly in measure theory, providing a mathematical unity to concepts that previously were disparate and vague.

With our knowledge of measure theory, we can easily state Kolmogorov's model for probability theory. But this model comes along with new terminology and interpretations that take some practice to acquire. The power and richness of this formulation will reveal itself gradually as we proceed.

7.1 Kolmogorov's model

A probabilistic experiment has an unpredictable result, although one often has prior knowledge about the probability of particular outcomes occurring.

Example 7.1 (Basic probabilistic experiments). We consider four examples:

1. **One coin:** We flip a fair coin once. Is the outcome heads or tails?
2. **First heads:** We flip a fair coin repeatedly until the first heads turns up. How many flips does it take?

3. **Linear darts:** We throw a dart at the unit interval $[0, 1]$ in the real line. The dart always strikes the interval, and all positions are "equally likely". Where does the dart hit?
4. **Square darts:** We throw a dart at the unit square $[0, 1]^2$ in the real plane. The dart always strikes the square, and all positions are "equally likely". Where does the dart hit?

In this lecture, we will use these experiments as running examples. ▪

Figure 7.1 (Linear darts). The dart strikes the unit interval at a uniformly random position.

Activity 7.2 (Mumblety-peg). Look up "mumblety-peg" in the dictionary. Please *do not* attempt to play mumblety-peg. ▪

We can treat the two examples involving coins using elementary notions of discrete probability, grounded in combinatorial reasoning. Each individual outcome can be assigned a probability, which gives rise to a distribution of probability over the individual outcomes of the experiment. Unfortunately, this is the *wrong way* to think about probability.

To find the right path, we look to the third experiment on linear darts. How do we make sense of a "uniform distribution" of probability? Indeed, the probability that the dart strikes any particular point is zero. Eventually, we realize the correct way to think about this experiment. The probability that the dart strikes a subset E of the interval $[0, 1]$ should equal the length of the set E. Once we achieve this enlightenment, it becomes clear that we need measure theory to formalize linear darts, because measure theory allows us to define length rigorously. The square darts experiment requires measure theory as well.

As Kolmogorov explains, in the early 20th century, mathematicians gradually realized that measure theory provides a grand unification of all of probability theory. The key insight is to shift our attention to *sets of outcomes* of a probabilistic experiment and to construct a distribution of probability mass over these sets of outcomes. This idea leads to the central definition in modern probability theory.

Definition 7.3 (Probability space). A *probability space* is a measure space $(\Omega, \mathcal{F}, \mathbb{P})$.

1. The *sample space* Ω is an abstract set of points, called *sample points*.
2. The *master σ-algebra* $\mathcal{F}$ contains some subsets of Ω, called *events*.
3. The *probability measure* $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a finite measure that satisfies $\mathbb{P}(\Omega) = 1$. It assigns a probability to each event.

In elementary examples, the sample space contains the possible outcomes of a probability experiment. The master σ-algebra contains sets of outcomes, called events. The probability measure reports how probable it is that the outcome of the experiment is contained in each event. Performing the experiment amounts to observing the particular outcome $\omega_0 \in \Omega$.

In the rest of this lecture, we will unpack this definition, providing concrete examples and interpretations. Probability theory replaces the dry language of measure theory with vivid words that suggest high-stakes wagers in the smoky back room of a Macau casino. One of the early challenges in probability theory is to become fluent with this new terminology.

7.2 The sample space

Probability theory is built on top of measure theory. We work with a fixed domain, which we glorify with its own name.

Definition 7.4 (Sample space). The *sample space* is a fixed set Ω whose elements ω are called *sample points* or *outcomes*. This notation is standard.

In the simplest examples, the sample space provides a direct description of a probability experiment. Each sample point $\omega \in \Omega$ is a possible outcome of the experiment. Performing the experiment amounts to distinguishing a particular outcome $\omega_0 \in \Omega$.

You can think about Tyche, the Greek goddess of chance, electing the outcome ω_0 according to her divine whim. Tyche's Roman counterpart is named Fortuna.

Example 7.5 (Basic experiments). In our three experiments, it is easy to identify the sample points and the sample space.

1. **One coin:** Recall that we flip a single fair coin. The sample space $\Omega = \{\mathrm{H}, \mathrm{T}\}$ consists of the outcomes H = heads and T = tails. The outcome ω_0 of the experiment is the (random) outcome of the coin flip.
2. **First heads:** We flip a fair coin until it turns up heads. The sample space $\Omega = \{1, 2, 3, \dots\} = \mathbb{N}$. The outcome ω_0 of the experiment is the (random) number of flips that we perform before we see the first heads.
3. **Linear darts:** We throw a dart, which strikes the unit interval $[0, 1]$ at a uniformly random location. The sample space $\Omega = [0, 1]$. The outcome ω_0 of the experiment is the random point in $[0, 1]$ where the dart strikes.
4. **Square darts:** We throw a dart, which strikes the unit square $[0, 1]^2$ at a uniformly random location. The sample space $\Omega = [0, 1]^2$. The outcome $\boldsymbol{\omega}_0$ of the experiment is the random point in $[0, 1]^2$ where the dart strikes.

In each of these cases, the sample space is simply the collection of possible outcomes of the experiment. ∎

It is productive to take a broader view of the sample space. In many circumstances, we are not performing a single experiment (or sequence of experiments) that we can easily describe with a list of concrete outcomes. Rather, we may want to think about a sample point ω as describing the complete state of the system we are studying. The sample space Ω contains all possible states that could occur. If we knew the actual state ω_0 of the system, then we would know the outcome of every possible experiment that we might perform.

Example 7.6 (Statistical mechanics). Suppose that we measure the temperature of my morning coffee at the beginning of class. This is a probabilistic experiment. One natural sample space is just the set of positive numbers, which corresponds to the temperature in degrees Kelvin.

But we might also want the sample space to include more information. For example, a sample point might list the position, momentum, and type of each molecule in the cup at each point in time. The sample space then consists of all possible sample points of this type. Of course, some sample points are more likely than others.

In a classic thought experiment, Maxwell's demon is able to move all of the molecules to to the top half of the cup. This is a possible, but extremely unlikely, state for the system.

Given a sample point, we can indeed compute the temperature of the coffee. We can also compute other thermodynamic quantities, such as the viscosity (!). The temperature alone gives us a very limited picture of the system, and we cannot ask more complicated questions if this is the only piece of information that we have.

Even though a sample point gives a fundamental description of the state of the system, it is not necessarily something that we can observe. ∎

Example 7.7 (Random number generator). A *random number generator* (RNG) is a mechanism that takes a finite input, called a *random seed*, and produces a long (but finite) sequence of pseudorandom bits. Computer scientists study the setting where the random seed contains a small number of truly random bits that the RNG expands

into a much longer sequence of pseudorandom bits. If we do not know the random seed, then a (computationally bounded) statistical test cannot distinguish the list of pseudorandom bits from an independent sequence of truly random bits.

We can model an RNG using a sample space Ω whose points are all possible values of the random seed. The RNG is a function that maps a random seed to a (finite) binary sequence. Once we select a particular random seed ω_0, the output of the RNG is deterministic. ▪

The RNG example suggests a useful way to think about other kinds of probability models. Tyche chooses the state ω_0 of the world. This determines the outcomes of all (classical) experiments. To the observer, who does not know the exact state, the experimental outcomes appear random. Performing a large number of experiments can provide information about the state, but we may never be able to determine the state completely from a limited family of experiments.

In elementary probability theory, the sample space plays a central role because it lists the specific outcomes of a well-defined experiment. In more advanced applications, the sample space recedes in importance. The identity ω_0 of the distinguished sample point also has limited significance (and we may not know ω_0 in any case).

7.3 The σ-algebra of events

The main insight behind modern probability theory is to assign probabilities to *sets of outcomes* of an experiment, not to individual outcomes. This shift in perspective is crucial.

When we studied measure theory on the real line, we learned that it is not possible to assign a consistent length to each subset of the real line. This led us to introduce the family of Borel sets in the real line, which are sets that have a well-defined length. Not every subset of the real line is a Borel set.

In the same way, we may not be able to assign a probability to every subset of the sample space. Instead, we isolate a collection of subsets of the sample space that will have well-defined probabilities. This collection must form a σ-algebra, so that we can define a (probability) measure on it.

Definition 7.8 (Master σ-algebra of events). The *master σ-algebra* $\mathcal{F}$ is a σ-algebra on the sample space Ω. The sets E that belong to $\mathcal{F}$ are called *events*. Each event E is a collection of sample points.

Warning: In general, not every subset of the sample space is an event. ▪

We use special language to talk about events. Suppose that $\omega_0 \in \Omega$ is the distinguished sample point. If $\omega_0 \in \mathsf{E}$, we say that the event E *occurs*. If $\omega_0 \notin \mathsf{E}$, we say that the event E *does not occur*. We will present more terminology in a minute.

Example 7.9 (Basic experiments). Let us describe the master σ-algebra $\mathcal{F}$ of events that we use for each of our experiments.

1. **One coin:** Recall that the sample space $\Omega = \{\mathrm{H}, \mathrm{T}\}$. The master σ-algebra $\mathcal{F}$ contains all subsets of the sample space:

$$\mathcal{F} = \big\{\emptyset, \{\mathrm{H}\}, \{\mathrm{T}\}, \{\mathrm{H}, \mathrm{T}\}\big\}.$$

   For example, if the outcome $\omega_0 = \mathrm{T}$, then the events $\{\mathrm{T}\}$ and $\{\mathrm{H}, \mathrm{T}\}$ occur. The events $\emptyset$ and $\{\mathrm{H}\}$ do not occur.
2. **First heads:** Recall that the sample space $\Omega = \mathbb{N}$. Once again, the master σ-algebra $\mathcal{F}$ contains all subsets of the sample space: $\mathcal{F} = \mathcal{P}(\mathbb{N})$. Particular events include

things like

$$\mathsf{E} = \{n \in \mathbb{N} : n \text{ is even}\} \quad \text{and} \quad \mathsf{F} = \{n \in \mathbb{N} : n \leq 10\}.$$

For example, if the outcome $\omega_0 = 7$, then the event F occurs but E does not.

3. **Linear darts:** The sample space $\Omega = [0, 1]$. The master σ-algebra of events is $\mathscr{F} = \mathscr{B}([0, 1])$, the collection of Borel sets in $[0, 1]$. In other words, events are subsets of $[0, 1]$ that have a well-defined length. The distinguished sample point $\omega_0 \in [0, 1]$ is the location where the dart strikes. An event E occurs when the dart lands in the set E; that is, $\omega_0 \in \mathsf{E}$.
4. **Square darts:** Now, the sample space $\Omega = [0, 1]^2$. The master σ-algebra of events is $\mathscr{F} = \mathscr{B}([0, 1]^2)$, the collection of Borel sets in $[0, 1]^2$. Events are those subsets of $[0, 1]^2$ that have a well-defined area. The distinguished sample point $\omega_0 \in [0, 1]^2$ is the location where the dart strikes. An event E occurs when the dart lands in the set E; that is, $\omega_0 \in \mathsf{E}$.

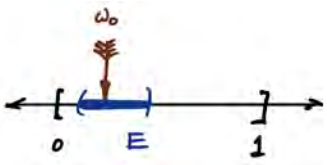


Figure 7.2 (An event in linear darts).

In these situations, we can identify the master σ-algebra using considerations from basic measure theory. It is not always so obvious. ▪

Since $\mathscr{F}$ is a σ-algebra, countable combinations of events are always events. When we talk about events, we replace the abstract language of set theory with more concrete terminology. Here are some of the most important examples. In the following, $\mathsf{E}, \mathsf{F} \in \mathscr{F}$ are events.

- $\emptyset$ is called the *impossible event*. This event cannot occur because the distinguished sample point $\omega_0 \notin \emptyset$.
- Ω is called the *certain event*. This event always occurs because the distinguished sample point $\omega_0 \in \Omega$.
- The event $\mathsf{E}^c := \Omega \setminus \mathsf{E}$ is the event that E does not occur.
- The event $\mathsf{E} \cap \mathsf{F}$ is the event that both E and F occur.
- If the events E and F are disjoint ($\mathsf{E} \cap \mathsf{F} = \emptyset$), then we say that E and F are *mutually exclusive*.
- The event $\mathsf{E} \cup \mathsf{F}$ is the event that E or F occurs. This includes the possibility that both E and F occur.
- The event $\mathsf{E} \triangle \mathsf{F}$ is the event that exactly one of E or F occurs.
- The event $\mathsf{E} \setminus \mathsf{F}$ is the event that E occurs but F does not.

Further terminology will be introduced as needed.

When constructing a probability space, we generally prefer the master σ-algebra $\mathscr{F}$ to be as large as possible. We want as many events as we can get. This choice helps us ensure that all sets that arise from our considerations remain events. This principle is limited by the fact that we also need to construct interesting probability measures.

The reason for making the master σ-algebra $\mathscr{F}$ as large as possible is so that we do not have to think about it very often. Later, we will see that smaller σ-algebras, contained in $\mathscr{F}$, play an important role in probability theory.

As a rule of thumb, if the sample space Ω is finite or countable, then the master σ-algebra $\mathscr{F} = \mathscr{P}(\Omega)$, the collection of all subsets of the sample space. If Ω is (a Borel subset of) the Euclidean space $\mathbb{R}^n$, then the master σ-algebra $\mathscr{F} = \mathscr{B}(\Omega)$, the Borel sets in $\mathbb{R}^n$ intersected with Ω.

In detail, for a Borel set $\Omega \subseteq \mathbb{R}^n$,

$$\mathscr{B}(\Omega) := \{\mathsf{B} \cap \Omega : \mathsf{B} \in \mathscr{B}(\mathbb{R}^n)\}.$$

Aside: More generally, if the sample space Ω is a separable metric space, then we can equip Ω with the master σ-algebra $\mathcal{F} = \mathcal{B}(\Omega)$ of Borel sets, generated by the open metric balls. In still more general settings, additional care may be warranted.

7.4 The probability measure

We have intimated that Tyche designates a "random" sample point $\omega_0 \in \Omega$. So far, we do not have a way to model what sample points are more or less likely. The probability measure $\mathbb{P}$ encapsulates this information.

Definition 7.10 (Probability measure). Let Ω be a sample space, equipped with a master σ-algebra $\mathcal{F}$. A *probability measure* $\mathbb{P} : \mathcal{F} \to [0, 1]$ is a function with three properties:

1. **Impossible event:** The probability $\mathbb{P}(\emptyset) = 0$.
2. **Certain event:** The probability $\mathbb{P}(\Omega) = 1$.
3. **Countable additivity:** For a countable sequence $(\mathsf{A}_i \in \mathcal{F} : i \in \mathbb{N})$ of *mutually exclusive* events,
$$\mathbb{P}\big(\dot{\bigcup}_{i=1}^{\infty} \mathsf{A}_i\big) = \sum_{i=1}^{\infty} \mathbb{P}(\mathsf{A}_i).$$

Warning: A probability measure assigns probabilities to events, not sample points! ▪

The probability of nothing happening is zero, or 0%. The probability of something happening is one, or 100%.

Activity 7.11 (Probability measures). Since a probability measure is just a finite measure, it satisfies all of the usual properties of a measure (Propositions 2.29 and 2.30). In particular, we can use monotonicity, the inclusion–exclusion law, the countable subadditivity property (called *Boole's law* in this context), and the results on limits of increasing and decreasing sequences of sets. Write out each of these results using probabilistic notation and language.

In addition, check that $\mathbb{P}(\mathsf{E}^c) = 1 - \mathbb{P}(\mathsf{E})$ for each event $\mathsf{E} \in \mathcal{F}$. What does this statement mean in words? ▪

For simple probability models, we can give an explicit description of the probability measure.

Example 7.12 (Basic experiments). Here are the natural probability measures for our experiments.

1. **One coin:** Recall that the sample space $\Omega = \{\mathrm{H}, \mathrm{T}\}$ and the master σ-algebra $\mathcal{F} = \mathcal{P}(\Omega)$. Since the coin is fair, the probability measure $\mathbb{P}$ assigns equal probability to each singleton outcome:
$$\mathbb{P}\{\mathrm{H}\} := \tfrac{1}{2} \quad \text{and} \quad \mathbb{P}\{\mathrm{T}\} := \tfrac{1}{2}.$$
Of course, $\mathbb{P}(\emptyset) = 0$ and $\mathbb{P}(\{\mathrm{H}, \mathrm{T}\}) = 1$.
2. **First heads:** Recall that $\Omega = \mathbb{N}$ and $\mathcal{F} = \mathcal{P}(\mathbb{N})$. Since the coin is fair, we can use combinatorial reasoning to determine the probability of each singleton outcome:
$$\mathbb{P}(\{n\}) := 2^{-n} \quad \text{for } n \in \mathbb{N}.$$
The probability of a general event $\mathsf{E} \subseteq \mathbb{N}$ is now determined by countable additivity:
$$\mathbb{P}(\mathsf{E}) = \sum\nolimits_{n \in \mathsf{E}} \mathbb{P}\{n\} = \sum\nolimits_{n \in \mathsf{E}} 2^{-n}.$$
You may want to verify that this approach leads to a well-defined probability measure on $\mathcal{F}$.

3. **Linear darts:** Recall that $\Omega = [0,1]$ and $\mathscr{F} = \mathscr{B}([0,1])$. For an event $\mathsf{E} \in \mathscr{F}$, the probability is determined by geometric reasoning:

$$\mathbb{P}(\mathsf{E}) := \frac{\text{length}(\mathsf{E})}{\text{length}([0,1])} = \lambda(\mathsf{E}).$$

   In other words, the probability that the dart lands in a (Borel) subset E of the interval $[0,1]$ is equal to its length $\lambda(\mathsf{E})$. Since the Lebesgue measure λ is translation invariant, this is the natural model for a uniform distribution of probability. Theorem 3.16, on the properties of the Lebesgue measure, ensures that $\mathbb{P}$ is a probability measure.
4. **Square darts:** Recall that $\Omega = [0,1]^2$ and $\mathscr{F} = \mathscr{B}([0,1]^2)$. For an event $\mathsf{E} \in \mathscr{F}$, the probability is again determined by geometric reasoning:

$$\mathbb{P}(\mathsf{E}) := \frac{\text{area}(\mathsf{E})}{\text{area}([0,1]^2)} = \lambda^2(\mathsf{E}).$$

   The probability that the dart lands in a (Borel) subset E of the unit square $[0,1]^2$ is equal to its area $\lambda^2(\mathsf{E})$. The Lebesgue measure λ^2 on the plane is translation invariant, so this is the natural model for a uniform distribution of probability. Theorem 6.14, on the product measure, ensures that $\mathbb{P}$ is a probability measure.

In each of these cases, the probability measure is determined by direct reasoning. ▪

The linear darts example confirms that we need to use measure theory to develop a rigorous account of probability. These examples also show that we can describe discrete probability models using exactly the same measure-theoretic framework. When viewed in this light, the difference between discrete and continuous probability blurs. Both cases are unified.

Keep in mind that the probability $\mathbb{P}$ is a *measure*, even though we are using a different notation now. As a consequence, you will sometimes see the probability of an event $\mathsf{E} \in \mathscr{F}$ written in terms of an integral:

$$\mathbb{P}(\mathsf{E}) = \int_\Omega \mathbb{1}_\mathsf{E}(\omega)\, \mathbb{P}(\mathrm{d}\omega) = \int_\Omega \mathbb{1}_\mathsf{E}\, \mathrm{d}\mathbb{P} = \int_\mathsf{E} \mathrm{d}\mathbb{P}.$$

You may encounter other similar notations.

Finally, one more piece of terminology. In probability theory, we generally replace the term "almost everywhere" with the term "almost sure". Thus, an event E with $\mathbb{P}(\mathsf{E}) = 1$ is called an *almost sure* event or, for emphasis, a $\mathbb{P}$-*almost sure* event.

> **Aside:** How much information do we need to determine a probability measure? Let $\mathscr{A}$ be an algebra of events that generates the master σ-algebra: $\sigma(\mathscr{A}) = \mathscr{F}$. If we can define a premeasure $\mathbb{P}_0$ on $\mathscr{A}$ that satisfies $\mathbb{P}_0(\Omega) = 1$, then the Hahn–Kolmogorov theorem (Appendix A) yields a unique probability measure $\mathbb{P}$ that extends $\mathbb{P}_0$ to $\mathscr{F}$. We will mostly construct probability measures from measures that we have already defined.

Quiz

You should take the opportunity to construct a few probability models of your own.

Activity 7.13 (Probability spaces). What are the natural probability spaces for describing the following probability experiments?

1. Roll one fair die.
2. Roll two fair dice.
3. Flip 10 fair coins.
4. Flip a fair coin a countably infinite number of times.
5. A lightbulb has an exponential lifetime, with mean lifetime of 1000 hours.

Make sure to list the sample space, the master σ-algebra of events, and a rule for determining the probability of every event. ▪

This activity suggests that simple discrete probability spaces are easy to identify. It is much less clear, however, how to construct a probability model for an infinite number of coin flips. (Each infinite sequence of outcomes HTHTHHTT... seems to have zero probability!) Similarly, you may not find it obvious how to define a probability measure that describes the exponential lifetime of a lightbulb. We will start working toward these goals in the next lecture.

Applications

Application 7.14 (Probabilistic method). The *probabilistic method* is a fundamental approach for establishing the existence of an object that satisfies some property. In this exercise, we will present some simple examples of this rich methodology. In particular, we will consider applications to coding theory and combinatorics.

1. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. For an event $\mathsf{E} \in \mathcal{F}$, show that the condition $\mathbb{P}(\mathsf{E}) > 0$ implies that $\mathsf{E} \neq \emptyset$. In other words, an event with strictly positive probability contains a sample point $\omega \in \mathsf{E}$ that witnesses the property described by the event.
2. **(Warmup: Street art).** Sometime during the night, Banksy paints the unit circle in the plane so that a (measurable) subset of 23% of the circle is red and the rest is blue. Regardless of the artistic quality, show that we can inscribe a square in the circle so that all four vertices are blue. **Hint:** Choose the square at random, and show that there is a positive probability that its vertices are all blue.

   The unit circle is the set $\{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 1\}$.
3. **(Kraft inequality).** A finite collection $\mathscr{C}$ of binary strings with finite lengths is called a *prefix-free code* if no string in $\mathscr{C}$ is a prefix of another string in $\mathscr{C}$. For each $i \in \mathbb{N}$, let N_i denote the number of strings of length i in $\mathscr{C}$. Establish the Kraft inequality, a limit on the number of codewords in a prefix-free code:

   $$\sum_{i \in \mathbb{N}} \frac{N_i}{2^i} \leq 1.$$

   Hint: Flip a fair coin until the first time that the sequence of outcomes appears as a string in $\mathscr{C}$. What is the probability that this event occurs on the ith flip?

   For example, the string "11" is a prefix of the strings "1100" and "1111".
4. **(**Kraft–McMillan inequality).** A finite collection $\mathscr{C}$ of binary strings with finite lengths is called a *uniquely decipherable code* if no pair of strings in $\mathscr{C}$ concatenates to form another string that appears in $\mathscr{C}$. For each $i \in \mathbb{N}$, let N_i denote the number of strings of length i in the code $\mathscr{C}$. Establish the Kraft–McMillan inequality, a limit on the number of codewords in a uniquely decipherable code:

   $$\sum_{i \in \mathbb{N}} \frac{N_i}{2^i} \leq 1.$$

   For example, the concatenation "11" + "00" = "1100". A uniquely decipherable code cannot contain all three of these strings.
5. **(*Diagonal Ramsey numbers).** Consider a complete graph K_n on n vertices. Assign each edge a color, either red or blue. A subgraph on r vertices is *monochromatic*

   The *complete graph* K_n is an undirected combinatorial graph on n vertices, and there is an edge connecting each pair of distinct vertices.

if its edges are all red or all blue. The *(diagonal) Ramsey number* $R(r)$ is the least value of n for which the graph must contain a monochromatic subgraph on r vertices, regardless of the choice of coloring. Ramsey (1929) showed that $R(r)$ is finite. We will develop a lower bound on the Ramsey number. Prove that

$$\binom{n}{r} < 2^{\binom{r}{2}-1} \quad \text{implies} \quad R(r) > n.$$

In particular, $R(r) > \lfloor 2^{r/2} \rfloor$. **Hint:** Choose the coloring at random. For each fixed set S of r vertices, consider the event that S is monochromatic.

Notes

The axiomatic foundation of probability theory can be traced back to Kolmogorov's work. You may find a similar account in any serious book on probability. For example, see Billingsley [Bil12] or Durrett [Dur19]. Our focus here is to connect probability theory with the measure theory foundations we have already developed. We also hope to build intuition about the role of the sample space, the master σ-algebra of events, and the probability measure.

For a survey of the probabilistic method, see the book of Alon & Spencer [AS16]. Most of our examples are drawn from their work. The geometric example of the probabilistic method is adapted from Grimmett & Stirzaker [GS01]. We will explore a few more examples in upcoming lectures as we develop additional tools. Unfortunately, the most interesting applications of the probabilistic method involve elaborate domain-specific reasoning that is outside the scope of this course.

8. Random Variables

Agenda:

1. Random variables
2. The law of a random variable
3. Distribution functions
4. Types of random variables

"I took the law and threw it away.
There's nothing wrong, it's just for play.
There's no law, no law anymore.
I want to steal from the rich and give to the poor."

—*Howling at the Moon (Sha-La-La)*, The Ramones (1984)

In the last lecture, we introduced the concept of a probability space. This is the arena where probability theory takes place. In this lecture, we will develop the concept of a real random variable, which you can think about as a single real-valued observation of a probabilistic system. In the next lecture, we will introduce the expectation, which gives the average value of a real random variable.

For motivation, recall that a probability space is a measure space $(\Omega, \mathcal{F}, \mathbb{P})$. The sample space Ω is a set of sample points. The master σ-algebra $\mathcal{F}$ contains events, which are sets of sample points. The probability measure $\mathbb{P}$ assigns a probability to each event.

In simple examples, the sample space is just the collection of possible outcomes of a concrete experiment. Before performing the experiment, we only have a probabilistic description of the outcomes, represented by the probability measure on events. Once we perform the experiment, we can observe the specific outcome $\omega_0 \in \Omega$. Tyche, the goddess of chance, elects the outcome ω_0 at random. The prior probability that ω_0 belongs to any particular event is governed by the probability measure $\mathbb{P}$.

In more sophisticated examples, the points in the sample space are interpreted as possible states of the world. Each sample point gives the complete description, or state, of a system that may be very complicated. *A priori*, the probability measure describes what subsets of states are more or less likely. As before, chance determines what state actually occurs. But the state itself may not be observable because of the complexity of the model.

So, what do we observe? It is productive to think about experiments as measurements that deliver pieces of information about the world. The outcomes of these experiments are determined by the (inaccessible) state. This intuition leads us to introduce the idea of a random variable, which is a function of the state. Since the state is random, the outcome of the experiment is also random. The difference is that we can observe the values of the random variables, but we may not be able to observe the underlying state. Nevertheless, we can learn something about the state by collecting observations.

8.1 Real random variables

We would like a formalism for describing a real-valued observable of a system that exhibits probabilistic behavior. The following definition captures this idea.

Definition 8.1 (Real random variable). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A *real random variable* is a *measurable* function $X : \Omega \to \mathbb{R}$.

In other words, a random variable X maps each sample point $\omega \in \Omega$ to a real value $X(\omega)$. Once Tyche designates a particular sample point $\omega_0 \in \Omega$, the value $X(\omega_0)$ of the random variable is completely determined. In fact, the choice ω_0 of the sample point determines the value of *every* random variable.

At first, this definition can be very confusing. A random variable is a fixed function on the sample space; there is nothing random about it. All of the randomness comes from the distribution of the sample point, which is modeled by the probability measure $\mathbb{P}$ on the sample space.

Example 8.2 (Basic experiments). For the basic probability experiments we discussed last time, each sample space can be placed in correspondence with a subset of the real line. Therefore, we do not have to look very far to find random variables.

1. **One coin:** The sample space $\Omega = \{\mathrm{H}, \mathrm{T}\}$, the σ-algebra $\mathcal{F} = \mathcal{P}(\Omega)$, and the probability measure $\mathbb{P}$ is the uniform measure ($\mathbb{P}\{\omega\}$ is constant for all $\omega \in \Omega$). We can define the random variable

$$X(\omega) := \begin{cases} 1, & \omega = \mathrm{H}; \\ 0, & \omega = \mathrm{T}. \end{cases}$$

   This is the indicator random variable of the event that the coin turns up heads.
2. **Head count:** Suppose we flip a fair coin n times. We may consider the sample space $\Omega = \{\mathrm{H}, \mathrm{T}\}^n$ with the σ-algebra $\mathcal{F} = \mathcal{P}(\Omega)$. The probability measure on Ω is uniform; it satisfies $\mathbb{P}\{\boldsymbol{\omega}\} = 2^{-n}$ for each $\boldsymbol{\omega} \in \Omega$. We can define a random variable

$$X(\boldsymbol{\omega}) = \#\{i \in \{1, \dots, n\} : \omega_i = \mathrm{H}\} \quad \text{for each } \boldsymbol{\omega} \in \Omega.$$

   This random variable reports the number of heads that turn up in n coin flips.
3. **First heads:** Suppose we flip a fair coin n times. The sample space $\Omega = \mathbb{N}$, and the σ-algebra $\mathcal{F} = \mathcal{P}(\mathbb{N})$. The probability measure satisfies $\mathbb{P}\{\omega\} = 2^{-\omega}$ for $\omega \in \mathbb{N}$. We can obviously define the random variable

$$X(\omega) = \omega \quad \text{for } \omega \in \Omega.$$

   This random variable counts the number of flips before the first heads turns up.
4. **Linear darts:** The probability space $([0, 1], \mathcal{B}([0, 1]), \lambda)$. We can obviously define the random variable

$$X(\omega) = \omega \quad \text{for } \omega \in [0, 1].$$

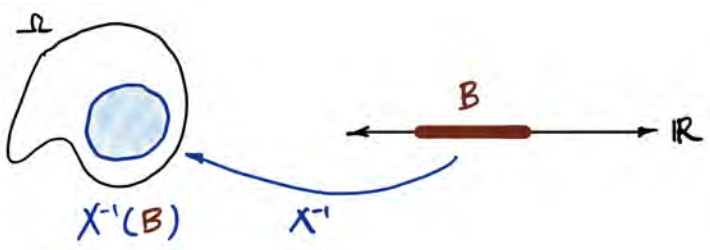
   This random variable reports the position where the dart strikes.

More interesting probability experiments lead to more interesting random variables. ▪

Measurability is a crucial feature of the definition of a random variable. Recall that a function $X : \Omega \to \mathbb{R}$ is measurable when

$$X^{-1}(\mathsf{B}) := \{\omega \in \Omega : X(\omega) \in \mathsf{B}\} \in \mathcal{F} \quad \text{for all Borel } \mathsf{B} \in \mathcal{B}(\mathbb{R}).$$

In words, the preimage of every Borel set is an event.



This construction will allow us to use the probability measure on the sample space to determine the distribution of values of the random variable.

Remark 8.3 **(*Random variables with extended values).** It is sometimes convenient to work with a random variable X that takes extended real values. That is, $X : \Omega \to \overline{\mathbb{R}}$ is measurable with respect to $\mathcal{B}(\overline{\mathbb{R}})$. In this course, random variables take finite values unless extended values are explicitly allowed.

> **Aside:** It is easy to extend the definition of a random variable beyond the real case. Let $(\mathsf{M}, \mathcal{G})$ be a measurable space. An M-valued random variable is a measurable function $X : \Omega \to \mathsf{M}$. That is,
>
> $$X^{-1}(\mathsf{G}) := \{\omega \in \Omega : X(\omega) \in \mathsf{G}\} \in \mathcal{F} \quad \text{for all } \mathsf{G} \in \mathcal{G}.$$
>
> The preimage of every $\mathcal{G}$-measurable set is an event. For the time being, when we say "random variable," we are always referring to a real random variable.

8.2 The law of a random variable

As we have noted, all of the randomness in a random variable comes from the randomness inherent in the selection of the sample point. The probability distribution on the sample space induces a distribution over the values of a random variable.

To see how this works, let $X : \Omega \to \mathbb{R}$ be a real random variable. For each Borel set $\mathsf{B} \in \mathcal{B}(\mathbb{R})$ in the real line, we can compute the probability

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

This is eminently reasonable because the preimage $X^{-1}(\mathsf{B})$ is an event, so it has a well-defined probability. This approach allows us to define a measure on the real line.

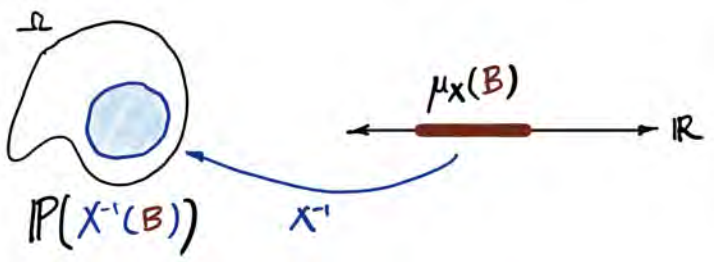
Definition 8.4 **(Law of a random variable).** Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $X : \Omega \to \mathbb{R}$ be a real-valued random variable. The *law* of the random variable is the Borel measure μ_X on the real line defined by

$$\mu_X(\mathsf{B}) := \mathbb{P}\big(X^{-1}(\mathsf{B})\big) = \mathbb{P}\{X \in \mathsf{B}\} \quad \text{for all Borel } \mathsf{B} \in \mathcal{B}(\mathbb{R}).$$

The law of the random variable is also called the *distribution* of the random variable.

Exercise 8.5 (The law is a probability measure). Check that the law μ_X of a real random variable X is a Borel probability measure. That is, μ_X is a Borel measure on $\mathbb{R}$ with $\mu_X(\mathbb{R}) = 1$.

Here is an illustration of the relationship between the probability measure $\mathbb{P}$ and the law μ_X of the random variable X:



The random variable X pushes the distribution $\mathbb{P}$ of probability on the sample space Ω forward to a distribution μ_X of probability on the real line $\mathbb{R}$. For each Borel set $\mathsf{B} \in \mathscr{B}(\mathbb{R})$, the law tells us the probability $\mu_X(\mathsf{B})$ that the random variable X takes a value in B.

This is an example of the push-forward of a measure that we encountered in Problem 5.44.

Since the measure of a set is the integral of the indicator, we can represent the probability of an event involving a random variable by integrating the law. For each Borel set $\mathsf{B} \in \mathscr{B}(\mathbb{R})$,

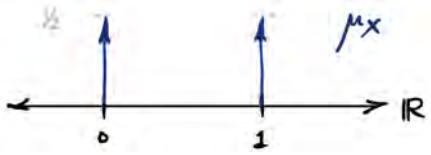
$$\mathbb{P}\{X \in \mathsf{B}\} = \mu_X(\mathsf{B}) = \int_{\mathbb{R}} \mathbb{1}_{\mathsf{B}}(x)\, \mu_X(\mathrm{d}x) = \int_{\mathbb{R}} \mathbb{1}_{\mathsf{B}}\, \mathrm{d}\mu_X = \int_{\mathsf{B}} \mathrm{d}\mu_X.$$

This relation is an example of the change of variables formula (5.8). Each of the integrals represents the same thing, and you may encounter any one of these notations out in the wild.

Example 8.6 (Basic experiments). Let us describe the probability laws for the random variables arising from our basic experiments.

1. **One coin:** The indicator random variable X that the coin turns up heads follows that BERNOULLI$(1/2)$ distribution:

$$\mu_X = \tfrac{1}{2}\delta_0 + \tfrac{1}{2}\delta_1$$

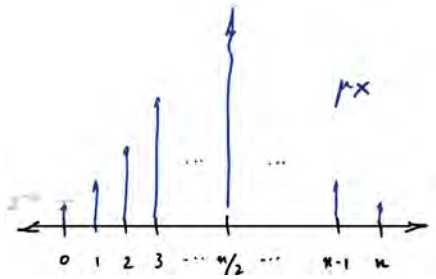


   The support of the law μ_X is the Borel set $\{0, 1\}$.

   See Definition 3.20 of the support of a Borel measure on the real line.

2. **Head count:** The random variable that counts the number of heads follows the BINOMIAL$(1/2, n)$ distribution.

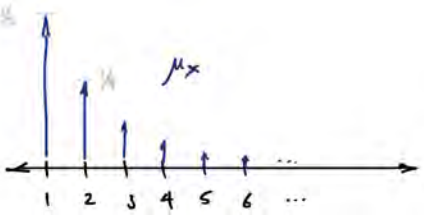
$$\mu_X = 2^{-n} \textstyle\sum_{i=0}^{n} \binom{n}{i} \delta_i$$



The support of the law μ_X is the Borel set $\{0, 1, 2, \ldots, n\}$.

3. **First heads:** The random variable X that reports the number of coin flips before we see the first heads follows the GEOMETRIC$(1/2)$ distribution.

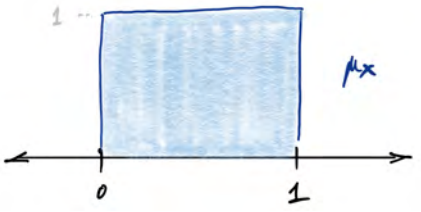
$$\mu_X = \sum_{n=1}^{\infty} 2^{-n} \delta_n$$



The support of the law μ_X is the Borel set $\mathbb{N}$.

4. **Linear darts:** The random variable X that describes the position of the dart follows the UNIFORM$[0, 1]$ distribution.

$$\mu_X = \lambda(\cdot \cap [0, 1])$$



The support of the law μ_X is the Borel set $[0, 1]$.

It usually takes more effort to ascertain the law, but it is always determined by pushing forward the probability measure by the random variable. ▪

In practice, it is often the case that the probability space $(\Omega, \mathcal{F}, \mathbb{P})$ fades into the background. Instead, we may focus on the random variable X and its law μ_X without worrying too much about the underlying probability model. Furthermore, it is usually not important to have an exact description of the function $X : \Omega \to \mathbb{R}$, because we can pass to the distribution μ_X on the real line and work there instead.

8.3 Distribution functions

There is an alternative way to represent the law of a real random variable that can be more convenient in some circumstances. Instead of working with the law of the random variable, we can work with the real-valued function that tabulates the cumulative distribution of probability.

Definition 8.7 (Distribution function). Let X be a real random variable. Define the function

$$F_X(a) := \mathbb{P}\{X \le a\} = \mu_X(-\infty, a] \quad \text{for } a \in \mathbb{R}.$$

The function F_X is called the *(cumulative) distribution function* of the random variable, often abbreviated *cdf* or *df*.

Proposition 8.8 (Distribution function: Properties). The distribution function $F_X : \mathbb{R} \to [0, 1]$ of a real random variable X has the following properties.

1. **Monotonicity:** If $a \le b$, then $F_X(a) \le F_X(b)$.
2. **Asymptotes:** We have $\lim_{a \downarrow -\infty} F_X(a) = 0$ and $\lim_{a \uparrow +\infty} F_X(a) = 1$.
3. **Right continuity:** We have $\lim_{x \downarrow a} F_X(x) = F_X(a)$ for each $a \in \mathbb{R}$.

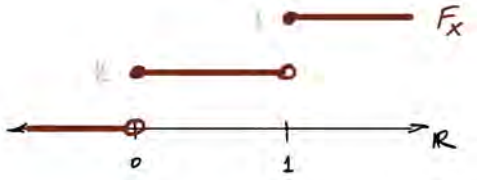
4. **Law:** For $a \leq b$, we have $\mu_X(a, b] = F_X(b) - F_X(a)$.

Exercise 8.9 (Distribution functions). Prove Proposition 8.8.

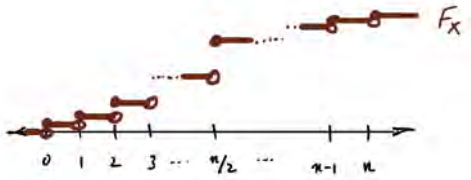
As we discussed in Lecture 3, every function that has properties (1)–(3) defines a unique Borel probability measure that satisfies (4). This claim requires a somewhat involved argument based on the Hahn–Kolmogorov theorem (Problem A.17).

Example 8.10 (Basic experiments). Here are the distribution functions associated with the random variables in our basic experiments.

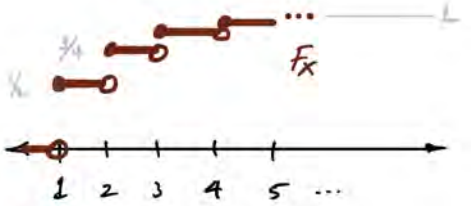
1. **One coin:** Here is an illustration of the distribution function F_X of the random variable X that indicates whether the coin comes up heads.



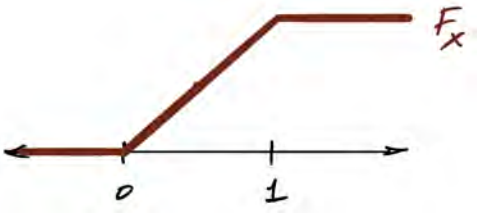
2. **Head count:** Here is an illustration of the distribution function F_X of the random variable X that counts the number of heads in n tosses.



3. **First heads:** Here is an illustration of the distribution function F_X of the random variable X that counts the number of flips before a coin turns up heads.



4. **Linear darts:** Here is an illustration of the distribution function F_X of the random variable X that describes where the dart strikes.



You can see that each of these distribution functions has the properties outlined in Proposition 8.8. ▪

In summary, the law provides a complete description of the distribution of a real random variable. The distribution function also provides a complete description of

the distribution of a real random variable. It is easy to write the distribution function in terms of the law. The distribution function also determines the law, but it only gives an explicit expression for the measure of a (half-open) interval. As such, you can always use the representation of the distribution that is most convenient for a particular problem.

8.4 Livestock

As we say, you cannot run a ranch without any cattle. In this section, we will describe the main breeds of random variables, and then we will introduce some specific animals from these varieties. You should be familiar with these examples from previous courses; they are listed here primarily for reference.

8.4.1 Flavors of random variables

The examples of random variables and distribution functions that we have seen suggest a taxonomy of basic real random variables.

1. **Indicator random variables:** A fundamentally important random variable is the indicator that an event occurs. Let $\mathsf{E} \in \mathcal{F}$ be an event. The associated indicator random variable is

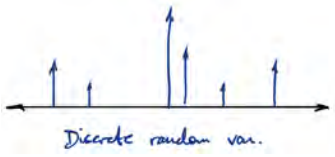
$$\mathbb{1}_{\mathsf{E}}(\omega) := \begin{cases} 1, & \omega \in \mathsf{E}; \\ 0, & \omega \notin \mathsf{E} \end{cases} \quad \text{for } \omega \in \Omega.$$

   An example of an indicator random variable is the indicator that a coin comes up heads. An indicator follows a $\textsc{bernoulli}(p)$ distribution where $p = \mathbb{P}(\mathsf{E})$.

2. **Discrete random variables:** A random variable X is *discrete* if its law μ_X can be written as a countable sum of Dirac point masses:

$$\mu_X = \sum_{i=1}^{\infty} p_i \delta_{a_i} \quad \text{where } a_i \in \mathbb{R} \text{ and } p_i \geq 0 \text{ and } \sum_{i=1}^{\infty} p_i = 1.$$

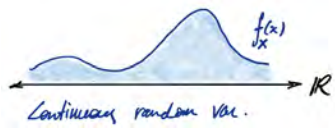
   For an example of a discrete random variable, consider the random variable that counts the number of flips before we see the first heads. Indicator random variables are also discrete.



3. **Absolutely continuous random variables:** A random variable X is *absolutely continuous* if its law μ_X has a density f_X with respect to the Lebesgue measure. That is,

$$\mu_X(\mathsf{B}) = \int_{\mathsf{B}} f_X(x)\, \lambda(\mathrm{d}x) \quad \text{for all } \mathsf{B} \in \mathcal{B}(\mathbb{R}), \tag{8.1}$$

   where $f_X : \mathbb{R} \to \mathbb{R}_+$ is a positive, measurable function with integral $\lambda(f_X) = 1$. For an example of an absolutely continuous random variable, consider the random variable that describes the position where the linear dart strikes the interval $[0, 1]$. It is common to refer simply to "continuous" random variables, even though this terminology is inaccurate.

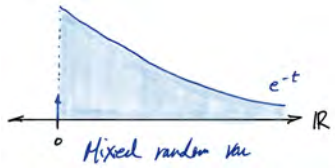


4. **Mixed random variables:** A random variable X is called *mixed* if it is a mixture of a discrete random variable and a continuous random variable. The law μ_X has the form

$$\mu_X = \alpha \mu_Y + (1 - \alpha)\mu_Z \quad \text{for } \alpha \in [0, 1],$$

   where μ_Y is the law of a discrete random variable and μ_Z is the law of a continuous random variable. For example, consider the random variable

$$\mu_X(\mathsf{B}) = 0.001\delta_0(\mathsf{B}) + 0.999 \int_{\mathsf{B}} \mathrm{e}^{-t} \mathbb{1}_{\mathbb{R}_+}(t)\, \lambda(\mathrm{d}t).$$



This random variable models an electronic component that, with probability 0.1%, burns out immediately when it is activated and otherwise has an exponential lifetime with mean one.

5. **Singular continuous:** Not all random variables have mixed distributions. A random variable X is said to be *singular continuous* if its distribution function F_X is continuous, but its law μ_X does *not* have a density with respect to the Lebesgue measure. For an example, consider the *Cantor distribution*. Singular continuous distributions also arise naturally in the study of particles undergoing Brownian motion. It is good to be aware that this type of random variable exists, but we will not discuss them again.

Let us emphasize that each of these random variables is modeled by its law, which is a Borel measure on the real line. In this sense, measure theory provides a unified view of discrete and continuous probability. Of course, when we make practical calculations with random variables, we may still rely on different methods for discrete and continuous examples.

8.4.2 Discrete random variables: Examples

We briefly introduce the main examples of discrete real random variables, with a few comments on their applications.

Example 8.11 (Bernoulli). Let $p \in [0,1]$. A real random variable $X \sim \text{BERNOULLI}(p)$ has the law

$$\mu_X = (1-p)\delta_0 + p\delta_1.$$

Bernoulli random variables model an experiment that succeeds with probability p. The coin flip experiment provides an example. Every indicator random variable has a Bernoulli distribution, so Bernoulli random variables arise when counting how many events occur.

It is also common to encounter a *signed Bernoulli* random variable Y, which has law $\mu_Y = (1-p)\delta_{-1} + p\delta_{+1}$. ∎

Example 8.12 (Discrete uniform). For a *finite* subset $\mathsf{S} \subseteq \mathbb{N}$, a random variable $X \sim \text{UNIFORM}(\mathsf{S})$ follows the law

$$\mu_X = \sum\nolimits_{k\in\mathsf{S}} \frac{1}{\#\mathsf{S}}\,\delta_k.$$

The uniform distribution models a situation where the random variable takes each value in the set S with equal probability. Think about rolling a die or drawing a card from a shuffled deck.

Among distributions supported on the finite set S, the distribution $\text{UNIFORM}(\mathsf{S})$ has the *maximum entropy*. ∎

The *entropy* of a distribution is a measure of how "random" it is; see Problem 8.33. Entropy arises in information theory and in statistical physics.

Example 8.13 (Binomial). Let $p \in [0,1]$ and $n \in \mathbb{N}$. A real random variable $X \sim \text{BINOMIAL}(n,p)$ has the law

$$\mu_X = \sum\nolimits_{k=0}^{n} p^k(1-p)^{n-k}\,\delta_k.$$

Binomial random variables model the total number of successes in a sequence of independent experiments, each with success probability p. The head count experiment provides an example. ∎

Example 8.14 (Geometric). Let $p \in (0,1)$. A random variable $X \sim \text{GEOMETRIC}(p)$ has the law

$$\mu_X = \sum_{k=1}^{\infty} p(1-p)^{k-1}\, \delta_k.$$

If we perform independent trials of an experiment with success probability p, the geometric random variable models the time at which the first success occurs. The first heads experiment provides an example.

Geometric random variables have the special property of being *memoryless*: the distribution of the waiting time for a success does not depend on how much time has already elapsed. Among distributions supported on $\mathbb{N}$ with mean $m > 0$, the $\text{GEOMETRIC}(1/m)$ distribution has the maximum entropy. ▪

A proper definition of the term "memoryless" requires the concept of conditioning. See Exercise 20.27.

Example 8.15 (Poisson). A Poisson random variable $X \sim \text{POISSON}(\beta)$ with mean $\beta > 0$ has the law

$$\mu_X = \sum_{k=0}^{\infty} \frac{\beta^k \mathrm{e}^{-\beta}}{k!}\, \delta_k.$$

A Poisson random variable models rare events. It describes the number of successes in a sequence of independent trials with success probability $p = \beta/n$ as the number n of trials increases.

Poisson random variables have the lovely *stability* property that a sum of independent Poisson random variables remains Poisson.

We will discuss methods for establishing stability in Lecture 21.

I cannot resist sharing some classic examples where the Poisson distribution arises. It has been used to describe the number of misprints on a page of the newspaper, the number of Prussian cavalry officers kicked to death by their horses in a given year, and the number of bombs that fell on a given district in London during the Blitz. ▪

8.4.3 Absolutely continuous random variables: Examples

Now, we turn to some of the main examples of continuous real random variables. Recall that a continuous variable is determined by its density (with respect to the Lebesgue measure); see the definition in (8.1).

Example 8.16 (Uniform). Let $\mathsf{S} \in \mathscr{B}(\mathbb{R})$ be a *bounded* Borel set. A real random variable $X \sim \text{UNIFORM}(\mathsf{S})$ has the density

$$f_X(x) = \frac{1}{\lambda(\mathsf{S})} \mathbb{1}_{\mathsf{S}}(x) \quad \text{for } x \in \mathbb{R}.$$

Uniform random variables model the situation when a value is equally likely to be anywhere in the set S. Linear darts provides an example.

Among continuous distributions supported on a bounded interval, say $[a, b]$, the $\text{UNIFORM}([a,b])$ distribution has the maximum entropy. ▪

See Problem 8.34.

Example 8.17 (Exponential and Laplace). A random variable $X \sim \text{EXPONENTIAL}(\beta)$ with rate $\beta > 0$ has the density

$$f_X(x) = \beta \mathrm{e}^{-\beta x}\, \mathbb{1}_{\mathbb{R}_+}(x) \quad \text{for } x \in \mathbb{R}.$$

The exponential distribution arises in queueing theory and other problems involving continuous waiting times. Exponential distributions have the elegant property of being *memoryless*: the distribution of the waiting time does not depend on how much time has already elapsed. Among continuous distributions supported on $\mathbb{R}_+$ with mean m, the $\text{EXPONENTIAL}(1/m)$ distribution has the maximum entropy.

See Exercise 20.27.

A random variable $Y \sim \text{LAPLACE}(\beta)$ with rate $\beta > 0$ has the density

$$f_Y(y) = \frac{\beta}{2} e^{-\beta|y|} \quad \text{for } y \in \mathbb{R}.$$

Laplace random variables can arise from Bayesian regression models, and they now play a role in the theory of differential privacy. ▪

Example 8.18 (Normal). A real random variable $X \sim \text{NORMAL}(m, \sigma^2)$ with mean $m \in \mathbb{R}$ and variance $\sigma^2 > 0$ has the density

$$f_X(x) = \frac{e^{-(x-m)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}} \quad \text{for } x \in \mathbb{R}.$$

A normal random variable is also called a *Gaussian* random variable. When $m = 0$ and $\sigma^2 = 1$, we refer to this distribution as the *standard normal* distribution.

The normal distribution is the single most important continuous distribution because of its role in the central limit theorem (Lecture 18). Normal random variables have the remarkable *stability* property that an independent sum of normal random variables remains normal. Among continuous distributions on $\mathbb{R}$ with mean m and variance σ^2, the distribution $\text{NORMAL}(m, \sigma^2)$ has the maximum entropy. ▪

Example 8.19 (Gamma and Beta). A random variable $X \sim \text{GAMMA}(\alpha, \beta)$ with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$ has the density

$$f_X(x) = \frac{x^{\alpha-1} e^{-\beta x} \beta^\alpha}{\Gamma(\alpha)} \mathbb{1}_{\mathbb{R}_+}(x) \quad \text{for } x \in \mathbb{R}.$$

As usual, Γ denotes Euler's gamma function, a kind of generalized factorial.

Gamma distributions can arise from sums of exponential random variables and from sums of squared Gaussian random variables, and they play a role in Bayesian inference.

A random variable $Y \sim \text{BETA}(\alpha, \beta)$ with shape parameters $\alpha, \beta > 0$ has the density

$$f_Y(y) = \frac{y^{\alpha-1}(1-y)^{\beta-1}}{\text{B}(\alpha, \beta)} \mathbb{1}_{(0,1)}(y) \quad \text{for } y \in \mathbb{R}.$$

As usual, B denotes the beta function, a kind of generalized binomial coefficient.

Beta distributions arise from geometric problems involving volumes of sections of Euclidean balls, and they also play a role in Bayesian inference. ▪

Example 8.20 (Cauchy). A random variable $X \sim \text{CAUCHY}(m, \gamma)$ with location $m \in \mathbb{R}$ and scale $\gamma > 0$ has the density

$$f_X(x) = \frac{1}{\pi\gamma} \cdot \frac{1}{1 + ((x-m)/\gamma)^2} \quad \text{for } x \in \mathbb{R}.$$

Cauchy random variables have very heavy tails. In fact, they do not even have a defined expectation, so they are an extreme example that is useful to keep in mind. They have the remarkable *stability* property that an independent sum of Cauchy random variables remains Cauchy. ▪

8.5 *Joint distributions

It is common to encounter multiple random variables at once, so we need a framework for studying them. For example, in the square darts example, we throw a dart at a square target $[0, 1]^2$ and record the position (X, Y) where it hits. Both the horizontal coordinate X and the vertical coordinate Y are real random variables. How are they

related? What is the analog of the law of a single random variable? How can we characterize the joint distribution?

In this section, we will show how to use measures to describe the distribution of a pair of random variables. The extension to more random variables is straightforward, at least in concept.

8.5.1 Pairs of random variables

We begin with the definition of a pair of random variables and some measurability properties.

Definition 8.21 (Pair of random variables). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Consider two random variables $X : \Omega \to \mathbb{R}$ and $Y : \Omega \to \mathbb{R}$. Then (X, Y) is called a *pair of real random variables*.

Although we only assumed that the individual random variables are real-valued measurable functions on the sample space, the pair is also a measurable function taking values in the plane.

Exercise 8.22 (*Pair of random variables: Measurability). Consider two functions $X, Y : \Omega \to \mathbb{R}$. Show that $(X, Y) : \Omega \to \mathbb{R}^2$ is a measurable function if and only if X and Y are both random variables. For the pair, measurability means that

$$\{(X, Y) \in \mathsf{B}\} := \{\omega \in \Omega : (X(\omega), Y(\omega)) \in \mathsf{B}\} \in \mathcal{F} \quad \text{for all } \mathsf{B} \in \mathcal{B}(\mathbb{R}^2).$$

That is, the preimage of a Borel set in the plane is an event. **Hint:** One direction depends on the fact that the coordinate projections are measurable. The other direction uses the fact that the σ-algebra $\mathcal{B}(\mathbb{R}^2)$ is generated by measurable rectangles (Proposition 6.7); see Proposition 4.2 for the pattern of argument.

8.5.2 Joint and marginal laws

Since (X, Y) is a measurable function, we are licensed to compute the probability of any Borel set in $\mathcal{B}(\mathbb{R}^2)$. This leads to the notion of a joint distribution.

Definition 8.23 (Pair of random variables: Joint law and marginal laws). Let (X, Y) be a pair of real random variables. The *joint law* is the Borel probability measure on $\mathcal{B}(\mathbb{R}^2)$ defined by

$$\mu_{XY}(\mathsf{B}) := \mathbb{P}\{(X, Y) \in \mathsf{B}\} \quad \text{for all Borel } \mathsf{B} \in \mathcal{B}(\mathbb{R}^2).$$

The law of the pair is also called the *joint distribution*. In this context, the distributions of the individual random variables are called the *marginal laws*:

$$\mu_X(\mathsf{B}) := \mathbb{P}\{X \in \mathsf{B}\} \quad \text{and} \quad \mu_Y(\mathsf{B}) := \mathbb{P}\{Y \in \mathsf{B}\} \quad \text{for each } \mathsf{B} \in \mathcal{B}(\mathbb{R}).$$

The marginal distributions are defined without reference to each other.

Example 8.24 (Square darts). Recall the setting for the square darts example. We equip the sample space $\Omega = [0, 1]^2$ with its Borel σ-algebra $\mathcal{F} = \mathcal{B}([0, 1]^2)$ and the Lebesgue measure λ^2 restricted to $[0, 1]^2$. The distinguished sample point $\boldsymbol{\omega}_0 \in \Omega$ describes the location where the dart strikes.

In this context, the most natural random variables are the coordinate projections,

which describe the horizontal and vertical position of the dart:

$$\begin{aligned} X(\boldsymbol{\omega}) &:= \pi_1(\boldsymbol{\omega}) = \omega_1; \\ Y(\boldsymbol{\omega}) &:= \pi_2(\boldsymbol{\omega}) = \omega_2 \end{aligned} \quad \text{for } \boldsymbol{\omega} = (\omega_1, \omega_2) \in [0,1]^2.$$

Since (X, Y) is the identity map on $[0,1]^2$, the joint law is simply the Lebesgue measure on the unit square:

$$\mu_{XY}(\mathsf{B}) = \lambda^2(\mathsf{B}) \quad \text{for all } \mathsf{B} \in \mathscr{B}([0,1]^2).$$

By elementary geometric reasoning, we can see that the marginal laws are the Lebesgue measures on the unit intervals:

$$\begin{aligned} \mu_X(\mathsf{B}) &= \lambda(\mathsf{B}); \\ \mu_Y(\mathsf{B}) &= \lambda(\mathsf{B}) \end{aligned} \quad \text{for all } \mathsf{B} \in \mathscr{B}([0,1]).$$

For instance, the probability that $X \in \mathsf{B}$ is the probability that $(X, Y) \in \mathsf{B} \times [0,1]$, which we can compute in terms of areas.

We can consider other pairs of random variables in this setting. For illustration, consider the pair that repeats the horizontal location of the dart twice:

$$(W(\boldsymbol{\omega}), Z(\boldsymbol{\omega})) := (\pi_1(\boldsymbol{\omega}), \pi_1(\boldsymbol{\omega})) \quad \text{for } \boldsymbol{\omega} = (\omega_1, \omega_2) \in [0,1]^2.$$

The joint law is a "diagonal" measure:

$$\mu_{WZ}(\mathsf{B}) = \lambda\{x \in \mathbb{R} : (x, x) \in \mathsf{B}\} \quad \text{for } \mathsf{B} \in \mathscr{B}([0,1]^2).$$

Once again, the marginal laws are both Lebesgue measure on the unit interval:

$$\mu_W(\mathsf{B}) = \lambda(\mathsf{B}) \quad \text{and} \quad \mu_Z(\mathsf{B}) = \lambda(\mathsf{B}) \quad \text{for all } \mathsf{B} \in \mathscr{B}([0,1]).$$

This discussion warns us that the marginal laws do not determine the joint law. ▪

8.5.3 Independence

As we saw in Example 8.24, the marginal laws are not enough to determine the joint law. Nevertheless, there is a special case that merits attention.

Definition 8.25 (Pair of random variables: Independence). Let (X, Y) be a pair of real random variables. We say that X and Y are *independent* if and only if the joint law is the product of the marginal laws:

$$\mu_{XY} = \mu_X \times \mu_Y.$$

In terms of the probability measure, independence means that

$$\mathbb{P}\{(X, Y) \in \mathsf{A} \times \mathsf{B}\} = \mathbb{P}\{X \in \mathsf{A}\} \cdot \mathbb{P}\{Y \in \mathsf{B}\} \quad \text{for } \mathsf{A}, \mathsf{B} \in \mathscr{B}(\mathbb{R}).$$

Example 8.26 (Square darts). In the square darts example, the horizontal position X and the vertical position Y of the dart are independent random variables. ▪

We will have much more to say about independence later (Lecture 13).

8.5.4 Probability and integral

As with a single random variable, the joint law is simply the push-forward of the probability measure by a function. As always, the measure of a set is the integral of the indicator. Thus, for each plane Borel set $\mathsf{B} \in \mathscr{B}(\mathbb{R}^2)$,

$$\begin{aligned} \mathbb{P}\{(X,Y) \in \mathsf{B}\} &= \mu_{XY}(\mathsf{B}) \\ &= \int_{\mathbb{R}^2} \mathbb{1}_{\mathsf{B}}(x,y)\, \mu_{XY}(\mathrm{d}x \times \mathrm{d}y) = \int_{\mathbb{R}^2} \mathbb{1}_{\mathsf{B}}\, \mathrm{d}\mu_{XY} = \int_{\mathsf{B}} \mathrm{d}\mu_{XY}. \end{aligned} \tag{8.2}$$

All the notations mean the same thing. In case the random variables X and Y are independent, Fubini–Tonelli (Theorem 6.23) helps us compute these integrals.

We can connect the marginal laws with the joint law by integrating over a cylinder. For example,

$$\mu_X(\mathsf{B}) = \int_{\mathsf{B}\times\mathbb{R}} \mathrm{d}\mu_{XY} \quad \text{for a real Borel set } \mathsf{B} \in \mathscr{B}(\mathbb{R}). \tag{8.3}$$

This type of formula leads to concrete tools for working with joint distributions.

8.5.5 Specifying the joint distribution

A measure on the plane is a complicated thing. So we may ask whether there are alternative mechanisms for specifying a joint distribution. In fact, there is a natural analog of the distribution function.

Definition 8.27 (Joint distribution function). Let (X,Y) be a pair of real random variables. The *joint distribution function* $F_{XY} : \mathbb{R}^2 \to \mathbb{R}_+$ is defined as

$$\begin{aligned} F_{XY}(a,b) &:= \mathbb{P}\{X \le a \text{ and } Y \le b\} \\ &= \mu_{XY}((-\infty, a] \times (-\infty, b]) \quad \text{for } a, b \in \mathbb{R}. \end{aligned}$$

Theorem 8.28 (Joint distribution function). The joint law μ_{XY} determines the joint distribution function F_{XY}. The joint distribution function F_{XY} is increasing and right-continuous in each variable separately. As $a, b \downarrow -\infty$, the limiting value of $F_{XY}(a,b)$ is zero. As $a, b \uparrow +\infty$, the limiting value is one.

Conversely, any function F with these properties determines a unique Borel probability measure μ on $\mathbb{R}^2$ with distribution function F.

**Proof.* We omit the proof. It follows from the Hahn–Kolmogorov theorem (Theorem A.12) in the same fashion as the construction of measures from distribution functions on the line (Problem A.17). In this setting, we start with the algebra generated by half-open geometric rectangles $(a,b] \times (c,d]$ because we can compute its measure easily from F_{XY}. ∎

Problems

Exercise 8.29 (Sums of measures). Measures are just positive functions on a σ-algebra, so we can form linear combinations. Suppose that μ, ν are Borel probability measures on $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$.

1. For $\alpha \in [0,1]$, show that $\alpha\mu + (1-\alpha)\nu$ is a Borel probability measure.

2. Show that the integral is positive-linear in the measure. For $\alpha, \beta \geq 0$,

$$\int_{\mathbb{R}} h(x)\,(\alpha\mu + \beta\nu)(\mathrm{d}x) = \alpha \int_{\mathbb{R}} h(x)\,\mu(\mathrm{d}x) + \beta \int_{\mathbb{R}} h(x)\,\nu(\mathrm{d}x).$$

What assumptions are required on the function $h : \mathbb{R} \to \mathbb{R}$?

Exercise 8.30 (Algebras generated by random variables). Let X be a real random variable on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. The σ-algebra generated by the random variable X is

$$\sigma(X) := \sigma\{X^{-1}(\mathsf{B}) : \mathsf{B} \in \mathcal{B}(\mathbb{R})\}.$$

We can decide whether each event in $\sigma(X)$ occurs, given only the value of $X(\omega)$. In other words, $\sigma(X)$ reflects the knowledge we gain about the sample point by observing the value of the random variable X.

1. Suppose that we flip a fair coin twice. What is the natural probability space?
2. Define a random variable $X = 1$ if the first coin comes up heads and $X = 0$ if the first coin comes up tails. What are the events in $\sigma(X)$? What are their probabilities?
3. Define a random variable $Y = 1$ if both coins show the same face and $Y = 0$ if the coins show different faces. What are the events in $\sigma(Y)$? What are their probabilities?
4. How do the observations in the last two parts support the interpretation in the problem statement?

Exercise 8.31 (Median). Let X be a real random variable on a probability space. A *median* of the random variable is a number $M \in \mathbb{R}$ with the property that

$$\mathbb{P}\{X \leq M\} \geq \tfrac{1}{2} \quad \text{and} \quad \mathbb{P}\{X \geq M\} \geq \tfrac{1}{2}.$$

As a warning, there could be more than one choice of M that satisfies this definition.

1. Compute the median value of a $\textsc{uniform}[0, 1]$ random variable.
2. For each $p \in [0, 1]$, compute a median value of a $\textsc{bernoulli}(p)$ random variable.
3. Prove that every real random variable has at least one median.
4. Show that the set of all medians of a random variable X composes an interval.

Problem 8.32 (*Skorokhod). Let X be a real random variable on an arbitrary probability space with (cumulative) distribution function F_X. We can also realize X as a random variable on the "universal" probability space $\mathsf{U} := ([0, 1], \mathcal{B}([0, 1]), \lambda)$, where λ is the Lebesgue measure restricted to $[0, 1]$. To do so, we define an (extended) real random variable on U by the formula

$$S(u) := \begin{cases} \inf\{a \in \mathbb{R} : F_X(a) \geq u\}, & 0 < u < 1; \\ -\infty, & u = 0; \\ +\infty, & u = 1. \end{cases}$$

This is essentially the functional inverse of the cdf F_X. We call S a *Skorokhod representation* of the random variable X.

1. Why is S a random variable? **Hint:** In the infimum, we can replace $\mathbb{R}$ with $\mathbb{Q}$.

2. Show that the distribution function F_S coincides with F_X. **Hint:** Do the endpoints matter? Note that the infimum is attained for $u \in (0,1)$, and S is an increasing function on $[0,1]$.
3. Let $U \sim \text{UNIFORM}[0,1]$. Explain why $S(U)$ has the same distribution as X.

Problem 8.33 (*Maximum entropy: Discrete case). Let X be a discrete, real random variable with law

$$\mu_X = \sum_{i \in \mathbb{Z}} p_i \delta_{a_i} \quad \text{for } a_i \in \mathbb{R} \text{ and } p_i \geq 0 \text{ and } \sum_{i \in \mathbb{Z}} p_i = 1.$$

The *entropy* of the discrete variable X is defined as

$$\text{entropy}(X) := -\sum_{i \in \mathbb{Z}} p_i \log p_i.$$

Throughout this course, the logarithm is the natural logarithm. We enforce the convention that $0 \log 0 = 0$.

Note that the entropy does not depend on the support, just the probabilities. As we will see, the entropy is a measure of the amount of "randomness" in the distribution.

1. Check that the entropy of a discrete random variable is a positive number.
2. **(*)** Show that $\boldsymbol{p} \mapsto -\sum_{i \in \mathbb{Z}} p_i \log p_i$ is a concave function for $\boldsymbol{p} \in \mathbb{R}^n_+$.
3. Compute the entropy of $X \sim \text{BERNOULLI}(p)$ for $p \in [0,1]$. For what choice of p is the entropy maximized? Minimized? Given an interpretation of these results.
4. Compute the entropy of $X \sim \text{UNIFORM}\{0, 1, 2, \ldots, n\}$.
5. Show that the uniform distribution is the maximum entropy distribution supported on $\{0, 1, 2, \ldots, n\}$. **Hint:** Reparameterize $p_i = \mathrm{e}^{\alpha_i}$ for $\alpha_i \in \mathbb{R}$, and use Lagrange multipliers.
6. Compute the entropy of $X \sim \text{GEOMETRIC}(p)$ for $p \in [0,1]$ in closed form. For what choice of p is the entropy maximized? Minimized?
7. Among those distributions supported on $\mathbb{N}$ that have mean $m > 0$, show that $\text{GEOMETRIC}(1/m)$ is the maximum entropy distribution.

Recall that the mean m of a distribution on $\mathbb{Z}_+$ is defined as

$$m := \sum_{i \in \mathbb{Z}_+} i \, p_i.$$

Problem 8.34 (*Maximum entropy: Continuous case). Let X be a continuous, real random variable with density $f_X : \mathbb{R} \to \mathbb{R}_+$. The *entropy* of the continuous variable X is defined as

$$\text{entropy}(X) := -\int_{\mathbb{R}} (f_X \log f_X) \, \mathrm{d}\lambda.$$

As before, the entropy does not depend on the support, just the density. As we will see, the entropy is a measure of the "randomness" in the distribution.

1. Compute the entropy of $X \sim \text{UNIFORM}([0,a])$ for $a > 0$.
2. Show by example that the entropy of a continuous random variable can be a negative number.
3. **(*)** Show that the uniform distribution is the maximum entropy distribution supported on $\{0, 1, 2, \ldots, n\}$. **Hint:** Reparameterize $f_X = \mathrm{e}^g$ for measurable $g : \mathbb{R} \to \mathbb{R}$, and use the Euler–Lagrange equations.
4. Compute the entropy of $X \sim \text{EXPONENTIAL}(\beta)$ for $\beta > 0$.
5. **(*)** Among continuous distributions supported on $\mathbb{R}_+$ with mean $m > 0$, show that the $\text{EXPONENTIAL}(1/m)$ distribution has the maximum entropy.
6. Compute the entropy of $X \sim \text{NORMAL}(m, \sigma^2)$ for $m \in \mathbb{R}$ and $\sigma^2 > 0$.
7. **(*)** Among continuous distributions supported on $\mathbb{R}$ with mean zero and variance one, show that the standard normal distribution, $\text{NORMAL}(0,1)$, is the maximum entropy distribution.

Recall that the mean m and variance σ^2 of a continuous distribution are defined as

$$m := \int_{\mathbb{R}} x \, f_X(x) \lambda(\mathrm{d}x);$$
$$\sigma^2 := \int_{\mathbb{R}} (x-m)^2 \, f_X(x) \lambda(\mathrm{d}x).$$

> **Aside:** It is possible to define entropy-like quantities for more general classes of real random variables. In particular, the information divergence between two random variables is defined whenever one distribution has a density with respect to the other (Exercise 6.27). We omit this development because it is outside the scope of our course.

Applications

A basic challenge in computational mathematics is to generate random variables that have a specified distribution. While a full discussion of this topic falls outside of the scope of this course, the material in this lecture gives us the tools we need to explore some of the basic methodologies for generating random variables.

Application 8.35 (Quantile sampling). Most programming languages have in-built functionality for generating uniform random variables: $U \sim \text{UNIFORM}[0, 1]$. In this problem, we will investigate how to use this source of randomness to generate other types of random variables.

According to Problem 8.32, if X is a real random variable with cumulative distribution function F_X, then the random variable $S(U)$ has the same distribution as X. Generating other types of (non-uniform) random variables thus reduces to finding a formula for the function S. This approach is called *quantile sampling*, and it is the preferred methodology when it can be implemented.

1. Find a formula for S for $X \sim \text{UNIFORM}([c, d])$ when $c < d$ and $c, d \in \mathbb{R}$.
2. Find S for a random variable $X \sim \text{BERNOULLI}(p)$ with success rate $p \in [0, 1]$.
3. A Cauchy random variable X is a continuous real random variable with density f_X and distribution function F_X:

$$f_X(x) = \frac{1}{\pi} \cdot \frac{1}{1 + x^2} \quad \text{and} \quad F_X(a) = \frac{1}{\pi} \arctan(a) \quad \text{for } x, a \in \mathbb{R}.$$

   Find S for a Cauchy random variable.
4. Find S for a Laplace random variable X, which has density $f_X(x) = \mathrm{e}^{-|x|}/2$ for $x \in \mathbb{R}$.
5. **(*)** Use quantile sampling to generate 1000 realizations of each one of these random variables, and plot a histogram of the values.

Application 8.36 (Getting used to rejection). Another standard methodology for generating random variables is the technique of rejection sampling. In this case, suppose that we would like to draw a sample from the distribution of a continuous, real random variable X with density f_Y and law μ_X.

Instead, we have access to samples from the distribution of another continuous, real random variable Y with density g_Y and law μ_Y. We require that the likelihood ratio satisfies a uniform bound. That is, for a constant $M > 0$,

$$\frac{f_X(a)}{g_Y(a)} \leq M \quad \text{for all } a \in \mathbb{R}.$$

For simplicity, you may assume that the proposal density is strictly positive: $g_Y > 0$.

Rejection sampling proceeds as follows. We independently draw a random variable Y and a uniform random variable $U \sim \text{UNIFORM}[0, 1]$ with law μ_U. Therefore, (Y, U) has the joint law $\mu_{YU} = \mu_Y \times \mu_U$. Define the acceptance event

$$\mathsf{E} := \{U \cdot M g_Y(Y) < f_X(Y)\}.$$

If the event E occurs, then we report the value Y of the proposal. Otherwise, we reject the sample from Y and start over. We will show that, in case of acceptance, the reported sample value Y follows the same distribution as the target random variable X.

1. Explain why the probability of the acceptance event E can be computed by evaluating the measure of a plane region under the joint law:
$$\begin{aligned}\mathbb{P}(\mathsf{E}) &= \mathbb{P}\{U \cdot M g_Y(Y) < f_X(Y)\} \\ &= \mu_{YU}\big\{(y,u) \in \mathbb{R}^2 : u \cdot M g_Y(y) < f_X(y)\big\}.\end{aligned}$$
Draw a sketch of this plane region.
2. Prove that $\mathbb{P}(\mathsf{E}) = 1/M$. **Hint:** Write the measure of the plane region as an integral, and use Fubini–Tonelli. Proposition 9.5 is also relevant.
3. Elementary notions of conditional probability yield a formula for the distribution of an accepted sample. For each Borel set $\mathsf{B} \in \mathscr{B}(\mathbb{R})$, prove that
$$\mathbb{P}\{Y \in \mathsf{B} \mid \mathsf{E}\} := \frac{\mathbb{P}\{Y \in \mathsf{B} \text{ and } (Y,U) \in \mathsf{E}\}}{\mathbb{P}(\mathsf{E})} = \mathbb{P}\{X \in \mathsf{B}\}.$$
In other words, an accepted sample has the same law as the target random variable. **Hint:** The pattern of argument is very similar to the computation of the acceptance probability $\mathbb{P}(\mathsf{E})$.
4. What is the probability that it takes exactly k repetitions of the rejection sampling procedure before we accept a sample? What is the expected number of repetitions required to accept a sample? What kind of random variable models this situation?
5. Suppose we wish to generate a standard normal variable X using the random variable $Y \sim \text{CAUCHY}(0,1)$ as the proposal. Compute the maximum value of the likelihood ratio. On average, what is the expected number of repetitions required to accept a sample?

The normal law is described in Example 8.18, and the Cauchy law is described in Example 8.20.

6. Suppose that we wish to generate a standard normal variable X using the random variable $Y \sim \text{LAPLACE}(1)$ as the proposal. Compute the maximum value of the likelihood ratio. On average, what is the expected number of repetitions required to accept a sample? Which proposal is better, Cauchy or Laplace?

The Laplace law is described in Example 8.17.

7. **(*)** Describe two algorithms that use uniform random variables to generate standard normal variables by combining quantile sampling (Application 8.35) and rejection sampling. Implement your algorithms. Use each one to draw 1000 realizations of a standard normal variable, and plot a histogram. In each case, how many uniform random variables did it take to complete the experiment?

Application 8.37 (*Box–Muller). Recall that a standard normal random variable Z variable is a continuous distribution with density
$$\varphi_Z(z) := \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-z^2/2} \quad \text{for } z \in \mathbb{R}.$$
As an illustration of the rejection sampling methodology (Application 8.36), we showed that it is possible to generate normal distributions from Cauchy or Laplace random variables. These methods are both a bit wasteful because we have to reject some of the samples. In this problem, we will describe a methodology for using two independent uniform random variables to generate two independent standard normal random variables. This is called the *Box–Muller transform*.

1. A pair (X, Y) of independent standard normal random variables has the *joint* probability density function

$$f_{XY}(x, y) := \frac{1}{2\pi} e^{-(x^2+y^2)/2} \quad \text{for } (x, y) \in \mathbb{R}^2.$$

   For the nonzero values of (X, Y), we can transform to polar coordinates $(R, \Theta) \in \mathbb{R}_{++} \times [0, 2\pi)$ by the rule $(X, Y) = (R\cos\Theta, R\sin\Theta)$. Find the joint law of (R, Θ) using the nonlinear transformation rule (Problem 6.29). Don't forget the Jacobian!
2. Changing variables again, calculate the joint probability density of the pair (U, Θ) where $U = \frac{1}{2}R^2$. What is the marginal distribution of U? The marginal distribution of Θ? Are they independent?
3. Consider an independent pair (U_1, U_2) of uniform random variables: $U_i \sim \text{UNIFORM}([0, 1])$ for $i = 1, 2$. Explain how to transform (U_1, U_2) to obtain a pair (X, Y) of independent standard normal variables.

Application 8.38 (Generative modeling). By pushing a random variable forward through a function, we can obtain very complicated distributions. In particular, we have seen that we can transform $U \sim \text{UNIFORM}([0, 1])$ to any real probability distribution on the line if we know its quantile function S; see Application 8.35.

The idea behind *generative modeling* is to use data to learn a function h that transforms a simple random variable (e.g., U) to a sample from the (approximate) empirical distribution of the data. In this problem, we will explore a rudimentary version of this idea. We will see how a simple artificial neural network can represent a histogram distribution, but we will not consider the problem of training the network to learn the histogram.

We say that a random variable X has a (finite) *histogram distribution* if it is a continuous random variable with a piecewise constant density f_X with a finite number of pieces. That is, there is an increasing real sequence $a_1 < a_2 < \cdots < a_{n+1}$ and a positive real sequence $c_i \geq 0$ with $\sum_{i=1}^n c_i = 1$ for which

$$f_X(x) := \frac{c_i}{a_{i+1} - a_i} \quad \text{for } x \in (a_i, a_{i+1}] \text{ and } i = 0, 1, 2, \ldots, n.$$

We set $f_X(x) = 0$ for $x \leq a_1$ and $x > a_{n+1}$, so the density has compact support. Each interval $(a_i, a_{i+1}]$ is called a *bin*, and we allow the bins to have different widths. The height c_i reflects the frequency with which we see items in the ith bin, relative to the width of the bin.

An *artificial neural network* is simply a composition of structured functions, called *layers*. In a basic feed-forward neural network, it is common that each layer is the composition of an affine function with a nonlinearity. In this problem, we consider a two-layer ReLU neural network with a single real input and a single real output, which can be written compactly as

An *affine function* is the composition of a linear map and a translation.

$$g(x) := \alpha + \sum_{i=1}^k \beta_i(\gamma_i(x - m_i))_+. \tag{8.4}$$

In this expression, $\alpha, \beta_i, \gamma_i, m_i \in \mathbb{R}$ for $i = 1, \ldots, k$.

1. Sketch an example of a histogram distribution.
2. Find the distribution function F_X of a histogram random variable. Note that F_X is piecewise affine, increasing, and continuous.

3. Compute the Skorokhod representation S of the the histogram random variable; see Problem 8.32. Note that S is piecewise affine and increasing. Moreover, when the frequencies c_i are *strictly* positive, the function S is continuous.
4. **(*)** Argue the following converse: If $h : [0,1] \to \mathbb{R}$ is a piecewise affine function with finitely many pieces, then $h(U)$ follows a finite histogram distribution. (This is true regardless of whether h is increasing or continuous.)
5. Show that (8.4) describes a continuous, piecewise affine function on $\mathbb{R}$ with a finite number of pieces. How many?
6. **(*)** Show that every *continuous*, piecewise affine function on $\mathbb{R}$ with a finite number of pieces can be written in the form (8.4) for some $k \in \mathbb{N}$. What is the minimal value of k possible? **Hint:** Observe that we can represent each of the following *hinge functions* using a single term from the sum:

$$q_1(x) := \begin{cases} 0, & x \le a_i; \\ r(x - a_i), & x > a_i; \end{cases} \quad \text{and} \quad q_2(x) := \begin{cases} r(x - a_i), & x \le a_i; \\ 0, & x > a_i. \end{cases}$$

   In this expression, $r \in \mathbb{R}$ and the a_i are the edges of the bins in the histogram.
7. Conclude that we can represent each histogram distribution X with *strictly positive frequencies* by passing a uniform variable U through an appropriate neural network of the form (8.4).
8. **(*)** Conclude that we can *approximate* any histogram distribution X arbitrarily well by passing a uniform variable U through a neural network (8.4). Can we bound the coefficients β_i and γ_i?

Notes

All of the material in this lecture is standard, and you will find similar presentations in any book on probability theory. See Cover & Thomas [CT06] for an introduction to information theory and entropy. The problems on quantile sampling, rejection sampling, and the Box–Muller transform were adapted from Owen's manuscript [Owe13] by Rob Webber and Ethan Epperly. I learned about the generative modeling idea from Helmut Bölcskei. For the result that two-layer ReLU networks can represent an arbitrary piecewise affine function, see the paper [Aro+18], for example.

Lecture bibliography

[Aro+18] R. Arora et al. "Understanding Deep Neural Networks with Rectified Linear Units". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL: https://openreview.net/forum?id=B1J_rgWRW.

[CT06] T. M. Cover and J. A. Thomas. *Elements of information theory*. Second. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2006.

[Owe13] A. B. Owen. "Monte Carlo theory, methods and examples". Available online. 2013. URL: https://artowen.su.domains/mc/.

9. Expectation & Jensen's Inequality

Agenda:
1. Expectation
2. Convex functions
3. Jensen's inequality
4. Beyond the real line

"Take nothing on its looks; take everything on evidence. There's no better rule."

—*Great Expectations*, Charles Dickens

In the last lecture, we introduced the concept of a real random variable, which we can think of as a single numerical observation of a probabilistic system. We will review these ideas again at the beginning of this lecture.

In this lecture, we turn to another important question: What is the average value of a real random variable? The average sums the values of the random variable, weighted by the probability that it takes on a particular value. As a consequence, defining this average properly requires an integral.

In probability theory, we refer to the *expectation* of a real random variable, rather than the integral of the random variable. We will introduce the concept of expectation for real random variables, and we will explore its properties. For the most parts, facts about the expectation are simply translations of analogous facts about the integral.

Since the expectation is a type of weighted *average*, it enjoys some extra features that a general integral does not. In pursuit of these results, we will define convex functions and develop some of their basic properties. Then we will present Jensen's inequality, which describes how expectation interacts with a convex function.

9.1 Recap

A probability space $(\Omega, \mathcal{F}, \mathbb{P})$ is a triple, consisting of a sample space Ω, a σ-algebra $\mathcal{F}$ of events, and a probability measure $\mathbb{P}$ defined on these events. It is fruitful to think about a probability space as a model for a very complex system that has unpredictable behavior. The sample points $\omega \in \Omega$ capture all the possible states of the system, and the probability measure $\mathbb{P}$ describes which sets of states are more or less likely.

A real random variable $X : \Omega \to \mathbb{R}$ is a real-valued measurable function on the sample space. We can think about X as a real-valued observable of the complex system. Since X is a function, the value $X(\omega)$ of a random variable is determined by the sample point ω. Since we think about the sample point ω as random, we also think about the value $X(\omega)$ as being random. When you see the symbol X, you should imagine a randomly distributed real number.

To describe the distribution of a real random variable X, we introduced the law $\mu_X : \mathcal{F} \to [0, 1]$ of the random variable. The law is a Borel probability measure on the real line that indicates what sets of real values are more or less likely:

$$\mu_X(\mathsf{B}) := \mathbb{P}\{X \in \mathsf{B}\} \quad \text{for each Borel set } \mathsf{B} \in \mathcal{B}(\mathbb{R}) \text{ in the real line.}$$

Once we know the law μ_X of a random variable, we can make probability calculations involving the (individual) random variable X without reference back to the original probability space.

The distribution function $F_X : \mathbb{R} \to [0, 1]$ provides an alternative representation of the law of the random variable. It is defined as

$$F_X(a) := \mathbb{P}\{X \le a\} = \mu_X(-\infty, a] \quad \text{for each } a \in \mathbb{R}.$$

Like the law, the distribution function also contains a complete description of the distribution of the random variable.

In case the random variable X is continuous, it also has a density $f_X : \mathbb{R} \to \mathbb{R}_+$ that models the amount of probability mass per unit length on the real line. The density provides another formula for the law:

$$\mu_X(\mathsf{B}) = \int_{\mathsf{B}} f_X(x)\, \lambda(\mathrm{d}x). \quad \text{for each Borel set } \mathsf{B}.$$

Do not confuse the density function f_X with the distribution function F_X. You should also be alert that a random variable that is not continuous does not admit a density.

On many probability spaces, we have an abundant collection of real random variables. In particular, for every event $\mathsf{E} \in \mathcal{F}$, we can define the indicator random variable $\mathbb{1}_{\mathsf{E}}$. Random variables are just (measurable) real-valued functions, so we can scale them, add them, and multiply them together to produce more random variables.

In general, two real random variables X, Y can interact in complicated ways. To describe the interaction completely, it is not enough to consider the marginal laws μ_X and μ_Y. Rather, we need to evaluate the joint law

$$\mu_{XY}(\mathsf{B}) := \mathbb{P}\{(X, Y) \in \mathsf{B}\} \quad \text{for each Borel set } \mathsf{B} \in \mathcal{B}(\mathbb{R}^2) \text{ in the plane.}$$

The joint law contains all of the information we need to understand the distribution of outcomes of the pair (X, Y). We can define the joint law of any (finite) family of real random variables in a similar fashion.

9.2 Expectation

What is the average value of a random variable? In analogy with a mechanical system, we want to sum up the values of the random variable weighted by the probability that it takes a particular value. In other words, we want to compute an integral.

9.2.1 Expectation and integration

By definition, a real random variable $X : \Omega \to \mathbb{R}$ is a measurable function on the sample space Ω. Therefore, we can integrate the random variable with respect to the probability measure $\mathbb{P}$ using the Lebesgue integral defined in Lecture 2.

Definition 9.1 (Expectation). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and let $X : \Omega \to \overline{\mathbb{R}}$ be a real random variable that may take extended values. The *expectation* of the random variable is the number

$$\mathbb{E}[X] := \mathbb{P}(X) := \int_\Omega X(\omega)\, \mathbb{P}(\mathrm{d}\omega) := \int_\Omega X \,\mathrm{d}\mathbb{P},$$

provided either that X is positive or that X is finite and $\mathbb{P}$-integrable.

We often write expectation without brackets ($\mathbb{E}\, X$). In this case, nonlinear functions bind before the expectation. For example, $\mathbb{E}\, X^2 := \mathbb{E}[X^2]$. The notation $\mathbb{P}(X)$ is analogous to the functional notation for integrals, but we prefer to write $\mathbb{E}[X]$.

Keep in mind that Definition 9.1 involves a Lebesgue integral. Therefore, for positive random variables,

$$\mathbb{E}[X] := \int_0^\infty \mathbb{P}\{X > t\}\, \mathrm{d}t = \int_{\mathbb{R}_+} \mathbb{P}\{X > t\}\, \lambda(\mathrm{d}t) \quad \textit{when } X \ge 0. \tag{9.1}$$

Proposition 4.39 confirms the equivalence of the Riemann and Lebesgue integrals in the last display. For finite-valued, integrable random variables,

$$\mathbb{E}[X] := \mathbb{E}[X_+] - \mathbb{E}[X_-] \quad \textit{when } \mathbb{E}\,|X| < +\infty. \tag{9.2}$$

Following our standard practice, let us carve out the class of random variables that have finite expectation.

Definition 9.2 (Integrable random variables). We define the linear space of integrable random variables:

$$\mathsf{L}_1 := \mathsf{L}_1(\Omega, \mathcal{F}, \mathbb{P}) := \big\{X : \Omega \to \mathbb{R} \text{ measurable} : \mathbb{E}\,|X| < +\infty\big\}.$$

Warning 9.3 (Non-integrable random variables). Not every real random variable has an expectation! For instance, a Cauchy random variable (Example 8.20) is not integrable because the distribution has too much mass away from zero. In other terms, a Cauchy variable has very heavy tails. Non-integrable random variables have some unintuitive behavior, but they are not just a curiosity because they arise in many applied problems. Nevertheless, we will focus on integrable random variables in this course. ▪

9.2.2 Change of variables

Integration on an abstract space probably remains a little mysterious, but we can shift our attention to the real line by means of the following result.

Proposition 9.4 (Law of the unconscious statistician). Let X be a real random variable defined on a probability space, and let $h : \mathbb{R} \to \overline{\mathbb{R}}$ be a measurable function. Then

Less frivolous authors call this result the *change of variables* formula.

$$\mathbb{E}[h(X)] = \int_{\mathbb{R}} h(x)\, \mu_X(\mathrm{d}x),$$

provided either that h is positive or that h is μ_X-integrable. See Figure 9.1 for an illustration.

Proof. This is just the change of variables formula from Problem 5.44. Now is the time to make sure the details are clear.

First, we assume that $h : \mathbb{R} \to \overline{\mathbb{R}}_+$ is a positive, measurable function. The composition $h \circ X$ is positive and measurable, hence a positive random variable. By definition, the integral on the right-hand side satisfies

$$\begin{aligned}
\int_{\mathbb{R}} h(x)\, \mu_X(\mathrm{d}x) &= \int_0^\infty \mu_X\{x \in \mathbb{R} : h(x) > t\}\, \mathrm{d}t \\
&= \int_0^\infty \mathbb{P}\big(X^{-1}\{x \in \mathbb{R} : h(x) > t\}\big)\, \mathrm{d}t \\
&= \int_0^\infty \mathbb{P}\left\{\omega \in \Omega : h(X(\omega)) > t\right\}\, \mathrm{d}t \\
&= \int_0^\infty \mathbb{P}\left\{h(X) > t\right\}\, \mathrm{d}t = \mathbb{E}[h(X)].
\end{aligned}$$

We have used Definition 8.4 of the law μ_X and the definition (9.1) of the expectation of the positive random variable $h(X)$.

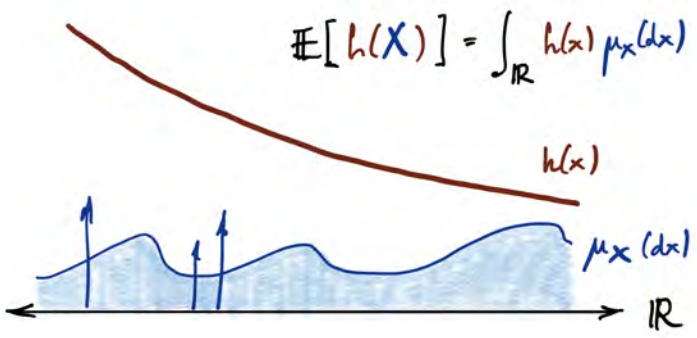


Figure 9.1 (Change of variables). To compute the expectation of a function $h(X)$ of a random variable X, we can integrate the function h with respect to the law μ_X.

In case $h : \mathbb{R} \to \mathbb{R}$ is μ_X-integrable, its positive part h_+ and negative part h_- have finite integrals with respect to μ_X. By definition of the integral of a signed function,

$$\begin{aligned}\int_{\mathbb{R}} h(x)\,\mu_X(\mathrm{d}x) &:= \int_{\mathbb{R}} h_+(x)\,\mu_X(\mathrm{d}x) - \int_{\mathbb{R}} h_-(x)\,\mu_X(\mathrm{d}x) \\ &= \mathbb{E}[h_+(X)] - \mathbb{E}[h_-(X)] = \mathbb{E}[h(X)].\end{aligned}$$

We have applied the change of variables formula from the last paragraph twice, once for h_+ and once for h_-. Last, we use the definition (9.2) of the expectation of the signed random variable $h(X)$. ∎

For a striking example of Proposition 9.4, we can choose $h : x \mapsto x$ to obtain

$$\mathbb{E}[X] = \int_{\mathbb{R}} x\,\mu_X(\mathrm{d}x). \tag{9.3}$$

This accords with our familiar notions of expectation on the line. Heuristically, we sum the value x of the random variable weighted by the probability $\mu_X(\mathrm{d}x)$ that the random variable takes the value x. See Figure 9.1.

It is instructive to instantiate the formula (9.3) for particular types of random variables. When the random variable X is discrete,

$$\mu_X = \sum\nolimits_{i=1}^{\infty} p_i \delta_{a_i} \quad \text{implies} \quad \mathbb{E}[X] = \sum\nolimits_{i=1}^{\infty} a_i p_i. \tag{9.4}$$

When the random variable X is continuous with density f_X,

$$\mu_X(\mathsf{B}) = \int_{\mathsf{B}} f_X(x)\,\lambda(\mathrm{d}x) \quad \text{implies} \quad \mathbb{E}[X] = \int_{\mathbb{R}} x\,f_X(x)\,\lambda(\mathrm{d}x). \tag{9.5}$$

This calculation requires Proposition 9.5 below (see also Exercise 9.6). In other words, (9.3) captures the familiar formulas from elementary probability and more.

9.2.3 Continuous random variables

For continuous random variables, there is an extension of Proposition 9.4 that displays the role of the density more clearly.

Proposition 9.5 (Expectation: Continuous random variable). Let X be a continuous real random variable with density f_X and with law μ_X. For a measurable function $h : \mathbb{R} \to \mathbb{R}$,

$$\mathbb{E}[h(X)] = \int_{\mathbb{R}} h(x)\, \mu_X(\mathrm{d}x) = \int_{\mathbb{R}} h(x)\, f_X(x)\, \lambda(\mathrm{d}x),$$

provided that h is positive or that the integral on the right-hand side is finite.

**Proof.* We may assume that $h : \mathbb{R} \to \mathbb{R}_+$ is a positive, measurable function. The result for signed functions follows by splitting h into its positive and negative parts.

By definition (8.1) of the law μ_X in terms of the density f_X, we have

$$\int_{\mathbb{R}} \mathbb{1}_{\mathsf{B}}(x)\, \mu_X(\mathrm{d}x) = \mu_X(\mathsf{B}) = \int_{\mathbb{R}} \mathbb{1}_{\mathsf{B}}(x)\, f_X(x)\, \lambda(\mathrm{d}x).$$

By linearity of the integral, the same relation holds for each positive simple function $s : \mathbb{R}_+ \to \mathbb{R}$:

$$\int_{\mathbb{R}} s(x)\, \mu_X(\mathrm{d}x) = \int_{\mathbb{R}} s(x)\, f_X(x)\, \lambda(\mathrm{d}x).$$

Recall that a positive simple function is a (finite) linear combination of indicator functions with positive coefficients.

Last, we use the staircase maps (5.2) to approximate h by an increasing limit of simple functions: $(Q_j \circ h) \uparrow h$. Apply the last display to the simple function $s = Q_j \circ h$. Invoke monotone convergence (Theorem 5.18) on reach the result for h. ∎

Exercise 9.6 (*Expectation: Continuous random variable). Give an alternative proof of Proposition 9.5 based on the definition (4.5) of the integral, the definition (8.1) of the law μ_X, and Fubini–Tonelli (Theorem 6.23).

9.2.4 Properties of expectation

Since expectation is just a Lebesgue integral, it inherits all of the basic properties of the Lebesgue integral.

Theorem 9.7 (Expectation: Properties). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space, and consider *integrable* real random variables $X, Y \in \mathsf{L}_1$.

1. **Indicators:** The expectation of the indicator of an event equals the probability of the event:
$$\mathbb{E}[\mathbb{1}_{\mathsf{E}}] = \mathbb{P}(\mathsf{E}) \quad \text{for each event } \mathsf{E} \in \mathcal{F}.$$
2. **Unital:** $\mathbb{E}[1] = 1$.
3. **Positive:** If $X \geq 0$, then $\mathbb{E}[X] \geq 0$.
4. **Monotone:** If $X \leq Y$, then $\mathbb{E}[X] \leq \mathbb{E}[Y]$.
5. **Linear:** For all scalars $\alpha, \beta \in \mathbb{R}$,
$$\mathbb{E}[\alpha X + \beta Y] = \alpha\, \mathbb{E}[X] + \beta\, \mathbb{E}[Y].$$
6. **Negligible sets:** If $X = Y$ $\mathbb{P}$-almost surely, then $\mathbb{E}[X] = \mathbb{E}[Y]$.

Recall that probabilists say "almost sure" instead of "almost everywhere".

Proof. This is just a restatement of Theorem 5.14. ∎

The only statement here that is special for the expectation is the unital property (2) that the expectation reproduces the constant 1. This point reflects the interpretation of the expectation is a (weighted) average. Another distinctive property of the expectation is Jensen's inequality, which we will discuss in Section 9.4.

The most important single property of expectation is linearity (4). Let us emphasize that this result holds for all (integrable) random variables, regardless of how they are related to each other. This innocuous result has extensive implications.

The statement (1) that the expectation of the indicator of an event is the probability of the event also has many useful consequences. For *arbitrary* events $\mathsf{E}_i \in \mathcal{F}$ with $i \in \mathbb{N}$, we can compute

$$\mathbb{E}\left[\#\{i \in \mathbb{N} : \mathsf{E}_i \text{ occurs}\}\right] = \mathbb{E}\left[\sum_{i=1}^{\infty} \mathbb{1}_{\mathsf{E}_i}\right] = \sum_{i=1}^{\infty} \mathbb{P}(\mathsf{E}_i).$$

The first relation is a reinterpretation, and the second relation is Tonelli's theorem for sums. In other words, we can use indicators to count how many events in a given family actually occur.

Activity 9.8 (Expectation: Positive random variables). As usual, results for the expectation can be divided into results for positive random variables and results for integrable random variables. We have not stated the specialized results for positive random variables here, but they can be extracted from Theorem 5.14. Write out these results using probabilistic language and terminology. ▪

Exercise 9.9 (Expectation: Range). Fix extended real numbers $a, b \in \overline{\mathbb{R}}$. Let $X : \Omega \to [a, b]$ be a real random variable that takes values in an interval, possibly infinite. Assuming that the expectation $\mathbb{E}\, X$ is defined, show that

$$a \le \mathbb{E}\, X \le b.$$

In other words, the expectation of X remains inside the range of possible values of X. **Hint:** Use monotonicity of the expectation.

9.2.5 Convergence theorems

Since the expectation is just a Lebesgue integral, it also comes equipped with a family of convergence theorems. These results follow from the analogous results for Lebesgue integrals with only a change of notation.

Theorem 9.10 (Expectation: Monotone convergence). Let $(X_i : i \in \mathbb{N})$ be a pointwise *increasing* sequence of *positive* random variables that may take extended values. Then

$$X_i \uparrow X \quad \text{implies} \quad \mathbb{E}[X_i] \uparrow \mathbb{E}[X].$$

In detail, $X_{i+1}(\omega) \ge X_i(\omega)$ for each $i \in \mathbb{N}$ and each $\omega \in \Omega$. The random variable $X : \Omega \to \overline{\mathbb{R}}_+$ is defined as the limit of the increasing sequence.

Theorem 9.10 is just a restatement of Theorem 5.18. It holds without any further qualification on the random variables. Indeed, an increasing sequence of positive random variables has a limit, which is always a positive random variable (that may take the value $+\infty$).

Theorem 9.11 (Expectation: Fatou's lemma). Let $(X_i : i \in \mathbb{N})$ be a sequence of *positive* random variables that may take extended values. Then

$$\liminf_{i\to\infty} \mathbb{E}[X_i] \ge \mathbb{E}\left[\liminf_{i\to\infty} X_i\right].$$

Theorem 9.11 is just a restatement of Theorem 5.20. It holds without any qualification, except that the random variables must be positive.

Theorem 9.12 (Expectation: Dominated convergence). Let $(X_i : i \in \mathbb{N})$ be a sequence of random variables. Assume that the sequence is *dominated by a fixed, integrable*

random variable:

$$|X_i| \le |Y| \quad \text{for all } i \in \mathbb{N}, \text{ where } Y \in \mathsf{L}_1(\mathbb{P}).$$

Then

$$X_i \to X \quad \text{implies} \quad \mathbb{E}[X_i] \to \mathbb{E}[X].$$

Warning: The dominating random variable Y cannot depend on the index i. ■

Theorem 9.12 is a specialization of Theorem 5.22. It is essential that the random variables be dominated by a single random variable Y that is integrable.

For the expectation, there is a special case of Theorem 9.12 that is often easier to use.

Corollary 9.13 (Expectation: Bounded convergence). Let $(X_i : i \in \mathbb{N})$ be a sequence of random variables. Assume that the sequence is *uniformly bounded by a constant*:

$$|X_i| \le M \quad \text{for all } i \in \mathbb{N}, \text{ where } M \in \mathbb{R}_+.$$

Warning: The constant M cannot depend on the index i. ■

Then

$$X_i \to X \quad \text{implies} \quad \mathbb{E}[X_i] \to \mathbb{E}[X].$$

Proof. Apply Theorem 9.12 with the random variable $Y(\omega) = M$ for all $\omega \in \Omega$. Constants are integrable with respect to a probability measure. ■

All of these convergence theorems have counterparts, assuming only convergence $\mathbb{P}$-almost surely. We omit explicit statements.

9.2.6 *Expectation in the plane

We have defined the expectation of a real-valued random variable. The definition can be extended to pairs of real-valued random variables in an obvious way.

Definition 9.14 (Pairs of real random variables: Expectation). Let (X, Y) be a pair of (finite-valued) real random variables. If both X and Y are $\mathbb{P}$-integrable, we define

$$\mathbb{E}(X, Y) := (\mathbb{E}\, X, \mathbb{E}\, Y).$$

That is, the expectation of the random vector $(X, Y) \in \mathbb{R}^2$ is simply the vector of the expectations.

Given a pair (X, Y) of real random variables with joint law μ_{XY}, we can compute the expectation of a bivariate measurable function $h : \mathbb{R}^2 \to \mathbb{R}$ using the law of the unconscious statistician:

$$\mathbb{E}[h(X, Y)] = \int_{\mathbb{R}^2} h(x, y)\, \mu_{XY}(\mathrm{d}x \times \mathrm{d}y). \tag{9.6}$$

This formula is valid when h is positive or when h is μ_{XY}-integrable. The proof is the same as Proposition 9.4.

When (X, Y) is an independent pair, the joint law $\mu_{XY} = \mu_X \times \mu_Y$. Therefore, we can invoke Fubini–Tonelli (Theorem 6.23) to pass from the double integral to an iterated integral.

Similar results are valid for any random vector $(X_1, \dots, X_n) \in \mathbb{R}^n$ whose components X_i are real random variables.

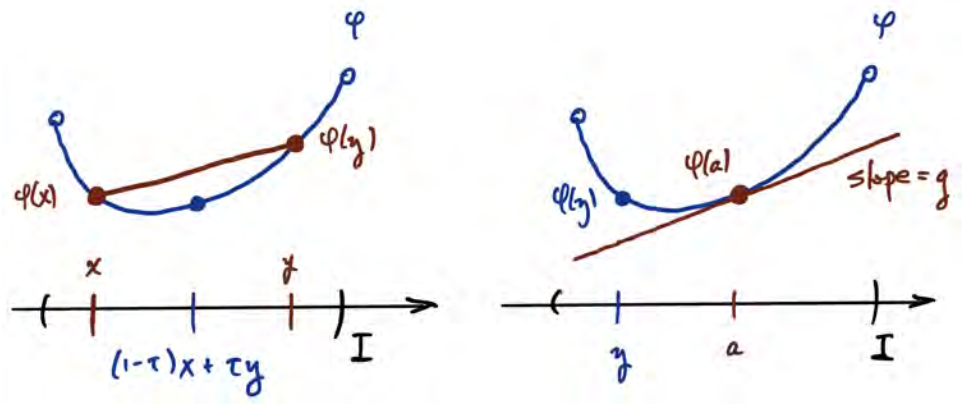


Figure 9.2 (Convex functions). Left: A convex function φ lies below its secants. **Right:** A convex function φ lies above its tangents.

9.3 Convex functions on the real line

As we have seen, the expectation of a random variable is just the integral with respect to the measure of probability. The expectation inherits all of the properties of the integral, but it enjoys a few additional perquisites because it is a weighted average. For example, we have already shown that the expectation reproduces constants.

For related reasons, expectation also interacts elegantly with convex functions. To develop this idea, we first define the concept of a convex function, and we establish some of the key properties. The proofs are included here for completeness, but the reader may focus on the statements rather than the arguments.

9.3.1 Convex functions

Linear functions preserve linear equality relations. When we consider linear *inequality* relations, we soon encounter convex functions. The following humble definition is central to a large part of applied mathematics.

Definition 9.15 (Convex function). Let $\mathsf{I} \subseteq \mathbb{R}$ be an *interval* of the real line, not necessarily finite. A (finite-valued) function $\varphi : \mathsf{I} \to \mathbb{R}$ is *convex* if

$$\varphi\big((1-\tau)x + \tau y\big) \le (1-\tau)\,\varphi(x) + \tau\,\varphi(y) \quad \text{for all } \tau \in [0,1]$$

and all points $x, y \in \mathsf{I}$. A function $\psi : \mathsf{I} \to \mathbb{R}$ is called *concave* if its negation $-\psi$ is convex.

In words, a convex function lies below its secants. Dually, a convex function lies above its tangents, as we will prove below in Proposition 9.19. The illustrations in Figure 9.2 capture these two ideas.

Example 9.16 (Convex functions). Many familiar functions are convex or concave.

1. **Affine functions:** For $a, b \in \mathbb{R}$, the affine function $t \mapsto a + bt$ is convex on $\mathbb{R}$.
2. **Absolute powers:** For $p \ge 1$, the function $t \mapsto |t|^p$ is convex on $\mathbb{R}$.
3. **Inverse powers:** For $p > 0$, the function $t \mapsto t^{-p}$ is convex on $\mathbb{R}_{++}$.
4. **Exponential:** For $\theta \in \mathbb{R}$, the function $t \mapsto \mathrm{e}^{\theta t}$ is convex on $\mathbb{R}$.
5. **Powers:** For $0 < p \le 1$, the function $t \mapsto t^p$ is concave on $\mathbb{R}_+$.

6. **Logarithm:** The function $t \mapsto \log t$ is concave on $\mathbb{R}_{++}$.
7. **Entropy:** The function $t \mapsto -t \log t$ is concave on $\mathbb{R}_{+}$.

You can verify these results using the tools from Section 9.3.4. We will encounter more examples of convex functions later. ▪

> **Aside:** We say that a function $\varphi : \mathsf{I} \to \mathbb{R}$ is *strictly convex* when
>
> $$\varphi\big((1-\tau)x + \tau y\big) < (1-\tau)\,\varphi(x) + \tau\,\varphi(y) \quad \text{for all } \tau \in [0,1]$$
>
> and all points $x, y \in \mathsf{I}$. Strict convexity means that the function does not have any affine segments. Which of the functions in Example 9.16 are convex?

9.3.2 Continuity and subgradients

The key property of a convex function on the real line is that its secants are increasing. This point is evident from Figure 9.2.

Lemma 9.17 (*Convex function: Secants). Let $\varphi : \mathsf{I} \to \mathbb{R}$ be a convex function on an interval I of the real line. For each $a \in \mathsf{I}$, introduce the secant function:

$$(\mathrm{D}\varphi)(a; h) := \frac{\varphi(a+h) - \varphi(a)}{h} \quad \text{when } h \neq 0 \text{ and } a + h \in \mathsf{I}.$$

For fixed $a \in \mathsf{I}$, the function $h \mapsto \mathrm{D}(a; \cdot)$ is increasing:

$$h \le h' \quad \text{implies} \quad (\mathrm{D}\varphi)(a; h) \le (\mathrm{D}\varphi)(a; h').$$

**Proof.* This statement is an algebraic consequence of Definition 9.15. For example, when $h > 0$, choose $x = a$ and $y = a + h'$, and set $\tau = h/h'$ so that $(1-\tau)x + \tau y = a + h$. The other cases are similar. ∎

Lemma 9.17 has several significant consequences. First of all, convex functions are essentially continuous.

Proposition 9.18 (Convex function: Continuity). A convex function $\varphi : \mathsf{U} \to \mathbb{R}$ on an *open* interval $\mathsf{U} \subseteq \mathbb{R}$ is continuous. In particular, φ is Borel measurable.

> **Warning:** A convex function on a closed interval need not be continuous at the endpoints. ▪

**Proof.* Fix $a \in \mathsf{U}$, and let $k > h > 0$ be sufficiently small that $a \pm k \in \mathsf{U}$. Lemma 9.17 implies that

$$(\mathrm{D}\varphi)(a; -k) \le (\mathrm{D}\varphi)(a; h) \le (\mathrm{D}\varphi)(a; +k).$$

It follows that

$$|(\mathrm{D}\varphi)(a; h)| \le \max\big\{|(\mathrm{D}\varphi)(a; +k)|,\ |(\mathrm{D}\varphi)(a; -k)|\big\} =: C.$$

Writing out the secant function on the left-hand side and rearranging, we see that φ is right-continuous at a. That is,

$$|\varphi(a+h) - \varphi(a)| \le Ch \quad \text{as } h \downarrow 0.$$

A similar argument shows that φ is left-continuous at a. ∎

Second, let us demonstrate that a convex function lies above its tangents in the following sense.

Proposition 9.19 (Convex function: Subgradient inequality). Let $\varphi : \mathsf{U} \to \mathbb{R}$ be a convex function on an *open* interval $\mathsf{U} \subseteq \mathbb{R}$. For each $a \in \mathsf{U}$, there is a number $g \in \mathbb{R}$ such that

$$\varphi(y) \geq \varphi(a) + g \cdot (y - a) \quad \text{for all } y \in \mathsf{U}. \tag{9.7}$$

The number g appearing in (9.7) is called a *subgradient* of the convex function φ at the point a. If φ is differentiable at a, then $g = \varphi'(a)$ is uniquely determined.

The subgradient g is not necessarily unique.

**Proof.* Fix a point $a \in \mathsf{U}$. Let $k > h > 0$ be sufficiently small that $a \pm k \in \mathsf{U}$. Lemma 9.17 implies that

$$(\mathrm{D}\varphi)(a; -k) \leq (\mathrm{D}\varphi)(a; h) \leq (\mathrm{D}\varphi)(a; +k).$$

As h decreases to zero, the secant $(\mathrm{D}\varphi)(a; h)$ is decreasing and bounded below by the left-hand side. Therefore,

$$g_+ := \inf_{h>0} (\mathrm{D}\varphi)(a; h) = \lim_{h \downarrow 0} (\mathrm{D}\varphi)(a; h) \leq (\mathrm{D}\varphi)(a; +k).$$

Writing out the secant $(\mathrm{D}\varphi)(a; +k)$ and rearranging,

$$\varphi(a + k) \geq \varphi(a) + g_+ k \quad \text{for all } k > 0 \text{ with } a + k \in \mathsf{U}.$$

A parallel argument shows that $g_- := \lim_{h \uparrow 0} (\mathrm{D}\varphi)(a; h)$ satisfies

$$\varphi(a - k) \geq \varphi(a) - g_- k \quad \text{for all } k > 0 \text{ with } a - k \in \mathsf{U}.$$

Lemma 9.17 also guarantees that $g_- \leq g_+$. Combining the last two displays, we realize that the inequality (9.7) holds for any number $g \in [g_-, g_+]$. ∎

9.3.3 Dual representation

The supremum of affine functions is always a convex function.

Proposition 9.20 (Supremum of affine functions). Let $a_j, b_j \in \mathbb{R}$ for each $j \in \mathsf{J}$ in an arbitrary index set. Define a function $\varphi : \mathbb{R} \to \overline{\mathbb{R}}$ by the rule

$$\varphi(x) := \sup\{a_j + b_j x : j \in \mathsf{J}\} \quad \text{for each } x \in \mathbb{R}.$$

Then φ is a convex function on its domain.

The *domain* of an extended function is the set of places where it is finite:

$$\mathrm{dom}(\varphi) := \{x \in \mathbb{R} : |\varphi(x)| < +\infty\}.$$

**Proof.* Choose $x, y \in \mathrm{dom}(\varphi)$ and $\tau \in [0, 1]$. Calculate that

$$\begin{aligned} \varphi((1-\tau)x + \tau y) &= \sup\{(1-\tau)(a_j + b_j x) + \tau(a_j + b_j y) : j \in \mathsf{J}\} \\ &\leq (1-\tau)\sup\{a_j + b_j x : j \in \mathsf{J}\} + \tau \sup\{a_j + b_j y : j \in \mathsf{J}\} \\ &= (1-\tau)\varphi(x) + \tau\varphi(y). \end{aligned}$$

Indeed, the supremum is subadditive and positively homogeneous. ∎

We may ask whether the converse holds. Can we write a convex function as the supremum of affine functions? We can answer this question in the affirmative by invoking the subgradient inequality. This result will play a role in our discussion of conditional expectation.

Corollary 9.21 (Convex function: Dual representation). Let $\varphi : \mathsf{U} \to \mathbb{R}$ be a convex function on an open interval U of $\mathbb{R}$. Then there is a *countable* set $\mathsf{J} \subseteq \mathbb{R}^2$ for which

$$\varphi(y) = \sup\{\varphi(a) + g \cdot (y - a) : (a, g) \in \mathsf{J}\} \quad \text{for all } y \in \mathsf{U}.$$

Proof. For each *rational* $a \in \mathsf{U} \cap \mathbb{Q}$, the subgradient inequality (9.7) furnishes a subgradient g_a for which

$$\varphi(y) \geq \varphi(a) + g_a \cdot (y - a) \quad \text{for all } y \in \mathsf{U}.$$

For rational $y \in \mathsf{U} \cap \mathbb{Q}$, we see that

$$\varphi(y) = \sup\{\varphi(a) + g_a \cdot (y - a) : a \in \mathsf{U} \cap \mathbb{Q}\}.$$

Proposition 9.18 states that φ is continuous, so the same formula is valid for all $y \in \mathsf{U}$ because the rationals are dense in U. ∎

9.3.4 Sufficient conditions

How can we confirm that a real-valued function is convex? From a vast arsenal of methods, let us select a few that are particularly useful. If you want, you can develop proofs of these results from the material presented above.

Fact 9.22 (Convex functions: Closed convex cone). Let $\mathsf{I} \subseteq \mathbb{R}$ be an interval. If $f, g : \mathsf{I} \to \mathbb{R}$ are convex functions and $\alpha, \beta \geq 0$, then $\alpha f + \beta g$ is a convex function. Suppose that $f := \lim_i f_i$ is the pointwise limit of a sequence $f_i : \mathsf{I} \to \mathbb{R}$ of convex functions for $i \in \mathbb{N}$. Then the limit $f : \mathsf{I} \to \mathbb{R}$ remains a convex function. ▪

Fact 9.23 (Convex function: First-derivative test). Suppose that $\varphi : \mathsf{U} \to \mathbb{R}$ is a differentiable function on an open interval U of the real line. If the derivative $\varphi' : \mathsf{U} \to \mathbb{R}$ is an increasing function, then φ is convex. ▪

Fact 9.24 (Convex function: Second-derivative test). Suppose that $\varphi : \mathsf{U} \to \mathbb{R}$ is a twice-differentiable function on an open interval U of the real line. If the second derivative $\varphi'' : \mathsf{U} \to \mathbb{R}$ is a positive-valued function, then φ is convex. ▪

Another powerful method is to exhibit a dual representation of a function as the supremum of affine functions. Then Proposition 9.20 guarantees convexity.

Exercise 9.25 (Convex functions: Examples). Use these methods to confirm that each of the functions listed in Example 9.16 is a convex function.

9.4 Jensen's inequality

We are now prepared to prove Jensen's inequality, which describes how the expectation interacts with a convex function.

The definition of a convex function $\varphi : \mathsf{U} \to \mathbb{R}$ involves a two-point inequality. In particular,

$$\tfrac{1}{2}\varphi(a) + \tfrac{1}{2}\varphi(b) \geq \varphi\big(\tfrac{1}{2}a + \tfrac{1}{2}b\big) \quad \text{for all } a, b \in \mathsf{U}. \tag{9.8}$$

That is, the simple average of function values exceeds the function at the simple average of the arguments.

Jensen's inequality shows that the definition of convexity is self-improving. For every probability measure, the expectation of the function exceeds the function of the expectation. This result is an immediate consequence of the subgradient inequality.

Theorem 9.26 (Jensen's inequality). Let $\varphi : \mathsf{U} \to \mathbb{R}$ be a convex function on an open interval $\mathsf{U} \subseteq \mathbb{R}$, and assume that φ is *bounded below*. For each $\mathbb{P}$*-integrable* random variable $X : \Omega \to \mathsf{U}$ that takes values in U,

$$\mathbb{E}\,\varphi(X) \geq \varphi(\mathbb{E}\,X). \tag{9.9}$$

Jensen's inequality requires mild conditions to ensure that all of the expectations are defined. As we will see, there are several ways to achieve this goal.

It is possible that the left-hand side equals $+\infty$.

The two-point inequality (9.8) is exactly the statement of Jensen's inequality (9.9) for the random variable X with the discrete law $\mu_X = \frac{1}{2}\delta_a + \frac{1}{2}\delta_b$.

Proof. Without loss of generality, we may assume that the convex function $\varphi : \mathsf{U} \to \mathbb{R}_+$ takes positive values by adding a constant, since it is bounded below. Thus, the expectation $\mathbb{E}\,\varphi(X)$ is defined, although it may equal $+\infty$.

Instantiate the subgradient inequality (9.7) with the value $a = \mathbb{E}\,X$. There is a fixed number $g \in \mathbb{R}$, depending on a, for which

$$\varphi(X(\omega)) \geq \varphi(a) + g \cdot (X(\omega) - a) \quad \text{for each } \omega \in \Omega.$$

Take the expectation of this inequality using monotonicity and linearity:

$$\begin{aligned} \mathbb{E}\,\varphi(X) &\geq \mathbb{E}[\varphi(a) + g \cdot (X - a)] \\ &= \varphi(a) + g \cdot (\mathbb{E}[X] - a) = \varphi(\mathbb{E}\,X). \end{aligned}$$

We have used the fact that $\mathbb{E}\,X = a$, and the expectation preserves constants. ∎

For concave functions, Jensen's inequality is reversed.

Corollary 9.27 (Jensen's inequality: Concave function). Let $\psi : \mathsf{U} \to \mathbb{R}$ be a concave function on an open interval $\mathsf{U} \subseteq \mathbb{R}$, and assume that ψ is *bounded above*. For each $\mathbb{P}$-*integrable* random variable $X : \Omega \to \mathsf{U}$ that takes values in U,

$$\mathbb{E}\,\psi(X) \leq \psi(\mathbb{E}\,X).$$

Exercise 9.28 (Jensen's inequality: Concave function). Prove Corollary 9.27. Show that Jensen's inequality for concave functions is also valid under the assumption that ψ is *bounded below*, rather than bounded above.

Exercise 9.29 (Jensen's inequality: Integrable case). Assume that $\varphi : \mathsf{U} \to \mathbb{R}$ is a convex function on an open interval. Let $X : \Omega \to \mathsf{U}$ be a $\mathbb{P}$-integrable random variable, and further assume that $\varphi(X)$ is $\mathbb{P}$-integrable. Establish that (9.9) is valid. What is the analogous result for concave functions?

Exercise 9.30 (Jensen's strict inequality). Jensen's inequality holds with a strict inequality if we add some additional assumptions. Assume that $\varphi : \mathsf{U} \to \mathbb{R}$ is *strictly* convex, and the support of the law μ_X of the random variable X contains *at least two values*. Then

$$\mathbb{E}\,\varphi(X) > \varphi(\mathbb{E}\,X),$$

provided that all of the expectations exist. Prove it.

Aside: The proof of Jensen's inequality relies on the observation that a convex function lies above one of its tangents. It is occasionally useful to extract this argument in its pure form. Consider a function $\varphi : \mathsf{U} \to \mathbb{R}$ and a $\mathbb{P}$-integrable random variable $X : \Omega \to \mathsf{U}$. Suppose that there exists $g \in \mathbb{R}$ for which

$$\varphi(y) \geq \varphi(\mathbb{E}\,X) + g \cdot (y - \mathbb{E}\,X) \quad \text{for all } y \in \mathsf{U}.$$

Then $\mathbb{E}\,\varphi(X) \geq \varphi(\mathbb{E}\,X)$.

9.4.1 Example: The GM–AM inequality

As a first example of the power of Jensen's inequality, we will establish the basic inequality between the generalized geometric mean (GM) and the arithmetic mean (AM) of two numbers.

Proposition 9.31 (GM–AM inequality). Fix *positive* numbers $p_1, \dots, p_n \geq 0$ with sum $\sum_{i=1}^n p_i = 1$. Then

$$\prod_{i=1}^n x_i^{p_i} \leq \sum_{i=1}^n p_i x_i \quad \text{for all } \textit{positive } x_i \in \mathbb{R}_+. \tag{9.10}$$

The left-hand side is a generalized geometric mean, while the right-hand side is a generalized arithmetic mean.

In particular, for $\tau \in [0, 1]$, we deduce that

$$x^\tau y^{1-\tau} \leq \tau x + (1-\tau) y \quad \text{for all } x, y \geq 0. \tag{9.11}$$

This is the usual statement of the GM–AM inequality for two variables.

Proof. The proof combines the fact that the exponential function is convex (Example 9.16) with Jensen's inequality (Theorem 9.26).

If $x_i = 0$ for any index i, then the inequality (9.10) obviously holds. Therefore, we may assume that each $x_i > 0$. Let us rewrite the quantity of interest:

$$\prod_{i=1}^n x_i^{p_i} = \exp\left(\sum_{i=1}^n p_i \log(x_i) \right).$$

In light of this expression, we introduce the random variable X that takes the values $\log(x_i)$ with probability p_i. The law is $\mu_X = \sum_{i=1}^n p_i \delta_{\log(x_i)}$.

Using the formula (9.4) for the expectation of a discrete random variable, we can reinterpret the last expression:

$$\exp\left(\sum_{i=1}^n p_i \log(x_i) \right) = \exp\left(\mathbb{E}\, X \right) \leq \mathbb{E} \exp(X) = \sum_{i=1}^n p_i \mathrm{e}^{\log x_i}.$$

The inequality is Jensen's. Simplify both sides to reach the stated result. ∎

Exercise 9.32 (Young's inequality). From (9.11), deduce Young's inequality. Let $p > 1$, and define $q > 1$ by the conjugacy relation $1/p + 1/q = 1$. Then

$$|xy| \leq \frac{1}{p} \cdot |x|^p + \frac{1}{q} \cdot |y|^q \quad \text{for all } x, y \in \mathbb{R}.$$

Exercise 9.33 (Continuous GM–AM inequality). Let X be a strictly positive real random variable. Check that

$$\exp(\mathbb{E} \log(X)) \leq \mathbb{E}\, X.$$

Explain why this is a "continuous" analog of the GM–AM inequality.

9.5 *Convexity: Beyond the real line

We have introduce convex functions and Jensen's inequality on the real line, but these ideas have a wider ambit. In this section, we give a brief presentation of convex functions on the n-dimensional Euclidean space $\mathbb{R}^n$. Then we note that the analog of Jensen's inequality holds in this setting as well.

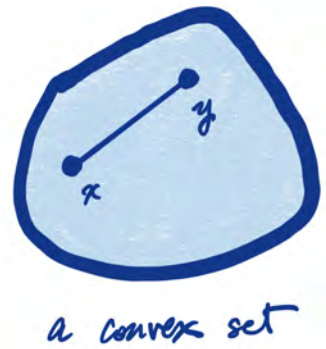
9.5.1 Convex sets and functions

On the real line, we defined a convex function on an interval. In higher dimensions, a convex function is defined on a domain called a convex set.

Definition 9.34 (Convex set). Let $\mathsf{C} \subseteq \mathbb{R}^n$. The set C is *convex* when

$$(1-\tau)\boldsymbol{x} + \tau\boldsymbol{y} \in \mathsf{C} \quad \text{for each } \tau \in [0,1] \text{ and all } \boldsymbol{x}, \boldsymbol{y} \in \mathsf{C}.$$

In other words, the set C contains the line segment connecting each pair of points in the set.



There are many familiar examples of convex sets. For instance, in the plane, a solid rectangle and a solid disc are both convex. In $\mathbb{R}^n$, the set of vectors with positive entries is convex.

As in the univariate setting, a multivariate convex function is a function on a convex set that lies below its secants.

Definition 9.35 (Convex function: Euclidean space). Let $\mathsf{C} \subseteq \mathbb{R}^n$ be a convex set. A real-valued function $\varphi : \mathsf{C} \to \mathbb{R}$ is called *convex* if

$$\varphi\big((1-\tau)\boldsymbol{x} + \tau\boldsymbol{y}\big) \le (1-\tau)\varphi(\boldsymbol{x}) + \tau\varphi(\boldsymbol{y}) \quad \text{for all } \tau \in [0,1] \text{ and all } \boldsymbol{x}, \boldsymbol{y} \in \mathsf{C}.$$

A function $\psi : \mathsf{C} \to \mathbb{R}$ is called *concave* if its negation $-\psi$ is convex.

Example 9.36 (Convex functions: Euclidean space). Many popular functions on $\mathbb{R}^n$ are convex or concave. Here are some examples, without proof.

1. **Affine functions:** For $a \in \mathbb{R}$ and $\boldsymbol{b} \in \mathbb{R}^n$, the affine function $\boldsymbol{x} \mapsto a + \boldsymbol{b}^{\mathsf{T}}\boldsymbol{x}$ is convex on $\mathbb{R}^n$.
2. **Positive quadratic forms:** For a *positive-semidefinite* matrix $\boldsymbol{A} \in \mathbb{R}^{n \times n}$, the quadratic function $\boldsymbol{x} \mapsto \boldsymbol{x}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{x}$ is convex on $\mathbb{R}^n$.
3. **Norm powers:** For a power $p \ge 1$, the function $\boldsymbol{x} \mapsto \|\boldsymbol{x}\|_2^p$ is convex on $\mathbb{R}^n$.
4. **Sum of exponentials:** For $\alpha_i \in \mathbb{R}$, the function $\boldsymbol{x} \mapsto \sum_{i=1}^n \mathrm{e}^{\alpha_i x_i}$ is convex on $\mathbb{R}^n$.
5. **Entropy:** The function $\boldsymbol{x} \mapsto -\sum_{i=1}^n x_i \log x_i$ is concave on $\mathbb{R}^n_+$.
6. **Geometric mean:** For $\tau \in [0,1]$, the function $(x, y) \mapsto x^\tau y^{1-\tau}$ is concave on $\mathbb{R}^2_+$.

Recall that $\boldsymbol{A} \in \mathbb{R}^{n \times n}$ is positive-semidefinite if and only if $\boldsymbol{u}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{u} \ge 0$ for each $\boldsymbol{u} \in \mathbb{R}^n$.

The menagerie of additional examples is vast. ∎

9.5.2 Dual representation

As in the univariate case, every multivariate convex function lies above its tangents, and these tangents provide a dual representation of the function.

Fact 9.37 (Convex function: Subgradients and dual representation). Let $\varphi : \mathsf{C} \to \mathbb{R}$ be a convex function on an *open* convex set $\mathsf{C} \subseteq \mathbb{R}^n$. For each $\boldsymbol{a} \in \mathsf{C}$, there is a subgradient vector $\boldsymbol{g} \in \mathbb{R}^n$ such that

$$\varphi(\boldsymbol{y}) \ge \varphi(\boldsymbol{a}) + \boldsymbol{g}^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{a}) \quad \text{for all } \boldsymbol{y} \in \mathsf{C}.$$

Furthermore, there is a *countable* set $\mathsf{J} \subseteq \mathbb{R}^{2n}$ for which

$$\varphi(\boldsymbol{y}) = \sup\{\varphi(\boldsymbol{a}) + \boldsymbol{g}^{\mathsf{T}}(\boldsymbol{y} - \boldsymbol{a}) : (\boldsymbol{a}, \boldsymbol{g}) \in \mathsf{J}\}.$$

This expression represents the convex function as a supremum of affine functions. ∎

9.5.3 Jensen's inequality

Jensen's inequality remains valid for multivariate convex functions.

Theorem 9.38 (Jensen's inequality). Let $\varphi : \mathsf{C} \to \mathbb{R}$ be a convex function on an open convex set $\mathsf{C} \subseteq \mathbb{R}^n$, and assume that φ is *bounded below*. For each $\mathbb{P}$*-integrable* random vector $\boldsymbol{X} = (X_1, \dots, X_n) : \Omega \to \mathsf{C}$ that takes values in C,

$$\mathbb{E}\, \varphi(\boldsymbol{X}) \geq \varphi(\mathbb{E}\, \boldsymbol{X}).$$

It is possible that the left-hand side equals $+\infty$ or that both sides equal $+\infty$.

**Proof sketch.* Follow the pattern of the proof of Theorem 9.38. Use the multivariate subgradient inequality (Fact 9.37) in place of the univariate form. ∎

Problems

Linearity of expectation is a powerful tool for solving problems. We begin with several examples where this principle allows us to make short work of an potentially challenging computation.

Exercise 9.39 (Inclusion–exclusion again). Let $\mathsf{E}_1, \dots, \mathsf{E}_n$ be events in a probability space. From Exercise 2.47, recall the identity

$$\mathbb{1}_{\bigcup_{i=1}^n \mathsf{E}_i} = \sum\nolimits_{k=1}^{n} (-1)^{k+1} \sum\nolimits_{i_1 < \cdots < i_k} \mathbb{1}_{\mathsf{E}_{i_1} \cap \cdots \cap \mathsf{E}_{i_k}}.$$

Prove that

$$\mathbb{P}\big(\bigcup\nolimits_{i=1}^n \mathsf{E}_i\big) = \sum\nolimits_{k=1}^{n} (-1)^{k+1} \sum\nolimits_{i_1 < \cdots < i_k} \mathbb{P}(\mathsf{E}_{i_1} \cap \dots \mathsf{E}_{i_k}).$$

Hint: Expectation is linear.

Exercise 9.40 (Derangements). A permutation $\pi : \{1, \dots, n\} \to \{1, \dots, n\}$ is a bijective function. A *fixed point* of a permutation π is a letter i for which $\pi(i) = i$. By elementary counting arguments, there are a total of $n!$ permutations on n letters. We may choose one of these $n!$ permutations uniformly at random.

1. For a random permutation on n letters, what is the expected number of fixed points? **Hint:** For each $i = 1, \dots, n$, consider the indicator of the event that i is a fixed point.
2. What is the probability that a random permutation on n letters has no fixed point? **Hint:** This is the keys problem (Problem 2.48).
3. What is the probability that a random permutation on n letters has k fixed points?
4. ***Ménages:** Suppose that n man–woman couples attend a dinner party. They are seated around a circular table, alternating between men and women. If one such configuration is chosen uniformly at random, what is the probability that no person is seated next to their partner?

Exercise 9.41 (Collect them all!). A certain brand of confection prints a joke on the inside of each wrapper. You eagerly purchase one confection every day after lunch so that you can share the joke with your officemates (who are equally enthusiastic about your obsession, I'm sure). There are n distinct jokes, and each confection is equally likely to contain each one of the jokes.

1. Compute the expected number of days between reading the ith novel joke and the $(i+1)$th novel joke for each $i = 0, \dots, n-1$.
2. Compute the expected number of days that it takes to encounter all n jokes.

Exercise 9.42 (Expectation and values). Information about the expectation of a random variable can translate into information about its values. Let X be a *positive*, real random variable.

1. Assume that $\mathbb{E}\, X = 0$. Deduce that $X = 0$ almost surely. **Hint:** This is a transliteration of one of the integral properties (Theorem 5.14).
2. Assume that $\mathbb{E}\, X < +\infty$. Deduce that $X < +\infty$ almost surely.

Recall that an event E is almost sure when $\mathbb{P}(\mathsf{E}) = 1$.

Problem 9.43 (Borel–Cantelli I). Let $(X_n : n \in \mathbb{N})$ be an arbitrary sequence of *positive* random variables. If the sum of expectations is finite, we can reach very strong conclusions about the limit of the sequence. This result is a core tool for proving almost-sure convergence.

1. Establish that
$$\sum\nolimits_{n=1}^{\infty} \mathbb{E}[X_n] < +\infty \quad \text{implies} \quad \mathbb{E}\left[\limsup\nolimits_{n\to\infty} X_n\right] = 0.$$
In particular, $\limsup_{n\to\infty} X_n = 0$ almost surely. **Hint:** Recall that $\limsup_{n\to\infty} X_n = \inf_{n\in\mathbb{N}} \sup_{i\geq n} X_i$. It is also convenient to make the simple bound $\sup_{i\geq n} X_i \leq \sum_{i\geq n} X_i$.
2. The classic formulation of the first Borel–Cantelli lemma follows when the random variables are indicators. Translate the result for indicators into a statement about events and the probability of the limit superior of a sequence of events. What does this result mean in words? (How many of the events can occur?)

Exercise 9.44 (Mengoli). Use Jensen's inequality to prove Mengoli's inequality:
$$\frac{1}{x-1} + \frac{1}{x} + \frac{1}{x+1} > \frac{3}{x} \quad \text{for } x > 1.$$

(*) Define the harmonic number $\mathrm{H}_n := \sum_{i=1}^{n} i^{-1}$ for $n \in \mathbb{N}$. Use Mengoli's inequality to prove that $\mathrm{H}_n \to \infty$ as $n \to \infty$.

Exercise 9.45 (*Isoperimetry for rectangles). The GM–AM inequality (Proposition 9.31) admits an elegant geometric interpretation.

1. Among all plane rectangles with fixed perimeter, prove that a square has the maximum area.
2. Equivalently, among all plane rectangles with fixed area, prove that a square has the least perimeter.
3. Among all rectangular parallelopipeds in $\mathbb{R}^n$ whose total side length is fixed, prove that a regular cube has the maximum volume.
4. Equivalently, among all rectangular parallelopipeds with fixed volume, prove that a regular cube has the minimum total side length.

Problem 9.46 (*Isoperimetry for polygons). Fix a natural number $n \geq 3$. Consider a convex polygon with n sides with all n vertices on the unit circle. Among all such polygons, prove that the regular n-gon has the maximum area. **Hint:** Use elementary geometry, trigonometric identities, and Jensen's inequality.

Next, we discuss some other results about convex functions.

Problem 9.47 (*Fenchel and Young). Let $h : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$ be an arbitrary function. Define the *Fenchel–Young conjugate*:

$$h^*(s) := \sup\{sx - h(x) : x \in \mathbb{R}\} \quad \text{for } s \in \mathbb{R}.$$

The choice of the letter s reflects its role as the *slope* of a line.

We explicitly allow both h and h^* to take on the value $+\infty$. This construction can be extended to functions in higher dimensions, where it is also an invaluable tool.

1. For every function h, prove that h^* is a lower-semicontinuous convex function. As a consequence, if we can show that a function φ is the Fenchel–Young conjugate of some other function, then it must be the case that φ is convex.
2. For $a, b \in \mathbb{R}$, consider the affine function $h(x) = a + bx$ for $x \in \mathbb{R}$. Compute the conjugate h^*.
3. For $p > 1$, consider the power function $h(x) = |x|^p$ for $x \in \mathbb{R}$. Compute h^*.
4. For $\theta \in \mathbb{R}$, consider the exponential function $h(x) = \exp(\theta x)$. Compute h^*.
5. Consider the negative entropy function $h(x) = x \log x$ for $x \geq 0$; set $h(x) = +\infty$ for $x < 0$. Compute h^*.
6. Establish the Fenchel–Young inequality. For every function $h : \mathbb{R} \to \mathbb{R} \cup \{+\infty\}$,
$$sx \leq h^*(s) + h(x) \quad \text{for all } s, x \in \mathbb{R}. \tag{9.12}$$
(*) When h is a differentiable convex function, find conditions under which (9.12) holds with equality.
7. Instantiate the Fenchel–Young inequality (9.12) for the power function $h(x) = |x|^p$ with $p > 1$. Compare with Exercise 9.32.
8. Instantiate the Fenchel–Young inequality (9.12) for the negative entropy function $h(x) = x \log x$. The result is not so interesting in one dimension, but it is a precursor to the Boltzmann–Gibbs variational principle.
9. For every function h, confirm that $h^{**} := (h^*)^* \leq h$.
10. **(**)** If h is a lower-semicontinuous (lsc) convex function, prove that $h^{**} = h$. **Hint:** There is a graphical proof based on studying the epigraphs.

Problem 9.48 (*Perspectives). Let $\varphi : \mathsf{I} \to \mathbb{R}$ be a function on an interval I of the real line. We can define a bivariate function

$$h_\varphi(x; s) := s \cdot \varphi(x/s) \quad \text{for } x \in \mathsf{I} \text{ and } s \in \mathbb{R}_{++}.$$

The function h is called the *perspective transform* of φ. The perspective transform can also be extended to higher dimensions.

1. Assuming that φ is convex, prove that the perspective transform h_φ is a convex function on $\mathsf{I} \times \mathbb{R}_{++}$. In detail, show that
$$h_\varphi(\tau x + (1-\tau)y; \tau s + (1-\tau)t) \leq \tau \cdot h_\varphi(x; s) + (1-\tau) \cdot h_\varphi(y; t)$$
for all $\tau \in [0, 1]$, for all $x, y \in \mathsf{I}$, and for all $s, t \in \mathbb{R}_{++}$.
2. Consider the quadratic-over-linear function $h : (x, y) \mapsto x^2/y$. Prove that h is convex on $\mathbb{R}_+ \times \mathbb{R}_{++}$.
3. Consider the divergence function $h : (x, y) \mapsto x(\log x - \log y)$. Prove that h is convex on $\mathbb{R}_{++} \times \mathbb{R}_{++}$.
4. Fix $\tau \in [0, 1]$. Consider the Heinz mean $h : (x, y) \mapsto x^\tau y^{1-\tau}$. Prove that h is *concave* on $\mathbb{R}_{++} \times \mathbb{R}_{++}$.
5. Fix $\tau \in [0, 1]$. Let X, Y be strictly positive random variables. Using Jensen's inequality (Theorem 9.38), deduce that
$$\mathbb{E}[X^\tau Y^{1-\tau}] \leq (\mathbb{E}\, X)^\tau (\mathbb{E}\, Y)^{1-\tau}.$$
Explain how to derive Hölder's inequality (Theorem 11.5) from this statement.

Algorithm 9.1 (Randomized Quicksort). A recursive, randomized algorithm that sorts a finite set of distinct real numbers in increasing order.

Input: The input $\mathsf{S} = \{a_1, \dots, a_m\}$ consists of m distinct real numbers
Output: The output $\boldsymbol{y} = (y_1, \dots, y_m)$ is a list of the input elements in increasing order

1 **function** RandQuicksort($\mathsf{S} = \{a_1, \dots, a_m\}$)
2 **if** $\mathsf{S} = \emptyset$ **then return** the empty list $\boldsymbol{y} = ()$.
3 Draw a random pivot $K \sim \textsc{uniform}\{1, \dots, m\}$
4 By comparing each element in S to a_K, form two subsets

$$\mathsf{S}_- = \{a_i : a_i < a_K\}$$
$$\mathsf{S}_+ = \{a_i : a_i > a_K\}$$

5 Recursively sort these two subsets:

$$\boldsymbol{y}_- = \text{RandQuicksort}(\mathsf{S}_-)$$
$$\boldsymbol{y}_+ = \text{RandQuicksort}(\mathsf{S}_+)$$

6 **return** the ordered list $\boldsymbol{y} = (\boldsymbol{y}_-, a_K, \boldsymbol{y}_+)$

Applications

Application 9.49 (First-moment method). Application 7.14 introduced the probabilistic method, a foundational technique that uses probability to establish the existence of a mathematical object with a distinguished property. In this application, we continue the development by introducing the *first-moment method*, a mechanism where we compute an expectation to demonstrate that a mathematical object exists. In these problems, linearity of expectation is a useful tool to keep in mind.

1. Let X be a real random variable, and fix $a \in \mathbb{R}$. Prove that $\mathbb{E}\, X > a$ implies that $X(\omega) > a$ for some $\omega \in \Omega$. In other words, a bound on the expectation ensures the existence of a particular type of sample point.
2. With a more careful argument, prove that $\mathbb{E}\, X \geq a$ implies that $X(\omega) \geq a$ for some $\omega \in \Omega$.
3. Suppose that a circular corral has 17 fenceposts, but exactly 5 of them are rotten. Prove that each consecutive sequence of 7 fenceposts contains at least 3 rotten posts. **Hint:** For each k, define the indicator that post k is rotten. Consider a random set of 7 consecutive fenceposts.
4. **Graph cuts:** Let $\mathsf{G} = (\mathsf{V}, \mathsf{E})$ be an undirected combinatorial graph. A *cut* is a subset S of the vertices. The *weight* of a cut S is the number of edges $e = \{u, v\} \in \mathsf{E}$ with $u \in \mathsf{S}$ and $v \in \mathsf{S}^{\mathsf{c}}$ or vice versa. Show that there is a cut whose weight is at least $(\#\mathsf{E})/2$.
5. **Vector balancing:** Suppose that $(\boldsymbol{u}_1, \dots, \boldsymbol{u}_n)$ are unit-norm vectors in $\mathbb{R}^d$. Show that there is a sequence $(\varepsilon_1, \dots, \varepsilon_n) \in \{\pm 1\}^n$ of signs for which

$$\left\| \sum\nolimits_{i=1}^{n} \varepsilon_i \boldsymbol{u}_i \right\|_2 \leq \sqrt{n}.$$

Application 9.50 (Randomized Quicksort). For some computational problems, the most elegant approach may involve a randomized algorithm, a procedure that makes random choices during its execution. Surprisingly, randomized algorithms can arise in settings (apparently) distant from probability theory.

For instance, a fundamental challenge in computer science is to sort a list $(a_1, a_2, \ldots, a_n)$ of n real numbers in increasing order. For simplicity, we will assume that the numbers are all *distinct*. There is a beautiful randomized recursive algorithm for this problem, called *randomized quicksort*; see Algorithm 9.1. In this application, we will analyze the algorithm.

1. Explain why RandQuicksort is a correct algorithm for sorting. More precisely, suppose that the initial input is a set $\{x_1, \ldots, x_n\}$ of n distinct real numbers. Prove that the output $\boldsymbol{y} = (y_1, \ldots, y_n)$ is an ordered list with the properties that $y_1 < y_2 < \cdots < y_n$ and that each element $y_i = x_{\pi(i)}$ for a permutation π on $\{1, \ldots, n\}$.
2. Without loss of generality, you may assume that the initial input is already ordered: $x_1 < x_2 < \cdots < x_n$. Why?
3. Argue that the algorithm (including the full recursion) compares a given pair x_i and x_j with indices $i < j$ either zero times or one time.
4. To analyze RandQuicksort, we need to count the number R of comparisons that it makes. For indices $i < j$, let E_{ij} be the event that the algorithm compares x_i and x_j at some point during the execution. Verify that
$$\mathbb{E}[R] = \sum\nolimits_{i<j} \mathbb{P}(\mathsf{E}_{ij}).$$
Hint: Write R as a sum of indicators.
5. Establish that $\mathbb{P}(\mathsf{E}_{ij}) = 2/(j - i + 1)$ for each pair of indices with $i < j$. **Hint:** The algorithm compares x_i and x_j if and only if (a) both x_i and x_j are in the input of a recursive execution of the algorithm *and* (b) one of x_i or x_j is selected as the pivot.
6. Deduce that $\mathbb{E}[R] < 2n \log n$. **Hint:** The harmonic number $\mathrm{H}_n \leq \log(n) + 1$.
7. For context, what is the maximum number of comparisons that the algorithm can make if it is most unlucky in the choice of pivots?
8. **(*)** Implement RandQuicksort. For each $n = 2^i$ with $8 \leq i \leq 16$, generate a list of n distinct numbers. For each fixed choice of n, run randomized quicksort 100 times. Plot a histogram of the number of comparisons made. Compute the empirical mean $m(n)$ and standard deviation $s(n)$ of the number of comparisons. Now, as a function of n, plot error bands $m(n) \pm s(n)$. Compare against the trend $f(n) = 2n \log n$. A log–log scale is appropriate here.

Notes

The treatment of convex functions follows Gruber [Gru07, Chap. 1]. For more about convex functions, see books of Boyd & Vandenberghe [BV04] and Rockafellar [Roc70]. The treatment of Jensen's inequality is adapted from Williams's book [Wil91]. For more learning about heavy-tailed random variables, see the book of Nair, Wierman, & Zwart [NWZ22].

Some of the problems in this lecture are drawn from Alon & Spencer [AS16], from Grimmett & Stirzaker [GS01], and from Steele [Ste04]. For an overview of randomized algorithms in computer science, including randomized quicksort, see the book of Motwani & Raghavan [MR95].

10. Moments & Tails

Agenda:

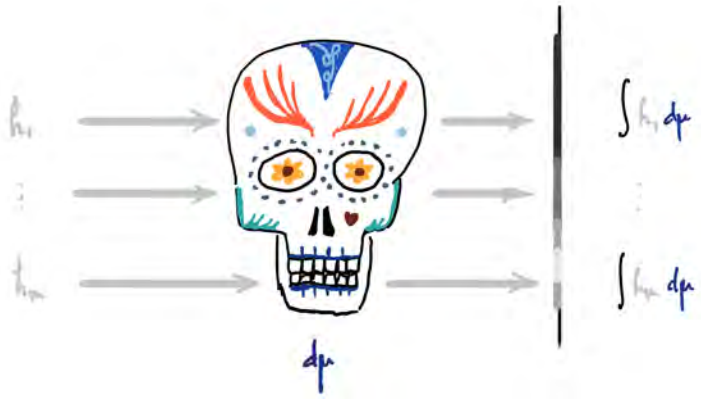
1. Moments
2. Tails
3. Markov's inequality
4. Integration by parts
5. *Duality

" 'Hallo, Pooh,' [Owl] said. 'How's things?'
'Terrible and Sad,' said Pooh, 'because Eeyore, who is a friend of mine, has lost his tail. And he's Moping about it. So could you very kindly tell me how to find it for him?'
'Well,' said Owl, 'the customary procedure in such cases is as follows.'
'What does Crustimoney Proseedcake mean?' said Pooh. 'For I am a Bear of Very Little Brain, and long words Bother me.' "

—*Winnie-The-Pooh*, A. A. Milne

A real random variable induces a distribution of probability on the real line. This distribution is described by its law, which is a Borel probability measure. Equivalently, the law is captured by its (cumulative) distribution function. These are complicated objects. The law assigns a probability to every Borel set, while the distribution function is a function on the entire real line. How can we acquire information about how the probability is distributed?

The basic idea is that we can collect data about a probability distribution by integrating it against test functions. This is a form of tomography:



Indeed, we can think about a skeleton as a distribution of bone mass in space. An X-ray machine sends a beam of particles from a source through the skeleton onto an exposure screen. The attenuation of the beam depends on the total amount of mass that it passes through along the way. In other words, we can model the intensity at a point on the exposure screen in terms of a line integral of the distribution.

In this lecture, we will see that we can perform a similar operation to collect information about the law of a random variable. To do so, we compute the expectations of functions of the random variable. Every such expectation provides a piece of data about the law of the random variable. In particular, we will see that certain expectations control the probability that the random variable takes large values.

10.1 Moments

We formalize this discussion using the concept of a moment of a random variable.

Definition 10.1 (Moment). Let X be a real random variable with law μ_X. A *moment* of the random variable X is the integral of a test function h against the law μ_X. Each moment takes the form

$$\mathbb{E}[h(X)] = \int_{\mathbb{R}} h(x)\, \mu_X(\mathrm{d}x) \quad \text{for measurable } h : \mathbb{R} \to \mathbb{R}. \tag{10.1}$$

In case $h \geq 0$, we allow the moment to take the value $+\infty$. For signed h, we require that h is μ_X-integrable.

Note that the formula (10.1) for the moment depends on Proposition 9.4, the law of the unconscious statistician. Section 10.4 offers a more refined interpretation of a moment as a linear functional of the law of the random variable.

10.1.1 Examples

Each moment provides a piece of information about the law μ_X of the random variable. The more moments we collect, the more data we have. Here are some examples of moments and what we can do with them.

Example 10.2 (Indicators). Let $h = \mathbb{1}_{\mathsf{B}}$ for a Borel set $\mathsf{B} \in \mathcal{B}(\mathbb{R})$. The associated moment

$$\mathbb{E}[\mathbb{1}_{\mathsf{B}}(X)] = \int_{\mathbb{R}} \mathbb{1}_{\mathsf{B}}(x)\, \mu_X(\mathrm{d}x) = \mu_X(\mathsf{B}) = \mathbb{P}\{X \in \mathsf{B}\}.$$

In other words, the moment reports the probability that the random variable takes a value in the Borel set B. On the Borel set B, the average amount of probability per unit length is $c = \mu_X(\mathsf{B})/\lambda(\mathsf{B})$. The number c provides a coarse approximation of the distribution over the set B. ∎

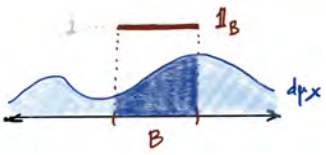


Figure 10.1 (Moment of an indicator).

Example 10.3 (Intervals). By extending Example 10.2, we can see that moments of intervals allow us to produce a piecewise constant approximation to the law of the random variable. For example, consider the function $h_n = \mathbb{1}_{[n,n+1)}$ for $n \in \mathbb{Z}$. Thus,

$$c_n := \mathbb{E}[\mathbb{1}_{[n,n+1)}(X)] = \mathbb{P}\{X \in [n, n+1)\}.$$

Using this information, we can form an approximation f of the law:

$$f := \sum\nolimits_{n \in \mathbb{N}} c_n \mathbb{1}_{[n,n+1)}.$$

On each interval, the value $f(n) = c_n$ of the approximation equals the probability that $X \in [n, n+1)$. The piecewise constant approximation of the law is usually called a *histogram*. ∎

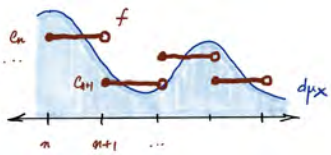


Figure 10.2 (Approximation via moments).

The most important classical moment is the first moment, or the center of mass.

Example 10.4 (First moment). The moment associated with the identity function $h(x) = x$ is often called the *first moment* of the random variable:

$$m_1 = \mathbb{E}[X] = \int_{\mathbb{R}} x\, \mu_X(\mathrm{d}x).$$

Note that the first moment may not be defined if $x \mapsto x$ is not μ_X-integrable.

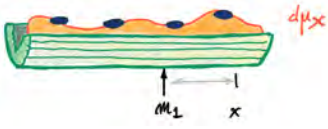


Figure 10.3 (Ants on a log).

The first moment has a mechanical interpretation. It is the point where we need to place a fulcrum to balance the distribution of mass. Indeed, the local mass $\mu_X(\mathrm{d}x)$ at a point x induces torque $(x - m_1)\,\mu_X(\mathrm{d}x)$ around the point m_1. The system balances because the total torque at m_1 is zero:

$$\int_{\mathbb{R}} (x - m_1)\,\mu_X(\mathrm{d}x) = \int_{\mathbb{R}} x\,\mu_X(\mathrm{d}x) - m_1 = m_1 - m_1 = 0.$$

In sequence, we have used the linearity of expectation, the fact that the expectation reproduces constants, and the definition of the first moment. ∎

10.1.2 Moments in probability theory

There are a number of other moments that play an important role in probability theory.

Example 10.5 (nth polynomial moment). For a natural number $n \in \mathbb{N}$, the nth *polynomial moment* of a real random variable X is

$$\mathbb{E}[X^n] = \int_{\mathbb{R}} x^n\,\mu_X(\mathrm{d}x).$$

This moment may not be defined if the Lebesgue integral does not exist. It is common to refer the the nth polynomial moment simply as the *nth moment* of the random variable. ∎

Example 10.6 (pth absolute polynomial moment). For a real number $p > 0$, the pth *absolute polynomial moment* of a real random variable X is

$$\mathbb{E}[|X|^p] = \int_{\mathbb{R}} |x|^p\,\mu_X(\mathrm{d}x).$$

This moment is always defined since $|X|^p \geq 0$, but it may take the value $+\infty$. ∎

Example 10.7 (Exponential moment). For a parameter $\theta \in \mathbb{R}$, the *exponential moment* of a real random variable X is

$$\mathbb{E}[\mathrm{e}^{\theta X}] = \int_{\mathbb{R}} \mathrm{e}^{\theta x}\,\mu_X(\mathrm{d}x).$$

This moment is always defined since $\mathrm{e}^{\theta X} \geq 0$, but it may take the value $+\infty$. ∎

You may have encountered the exponential moment in the guise of the *moment generating function*, which we will discuss later. The following exercise justifies this terminology.

Exercise 10.8 (Moment generating function). Let X be a *bounded* real random variable. For all $\theta \in \mathbb{R}$, prove that

$$\mathbb{E}[\mathrm{e}^{\theta X}] = \sum_{n=0}^{\infty} \frac{\theta^n}{n!} \cdot \mathbb{E}[X^n].$$

In combinatorics, the right-hand side is called the exponential generating function of the polynomial moments. **Hint:** Use dominated convergence.

There are some other fundamental classes of moments that are used to characterize the distribution of a real random variable. These examples will not play a central role in this class, but you may wish to be aware of the definitions.

Example 10.9 (*Characteristic function). The *characteristic function* of a real random variable X is a complex-valued function $\chi_X : \mathbb{R} \to \mathbb{C}$ on the real line. It is defined as

$$\chi_X(\theta) := \mathbb{E}[e^{i\theta X}] := \int_{\mathbb{R}} e^{i\theta x}\, \mu_X(dx)$$
$$:= \int_{\mathbb{R}} \cos(\theta x)\, \mu_X(dx) + i \int_{\mathbb{R}} \sin(\theta x)\, \mu_X(dx).$$

The characteristic function $\chi_X(\theta)$ is defined and finite for all $\theta \in \mathbb{R}$ because the sine and cosine functions are bounded and measurable. The characteristic function describes the global "frequency content" of the law μ_X at real frequencies $\theta \in \mathbb{R}$. Like the distribution function F_X, the characteristic function contains enough information to determine the law μ_X. We postpone a full discussion until Lecture 21. ▪

Example 10.10 (*Stieltjes transform). The *Stieltjes transform* of a real random variable X is a complex-valued function $G_X : \mathbb{C} \to \mathbb{C}$ on the complex plane. It is defined as

$$G_X(z) := \mathbb{E}[(X - z)^{-1}] := \int_{\mathbb{R}} (x - z)^{-1}\, \mu_X(dx) \quad \text{for } z \in \mathbb{C}.$$

The Stieltjes transform is finite for all $z \in \mathbb{C}$ whose imaginary part is nonzero. To understand this function, observe that its imaginary part satisfies

$$\frac{1}{\pi} \operatorname{Im} G_X(s + i\eta) = \frac{1}{\pi} \int_{\mathbb{R}} \frac{\eta}{(s - x)^2 + \eta^2}\, \mu_X(dx) \quad \text{for } s, \eta \in \mathbb{R}.$$

The integrand is a Cauchy density centered at s with scale η, so the integral is a convolution of this Cauchy distribution with the law μ_X. In other words, you can think about the Stieltjes transform as a family of smoothed versions of the law. The Stieltjes transform plays an important role in random matrix theory, but we will not discuss it further. ▪

> **Aside:** Some authors reserve the term "moment" specifically for polynomial moments. Throughout this course, we use the more general Definition 10.1.

10.1.3 Tails

Among many questions we may ask about a random variable, we can investigate the probability that it takes a large value.

Definition 10.11 (Tails). Let X be a real random variable.

- The *right tail probability* at level $t \in \mathbb{R}$ is $\mathbb{P}\{X \geq t\}$.
- The *left tail probability* at level $t \in \mathbb{R}$ is $\mathbb{P}\{X \leq -t\}$.
- The *tail probability* at level $t \in \mathbb{R}_+$ is $\mathbb{P}\{|X| \geq t\}$.

The right tail probability at level t is the moment associated with the indicator $\mathbb{1}_{[t,+\infty)}$. Similar interpretations apply to the other tail probabilities. See Figure 10.4.

Example 10.12 (Tails: Earthquakes). The right tail $\mathbb{P}\{X \geq t\}$ describes the probability that the magnitude exceeds a level t. Let X be a random variable that models the magnitude of an earthquake in Southern California, measured on the Richter scale.

For both scientists and insurance conglomerates, it is a matter of significant interest to understand how the tail probability decays as the level t increases. Do the tail

The Richter–Gutenberg scale for measuring the magnitude of an earthquake was invented by Charles Francis Richter and Beto Gutenberg in 1935 as part of a project sponsored by the Caltech Seismology laboratory. Richter became Professor at Caltech in 1952.

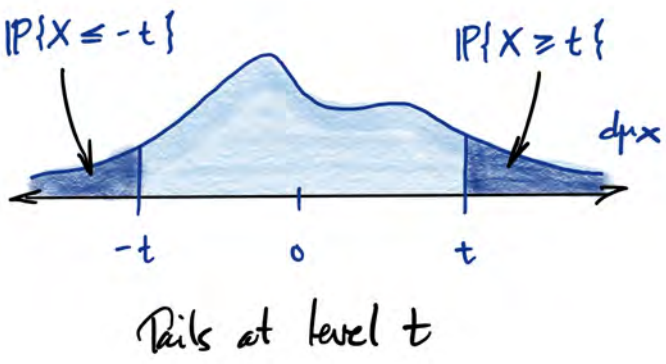


Figure 10.4 (Tails). The tails of a random variable capture the probability that the random variable takes on a large value.

probabilities decay exponentially, so that the probability of seeing an earthquake with large magnitude is exceptionally rare? Or do the tail probabilities decay polynomially (according to a power law), so that the probability of a large-magnitude earthquake is significant? ▪

The object of today's lecture is to establish a near-equivalence between polynomial decay of tail probabilities and the size of absolute polynomial moments:

1. Polynomial moments control tail decay.
2. Tail decay controls polynomial moments.

These claims depend on some very important tools that have wide application in probability theory. We establish the first statement in Section 10.2.1 as a consequence of Markov's inequality. We establish the second statement in Section 10.3.1 as a consequence of the integration by parts formula.

10.2 From moments to tails

In this section, we will show how to use moment information to extract information about tail decay.

10.2.1 Markov's inequality

There is a very simple technique for bounding the right tail probability of a random variable using the expectation of its positive part.

Theorem 10.13 (Markov's inequality). For every real random variable X, the right tail probability satisfies

$$\mathbb{P}\{X \geq t\} \leq \frac{\mathbb{E}[X_+]}{t} \quad \text{for all } t > 0.$$

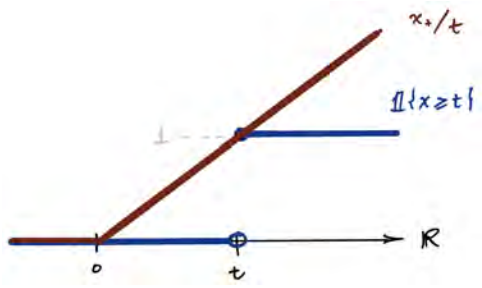
Note that the right-hand side is always defined, but may equal $+\infty$.

Markov's inequality is often stated for *positive* random variables, in which case it is redundant to take the positive part.

The result is named after A. A. Markov, although it was discovered by his adviser P. L. Chebyshev. The terminology, however, has become standard. Markov's inequality

is what a Russian mathematician would call a "trivial but useful observation", which is considered high praise.

Proof. The idea behind the proof of Markov's inequality is best captured with a picture:



In words, the indicator $\mathbb{1}\{X \geq t\}$ is dominated by the hinge function X_+/t.

More formally, let us fix a strictly positive level $t > 0$. We calculate that

$$\mathbb{P}\{X \geq t\} = \mathbb{E}[\mathbb{1}\{X \geq t\}] \leq \mathbb{E}[X_+/t] = \mathbb{E}[X_+]/t.$$

We have used the basic property that the expectation of an indicator coincides with the probability of the event. The second bound follows from the monotonicity of expectation and the fact that the indicator is bounded by the linear function. Last, we use the linearity of expectation. ∎

Markov's inequality can be used directly, but its full power arises when we apply it to (monotone) transformations of a random variable. This result bounds tail probabilities in terms of moments.

Corollary 10.14 (Markov's inequality). Let X be a random variable, and let $\varphi : \mathbb{R} \to \mathbb{R}_+$ be an *increasing, positive* function. Then

$$\mathbb{P}\{X \geq t\} \leq \frac{\mathbb{E}[\varphi(X)]}{\varphi(t)} \quad \text{when } \varphi(t) > 0.$$

Proof. Since the function φ is increasing, we have the containment of events

$$\{X \geq t\} \subseteq \{\varphi(X) \geq \varphi(t)\} \quad \text{for each } t \in \mathbb{R}.$$

By monotonicity of the probability measure,

$$\mathbb{P}\{X \geq t\} \leq \mathbb{P}\{\varphi(X) \geq \varphi(t)\} \quad \text{for each } t \in \mathbb{R}.$$

Apply Theorem 10.13 to the (positive) random variable $\varphi(X)$ at any strictly positive level $\varphi(t) > 0$. ∎

Exercise 10.15 (Markov: Extreme examples). Find a nontrivial random variable for which Markov's inequality holds with equality. **Hint:** Look closely at the graphical proof.

10.2.2 Polynomial moments control tail decay

We will use a particular instance of Markov's inequality. Let X be any real random variable. For all $p > 0$,

$$\mathbb{P}\{|X| \geq t\} \leq \frac{\mathbb{E}|X|^p}{t^p} \quad \text{for all } t > 0. \tag{10.2}$$

This result follows when we apply Corollary 10.14 to the positive random variable $|X|$ with the increasing function $\varphi : t \mapsto (t_+)^p$.

Let us reinterpret the inequality (10.2):

If $\mathbb{E}\,|X|^p < +\infty$, then the tail probability $\mathbb{P}\{|X| \geq t\}$ decays at least as fast as $\text{Const} \cdot t^{-p}$ as $t \to \infty$.

In other words, the absolute polynomial moments of a random variable give upper bounds on the rate at which the tail probability decays. In fact, the converse of this statement is almost true. This is the object of the next section.

10.3 From tails to moments

In the last section, we used Markov's inequality to bound a tail probability in terms of a polynomial moment. In this section, we will show how to bound a polynomial moment in terms of the tail probability.

10.3.1 Integration by parts

The key to this argument is the integration by parts formula. This result expresses moments of a random variable in terms of tail probabilities.

Theorem 10.16 (Integration by parts). Let X be a *positive* real random variable. Let $\varphi : \mathbb{R}_+ \to \mathbb{R}$ be an increasing, continuously differentiable function. Then

$$\mathbb{E}[\varphi(X)] = \varphi(0) + \int_{\mathbb{R}_+} \mathbb{P}\{X \geq t\}\, \varphi'(t)\, \lambda(\mathrm{d}t).$$

We can remove the assumption that φ is increasing, provided that the integral on the right-hand side is defined.

Proof. This result is an easy exercise. See Problem 6.26 for a precursor to this result. We will provide a complete proof because of its importance for us.

We can compute the expectation of $\mathbb{E}[\varphi(X)]$ via the law of the unconscious statistician (Proposition 9.4). As usual, we write μ_X for the law of X. Then

$$\begin{aligned}
\mathbb{E}[\varphi(X)] = \int_{\mathbb{R}} \varphi(x)\, \mu_X(\mathrm{d}x) &= \varphi(0) + \int_{\mathbb{R}} [\varphi(x) - \varphi(0)]\, \mu_X(\mathrm{d}x) \\
&= \varphi(0) + \int_{\mathbb{R}} \left[\int_{[0,x]} \varphi'(t)\, \lambda(\mathrm{d}t) \right] \mu_X(\mathrm{d}x) \\
&= \varphi(0) + \int_{\mathbb{R}} \left[\int_{\mathbb{R}_+} \mathbb{1}_{\{t \leq x\}}\, \varphi'(t)\, \lambda(\mathrm{d}t) \right] \mu_X(\mathrm{d}x) \\
&= \varphi(0) + \int_{\mathbb{R}_+} \left[\int_{\mathbb{R}} \mathbb{1}_{\{x \geq t\}}\, \mu_X(\mathrm{d}x) \right] \varphi'(t)\, \lambda(\mathrm{d}t) \\
&= \varphi(0) + \int_{\mathbb{R}_+} \mu_X\{X \geq t\}\, \varphi'(t)\, \lambda(\mathrm{d}t).
\end{aligned}$$

The notation for the indicator functions is abbreviated for legibility.

In the first line, we have added and subtracted $\varphi(0)$, using the linearity of the integral and the fact that $\mu_X(\mathbb{R}) = 1$. To pass to the second line, we apply the fundamental theorem of calculus. We introduce an indicator function to represent the domain $[0, x]$ of integration for the variable t. Next, we invoke Fubini–Tonelli (Theorem 6.23) to interchange the integrals, which is justified when φ' is positive or the resulting integral is finite. Finally, we rewrite the indicator function in terms of the variable x and use the fact that the integral of an indicator is the measure of the set. ∎

Observe that the special case $\varphi(x) = x$ coincides with the definition of the expectation of a positive random variable:

$$\mathbb{E}[X] = \int_{\mathbb{R}_+} \mathbb{P}\{X \geq t\} \, \lambda(\mathrm{d}t).$$

There is a typographical difference between this expression and the definition (9.1) of the expectation, owing to the change from a strict inequality ($>$) to a weak inequality ($\geq$) in the tail. In fact, both expressions are equivalent because the Lebesgue measure is insensitive to the values of the integrand on singletons. Have a close look at the proof to confirm this point.

Problem 10.17 (*Integration by parts: Without derivatives). Find an extension of the integration by parts formula (Theorem 10.16) that holds when $\varphi : \mathbb{R}_+ \to \mathbb{R}_+$ is positive and increasing, but not necessarily differentiable. **Hint:** Consider the Borel measure ν on $\mathbb{R}_+$ that satisfies $\nu([0, t]) = \varphi(t)$ for $t \geq 0$.

10.3.2 From tail decay to polynomial moments

We will use another consequence of integration by parts. For an arbitrary real random variable X and a positive number $q > 0$, Theorem 10.16 implies that

$$\mathbb{E}[|X|^q] = \int_{\mathbb{R}_+} \mathbb{P}\{|X| \geq t\} \cdot q t^{q-1} \, \lambda(\mathrm{d}t). \tag{10.3}$$

This statement follows by considering the positive random variable $|X|$ and the increasing function $\varphi : t \mapsto t^q$. It is possible that both sides of this expression equal $+\infty$ in case the tail probability decays slowly.

Now, suppose that X is a random variable whose tail probability satisfies

$$\mathbb{P}\{|X| \geq t\} \leq \text{Const} \cdot t^{-p} \quad \text{for all } t > 0. \tag{10.4}$$

Let us calculate the qth absolute moment, assuming that $q < p$.

$$\begin{aligned}
\mathbb{E}\,|X|^q &= \int_{\mathbb{R}_+} \mathbb{P}\{|X| \geq t\} \cdot q t^{q-1} \, \mathrm{d}t \\
&= \int_{[0,1]} \mathbb{P}\{|X| \geq t\} \cdot q t^{q-1} \, \mathrm{d}t + \int_{(1,+\infty)} \mathbb{P}\{|X| \geq t\} \cdot q t^{q-1} \, \mathrm{d}t \\
&\leq \int_{[0,1]} 1 \cdot q t^{q-1} \, \mathrm{d}t + \int_{(1,+\infty)} \text{Const} \cdot t^{-p} \cdot q t^{q-1} \, \mathrm{d}t \\
&= 1 + \frac{\text{Const} \cdot q}{p - q}.
\end{aligned}$$

The first relation is (10.3). In the second, we split the domain of integration at $t = 1$. In the first integral, we use the trivial bound $\mathbb{P}\{|X| \geq t\} \leq 1$; in the second integral, we use the assumption (10.4) about the tail decay. To evaluate the integrals, we use standard calculus (antiderivatives and the fundamental theorem of calculus).

Let us reinterpret the computation in the last paragraph.

> If the tail probability $\mathbb{P}\{|X| \geq t\}$ decays as least as fast as $\text{Const} \cdot t^{-p}$ as $t \to \infty$, then the qth absolute moment $\mathbb{E}\,|X|^q < +\infty$ for all $q < p$.

In other words, control on the rate of tail decay implies that some absolute polynomial moments are finite.

Along with the discussion in Section 10.2.1, we see that polynomial tail decay at rate t^{-p} is almost equivalent to the moment $\mathbb{E}\,|X|^p$ being finite. We will build on this insight in the next lecture, on L_p spaces.

Another way to understand these facts is to think about the alternative representation of the absolute moment as an integral:

$$\mathbb{E}\,|X|^p = \int_{\mathbb{R}} |x|^p \, \mu_X(\mathrm{d}x).$$

The integral is finite precisely when the tail probability $t \mapsto \mu_X\{|x| \geq t\}$ of the random variable decays fast enough to counteract the growth of the polynomial function $x \mapsto |x|^p$ as $|x| \to \infty$. We need Markov's inequality (10.2) and the integration by parts formula (10.3) to make this intuition rigorous.

10.4 *Duality between functions and measures

We conclude with some additional context for the concept of a moment. Later, these ideas will resurface when we talk about how to define distances between probability distributions.

10.4.1 A measure induces a linear functional on functions

Consider a measurable space $(\mathsf{X}, \mathcal{F})$. In our study of integration, we have seen that each (finite) measure μ on $\mathcal{F}$ defines a real-valued functional on the linear space of measurable, μ-integrable functions:

$$\langle \mu,\ \cdot \rangle : h \mapsto \int_{\mathsf{X}} h \, \mathrm{d}\mu \quad \text{for } \mu\text{-integrable } h : \mathsf{X} \to \mathbb{R}.$$

This notation is similar to our notation for integrals, $\mu(h) := \int_{\mathsf{X}} h \, \mathrm{d}\mu$. Theorem 5.14 shows that the integral is a *linear* functional on the class of μ-integrable functions. Using the bracket, we can write

$$\langle \mu,\ \alpha g + \beta h \rangle = \alpha \langle \mu,\ g \rangle + \beta \langle \mu,\ h \rangle \quad \text{for } \alpha, \beta \in \mathbb{R} \text{ and } f, g \in \mathsf{L}_1(\mu).$$

Note that the class $\mathsf{L}_1(\mu)$ of integrable functions depends on the measure μ, so the linearity property is not valid for the same functions for all measures.

10.4.2 A function induces a positive-linear functional on measures

Dually, a measurable function induces a real-valued functional on finite measures. Formally, a measurable function $h : \mathsf{X} \to \mathbb{R}$ defines a map

$$\langle \cdot,\ h \rangle : \mu \mapsto \int_{\mathsf{X}} h \, \mathrm{d}\mu \quad \text{for finite Borel measures } \mu : \mathcal{F} \to \mathbb{R}_+.$$

This construction requires more thought, however, because the integral may not be defined for all such measures.

To that end, we need to restrict our attention to a smaller class of measurable functions. Consider the linear space

$$\mathsf{C}_b := \mathsf{C}_b(\mathsf{X}; \mathbb{R}) := \{h : \mathsf{X} \to \mathbb{R} \text{ bounded and measurable}\}.$$

The notation C_b is temporary, and it is at variance with standard notations!

For every function $h \in \mathsf{C}_b(\mathsf{X}; \mathbb{R})$, we can reliably define $\mu(h)$ for every finite Borel measure μ on $\mathcal{F}$. (Why?) In this case, $\mu(h)$ must take a finite value.

As we saw in Exercise 8.29, measures are just functions that take positive real values, so we can scale them by positive numbers and add them. For finite Borel measure $\mu, \nu : \mathcal{F} \to \mathbb{R}_+$,

$$(\alpha\mu + \beta\nu)(\mathsf{E}) = \alpha\mu(\mathsf{E}) + \beta\nu(\mathsf{E}) \quad \text{for } \alpha, \beta \geq 0 \text{ and } \mathsf{E} \in \mathcal{F}.$$

Furthermore, we can extend this relation from sets to functions. Using the bracket, we can write

$$\langle \alpha\mu + \beta\nu,\ h \rangle = \alpha\langle \mu,\ h \rangle + \beta\langle \nu,\ h \rangle \quad \text{for } \alpha, \beta \geq 0 \text{ and } h \in \mathsf{C}_b(\mathsf{X}; \mathbb{R}).$$

In other words, a function $h \in \mathsf{C}_b(\mathsf{X}; \mathbb{R})$ induces a positive-linear functional $\langle \cdot,\ h \rangle$ on finite Borel measures.

> **Aside:** If the test function h is positive, we can even drop the requirement that it is bounded. In this case, the functional $\langle \cdot,\ h \rangle$ on measures takes only positive values, including perhaps $+\infty$.

10.4.3 The linear space of signed measures

For many purposes, the results in the last paragraph are sufficient. Nevertheless, they are not entirely satisfactory because they tempt us to form general linear combinations of measures, such as $\alpha\mu + \beta\nu$ for real numbers $\alpha, \beta \in \mathbb{R}$. This is problematic because the resulting object may assign *negative* values to some measurable sets.

To patch this hole, we must generalize the definition of a measure to allow it to take positive and negative values.

Definition 10.18 (Signed measure). Let $(\mathsf{X}, \mathcal{F})$ be a measurable space. A *signed measure* is a function $\mu : \mathcal{F} \to \mathbb{R}$ taking finite real values and with the properties

1. **Empty set:** $\mu(\emptyset) = 0$.
2. **Countable additivity:** For each sequence $(\mathsf{A}_i \in \mathcal{F} : i \in \mathbb{N})$ of *disjoint* measurable sets,
$$\mu\big(\dot{\bigcup}_{i=1}^{\infty} \mathsf{A}_i\big) = \sum_{i=1}^{\infty} \mu(\mathsf{A}_i).$$

A signed measure that only takes positive values is called a *positive measure*.

> **Warning:** A signed measure can only take finite values. ■

We require the series on the right-hand side to converge absolutely.

A positive measure is what we formerly called a finite measure.

A signed measure describes a distribution of mass on measurable sets, but it can place negative mass on some sets. It may be helpful to think about a signed measure as a model for a distribution of electric charge, since charge can be positive or negative.

It is easy to see that the signed measures form a (real) linear space under the usual rules for scaling and addition of functions:

$$(\alpha\mu + \beta\nu)(\mathsf{E}) = \alpha\mu(\mathsf{E}) + \beta\nu(\mathsf{E}) \quad \text{for } \alpha, \beta \in \mathbb{R} \text{ and } \mathsf{E} \in \mathcal{F}.$$

We introduce notation for the linear space of signed measures:

$$\mathsf{M}_1 := \mathsf{M}_1(\mathsf{X}; \mathbb{R}) := \{\mu : \mathcal{F} \to \mathbb{R} \text{ is a signed measure}\}.$$

This notation is fairly common, but not universal.

What does a signed measure look like? The answer is quite simple.

Theorem 10.19 (Hahn–Jordan). Let μ be a signed measure on measurable space $(\mathsf{X}, \mathscr{F})$. Then the signed measure can be decomposed as $\mu = \mu_+ - \mu_-$ where both μ_+ and μ_- are (finite) positive measures on $(\mathsf{X}, \mathscr{F})$.

Problem 10.20 (Hahn–Jordan decomposition). Prove Theorem 10.19. **Hint:** Construct the "maximal" set $\mathsf{P} \in \mathscr{F}$ for which $\mu(\mathsf{P}) \geq 0$. Define

$$\mu_+(\mathsf{E}) := \mu(\mathsf{E} \cap \mathsf{P}) \quad \text{and} \quad \mu_-(\mathsf{E}) := \mu(\mathsf{E} \cap \mathsf{P}^c) \quad \text{for all measurable } \mathsf{E} \in \mathscr{F}.$$

The decomposition is essentially unique in the sense that any other maximal positive set $\mathsf{P}' \in \mathscr{F}$ has the property that $\mu(\mathsf{P} \triangle \mathsf{P}') = 0$.

With this fact at hand, we can define the integral with respect to a signed measure μ. For all $h \in \mathsf{C}_b(\mathsf{X}; \mathbb{R})$,

$$\int_{\mathsf{X}} h \, \mathrm{d}\mu := \mu(h) := \mu_+(h) - \mu_-(h),$$

where $\mu_\pm$ are the positive and negative parts of the signed measure μ provided by the Hahn–Jordan decomposition. Since the decomposition is unique up to negligible sets, the integral is well defined.

It is also easy to check that the integral with respect to a signed measure $\mu \in \mathsf{M}_1(\mathsf{X}; \mathbb{R})$ is a linear function of the integrand:

$$\mu(\alpha g + \beta h) = \alpha \mu(g) + \beta \mu(h) \quad \text{for all } \alpha, \beta \in \mathbb{R} \text{ and } g, h \in \mathsf{C}_b(\mathsf{X}; \mathbb{R}).$$

Dually, for a function $h \in \mathsf{C}_b(\mathsf{X}; \mathbb{R})$,

$$(\alpha \mu + \beta \nu)(h) = \alpha \mu(h) + \beta \nu(h) \quad \text{for all } \alpha, \beta \in \mathbb{R} \text{ and } \mu, \nu \in \mathsf{M}_1(\mathsf{X}; \mathbb{R}).$$

You should verify that this statement is correct.

In other words, we can define a *duality pairing*:

$$\langle \mu, \, h \rangle := \mu(h) := \int_{\mathsf{X}} h \, \mathrm{d}\mu \qquad \begin{array}{l} \text{where } \mu \in \mathsf{M}_1(\mathsf{X}; \mathbb{R}) \text{ is a signed measure;} \\ \text{where } h \in \mathsf{C}_b(\mathsf{X}; \mathbb{R}) \text{ is a function.} \end{array}$$

For each signed measure, $\langle \mu, \, \cdot \rangle$ is a linear functional on functions. For each function, $\langle \cdot, \, h \rangle$ is a linear functional on signed measures.

Aside: This presentation describes an algebraic duality of linear spaces. To extend to a topological duality, we would need to equip the spaces with a notion of convergence.

10.4.4 Perspective

We now have a more complete appreciation for the idea that a moment is a linear functional of the law of a random variable. Indeed, with our new notation, a moment takes the form $\langle \cdot, \, h \rangle$ for a function h. We have also seen that this construction is dual to the fact that an integral with respect to a (signed) measure μ is a linear functional $\langle \mu, \, \cdot \rangle$ that acts on functions. Both of these perspectives are fundamental.

First, for a fixed *positive* measure μ, we can use the integral with respect to the measure μ to define a distance on μ-integrable functions $g, h \in \mathsf{L}_1(\mu)$. For example,

As before, we warn the reader that this is a pseudonorm, not a norm.

$$\|g - h\|_{\mathsf{L}_1(\mu)} := \int_{\mathbb{R}} |g - h| \, \mathrm{d}\mu.$$

In the next lecture, we will generalize this idea to define other kinds of distances between functions.

Second, by considering an appropriate class of test functions h, we can also use integrals with respect to test functions to define a distance between signed measures $\mu, \nu \in \mathsf{M}_1$. This requires some work, and we will take up this challenge in Lecture 17.

The alert reader may wonder why we have restricted our attention to the class C_b of bounded functions, even though the definition of a moment explicitly allows more general functions. The simple reason is that we need $h \in \mathsf{C}_b$ to define $\langle \mu, h \rangle$ for every signed measure $\mu \in \mathsf{M}_1$.

If we consider test functions h from a *larger* class, then we must restrict our attention to a *smaller* family of signed measures. This is precisely the point of the discussion in this lecture. If the test function h grows like $|x|^p$ as $|x| \to \infty$, then the associated moment $\int h \, d\mu_X$ is finite when the tail probability $\mu_X\{|x| \geq t\}$ decays at least as fast as $|t|^{-p}$. In other words, test functions that grow are paired with measures that decay at infinity at a complementary rate.

Problems

Exercise 10.21 (Normal tails). It is useful to have simple and accurate approximations for the upper tail of the normal random variable. Let $Z \sim \text{NORMAL}(0, 1)$.

1. Show that $\mathbb{P}\{Z \geq t\} \leq \frac{1}{2} e^{-t^2/2}$ for $t \geq 0$. **Hint:** Write the left-hand side as an integral. Maximize the difference between the left-hand side and right-hand side using calculus.
2. Show that $\mathbb{P}\{Z \geq t\} \leq \frac{1}{t\sqrt{2\pi}} e^{-t^2/2}$ for $t > 0$. **Hint:** Write the left-hand side as an integral, and introduce the extra factor $1 \wedge t$.

Exercise 10.22 (Gaussian IBP). This innocuous exercise is very important for some parts of probability theory. Let $Z \sim \text{NORMAL}(0, \sigma^2)$.

1. Let $f : \mathbb{R} \to \mathbb{R}$ be a bounded function with a bounded, continuous derivative. Prove the Gaussian integration by parts formula:
$$\mathbb{E}[Zf(Z)] = \sigma^2 \cdot \mathbb{E}[f'(Z)]. \tag{10.5}$$
Hint: This is really just ordinary integration by parts from calculus.
2. **(*)** Extend the formula (10.5) to all f with $f' \in \mathsf{L}_1(\gamma)$.
3. For $p \in \mathbb{N}$, evaluate $\mathbb{E}[Z^{2p}] = \mathbb{E}[Z \cdot Z^{2p-1}]$ by iterative application of Gaussian integration by parts.
4. Deduce that $(\mathbb{E}\,|Z|^p)^{1/p} \leq \text{Const} \cdot \sigma\sqrt{p}$ for all even $p = 2, 4, 6, \ldots$. **Hint:** Stirling.
5. **(*)** Extend the bound in the last part to all $p > 0$. **Hint:** Use the fact that homogeneous moments are increasing (Theorem 11.4).

Problem 10.23 (Reconstruction). If we collect enough moment information, then we can sometimes determine the distribution of a random variable completely.

1. Let X be a real random variable that takes values in $\{0, 1, 2, \ldots, n\}$. Suppose that we know $\mathbb{E}\,X^p$ for $p = 1, 2, \ldots, n$. Explain how to reconstruct the distribution of X. **Hint:** A moment is a *linear* functional of the law μ_X. When is a Vandermonde matrix nonsingular?
2. Continue with the assumptions from the previous part. Define the moment generating function $m_X(\theta) := \mathbb{E}\, e^{\theta X}$ for $\theta \in \mathbb{R}$. Explain how to reconstruct the distribution of X from the function m_X. **Hint:** Look at the derivatives at zero!

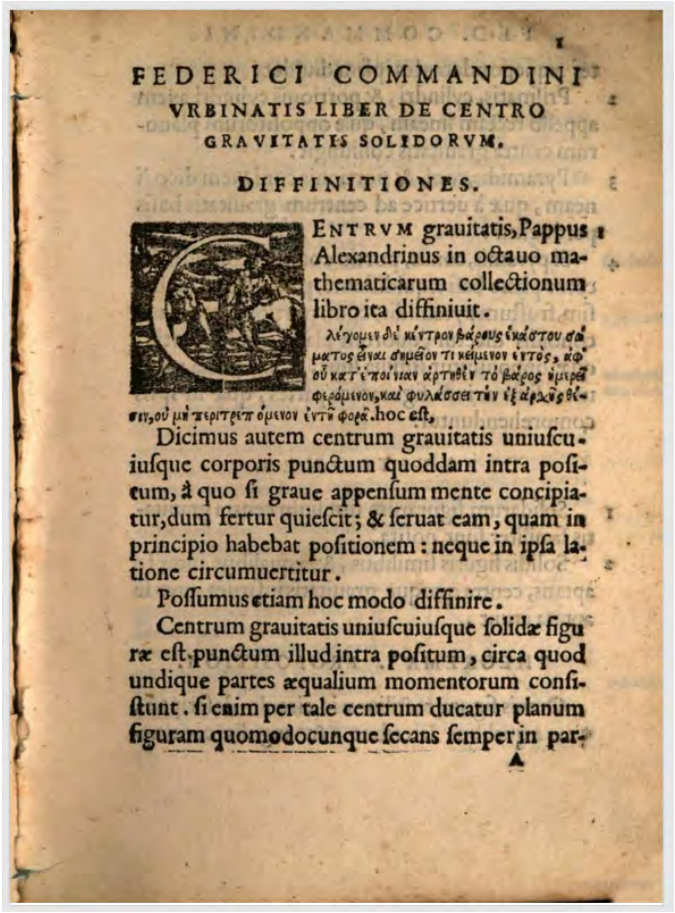
Exercise 10.24 (Moment growth and tails). The rate of tail decay can also be captured by the *growth* of polynomial moments. Here is an important example that arises in high-dimensional probability and geometry.

1. Let X be a real random variable whose homogeneous moments grow at a rate no faster than the square root. That is, assume there is a constant $C_1 > 0$ for which
$$(\mathbb{E}\,|X|^p)^{1/p} \le C_1\sqrt{p} \quad \text{for each } p > 0.$$
Develop a bound for $\mathbb{P}\{|X| \ge t\}$ using Markov's inequality with the best choice of the power p.
2. Conversely, suppose that X is a real random variable whose tail probability satisfies the bound
$$\mathbb{P}\{|X| \ge t\} \le C_2 \cdot \mathrm{e}^{-c_3 \cdot t^2} \quad \text{for all } t \ge 0.$$
The constants $C_2, c_3 > 0$ do not depend on t, but will depend on X. Using integration by parts, prove that the homogeneous moments of $|X|$ satisfy the bound in (a). **Hint:** Look up Euler's integral for the gamma function, and make a change of variables.
3. Explain why the random variables in (a)–(b) are called *subgaussian*. Give an example of a subgaussian random variable that is not a Gaussian (or a constant).
4. **(*)** Show that the equivalent conditions in (a) and (b) are also equivalent to a bound on the mgf:
$$m_{|X|}(\theta) := \mathbb{E}\,\mathrm{e}^{\theta|X|} \le C_3 \cdot \mathrm{e}^{C_4 \cdot \theta^2}.$$
The constants $C_3, C_4 > 0$ are independent of θ, but depend on X.
5. **(*)** Repeat parts (1) and (2) under the alternative assumption that
$$(\mathbb{E}\,|X|^p)^{1/p} \le Cp \quad \text{for each } p > 0.$$
These random variables are called *subexponential*. What is the analog of (4) in this case?

Notes

In the literature, there is some inconsistency in the definition of the term "moment", and we have opted for the most expansive definition. The definition of the term "tail" also varies somewhat, but the general idea is to capture the probability that a random variable takes large or small values. The material on Markov's inequality and integration by parts is standard fare for a probability course. The discussion of duality between measures and functions is adapted from the functional analysis literature; for example, see Rudin [Rud91].

The term "moment" appears to derive from a 1565 Latin translation of Archimedes by Federico Commandino; see Figure 10.5. Commandino writes, "The center of gravity of each solid figure is that point within it, about which, on all sides, parts of equal moment stand." Moment is being used in the sense of importance or "momentousness". Jeremy Bernstein tracked down this etymology while preparing notes for the 2019 implementation of this class.



FEDERICI COMMANDINI
VRBINATIS LIBER DE CENTRO
GRAVITATIS SOLIDORVM.

DIFFINITIONES.

CENTRVM grauitatis, Pappus Alexandrinus in octauo mathematicarum collectionum libro ita diffiniuit.

λέγομεν δὲ κέντρον βάρους ἑκάστου σώματος εἶναί σημεῖόν τι κείμενον ἐντὸς, ἀφ' οὗ κατ' ἐπίνοιαν ἀρτηθὲν τὸ βάρος ὑμιρεῖ φερόμενον, καὶ φυλάσσει τὴν ἐξ ἀρχῆς θέσιν, οὐ μὴ περιτρεπόμενον ἐντῇ φορᾷ. hoc est,

Dicimus autem centrum grauitatis uniuscuiusque corporis punctum quoddam intra positum, à quo si graue appensum mente concipiatur, dum fertur quiescit; & seruat eam, quam in principio habebat positionem: neque in ipsa latione circumuertitur.

Possumus etiam hoc modo diffinire.

Centrum grauitatis uniuscuiusque solidæ figuræ est punctum illud intra positum, circa quod undique partes æqualium momentorum consistunt. si enim per tale centrum ducatur planum figuram quomodocunque secans semper in par-

A

Figure 10.5 (Moment: Etymology). The term "moment" originates from Federico Commandino's 1565 book *Liber de centro gravitatis solidorum*. The sentence of interest is "Centrum gravitatis ... momentorum consistunt."

11. L_p Spaces

Agenda:

1. L_p spaces
2. Inequalities
3. Convergence
4. Completeness

> "My reputation is far bigger than my sales. . . I was talking to Lou Reed the other day, and he said that the first Velvet Underground record sold only 30,000 copies in its first five years. Yet, that was an enormously important record for so many people. I think everyone who bought one of those 30,000 copies started a band! So I console myself in thinking that some things generate their rewards in second-hand ways."
>
> —Brian Eno, qtd. in "Lots of aura, no air play" by Kristine McKenna, *Los Angeles Times*, 23 May 1982

In the last lecture, we saw that the polynomial moments of a random variable are closely related to the decay of tail probabilities. This observation leads us to define the space L_p, which consists of all random variables whose pth moment is finite.

These spaces play a central role in probability theory. We have already encountered the L_1 space of integrable random variables. As we will see in the next lecture, the L_2 space is very important because it allows us to place the notions of variance and covariance in an appropriate context. Other L_p spaces also arise from time to time, and it is valuable to understand their properties all at once.

To begin our study, we will use Jensen's inequality repeatedly to derive some basic inequalities that are used to understand the structure of the L_p spaces. Afterward, we will develop the a notion of convergence in L_p, and we will argue that L_p spaces are complete.

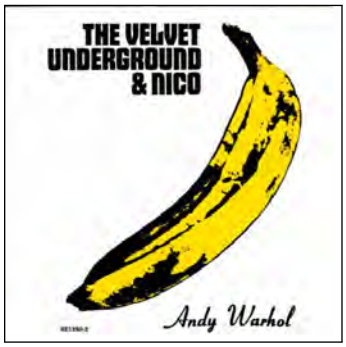


Figure 11.1 (An LP space). This is the first Velvet Underground record.

11.1 L_p spaces

In the last lecture, we proved that polynomial moments are related to the decay of tail probabilities. Let X be a real random variable, and let $p > 0$.

- If the polynomial moment $\mathbb{E}\,|X|^p < +\infty$, then the tail probability decays like $\mathbb{P}\{|X| \geq t\} \leq \text{Const} \cdot t^{-p}$ for all $t > 0$.
- If the tail probability decays like $\mathbb{P}\{|X| \geq t\} \leq \text{Const} \cdot t^{-p}$, then the polynomial moment $\mathbb{E}\,|X|^q < +\infty$ when $0 < q < p$.

In this section, we study the class of random variables whose pth moment is finite.

11.1.1 The space of p-integrable random variables

We begin by carving out some collections of random variables.

Definition 11.1 (L_p space). For $p > 0$, the space $L_p := L_p(\mathbb{P}) := L_p(\Omega, \mathcal{F}, \mathbb{P})$ is

$$L_p := \{X : \Omega \to \mathbb{R} \text{ is a random variable with } \mathbb{E}\,|X|^p < +\infty\}.$$

We call L_p the space of *p-integrable random variables* or the space of *random variables with p moments*.

We have already encountered the space L_1, which is now subsumed under Definition 11.1. By the considerations above, the random variables in L_p have tail decay with rate at least t^{-p}. Conversely, every random variable with tail decay t^{-q} for $q > p$ belongs to L_p. The most basic fact is that L_p is a linear space.

Proposition 11.2 (L_p is a linear space). For any probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and for all $p > 0$, the set $\mathsf{L}_p(\Omega, \mathcal{F}, \mathbb{P})$ becomes a linear space with the standard scalar multiplication and addition of functions.

Proof. Let $X, Y \in \mathsf{L}_p$. For $\alpha \in \mathbb{R}$, it is clear that $\alpha X \in \mathsf{L}_p$ because

$$\mathbb{E}\,|\alpha X|^p = |\alpha| \cdot \mathbb{E}\,|X|^p < +\infty.$$

Now, to verify that the sum $X + Y \in \mathsf{L}_p$, we first calculate that

$$|X + Y|^p \le \big(2 \max\{|X|, |Y|\}\big)^p = 2^p \max\{|X|^p, |Y|^p\} \le 2^p \cdot \big(|X|^p + |Y|^p\big).$$

This inequality holds pointwise. Using linearity, we take the expectation:

$$\mathbb{E}\,|X + Y|^p \le 2^p \cdot \big(\mathbb{E}\,|X|^p + \mathbb{E}\,|Y|^p\big) < +\infty.$$

Thus, L_p is stable under scalar multiplication and sums. It is a linear space. ∎

Aside: Many authors write L^p for the space of p-integrable random variables, with the power p in the superscript. Some mathematicians (including the ones who taught me) reserve superscripts in function spaces for differentiability properties, while subscripts reflect integrability properties. These notes follow the latter convention.

11.1.2 Homogeneous moments

It is productive to introduce a measure of the size of a random variable that belongs to L_p. To do so, we will adjust the pth moment to obtain a positively homogeneous functional.

Definition 11.3 (pth homogeneous moment). Let X be a real random variable. The pth homogeneous moment is

$$\|X\|_p := \|X\|_{\mathsf{L}_p} := (\mathbb{E}\,|X|^p)^{1/p}. \tag{11.1}$$

Warning: In spite of the notation, $\|\cdot\|_p$ is only a pseudonorm. See Corollary 11.13. ∎

For every real random variable X, this functional is *positively* homogeneous:

$$\|\alpha X\|_p = |\alpha| \cdot \|X\|_p \quad \text{for all } \alpha \in \mathbb{R}. \tag{11.2}$$

This result follows immediately from the linearity of expectation. We will collect some less trivial properties of the homogeneous moments and then present an omnibus result about the functional $\|\cdot\|_p$.

11.1.3 Monotonicity of homogeneous moments

Our next result describes the relationships among the homogeneous moments of different orders.

Theorem 11.4 (Homogeneous moments: Monotonicity). For $0 < p \le q$, we have the relation

$$\|X\|_p \le \|X\|_q \quad \text{for each random variable } X.$$

In particular, for $p \le q$, we have $\mathsf{L}_q \subseteq \mathsf{L}_p$.

Warning: The order of the norms is reverse to the order of inclusions. See Warning 11.22 for additional context. ■

In words, there are fewer random variables with tails that decay quickly than random variables with tails that decay slowly.

Proof. This result a direct consequence of Jensen's inequality. When $0 < p \le q$, the ratio $r := q/p \ge 1$. Recall that the function $\varphi(t) := |t|^r$ is convex for $r \ge 1$ and bounded below by zero. Theorem 9.26 implies that

$$\mathbb{E}\,|X|^q = \mathbb{E}\left[|X|^{p\cdot(q/p)}\right] \ge \left(\mathbb{E}\,|X|^p\right)^{q/p}.$$

This relation is valid, even when the left-hand side or both sides are infinite. Take the $1/q$ power to see that $\|X\|_q \ge \|X\|_p$. ■

11.1.4 Hölder's inequality

To better understand the behavior of the homogeneous moments, let us develop a bound for the homogeneous moments of a product.

Theorem 11.5 (Hölder's inequality). Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Consider positive numbers $p, q > 1$ with $p^{-1} + q^{-1} = 1$. For real random variables $X \in \mathsf{L}_p$ and $Y \in \mathsf{L}_q$, the product $XY \in \mathsf{L}_1$, and

$$\|XY\|_1 \le \|X\|_p \cdot \|Y\|_q < +\infty. \tag{11.3}$$

Hölder's inequality expresses a duality relation between L_p and L_q for dual indices $1/p + 1/q = 1$. We will give a standard proof based on a sequence of numerical inequalities.

Proof. We recall the GM–AM inequality from (9.11):

$$x^\tau y^{1-\tau} \le \tau x + (1-\tau)y \quad \text{for } \tau \in [0,1] \text{ and } x, y \ge 0.$$

This result is an instant consequence of the fact that $\exp(\cdot)$ is a convex function. By the change of variables $\tau = 1/p$ and $1 - \tau = 1/q$, we arrive at Young's inequality:

$$|xy| \le \frac{1}{p}\cdot|x|^p + \frac{1}{q}\cdot|y|^q \quad \text{for } x, y \in \mathbb{R}. \tag{11.4}$$

Hölder's inequality follows easily from this statement.

The key idea is to rescale the random variables X and Y to make full use of Young's inequality (11.4). Indeed, we have the pointwise inequality

$$\frac{|XY|}{\|X\|_p\|Y\|_q} \le \frac{1}{p}\cdot\frac{|X|^p}{\|X\|_p^p} + \frac{1}{q}\cdot\frac{|Y|^q}{\|Y\|_q^q}.$$

Since both sides are positive, we may take the expectation without further justification:

$$\begin{aligned}\mathbb{E}\left[\frac{|XY|}{\|X\|_p\|Y\|_q}\right] &\le \mathbb{E}\left[\frac{1}{p}\cdot\frac{|X|^p}{\|X\|_p^p} + \frac{1}{q}\cdot\frac{|Y|^q}{\|Y\|_q^q}\right] \\ &= \frac{1}{p}\cdot\frac{\mathbb{E}\,|X|^p}{\|X\|_p^p} + \frac{1}{q}\cdot\frac{\mathbb{E}\,|Y|^q}{\|Y\|_q^q} = 1.\end{aligned}$$

We have applied linearity of expectation and the definition (11.1) of the pth homogeneous moment. By assumption, $1/p + 1/q = 1$. To complete the argument, we invoke linearity of expectation on the left-hand side of the previous display, and we multiply through by $\|X\|_p \|Y\|_q$. ∎

Exercise 11.6 (Hölder's equality). Under what conditions does (11.3) hold with equality?

Exercise 11.7 (Geometric means: Concavity). Let X, Y be positive random variables. For each $\tau \in [0, 1]$, show that

$$\mathbb{E}[X^\tau Y^{1-\tau}] \le (\mathbb{E}\, X)^\tau \cdot (\mathbb{E}\, Y)^{1-\tau}.$$

Deduce that $(x, y) \mapsto x^\tau y^{1-\tau}$ is a concave function on $\mathbb{R}_+ \times \mathbb{R}_+$. **Hint:** Change variables in Hölder's inequality.

Exercise 11.8 (Moments: Log-Convexity). Derive *Lyapunov's inequality*. Let X be a positive random variable. For real parameters $0 < p \le r \le q < +\infty$,

$$\mathbb{E}\, X^r \le (\mathbb{E}\, X^p)^\tau \cdot (\mathbb{E}\, X^q)^{1-\tau} \quad \text{where} \quad r = \tau p + (1 - \tau) q.$$

Equivalently, the function $p \mapsto \log(\mathbb{E}\, X^p)$ is convex for $p > 0$.

(*) Deduce that the logarithm of Euler's gamma function is convex. **Hint:** Consider an exponential random variable.

11.1.5 Minkowski's inequality

Next, we consider how the homogeneous moments of a sum compare with the homogeneous moments of the summands.

Theorem 11.9 (Minkowski's inequality). Fix $p \ge 1$. Let $X, Y \in L_p$. Then

$$\|X + Y\|_p \le \|X\|_p + \|Y\|_p. \tag{11.5}$$

In other words, the pth homogeneous moment obeys the triangle inequality when $p \ge 1$.

Warning: This result does not hold for $0 < p < 1$. ∎

Proof. Following Riesz, we derive Minkowski's inequality from Hölder's inequality. The triangle inequality for real numbers implies the pointwise inequality

$$|X + Y|^p = |X + Y| \cdot |X + Y|^{p-1} \le |X| \cdot |X + Y|^{p-1} + |Y| \cdot |X + Y|^{p-1}.$$

Take the expectation to obtain

$$\|X + Y\|_p^p \le \mathbb{E}\left[|X| \cdot |X + Y|^{p-1}\right] + \mathbb{E}\left[|Y| \cdot |X + Y|^{p-1}\right].$$

Apply Hölder's inequality to each term on the right, noting that the exponent $q = p/(p - 1)$ is conjugate to the exponent p. We reach

$$\|X + Y\|_p^p \le \|X\|_p \|X + Y\|_p^{p-1} + \|Y\|_p \|X + Y\|_p^{p-1}.$$

This statement readily implies Minkowski's inequality (11.5). ∎

Exercise 11.10 (Triangle equality). Under what conditions does (11.5) hold with equality?

Exercise 11.11 (Lower triangle inequality). Fix $p \ge 1$. Let $X, Y \in L_p$. Verify that

$$\|X + Y\|_p \ge \left| \|X\|_p - \|Y\|_p \right|.$$

Problem 11.12 **(*The quasi-triangle inequality).** For $0 < p < 1$, the homogeneous moment $\|\cdot\|_p$ does not satisfy the triangle inequality. Nevertheless, it does satisfy a *quasi-triangle inequality*:

$$\|X + Y\|_p \leq 2^{(1/p)-1}\big(\|X\|_p + \|Y\|_p\big).$$

Verify this statement. **Hint:** Show that $|x + y|^p \leq |x|^p + |y|^p$ for $x, y \in \mathbb{R}$.

11.1.6 The L_p pseudonorm

We can summarize our discussion about the homogeneous moments for $p \geq 1$ in the following result.

Corollary 11.13 **(L_p pseudonorm).** Fix $p \geq 1$. Then $\|\cdot\|_p : L_p \to \mathbb{R}_+$ satisfies the properties of a pseudonorm, (1)–(3) below, for all random variables $X, Y \in L_p$.

1. **Positive semidefinite:** $\|X\|_p \geq 0$ and $\|0\|_p = 0$.
2. **Positive homogeneous:** $\|\alpha X\|_p = |\alpha| \cdot \|X\|_p$ for all $\alpha \in \mathbb{R}$.
3. **Triangle inequality:** $\|X + Y\|_p \leq \|X\|_p + \|Y\|_p$.
4. **Almost positive:** If $\|X\|_p = 0$, then $X = 0$ almost surely.

Proof. The first and last statements are a direct consequence of expectation properties (Theorem 9.7). The second statement appears in (11.2). The third statement is Minkowski's inequality (Theorem 11.9). ∎

As usual, the pseudonorm induces a pseudometric on random variables in L_p. For $X, Y \in L_p$, we can consider the distance

$$\operatorname{dist}_p(X, Y) := \|X - Y\|_p.$$

This function satisfies most of the properties (positivity, symmetry, triangle inequality) of a metric. On the other hand, $\operatorname{dist}_p(X, Y) = 0$ only implies that $X = Y$ almost surely.

As we will discuss in Section 11.2, the L_p pseudonorm allows us to make the linear space L_p into a pseudo-Banach space.

11.1.7 L_∞ spaces

We supplement our collection of $L_p(\Omega, \mathcal{F}, \mathbb{P})$ with one more important example.

Definition 11.14 **(Essential supremum; essentially bounded).** For a real random variable X, define the *essential supremum*:

$$\operatorname{ess\,sup}(X) := \inf\{M \in \mathbb{R} : X \leq M \text{ almost surely}\}.$$

A random variable with $\operatorname{ess\,sup}(|X|) < +\infty$ is said to be *essentially bounded*.

Definition 11.15 **(L_∞ space; pseudonorm).** The space $L_\infty := L_\infty(\mathbb{P}) := L_\infty(\Omega, \mathcal{F}, \mathbb{P})$ consists of random variables that are essentially bounded:

$$L_\infty := \{X : \Omega \to \mathbb{R} \text{ is a random variable with } \operatorname{ess\,sup}(|X|) < +\infty\}.$$

For a real random variable X, we write

$$\|X\|_\infty := \|X\|_{L_\infty} := \operatorname{ess\,sup}(|X|).$$

Thus, $\mathsf{L}_\infty = \{X : \|X\|_\infty < +\infty\}$.

Exercise 11.16 **(L_∞ is a linear space).** By direct argument, confirm that L_∞ is a linear space of real random variables.

The next problem explains the notation and the relationship between L_∞ and the other L_p spaces.

Problem 11.17 **(Essential supremum).** Prove that the essential supremum of the absolute value is the limit of the L_p pseudonorms:

$$\|X\|_\infty = \lim_{p\to\infty} \|X\|_p.$$

Deduce that $\|\cdot\|_\infty : \mathsf{L}_\infty \to \mathbb{R}_+$ is a pseudonorm on the linear space L_∞.

The space L_∞ is conjugate to the space L_1. This is the extreme case of Hölder's inequality.

Exercise 11.18 **(Hölder's inequality).** For random variables $X \in \mathsf{L}_1$ and $Y \in \mathsf{L}_\infty$, show that $XY \in \mathsf{L}_1$, and

$$\|XY\|_1 \le \|X\|_1 \|Y\|_\infty.$$

11.1.8 *More L_p spaces

There are some other spaces of random variables that are related to L_p spaces. You may encounter these from time to time, but they are less important in practice. First, we introduce weak L_p spaces, which are defined in terms of tail decay.

Definition 11.19 **(Weak L_p space).** For $p > 0$, the *weak* L_p *space* $\mathsf{L}_{p,\infty} := \mathsf{L}_{p,\infty}(\Omega, \mathcal{F}, \mathbb{P})$ consists of the random variables $X : \Omega \to \mathbb{R}$ whose tail probability decays at the rate t^{-p}.

$$\mathsf{L}_{p,\infty} := \left\{X : \sup_{t\ge 0} t^p \, \mathbb{P}\{|X| \ge t\} < +\infty\right\}.$$

The condition that defines the space does not satisfy the triangle inequality, so it is not a pseudonorm.

Exercise 11.20 **(Weak L_p spaces).** Fix $p > 0$. Show that $\mathsf{L}_p \subseteq \mathsf{L}_{p,\infty} \subseteq \mathsf{L}_q$ for all $q < p$.

Aside: You will sometimes encounter the space $\mathsf{L}_0 := \mathsf{L}_0(\Omega, \mathcal{F}, \mathcal{P})$, which consists of all random variables on the probability space $(\Omega, \mathcal{F}, \mathcal{P})$. The notation is problematic, and it is probably best avoided. In particular, there seems to be a range of opinions about what topology L_0 carries.

11.1.9 *Probability versus functional analysis: Two warnings

If you have taken a functional analysis course, you have encountered variants of the L_p spaces that we have discussed. There are some intrinsic differences in these definitions that can cause dizziness, confusion, and blood loss. Let us present two warnings that describe the main differences between our approach to L_p spaces and the functional analysis approach.

Warning 11.21 **(Functions versus equivalence classes).** In functional analysis, it is common to identify functions that are equal almost everywhere as the same function. In the present context, here is the analogous approach. Define an equivalence class of real random variables:

$$[X] := \{Y : \Omega \to \mathbb{R} \text{ is a random variable} : Y = X \text{ almost surely}\}.$$

We can collect the equivalence classes of p-integrable random variables:

$$\mathscr{L}_p := \{[X] : \mathbb{E}\,|X|^p < +\infty\}.$$

The collection $\mathscr{L}_p$ is a linear space (formed as a quotient of L_p). We can define

$$\|[X]\|_p := (\mathbb{E}\,|X|^p)^{1/p} \quad \text{for } [X] \in \mathscr{L}_p.$$

On $\mathscr{L}_p$, the function $\|\cdot\|_p$ is actually a norm, rather than a pseudonorm.

Many probabilists prefer to treat random variables as ordinary functions (not equivalence classes). We will do so in this class, in large part because functions are more concrete objects for us to think about. This choice has some consequences. For instance, we have to live with linear spaces equipped with pseudonorms, rather than norms. Furthermore, many statements in probability have to be understood as holding almost surely. ▪

Warning 11.22 (Inclusion of L_p spaces). In functional analysis, we encounter sequence spaces and function spaces that superficially resemble our L_p spaces. For $p > 0$, we may define

$$\ell_p(\mathbb{N}) := \left\{\boldsymbol{a} : \mathbb{N} \to \mathbb{R} \text{ with } \sum\nolimits_{i=1}^{\infty} |a_i|^p < +\infty\right\};$$
$$L_p(\mathbb{R}) := \left\{f : \mathbb{R} \to \mathbb{R} \text{ with } \int_{\mathbb{R}} |f|^p \,\mathrm{d}\lambda < +\infty\right\}.$$

In words, the sequence space $\ell_p(\mathbb{N})$ contains sequences that are p-integrable with respect to the counting measure. The function space $L_p(\mathbb{R})$ contains real functions that are p-integrable with respect to the Lebesgue measure. Neither the counting measure nor the Lebesgue measure is a probability measure, so these two spaces have a different flavor from the L_p spaces that arise in probability.

Indeed, the sequence spaces satisfy an inclusion that is reverse to the one in Theorem 11.4:

$$0 < p \le q \quad \text{implies} \quad \ell_p(\mathbb{N}) \subseteq \ell_q(\mathbb{N}).$$

The function spaces satisfy no containments at all: $L_p(\mathbb{R}) \not\subseteq L_q(\mathbb{R})$ for any $p, q > 0$. Be careful! ▪

11.2 Convergence in L_p spaces

For $p \ge 1$, the L_p pseudonorm allows us to equip L_p with a notion of convergence, just as we do in a normed space. In this section, we briefly discuss what this type of convergence means. Then we turn to the problem of showing that L_p is complete: every Cauchy sequence converges.

11.2.1 Convergence

We begin with the definition of convergence in L_p. From now on, we focus on the case $p \in [1, \infty]$ so that we can benefit from the pseudonorm structure.

Definition 11.23 (L_p convergence). Fix $p \in [1, \infty]$. A sequence $(X_j \in \mathsf{L}_p : j \in \mathbb{N})$ of random variables *converges in* L_p when there is a random variable $X \in \mathsf{L}_p$ for which

$$\|X_j - X\|_p \to 0 \quad \text{as } j \to \infty.$$

We may also write $X_j \to X$ in L_p.

By monotonicity (Theorem 11.4), it is easy to see that

$$\|X_j - X\|_p \to 0 \quad \text{implies} \quad \|X_j - X\|_q \to 0 \quad \text{for all } q \le p.$$

In other words, convergence in L_p implies convergence in L_q for all $q \le p$. Convergence in L_p is sometimes called *convergence of pth moments*.

Activity 11.24 (Convergence in L_p). What does convergence in L_p mean? Draw some pictures to illustrate the concept. **Hint:** Recall that moments control tail decay and vice versa. ▪

Unfortunately, convergence in L_p is incomparable with the notions of pointwise convergence and almost-sure convergence.

Problem 11.25 (*Convergence failures). Let us consider the "universal" probability space $([0,1], \mathcal{B}([0,1]), \lambda)$.

1. Construct a sequence $(X_j : j \in \mathbb{N})$ of random variables that converges almost surely but does not converge in L_1.
2. Construct a sequence $(Y_j : j \in \mathbb{N})$ of random variables that converges in L_1 but does not converge almost surely.

Recall that pointwise convergence always implies almost-sure convergence.

Warning 11.26 (L_p limits are not necessarily unique). In contrast to the situation in functional analysis, the limit of a convergent sequence in L_p may not be unique. Indeed, it is possible that $X_j \to Y$ and $X_j \to Y'$ in L_p, where the limits $Y \neq Y'$. Nevertheless, by the triangle inequality, we can quickly verify that the two limits satisfy

$$\|Y - Y'\|_p = 0.$$

As a consequence, $Y = Y'$ almost surely. ▪

11.2.2 Cauchy sequences

Next, we introduce a class of sequences of random variables with the property that the tail of the sequence eventually enters an arbitrarily small ball and remains there. These sequences are candidates for convergent sequences, but the definition does not require us to identify the actual limit.

Definition 11.27 (L_p Cauchy sequence). Fix $p \in [1, \infty]$. Let $(X_j : j \in \mathbb{N})$ be a sequence of random variables in L_p. We say that the sequence is *Cauchy* when

$$\sup_{i,j \ge N} \|X_i - X_j\|_p \to 0 \quad \text{as } N \to \infty.$$

See Figure 11.2 for an illustration.

It is common to say that a Cauchy sequence "converges inside itself" or that it "wants to converge." The next exercise justifies the language.

Exercise 11.28 (Convergent sequences are Cauchy). Let $(X_j : j \in \mathbb{N})$ be a sequence of random variables in L_p that converges to a limit $X \in \mathsf{L}_p$. That is, $\|X_j - X\|_p \to 0$. Show that (X_j) is a Cauchy sequence.

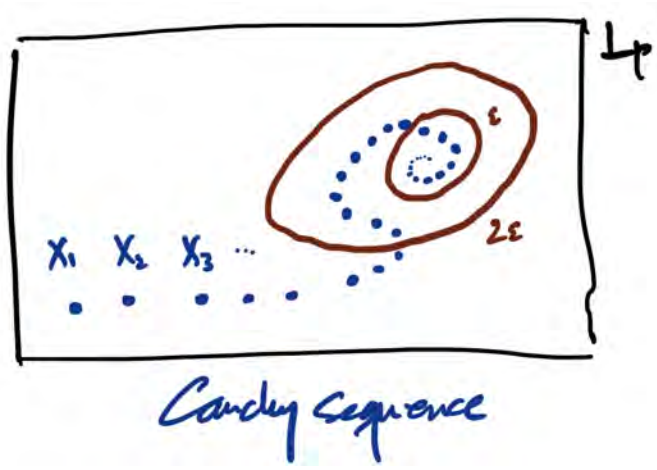


Figure 11.2 (Cauchy sequence). For each $\varepsilon > 0$, a Cauchy sequence $(X_j : j \in \mathbb{N})$ in L_p eventually enters an L_p ball of radius ε about an element X_k where $k = k(\varepsilon)$. The entire tail of the sequence $(X_i : i \geq k)$ remains in that ball.

11.2.3 Completeness

In a general (pseudo)normed space, it is not always the case that a Cauchy sequence converges. Instead, it is possible for the sequence to "slip through the cracks," approaching a limit that does not belong to the space. For example, the decimal approximations $1, 1.4, 1.41, 1.414, \ldots$ are rational numbers that approach $\sqrt{2}$. This Cauchy sequence does not converge in $\mathbb{Q}$, but it does converge in $\mathbb{R}$.

This observation motivates the next definition, which is a standard concept from functional analysis.

Definition 11.29 (Completeness). We say that a (pseudo)normed linear space is *complete* when every Cauchy sequence converges to a limit in the space. A complete (pseudo)normed space is called a *(pseudo-)Banach space*.

Theorem 11.30 (L_p is complete). Fix $p \in [1, \infty]$. Let $(X_j \in \mathsf{L}_p : j \in \mathbb{N})$ be a Cauchy sequence in L_p. Then there is a random variable $Y \in \mathsf{L}_p$ for which $\|X_j - Y\|_p \to 0$. In other words, L_p is complete.

"You complete me."

—*Jerry Maguire* (1991)

**Proof.* We will extract a subsequence $(X_{k_n} : n \in \mathbb{N})$ from the original sequence that converges within itself so fast that it is easy to produce a limit $X_{k_n} \to Y$ in L_p. Then we will demonstrate that the entire sequence converges to this limit: $X_j \to Y$ in L_p.

Since $(X_j : j \in \mathbb{N})$ is Cauchy, we can select an increasing sequence $(k_n : n \in \mathbb{N})$ of indices for which

$$\|X_i - X_j\|_p \leq 2^{-n} \quad \text{for all } i, j \geq k_n.$$

Then

$$\mathbb{E}\,|X_{k_{n+1}} - X_{k_n}| = \|X_{k_{n+1}} - X_{k_n}\|_1 \leq \|X_{k_{n+1}} - X_{k_n}\|_p \leq 2^{-n}.$$

By Tonelli's theorem for series (Exercise 5.39),

$$\mathbb{E}\left[\sum\nolimits_{n=1}^{\infty} |X_{k_{n+1}} - X_{k_n}|\right] = \sum\nolimits_{n=1}^{\infty} \mathbb{E}\,|X_{k_{n+1}} - X_{k_n}| < +\infty.$$

In particular, the series on the left-hand side converges almost surely because every integrable (positive) random variable is almost surely finite (Exercise 9.42). As a consequence,

$$\sum\nolimits_{n=1}^{\infty} (X_{k_{n+1}} - X_{k_n}) \quad \text{converges (absolutely) almost surely.}$$

The sum telescopes, and we see that $\lim_{n\to\infty} X_{k_n}(\omega)$ exists for almost every $\omega \in \Omega$. Define the random variable

$$Y(\omega) := \limsup\nolimits_{n\to\infty} X_{k_n}(\omega) \quad \text{for each } \omega \in \Omega.$$

Indeed, the limit superior yields a measurable function. We recognize that $X_{k_n} \to Y$ almost surely because the limit coincides with the limit superior whenever the limit exists.

We have now produced a tentative limit Y for the original sequence $(X_j : j \in \mathbb{N})$. Our next task is to verify that the random variable Y belongs to L_p and is indeed the limit of the original sequence. By construction,

$$\mathbb{E}\,|X_j - X_{k_n}|^p = \|X_j - X_{k_n}\|_p^p \le 2^{-mp} \quad \text{for all } j \ge k_n \text{ and } n \ge m.$$

Using Fatou's lemma (Theorem 9.11), for fixed $i \ge k_m$,

$$\begin{aligned} 2^{-mp} &\ge \liminf\nolimits_{m\to\infty} \mathbb{E}\,|X_j - X_{k_n}|^p \\ &\ge \mathbb{E}\left[\liminf\nolimits_{m\to\infty} |X_j - X_{k_n}|^p\right] = \mathbb{E}\,|X_j - Y|^p. \end{aligned}$$

Indeed, $X_{k_n} \to Y$ almost surely, and the expectation is insensitive to the negligible set where the sequence does not converge. This calculation ensures that $X_j - Y \in \mathsf{L}_p$ for each $j \in \mathbb{N}$. From Minkowski's inequality, we deduce that $Y \in \mathsf{L}_p$. Finally, we observe that

$$\limsup\nolimits_{j\to\infty} \|X_j - Y\|_p \le 2^{-m}.$$

Take $m \to \infty$ to confirm that $X_j \to Y$ in L_p. ∎

Exercise 11.31 (*L_p is complete for $0 < p < 1$.). Extend Theorem 11.30 to the case $0 < p < 1$. **Hint:** You only need to use the quasi-triangle inequality in place of the triangle inequality.

Problem 11.32 (Closure and completeness). A linear subspace $\mathsf{K} \subseteq \mathsf{L}_p$ is *closed* in L_p if it contains all its limit points. More precisely, suppose $(X_i \in \mathsf{K} : i \in \mathbb{N})$ is a sequence in K that converges in L_p to a limit point $X \in \mathsf{L}_p$. Then K is closed if and only if the limit $X \in \mathsf{K}$. Prove that a subspace K of L_p is closed if and only if the subspace K is complete. **Hint:** You need to use the fact that convergent sequences are Cauchy and the theorem that L_p itself is complete.

Problems

Exercise 11.33 (Moments: Interpolation). Consider real parameters $1 \le p \le r \le q \le +\infty$. Derive *Littlewood's inequality*:

$$\|X\|_r \le \|X\|_p^{\theta} \cdot \|X\|_q^{1-\theta} \quad \text{where} \quad \frac{1}{r} = \frac{\theta}{p} + \frac{1-\theta}{q}.$$

In other words, a homogeneous moment whose order r lies between p and q is bounded by an appropriate geometric mean of the pth and qth homogeneous moments. The weight θ in the geometric mean is computed by writing $1/r$ as a weighted average of $1/p$ and $1/q$.

Problem 11.34 (L_p pseudonorm: Duality). For $p \geq 1$, we can find an alternative "dual" representation for the L_p pseudonorm. Prove that

$$\|X\|_p = \sup\{\mathbb{E}[XY] : \|Y\|_q \leq 1\} \quad \text{where} \quad \frac{1}{p} + \frac{1}{q} = 1.$$

Hint: One direction follows from Hölder's inequality. The other direction requires the selection of an appropriate random variable Y that depends on X.

Problem 11.35 (Hölder: Beyond two factors). Hölder's inequality can be extended to products of more than two random variables.

1. Consider a family $(X_1, X_2, \dots, X_n)$ of real random variables. Fix numbers $p_i \geq 1$ with the property that $p_1^{-1} + \cdots + p_n^{-1} = 1$. Establish that

   $$\|X_1 X_2 \cdots X_n\|_1 \leq \|X_1\|_{p_1} \cdot \|X_2\|_{p_2} \cdots \|X_n\|_{p_n}.$$

   Hint: You can prove this quickly by induction.
2. Deduce that the weighted geometric mean is a concave function. For positive numbers $\tau_i \geq 0$ with $\sum_{i=1}^n \tau_i = 1$, show that

   $$(x_1, x_2, \dots, x_n) \mapsto x_1^{\tau_1} x_2^{\tau_2} \cdots x_n^{\tau_n}$$

   is a concave function on the set $\mathbb{R}_+^n$ of positive vectors. **Hint:** Change variables in Hölder's inequality, and interpret the statement as an invocation of Jensen's inequality.

Problem 11.36 (Uniform smoothness and convexity).** There are some beautiful geometric inequalities that hold for random variables in L_p spaces. These results are related to the familiar parallelogram law in Euclidean space (Exercise 12.10), but they reflect the fact that L_p balls have varying curvature. The properties encoded by these results are called *uniform smoothness* ($p \geq 2$) and *uniform convexity* ($1 \leq p \leq 2$). They play an important role in more advanced studies of martingales.

1. For $p \geq 2$, prove the Gross's *two-point inequality*: For all $x, y \in \mathbb{R}$,

   $$\left[\tfrac{1}{2} \cdot \left(|x + y|^p + |x - y|^p\right)\right]^{2/p} \leq x^2 + (p - 1) \cdot y^2$$

   Show that the inequality is reversed for $1 \leq p \leq 2$. **Hint:** Define the function

   $$u(t) := \tfrac{1}{2} \cdot \left(|x + \sqrt{t}y|^p + |x - \sqrt{t}y|^p\right) \quad \text{for } t \in [0, 1].$$

   To control $u(1) - u(0)$, develop a bound on the derivative $u'(t)$. Exploit the fact that $t \mapsto t^{p-1}$ is convex for $p \geq 2$.
2. For $p \geq 2$, extend this result to real random variables $X, Y \in \mathsf{L}_p$:

   $$\left[\tfrac{1}{2} \cdot \left(\|X + Y\|_p^p + \|X - Y\|_p^p\right)\right]^{2/p} \leq \|X\|_p^2 + (p - 1) \cdot \|Y\|_p^2.$$

   Show that the inequality is reversed for $1 \leq p \leq 2$.
3. For $p \geq 2$, use Jensen's inequality to derive that

   $$\tfrac{1}{2} \cdot \left(\|X + Y\|_p^2 + \|X - Y\|_p^2\right) \leq \|X\|_p^2 + (p - 1) \cdot \|Y\|_p^2.$$

   Draw a picture to illustrate what this inequality means.

Notes

Our development of L_p spaces is inspired by Williams [Wil91, Chap. 6]; in particular, we have used his proof of Theorem 11.30.

For more results on L_p spaces, see the books of Garling [Gar07], Lieb & Loss [LL01], and Steele [Ste04]. The proof strategy for the uniform smoothness inequality is drawn from [Tro22].

Lecture bibliography

[Gar07] D. J. H. Garling. *Inequalities: a journey into linear analysis*. Cambridge University Press, 2007. DOI: 10.1017/CBO9780511755217.

[LL01] E. H. Lieb and M. Loss. *Analysis*. 2nd ed. American Mathematical Society, 2001. DOI: 10.1090/gsm/014.

[Ste04] J. M. Steele. *The Cauchy-Schwarz master class*. An introduction to the art of mathematical inequalities. Mathematical Association of America / Cambridge Univ. Press, 2004. DOI: 10.1017/CBO9780511817106.

[Tro22] J. A. Tropp. *ACM 204: Matrix Analysis*. CMS Lecture Notes 2022-01. Caltech, 2022. DOI: 10.7907/m421-yb89.

[Wil91] D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

12. L_2 Spaces & Orthogonality

"**Scarecrow:** The sum of the square roots of any two sides of an isosceles triangle is equal to the square root of the remaining side. Oh joy, rapture! I've got a brain! How can I ever thank you enough?

"**Wizard:** Well, you can't."

—*The Wizard of Oz*, 1939

Agenda:
1. L_2 spaces
2. Cauchy–Schwarz
3. Inner-product geometry
4. Covariance
5. Orthogonal projection

In the last lecture, we introduced the scale of $L_p(\Omega; \mathcal{F}; \mathbb{P})$ spaces associated with a probability space (where $p > 0$). These are linear spaces containing random variables that have finite pth absolute polynomial moments. We saw that these spaces decrease in size as the power p increases.

For $p \geq 1$, the L_p space is equipped with a pseudonorm $\|\cdot\|_p$. The pseudonorm induces a notion of convergence, and we saw that each L_p space is complete. That is, every Cauchy sequence in L_p converges to a limit in L_p.

In this lecture, we undertake a deeper investigation of the space L_2. In addition to all the properties outlined above, the L_2 space is equipped with a pseudo-inner product. This structure allows us to define orthogonal random variables, and it is the right context for introducing the variance and covariance.

The inner-product geometry also permits us to define the orthogonal projection of a random variable onto a subspace. Later, we will see that the orthogonal projection serves as the foundation for the concept of conditional expectation, which is one of the most important ideas in probability theory.

12.1 Square-integrable random variables

To begin, let us set the stage with the formal definitions, which specialize the more general notions from Lecture 11.

Definition 12.1 (L_2 space). The space $L_2 := L_2(\mathbb{P}) := L_2(\Omega, \mathcal{F}, \mathbb{P})$ is defined as

$$L_2 := \{X : \Omega \to \mathbb{R} \text{ is a random variable with } \mathbb{E}\,|X|^2 < +\infty\}.$$

For a random variable $X \in L_2$, the L_2 pseudonorm is defined as

$$\|X\|_2 := \|X\|_{L_2} := (\mathbb{E}\,|X|^2)^{1/2}.$$

We refer to L_2 as the space of *square-integrable random variables*.

According to Proposition 11.2, the space L_2 is a linear space. Corollary 11.13 ensures that $\|\cdot\|_2$ is a pseudonorm on L_2; in particular, it satisfies the triangle inequality. Theorem 11.30 states that L_2 is complete: every Cauchy sequence in L_2 converges to a random variable in L_2.

12.2 The Cauchy–Schwarz inequality

The most basic fact about the space L_2 is an inequality that gives a bound for the expectation of a product of square-integrable random variables. This result specializes Hölder's inequality (Theorem 11.5). Because of its importance, we will give an independent proof.

Warning: "Schwarz" is **not** spelled with a "t"! ∎

Theorem 12.2 (Cauchy–Schwarz). For square-integrable, real random variables $X, Y \in \mathsf{L}_2$, the product $XY \in \mathsf{L}_1$. Furthermore,

$$|\mathbb{E}[XY]| \le \mathbb{E}\,|XY| = \|XY\|_1 \le \|X\|_2 \|Y\|_2. \tag{12.1}$$

Proof. This argument is due to Schwarz. For any parameter $\xi \in \mathbb{R}$, observe that

$$0 \le \mathbb{E}\,(\xi X + Y)^2 = \xi^2 \cdot (\mathbb{E}\,X^2) + 2\xi \cdot \mathbb{E}[XY] + (\mathbb{E}\,Y^2).$$

The left-hand inequality holds because the expectation of a positive random variable is positive. The expectation is finite because L_2 is a linear space. To obtain the equality, we expand the square and invoke the linearity of expectation.

According to the quadratic formula, a quadratic polynomial $\xi^2 a + 2\xi b + c \ge 0$ for all $\xi \in \mathbb{R}$ if and only if the discriminant $(2b)^2 - 4ac \le 0$. That is,

$$(\mathbb{E}[XY])^2 \le (\mathbb{E}\,X^2) \cdot (\mathbb{E}\,Y^2)$$

To obtain the (stricter) result with absolute values, simply make the change of variables $X \mapsto |X|$ and $Y \mapsto |Y|$. ∎

Exercise 12.3 (Cauchy–Schwarz: Equality). Suppose that the Cauchy–Schwarz inequality holds with equality:

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

What can we deduce about the relationship between X and Y? **Hint:** Under what circumstances is $\mathbb{E}(\xi X + Y)^2 > 0$ for all $\xi \in \mathbb{R}$?

Exercise 12.4 (L_2 triangle inequality). Use the Cauchy–Schwarz inequality (Theorem 12.2) to verify that

$$\|X + Y\|_2 \le \|X\|_2 + \|Y\|_2 \quad \text{for } X, Y \in \mathsf{L}_2.$$

Hint: Specialize the proof of Minkowski's inequality (Theorem 11.9).

12.3 The L_2 pseudo-inner product

The Cauchy–Schwarz inequality (Theorem 12.2) allows us to introduce an inner-product geometry on the space of square-integrable random variables.

Definition 12.5 (L_2 pseudo-inner product). For square-integrable, real random variables $X, Y \in \mathsf{L}_2$, define

$$\langle X,\ Y\rangle := \langle X,\ Y\rangle_{\mathsf{L}_2} := \mathbb{E}[XY].$$

In particular, $\langle X,\ X\rangle = \|X\|_2^2$.

The Cauchy–Schwarz inequality ensures that the pairing $\langle \cdot,\ \cdot\rangle : \mathsf{L}_2 \times \mathsf{L}_2 \to \mathbb{R}$ is defined for all square-integrable random variables.

Exercise 12.6 (L_2 pseudo-inner product: Properties). Show that the pairing $\langle \cdot,\ \cdot\rangle$ on L_2 meets the definition of a pseudo-inner product, (1)–(3) below. Let $X, Y, Z \in \mathsf{L}_2$.

1. **Positive semidefinite:** $\langle X,\ X\rangle \geq 0$.
2. **Symmetric:** $\langle X,\ Y\rangle = \langle Y,\ X\rangle$.
3. **Bilinear:** For all real scalars $\alpha, \beta \in \mathbb{R}$,

$$\langle \alpha X + \beta Y,\ Z\rangle = \alpha\langle X,\ Z\rangle + \beta\langle Y,\ Z\rangle;$$
$$\langle Z,\ \alpha X + \beta Y\rangle = \alpha\langle Z,\ X\rangle + \beta\langle Z,\ Y\rangle.$$

4. **Almost positive:** If $\langle X,\ X\rangle = 0$, then $X = 0$ almost surely.

Exercise 12.6 tells us that $\langle \cdot,\ \cdot\rangle$ behaves almost exactly like the inner products that we encounter in linear algebra and functional analysis. The only caveat is that $\langle X,\ X\rangle = 0$ only allows us to conclude that $X = 0$ almost surely.

Aside from the latter point, we can now think about random variables in L_2 geometrically, exploiting intuitions we have already developed. Indeed, the pseudo-inner product gives us a notion of the "alignment" between two random variables. This brings us to the next definition.

Definition 12.7 (Orthogonality). Let $X, Y \in L_2$. If $\langle X,\ Y\rangle = 0$, then we say that X and Y are *orthogonal* random variables. We may write $X \perp Y$ to denote orthogonality.

Warning: Orthogonality random variables need not be "independent". ▪

Example 12.8 (Indicators: Orthogonality). Suppose that A and B are mutually exclusive events. Then the indicator random variables are orthogonal:

$$\mathbb{E}[\mathbb{1}_A \mathbb{1}_B] = \mathbb{E}[\mathbb{1}_\emptyset] = 0.$$

More generally, if $A \cap B = E$, then the indicators $\mathbb{1}_A$ and $\mathbb{1}_B$ are orthogonal if and only if E is a negligible event: $\mathbb{P}(E) = 0$. ▪

Geometrically, orthogonal random variables behave like vectors at right angles to each other. In particular, they enjoy a Pythagorean relation.

Exercise 12.9 (L_2 Pythagorean theorem). Let $X, Y \in L_2$. Then

$$\langle X,\ Y\rangle = 0 \quad \text{implies} \quad \|X + Y\|_2^2 = \|X\|_2^2 + \|Y\|_2^2.$$

Hint: Use bilinearity of the pseudo-inner product.

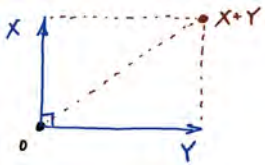


Figure 12.1 (Pythagorean relation).

Even without orthogonality, we have an identity for the squared lengths of two random variables.

Exercise 12.10 (L_2 parallelogram law). Let $X, Y \in L_2$. Then

$$\|X + Y\|_2^2 + \|X - Y\|_2^2 = 2\|X\|_2^2 + 2\|Y\|_2^2.$$

In other words, the total squared length of the diagonals of a parallelogram equals the total squared length of the four sides.

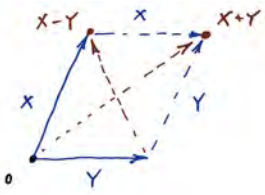


Figure 12.2 (Parallelogram law).

The Pythagorean relation and the parallelogram law depend crucially on the fact that we are using the L_2 pseudonorm to measure magnitudes. These results fail in other L_p spaces ($p \neq 2$), although there are some very interesting substitutes (see Problem 11.36).

We have already seen that L_2 is complete (Theorem 11.30), so the pseudo-inner product structure makes L_2 into a *pseudo-Hilbert space*.

12.4 Covariance and variance

The space L_2 of square-integrable random variables supports another pseudo-inner product structure that plays a special role in probability theory. One reason that two random variables may have a large pseudo-inner product is that they both have large expectations. It can be valuable to subtract the expectations before asking how closely the random variables are aligned with each other.

Definition 12.11 (Covariance). Let $X, Y \in \mathsf{L}_2$. Define the *covariance*

$$\operatorname{Cov}(X, Y) := \langle X - \mathbb{E}\, X,\ Y - \mathbb{E}\, Y \rangle = \mathbb{E}[XY] - (\mathbb{E}\, X)(\mathbb{E}\, Y).$$

If $\operatorname{Cov}(X, Y) = 0$, then we say that X and Y are *uncorrelated*.

Warning: Uncorrelated random variables need not be "independent", which is a stronger requirement. See Lecture 13. ▪

Example 12.12 (Indicators: Covariance). For two events A and B, the covariance of the indicators satisfies

$$\begin{aligned}\operatorname{Cov}(\mathbb{1}_{\mathsf{A}}, \mathbb{1}_{\mathsf{B}}) &= \mathbb{E}[\mathbb{1}_{\mathsf{A}\cap\mathsf{B}}] - \mathbb{E}[\mathbb{1}_{\mathsf{A}}]\,\mathbb{E}[\mathbb{1}_{\mathsf{B}}] \\ &= \mathbb{P}(\mathsf{A} \cap \mathsf{B}) - \mathbb{P}(\mathsf{A}) \cdot \mathbb{P}(\mathsf{B}).\end{aligned}$$

If A and B are mutually exclusive events, then the covariance is always negative. If $\mathsf{A} \subseteq \mathsf{B}$, then the covariance is always positive. ▪

Exercise 12.13 (Covariance pseudo-inner product). Show that Cov defines a pseudo-inner product on L_2. That is, Cov is positive semidefinite, symmetric, and bilinear. What can we conclude about X when $\operatorname{Cov}(X, X) = 0$?

The pseudonorm associated with the covariance form should be familiar to you.

Definition 12.14 (Variance). Let $X \in \mathsf{L}_2$. The *variance* of X is defined as

$$\operatorname{Var}[X] := \operatorname{Cov}(X, X).$$

Exercise 12.15 (Variance). Check that the variance satisfies the familiar definitions:

$$\operatorname{Var}[X] = \mathbb{E}[(X - \mathbb{E}\, X)^2] = \mathbb{E}\, X^2 - (\mathbb{E}\, X)^2.$$

Confirm that the square-root of the variance is a pseudonorm on L_2. Establish the variational formulation:

$$\operatorname{Var}[X] = \inf_{\alpha\in\mathbb{R}} \mathbb{E}[(X - \alpha)^2] = \inf_{\alpha\in\mathbb{R}} \|X - \alpha\|_2^2. \tag{12.2}$$

In words, the variance computes the expected squared deviation of a random variable X from its average value $\mathbb{E}\, X$. So the variance is a summary quantity that describes how much a random variable fluctuates around its mean value. The relation (12.2) gives us an important new insight about the expectation. Indeed, $\mathbb{E}\, X$ is the constant (random variable) that best approximates the random variable X with respect to the L_2 pseudo-norm. This innocent observation will turn out to be very important.

Since the variance is the quadratic form induced by a bilinear form, we can decompose the variance of a sum in terms of the covariances of the summands.

Proposition 12.16 (Variance: Pythagorean relation). Consider square-integrable real random variables $X_1, \dots, X_n \in \mathsf{L}_2$. Then

$$\operatorname{Var}\left[\sum\nolimits_{i=1}^{n} X_i\right] = \sum\nolimits_{i=1}^{n} \operatorname{Var}[X_i] + 2 \sum\nolimits_{1\le i<j\le n} \operatorname{Cov}(X_i, X_j).$$

In particular,

$$\operatorname{Cov}(X_i, X_j) = 0 \text{ when } i \neq j \quad \text{implies} \quad \operatorname{Var}\left[\sum\nolimits_{i=1}^{n} X_i\right] = \sum\nolimits_{i=1}^{n} \operatorname{Var}[X_i].$$

In other words, the variance of a sum of *mutually uncorrelated* random variables equals the sum of the variances.

Exercise 12.17 (Variance: Pythagorean relation). Verify Proposition 12.16.

A valuable feature of the Pythagorean relation (Proposition 12.16) is the relatively weak hypothesis of mutual uncorrelation. It does not require "independence", a much stronger property that we will discuss in the next lecture.

Since the covariance is a pseudo-inner product, it satisfies its own Cauchy–Schwarz relation.

Exercise 12.18 (Covariance: Cauchy–Schwarz). Let $X, Y \in \mathsf{L}_2$. Prove the Cauchy–Schwarz inequality for the covariance.

$$|\operatorname{Cov}(X, Y)|^2 \leq \operatorname{Var}[X] \cdot \operatorname{Var}[Y].$$

Hint: This is an immediate consequence of Theorem 12.2.

Last, we define the correlation between two random variables. This quantity reflects whether two random variables have the same trend, opposite trends, or no relationship on average.

Definition 12.19 (Correlation). Let $X, Y \in \mathsf{L}_2$. The *correlation* between X and Y is

$$\rho(X, Y) := \frac{\operatorname{Cov}(X, Y)}{\sqrt{\operatorname{Var}[X] \cdot \operatorname{Var}[Y]}} \in [-1, +1].$$

We use the convention that $\rho(X, Y) = 0$ if either $\operatorname{Var}[X] = 0$ or $\operatorname{Var}[Y] = 0$.

The upper and lower bounds on correlation derive from the Cauchy–Schwarz inequality (Exercise 12.18) for the covariance.

Activity 12.20 (Correlation). If $|\rho(X, Y)| = 1$, what can we deduce about X and Y? Sketch pairs of random variables whose correlation is maximal (+1) and minimal (−1). Draw a picture of two uncorrelated random variables, each with nonzero variance. ▪

12.5 Orthogonal projection

In Exercise 12.15, we saw that the expectation $\mathbb{E} X$ is the constant random variable that is closest to X with respect to the L_2 pseudonorm. In this section, we generalize this observation by showing that every (complete) linear subspace in L_2 contains a random variable Y at minimal distance from X.

Theorem 12.21 (L_2 orthogonal projection). Let $\mathsf{K} \subseteq \mathsf{L}_2$ be a *complete* linear subspace. For each random variable $X \in \mathsf{L}_2$, there is a random variable $Y \in \mathsf{K}$ with two properties:

1. **Minimal distance:** $\|X - Y\|_2 = \inf\{\|X - W\| : W \in \mathsf{K}\}$.
2. **Orthogonal error:** $X - Y \perp Z$ for all $Z \in \mathsf{K}$.

Moreover, each of the properties (1) and (2) implies the other.

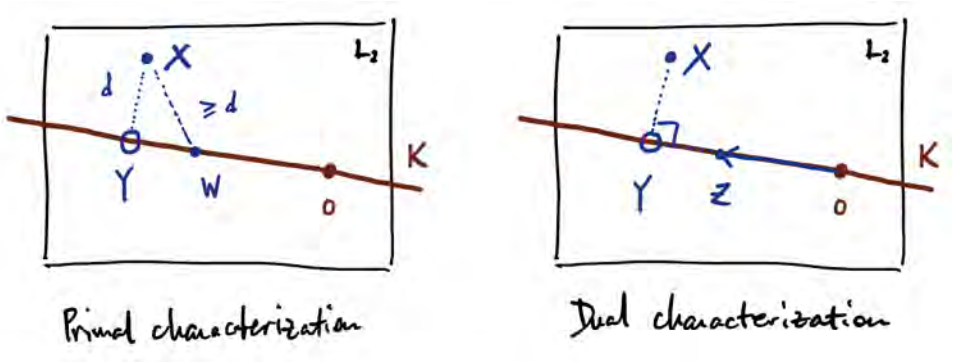


Figure 12.3 (Orthogonal projection). The primal characterization (1) states that an orthogonal projection achieves the minimal distance d from X to the subspace K. The dual characterization (2) states that the residual vector $X - Y$ from an orthogonal projection is orthogonal to the subspace K.

The random variable $Y \in \mathsf{K}$ promised by Theorem 12.21 is called (a *version* of) the *orthogonal projection* of X onto the subspace K. In general, there may be many versions of the orthogonal projection. Fortunately, if Y, Y' are both versions, then $\|Y - Y'\|_2 = 0$ so that $Y = Y'$ almost surely.

Warning: Orthogonal projections are not necessarily unique! ▪

Modulo the latter point, Theorem 12.21 indicates that the geometry of orthogonal projection in L_2 is similar to orthogonal projection in inner-product spaces you have encountered before. See Figure 12.3.

Warning 12.22 (Completeness). We must confirm that the subspace K is complete before we can be confident that the orthogonal projection onto K exists. In our context, the easiest way to do so is to argue that $\mathsf{K} = \mathsf{L}_p(\Omega', \mathcal{F}', \mathbb{P}')$ for some other probability space $(\Omega', \mathcal{F}', \mathbb{P}')$ and invoke Theorem 11.30. ▪

12.5.1 Proof of the orthogonal projection theorem

The main challenge in proving the result on orthogonal projection, Theorem 12.21, is to find a candidate Y for an orthogonal projection of X onto the subspace K. To do so, we construct a *minimizing sequence*. Then we prove that this minimizing sequence is Cauchy, and we extract its limit.

A minimizing sequence

Define the distance $d := \inf\{\|X - W\|_2 : W \in \mathsf{K}\}$ from X to the subspace K. By definition of the infimum, there is a sequence $(Y_n \in \mathsf{K} : n \in \mathbb{N})$ that approaches the minimum distance from X to K. That is,

$$\|X - Y_n\|_2 \downarrow \inf\{\|X - W\|_2 : W \in \mathsf{K}\} = d \quad \text{as } n \to \infty.$$

See Figure 12.4. Our first task is to prove that the minimizing sequence (Y_n) has a limit in K.

The minimizing sequence is Cauchy

Let us argue that the minimizing sequence is Cauchy. We need to obtain detailed information about the distance $\|Y_i - Y_j\|_2$ for large indices i, j. Since we only have information about the distances $\|X - Y_i\|_2$ and $\|X - Y_j\|_2$, it is natural to invoke the

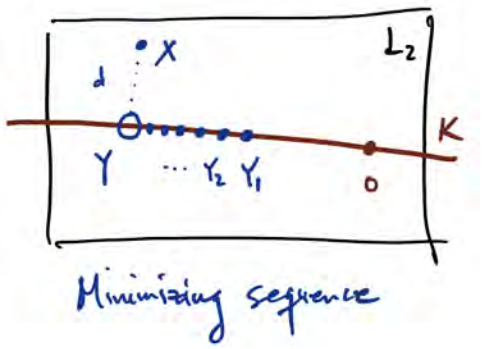


Figure 12.4 (Minimizing sequence). A minimizing sequence $(Y_1, Y_2, Y_3, \dots)$ has the property that $\|X - Y_n\|_2$ approaches the infimal value $d := \inf_{Y \in \mathsf{K}} \|X - Y\|_2$.

parallelogram law (Exercise 12.10) to understand the geometry of these random variables.

Fix indices $i, j \in \mathbb{N}$. Apply the parallelogram law to the random variables $\frac{1}{2}(X - Y_i)$ and $\frac{1}{2}(X - Y_j)$ to obtain

$$\left\|\tfrac{1}{2}(Y_i - Y_j)\right\|_2^2 = \underbrace{\tfrac{1}{2}\|X - Y_i\|_2^2}_{\to d^2/2} + \underbrace{\tfrac{1}{2}\|X - Y_j\|_2^2}_{\to d^2/2} - \underbrace{\left\|X - \tfrac{1}{2}(Y_i + Y_j)\right\|_2^2}_{\geq d^2}.$$

The first two members of the right-hand side tend to $d^2/2$ as $i, j \to \infty$ because (Y_n) is a minimizing sequence. To see that the last member exceeds d^2, note that $\frac{1}{2}(Y_i - Y_j) \in \mathsf{K}$ because Y_i and Y_j belong to the subspace K. The number d^2 is the minimum squared distance from X to any point in K. Thus,

$$\sup_{i,j \geq N} \|Y_i - Y_j\|_2 \to 0 \quad \text{as } N \to \infty.$$

Indeed, the previous argument shows that this limit cannot be strictly positive, but it most certainly cannot be strictly negative.

The limit of the minimizing sequence

We determine that $(Y_n : n \in \mathbb{N})$ is a Cauchy sequence contained in the complete subspace K. As a consequence, there is a random variable $Y \in \mathsf{K}$ for which $\|Y_n - Y\|_2 \to 0$ as $n \to \infty$. We must demonstrate that this random variable Y is a version of the orthogonal projection of X onto K.

The limit achieves the minimal distance

First, let us show that Y achieves the minimum distance from X to the subspace K. This is intuitively clear; the proof involves the triangle inequality:

$$\|X - Y\|_2 \leq \|X - Y_n\|_2 + \|Y_n - Y\|_2 \to d \quad \text{as } n \to \infty.$$

Indeed, the quantity $\|X - Y_n\|_2 \to d$ because (Y_n) is a minimizing sequence. The quantity $\|Y_n - Y\|_2 \to 0$ because Y is an L_2 limit of (Y_n). In summary, $\|X - Y\|_2 = d$.

Minimal distance implies orthogonality of error

Next, we check that the residual $X - Y$ is orthogonal to every random variable $Z \in \mathsf{K}$. This requires a variational argument. We perturb Y slightly in the direction Z, and we

note that the distance to X must increase. By a close analysis of the change in distance, we can extract an orthogonality relation.

Fix $Z \in \mathsf{K}$. For each real number $\xi \in \mathbb{R}$, the random variable $Y + \xi Z \in \mathsf{K}$ because K is a subspace. Since Y minimizes the distance from X to K,

$$\|X - (Y + \xi Z)\|_2^2 \geq \|X - Y\|_2^2.$$

Write the squared pseudonorms as pseudo-inner products, and use bilinearity to expand and cancel terms. By choosing $\operatorname{sgn}(\xi) = \operatorname{sgn}(\langle Z,\ X - Y\rangle)$, we arrive at the relation

$$-2|\xi| \cdot |\langle Z,\ X - Y\rangle| + |\xi|^2 \cdot \|Z\|_2^2 \geq 0.$$

Divide through by $|\xi|$, and take the limit as $|\xi| \downarrow 0$ to determine that $|\langle Z,\ X - Y\rangle| = 0$.

Orthogonality of residual implies minimal distance

To show that the primal and dual characterizations of the orthogonal projection are equivalent, it remains to show that the dual characterization implies the primal. This result uses the Pythagorean relation (Exercise 12.9) to exploit the orthogonality between the residual and the subspace.

Suppose that $Y \in \mathsf{K}$ and $Z \perp (X - Y)$ for all $Z \in \mathsf{K}$. For each random variable $W \in \mathsf{K}$,

$$\|X - W\|_2^2 = \|(X - Y) + (Y - W)\|_2^2 = \|X - Y\|_2^2 + \|Y - W\|_2^2 \geq \|X - Y\|_2^2.$$

Indeed, since K is a subspace, $Z := Y - W \in \mathsf{K}$. By assumption, $Z \perp X - Y$. Therefore, we can apply the Pythagorean relation. The last inequality holds because the L_2 pseudonorm is positive. Finally, by taking the infimum of the last display over $W \in \mathsf{K}$, we realize that

$$\inf\{\|X - W\|_2 : W \in \mathsf{K}\} \geq \|X - Y\|_2.$$

This is the primal characterization of the orthogonal projection.

Uniqueness

Finally, we must show that every version of the orthogonal projection is equal almost surely. To that end, suppose that there are two distinct versions Y, Y' of the orthogonal projection of X onto K. Define the residual random variables $E := X - Y$ and $E' := X - Y'$. According to the parallelogram law (Exercise 12.10),

$$\begin{aligned} d^2 &= 2\|\tfrac{1}{2}E\|_2^2 + 2\|\tfrac{1}{2}E'\|_2^2 = \|\tfrac{1}{2}(E + E)\|_2^2 + \|\tfrac{1}{2}(E - E')\|_2^2 \\ &= \|X - \tfrac{1}{2}(Y + Y')\|_2^2 + \|\tfrac{1}{2}(Y - Y')\|_2^2 \geq d^2 + \tfrac{1}{4}\|Y - Y'\|_2^2. \end{aligned}$$

Indeed, $\|E\|_2^2 = d^2 = \|E'\|_2^2$ by the primal characterization of the orthogonal projection. Since $\frac{1}{2}(Y + Y') \in \mathsf{K}$, this random variable lies at least a distance of d away from X. We must conclude that

$$\|Y - Y'\|_2 = 0.$$

As a consequence, $Y = Y'$ almost surely.

Problems

Exercise 12.23 (Chebyshev i druz'ya). Beyond Markov's inequality, there are many useful probability inequalities. Here are some basic results that often arise in practice.

1. **Chebyshev:** Use Markov's inequality to prove Chebyshev's inequality. For each $X \in \mathsf{L}_2$,
$$\mathbb{P}\{|X - \mathbb{E}X| \geq t\} \leq \frac{\operatorname{Var}[X]}{t^2} \quad \text{for each } t > 0.$$
(*) For fixed $t > 0$, find a random variable with $\mathbb{E}X = 0$ where Chebyshev's inequality holds with equality. **Hint:** Look at the graphical proof of Markov's inequality.
2. For each *positive* random variable $X \in \mathsf{L}_2$ with $\mathbb{E}X > 0$, use Chebyshev's inequality to deduce that
$$\mathbb{P}\{X = 0\} \leq \frac{\operatorname{Var}[X]}{(\mathbb{E}X)^2}.$$
3. ***Paley–Zygmund:** Prove the Paley–Zygmund inequality, which sharpens the last result. For positive $X \in \mathsf{L}_2$,
$$\mathbb{P}\{X > \theta(\mathbb{E}X)\} \geq \frac{(1-\theta)^2(\mathbb{E}X)^2}{\mathbb{E}[X^2]} \quad \text{for } \theta \in [0,1].$$
Hint: Consider complementary events $\{X > \theta\,\mathbb{E}X\}$ and $\{X \leq \theta\,\mathbb{E}X\}$.
4. ***Chebyshev–Cantelli:** For $X \in \mathsf{L}_2$, establish a one-sided version of Chebyshev's inequality:
$$\mathbb{P}\{X - \mathbb{E}X \geq t\} \leq \frac{\operatorname{Var}[X]}{\operatorname{Var}[X] + t^2} \quad \text{for } t \geq 0.$$

Problem 12.24 (*Reverse Cauchy–Schwarz). In general, there is no complementary lower bound to the Cauchy–Schwarz inequality. Nevertheless, if we assume that two random variables are almost proportional, it is possible to obtain a satisfactory reversal.

1. Let $X, Y \in \mathsf{L}_2$ be *positive* random variables that satisfy the pointwise inequality
$$0 < m \leq \frac{X}{Y} \leq M \quad \text{for fixed numbers } m, M > 0.$$
Prove the reverse Cauchy–Schwarz inequality:
$$\mathbb{E}[XY] \geq \frac{G}{A} \cdot \|X\|_2 \cdot \|Y\|_2.$$
where $A := (m+M)/2$ and $G := \sqrt{mM}$. Recall that the ratio $G/A \leq 1$ by the GM–AM inequality (9.11). **Hint:** The key observation is the positivity relation $(M - X/Y)(X/Y - m) \geq 0$.
2. Deduce Kantorovich's inequality: For a *positive* random variable Z that satisfies the pointwise inequality $0 < m \leq Z \leq M$,
$$\mathbb{E}[Z] \cdot \mathbb{E}[Z^{-1}] \leq \left(\frac{A}{G}\right)^2$$
where $A = (m+M)/2$ and $G = \sqrt{mM}$, as above.

Problem 12.25 (*Jensen defect). In general, there is no lower bound complementary to Jensen's inequality. Nevertheless, for functions that are "moderately" convex, we can obtain an elegant expression for the gap between the two sides of Jensen's inequality. This result is due to Hölder.

Let $\varphi : \mathsf{U} \to \mathbb{R}$ be a twice-differentiable convex function on an open interval U of the real line. Assume that

$$m \le f''(t) \le M \quad \text{for all } t \in \mathsf{U}.$$

Prove that there is a value $\xi \in [m, M]$ for which

$$\mathbb{E}[\varphi(X)] - \varphi(\mathbb{E}\, X) = \tfrac{1}{2}\xi \cdot \operatorname{Var}[X].$$

Hint: Apply Jensen's inequality to the two convex functions

$$\varphi_0(t) := \tfrac{1}{2}Mt^2 - \varphi(t);$$
$$\varphi_1(t) := \varphi(t) - \tfrac{1}{2}mt^2.$$

Problem 12.26 (Madame Covary). In this problem, we will explore the geometry of covariance and correlation. Let $\mathsf{K} \subseteq \mathsf{L}_2$ be the set of real random variables in L_2 with expectation zero.

1. Explain why K is a linear subspace. **(*)** Argue that K is complete.
2. For $X \in \mathsf{L}_2$, find an explicit, simple formula for an orthogonal projection of X onto K.
3. Use (2) to define an orthogonal projection map $\boldsymbol{P} : \mathsf{L}_2 \to \mathsf{K}$. Verify that $\boldsymbol{P}$ is linear, idempotent ($\boldsymbol{P}^2 = \boldsymbol{P}$), and self-adjoint with respect to the L_2 pseudo-inner product ($\langle \boldsymbol{P}X,\ Y\rangle = \langle X,\ \boldsymbol{P}Y\rangle$).
4. For $X, Y \in \mathsf{L}_2$, express the covariance $\operatorname{Cov}(X, Y)$ and variance $\operatorname{Var}[X]$ in terms of the L_2 pseudo-inner product and the map $\boldsymbol{P}$.
5. Sketch a pair of random variables with strictly positive variance that have $\varrho(X, Y) = 0$.
6. Confirm that $|\varrho(X, Y)| \le 1$. **(*)** Determine conditions for equality.

Problem 12.27 (Pavlov). Orthogonal projection is closely connected with conditional expectation. This problem begins to develop your intuition. Let $X, Y \in \mathsf{L}_2$. Define

$$\mathsf{K} := \mathsf{K}_Y := \{Z \in \mathsf{L}_2 : Z = h(Y) \text{ for measurable } h : \mathbb{R} \to \mathbb{R}\}.$$

The orthogonal projection of X onto K_Y provides a first definition of the conditional expectation $\mathbb{E}[X \mid Y]$. Technical details will appear in Lecture 19.

1. Show that K is a complete linear subspace of L_2. **Hint:** Show that K_Y is isomorphic to $\mathsf{L}_2(\mu_Y)$, which is complete.
2. Let $X = g(Y) \in \mathsf{L}_2$ for a measurable function $g : \mathbb{R} \to \mathbb{R}$. Find an orthogonal projection of X onto K.
3. If the pair (X, Y) is independent, find an orthogonal projection of X onto K.
4. **(*)** Assume that X is a discrete random variable that takes integer values. Then X has the law $\mu_X = \sum_{i\in\mathbb{Z}} p_i \delta_i$ where $\sum_{i\in\mathbb{Z}} p_i = 1$ and $p_i \ge 0$. Consider the random variable $Y = |X|$. Find an orthogonal projection of X onto K.
5. **(*)** Let $Y = g(X)$ for measurable g. Find an orthogonal projection of X onto K.

Problem 12.28 (Orthogonal projection: Properties). Orthogonal projection of random variables shares many properties with orthogonal projection in a finite-dimensional linear space. Let K be a complete linear subspace of L_2, and define the orthogonal complement

$$\mathsf{K}^\perp := \{Y \in \mathsf{L}_2 : \langle Y,\ X\rangle = 0 \text{ for all } X \in \mathsf{K}\}.$$

For a random variable $X \in \mathsf{L}_2$, let $\boldsymbol{P}(X) := \boldsymbol{P}_{\mathsf{K}}(X)$ denote any version of the orthogonal projection of X onto K. Define $\boldsymbol{P}^\perp(X) := X - \boldsymbol{P}(X)$ to be a version of the error in the orthogonal projection.

Properly speaking, $\boldsymbol{P}$ and $\boldsymbol{P}^\perp$ are not functions, so you should use this notation with caution.

1. **Orthogonal complement:** Show that $\mathsf{K}^\perp$ is a complete linear subspace of L_2. Show that $\boldsymbol{P}^\perp(X)$ is a version of the orthogonal projection of X onto $\mathsf{K}^\perp$.
2. **Linearity:** Let $X, Y \in \mathsf{L}_2$ and $\alpha, \beta \in \mathbb{R}$. Show that $\alpha\boldsymbol{P}(X) + \beta\boldsymbol{P}(Y)$ is a version of the orthogonal projection of $\alpha X + \beta Y$ onto K. Show that $\alpha\boldsymbol{P}^\perp(X) + \beta\boldsymbol{P}^\perp(Y)$ is a version of the orthogonal projection of $\alpha X + \beta Y$ onto $\mathsf{K}^\perp$.
3. **Idempotency:** Show that $\boldsymbol{P}(\boldsymbol{P}(X)) = \boldsymbol{P}(X)$ almost surely. Establish the same property for $\boldsymbol{P}^\perp$.
4. **Orthogonality:** Show that $\langle \boldsymbol{P}(X),\ \boldsymbol{P}^\perp(X)\rangle = 0$.
5. **Orthogonal decomposition:** Confirm that $\boldsymbol{P}(X) + \boldsymbol{P}^\perp(X) = X$ almost surely.
6. **Pythagorean theorem:** Show that $\|\boldsymbol{P}(X)\|_2^2 + \|\boldsymbol{P}^\perp(X)\|_2^2 = \|X\|_2^2$.
7. **Contraction:** Deduce that
$$\sup\{\|\boldsymbol{P}(X)\|_2 : \|X\|_2 \le 1\} = 1;$$
$$\sup\{\|\boldsymbol{P}^\perp(X)\|_2 : \|X\|_2 \le 1\} = 1.$$
8. **Nesting:** Suppose that $\mathsf{M} \subseteq \mathsf{K}$ is a complete linear subspace. Show that the orthogonal projection of X onto M coincides with the orthogonal projection of $\boldsymbol{P}_\mathsf{K}(X)$ onto M. That is, $\boldsymbol{P}_\mathsf{M} \circ \boldsymbol{P}_\mathsf{K} = \boldsymbol{P}_\mathsf{M}$. In addition, show that $\boldsymbol{P}_\mathsf{K} \circ \boldsymbol{P}_\mathsf{M} = \boldsymbol{P}_\mathsf{M}$.

Problem 12.29 (Riesz representation). A *linear functional* on L_2 is a linear map $\varphi : \mathsf{L}_2 \to \mathbb{R}$ that takes real values. The Riesz representation theorem states that every bounded linear functional can be represented as an inner product. That is, there is a random variable $Y \in \mathsf{L}_2$, depending only on φ, for which $\varphi(X) = \langle Y,\ X\rangle$. The purpose of this problem is to establish this important fact.

1. **Uniqueness:** Suppose that there are random variables $Y, Y' \in \mathsf{L}_2$ that both represent the linear functional: $\varphi(X) = \langle Y,\ X\rangle$ and $\varphi(X) = \langle Y',\ X\rangle$. Show that $Y = Y'$ almost surely. **Hint:** Consider $X = Y - Y'$.
2. **Trivial case:** If $\varphi(X) = 0$ for all $X \in \mathsf{L}_2$, find a representation of φ as an inner product. From now on, assume that $\varphi \neq 0$.
3. **Continuity:** The norm of the linear functional is defined as
$$\|\varphi\| := \sup\{\|\varphi(X)\|_2 : \|X\|_2 \le 1\}.$$
We say that φ is *bounded* when $\|\varphi\| < +\infty$. Show that φ is continuous if and only if it is bounded.
4. **Null space:** Introduce the null space $\mathsf{K} := \{X \in \mathsf{L}_2 : \varphi(X) = 0\}$ of the linear functional. Show that K is a complete linear subspace of L_2. **Hint:** K is the inverse image of a closed set under a bounded linear map.
5. **Orthogonal complement:** Show that $\mathsf{K}^\perp$ is a complete linear subspace of L_2. Argue that $\mathsf{K}^\perp$ contains a nontrivial random variable Z. More precisely, "nontrivial" means that $\{Z \neq 0\}$ is an event with strictly positive probability.
6. **A projection onto the kernel:** Fix a random variable $Z \in \mathsf{K}^\perp$ with $\|Z\|_2 = 1$. For an arbitrary random variable $X \in \mathsf{L}_2$, define the random variable $W := W(X) := \varphi(X)Z - \varphi(Z)X$. Show that $W \in \mathsf{K}$. (*) In fact, $W(X)$ is a version of the orthogonal projection of X onto K. Why?
7. **Representation:** Continuing from the last part, use the observation that $\langle Z,\ W\rangle = 0$ to deduce that $\varphi(X) = \langle Y,\ X\rangle$ for a random variable $Y \in \mathsf{L}_2$ that does not depend on X.
8. **Norm:** Show that $\|Y\|_2 = \|\varphi\|$.

Problem 12.30 (*Orthogonal projection: Convex set). A set $\mathsf{K} \subseteq \mathsf{L}_2$ of random variables is *convex* if
$$X, Y \in \mathsf{K} \quad \text{implies} \quad (1-\tau)X + \tau Y \in \mathsf{K} \quad \text{for all } \tau \in [0,1].$$

We say that a (convex) subset $\mathsf{K} \subseteq \mathsf{L}_2$ is *closed* if every Cauchy sequence in K converges to a limit in K.

Let $X \in \mathsf{L}_2$ be a random variable, and let $\mathsf{K} \subseteq \mathsf{L}_2$ be closed and convex. A random variable $Y \in \mathsf{K}$ is called an *orthogonal projection* of X onto K if

$$\|X - Y\|_2 = \inf\{\|X - W\|_2 : W \in \mathsf{K}\}.$$

1. Show that a linear subspace in L_2 is convex.
2. Show that $\{Z \in \mathsf{L}_2 : \|Z\|_p \leq 1\}$ is convex and closed for $p \geq 2$.
3. If K is closed and convex, prove that every random variable $X \in \mathsf{L}_2$ has an orthogonal projection onto K.
4. What is the dual characterization of an orthogonal projection? In other words, what can we say about $\langle Z,\ X - Y \rangle$ for $Z \in \mathsf{K}$?
5. Show that any point $Y \in \mathsf{K}$ that satisfies the dual characterization is an orthogonal projection.

Applications

Application 12.31 (Gauss & Markov). Suppose that we observe paired real-valued data $((x_i, y_i) : i = 1, \dots, n)$. For unknown $a_\star \in \mathbb{R}$, suppose that the responses follow the linear model

$$y_i = a_\star x_i + \eta_i \quad \text{for } i = 1, \dots, n.$$

We assume that the random errors satisfy $\mathbb{E}[\eta_i] = 0$ and $\mathrm{Var}[\eta_i] = \sigma^2$ for all i and $\mathrm{Cov}(\eta_i, \eta_j) = 0$ for all $i \neq j$. We can estimate the true model by means of an ordinary least-squares (OLS) problem:

$$\mathsf{minimize}_{a \in \mathbb{R}} \quad \frac{1}{2} \sum\nolimits_{i=1}^{n} (y_i - a x_i)^2.$$

Let $\hat{a}$ be the solution to the OLS problem. This problem explores the motivation for using OLS.

1. By direct calculation, confirm that the estimator $\hat{a}$ is a linear function of the observed responses (y_i). In this context, we think about the covariates (x_i) as fixed numbers.
2. Check that the estimator is unbiased: $\mathbb{E}[\hat{a}] = a_\star$.
3. Among unbiased estimators for $a_\star$ as a *linear* function of (y_i), prove that the OLS estimator $\hat{a}$ has minimum variance. This fact is called the Gauss-Markov theorem.

Let us upgrade to a multivariate linear model. Consider paired data $(\boldsymbol{x}_i, y_i) \in \mathbb{R}^n \times \mathbb{R}$ for $i = 1, \dots, n$. For an (unknown) vector $\boldsymbol{a}_\star \in \mathbb{R}^n$, suppose that the responses follow the linear model

$$y_i = \boldsymbol{a}_\star^{\mathsf{T}} \boldsymbol{x}_i + \eta_i \quad \text{for } i = 1, \dots, n.$$

Maintain the same assumptions on the random errors η_i. We can estimate the true model by means of the ordinary least-squares problem:

$$\mathsf{minimize}_{\boldsymbol{a} \in \mathbb{R}^n} \quad \frac{1}{2} \sum\nolimits_{i=1}^{n} (y_i - \boldsymbol{a}^{\mathsf{T}} \boldsymbol{x}_i)^2.$$

Let $\hat{\boldsymbol{a}}$ be the solution to the OLS problem.

4. Show that $\hat{\boldsymbol{a}}$ is the minimum-variance unbiased estimator of $\boldsymbol{a}_\star$.

Application 12.32 (*Cramér, Rao, and Fisher). In statistics, we try to construct methods for making inferences from data. Consider a distribution that is specified by one or more parameters. For instance, a normal distribution $\textsc{normal}(m, \sigma^2)$ depends on the mean $m \in \mathbb{R}$ and the variance $\sigma^2 > 0$. Given some samples drawn from the distribution, we may try to estimate one of the parameters of the distribution.

There are two simple quantities that reflect the quality of an estimator. The *bias* measures how far, on average, the parameter estimate lies from the true parameter value. The *variance* measures how much the parameter estimate fluctuates, on average, over the choice of a random sample. Among all estimators with a given bias, we prefer the one with the lowest variability. Therefore, to evaluate the quality of a particular estimator, it is helpful to have a lower bound on the variance of the estimator. This problem explores a fundamental method for producing such a bound.

Consider a parameterized family of probability density functions $f_\theta : \mathbb{R}^d \to \mathbb{R}_+$ For simplicity, we assume that the parameter $\theta \in \mathsf{U}$, an open interval of the real line, and that $\theta \mapsto f_\theta(\boldsymbol{x})$ is differentiable for each $\boldsymbol{x} \in \mathbb{R}^d$. Define the *score function*

$$s(\theta; \boldsymbol{x}) := \partial_\theta \big(\log f_\theta(\boldsymbol{x}) \big) \quad \text{for } \theta \in \mathsf{U} \text{ and } \boldsymbol{x} \in \mathbb{R}^d.$$

The logarithm of the density reflects the likelihood that the parameter value is θ, given an observed value $\boldsymbol{x} \in \mathbb{R}^d$. The score function measures how quickly this likelihood changes as we vary the parameter θ.

1. Let $W, S \in \mathsf{L}_2$. Show that
$$\operatorname{Var}[W] \geq \frac{|\operatorname{Cov}(W, S)|}{\operatorname{Var}[S]}.$$
2. Compute the score function of the density f_θ of a $\textsc{normal}(\theta, \sigma^2)$ distribution on $\mathbb{R}$. Compute the score function of the density of a $\textsc{normal}(m, \theta)$ distribution on $\mathbb{R}$ where $\theta > 0$.
3. The score function is very useful for working with complicated densities that are hard to normalize. For a bounded, measurable function $h : \mathbb{R} \to \mathbb{R}$, consider the *Gibbs distribution*:
$$f_\theta(x) := \frac{1}{Z_\theta} \mathrm{e}^{-\theta h(x)} \quad \text{for } x \in \mathbb{R}.$$
The constant Z_θ is chosen to ensure that f_θ is a density. Compute the score function.
4. Consider a family $(\boldsymbol{X}_\theta : \theta \in \mathsf{U})$ of random variables taking values in $\mathbb{R}^d$. Suppose that $\boldsymbol{X}_\theta$ has density $f_\theta : \mathbb{R}^d \to \mathbb{R}_+$. Under appropriate regularity conditions, show that the expected score is zero:
$$\mathbb{E}[s(\theta; \boldsymbol{X}_\theta)] = 0.$$
Hint: Differentiate the relation $\int f_\theta(\boldsymbol{x}) \, \lambda^d(\mathrm{d}\boldsymbol{x}) = 1$, and use dominated convergence to draw the derivative through the integral.
5. Define the *Fisher information* that $\boldsymbol{X}_\theta$ contains about the parameter θ:
$$I(\theta) := \operatorname{Var}[s(\theta; \boldsymbol{X}_\theta)].$$
Confirm that
$$I(\theta) = \int_{\mathbb{R}^d} \left(\frac{(\partial_\theta f_\theta)(\boldsymbol{x})}{f_\theta(\boldsymbol{x})} \right)^2 f_\theta(\boldsymbol{x}) \, \lambda^d(\mathrm{d}\boldsymbol{x}).$$

6. Calculate the Fisher information for $X_\theta \sim \text{NORMAL}(\theta, \sigma^2)$. Calculate the Fisher information for $X_\theta \sim \text{NORMAL}(m, \theta)$.
7. Let $W : \mathbb{R}^d \to \mathbb{R}$ be a measurable function. Under appropriate regularity conditions, show that

$$\text{Cov}\big(W(\boldsymbol{X}_\theta), s(\theta; \boldsymbol{X}_\theta)\big) = \partial_\theta\big(\mathbb{E}[W(\boldsymbol{X}_\theta)]\big).$$

   This is essentially just integration by parts. You can discover the definitions of score functions and Fisher information if you start at the right-hand side of this relation and work backward.
8. Define the mean function $\tau(\theta) := \mathbb{E}[W(\boldsymbol{X}_\theta)]$. Establish the Cramér–Rao inequality:

$$\text{Var}[W(\boldsymbol{X}_\theta)] \geq \frac{|\partial_\theta \tau(\theta)|}{I(\theta)}.$$

What does this have to do with statistics? Suppose that $W(\boldsymbol{X}_\theta)$ is an *estimator* for the parameter θ. That is, given an observation $\boldsymbol{X}_\theta$, the function $W(\boldsymbol{X}_\theta)$ produces an approximation for the parameter θ. Since the data $\boldsymbol{X}_\theta$ is random, the value $W(\boldsymbol{X}_\theta)$ of the estimator is random too. Define the *bias* of the estimator:

$$\text{bias}(W(\boldsymbol{X}_\theta)) := \tau(\theta) - \theta.$$

We say that $W(\boldsymbol{X}_\theta)$ is an *unbiased estimator* of θ if the bias equals zero.

9. For an unbiased estimator, find a simplification of the Cramér–Rao inequality.
10. Let $X \sim \text{NORMAL}(\theta, \sigma^2)$ where σ^2 is known. Given this single observation, find an unbiased estimator $W(X)$ for the mean θ. Compute the variance, $\text{Var}[W(X)]$, of the estimator. Compare the result with the Cramér–Rao bound.
11. For $\theta > 0$, suppose we know that $X \sim \text{NORMAL}(0, \theta)$. Find an unbiased estimator $W(X)$ for the variance parameter θ. Compute the variance $\text{Var}[W(X)]$ of the estimator. Compare the result with the Cramér–Rao bound.
12. **(*)** Suppose that X is a continuous real random variable with density f_θ. Let $(X_1, \dots, X_n)$ be an independent family where each X_i has the same distribution as X. Show that $I(X_1, \dots, X_n) = n \cdot I(X)$. In particular, we can simplify the Cramér–Rao bound when we collect i.i.d. data.
13. **(*)** There is nothing special about continuous random variables. Develop an analog of this theory in case that $\boldsymbol{X}_\theta$ has a density with respect to some other measure μ on $\mathbb{R}^d$.

Recall that "i.i.d." stands for independent and identically distributed.

In Application 7.14, we saw that probability can be used to verify the existence of objects that have distinguished properties. Application 9.49 showed how we can implement this program by computing the expectation of a random variable. There is a further extension called the *second-moment method*, which uses information about the first two moments in combination with the inequalities from Exercise 12.23. The second-moment method is a useful tool in graph theory, number theory, additive combinatorics, algorithms (e.g., constraint satisfaction problems), and other areas. Most interesting applications of the second-moment method are a bit complicated, but we can offer an elegant example from analytic number theory.

Application 12.33 (Second-moment method: Prime factors). A powerful intuition from number theory is that prime numbers are "randomly distributed". Of course, this statement is not literally true, but it underlies applications of prime numbers in computer science (including crytography, fingerprinting, etc.). It also invites the use of probabilistic methods in number theory.

In this problem, we establish a special case of the Turán–Kubilius inequality. Given a random integer from a finite interval, this result provides an estimate for the number of prime factors contained in a given set.

Define the set $\mathsf{N} := \{1, 2, 3, \dots, N\}$ of natural numbers; assume $N \geq 6$. Fix an arbitrary finite set $\mathsf{P} \subset \mathbb{N}$ of prime numbers, each smaller than P, where $P \leq N$. We define the *logarithmic size* of the set P to be

$$\ell(\mathsf{P}) := \sum\nolimits_{p \in \mathsf{P}} p^{-1}.$$

Draw a random number $I \sim \textsc{uniform}(\mathsf{N})$. The random number I has a random number W of prime factors in the set P:

$$W := \#\{p \in \mathsf{P} : p \mid I\}.$$

For natural numbers $a, b \in \mathbb{N}$, recall that $a \mid b$ means that a divides b.

Our goal is to study the behavior of W and deduce number-theoretic conclusions.

1. For each natural number $d \leq N$, define the indicator random variable $X_d := \mathbb{1}_{d \mid I}$. Prove that
$$\left|\mathbb{E}[X_d] - d^{-1}\right| \leq \frac{1}{N};$$
$$\left|\mathrm{Var}[X_d] - d^{-1}(1 - d^{-1})\right| \leq \frac{1}{N}.$$
Hint: In the interval N, how many numbers are divisible by d?
2. Observe that the random variable $W = \sum_{p \in \mathsf{P}} X_p$. Calculate that the expectation satisfies the bound
$$|\mathbb{E}\, W - \ell(\mathsf{P})| \leq \frac{P}{N}.$$
Recall that $\ell(\mathsf{P})$ is the logarithmic size of the set P.
3. Show that there exists a number $n \in \mathsf{N}$ with at least $\ell(\mathsf{P}) - 1$ prime factors from the set P. Show that there is a number $n \in \mathsf{N}$ with at most $\ell(\mathsf{P}) + 1$ prime factors from P.
4. For *distinct prime* numbers $p, q \in \mathbb{N}$, show that
$$|\mathrm{Cov}(X_p, X_q)| \leq \frac{1}{N}.$$
Hint: Note that $p \mid I$ and $q \mid I$ if and only if $pq \mid I$.
5. Calculate the variance of W:
$$\mathrm{Var}[W] \leq \ell(\mathsf{P}) + \frac{P^2}{2N}.$$
Hint: Use the fact that W is a sum to write the variance as a (double) sum of covariances.
6. For $t > 0$, verify that
$$\mathbb{P}\left\{|W - \ell(\mathsf{P})| \geq t\sqrt{\ell(\mathsf{P})}\right\} \leq \frac{1}{t^2}\left[1 + \frac{P^2}{2N\ell(\mathsf{P})}\right].$$
7. Suppose that $P^2 \leq N \leq P^3$. Argue that an integer $n \leq N$ can have at most two prime factors larger than P.
8. **(***Mertens)** Consider the set $\mathsf{P} = \{2, 3, 5, 7, 11, \dots, P\}$. Prove Mertens's second theorem:
$$|\ell(\mathsf{P}) - \log\log P| \leq \text{Const}.$$
Hint: There is an elementary argument; see [YY87, Probs. 171–174].

9. For large N, deduce that most numbers in $\{1, \dots, N\}$ have about $\log \log N$ distinct prime factors. Give a formal mathematical statement of this result.
10. **(*)** There is nothing special about the set $\mathsf{N} = \{1, \dots, N\}$. Extend these results to any *interval* of the natural numbers with cardinality N.

Notes

Our development of L_2 spaces is also inspired by Williams [Wil91, Chap. 6]. The proof of Theorem 12.21 is drawn from his book, although the argument is standard.

For more on inequalities in L_2 spaces, see the books of Garling [Gar07], Lieb & Loss [LL01], and Steele [Ste04]. The treatment of the Cramér–Rao inequality is adapted from [CB90; LC98]. The example of the second-moment method is adapted from treatments by Alon & Spencer [AS16] and by Tao [Taob].

Lecture bibliography

[AS16] N. Alon and J. H. Spencer. *The probabilistic method*. Fourth. John Wiley & Sons, 2016.

[CB90] G. Casella and R. L. Berger. *Statistical inference*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1990.

[Gar07] D. J. H. Garling. *Inequalities: a journey into linear analysis*. Cambridge University Press, 2007. DOI: 10.1017/CBO9780511755217.

[LC98] E. L. Lehmann and G. Casella. *Theory of point estimation*. Second. Springer-Verlag, New York, 1998.

[LL01] E. H. Lieb and M. Loss. *Analysis*. 2nd ed. American Mathematical Society, 2001. DOI: 10.1090/gsm/014.

[Ste04] J. M. Steele. *The Cauchy-Schwarz master class*. An introduction to the art of mathematical inequalities. Mathematical Association of America / Cambridge Univ. Press, 2004. DOI: 10.1017/CBO9780511817106.

[Taob] T. Tao. *Second-moment and entropy methods*. URL: https://terrytao.wordpress.com/2019/11/12/254a-notes-9-second-moment-and-entropy-methods/.

[Wil91] D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

[YY87] A. M. Yaglom and I. M. Yaglom. *Challenging mathematical problems with elementary solutions. Vol. II*. Problems from various branches of mathematics, Translated from the Russian by James McCawley, Jr., Reprint of the 1967 edition. Dover Publications, Inc., New York, 1987.

13. Independence

> "When in the Course of human events, it becomes necessary for one people to dissolve the political bands which have connected them with another, and to assume among the powers of the earth, the separate and equal station to which the Laws of Nature and of Nature's God entitle them, a decent respect to the opinions of mankind requires that they should declare the causes which impel them to the separation."
>
> —*The Declaration of Independence*, 4 July 1776

Agenda:

1. Elementary independence
2. Independence of σ-algebras
3. Independence and product measure
4. Kolmogorov extension theorem

One of the features of probability theory that distinguishes it from ordinary measure theory is the idea of independence and the related idea of conditioning. In this lecture, we will begin to explore what it means for a collection of probabilistic experiments to be independent from each other.

So, what do we mean when we say that two experiments are independent? Heuristically, the outcome of one experiment has no bearing on the outcome of the second. If we know the outcome of the first experiment, the distribution of outcomes of the second experiment remains unchanged.

More concretely, suppose that I flip a coin and you roll a die. If these experiments are independent, then we anticipate that...

1. The face of the coin has no influence on the value of the die.
2. No event involving the coin informs us about any event involving the die.
3. No random variable determined only by the outcome of the coin flip is correlated with any random variable determined only by the outcome of the die.

You can start to appreciate that these desiderata are much stronger than the assumption that two particular random variables are uncorrelated.

In this lecture, we first summarize the elementary notions of independence from basic probability theory. Afterward, we discuss how to generalize these ideas to reach a definition of what it means for two collections of events to be independent from each other. We will see that this general definition subsumes all of the elementary notions of independence. Last, we will explore the connection between independence of random variables and product measures.

13.1 Elementary independence

We begin with the definition of the elementary conditional probability, which is the probability that one event occurs, given that another event has occurred. Two events are independent if the occurrence of one event does not change the probability that the other event occurs. Afterward, we discuss what it means for two random variables to be independent.

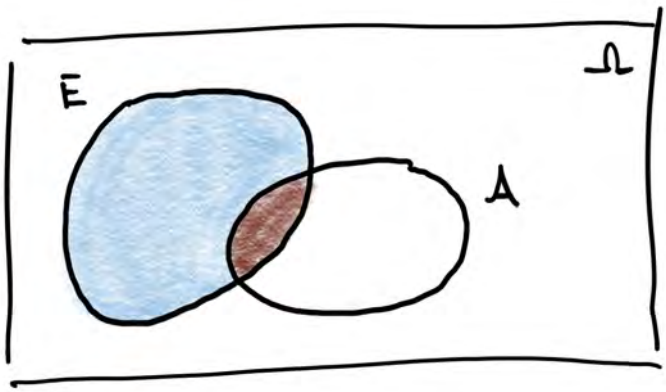


Figure 13.1 (Conditional probability). Given that the event E occurs, the event A occurs if and only if $\mathsf{A} \cap \mathsf{E}$ occurs (red). Meanwhile, the event A does not occur if and only if $\mathsf{A}^{\mathsf{c}} \cap \mathsf{E}$ occurs (blue).

13.1.1 Conditional probability

As usual, we fix a master probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Consider an event $\mathsf{E} \in \mathcal{F}$ with strictly positive probability: $\mathbb{P}(\mathsf{E}) > 0$. Suppose that we have knowledge that the event E occurs. That is, the distinguished sample point $\omega_0 \in \mathsf{E}$. In general, this piece of information is not sufficient to determine the sample point ω_0 completely. Nevertheless, it does allow us to update our prior knowledge of the probability that another event, say $\mathsf{A} \in \mathcal{F}$, occurs. If E occurs, then the only way for A to occur is for $\mathsf{A} \cap \mathsf{E}$ to occur. Likewise, the only way for A not to occur is for $\mathsf{A}^{\mathsf{c}} \cap \mathsf{E}$ to occur. In other words, we want to restrict our attention to events of the form $\mathsf{B} \cap \mathsf{E}$ for $\mathsf{B} \in \mathcal{F}$. See Figure 13.1.

This observation suggests that we need to consider a new distribution of probability over events $\mathsf{B} \cap \mathsf{E}$ with $\mathsf{B} \in \mathcal{F}$. Each of these events has an *a priori* probability $\mathbb{P}(\mathsf{B} \cap \mathsf{E})$. But these values may not compose a probability distribution because $\mathbb{P}(\Omega \cap \mathsf{E}) = \mathbb{P}(\mathsf{E})$, which may not equal one. Therefore, we need to rescale the prior probabilities to obtain a new probability distribution over the events restricted to E. These arguments lead to an important elementary definition.

Definition 13.1 (Elementary conditional probability: Events). Let $\mathsf{E} \in \mathcal{F}$ be an event with strictly positive probability: $\mathbb{P}(\mathsf{E}) > 0$. The probability that an event $\mathsf{A} \in \mathcal{F}$ occurs, given that E occurs, is defined as

$$\mathbb{P}(\mathsf{A} \,|\, \mathsf{E}) := \frac{\mathbb{P}(\mathsf{A} \cap \mathsf{E})}{\mathbb{P}(\mathsf{E})}.$$

In other words, the conditional probability $\mathbb{P}(\mathsf{A} \,|\, \mathsf{E})$ is the proportion of the probability $\mathbb{P}(\mathsf{E})$ that is attributable to the event $\mathbb{P}(\mathsf{A} \cap \mathsf{E})$ occurring.

Exercise 13.2 (Elementary conditional probability: Events). With the assumptions of Definition 13.1, confirm that $\{\mathsf{B} \cap \mathsf{E} : \mathsf{B} \in \mathcal{F}\}$ is a σ-algebra contained in $\mathcal{F}$, called the *restriction* of $\mathcal{F}$ to the event E. Check that $\mathbb{P}(\cdot \,|\, \mathsf{E})$ is a probability distribution on the restricted σ-algebra.

13.1.2 Independence of events

The definition of conditional probability leads directly to the notion of independence. Two events A, B with strictly positive probability are independent when knowledge that B occurs does not change the probability that A occurs, and conversely:

$$\mathbb{P}(\mathsf{A} \,|\, \mathsf{B}) = \mathbb{P}(\mathsf{A}) \quad \text{and} \quad \mathbb{P}(\mathsf{B} \,|\, \mathsf{A}) = \mathbb{P}(\mathsf{B}).$$

Using Definition 13.1, we quickly see that each one of these relations is equivalent to the condition that $\mathbb{P}(\mathsf{A} \cap \mathsf{B}) = \mathbb{P}(\mathsf{A}) \cdot \mathbb{P}(\mathsf{B})$. We enshrine this formula in a definition.

Definition 13.3 (Elementary independence: Events). Two events $\mathsf{A}, \mathsf{B} \in \mathscr{F}$ are *independent* when

$$\mathbb{P}(\mathsf{A} \cap \mathsf{B}) = \mathbb{P}(\mathsf{A}) \cdot \mathbb{P}(\mathsf{B}). \tag{13.1}$$

Note that Definition 13.3 no longer requires an assumption that the probability of each event is strictly positive. Indeed, if either event has zero probability, then both sides of the relation (13.1) equal zero. Definition 13.3 agrees with the elementary notion of independence of events.

Exercise 13.4 (Independence: Complements). Show that the events A, B are independent if and only if the events A, B^c are independent.

13.1.3 Independence of random variables

Next, we turn to a definition of independence for random variables. To say that two random variables are independent, we want to make sure that neither one of the random variables provides information about the value of the other. We can formulate this idea in terms of the events associated with the random variables.

Definition 13.5 (Elementary independence: Random variables). Let X, Y be real random variables on a probability space. The random variables X and Y are *independent* when

$$\mathbb{P}\{X \in \mathsf{A} \text{ and } Y \in \mathsf{B}\} = \mathbb{P}\{X \in \mathsf{A}\} \cdot \mathbb{P}\{Y \in \mathsf{B}\} \quad \text{for all } \mathsf{A}, \mathsf{B} \in \mathscr{B}(\mathbb{R}). \tag{13.2}$$

We sometimes write $X \perp\!\!\!\perp Y$ to mean that X and Y are independent random variables. See Figure 13.2.

In Section 13.2, we will connect this definition with the previous Definition 8.25 of independent random variables.

Warning: Do not conflate orthogonality ($X \perp Y$) with independence ($X \perp\!\!\!\perp Y$)! ■

In other words, two random variables X and Y are independent when the probability that $(X, Y) \in \mathsf{A} \times \mathsf{B}$ equals the product of the probabilities that $X \in \mathsf{A}$ and $Y \in \mathsf{B}$ for all Borel sets A, B. From this fact, independence appears to be related to product measures; we will pursue this observation in Section 13.2.

Definition 13.5 involves a lot of events, but this is inevitable because we must be sure that no set of values of X informs us about the probability of any set of values of Y occurring. As a particular consequence,

$$\mathbb{P}\{X \le a \text{ and } Y \le b\} = \mathbb{P}\{X \le a\} \cdot \mathbb{P}\{Y \le b\} \quad \text{for all } a, b \in \mathbb{R}. \tag{13.3}$$

The formulation (13.3) is the standard way of defining independence of random variables in introductory courses.

In fact, the relation (13.3) implies that the apparently stronger relation (13.2) holds. This claim requires Dynkin's theorem on intersection-stable systems; see Example E.7.

Exercise 13.6 (Independence: Indicators). Check that two events A, B are independent if and only if their indicator random variables $\mathbb{1}_{\mathsf{A}}$ and $\mathbb{1}_{\mathsf{B}}$ are independent.

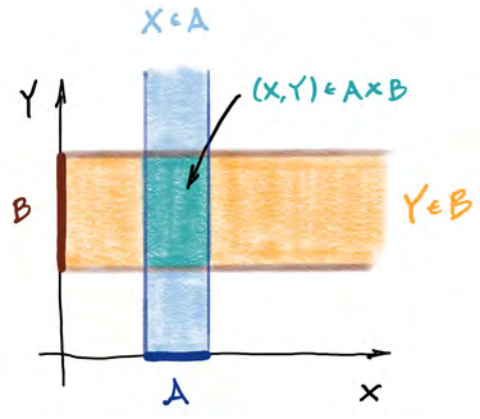


Figure 13.2 (Independence: Random variables). A pair (X, Y) of random variables is *independent* when the probability that $(X, Y) \in \mathsf{A} \times \mathsf{B}$ equals the product of the probability that $X \in \mathsf{A}$ and the probability that $Y \in \mathsf{B}$.

13.2 Independence and product measures

The elementary definition of independence for random variables (Definition 13.5) suggests a connection between independence and product measure. This insight is valid, and it works both directions.

13.2.1 The joint distribution of an independent pair of random variables

Suppose that X and Y are independent random variables. According to Definition 13.5, this statement means precisely that

$$\mathbb{P}\{X \in \mathsf{A}, Y \in \mathsf{B}\} = \mathbb{P}\{X \in \mathsf{A}\} \cdot \mathbb{P}\{Y \in \mathsf{B}\} \quad \text{for all } \mathsf{A}, \mathsf{B} \in \mathscr{B}(\mathbb{R}).$$

This relation determines the (joint) distribution μ_{XY} of the pair $(X, Y) \in \mathbb{R}^2$ on rectangles in terms of the marginal distributions μ_X and μ_Y:

$$\mu_{XY}(\mathsf{A} \times \mathsf{B}) = \mu_X(\mathsf{A}) \cdot \mu_Y(\mathsf{B}) \quad \text{for all } \mathsf{A}, \mathsf{B} \in \mathscr{B}(\mathbb{R}).$$

Theorem 6.14 states that the distribution μ_{XY} has a unique extension to the product σ-algebra $\mathscr{B}(\mathbb{R}) \times \mathscr{B}(\mathbb{R}) = \mathscr{B}(\mathbb{R}^2)$. This extension is the product measure given by the marginal distributions. That is,

$$\mu_{XY} = \mu_X \times \mu_Y.$$

We encapsulate this argument in a proposition.

Proposition 13.7 (Independent random variables: Joint distribution). Consider *independent* real random variables X and Y with distributions μ_X and μ_Y on the Borel sets of the real line. Then the joint distribution of the pair (X, Y) is the product measure $\mu_{XY} = \mu_X \times \mu_Y$ on the Borel sets in $\mathbb{R}^2$. That is, Definition 8.25 is consistent with Definition 13.5.

13.2.2 The product measure defines an independent pair of random variables

Conversely, suppose that μ_X and μ_Y are probability distributions on the Borel sets of the real line. Our goal is to build a probability space that supports two *independent* random variables X and Y with marginal laws μ_X and μ_Y.

To do so, we simply construct the product probability space $(\mathbb{R}^2, \mathscr{B}(\mathbb{R}^2), \mu_X \times \mu_Y)$. Introduce the coordinate functions:

$$X = \pi_1(\omega_1, \omega_2) = \omega_1 \quad \text{and} \quad Y = \pi_2(\omega_1, \omega_2) = \omega_2 \quad \text{for } \boldsymbol{\omega} \in \mathbb{R}^2.$$

You should check that the marginal distribution of X is μ_X, while the marginal distribution of Y is μ_Y. The pair (X, Y) obviously has the joint distribution $\mu_{XY} = \mu_X \times \mu_Y$, so

$$\mathbb{P}\{(X, Y) \in \mathsf{A} \times \mathsf{B}\} = \mathbb{P}\{X \in \mathsf{A}\} \cdot \mathbb{P}\{Y \in \mathsf{B}\} \quad \text{for all } \mathsf{A}, \mathsf{B} \in \mathscr{B}(\mathbb{R}).$$

In other words, the random variables X and Y are independent. We will generalize this result below in Theorem 13.24.

13.2.3 Expectation and independence

The connection between independence and product measures tells us how to compute the expectation of a function $h(X, Y)$. In particular, we have an elegant result for the expectation of a product function.

Proposition 13.8 (Independent random variables: Expectation). Suppose that (X, Y) is an independent pair of real random variables with marginal laws μ_X and μ_Y. Suppose that $f \in \mathsf{L}_1(\mu_X)$ and $g \in \mathsf{L}_1(\mu_Y)$. Then

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)] \cdot \mathbb{E}[g(Y)]. \tag{13.4}$$

Proof. The result is just an application of Fubini–Tonelli (Theorem 6.23). Indeed,

$$\begin{aligned}
\mathbb{E}[f(X)g(Y)] &= \int_{\mathbb{R}^2} f(x)g(y)\, \mu_{XY}(\mathrm{d}x \times \mathrm{d}y) \\
&= \int_{\mathbb{R}^2} f(x)g(y)\, (\mu_X \times \mu_Y)(\mathrm{d}x \times \mathrm{d}y) \\
&= \left(\int_{\mathbb{R}} f(x)\, \mu_Y(\mathrm{d}y) \right) \left(\int_{\mathbb{R}} g(y)\, \mu_Y(\mathrm{d}y) \right) = \mathbb{E}[f(X)] \cdot \mathbb{E}[g(Y)].
\end{aligned}$$

The first relation is (9.6), the (multivariate) law of the unconscious statistician. We have used Proposition 13.7 to see that the joint distribution is the product of the marginal distributions. Then we invoked Fubini–Tonelli to replace the integral over the product measure with an iterated integral. The last relation is Proposition 9.4, the law of the unconscious statistician. ∎

Exercise 13.9 (Independence and expectation). Let X and Y be real random variables with marginal laws μ_X and μ_Y. For all bounded, measurable functions $f, g : \mathbb{R} \to \mathbb{R}$, suppose that

$$\mathbb{E}[f(X)g(Y)] = \mathbb{E}[f(X)] \cdot \mathbb{E}[g(Y)]$$

Prove that X and Y are independent. **Hint:** Consider indicator functions.

Exercise 13.10 (Independence and functions). Let X and Y be real random variables. Prove that the pair (X, Y) is independent if and only if the pair $(f(X), g(Y))$ is independent for all measurable functions $f, g : \mathbb{R} \to \mathbb{R}$.

13.2.4 Independence versus uncorrelation

For comparison, recall that two real random variables $X, Y \in \mathsf{L}_2$ are *uncorrelated* when

$$\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y].$$

This is the very simplest case of the calculation (13.4). On the other hand, uncorrelation does not allow us to extend this result to any other functions. We can now appreciate that it is far easier for two random variables to be uncorrelated than for two random variables to be independent.

This discussion hints at the strength of the independence assumption. Even if we process each of the independent random variables in an arbitrary way (but without reference to the other), we cannot make them correlated with each other!

Exercise 13.11 (Independence: Expectation of a product). Let $X, Y \in \mathsf{L}_1$ be independent random variables, not necessarily in L_2. Show that $\mathbb{E}[XY] = \mathbb{E}[X] \cdot \mathbb{E}[Y]$.

Exercise 13.12 (Independence: Variance). Suppose that $X, Y \in \mathsf{L}_2$ are independent random variables. Show that $\operatorname{Var}[X + Y] = \operatorname{Var}[X] + \operatorname{Var}[Y]$.

13.3 Independence and σ-algebras

We can think about σ-algebras as carrying information. If we know whether each event in a σ-algebra occurs, then we have acquired some information about the distinguished sample point ω_0. It is natural to extend the concept of independence from events and random variables to a general notion of independence of σ-algebras. This definition is flexible enough to subsume the elementary definitions of independence from before—and more things besides.

13.3.1 Example: Coin flips

Consider the elementary probability experiment where we flip two fair coins. To model this experiment, we introduce the sample space $\Omega = \{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}$, the σ-algebra $\mathcal{F} = \mathcal{P}(\Omega)$, and the uniform probability measure $\mathbb{P}$ on the sample space.

Let us extract a sub-σ-algebra that contains the events that are determined by the first coin flip:

$$\mathcal{G}_1 := \{\emptyset, \{\mathrm{HH}, \mathrm{HT}\}, \{\mathrm{TH}, \mathrm{TT}\}, \Omega\}.$$

You may check that $\mathcal{G}_1$ is a σ-algebra contained in $\mathcal{F}$. If we see that the first coin turns up H, say, we can decide whether each of the events in $\mathcal{G}_1$ has occurred or not. So $\mathcal{G}_1$ captures the knowledge we attain by observing the value of the first coin flip.

Likewise, we can extract a sub-σ-algebra that contains the events that are determined by the second coin flip:

$$\mathcal{G}_2 := \{\emptyset, \{\mathrm{HH}, \mathrm{TH}\}, \{\mathrm{HT}, \mathrm{TT}\}, \Omega\}.$$

If we observe that the second coin turns up T, say, we can decide whether each of the events in $\mathcal{G}_2$ has occurred or not. The knowledge about the second coin does not determine whether either of the nontrivial events in $\mathcal{G}_1$ has occurred.

Now, consider a pair of events, each drawn from one of the sub-σ-algebras. For instance, let $\mathsf{G}_1 = \{\mathrm{HH}, \mathrm{HT}\}$ and $\mathsf{G}_2 = \{\mathrm{HH}, \mathrm{TH}\}$. Observe that

$$\mathbb{P}\{\mathsf{G}_1 \cap \mathsf{G}_2\} = \tfrac{1}{4} = \mathbb{P}\{\mathsf{G}_1\} \cdot \mathbb{P}\{\mathsf{G}_2\}.$$

In other words, the events G_1 and G_2 are independent. By further investigation, we can see that every event in $\mathcal{G}_1$ is independent from every event in $\mathcal{G}_2$. In other words, $\mathcal{G}_1$ and $\mathcal{G}_2$ provide independent pieces of information about the experiment.

13.3.2 Independent σ-algebras

We want to capture the idea that one σ-algebra provides no information about the events in another σ-algebra.

Definition 13.13 (Independence: σ-algebras). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let $\mathcal{G}_i \subseteq \mathcal{F}$ be a σ-algebra contained in $\mathcal{F}$ for each $i = 1, 2$. We say that the two σ algebras $\mathcal{G}_1$ and $\mathcal{G}_2$ are *independent* when

$$\mathbb{P}(\mathsf{G}_1 \cap \mathsf{G}_2) = \mathbb{P}(\mathsf{G}_1) \cdot \mathbb{P}(\mathsf{G}_2) \quad \text{for all } \mathsf{G}_1 \in \mathcal{G}_1 \text{ and all } \mathsf{G}_2 \in \mathcal{G}_2.$$

In words, independence means that every event in the first σ-algebra $\mathcal{G}_1$ is independent from every event in the second σ-algebra $\mathcal{G}_2$. Let us emphasize that Definition 13.13 involves not only events but also the probability measure. Independence reflects how the probabilities are assigned to events in the two σ-algebras.

13.3.3 Independent events

As a first application of Definition 13.13, let us explain how it captures the elementary notion of independence of events. Recall that an event $\mathsf{A} \in \mathcal{F}$ generates the σ-algebra $\sigma(\{\mathsf{A}\}) = \{\emptyset, \mathsf{A}, \mathsf{A}^c, \Omega\}$.

Definition 13.14 (Independence: Events). Let $\mathsf{A}, \mathsf{B} \in \mathcal{F}$ be events. We say that the events A and B are *independent* when the events generate σ-algebras $\sigma(\{\mathsf{A}\})$ and $\sigma(\{\mathsf{B}\})$ that are independent.

Exercise 13.15 (Independence: Events). Confirm that Definition 13.14 is equivalent with the elementary Definition 13.3.

Activity 13.16 (Sigma-algebra generated by events). Propose a definition of the σ-algebra generated by a countable family $(\mathsf{A}_i : i \in \mathbb{N})$ of events. ▪

13.3.4 Sigma-algebras generated by random variables

Before we continue, let us give a formal definition of the σ-algebra generated by some random variables.

Definition 13.17 (Sigma-algebra generated by a random variable). For a real random variable X, define

$$\sigma(X) := \sigma(\{X^{-1}(\mathsf{B}) : \mathsf{B} \in \mathcal{B}(\mathbb{R})\}).$$

More generally, for a countable family $(X_i : i \in \mathbb{N})$ of real random variables, define

$$\sigma(X_i : i \in \mathbb{N}) := \sigma\{X_i^{-1}(\mathsf{B}) : \mathsf{B} \in \mathcal{B}(\mathbb{R}) \text{ and } i \in \mathbb{N}\}.$$

In other words, the σ-algebra generated by a random variable X consists of all events that are the preimages of Borel sets. If we know the value $X(\omega)$ of the random variable, then we can determine whether or not an arbitrary event in $\sigma(X)$ has occurred. The general definition is similar in spirit.

13.3.5 Independent random variables

Definition 13.13 also contains the notion of independence for random variables.

Definition 13.18 (Independence: Random variables). Let X, Y be real random variables.

We say that the random variables X and Y are *independent* when they generate σ-algebras $\sigma(X)$ and $\sigma(Y)$ that are independent.

Exercise 13.19 (Independence: Random variables). Confirm that Definition 13.18 is equivalent with the elementary Definition 13.5.

Problem 13.20 (Dependence and measurability). Let X, Y be real random variables. Show that Y is measurable with respect to the σ-algebra $\sigma(X)$ if and only if $Y = f(X)$ for a Borel measurable function $f : \mathbb{R} \to \mathbb{R}$.

13.3.6 Independent families of σ-algebras

We can generalize Definition 13.13 further to address the independence of larger families of σ-algebras.

Definition 13.21 (Independence: Sequence of σ-algebras). Consider a *countable* collection $(\mathcal{G}_i \subseteq \mathcal{F} : i \in \mathbb{N})$ of sub-σ-algebras in $\mathcal{F}$. These σ-algebras are *independent* when

$$\mathbb{P}(\mathsf{G}_{i_1} \cap \cdots \cap \mathsf{G}_{i_n}) = \prod_{j=1}^{n} \mathbb{P}(\mathsf{G}_{i_j})$$

whenever $n \in \mathbb{N}$, and $i_1 < \cdots < i_n$ are *distinct* indices, and $\mathsf{G}_{i_j} \in \mathcal{G}_{i_j}$ for each $j = 1, \dots, n$.

Activity 13.22 (Independence: Countable collections). Develop a definition of what it means for a countable family of events to be independent. Develop a definition of what it means for a countable family of random variables to be independent. ▪

Warning 13.23 (Pairwise independence). Consider a family $(X_i : i \in \mathsf{I})$ of real random variables. Suppose that we know that each pair (X_i, X_j) is independent. This property is called *pairwise independence*. Unfortunately, pairwise independence does not even imply that a triple (X_i, X_j, X_ℓ) of distinct random variables is independent!

Similarly, for $k \geq 3$, the assumption that each sub-family of k random variables is independent does not imply that any set of $(k+1)$ random variables is independent. In particular, if you want to know that the entire collection of random variables is independent, you must enforce this property explicitly.

Although this may seem like a technicality, it is a common source of errors. Moreover, the concept of k-wise independence plays an important role in the theory of algorithms. For example, see Application 13.36. ▪

13.3.7 Why?

This abstract perspective would be sterile if it only allowed us to talk about things that we already understand, such as independent events and independent random variables. Even without further applications, it is useful for us to start thinking about σ-algebras as carrying information about the state of the world. From this point of view, independent σ-algebras carry independent information.

In fact, there are relatively simple things that are hard to describe accurately without this machinery. For example, consider three real random variables X, Y, Z. We can easily define what it means for the pair (X, Y) to be independent from the random variable Z. Indeed, we just require that $\sigma(X, Y)$ is independent from $\sigma(Z)$.

In fact, there are even settings where independence of σ-algebras models something that we cannot easily describe using only events or random variables. For an example, see Section E.2, on the Kolmogorov 0–1 law.

13.4 Kolmogorov's extension theorem

In Section 13.2.2, we saw that it is possible to build a probability space that supports two independent random variables with specified marginal laws. We would like to perform the same feat with a countable list of marginal laws to construct an independent sequence of random variables. The next result asserts that we can always achieve this goal. It provides a technical foundation for the theory of discrete-time stochastic processes.

Theorem 13.24 (Kolmogorov extension). Let $(\mu_1, \mu_2, \mu_3, \dots)$ be a sequence of probability measures defined on the Borel sets of the real line. There exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ on which we can define an *independent* sequence $(X_1, X_2, X_3, \dots)$ of real random variables where the law of X_i is μ_i for each index $i \in \mathbb{N}$. That is,

$$\mathbb{P}\left\{(X_{i_1}, \dots, X_{i_n}) \in \mathsf{B}_{i_1} \times \cdots \times \mathsf{B}_{i_n}\right\} = \prod\nolimits_{j=1}^{n} \mathbb{P}\left\{X_{i_j} \in \mathsf{B}_{i_j}\right\} = \prod\nolimits_{j=1}^{n} \mu_{i_j}(\mathsf{B}_{i_j})$$

for all $n \in \mathbb{N}$, and distinct indices $i_1 < \cdots < i_n \in \mathbb{N}$, and Borel sets $\mathsf{B}_{i_j} \in \mathcal{B}(\mathbb{R})$ for $j = 1, \dots, n$.

See Appendix E for the proof of the Kolmogorov extension theorem.

The key point here is that Theorem 13.24 furnishes a probability space containing a (countable) sequence of independent random variables. The precise construction is not important—just the fact that the probability space exists. The probability measure $\mathbb{P}$ packs up all the information about the joint distribution of the random variables, so we can use $\mathbb{P}$ to compute the probability of any event in $\mathcal{F}$, which happens to be the product σ-algebra $\mathcal{B}(\mathbb{R})^{\mathbb{N}}$. At the same time, when we study an independent sequence, we will typically focus on the individual random variables X_i and their laws μ_i, rather than the underlying probability space.

Example 13.25 (A sequence of coin flips). Suppose that we want to exhibit a model for a countable sequence of independent fair coin flips. In this case, we consider laws $\mu_i \sim \textsc{bernoulli}(1/2)$ for each $i \in \mathbb{N}$. Kolmogorov's extension theorem yields a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that supports an independent family $(X_i : i \in \mathbb{N})$ of real random variables, where each $X_i \sim \textsc{bernoulli}(1/2)$. ∎

Example 13.26 (A sequence of normal variables). Another important example involves the sequence $\mu_i \sim \textsc{normal}(0, 1)$ for each $i \in \mathbb{N}$. In this case, Kolmogorov's extension theorem yields a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that supports an independent family $(Z_i : i \in \mathbb{N})$ of real random variables where $Z_i \sim \textsc{normal}(0, 1)$ for each $i \in \mathbb{N}$. ∎

Warning 13.27 (*Countable product of Borel σ-algebra). Unlike the case of a finite product, $\mathcal{B}(\mathbb{R}^{\mathbb{N}}) \neq \mathcal{B}(\mathbb{R})^{\mathbb{N}}$. The reason is a mismatch between the definitions of the product topology and the product σ-algebra. Indeed, the product σ-algebra $\mathcal{B}(\mathbb{R})^{\mathbb{N}}$ contains all countable intersections of measurable cylinders. But the product topology on $\mathbb{R}^{\mathbb{N}}$ only contains finite intersections of open cylinders. The Borel σ-algebra $\mathcal{B}(\mathbb{R}^{\mathbb{N}})$ is generated by the product topology, and the product topology has only a limited stock of open sets. ∎

Problems

Exercise 13.28 (Mixture). Let X and Y be real random variables on a probability space. For $\alpha \in [0, 1]$, let I be a $\textsc{bernoulli}(\alpha)$ random variable that is independent from X

and Y. Define the random variable

$$Z = \begin{cases} X, & \text{if } I = 1; \\ Y, & \text{if } I = 0. \end{cases}$$

Express the law of Z in terms of the laws of X and Y.

Problem 13.29 (Poisson coin flips). Suppose that we flip a fair coin N times, where $N \sim \textsc{poisson}(\beta)$. Let X denote the number of heads that turn up, and let Y denote the number of tails that turn up. Show that the pair (X, Y) is independent. Bizarrely, if we are given the number N of flips, then the pair (X, Y) is no longer independent.

Exercise 13.30 (Minimum and maximum). Consider an i.i.d. family $(X_1, \dots, X_n)$ of random variables where $X_i \sim X$. In this exercise, we explore the distribution of the maximum and minimum.

1. Let $Y = \max_i X_i$. Show that the distribution function F_Y of the maximum satisfies
$$F_Y(a) = F_X(a)^n \quad \text{for } a \in \mathbb{R}.$$
2. Let $Z = \min_i X_i$. Show that the distribution function F_Z of the minimum satisfies
$$F_Z(a) = 1 - (1 - F_X(a))^n \quad \text{for } a \in \mathbb{R}.$$
3. Suppose that $X \sim \textsc{exponential}(\beta)$, an exponential random variable with rate $\beta > 0$. Determine the distribution of the maximum and the minimum of n i.i.d. copies of X.

Exercise 13.31 (Variance representations). There are several ways to write the variance of a random variable $X \in \mathsf{L}_2$ by introducing an *independent copy* Y of the random variable X. Verify that

That is, Y has the same law as X, and the pair (X, Y) is independent.

$$\operatorname{Var}[X] = \tfrac{1}{2}\,\mathbb{E}[(X - Y)^2] = \mathbb{E}[(X - Y)_+^2] = \mathbb{E}[(X - Y)_-^2].$$

Problem 13.32 (*Mutual information). Let X and Y be discrete random variables taking values in $\mathbb{N}$. Write f_X and f_Y for the marginal probability mass functions and f_{XY} for the joint mass function. Define the *mutual information*

$$I(X; Y) := \mathbb{E} \log \left(\frac{f_{XY}(X, Y)}{f_X(X) \cdot f_Y(Y)} \right).$$

We can interpret the mutual information as the amount of randomness in Y that is explained by X, or vice versa.

1. Observe that $I(X, Y) = 0$ when the pair (X, Y) is independent.
2. Show that
$$a(\log a - \log h) \geq a - h \quad \text{for all } a, h > 0.$$
Under what condition does equality hold? **Hint:** Apply the subgradient inequality (Proposition 9.19) to the negative logarithm.
3. Argue that $I(X, Y) \geq 0$ with equality only if (X, Y) is independent.

Exercise 13.33 (Chebyshev correlation inequalities). Let X be a real random variable with law μ_X. Consider μ_X-integrable functions $f, g : \mathbb{R} \to \mathbb{R}$. In this problem, we will show that it is possible to bound expectations $\mathbb{E}[f(X)g(X)]$ when the functions f, g are monotone. These inequalities are often used in statistical physics applications. The results appear here because the proof relies on independence.

These results are often called Chebyshev's "other" inequalities.

1. First, assume that f and g both are increasing. Let Y be an independent copy of X, and verify that
$$\mathbb{E}\left[(f(X) - f(Y)) \cdot (g(X) - g(Y))\right] \geq 0.$$
2. Deduce the Chebyshev correlation inequality:
$$\mathbb{E}[f(X)g(X)] \geq \mathbb{E}[f(X)] \cdot \mathbb{E}[g(X)].$$
3. Now, assume that f is increasing and g is decreasing. Establish a complementary rearrangement inequality:
$$\mathbb{E}[f(X)g(X)] \leq \mathbb{E}[f(X)] \cdot \mathbb{E}[g(X)].$$
4. As an example, confirm that
$$\mathbb{E}[Xe^{-\theta X}] \leq \mathbb{E}[X] \cdot \mathbb{E}[e^{-\theta X}] \quad \text{for } \theta > 0.$$
This inequality arises when bounding the derivative of a moment generating function.

Problem 13.34 (Generalized Minkowski). Consider independent real random variables X and Y with laws μ_X and μ_Y. For a measurable function $f : \mathbb{R}^2 \to \mathbb{R}$ and a power $p \geq 1$, the generalized Minkowski inequality states that
$$\left[\int_{\mathbb{R}} \left| \int_{\mathbb{R}} f(x,y)\, \mu_Y(\mathrm{d}y) \right|^p \mu_X(\mathrm{d}x)\right]^{1/p} \leq \int_{\mathbb{R}} \left(\int_{\mathbb{R}} |f(x,y)|\, \mu_X(\mathrm{d}x) \right)^{1/p} \mu_Y(\mathrm{d}y).$$
This relation holds whenever the integrals are defined.

1. If we define the partial expectation $\mathbb{E}_X$ with respect to X and the partial expectation $\mathbb{E}_Y$ with respect to Y, show that we can rewrite the inequality in the compact form
$$\left[\mathbb{E}_X |\mathbb{E}_Y f(X,Y)|^p\right]^{1/p} \leq \mathbb{E}_Y \left[\left(\mathbb{E}_X |f(X,Y)|^p\right)^{1/p}\right].$$
2. Show how to derive Minkowski's inequality (Theorem 11.9) as a consequence of this relation. **Hint:** Let Y be a Bernoulli random variable.
3. Establish the generalized Minkowski inequality. **Hint:** Introduce an independent copy Y' of Y, and note that
$$|\mathbb{E}_Y f(X,Y)|^p = |\mathbb{E}_Y f(X,Y)|^{p-1} \cdot |\mathbb{E}_{Y'} f(X,Y')|.$$
Proceed in the same manner as Riesz's proof of the simpler Minkowski inequality.

Problem 13.35 (*Pairwise independence: Subset sums). In this problem, we show that it is possible to construct $2^m - 1$ random variables that are pairwise independent, given a family of m fully independent random variables.

Consider an i.i.d. family $(X_1, \dots, X_n)$ of $\textsc{bernoulli}(1/2)$ random variables. For each nonempty subset $\mathsf{S} \subseteq \{1, 2, 3, \dots, n\}$, define
$$Y_{\mathsf{S}} := \left(\sum\nolimits_{i \in \mathsf{S}} X_i\right) \bmod 2.$$
It is clear that each random variable Y_{S} is Bernoulli because it takes values in $\{0, 1\}$.

1. For each set S, prove that $Y_{\mathsf{S}} \sim \textsc{bernoulli}(1/2)$.
2. For two distinct sets S, T, prove that the pair $(Y_{\mathsf{S}}, Y_{\mathsf{T}})$ is independent.
3. Find three random variables $(Y_{\mathsf{R}}, Y_{\mathsf{S}}, Y_{\mathsf{T}})$ that are not independent. **Hint:** Consider the case $\mathsf{R} \dot\cup \mathsf{S} = \mathsf{T}$.

Applications

Application 13.36 (*Hashing and pairwise independence). Hashing is a method for taking a long string and mapping it to a short "key" that can be used as a summary of the string. Hashing maps are often chosen randomly with the intention that two input strings are very unlikely to map to the same key value. At the same time, it is important to construct and apply the hashing map quickly. These desiderata lead us to consider random variables that have limited randomness and limited independence.

In this problem, we use a small dose of number theory to construct pairwise independent random variables and an associated hashing scheme. Fix a (large) *prime* number p. Let $X_1, X_2 \sim \text{UNIFORM}\{0, 1, 2, 3, \ldots, p-1\}$ be independent. Construct the random variables

$$Y_j = (X_1 + jX_2) \bmod p \quad \text{for } j = 0, 1, 2, \ldots, p-1.$$

It takes approximately $2 \log_2 p$ random bits to form X_1 and X_2, so we obtain p random variables Y_j from a modest number of coin flips. As we will see, these random variables can be used to construct a hashing scheme.

1. Show that $Y_j \sim \text{UNIFORM}\{0, 1, 2, \ldots, p-1\}$.
2. For each pair (i, j) of distinct indices, show that the pair (Y_i, Y_j) is independent.
3. Show that there is a triple (Y_i, Y_j, Y_k) that is not independent.

What does this have to do with hashing? We would like to summarize each potential input value in $\{0, 1, 2, \ldots, p-1\}$ by a short key. To that end, fix a natural number $n \ll p$. Let H be a (small) family of functions with domain $\{0, 1, 2, \ldots, p-1\}$ and codomain $\{0, 1, 2, \ldots, n-1\}$. The functions in H are called *hash functions*.

Draw a hash function h uniformly at random from the family H. For this choice of h, we say that a pair (i, j) of inputs *collides* when $h(i) = h(j)$. A good family of hash functions will limit the probability of a collision.

We say that the family H of hash functions is *2-universal* when each pair (i, j) of values has a small collision probability:

$$\mathbb{P}\{h(i) = h(j)\} \leq \frac{1}{n} \quad \text{for all distinct } i, j \in \{0, 1, 2, \ldots, p-1\}.$$

The idea is that the hash is very likely to map two distinct inputs i, j to two distinct keys $h(i), h(j)$. Therefore, we can use the key as a summary of the input.

4. Consider a 2-universal family H. Let $\mathsf{S} \subseteq \{0, 1, 2, \ldots, p-1\}$ be a fixed, but unknown, set of inputs. Choose an input $i \in \mathsf{S}$, and bound the expected number of $j \in \mathsf{S}$ that collide with i if we pick $h \in \mathsf{H}$ uniformly at random. How big should the set $\{0, 1, 2, \ldots, n-1\}$ of keys be in comparison with $\#\mathsf{S}$?

Here is a simple construction of a 2-universal hash function, based on pairwise independence. We construct a family H of hash functions that is indexed by values $a, b \in \{0, 1, 2, \ldots, p-1\}$. The associated function is defined as

$$h_{ab}(j) := [(a + jb) \bmod p] \bmod n.$$

We draw a hash h randomly by picking a, b independently and uniformly at random, with the constraint that $b \neq 0$.

5. Show that H is 2-universal.

Application 13.37 (*Second-moment method: Graph thresholds). As another application of the second-moment method, we will show that a certain class of random graphs is either very likely or very unlikely to contain a clique containing four vertices, depending on the exact parameter choices.

An *Erdős–Rényi graph* $\mathsf{G}(n, p)$ is an undirected, combinatorial graph drawn at random from the following distribution. The graph has n vertices, and each edge appears *independently* with probability p. A *clique* is a set C of vertices that is completely connected; that is, the graph contains the edge connecting each pair of vertices in the clique C.

We may ask about whether it is probable that an Erdős–Renyí graph $\mathsf{G}(n, p)$ contains *a clique on four vertices*. The answer depends on how the edge probability p scales with the number n of vertices. Problems like this arise when studying the connectivity properties of networks. For example, is it likely that there are four individuals in a social network where each pair is friends with each other?

1. Let S be a fixed set of four vertices in $\mathsf{G}(n, p)$. Show that the probability that S is a clique is exactly p^6.
2. Let X_n denote the number of cliques on four vertices in a random instance of $\mathsf{G}(n, p)$. Deduce that
$$\mathbb{E}[X_n] = \binom{n}{4} \cdot p^6 \approx \frac{n^4 p^6}{24}.$$
In particular, if $p \ll n^{2/3}$, then it is likely that the graph contains no clique on four vertices.
3. **(*)** Consider the asymptotic setting where $p = p(n)$. Assume that $p/n^{2/3} \to 0$. Argue that $\mathbb{P}\{X_n = 0\} \to 1$ as $n \to \infty$. What does this statement mean?
4. Suppose that Z is a positive, real random variable. If $\operatorname{Var}[Z] \ll (\mathbb{E} Z^2)$, show that $\mathbb{P}\{Z > 0\} \gg 0$.
5. **(**)** Assume that $p \gg n^{2/3}$. Show that $\operatorname{Var}[X_n] \ll (\mathbb{E} X_n)^2$. **Hint:** Subsets sharing two or more vertices are not independent, so you have to make conditional variance computations. Unfortunately, this is not trivial.
6. If $p \gg n^{2/3}$, deduce that $\mathsf{G}(n, p)$ is very likely to contain a clique on four vertices.
7. **(*)** In the asymptotic setting where $p/n^{2/3} \to \infty$, show that $\mathbb{P}\{X_n \geq 1\} \to 1$ as $n \to \infty$. What does this statement mean? In fact, $X_n \approx \mathbb{E} X_n$ for large n.

The notation $\ll$ means "much less than", but it does have a formal definition.

Notes

The overarching discussion of independence is adapted from Williams [Wil91]. See Motwani & Raghavan [MR95] for some discussion of k-wise independence and its applications. For more information about correlation inequalities, see [AS16]. Some of the problems are drawn from Grimmett & Stirzaker [GS01]. The example of the second-moment method is from Alon & Spencer [AS16].

Lecture bibliography

[AS16] N. Alon and J. H. Spencer. *The probabilistic method*. Fourth. John Wiley & Sons, 2016.

[GS01] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. 3rd ed. Oxford University Press, 2001.

[MR95] R. Motwani and P. Raghavan. *Randomized algorithms*. Cambridge University Press, Cambridge, 1995. DOI: 10.1017/CBO9780511814075.

[Wil91] D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

III. independent sums

14. Independent Sums

Agenda:
1. Stochastic processes
2. Independent sums
3. Model applications
4. Empirical behavior
5. Weak law of large numbers

> "Les dieux avaient condamné Sisyphe à rouler sans cesse un rocher jusqu'au sommet d'une montagne d'où la pierre retombait par son propre poids. Ils avaient pensé avec quelque raison qu'il n'est pas de punition plus terrible que le travail inutile et sans espoir...
>
> "La lutte elle-même vers les sommets suffit à remplir un couer d'homme. Il faut imaginer Sisyphe heureux."
>
> "The gods had condemned Sisyphus to ceaselessly roll a boulder up to the summit of a mountain, at which point the stone would fall back down because of its own weight. They had thought, with some reason, that there is no more terrible punishment than useless and hopeless labor...
>
> "The struggle itself toward the heights suffices to fill a man's heart. One must imagine Sisyphus happy."
>
> —Albert Camus, *Le Mythe de Sisyphe*, 1942

So far, we have been focusing on measure theory and probability foundations. In this lecture, we begin our study of stochastic processes, which are collections of (dependent) random variables. Stochastic processes are used to model probabilistic phenomena in computational mathematics, engineering, statistics, and other disciplines. Therefore, insights about the behavior of stochastic processes have a wide range of implications.

This course covers three types of sequential stochastic processes. In this lecture, we will start to investigate the behavior of the partial sums of a sequence of independent random variables, also known as an independent sum. Among other things, independent sums can be used to model random walks, statistical experiments, and Monte Carlo integration. Later, we will consider a more sophisticated type of stochastic process, called a martingale, that models the payoff in a repeated sequence of fair games, where the strategy can evolve depending on historical outcomes. Finally, we will turn to Markov chains, random sequences where the distribution of the next value depends only on the current value.

14.1 Stochastic processes

First, we introduce the concept of a general stochastic process.

The Greek word *stokhazomai* means "to aim at a target."

Definition 14.1 (Stochastic process). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. A *stochastic process* is a family $(X_t : t \in \mathsf{T})$ of real random variables defined on the probability space. The set T is called the *index set*. Stochastic processes are often called *random processes*.

Let us emphasize that the random variables that compose a stochastic process are typically not independent from each other. It is hard to say much about a stochastic process without adding some kind of additional structure. To that end, we will impose assumptions about how the random variables interact with each other. These assumptions allow us to model particular phenomena, and they allow us to develop a richer understanding of these examples.

Historically, the index set was denoted by the letter T because it usually models time. In particular, we may consider discrete-time and continuous-time models, but there are other possibilities.

- **Discrete-time process:** The index set T equals $\mathbb{N}$ or $\mathbb{Z}_+$ or $\mathbb{Z}$.
- **Continuous-time process:** The index set T equals $\mathbb{R}_+$ or $\mathbb{R}$.
- **Spatial process:** The index set $\mathsf{T} \subseteq \mathbb{R}^n$. These examples are often called *random fields*.

In this course, we will focus on discrete-time stochastic processes, which we will usually denote by $(X_n : n \in \mathbb{N})$ or something similar. The letter n reminds us that the index is an integer. Continuous-time processes require much more technical machinery, and it is best to acquire a firm understanding of the discrete-time setting before turning to this subject (e.g., in ACM 118). Stochastic processes with a spatial index, such as Gaussian processes, are also very important in applications (ACM 118 or ACM 217).

14.2 Independent sums

We begin our study of discrete-time stochastic processes with a particularly simple class of models, based on independent random variables.

14.2.1 The probability space

First, let us recall Kolmogorov's extension theorem (Theorem 13.24). Consider a sequence $(\mu_i : i \in \mathbb{N})$ of probability distributions on the Borel sets in $\mathbb{R}$. Then there exists a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ that supports an *independent* sequence $(Y_i : i \in \mathbb{N})$ of real random variables, where the marginal law of Y_i is μ_i for each index $i \in \mathbb{N}$.

This fact serves as the technical foundation for the theory of discrete-time stochastic processes, because it ensures that we can construct an independent sequence of random variables with arbitrary distributions. We usually do not lavish much attention on the underlying probability space, preferring to operate with the random variables themselves.

14.2.2 Some random processes

Given an independent sequence of real random variables, we can construct a new random sequence from the partial sums.

Definition 14.2 (Independent sum). Consider an independent sequence $(Y_i : i \in \mathbb{N})$ of real random variables. Define the random variables

$$X_0 := 0 \quad \text{and} \quad X_n := \sum\nolimits_{i=1}^{n} Y_i \quad \text{for } n \in \mathbb{N}.$$

That is, $(X_n : n \in \mathbb{Z}_+)$ is the sequence of partial sums of the sequence $(Y_i : i \in \mathbb{N})$. The family $(X_n : n \in \mathbb{Z}_+)$ is a discrete-time stochastic process, called an *independent sum process* or a *partial sum process*.

The random variables in a partial sum process are not independent from each other. Indeed, X_n and X_k both involve the summands Y_i for $i \le n \wedge k$. Nevertheless, the dependency among the X_n is both simple and manageable.

It is also productive to consider some related random processes, obtained by rescaling the partial sum process. In particular, we define the running average.

Definition 14.3 (Running average). Consider an independent sequence $(Y_i : i \in \mathbb{N})$ of real random variables. Define the random variables

$$\bar{X}_n := \frac{1}{n} \sum_{i=1}^{n} Y_i \quad \text{for } n \in \mathbb{N}.$$

Then $(\bar{X}_n : n \in \mathbb{N})$ is a discrete-time stochastic process, called the *running average* of the process $(Y_i : i \in \mathbb{N})$.

14.3 Applications

Independent sums and running averages are simple models, but they have a wide range of applications. In this section, we outline some of the main examples.

14.3.1 Random walks

Consider a particle that begins its life at the origin of the real line. Every second, the particle jumps to a new location by adding a random increment to its current location. In the most basic setting, each increment is independent from all previous increments and from the current location of the particle. This stochastic process is called a *random walk* on the real line.

Formally, we consider an independent sequence $(Y_i : i \in \mathbb{N})$ of random increments. Then the partial sum process $X_n = \sum_{i=1}^{n} Y_i$ describes the position of the particle after n steps.

In the simplest case, the increments are chosen to be independent and identically distributed (i.i.d.) random variables. It is common to express this condition by writing $Y_i \sim Y$ i.i.d., where Y is some fixed random variable. We also say that the Y_i are i.i.d. copies of Y.

The symbol $\sim$ means "has the distribution."

In particular, we may consider the distribution $Y \sim \text{UNIFORM}\{\pm 1\}$ for the increments. Then the particle travels on the integers, at each time moving randomly left or right by one position. This process is called the *simple random walk* on the integers $\mathbb{Z}$. At time n, the random location X_n of the particle has the marginal distribution $X_n \sim 2\,\text{BINOMIAL}(n, 1/2) - n$.

14.3.2 Renewals

Consider (real-world) events that happen periodically after some random interval of time. For example, a decaying radioactive mass might omit a photon. Or a shopper might arrive at the cash register of a package store to purchase a lottery ticket. These are examples of *renewal processes*.

We can model the arrivals by an independent sequence $(Y_i : i \in \mathbb{N})$. The random variable Y_i describes the time between events i and $i - 1$. The partial sum process $X_n = \sum_{i=1}^{n} Y_i$ is the total amount of time that elapses before n events occur.

The most common model for renewals takes the interarrival times Y_i to be i.i.d. copies of an exponential random variable $Y \sim \text{EXPONENTIAL}(\lambda)$ with rate $\lambda \in \mathbb{R}_+$. In this case, the partial sum process X_n has the marginal distribution $X_n \sim \text{GAMMA}(n, \lambda)$ with shape parameter $n \in \mathbb{N}$ and rate λ.

14.3.3 Independent experiments

Consider a sequence of independent experiments that may succeed or fail. For example, your instructor may shoot a basketball repeatedly, declaring success each time he achieves nothing but net. Another example: every night, your instructor asks his four-year-old to put away her Duplos before bedtime, with occasional success. Another example: Every day, a man named Sisyphus attempts to push a large boulder to the top of a hill. In these cases (and many others), the assumption of independence is perhaps questionable.

We can model the success or failure of these trials using an independent sequence Y_i of Bernoulli random variables, where $Y_i = 1$ if the ith trial succeeds and $Y_i = 0$ if the ith trial fails. The partial sum process $X_n = \sum_{i=1}^{n} Y_i$ describes the total number of successes in the first n trials. Similarly, the running average process $\bar{X}_n = n^{-1} \sum_{i=1}^{n} Y_i$ describes the proportion of successes during the first n trials.

When we repeat the same experiment over and over again, then we can model the successes Y_i as i.i.d. copies of a Bernoulli random variable $Y \sim \textsc{bernoulli}(p)$. In this case, the partial sum process X_n follows the marginal distribution $X_n \sim \textsc{binomial}(n, p)$. The running average $\bar{X}_n$ has expectation p, and it serves as an empirical estimate for the success probability p.

14.3.4 Statistical estimation

Consider a statistical experiment, where we randomly select members from a population and enter them into a study. Perhaps, we administer each participant an experimental vaccine for COVID-19 and measure her antibody levels after 72 hours. Perhaps the tech overlord administering an influential website would like to assess how much time a typical visitor spends reading wackadoodle conspiracy theories in his newsfeed. Perhaps, we want to ascertain how long it takes a lab rat to navigate to the center of a dangerous labyrinth, overseen by David Bowie.

When the experiment has a real-valued outcome, we can model the response Y_i of each subject as a real-valued random variable. Since we have randomized the choice of participants in the study, the responses $(Y_i : i \in \mathbb{N})$ form an independent sequence. The running average $\bar{X}_n = n^{-1} \sum_{i=1}^{n}$ serves as an empirical estimate for the expected response. When participants are registered sequentially, the running average provides an evolving picture of our current estimate for the expected response.

In statistical estimation, it is also common to model the responses as i.i.d. copies of a fixed random variable Y. In this setting, however, we may not have strong prior knowledge about the properties of the random variable Y. Therefore, it is desirable to develop results about the running average that hold under weak assumptions.

14.3.5 Monte Carlo integration

Independent sums also arise in the design of computer algorithms. Here is a basic example that is the starting point for a very important class of techniques in computational mathematics and statistics.

Suppose that we wish to approximate the integral $\int_\Omega f \, \mathrm{d}\mu$, where μ is a probability measure on Ω and $f \in \mathsf{L}_1(\mu)$ is integrable. This problem is called *numerical quadrature*. When $\Omega = \mathbb{R}$ and μ is a standard distribution (e.g., Gaussian or Laplace), there are very effective quadrature methods based on deterministic rules. On the other hand, when Ω is a high-dimensional space or μ is a complicated distribution, it can be tricky to evaluate the integral using a fixed rule. Instead, we may turn to a probabilistic method called *Monte Carlo integration*.

Draw an independent sequence $(Z_i \in \Omega : i \in \mathbb{N})$ of random variables, each with the distribution μ. In practice, this step can be very challenging, but we shall assume that it has been accomplished.

For each index $i \in \mathbb{N}$, we compute $Y_i = f(Z_i)$. The sequence $(Y_i : i \in \mathbb{N})$ is independent, and the marginal distribution of Y_i is the push-forward of μ by the function f. (See Problem 5.44.) Therefore, the running average of the Y_i serves as an empirical approximation of the integral:

$$\bar{X}_n = \frac{1}{n} \sum\nolimits_{i=1}^{n} Y_i \approx \int_\Omega f \, \mathrm{d}\mu.$$

The question, of course, is how many samples n we need to approximate the integral to a certain tolerance and with a specified probability of error. The answer depends on both the probability measure μ and the function f.

14.4 Empirical behavior of independent sums

Now that we are persuaded of the potential utility of the partial sum model, we can start to ask how this stochastic process behaves. The independence of the underlying sequence leads to some simple observations. We will also exhibit some experiments that motivate the technical questions we will pursue.

14.4.1 Mean and variance

Consider a sequence $(Y_i : i \in \mathbb{N})$ consisting of independent copies of a real random variable Y. Form the partial sum process $X_n = \sum_{i=1}^n Y_i$ for $i \in \mathbb{Z}_+$. Using linearity, we quickly determine the expectation:

$$\mathbb{E}[X_n] = \sum\nolimits_{i=1}^{n} \mathbb{E}[Y_i] = n \cdot \mathbb{E}[Y] \quad \text{if } Y \geq 0 \text{ or } Y \in \mathsf{L}_1.$$

Since independent random variables are uncorrelated (Exercise 13.11), the Pythagorean relation (Proposition 12.16) implies that the variance of the independent sum is additive. We find that

$$\operatorname{Var}[X_n] = \sum\nolimits_{i=1}^{n} \operatorname{Var}[Y_i] = n \cdot \operatorname{Var}[Y] \quad \text{if } Y \in \mathsf{L}_2.$$

In other words, the average value of the nth partial sum X_n is just n times the average of the increment Y. The standard deviation of X_n is just $\sqrt{n}$ times the standard deviation $\operatorname{stdev}(Y)$ of the increment Y; this is the typical scale for fluctuations around the mean.

Recall that $\operatorname{stdev}(Y) := \sqrt{\operatorname{Var}[Y]}$.

We remark that neither of the calculations uses the full power of the independence assumption. To compute the mean, we do not need any assumptions beyond integrability. To compute the variance, we only require mutual uncorrelation, which is far weaker than independence. In later investigations, however, independence will play a stronger role.

14.4.2 Sample paths

Probabilists use sample paths to picture the evolution of a random process. Recall that a real random variable is a real-valued function on the sample space. Therefore, each sample point $\omega \in \Omega$ determines the trajectory of the partial sum process for all times: $(X_n(\omega) : n \in \mathbb{N})$. *A priori*, the sample point is random, and so the trajectory of the stochastic process is random. Once Tyche distinguishes a sample point ω_0, the entire history of the partial sum process is sealed.

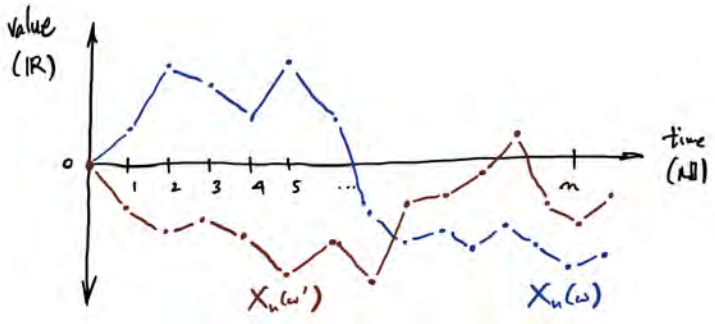


Figure 14.1 (Sample paths). The trajectory $(X_n(\omega) : n \in \mathbb{N})$ of a discrete-time random process is a function of the sample point $\omega \in \Omega$. The (random) trajectory is called a sample path. You can see that different sample points ω and ω' result in different trajectories.

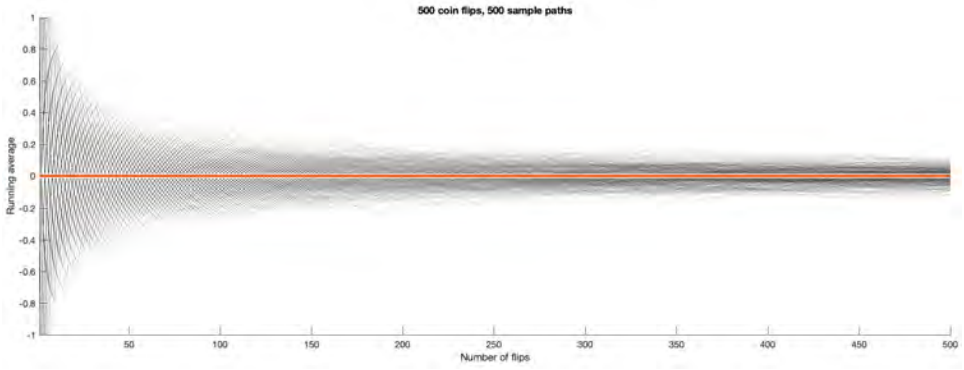


Figure 14.2 (Sample paths for the running average). This plot displays the first 500 time steps of 500 sample paths of the running average $\bar{X}_n$ of a series of coin flips, taking values ± 1. The orange line marks the expected value: $\mathbb{E}[\bar{X}_n] = 0$.

For discrete-time processes, it is convenient to illustrate the sample paths using a piecewise linear interpolant of the discrete values. Figure 14.1 contains an illustration.

14.4.3 Sample paths of the running average

We can perform computer experiments to get a picture of the random distribution of sample paths. This project requires a particular choice of distribution for the increment Y. Let us consider $Y \sim \text{UNIFORM}\{\pm 1\}$, which we can regard as a model for a single flip of a fair coin ($+1$ = heads and -1 = tails). More topically, it models a vote in an election where each candidate is equally favored ($+1$ = Gryffindor and -1 = Slytherin).

The running average $\bar{X}_n = n^{-1} \sum_{i=1}^{n} Y_i$ is the total of the first n outcomes, relative to the number n of trials. For coins, $\bar{X}_n$ is the difference between the proportion of heads and the proportion of tails in the first n coin flips. For votes, $\bar{X}_n$ is the difference between the proportion of votes in favor of Gryffindor and the proportion of votes in favor of Slytherin after n ballots have been cast.

Figure 14.2 displays the first 500 time steps of 500 random sample paths of the running average. The orange line marks the expected value: $\mathbb{E}[\bar{X}_n] = 0$ for all $n \in \mathbb{N}$.

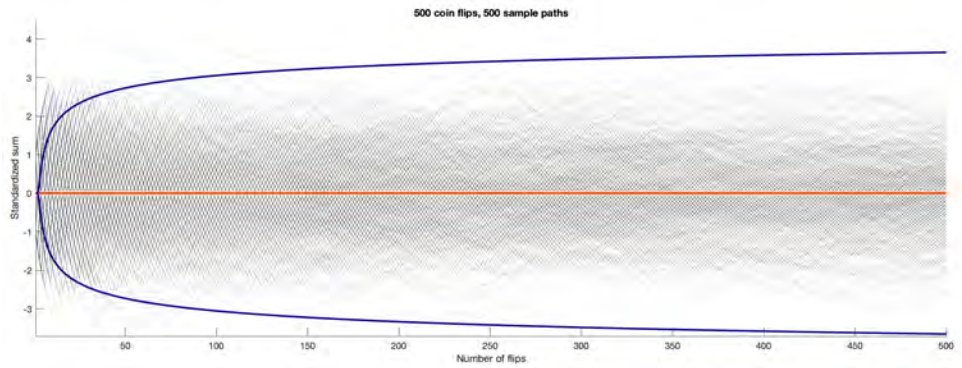


Figure 14.3 (Sample paths of the standardized sum). This plot displays the first 500 time steps of 500 sample paths of the standardized sum T_n of a series of coin flips, taking values ± 1. The orange line marks the expected value: $\mathbb{E}[T_n] = 0$. The blue envelope marks the long-term extreme values, as predicted by theory.

We can see that the collection of sample paths forms a funnel that tapers inward toward the expectation as the time horizon n increases. This plot raises some questions:

- **Limits:** As $n \to \infty$, it appears that most of the sample paths tend toward zero. Does the limit of the running average indeed approach the expectation? In what sense do we understand this limit?
- **Concentration:** At a given time n, how unlikely is it to see a value of $\bar{X}_n$ that differs substantially from the expectation? What is the probability that $\bar{X}_n$ takes a value in the typical range?

Our challenge is to produce good concentration bounds for the running average at a given time n and to prove limit theorems as $n \to \infty$.

14.4.4 Sample paths of the standardized sum

By choosing an alternative scaling of the partial sum process, we can explore how much the partial sums fluctuate around the mean value. We continue to work with the distribution $Y \sim \textsc{uniform}\{\pm 1\}$. Observe that

$$\operatorname{Var}[X_n] = \sum_{i=1}^{n} \operatorname{Var}[Y_i] = n.$$

In other words, the typical scale for fluctuations of the nth partial sum about the mean is $\sqrt{n}$. Therefore, we consider the *standardized sum*

$$T_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} Y_i \quad \text{for } n \in \mathbb{N}.$$

By construction, $\mathbb{E}[T_n] = 0$ and $\operatorname{Var}[T_n] = 1$. Thus, each of the standardized sums has the same scale, and they are comparable with each other.

Figure 14.3 displays the first 500 time steps of 500 random sample paths of the standardized sum. The orange line marks the expectation value: $\mathbb{E}[T_n] = 0$ for all $n \in \mathbb{N}$. We can see that the collection of sample paths quickly settle down to a consistent profile. At a given time n, the shade of gray reflects how many sample paths are passing through a given value. There are more near the expectation, and fewer farther away. This density appears to be stable as time evolves. This plot raises some questions:

- **Distribution:** As $n \to \infty$, does the density of sample paths settle down to a limiting distribution?
- **Envelope:** Sample paths occasionally stray away from the bulk, taking unusually large values. Can we quantify the extreme values that a sample path is likely to achieve?

To answer the first question, we need to decide what it means for a sequence of distributions to converge to another distribution. The second question has a beautiful answer (called the law of the iterated logarithm), but it requires a delicate analysis that we will not pursue in this class (but see Lecture 26).

14.5 Independent sums: Overview

In the last section, we have raised a number of questions about independent sums. In this part of the class, we will develop good (but not always optimal) answers to these questions. The investigation splits into two parts:

1. **Nonasymptotic results:** We will be interested in methods for describing the behavior of a particular sum X_n for a fixed value of n. These results are very useful in practice because they apply to explicit sums that we have in our hands.
2. **Asymptotic results:** We will also derive information about the long-time behavior of a rescaled sum, such as the running average $\bar{X}_n$ or the standardized sum T_n, as the time horizon n tends to infinity. These results are valuable because they are very clean, and they provide powerful heuristics for thinking about independent sums with many terms.

As it happens, there is a deep relationship between nonasymptotic results and asymptotic results. Indeed, our primary strategy for proving limit laws is to establish an appropriate result for a finite sum and take the limit. This idea is summarized in a quotation that is attributed to Kolmogorov:

> "Behind every limit theorem is an inequality." —A. N. Kolmogorov

Our results for independent sums share another common feature that is worth emphasis. Because of the independence assumption, we can use simple properties of the individual summands to derive strong conclusions about the behavior of the entire sum. This approach is exemplified in the computation of the mean and variance in Section 14.4.1, but it holds more widely. You can think about this idea as a kind of "local to global" principle. It highlights the power of the independence assumption.

Problems

Exercise 14.4 (Independent sums). Let X and Y be independent, real random variables. Suppose that X and Y have marginal laws μ_X and μ_Y. Define the sum $Z = X + Y$.

1. Show that the law of Z satisfies
$$\mu_Z(\mathsf{B}) = \int_{\mathbb{R}} \left[\int_{\mathbb{R}} \mathbb{1}_{\mathsf{B}}(x+y)\, \mu_Y(\mathrm{d}y) \right] \mu_X(\mathrm{d}x) \quad \text{for all } \mathsf{B} \in \mathscr{B}(\mathbb{R}).$$
2. Deduce that the distribution function of Z satisfies
$$F_Z(a) = \int_{\mathbb{R}} F_Y(a-x)\, \mu_X(\mathrm{d}x) = \int_{\mathbb{R}} F_X(a-y)\, \mu_Y(\mathrm{d}y) \quad \text{for } a \in \mathbb{R}.$$

3. Now, assume that Y is a continuous random variables with density f_Y. Show that Z is a continuous random variable with density

$$f_Z(a) = \int_{\mathbb{R}} f_Y(a - x)\, \mu_X(\mathrm{d}x) \quad \text{for } a \in \mathbb{R}.$$

   Hint: Start with (1) and invoke Fubini–Tonelli (Theorem 6.23). Although you might be tempted to differentiate the formula in (2), this approach requires a hard technical argument.
4. Specialize the last formula to the case where both X and Y are continuous.

Exercise 14.5 (Some stable random variables). There are some very special classes of random variables that are stable under addition. In this problem, you are invited to confirm stability in three cases. These results can be obtained more simply using the methods from Lecture 21.

1. Let X and Y be independent random variables with $\textsc{poisson}(1)$ distributions. Show that $Z = X + Y$ follows the $\textsc{poisson}(2)$ distribution.
2. Let X and Y be independent random variables with $\textsc{normal}(0, 1)$ distributions. Show that $Z = X + Y$ follows the $\textsc{normal}(0, 2)$ distribution.
3. **(*)** Let X and Y be independent random variables with $\textsc{cauchy}(0, 1)$ distributions. Show that $Z = X + Y$ follows the $\textsc{cauchy}(0, 2)$ distribution.

Notes

Independent sums are one of the primary objects of study in elementary probability and in probability theory. You will find some presentation of this material in almost any probability book that you open.

15. The Law of Large Numbers

"A-breaking rocks in the hot sun.
I fought the law, and the law won."

—*I Fought the Law* by Sonny Curtis
The Crickets (1960); Bobby Fuller Four (1966); The Clash (1977)

Agenda:
1. Chebyshev's weak law
2. Almost-sure convergence
3. Kolmogorov's strong law
4. Cantelli's strong law

In many contexts, we perform repeated trials of an experiment that produces a real-valued result. The goal of this process is to observe the outcomes of experiments so that we can make inferences about the probability distribution underlying the sequence of outcomes. In particular, we may want to estimate the expected value of the distribution by averaging the observed outcomes. The law of large numbers asserts that this approach is valid.

15.1 The law of large numbers

Suppose that Y is a real random variable. Let $(Y_i : i \in \mathbb{N})$ be *independent, identical copies* of Y. Our goal is to estimate the expectation $\mathbb{E}\, Y$ from the observed values $Y_1, Y_2, Y_3, \dots$.

A natural approach is to form the running average $\bar{X}_n := n^{-1} \sum_{i=1}^{n} Y_i$ of the first n observations. For a fixed number n of observations, statisticians often call $\bar{X}_n$ the *sample average*. Using linearity of expectation, it is easy to see that

$$\mathbb{E}\, \bar{X}_n = \mathbb{E}\, Y \quad \text{for each } n \in \mathbb{N}, \text{ provided that } Y \in \mathsf{L}_1.$$

Using additivity of variance for independent random variables, we find that

$$\operatorname{Var}[\bar{X}_n] = \frac{1}{n} \operatorname{Var}[Y] \quad \text{for each } n \in \mathbb{N}, \text{ provided that } Y \in \mathsf{L}_2.$$

In other words, the running average is an unbiased estimator for the expectation of Y, and the variance of the estimator declines as the number n of observations increases. These arguments suggest that the running average serves as an increasingly accurate estimate for $\mathbb{E}\, Y$ as we increase the number n of observations.

We would like to quantify this intuition about the long-run behavior of the running average with a result like "$\bar{X}_n \to \mathbb{E}\, Y$ as $n \to \infty$". A statement of this form is called a *law of large numbers* (LLN) for the running average.

There are many different kinds of LLNs for the running average. One major dichotomy reflects the type of convergence that we establish. *Weak LLNs* assert that the running average converges in probability to the expectation. *Strong LLNs* assert that the running average converges almost surely to the expectation. We will elaborate on this distinction below.

LLNs also differ in the precise assumptions that they place on the underlying distribution of the family $(Y_i : i \in \mathbb{N})$. For example, some LLNs concern the case where each of the summands follows the same distribution, while other results allow the summands to follow dissimilar distributions. LLNs also proceed from particular hypotheses about the integrability of the summands (e.g., L_1 or L_2), and the difficulty of the proof is usually inverse to the strength of the assumptions.

15.2 Chebyshev's weak law of large numbers

The weak law of large numbers (WLLN) is a limit theorem that describes the running average of a sequence of i.i.d. copies of a random variable Y. This result tells us that the running average $\bar{X}_n$ of the sequence provides a sequence of point estimates for the expectation $\mathbb{E}\, Y$.

15.2.1 Convergence in probability

To state the result, let us introduce a new mode of convergence for random variables.

Definition 15.1 (Convergence in probability). A sequence $(W_n : n \in \mathbb{N})$ of real random variables converges in probability to a real random variable W when

$$\lim_{n\to\infty} \mathbb{P}\{|W_n - W| > t\} = 0 \quad \text{for each } t > 0.$$

Roughly, convergence in probability means that is eventually unlikely for W_n to differ from W by more than any positive threshold $t > 0$.

> **Aside:** Convergence in probability is closely related to convergence in L_1. In general, these two modes of convergence are incomparable. Nevertheless, under an additional assumption that the random variables $(W_n : n \in \mathbb{N})$ form a "uniformly integrable" family, the two modes of convergence coincide. See Lecture 25.

15.2.2 Chebyshev's variance inequality

The key to proving our first LLN is a fundamental variance inequality. This result was originally framed by Bienaymé. Chebyshev apparently provided the first proof, and his name is now associated with the statement. Chebyshev's inequality controls the tail decay of a random variable in terms of its variance. This result is a powerful tool for studying independent sums because the variance of an independent sum is additive.

Proposition 15.2 (Chebyshev's variance inequality). Let $X \in \mathsf{L}_2$ be a real random variable. Then

$$\mathbb{P}\{|X - \mathbb{E}\, X| \geq t\} \leq \frac{\operatorname{Var}[X]}{t^2} \quad \text{for all } t > 0.$$

Proof. This is an instant consequence of Markov's inequality (Theorem 10.13). ∎

15.2.3 The weak law

The simplest version of the WLLN follows as an immediate consequence of Chebyshev's inequality.

Theorem 15.3 (Chebyshev's WLLN). Let $Y \in \mathsf{L}_2$ be a real random variable, and consider an i.i.d. sequence $(Y_i : i \in \mathbb{N})$ of copies of Y. The running averages $\bar{X}_n$ of the

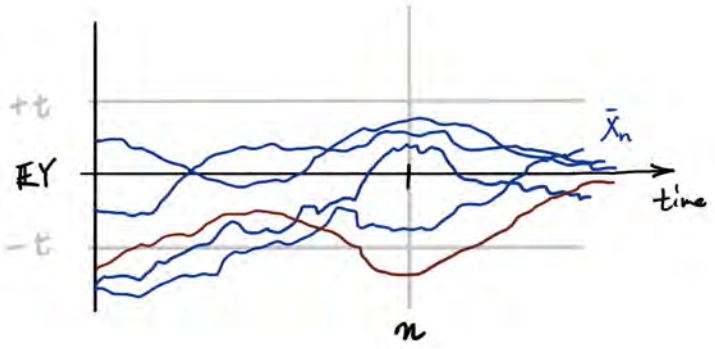


Figure 15.1 (Weak law of large numbers). The weak law of large numbers concerns the running average $\bar{X}_n$ of an i.i.d. sequence of copies of Y. It asserts that the sample paths of $\bar{X}_n$ converge in probability to the expectation $\mathbb{E}\,Y$. The blue sample paths lie within the band $\mathbb{E}[Y] \pm t$. The red sample path has escaped from the band.

sequence converge *in probability* to $\mathbb{E}\,Y$. Explicitly,

$$\lim_{n\to\infty} \mathbb{P}\left\{|\bar{X}_n - \mathbb{E}\,Y| > t\right\} = 0 \quad \text{for each } t > 0.$$

The argument depends heavily on the assumption that $Y \in \mathsf{L}_2$, so that we can compute variances.

Proof. We can apply Chebyshev's variance inequality to the running average $\bar{X}_n$ using the calculations at the start of Section 15.1. Indeed,

$$\mathbb{E}\,\bar{X}_n = \mathbb{E}\,Y \quad \text{and} \quad \operatorname{Var}[\bar{X}_n] = \frac{1}{n}\operatorname{Var}[Y].$$

Invoke Proposition 15.2 to reach the bound

$$\mathbb{P}\left\{|\bar{X}_n - \mathbb{E}\,Y| > t\right\} \le \frac{\operatorname{Var}[Y]}{nt^2} \quad \text{for all } t > 0. \tag{15.1}$$

Last, take the limit of (15.1) as $n \to \infty$. ∎

See Figure 15.1 for an illustration of what the WLLN means for the sample paths of the running average. In words, we fix a level $t > 0$ for the deviations of the running average away from the expectation $\mathbb{E}\,Y$. At a given time n, a large proportion (blue) of the sample paths are likely to fall within the band $\mathbb{E}\,Y \pm t$, but a small proportion (red) may escape. As the time horizon $n \to \infty$, the proportion of paths within the band at time n increases to 100%.

On the positive side, the proof of Theorem 15.3 is very easy. On the negative side, the assumption that $Y \in \mathsf{L}_2$ seems unnecessarily strict (maybe $Y \in \mathsf{L}_1$ is enough?). Furthermore, convergence in probability is not a very impressive type of convergence. In the next section, we will discuss an improvement.

Problem 15.4 (*WLLN: Integrable distribution). Prove that the conclusion of Theorem 15.3 holds when $Y \in \mathsf{L}_1$. **Hint:** For each level $t > 0$, you can approximate the random variable Y by a bounded random variable Y_B that satisfies $|Y_B| \le B$.

Problem 15.5 (WLLN: Non-identical distributions). Formulate and prove a WLLN for the running average that holds when the family $(Y_i : i \in \mathbb{N})$ is not necessarily identically distributed. What are natural assumptions on the variances?

Exercise 15.6 **(*WLLN: Pairwise independence).** Show that the WLLN holds under the weaker assumption that $(Y_i : i \in \mathbb{N})$ is pairwise independent. That is, the pair (Y_i, Y_j) is independent for all i, j. See Warning 13.23 for a brief discussion.

15.3 Kolmogorov's strong law of large numbers

In this section, we will discuss a better class of result, called the strong law of large numbers (SLLN).

15.3.1 Almost-sure convergence

Before we can present the strong law, it is productive to elaborate on the type of convergence involved.

Definition 15.7 **(Almost-sure event).** Let $\mathsf{A} \in \mathcal{F}$ be an event with $\mathbb{P}(\mathsf{A}) = 1$. This kind of event is called $\mathbb{P}$-*almost sure* or *almost sure* or just *a.s.*

The notion of an almost-sure event in probability theory is the companion to the notion of an almost-everywhere set in measure theory. The terminology changes to reflect the probabilistic setting.

Of course, the certain event Ω is an almost-sure event. In contrast, an almost-sure event does not necessarily occur. We usually do not specify the probability distribution $\mathbb{P}$ when discussing almost-sure events unless it is required for clarity. The concept leads to a mode of convergence.

Definition 15.8 **(Almost-sure convergence).** Consider a sequence $(W_n : n \in \mathbb{N})$ of real random variables defined on the same probability space $(\Omega, \mathcal{F}, \mathbb{P})$. We say that W_n *converges almost surely* to a random variable W when

$$\mathbb{P}\{\omega \in \Omega : W_n(\omega) \to W(\omega)\} = 1.$$

We often write $W_n \to W$ a.s. to denote this type of convergence.

In other words, for a randomly chosen sample point ω, there is a 100% chance that the associated sample path $W_n(\omega)$ converges to the limiting value $W(\omega)$. Equivalently, almost-sure convergence asserts that

$$\mathbb{P}\{\omega \in \Omega : W_n(\omega) \not\to W(\omega)\} = 0.$$

We can also express this relation as

$$\mathbb{P}\{\limsup_{n\to\infty} |W_n - W| > 0\} = 0.$$

The limit superior is defined as

$$\begin{aligned}\limsup_{n\to\infty} a_n &:= \lim_{n\to\infty} \sup_{j\geq n} a_j \\ &= \inf_{n\in\mathbb{N}} \sup_{j\geq n} a_j.\end{aligned}$$

The compact set-builder notation suppresses the role of the sample point in the last expression.

From this relation, we can start to see that almost-sure convergence is a stronger notion that convergence in probability. Recall that $W_n \to W$ in probability when

$$\sup_{t>0} \lim_{n\to\infty} \mathbb{P}\{|W_n - W| > t\} = 0.$$

The time variable n and the level t appear outside the probability, whereas they appear inside for almost-sure convergence.

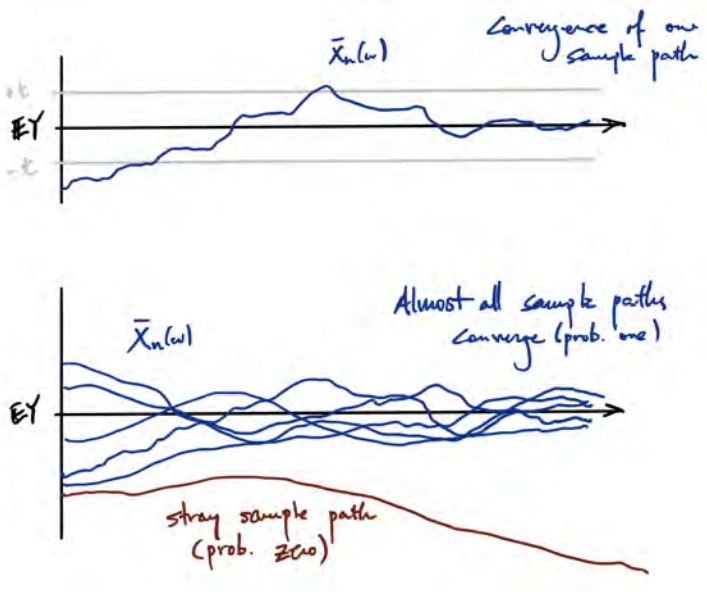


Figure 15.2 (Strong law of large numbers). The strong law of large numbers concerns the running average $\bar{X}_n$ of an i.i.d. sequence of copies of Y. It asserts that the running average converges almost surely to the expectation $\mathbb{E}\,Y$. There is a zero probability that a sample path fails to converge to the expectation.

Problem 15.9 (Convergence: Implications). Show that $W_n \to W$ pointwise implies that $W_n \to W$ almost surely. Show that $W_n \to W$ almost surely implies $W_n \to W$ in probability. By example, argue that neither of these implications can be reversed in general.

15.3.2 The SLLN

With this preparation, we can state an optimal version of the strong law of large numbers.

Theorem 15.10 (Kolmogorov's SLLN). Let $Y \in \mathsf{L}_1$ be a real random variable, and consider an *i.i.d. sequence* $(Y_i : i \in \mathbb{N})$ of copies of Y. The running averages $\bar{X}_n = n^{-1} \sum_{i=1}^{n} Y_i$ of the sequence converge *almost surely* to $\mathbb{E}\,Y$. Explicitly,

$$\mathbb{P}\left\{\bar{X}_n \to \mathbb{E}\,Y\right\} = 1.$$

Figure 15.2 illustrates what almost-sure convergence means in this context. In words, the probability of encountering a stray sample path that does not converge to the expectation is zero.

A direct proof of Theorem 15.10 is delicate, and we prefer to spend our effort on other things. You may find the argument in many textbooks on probability theory, such as [Shi96, Sec. IV.3]. For a sketch of the proof, see Problem 15.21. We will prove a variant (Theorem 15.11) of the SLLN later in this lecture.

Instead, let us discuss what Kolmogorov's SLLN means. First of all, the assumption that $Y \in \mathsf{L}_1$ is necessary to ensure that its expectation $\mathbb{E}\,Y$ is finite. Under this minimal assumption, the result states that the running averages converge almost surely to the expectation. For a randomly chosen sample point $\omega \in \Omega$, there is a 100% chance

that the sample path $\bar{X}_n(\omega) \to \mathbb{E}\, Y$. (In particular, the running averages converge in probability because of Exercise 15.9.)

15.3.3 Implications of the SLLN

The SLLN has important implications for statistics and for probability theory.

The sample average estimator

First, the SLLN gives an asymptotic justification for the sample average estimator for the population mean. Let Y be a real random variable that describes the distribution of some variable associated with a population. Suppose that we sample n individuals independently at random from the population (with replacement), and we record the values Y_i of their responses. By construction of the sample, the Y_i are i.i.d. copies of Y.

For example, Y might model the number of "friends" that a member of a social networking site has. We select n random members of the site, and we inquire about the number Y_i of "friends" that each of these individuals has.

We can estimate the population mean $\mathbb{E}\, Y$ using the sample mean $\bar{X}_n = n^{-1} \sum_{i=1}^{n} Y_i$. The SLLN tells us that, as we sample more and more individuals, we can have 100% confidence that $\bar{X}_n \to \mathbb{E}\, Y$. In other words, the SLLN gives a long-run guarantee that the sample average estimator tends to the population mean under the weakest possible assumption ($Y \in \mathsf{L}_1$).

> **Aside:** In some applications, we may encounter "heavy-tailed" random variables that do not have an expectation. The SLLN does not apply in these cases. Potential examples include things like the magnitude of earthquakes or the value of certain financial assets.

Frequentist interpretation of probabilities

Second, the SLLN justifies the frequentist interpretation of probabilities. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Let A be an event with probability $p = \mathbb{P}(\mathsf{A})$. How can we understand what this probability means? Here is one approach.

Imagine that we can perform an experiment and observe whether or not the event A occurs. Let $Y = \mathbb{1}_{\mathsf{A}}$ be the indicator random variable that the event occurs. Of course, $\mathbb{E}\, Y = \mathbb{P}(\mathsf{A}) = p$.

Now, suppose that we can perform repeated independent trials of this experiment and observe whether A occurs in each trial. If so, we obtain a sequence $(Y_i : i \in \mathbb{N})$ of i.i.d. copies of the indicator $Y = \mathbb{1}_{\mathsf{A}}$. The running average $\bar{X}_n$ gives the proportion of the first n trials in which the event A occurs. The SLLN states that, with probability one, $\bar{X}_n \to \mathbb{E}\, Y = p$. In other words, we can think about the probability p of the event A as the long-run proportion of times that the event occurs.

> **Aside:** Some philosophers question whether it even makes sense to talk about repeated experiments. For example, the 2020 US presidential election only happened one time, in all its awful glory. Does it make sense to talk about the "probability" of an outcome of the election in the frequentist sense? Bayesians would argue that, instead, probabilities reflect our prior assumptions, updated based on available evidence. There is also a school of thought that probabilities are "degrees of belief." Other authors regard probabilities as a reflection of one's willingness to wager on the outcome. These debates are extra-mathematical.

15.4 Cantelli's SLLN

We will prove another strong law of large numbers, due to Cantelli.

Theorem 15.11 (Cantelli's SLLN). Let $Y \in \mathsf{L}_4$, and consider an i.i.d. sequence $(Y_i : i \in \mathbb{N})$ of copies of Y. Then the running average $\bar{X}_n \to \mathbb{E}\, Y$ *almost surely*.

In contrast with Kolmogorov's SLLN (Theorem 15.10), the hypotheses of Cantelli's SLLN are more generous. The assumption that $Y \in \mathsf{L}_4$ allows us to give a short, transparent proof. The approach extends the argument behind Chebyshev's WLLN (Theorem 15.3) by using some stronger tail bounds.

Problem 15.12 (Cantelli: Non-identical distributions). Formulate and prove a version of Cantelli's strong law when $(Y_i : i \in \mathbb{N})$ are independent but may not have the same distribution.

Exercise 15.13 (*Cantelli: Four-wise independence). Show that Cantelli's strong law holds when $(Y_i : i \in \mathbb{N})$ under the weaker assumption of four-wise independence. That is, (Y_i, Y_j, Y_k, Y_ℓ) is independent for all indices i, j, k, ℓ. See Warning 13.23 for some discussion.

15.4.1 The lemmata of Borel & Cantelli

An important ingredient in the proof of Cantelli's strong law is a classic result, called the first Borel–Cantelli lemma.

Proposition 15.14 (Borel–Cantelli I). Let $(\mathsf{A}_n : n \in \mathbb{N})$ be a sequence of events. Then

$$\sum\nolimits_{n=1}^{\infty} \mathbb{P}(\mathsf{A}_n) < +\infty \quad \text{implies} \quad \mathbb{P}(\limsup\nolimits_{n\to\infty} \mathsf{A}_n) = 0.$$

Recall that $\limsup_{n\to\infty} \mathsf{A}_n := \bigcap_{n=1}^{\infty} \bigcup_{i\geq n} \mathsf{A}_i$.

The limit superior, $\limsup_{n\to\infty} \mathsf{A}_n$, is the event that an infinite number of the events A_n occur. The Borel–Cantelli lemma concerns the case where the total probability $\sum_n \mathbb{P}(\mathsf{A}_n)$ is finite. In this situation, with probability one, only a finite number of the events A_i occur.

Proof. This result follows when we apply Problem 5.41 to the indicator random variables of the events. We can also give a direct proof.

Fix an index $n \in \mathbb{N}$. By definition of the limit superior and the union bound,

$$\mathbb{P}\big(\limsup\nolimits_{n\to\infty} \mathsf{A}_n\big) \leq \mathbb{P}\big(\bigcup\nolimits_{i\geq n} \mathsf{A}_i\big) \leq \sum\nolimits_{i\geq n} \mathbb{P}(\mathsf{A}_i).$$

As $n \to \infty$, the latter sum converges to zero because the entire sequence of probabilities is summable. ∎

Problem 15.15 (*Borel–Cantelli II). There is a partial converse of Proposition 15.14 under an additional independence assumption.

1. For numbers $0 \leq p_n < 1$, demonstrate that

$$\prod\nolimits_{n=1}^{\infty} (1 - p_n) = 0 \quad \text{if and only if} \quad \sum\nolimits_{n=1}^{\infty} p_n = +\infty.$$

2. Assume that the family $(\mathsf{A}_n : n \in \mathbb{N})$ of events is *independent*. Prove the second Borel–Cantelli lemma:

$$\sum\nolimits_{n=1}^{\infty} \mathbb{P}(\mathsf{A}_n) = +\infty \quad \text{implies} \quad \mathbb{P}\big(\limsup\nolimits_{n\to\infty} \mathsf{A}_n\big) = 1.$$

Hint: Take complements and use the first part.

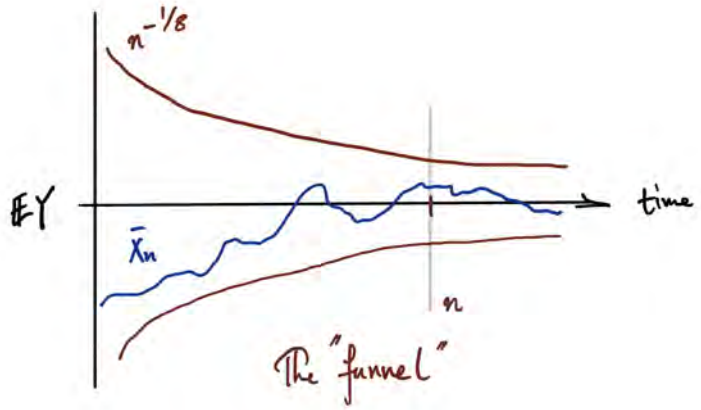


Figure 15.3 (The funnel). At time $n \in \mathbb{N}$, a sample path $\bar{X}_n(\omega)$ of the running average is very likely to lie inside the "funnel" with envelope $\pm n^{-1/8}$.

15.4.2 Proof of Cantelli's strong law

To simplify matters, we can and will assume that the random variable Y is centered.

Exercise 15.16 (Cantelli: Centering). Prove that we can take $\mathbb{E}\,Y = 0$ without loss of generality.

The main technical ingredient in the proof of Cantelli's theorem is the following claim, which we establish in Section 15.4.3.

Claim 15.17 (Cantelli: Tail bounds). Assume that $\mathbb{E}\,Y = 0$, and abbreviate $M := \mathbb{E}\,|Y|^4$. For all $n \in \mathbb{N}$, we have

$$\mathbb{P}\left\{|\bar{X}_n| \geq n^{-1/8}\right\} \leq 3M \cdot n^{-3/2}.$$

In other words, as the time n increases, the running average $\bar{X}_n$ is increasingly unlikely to fall outside a "funnel" around zero. See Figure 15.3. The key facts are that the funnel narrows to zero and that the probabilities are summable over time. You should compare this bound with the one that arises in the proof of Chebyshev's WLLN.

Proof of Theorem 15.11. Without loss, assume that $\mathbb{E}\,Y = 0$, and grant that Claim 15.17 is valid. We must show that $\bar{X}_n \to 0$ almost surely.

For each $n \in \mathbb{N}$, define the event A_n that the sample path $\big(\bar{X}_k(\omega) : k \in \mathbb{N}\big)$ lies outside the "funnel" at time n. That is,

$$\mathsf{A}_n := \left\{\omega \in \Omega : |\bar{X}_n(\omega)| \geq n^{-1/8}\right\}.$$

Claim 15.17 implies that

$$\sum\nolimits_{n=1}^{\infty} \mathbb{P}(\mathsf{A}_n) \leq 3M \sum\nolimits_{n=1}^{\infty} n^{-3/2} < +\infty.$$

The Borel–Cantelli lemma (Proposition 15.14) yields

$$\mathbb{P}\big(\limsup\nolimits_{n\to\infty} \mathsf{A}_n\big) = 0. \tag{15.2}$$

With probability one, only a finite number of the events A_n occur. As a consequence, $\bar{X}_n \to 0$ almost surely.

Here are the details. For each sample point $\omega \in \Omega$, we can define the number

$$N(\omega) := \sup\{n \in \mathbb{N} : \omega \in \mathsf{A}_n\} = \sup\{n \mathbb{1}_{\mathsf{A}_n}(\omega) : n \in \mathbb{N}\}.$$

We explicitly allow $N(\omega) = +\infty$.

In other words, $N(\omega)$ is the last time n that the sample point ω belongs to an event A_n. The function $\omega \mapsto N(\omega)$ is measurable because it is a countable supremum of measurable functions. The relation (15.2) ensures that the (extended) random variable $N < +\infty$ with probability one. Furthermore,

$$\mathbb{P}\left\{\omega \in \Omega : N(\omega) < +\infty \text{ and } |\bar{X}_n(\omega)| < n^{-1/8} \text{ for all } n > N(\omega)\right\} = 1.$$

As a particular consequence of the last display, it holds that

$$\mathbb{P}\left\{\omega \in \Omega : |\bar{X}_n(\omega)| \to 0\right\} = 1.$$

We conclude that $\bar{X}_n \to 0$ almost surely. ∎

15.4.3 Cantelli tail bounds: Proof

Finally, we must establish Claim 15.17. This result uses Markov's inequality to convert moment information into tail information, and it relies on independence to obtain a good bound for the moment.

Assume that $\mathbb{E}\, Y = 0$, and write $M := \mathbb{E}\,|Y|^4$. Markov's inequality (Theorem 10.13) applied to $|\bar{X}_n|^4$ ensures that

$$\mathbb{P}\left\{|\bar{X}_n| \geq n^{-1/8}\right\} \leq n^{1/2} \cdot \mathbb{E}\,|\bar{X}_n|^4 = n^{-3.5} \cdot \mathbb{E}\,|X_n|^4,$$

where $X_n := \sum_{i=1}^n Y_i$ is the (unnormalized) partial sum.

We bound the fourth moment of the partial sum by direct computation:

$$\begin{aligned}
\mathbb{E}\,|X_n|^4 &= \mathbb{E}\left[\sum\nolimits_{i,j,k,\ell=1}^n Y_i Y_j Y_k Y_\ell\right] \\
&= \sum\nolimits_{i=1}^n \mathbb{E}\, Y_i^4 + 6 \sum\nolimits_{i<k} \mathbb{E}[Y_i^2 Y_k^2] \\
&= n \cdot \mathbb{E}\, Y^4 + 3n(n-1) \cdot (\mathbb{E}\, Y^2)^2 \\
&\leq n \cdot \mathbb{E}\, Y^4 + 3n(n-1) \cdot \mathbb{E}\, Y^4 \leq 3n^2 M.
\end{aligned}$$

To reach the first relation, we expand the fourth power of the sum. Since the sequence $(Y_i : i \in \mathbb{N})$ consists of independent random variables with mean zero, the summands with an unpaired index have expectation zero. What remains are the terms where all indices are the same: $i = j = k = \ell$. Also remaining are the terms where $i = j < k = \ell$ or one of the other five other permutations of the letters in this formula. By independence and identical distribution, since $i < k$, we can see that $\mathbb{E}[Y_i^2 Y_k^2] = (\mathbb{E}\, Y^2)^2 \leq \mathbb{E}\, Y^4$, where the last relation is Jensen's inequality. Finally, we write $M := \mathbb{E}\, Y^4$ and combine terms.

Problems

Problem 15.18 (Kolmogorov's maximal inequality). Kolmogorov improved Chebyshev's inequality as follows. Consider an independent family $(Y_1, \dots, Y_n)$ of zero-mean random variables in L_2. Introduce the partial sums $X_k = \sum_{i=1}^k X_i$ for $k = 1, \dots, n$. Then

$$\mathbb{P}\{\max\nolimits_{k \leq n} |X_k| \geq t\} \leq \frac{1}{t^2} \cdot \mathrm{Var}[X_n] \quad \text{for } t > 0.$$

This bound is called *Kolmogorov's maximal inequality*. In Lecture 26, we will develop some far-reaching generalizations using martingale methods.

1. Use the union bound and Chebyshev's inequality to obtain a (much) weaker tail bound for $\max_{k \le n} |X_k|$.
2. For each $k = 1, \dots, n$, define the event E_k that $|X_k| \ge t$ while $|X_j| < t$ for all $j < k$. Show that these events are mutually exclusive.
3. Argue that $(X_n - X_k)$ is independent from the family $(Y_1, \dots, Y_k)$. Hence, $(X_n - X_k)$ is independent from X_k and from E_k.
4. Verify that $\mathbb{E}[\mathbb{1}_{\mathsf{E}_k} X_k (X_n - X_k)] = 0$.
5. For each $k = 1, \dots n$, show that

$$\mathbb{P}(\mathsf{E}_k) \le \frac{1}{t^2} \mathbb{E}[\mathbb{1}_{\mathsf{E}_k} X_k^2] \le \frac{1}{t^2} \mathbb{E}[\mathbb{1}_{\mathsf{E}_k} X_n^2].$$

6. Sum these inequalities to conclude that the maximal inequality holds.

Problem 15.19 (*Independent sums in L_2: Convergence theory). Consider an *independent* sequence $(Y_1, Y_2, Y_3, \dots) \in \mathsf{L}_2$ of *square-integrable* random variables. We assume that the random variables are centered and have controlled variance:

$$\mathbb{E}[Y_i] = 0 \quad \text{for } i \in \mathbb{N} \quad \text{and} \quad \sum\nolimits_{i=1}^{\infty} \operatorname{Var}[Y_i] < +\infty.$$

Define the partial sums $X_n := \sum_{i=1}^n Y_i$ for $n \in \mathbb{N}$. By direct arguments, we can show that the partial sums converge to a limit in L_2 and almost surely. Later, in Lecture 25, we will learn how to obtain the same results from a martingale argument.

1. **L_2 convergence:** Show that $(X_n : n \in \mathbb{N})$ is a Cauchy sequence in L_2, so it converges to a limit in L_2.
2. ***Almost-sure convergence:** A fortiori, demonstrate that $(X_n : n \in \mathbb{N})$ converges almost surely. **Hint:** It suffices to prove that $\max_{m,n \ge N} |X_m - X_n| \to 0$ almost surely as $N \to \infty$. For this purpose, you can apply the Kolmogorov maximal inequality (Problem 15.18) to $\max_{k \le n} |X_{N+k} - X_N|$ and take a limit as $n \to \infty$.

Problem 15.20 (*Kronecker's lemma). Consider two sequences $(a_n : n \in \mathbb{N})$ and $(x_n : n \in \mathbb{N})$ of real numbers. Assume that $0 < a_n \uparrow +\infty$. Prove the following statement:

$$\sum\nolimits_{i=1}^{\infty} \frac{x_i}{a_i} \quad \text{converges} \quad \text{implies} \quad \frac{1}{a_n} \sum\nolimits_{i=1}^{n} x_i \to 0 \quad \text{as } n \to \infty.$$

Hint: Use summation by parts.

Problem 15.21 (Slytherins).** In this problem, we will establish two versions of the strong law of large numbers (SLLN). These results have weaker hypotheses than Cantelli's SLLN (Theorem 15.11), but the proofs are commensurately harder. See Lecture 25 for some alternative approaches using martingale methods.

Consider an *independent* sequence $(Y_1, Y_2, Y_3, \dots) \in \mathsf{L}_2$ of *square-integrable* random variables. We assume that the random variables are centered and have controlled variance:

$$\mathbb{E}[Y_i] = 0 \quad \text{for } i \in \mathbb{N} \quad \text{and} \quad \sum\nolimits_{i=1}^{\infty} \frac{\operatorname{Var}[Y_i]}{i^2} < +\infty.$$

We do *not* assume identical distribution at this point. Define the partial sums $X_n := \sum_{i=1}^n Y_i$ for $n \in \mathbb{N}$.

1. **L_2 SLLN:** Use Problem 15.19 to establish that $\sum_{i=1}^n Y_i / i$ converges almost surely as $n \to \infty$. Apply Kronecker's lemma (Problem 15.20) to conclude that $X_n / n \to 0$ almost surely. Formulate a SLLN for the running average of independent, square-integrable real random variables.

The L_2 SLLN is about as good as it gets unless we make stronger hypotheses to link the summands, such as an i.i.d. assumption.

2. ****Kolmogorov's SLLN:** Now, assume that $(Y_i : i \in \mathbb{N})$ are *i.i.d.* copies of an integrable random variable $Y \in \mathsf{L}_1$. Prove Kolmogorov's SLLN (Theorem 15.10). **Hint:** Apply the L_2 SLLN to the truncated random variables $Y_i \mathbb{1}_{|Y_i| \le i}$.

Problem 15.22 (Renewal theorem). Independent sums arise in the study of waiting times for (real-world) events to occur, such as emission of a photon from a radioactive mass or the completion of a task by a computer server. In this problem, we establish the *renewal theorem*, a fundamental result that describes how many events occur per unit time in the long run. This result has many practical applications (e.g., in queuing), as well as theoretical application (e.g., in the study of Markov chains).

Let $Y \in \mathsf{L}_1$ be *positive* random variable that models the waiting time, and write $m = \mathbb{E}\, Y$ for the expected waiting time. Now, consider an *i.i.d.* sequence $(Y_1, Y_2, Y_3, \dots)$ of copies Y. For each time $t \ge 0$, define a random variable N_t that counts the total number of events that have occurred up to time t:

$$N_t := \sup\{n \in \mathbb{N} : Y_1 + Y_2 + \cdots + Y_n \le t\}.$$

We will prove that

$$\frac{N_t}{t} \to \frac{1}{m} \quad \text{almost surely.} \tag{15.3}$$

In words, in the long-run, the number of events per unit time is the reciprocal of the wait time with probability one.

1. For each $t \in \mathbb{R}$, explain why N_t is a (finite) real random variable.
2. As usual, define the partial sums $X_n := \sum_{i=1}^n Y_i$. Explain why $X_n/n \to m$ almost surely as $n \to \infty$.
3. Show that $N_t \to \infty$ almost surely as $t \to \infty$. **Hint:** For each $n \in \mathbb{N}$, the event $\{N_t \ge n\} = \{X_n \le t\}$.
4. Deduce that
$$\mathbb{P}\left\{ \frac{X_{N_t}}{N_t} \to m \quad \text{and} \quad \frac{X_{1+N_t}}{1+N_t} \to m \quad \text{as } t \to \infty \right\} = 1.$$
5. Verify the pair of inequalities
$$\frac{X_{N_t}}{N_t} \le \frac{t}{N_t} < \frac{X_{1+N_t}}{N_t}.$$
6. Confirm that the upper and lower bounds in the last display converge almost surely to m. Conclude that the renewal theorem (15.3) is valid.

Applications

Application 15.23 (Monte Carlo integration). Monte Carlo integration is a fundamental computational method for approximating integrals. It is most suitable for high-dimensional integrals and for integrals with respect to a distribution that may be hard to sample directly. We explore the simplest form of the method.

Abstractly, suppose that μ is a probability measure on $\mathbb{R}^d$. Let $f : \mathbb{R}^d \to \mathbb{R}$ be a μ-integrable function. We would like to evaluate the integral

$$I = \int f \, \mathrm{d}\mu.$$

The Monte Carlo approach proceeds by drawing independent random samples $Y_1, Y_2, Y_3, \ldots$ from the distribution μ. Then we compute $X_k = f(Y_k)$, and we form the approximation

$$\hat{I}_n = \frac{1}{n} \sum_{k=1}^{n} X_k.$$

1. For each n, show that $\hat{I}_n$ is an unbiased estimator. That is, $\mathbb{E}\, \hat{I}_n = I$.
2. Use Kolmogorov's SLLN to explain what happens as $n \to \infty$. (Does Cantelli's SLLN apply?)
3. From now on, assume that f^2 is μ-integrable. Give a bound for the variance $\mathrm{Var}[\hat{I}_n]$. **(*)** Use the central limit theorem (Lecture 18) to describe the fluctuations of the error $|\hat{I}_n - I|$ as $n \to \infty$.
4. In practice, we only get a finite number of samples. For n samples, about how big do you anticipate the error $|\hat{I}_n - I|$ will be? Does the ambient dimension d play a direct role?
5. Explain how to use Monte Carlo integration to estimate two numerical constants via the formulas

$$\pi = 4 \int_{[0,1]^2} \mathbb{1}\{x^2 + y^2 \le 1\}\, \lambda^2(\mathrm{d}x \times \mathrm{d}y) \quad = 3.14159\,26535\,89793\,23846\ldots$$

$$\gamma = -\int_0^\infty \log(x)\, \mathrm{e}^{-x}\, \lambda(\mathrm{d}x) \quad = 0.57721\,56649\,01532\,86060\ldots.$$

6. For each integral, perform the following computer experiment. Let $n = 10^i$ for $i = 1, 2, 3, 4, 5$. Estimate the integral using n samples. Repeat 1000 times. Make a histogram of the estimates. Report the mean and variance of the distribution. Discuss.
7. For each integral, perform the following computer experiment. Draw 1000 samples X_k. Compute the sample path $n \mapsto \hat{I}_n$. Repeat this process 100 times. Plot all 100 sample paths on the same graph with translucent lines. Discuss.
8. **(*)** Approximate the integral for γ using Gauss–Laguerre quadrature with $n = 2^i$ samples for $i = 1, \ldots, 8$. Estimate the convergence rate. Discuss.

Notes

You will find similar material on laws of large numbers in any book on probability theory.

16. Concentration Inequalities

Agenda:
1. Chebyshev's inequality
2. Exponential moments
3. The Laplace transform method
4. Hoeffding's inequality
5. Bernstein's inequality

> "Concentrate all your thoughts upon the work at hand. The sun's rays do not burn until brought to a focus."
>
> —Alexander Graham Bell

A concentration inequality bounds the probability that a random variable takes a value that is significantly different from its expectation:

$$\mathbb{P}\{|X - \mathbb{E}\,X| \geq t\} \leq \dots.$$

Concentration inequalities are *nonasymptotic*: they deliver concrete information about particular random variables, so they are well suited for applications. Indeed, concentration inequalities are among the most widely used tools in modern statistics, mathematics of data science, and related fields.

In this lecture, we develop the basic theory of concentration inequalities for independent sums. These results give very strong bounds on the tail probabilities of an independent sum with a fixed number of terms. In many situations, we can obtain tail bounds with *exponential decay*—or sometimes even better.

16.1 Example: Chebyshev's inequality

We have already encountered the simplest concentration result. Indeed, Chebyshev's variance inequality provides a bound for the tail probability of a random variable in terms of its variance.

Recall that the variance may be defined as the expected squared deviation of a random variable from its expectation:

$$\mathrm{Var}[Y] := \mathbb{E}(Y - \mathbb{E}\,Y)^2.$$

Thus, we can think about the standard deviation, $\mathrm{stdev}(Y) := \sqrt{\mathrm{Var}[Y]}$, as the scale on which the random variable typically fluctuates around its expectation. Note that the standard deviation has the same units as the random variable. We can restate Chebyshev's inequality as follows.

Proposition 16.1 (Chebyshev). Let $X \in \mathsf{L}_2$ be a real random variable. Then

$$\mathbb{P}\{|X - \mathbb{E}\,X| \geq \mathrm{stdev}[X] \cdot t\} \leq 1 \wedge t^{-2} \quad \text{for all } t > 0.$$

Recall that $\wedge$ denotes the minimum of two numbers.

Chebyshev's inequality is a natural tool for studying a sum of mutually uncorrelated random variables. In this case, the variance of the sum is the sum of the variances.

Example 16.2 (Concentration: Uncorrelated sum). Consider a family $(Y_1, \dots, Y_n) \in \mathsf{L}_2$ of *mutually uncorrelated* real random variables that are square integrable. Introduce the ordinary sum $X := \sum_{i=1}^n Y_i$. By the Pythagorean relation (Proposition 12.16),

$$\sigma^2 := \operatorname{Var}[X] = \sum\nolimits_{i=1}^n \operatorname{Var}[Y_i].$$

Chebyshev's inequality yields the tail bound

$$\mathbb{P}\{|X - \mathbb{E}\,X| \geq \sigma \cdot t\} \leq 1 \wedge t^{-2} \quad \text{for } t > 0.$$

This is a concentration inequality for a sum of uncorrelated random variables. ▪

Let emphasize again that Proposition 16.2 holds whenever the summands are mutually uncorrelated. We need *neither* independence *nor* identical distribution. Furthermore, the standard deviation σ is a natural scale for studying the deviation of X from its expectation. On this scale, the tail probability decays at least as fast as t^{-2} for all $t > 0$. See Figure 16.1 for an illustration.

Example 16.3 (Concentration: Running average). As a particular example, consider an independent sequence $(Y_1, Y_2, Y_3, \dots)$ of copies of a random variable $Y \in \mathsf{L}_2$. Let $\bar{X}_n := n^{-1} \sum_{i=1}^n Y_i$ be the running average. Since independent random variables are uncorrelated (Exercise 13.11), Example 16.2 implies that

$$\mathbb{P}\left\{\left|\bar{X}_n - \mathbb{E}\,Y\right| \geq \frac{\operatorname{stdev}[Y] \cdot t}{\sqrt{n}}\right\} \leq \frac{1}{t^2} \quad \text{for } t > 0.$$

A striking feature of this formulation is that the scale for concentration *decreases* as the number n of summands grows. Making the change of variables $t \mapsto t\sqrt{n}$, we have the alternative expression

$$\mathbb{P}\left\{\left|\bar{X}_n - \mathbb{E}\,Y\right| \geq \operatorname{stdev}[Y] \cdot t\right\} \leq \frac{1}{nt^2} \quad \text{for } t > 0.$$

In other words, the probability of a fluctuation larger than a fixed size decreases as the number of summands grows. ▪

We may summarize the key points. Chebyshev's inequality is a concentration inequality: it gives a bound for the probability that a random variable is far from its expected value. For uncorrelated sums, Chebyshev's inequality exploits the fact that the variance is additive. The resulting bound only involves the coarsest features of the individual summands (that is, their variances). It operates under weak assumptions (square-integrability), and it is very easy to use.

On the other hand, Chebyshev's inequality gives rather limited information on the tail decay. If we want to improve, we need to pose further assumptions. In the rest of this lecture, we will see that we can obtain much steeper concentration if we require that the summands are independent and bounded.

16.2 Exponential moments

As we saw in Lecture 10, bounds on polynomial moments are roughly equivalent to polynomial bounds on tail decay. In this section, we take this idea to an extreme by introducing the concept of an *exponential* moment. Exponential moments provide a way to check that the tails of a random variable decay exponentially fast.

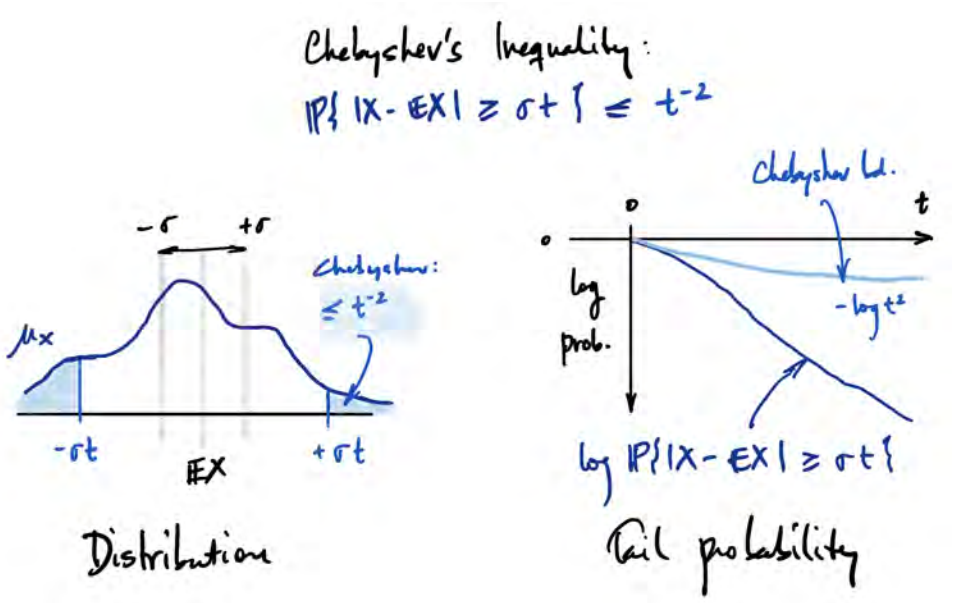


Figure 16.1 (Chebyshev's inequality). This diagram depicts the bound produced by Chebyshev's inequality. It shows that the probability of a deviation on the scale of the standard deviation σ decays at least quadratically.

16.2.1 Moment and cumulant generating functions

To begin, let us introduce two functions that pack up information about the exponential moments of a random variable.

Definition 16.4 (Mgf and cgf). Let X be a real random variable. Define the *moment generating function (mgf)*:

$$m_X(\theta) := \mathbb{E}\, \mathrm{e}^{\theta X} \quad \text{for each } \theta \in \mathbb{R}.$$

Define the *cumulant generating function (cgf)*:

$$\xi_X(\theta) := \log m_X(\theta) = \log \mathbb{E}\, \mathrm{e}^{\theta X} \quad \text{for each } \theta \in \mathbb{R}.$$

The mgf and cgf are always defined, but they can take extended real values.

The mgf and cgf are useful only for random variables that have very rapidly decaying tails. In particular, observe that $X \in \mathsf{L}_\infty$ implies that $m_X(\theta)$ and $\xi_X(\theta)$ take finite values for all $\theta \in \mathbb{R}$. The mgf and cgf are also valuable for random variables that have exponential tail decay, in which case they may not be finite on the whole real line (Exercise 16.11).

16.2.2 Properties of exponential moments

Let us summarize some of the basic properties of the mgf and cgf.

Exercise 16.5 (Mgf: Convexity). Prove that m_X is a positive, convex function. **Hint:** The mgf is an average of positive, convex functions.

Exercise 16.6 (Cgf: Shifts). Show that $\xi_{X+a}(\theta) = \xi_X(\theta) + a\theta$ for each scalar $a \in \mathbb{R}$.

Problem 16.7 (Cgf: Convexity and exponential means). For simplicity, assume that X is a bounded, nonconstant random variable. Similar results hold for random variables whose cgf $\xi_X(\theta)$ is finite on a neighborhood of $\theta = 0$.

1. Observe that $\xi_X(0) = 0$ without qualification.
2. Using bounded convergence, compute the first and second derivative of $\xi_X(\theta)$.
3. Deduce that $\xi_X'(0) = \mathbb{E}[X]$ and $\xi_X''(0) = \operatorname{Var}[X]$.
4. **(*)** Prove that ξ_X is a strictly convex function. **Hint:** Rewrite $\xi_X''(\theta)$ as the variance of a nonconstant random variable related to X.
5. **(*)** Using Jensen's inequality, verify that

$$\inf X \le \frac{1}{(-\theta)}\xi_X(-\theta) \le \mathbb{E}\, X \le \frac{1}{\theta}\xi_X(\theta) \le \sup X \quad \text{for } \theta > 0.$$

   Note that $\lim_{\theta\to 0} \theta^{-1}\xi_X(\theta) = \mathbb{E}[X]$.
6. **(*)** More generally, show that $\theta \mapsto \theta^{-1}\xi_X(\theta)$ is an increasing function. These results support the interpretation of $\theta^{-1}\xi_X(\theta)$ as an *exponential mean* of the random variable X, which is a type of weighted average parameterized by θ.

Recall that a function with two continuous derivatives is convex if and only if the second derivative is positive.

16.2.3 Examples

In this section, we undertake mgf and cgf calculations for some important classes of random variables. We typically use cgfs to study the concentration of a random variable about its mean, so it is often more natural to study the cgf of a centered random variable.

Exercise 16.8 (Cgf: Bernoulli distribution). Let $Y \sim \textsc{bernoulli}(p)$ be a Bernoulli random variable with mean $p \in [0, 1]$. Calculate that the mgf takes the form

$$m_Y(\theta) = 1 + p \cdot (\mathrm{e}^\theta - 1) \quad \text{for all } \theta \in \mathbb{R}.$$

Deduce that cgf satisfies the bound $\xi_Y(\theta) \le p \cdot (\mathrm{e}^\theta - 1)$ for all $\theta \in \mathbb{R}$.

Exercise 16.9 (Cgf: Poisson distribution). Let $Y \sim \textsc{poisson}(\beta)$ be a Poisson random variable with mean $\beta \in \mathbb{R}_+$. Calculate that the cgf of the *centered* variable $Z = Y - \mathbb{E}\, Y$ takes the form

$$\xi_Z(\theta) = \mathrm{e}^{\beta\theta} - \beta\theta - 1 \quad \text{for all } \theta \in \mathbb{R}.$$

This is an example of an unbounded random variable whose cgf is finite on the whole real line. How fast does $\xi_Z(\theta)$ grow as $\theta \to -\infty$ and as $\theta \to +\infty$?

Exercise 16.10 (Cgf: Normal distribution). Let $Z \sim \textsc{normal}(0, \sigma^2)$ be a centered normal random variable with variance $\sigma^2 \in \mathbb{R}_+$. Calculate that the cgf takes the form

$$\xi_Z(\theta) = \frac{\sigma^2\theta^2}{2} \quad \text{for all } \theta \in \mathbb{R}.$$

This is another example of an unbounded random variable whose cgf is finite on the whole real line. **Hint:** Complete the square in the exponential, change variables, and use the fact that the standard normal density has total mass one.

Exercise 16.11 (Cgf: Exponential distribution). Let $Y \sim \textsc{exponential}(\beta)$ be an exponential random variable with mean $\beta \in \mathbb{R}_+$. This is an example of an unbounded random variable whose cgf is only finite on part of the real line.

1. Calculate that the cgf takes the form

$$\xi_Y(\theta) = -\log(1 - \beta\theta) \quad \text{for all } \theta < \beta^{-1}.$$

2. Show that the cgf of the *centered* random variable $Z = Y - \mathbb{E}\, Y$ satisfies

$$\xi_Z(\theta) \le \frac{\beta^2\theta^2/2}{1-\beta\theta_+} \quad \text{for all } \theta < \beta^{-1}.$$

   The factor β^2 in the numerator is the variance of Z. **Hint:** For $\theta > 0$, compare the full Taylor series. For $\theta < 0$, use a second-order Taylor expansion with exact remainder.
3. Plot the cgf $\xi_Z(\theta)$ and the upper bound. Observe that they agree to second order at $\theta = 0$.

Exercise 16.12 (Tails control exponential moments). Let X be a real random variable. Suppose that there are strictly positive constants $a, b > 0$ for which

$$\mathbb{P}\{|X| \ge t\} \le a\mathrm{e}^{-bt} \quad \text{for all } t \ge 0.$$

Find an upper bound for $m_X(\theta)$, and deduce that $m_X(\theta)$ is finite on an open interval containing $\theta = 0$. **Hint:** Use integration by parts (Theorem 10.16).

16.2.4 Cumulants are additive

The cgf is an ideal tool for studying *independent sums* because the cgf of an independent sum is the sum of the cgfs. This property will serve as a powerful substitute for the additivity of the variance for an *uncorrelated sum* (Proposition 12.16).

Proposition 16.13 (Cgf: Additivity). Consider an *independent* family $(Y_1, \dots, Y_n)$ of real random variables, and form the sum $X = \sum_{i=1}^n Y_i$. Then

$$m_X(\theta) = \prod\nolimits_{i=1}^n m_{Y_i}(\theta) \quad \text{for each } \theta \in \mathbb{R}.$$

In particular, taking logarithms,

$$\xi_X(\theta) = \sum\nolimits_{i=1}^n \xi_{Y_i}(\theta) \quad \text{for each } \theta \in \mathbb{R}.$$

Proof. This result follows from a short, magical calculation:

$$m_X(\theta) = \mathbb{E}\,\mathrm{e}^{\theta X} = \mathbb{E}\prod\nolimits_{i=1}^n \mathrm{e}^{\theta Y_i} = \prod\nolimits_{i=1}^n \mathbb{E}\,\mathrm{e}^{\theta Y_i} = \prod\nolimits_{i=1}^n m_{Y_i}(\theta).$$

Of course, the exponential of a sum is the product of the exponentials. Since the family $(Y_i : i = 1, \dots, n)$ of random variables is independent, the expectation of a product of functions of independent random variables is the product of the expectations (Proposition 13.8). Since all the exponentials are positive, there are no concerns about integrability. ∎

Exercise 16.14 (Cgf: Binomial distribution). Using Exercise 16.8 and Proposition 16.13, deduce that the cgf of a random variable X with the BINOMIAL(n, p) distribution satisfies

$$\xi_X(\theta) \le np \cdot (\mathrm{e}^\theta - 1) \quad \text{for all } \theta \in \mathbb{R}.$$

Consider the centered binomial variable $Z = X - \mathbb{E}[X]$. Confirm that

$$\xi_Z(\theta) \le np \cdot (\mathrm{e}^\theta - \theta - 1).$$

The similarity with the Poisson cgf bound (Exercise 16.9) is no accident!

Exercise 16.15 (Cgf: Gamma distribution). Consider the gamma random variable $X \sim \textsc{gamma}(n, \beta)$ with shape $n \in \mathbb{N}$ and scale $\beta \in \mathbb{R}_+$. Form the centered random variable $Z = X - \mathbb{E}\, X$. Show that the cgf satisfies

$$\xi_Z(\theta) \le \frac{n\beta^2\theta^2/2}{1 - \beta\theta_+} \quad \text{for } \theta < \beta^{-1}.$$

Hint: When the shape parameter n is a natural number, the $\textsc{gamma}(n, \beta)$ random variable is a sum of i.i.d. exponential random variables with mean β.

16.2.5 *Generating functions

You may wonder about the "generating function" terminology. As you saw in Problem 10.23, the derivatives of the mgf at zero deliver the polynomial moments of the random variable:

$$(\mathrm{D}^p m_X)(0) = \mathbb{E}\, X^p \quad \text{for each } p \in \mathbb{Z}_+.$$

In words, m_X is the exponential generating function of the sequence $(\mathbb{E}\, X^p : p \in \mathbb{Z}_+)$ of polynomial moments. See also Exercise 10.8.

Similarly, we define the pth cumulant κ_p of the random variable to be

$$(\mathrm{D}^p \xi_X)(0) =: \kappa_p(X) \quad \text{for each } p \in \mathbb{N}.$$

In words, ξ_X is the exponential generating function of the sequence $(\kappa_p(X) : p \in \mathbb{N})$ of cumulants of X.

What are these cumulants? The first two are familiar: $\kappa_1(X) = \mathbb{E}\, X$ and $\kappa_2(X) = \mathrm{Var}[X]$. Higher-order cumulants are harder to write down and interpret. Although cumulants are less intuitive than moments, they have a number of algebraic properties that make them more fundamental objects. We have glimpsed this fact in Proposition 16.13. See also Problem 16.37.

16.3 The Laplace transform method

We have used Markov's inequality several times to obtain tail bounds when we control polynomial moments. Chebyshev's inequality provides one immediate example. Likewise, we can use Markov's inequality to obtain tail bounds when we control exponential moments.

16.3.1 Tail bounds via cgfs

Let us show how to use the cumulant generating function to derive tail bounds with exponential decay (or better!). To use this result, it suffices to have good upper bounds on the cgf.

Theorem 16.16 (Laplace transform method). Let X be a real random variable. Then, for each $t \in \mathbb{R}$,

$$\mathbb{P}\{X \ge t\} \le \exp\left(-\sup_{\theta>0}(\theta t - \xi_X(\theta))\right);$$
$$\mathbb{P}\{X \le t\} \le \exp\left(-\sup_{\theta<0}(\theta t - \xi_X(\theta))\right).$$

Proof. Fix a parameter $\theta > 0$. Note that the function $x \mapsto \mathrm{e}^{\theta x}$ is strictly increasing and strictly positive. Therefore,

$$\mathbb{P}\{X \ge t\} = \mathbb{P}\left\{\mathrm{e}^{\theta X} \ge \mathrm{e}^{\theta t}\right\} \le \mathrm{e}^{-\theta t} \cdot \mathbb{E}\, \mathrm{e}^{\theta X} = \mathrm{e}^{-\theta t} \cdot m_X(\theta).$$

The inequality is Markov's. Combining the terms on the right-hand side in the exponent and recognizing the cgf, we arrive at the bound

$$\mathbb{P}\{X \geq t\} \leq \exp\left(-(\theta t - \xi_X(\theta))\right).$$

Since the parameter $\theta > 0$ is arbitrary, we can take the infimum of the right-hand side over $\theta > 0$. This leads to the first inequality in the statement.

For the second inequality, fix $\theta < 0$, and note that $x \mapsto \mathrm{e}^{\theta x}$ is strictly decreasing and strictly positive. Thus,

$$\mathbb{P}\{X \leq t\} = \mathbb{P}\left\{\mathrm{e}^{\theta X} \geq \mathrm{e}^{\theta t}\right\}.$$

The rest of the argument is the same. ∎

As a first example, let us devise a clean tail bound for a normal random variable.

Example 16.17 (Normal random variable: Tail bounds). Let $Z \sim \textsc{normal}(0, \sigma^2)$ be a centered normal random variable with variance σ^2. Using the Laplace transform method (Theorem 16.16) and the normal cgf calculation (Exercise 16.10),

$$\mathbb{P}\{Z \geq t\sigma\} \leq \exp\left(-\sup_{\theta>0}(t\sigma\theta - \sigma^2\theta^2/2)\right) = \mathrm{e}^{-t^2/2} \quad \textit{for all } t > 0.$$

Indeed, the supremum is achieved when $\theta = t/\sigma$. We can obtain a parallel bound for the lower tail via the same method. Alternatively, note that Z and $-Z$ have the same distribution. Compare with Exercise 10.21. ▪

Exercise 16.18 (Exponential moments control tails). Suppose that $\xi_X(\theta)$ is finite on an open interval containing $\theta = 0$. Argue that the upper tail probability decays at least as fast as an exponential function. That is, there exist strictly positive constants $a, b > 0$ for which

$$\mathbb{P}\{X \geq t\} \leq a\mathrm{e}^{-bt} \quad \text{for all } t \geq 0.$$

Formulate and prove an analogous result for the lower tail. This is the converse of Exercise 16.12.

> **Aside:** The Laplace transform method is so called because the mgf is the Laplace transform of the distribution μ_X of the random variable X. The idea of using Laplace transforms to produce tail bounds appears in Bernstein's 1927 probability text (in Russian). In the West, this idea is often associated with the names Cramér, Chernoff, and Hoeffding, but the works that support this attribution were not written until somewhat later.

16.3.2 Tail bounds for independent sums

Owing to the additivity of cgfs, the Laplace transform method yields particularly elegant bounds for an independent sum.

Corollary 16.19 (Laplace transform: Independent sum). Consider an *independent* family $(Y_1, \dots, Y_n)$ of real random variables. Form the sum $X = \sum_{i=1}^n Y_i$. For $t \in \mathbb{R}$,

$$\mathbb{P}\{X \geq t\} \leq \exp\left(-\sup_{\theta>0}\left(\theta t - \textstyle\sum_{i=1}^n \xi_{Y_i}(\theta)\right)\right);$$
$$\mathbb{P}\{X \leq t\} \leq \exp\left(-\sup_{\theta<0}\left(\theta t - \textstyle\sum_{i=1}^n \xi_{Y_i}(\theta)\right)\right).$$

Proof. Combine Theorem 16.16 and Proposition 16.13. ∎

As a simple example, we can work out tail bounds for a centered binomial random variable. Related calculations are widely applicable (Problem 16.38).

Example 16.20 (Binomial random variable: Tail bounds). Let $X \sim \text{BINOMIAL}(n, p)$ be a binomial random variable, for which $\mathbb{E}[X] = np$. Using Corollary 16.19 and the binomial cgf bound (Exercise 16.14), we find that

$$\begin{aligned}\mathbb{P}\{X - \mathbb{E}[X] \geq t \cdot \mathbb{E}[X]\} &\leq \exp\left(-np \cdot \sup_{\theta>0}\left(\theta t - (\mathrm{e}^{\theta} - \theta - 1)\right)\right) \\ &= \exp\left(-np \cdot ((1+t)\log(1+t) - t)\right) \\ &= \left(\frac{\mathrm{e}^{t}}{(1+t)^{1+t}}\right)^{np} \quad \text{for } t > 0.\end{aligned}$$

The supremum is attained at $\theta = \log(1+t)$. Similarly,

$$\mathbb{P}\{X - \mathbb{E}[X] \leq -t \cdot \mathbb{E}[X]\} \leq \left(\frac{\mathrm{e}^{-t}}{(1-t)^{(1-t)}}\right)^{np} \quad \text{for } t \in (0,1).$$

These results provide elegant bounds on the probability that a binomial random variable is significantly larger or smaller than its expected value. It is no coincidence that the cgf of a centered Poisson variable (Exercise 16.9) appears in the calculation! ▪

16.3.3 *The rate function

For the Laplace transform method, the most common use case occurs when $\mathbb{E}\, X = 0$. Under this assumption, we can express the result in a more compact way. Define the *rate function* of the *centered* random variable X:

$$\Lambda_X(t) := \sup_{\theta \in \mathbb{R}}(\theta t - \xi_X(\theta)) \quad \text{for } t \in \mathbb{R}. \tag{16.1}$$

Then, for each $t \geq 0$,

$$\begin{aligned}\mathbb{P}\{X \geq +t\} &\leq \mathrm{e}^{-\Lambda_X(+t)}; \\ \mathbb{P}\{X \leq -t\} &\leq \mathrm{e}^{-\Lambda_X(-t)}.\end{aligned} \tag{16.2}$$

Thus, the rate function Λ_X contains a full spectrum of tail bounds for X.

Problem 16.21 (Rate function). Let X be a real random variable with $\mathbb{E}[X] = 0$. For simplicity, you may assume that X is bounded and nonconstant, so the cgf ξ_X is finite and strictly convex. This problem relies on the results from Problem 16.7.

1. Check that the supremum in the definition (16.1) of the rate function $\Lambda_X(t)$ is attained at the (unique) value $\theta = \theta(t)$ that solves $\xi_X'(\theta) = t$.
2. When $t > 0$, confirm that $\theta(t) > 0$. When $t < 0$, confirm that $\theta(t) < 0$. **Hint:** Recall that ξ_X' is increasing and $\xi_X'(0) = 0$.
3. Explain how (16.2) follows from Theorem 16.16 and item (2).
4. Show that the rate function Λ_X is a convex function. **Hint:** Note that Λ_X is a supremum of affine functions.
5. Deduce that Λ_X is a positive function whose minimal value $\Lambda_X(0) = 0$.
6. Argue that $\Lambda_X(t)/|t|$ is bounded away from zero as $t \to -\infty$ or $t \to +\infty$. **Hint:** See Exercise 16.18.

> **Aside:** If you are familiar with convex analysis, you will recognize that the rate function Λ_X is the Fenchel–Young conjugate (or the Legendre transform) of the cgf ξ_X. See Problem 9.47.

16.3.4 *Cramér's theorem

Although it may seem that the Laplace transform approach is merely a clever trick, it actually results in sharp (asymptotic) bounds for the tails of an i.i.d. sum. This is the foundational result in the study of *large-deviation principles*.

Theorem 16.22 (Cramér). Consider a sequence $(Y_i : i \in \mathbb{N})$ of *i.i.d. copies* of a bounded random variable $Y \in \mathsf{L}_\infty$ with $\mathbb{E}\, Y = 0$. Form the running averages $\bar{X}_n := n^{-1} \sum_{i=1}^n Y_i$ for $n \in \mathbb{N}$. For all $t \geq 0$,

$$\lim_{n\to\infty} \frac{1}{n} \log \mathbb{P}\left\{\bar{X}_n \geq t\right\} = -\Lambda_Y(t).$$

Parallel statements hold for the lower tail at each level $t < 0$.

Corollary 16.19 already yields the upper bound $-\Lambda_Y(t)$ for the limit appearing in Cramér's theorem. The lower bound takes a substantial amount of extra work. See the problems section of Lecture 18 for a proof sketch.

16.4 Example: Hoeffding's inequality

So far, our presentation does not make a very strong case for the utility of the Laplace transform method. To illustrate the wider implications of this approach, we present a powerful concentration inequality, due to Hoeffding. This result shows that an independent sum of bounded random variables has Gaussian tail decay.

16.4.1 The Hoeffding inequality

We begin with a statement of the result and some discussion. Then we turn to the proof.

Theorem 16.23 (Hoeffding). Consider an *independent* family $(Y_1, \dots, Y_n)$ of real random variables that satisfy uniform bounds

$$|Y_i - \mathbb{E}\, Y_i| \leq a_i \quad \text{for each index } i = 1, \dots, n.$$

Form the sum $X = \sum_{i=1}^n Y_i$. Then, for all $t \geq 0$,

$$\mathbb{P}\left\{|X - \mathbb{E}\, X| \geq t\sqrt{v}\right\} \leq 2\,\mathrm{e}^{-t^2/2}, \quad \text{where} \quad v := \sum\nolimits_{i=1}^n a_i^2.$$

Exercise 16.24 (Hoeffding: Centering). Without loss of generality, we can prove Hoeffding's inequality with the extra assumption that $\mathbb{E}\, Y_i = 0$ for each $i = 1, \dots, n$. Check this claim.

Hoeffding's inequality states that an independent sum X of bounded random variables is extremely unlikely to take a value far from its expectation. The tail probability of the variable X has a profile similar to a Gaussian random variable with expectation $\mathbb{E}\, X$ and variance v.

For this reason, the number v is sometimes called the *variance proxy*. It is always an upper bound for the actual variance:

$$\begin{aligned}\operatorname{Var}[X] = \sum\nolimits_{i=1}^n \operatorname{Var}[Y_i] &= \sum\nolimits_{i=1}^n \mathbb{E}(Y_i - \mathbb{E}\, Y_i)^2 \\ &\leq \sum\nolimits_{i=1}^n \|Y_i - \mathbb{E}\, Y_i\|_\infty^2 \leq \sum\nolimits_{i=1}^n a_i^2 = v.\end{aligned}$$

There are situations where the variance proxy coincides with the true variance. For example, if each $Y_i \sim \text{UNIFORM}\{\pm 1\}$, then $\text{Var}[X] = v$. On the other hand, the difference between $\text{Var}[X]$ and v can be arbitrarily large (when the variance of Y_i is much smaller than the upper bound a_i).

Example 16.25 (Concentration: Running average). Consider an independent sequence $(Y_1, Y_2, Y_3, \dots)$ of copies of a random variable $Y \in \mathsf{L}_\infty$. Let $\bar{X}_n := n^{-1} \sum_{i=1}^n Y_i$ be the running average. Then Theorem 16.23 implies that

$$\mathbb{P}\left\{\left|\bar{X}_n - \mathbb{E}\, Y\right| \geq t \cdot \|Y - \mathbb{E}\, Y\|_\infty\right\} \leq 2\mathrm{e}^{-nt^2/2} \quad \text{for } t > 0.$$

Assuming that $Y \in \mathsf{L}_\infty$, we achieve concentration on the scale of $\|Y - \mathbb{E}\, Y\|_\infty$ with Gaussian tail decay. Notice that the number n of summands appears in the exponent! Compare this result with Example 16.3, where we achieved concentration on the scale of $\sqrt{\text{Var}[Y]}$ with the paltry tail decay $(nt)^{-2}$. ▪

In the next section, we will establish Bernstein's inequality, which yields better control on the variance with worse control on the tails.

Warning 16.26 (Independence). In theoretical probability, it is quite common to place independence assumptions on families of random variables. These assumptions lead to very powerful outcomes, such as Gaussian tail behavior for an independent sum. This is the conclusion of Hoeffding's inequality (Theorem 16.23). For random variables that are uncorrelated (but not necessarily independent), we can only achieve the weaker conclusions of Chebyshev's inequality (Proposition 15.2).

In applications, one must take great care to check that independence assumptions are warranted. For example, the financial crisis of 2008 occurred, in part, because lenders assumed that mortgage failures were independent events, whereas the failures were actually strongly correlated with each other.

Personally, I believe that independence is a very dangerous hypothesis. It is quite difficult to check independence empirically. Yet we can easily gull ourselves into a false sense of confidence about theoretical predictions made using independence. This complaint is defanged only in settings whether it is possible to engineer independence (e.g., in randomized study design or in computer algorithms). ▪

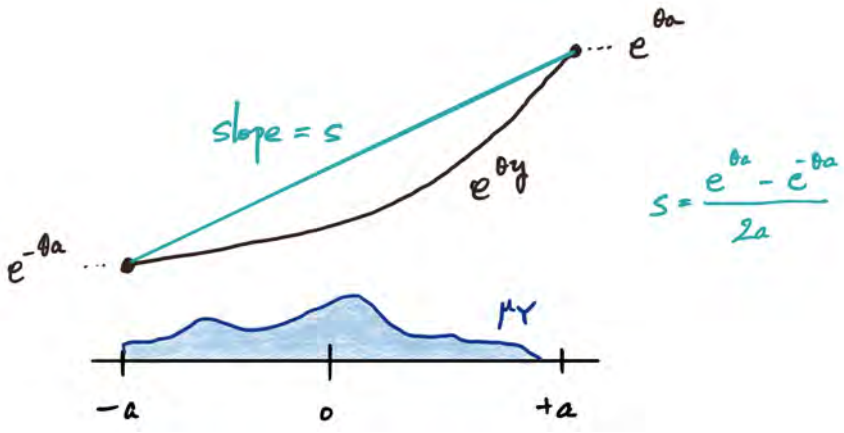
16.4.2 Hoeffding cgf bound

Aside from the Laplace transform method, the key technical ingredient in the proof of Hoeffding's theorem is an estimate for the cgf of a bounded random variable.

Lemma 16.27 (Hoeffding: Cgf bound). Let Y be a real random variable with $\mathbb{E}\, Y = 0$ and $|Y| \leq a$. Then $\xi_Y(\theta) \leq \theta^2 a^2/2$.

You can see a strong parallel between Lemma 16.27 and the cgf bound for a normal random variable (Exercise 16.10).

Proof. As with Markov's inequality, this result is best understood graphically. We bound the exponential function $y \mapsto \mathrm{e}^{\theta y}$ on the interval $[-a, +a]$ above by a straight line.



Let us use this idea to control the mgf:

$$\begin{aligned} m_Y(\theta) = \mathbb{E}\, \mathrm{e}^{\theta Y} &\le \mathbb{E}\left[\mathrm{e}^{-\theta a} + (Y + a) \cdot \frac{\mathrm{e}^{+\theta a} - \mathrm{e}^{-\theta a}}{2a}\right] \\ &= \mathrm{e}^{-\theta a} + (\mathbb{E}\, Y + a) \cdot \frac{\mathrm{e}^{+\theta a} - \mathrm{e}^{-\theta a}}{2a} \\ &= \mathrm{e}^{-\theta a} + \tfrac{1}{2}(\mathrm{e}^{+\theta a} - \mathrm{e}^{-\theta a}) = \cosh(\theta a). \end{aligned}$$

The first inequality is monotonicity of expectation. We have represented the line in point–slope form, using the left endpoint $-a$ of the interval. Then we applied linearity of expectation, along with the assumption that $\mathbb{E}\, Y = 0$. Finally, we recognize the hyperbolic cosine.

By comparing the Taylor series (exercise!), you may confirm that

$$m_Y(\theta) \le \cosh(\theta a) \le \exp(\theta^2 a^2/2).$$

Take the logarithm to complete the proof. ∎

Problem 16.28 **(*Asymmetric Hoeffding: Cgf bound).** The full statement of Hoeffding's inequality involves a more refined cgf bound. Suppose that $\mathbb{E}\, Y = 0$ and $a \le Y \le b$. Prove that $\xi_Y(\theta) \le (b-a)^2/8$. **Hint:** This computation requires some insight. You can realize the second derivative of the cgf as the variance of a bounded random variable and control the second derivative using an elementary bound for the variance.

16.4.3 Hoeffding inequality: Proof

We may now complete the proof of Hoeffding's inequality, Theorem 16.23.

Without loss of generality, assume that $\mathbb{E}\, Y_i = 0$ for each index i, so that $\mathbb{E}\, X = 0$. By the additivity of cgfs (Proposition 16.13) and the Hoeffding cgf bound (Lemma 16.27), we find that

$$\xi_X(\theta) = \sum\nolimits_{i=1}^{n} \xi_{Y_i}(\theta) \le \sum\nolimits_{i=1}^{n} \theta^2 a_i^2/2 = \theta^2 v/2.$$

For $t > 0$, we may apply the Laplace transform method (Theorem 16.16) to X to obtain

$$\mathbb{P}\{X \ge t\} \le \exp\left(-\sup\nolimits_{\theta>0} \left(\theta t - \theta^2 v/2\right)\right) = \mathrm{e}^{-\theta^2/(2v)}.$$

The supremum is attained when $\theta = t/v$. Similarly,

$$\mathbb{P}\{X \le -t\} \le \mathrm{e}^{-t^2/(2v)}.$$

Altogether, for $t > 0$,

$$\mathbb{P}\{|X| \geq t\} = \mathbb{P}\{X \geq t\} + \mathbb{P}\{X \leq -t\} \leq 2\mathrm{e}^{-t^2/(2v)}.$$

Make the change of variables $t \mapsto t\sqrt{v}$ to arrive at the stated result.

Exercise 16.29 (Asymmetric Hoeffding inequality). Formulate and prove a version of Hoeffding's inequality under the assumption that $a_i \leq Y_i \leq b_i$ for each index i.

16.5 Example: Bernstein's inequality

In this section, we present another powerful concentration inequality, due to Bernstein. This result gives an excellent tail bound for an independent sum of bounded random variables. Although there are many other concentration inequalities, this is the single most useful example.

16.5.1 The Bernstein inequality

We begin with the statement and some discussion. Then we turn to the proof.

Theorem 16.30 (Bernstein's inequality). Consider an *independent* family $(Y_1, \dots, Y_n)$ of real random variables, each subject to the uniform bound

$$|Y_i - \mathbb{E}\, Y_i| \leq B \quad \text{for each index } i = 1, \dots, n.$$

Form the sum $X = \sum_{i=1}^n Y_i$, and set $\sigma^2 = \mathrm{Var}[X] = \sum_{i=1}^n \mathrm{Var}[Y_i]$. Then

$$\mathbb{P}\{|X - \mathbb{E}\, X| \geq t\} \leq 2\exp\left(\frac{-t^2/2}{\sigma^2 + Bt}\right) \quad \text{for all } t \geq 0.$$

Exercise 16.31 (Bernstein: Centering). Without loss of generality, we can assume that $\mathbb{E}\, Y_i = 0$ for each $i = 1, \dots, n$ in the proof of Bernstein's inequality. Explain how this reduction works.

Like most concentration inequalities for independent sums, Bernstein's inequality exploits simple information about the summands (here, the mean, variance, and a uniform bound) to obtain incredibly strong information about the concentration of the sum around its mean. Figure 16.2 illustrates the bound from Theorem 16.30.

It is informative to rewrite the bound in Bernstein's inequality on the scale σ of the standard deviation:

$$\mathbb{P}\{|X - \mathbb{E}\, X| \geq t\sigma\} \leq 2\exp\left(\frac{-t^2/2}{1 + (B/\sigma)\cdot t}\right) \quad \text{for all } t \geq 0. \tag{16.3}$$

The expression (16.3) indicates that the tail decay can be separated into two regimes:

1. **Moderate deviations:** When the level $t \ll \sigma/B$, then the tail bound (16.3) is roughly equal to $\mathrm{e}^{-t^2/(2\sigma^2)}$. At this scale, the tail decay of the independent sum resembles that of a Gaussian random variable with the same variance.
2. **Large deviations:** When the level $t \gg \sigma/B$, then the tail bound (16.3) is roughly equal to $\mathrm{e}^{-t/(2B)}$. At this larger scale, the tail decay of the independent sum resembles that of an exponential random variable with mean $2B$.

The level $t = \sigma/B$ where we see the (soft) transition between the two regimes depends on the ratio between the standard deviation σ and the uniform bound B. Here is an

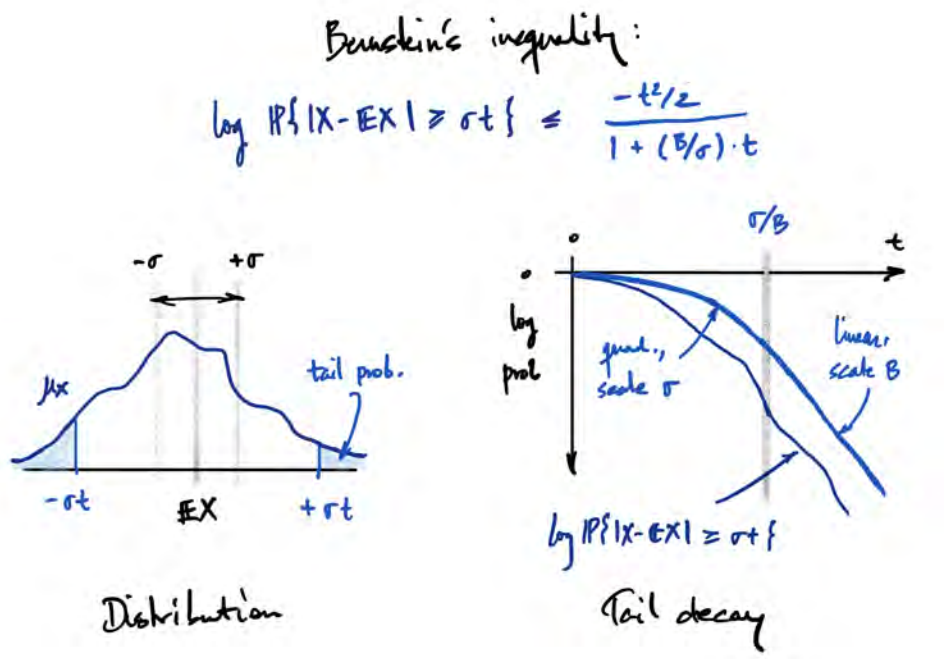


Figure 16.2 (Bernstein's inequality). This figure depicts Bernstein's inequality for an independent sum. For moderate deviations, it produces Gaussian tail decay on the scale of the standard deviation σ of the sum. For large deviations, it produces exponential tail decay on the scale of the upper bound B on the summands.

interpretation. An independent sum whose terms are all small relative to the variance behaves like a Gaussian random variable. On the other hand, when terms can be large relative to the variance, the sum behaves more like an exponential random variable. These are real phenomena that you can see in simulations.

16.5.2 The Bernstein cgf bound

Aside from the Laplace transform method, the key technical ingredient in the proof of Bernstein's inequality is an alternative estimate for the cgf of a bounded random variable.

Lemma 16.32 (Bernstein: Cgf bound). Let Y be a real random variable with $\mathbb{E}\, Y = 0$ and $|Y| \le B$. Then

$$\xi_Y(\theta) \le \frac{\theta^2/2}{1 - B|\theta|} \cdot \operatorname{Var}[Y] \quad \text{when } |\theta| < 1/B.$$

Observe the strong parallel between the Bernstein cgf bound and the cgf bound for a centered exponential random variable (Exercise 16.11). We can understand Bernstein's inequality as a comparison between an independent sum of bounded random variables and an independent sum of exponential random variables. Exercise 16.34 extends this approach to study an independent sum of random variables with exponential tail decay.

Proof. The idea behind this proof is to expand the exponential function in the mgf as a

Taylor series. Assume that $|\theta| < 1/B$. First,

$$m_Y(\theta) = \mathbb{E}[\mathrm{e}^{\theta Y}] = \mathbb{E}\left[1 + \theta Y + \sum\nolimits_{p=2}^{\infty} \frac{\theta^p}{p!} Y^p\right].$$

To continue, we make a simple pointwise bound on the sum:

$$\sum\nolimits_{p=2}^{\infty} \frac{\theta^p}{p!} Y^p \le \frac{\theta^2}{2} \sum\nolimits_{p=2}^{\infty} |\theta B|^{p-2} \cdot Y^2 = \frac{\theta^2/2}{1 - B|\theta|} \cdot Y^2.$$

We have identified a geometric series. Next, combine the last two displays. Using monotonicity and linearity of expectation,

$$m_Y(\theta) \le \mathbb{E}\left[1 + \theta Y + \frac{\theta^2/2}{1 - B|\theta|} \cdot Y^2\right] = 1 + \frac{\theta^2/2}{1 - B|\theta|} \cdot \mathrm{Var}[Y].$$

Take the logarithm to identify the cgf $\xi_Y(\theta)$. Finally, invoke the numerical inequality $\log(1 + a) \le a$, valid for $a > -1$. ∎

16.5.3 Bernstein's inequality: Proof

We may now complete the proof of Bernstein's inequality, Theorem 16.30.

Without loss of generality, assume that $\mathbb{E}\, Y_i = 0$ for each index i, so that $\mathbb{E}\, X = 0$. By the additivity of cgfs (Proposition 16.13) and the Bernstein cgf bound (Lemma 16.32), we find that

$$\xi_X(\theta) = \sum\nolimits_{i=1}^{n} \xi_{Y_i}(\theta) \le \frac{\theta^2/2}{1 - B|\theta|} \cdot \sum\nolimits_{i=1}^{n} \mathrm{Var}[Y_i] = \frac{\theta^2/2}{1 - B|\theta|} \cdot \sigma^2.$$

Now, apply the Laplace transform method (Theorem 16.16) to X to obtain

$$\mathbb{P}\{X \ge t\} \le \exp\left(-\sup_{0<\theta<1/B} \left(\theta t - \frac{\theta^2\sigma^2/2}{1 - B|\theta|}\right)\right) \quad \text{for } t > 0.$$

Although it is possible to evaluate the supremum exactly, we instead select a clever value of the parameter: $\theta = t/(\sigma^2 + B|t|) < 1/B$. After some algebra, it emerges that

$$\mathbb{P}\{X \ge t\} \le \exp\left(\frac{-t^2/2}{\sigma^2 + Bt}\right).$$

By essentially the same argument,

$$\mathbb{P}\{X \le -t\} \le \exp\left(\frac{-t^2/2}{\sigma^2 + Bt}\right).$$

Altogether, for $t > 0$,

$$\mathbb{P}\{|X| \ge t\} = \mathbb{P}\{X \ge t\} + \mathbb{P}\{X \le -t\} \le 2\exp\left(\frac{-t^2/2}{\sigma^2 + Bt}\right).$$

This is the required result.

16.5.4 *Refinements

There are several variants of Bernstein's inequality that follow from closely related arguments.

Exercise 16.33 (Bernstein cgf: Improvement). The bound in Lemma 16.32 may be refined to

$$\xi_Y(\theta) \le \frac{\theta^2/2}{1 - B|\theta|/3} \cdot \operatorname{Var}[Y] \quad \text{when } |\theta| < 3/B.$$

Verify this claim. What is the consequence for Bernstein's inequality?

Exercise 16.34 (Bernstein cgf: Exponential tails). Let Y be a real random variable with $\mathbb{E}\, Y = 0$ and exponential tail decay. For $\beta > 0$,

$$\mathbb{P}\{|Y| \ge t\} \le \mathrm{e}^{-t/\beta} \quad \text{for all } t \ge 0.$$

1. Using integration by parts, confirm that the polynomial moments satisfy
$$\mathbb{E}\,|Y|^p \le \beta^p p! \quad \text{for all } p \in \mathbb{N}.$$
2. Verify that the cgf satisfies
$$\xi_Y(\theta) \le \frac{\theta^2}{1 - \beta|\theta|} \quad \text{when } |\theta| < \beta^{-1}.$$
3. Formulate and prove an extension of Bernstein's inequality for an independent sum of random variables with uniform exponential tail decay.
4. Formulate and prove an extension of Bernstein's inequality for independent sum of random variables that satisfy the moment conditions in item (1).

Problems

Exercise 16.35 (Mean and median). From Exercise 8.31, recall that every real random variable has a median $M \in \mathbb{R}$, a number for which

$$\mathbb{P}\{X \ge M\} \ge \tfrac{1}{2} \quad \text{and} \quad \mathbb{P}\{X \le M\} \ge \tfrac{1}{2}.$$

For every square-integrable random variable, the mean and the median are close to each other.

1. Let $X \in \mathsf{L}_2$ be a real random variable. Show that
$$|M - \mathbb{E}\, X| \le \sqrt{2 \operatorname{Var}[X]}.$$
Hint: Use Chebyshev's inequality (Proposition 15.2).
2. **(*)** Remove the factor two from the right-hand side of the last result. **Hint:** Use the Chebyshev–Cantelli inequality (Exercise 12.23).

Problem 16.36 (Maximal inequalities). The mgf can be used to obtain elegant bounds for the expected maximum of a family of random variables with sufficiently fast tail decay. Surprisingly, these results do not require independence.

1. Consider a family $(X_1, \dots, X_n)$ of random variables. Show that
$$\mathbb{E} \max_{i=1,\dots,n} X_i \le \theta^{-1} \log\left(\sum\nolimits_{i=1}^{n} \mathbb{E}[\mathrm{e}^{\theta X_i}]\right) \quad \text{for } \theta > 0.$$
Hint: See Problem 16.7.

2. Consider the case where $Z_i \sim \text{NORMAL}(0, \sigma^2)$ for each $i = 1, \dots, n$. We do *not* assume that the family of random variables is independent. Show that
$$\mathbb{E} \max_{i=1,\dots,n} Z_i \le \sqrt{2\sigma^2 \log n}.$$
3. **(**)** Consider the case where $Z_i \sim \text{NORMAL}(0, \sigma^2)$ for each $i = 1, \dots, n$, and the family of random variables *is* independent. For sufficiently large n, show that
$$\mathbb{E} \max_{i=1,\dots,n} Z_i \ge \text{const} \cdot \sqrt{\sigma^2 \log n}.$$
Hint: By rescaling, assume that $\sigma^2 = 1$. Find a lower bound on the normal tail, say $\mathbb{P}\{Z_1 \ge t\} \ge \text{const} \cdot \mathrm{e}^{-t^2}$ for $t \ge 0$. You will also need Exercise 13.30.
4. **(**)** Consider the case where each $X_i \sim \text{EXPONENTIAL}(\beta)$, an exponential random variable with mean $\beta > 0$. Show that
$$\mathbb{E} \max_{i=1,\dots,n} (X_i - \beta) \le \beta\big(\log n + \sqrt{2 \log n}\big).$$

Problem 16.37 (Cumulants and independence).** In spite of the strange definition, cumulants are both beautiful and powerful. This problem explores some of the key properties of these objects that illuminates their significance. For a real random variable X, recall that the moments are defined by the series
$$\mathbb{E}[\mathrm{e}^{\theta X}] = \sum\nolimits_{p=0}^{\infty} \frac{\theta^p}{p!} \cdot \mathbb{E}[X^p] =: \sum\nolimits_{p=0}^{\infty} \frac{\theta^p}{p!} \cdot m_p(X).$$
The cumulants are defined by the series
$$\log \mathbb{E}[\mathrm{e}^{\theta X}] = \sum\nolimits_{p=1}^{\infty} \frac{\theta^p}{p!} \cdot \kappa_p(X).$$
These series may be interpreted formally, or you can require all random variables to be bounded to avoid convergence issues.

1. **Homogeneity:** Check that the cumulants are homogeneous:
$$\kappa_p(\alpha X) = \alpha^p \kappa_p(X) \quad \text{for } \textit{complex } \alpha \in \mathbb{C}.$$
2. **Independence:** If the pair (X, Y) is independent, show that
$$\kappa_p(X + Y) = \kappa_p(X) + \kappa_p(Y).$$
3. ***Cumulant–moment relation:** Show that the cumulants can be defined recursively in terms of the moments:
$$\kappa_p(X) = m_p(X) - \sum\nolimits_{i=1}^{p-1} \binom{p-1}{i-1} \cdot \kappa_i(X) m_{p-i}(X).$$
4. **Moment–cumulant relation:** Show that the moments can be represented in terms of the cumulants:
$$m_p(X) = \kappa_p(X) + \varphi_p(\kappa_1(X), \dots, \kappa_{p-1}(X))$$
where φ_p is a multivariate polynomial without a constant term. **(**)** Can you find an explicit form for the polynomial?

5. ***Good's formula:** Let $X^{(1)}, \dots, X^{(p)}$ be i.i.d. copies of X. Let ζ_p be a primitive pth root of unity. Define the random variable

$$X(\zeta_p) := \sum\nolimits_{i=1}^{p} \zeta_p^i X^{(i)}.$$

Verify that

$$\kappa_p(X) = \frac{1}{p}\,\mathbb{E}[X(\zeta_p)^p].$$

Hint: There is a short argument using all four of the previous statements. Recall that $\sum_{i=1}^{p} \zeta_p^m = 0$ when p does not divide m.

6. **Mixed cumulants:** Consider an arbitrary family $(X_1, \dots, X_p)$ of real random variables. Define the mixed cumulant

$$\kappa_p[X_1, \dots, X_p] := \frac{1}{p}\,\mathbb{E}[X_1(\zeta_p) \cdot X_2(\zeta_p) \cdots X_p(\zeta_p)].$$

Observe that the mixed cumulant is a multilinear function. This function can also be obtained by polarizing the p-homogeneous polynomial $\kappa_p(\cdot)$.

7. ***Mixed cumulants detect independence:** Suppose that there is a nontrivial, proper subset $\mathsf{S} \subset \{1, \dots, p\}$ where $(X_i : i \in \mathsf{S})$ and $(X_j : j \notin \mathsf{S})$ are independent. Prove that

$$\kappa_p[X_1, \dots, X_p] = 0.$$

8. **Mixed cumulants are additive:** Suppose that the family $(X_1, \dots, X_n)$ is independent from the family $(Y_1, \dots, Y_n)$. Show that

$$\kappa_p[X_1 + Y_1, \dots, X_p + Y_p] = \kappa_p[X_1, \dots, X_p] + \kappa_p[Y_1, \dots, Y_p].$$

Problem 16.38 (Chernoff inequalities). Chernoff's inequalities control the tails of an independent sum of bounded positive random variables. These results are commonly applied to study an independent sum of indicator random variables, which counts the total number of independent events that occur. When the events have different probabilities, we cannot use a binomial random variable as a model. Nevertheless, Chernoff's bounds show that these sums behave much like binomial random variables.

Consider an independent family $(Y_1, \dots, Y_n)$ of random variables that satisfy $0 \le Y_i \le 1$. In particular, indicator random variables fulfill the latter condition. Let $X = \sum_{i=1}^{n} Y_i$, and define $a = \mathbb{E}[X] = \sum_{i=1}^{n} \mathbb{E}\, Y_i$.

1. Establish the Chernoff mgf bound:

$$\log \mathbb{E}\, \mathrm{e}^{\theta Y_i} \le (\mathrm{e}^{\theta} - 1)(\mathbb{E}\, Y_i) \quad \text{for all } \theta \in \mathbb{R}.$$

Hint: Bound the exponential function above by its secant on $[0, 1]$.

2. Prove the upper Chernoff inequality via the Laplace transform method:

$$\mathbb{P}\left\{\sum\nolimits_{i=1}^{n} X_i \ge (1+t)a\right\} \le \left(\frac{\mathrm{e}^t}{(1+t)^{1+t}}\right)^a \le \left(\frac{\mathrm{e}}{1+t}\right)^{(1+t)a} \quad \text{for } t > 0.$$

The last inequality follows from a simple bound.

3. Apply the same ideas to obtain the lower Chernoff inequality:

$$\mathbb{P}\left\{\sum\nolimits_{i=1}^{n} X_i \le (1-t)a\right\} \le \left(\frac{\mathrm{e}^{-t}}{(1-t)^{1-t}}\right)^a \le \mathrm{e}^{-t^2 a/2} \quad \text{for } t \in [0, 1].$$

The last inequality follows from the more accurate tail bound by taking a second-order Taylor expansion with exact remainder.

4. Compare these results with Example 16.20.
5. **Balls & bins:** Suppose that we toss n balls independently at random into m bins, where each ball lands in a uniformly random bin. What is the expected number of balls in the first bin? Bound the probability that there are at least twice as many balls as expected. Bound the probability that there are fewer than half as many balls as expected. How do these results depend on m and n?
6. **Amplification:** Some decision problems (e.g., "Is a given integer composite?") admit efficient randomized algorithms. Imagine that a randomized algorithm returns the correct answer with probability p, where $p > 1/2$. We can run the algorithm repeatedly and take a majority vote to enhance the success probability. To obtain a failure probability below ε, how many times should we run the algorithm?

Problem 16.39 (*Gaussian chaos). Let $\boldsymbol{A} \in \mathbb{R}^{n\times n}$ be a symmetric matrix. Draw a vector $\boldsymbol{Z} = (Z_1, \dots, Z_n)$ whose entries are i.i.d. standard normal random variables. Define the random variable

$$X := \boldsymbol{Z}^{\mathsf{T}}\boldsymbol{A}\boldsymbol{Z} = \sum\nolimits_{i,j=1}^{n} a_{ij} Z_i Z_j.$$

This type of random variable is called a *second-order Gaussian chaos*. We will develop a concentration inequality using special properties of the Gaussian distribution.

1. Show that the standard normal vector $\boldsymbol{Z} = (Z_1, \dots, Z_n)$ is rotationally invariant. That is, $\boldsymbol{U}\boldsymbol{Z} \sim \boldsymbol{Z}$ for each fixed orthogonal matrix $\boldsymbol{U} \in \mathbb{R}^{n\times n}$. **Hint:** Use the multivariate change of variables formula (Problem 6.28).
2. The symmetric matrix $\boldsymbol{A}$ has an eigenvalue decomposition: $\boldsymbol{A} = \boldsymbol{U}\boldsymbol{D}\boldsymbol{U}^{\mathsf{T}}$ where $\boldsymbol{D} = \operatorname{diag}(d_1, \dots, d_n)$ and $\boldsymbol{U} \in \mathbb{R}^{d\times d}$ is orthogonal. Verify that the squared Frobenius norm $\|\boldsymbol{A}\|_{\mathrm{F}}^2 = \sum_{i=1}^n d_i^2$ and the ℓ_2 operator norm $\|\boldsymbol{A}\| = \max_i |d_i|$.
3. Deduce that the chaos has the same distribution as a simpler random variable:

$$X \sim \sum\nolimits_{i=1}^{n} d_i Z_i^2.$$

   In particular, $X - \mathbb{E}\, X \sim \sum_{i=1}^n d_i (Z_i^2 - 1)$
4. Show that the cgf of the centered random variable $Y = Z_1^2 - 1$ satisfies

$$\xi_Y(\theta) = -\tfrac{1}{2}\left[\log(1-2\theta) + 2\theta\right] \le \frac{\theta^2}{1-2\theta} \quad \text{when } 2\theta < 1.$$

5. Derive that

$$\xi_{X-\mathbb{E}X}(\theta) \le \frac{\theta^2 \|\boldsymbol{A}\|_{\mathrm{F}}^2}{1 - 2\theta\|\boldsymbol{A}\|} \quad \text{when } 2\theta\|\boldsymbol{A}\| < 1.$$

6. For all $t > 0$, establish the Bernstein-type concentration inequality

$$\mathbb{P}\left\{|X - \mathbb{E}\, X| \ge t\right\} \le \exp\left(\frac{-t^2/4}{\|\boldsymbol{A}\|_{\mathrm{F}}^2 + \|\boldsymbol{A}\| \cdot t}\right).$$

Problem 16.40 (Bennett's inequality). For an independent sum of random variables that admit one-sided bounds, we can obtain a refinement of Bernstein's inequality that yields slightly better tail behavior.

Consider an independent family $(Y_1, \dots, Y_n)$ of independent random variables, subject to $\mathbb{E}[Y_i] = 0$ and the one-sided bound $Y_i \le B$ for each index i. Let $X = \sum_{i=1}^n Y_i$, and define $\sigma^2 = \operatorname{Var}[X]$.

1. Establish the Bennett cgf inequality. For each index i,

$$\xi_{Y_i}(\theta) \le \frac{e^{\theta B} - \theta B - 1}{B^2} \cdot \mathrm{Var}[Y] \quad \text{for all } \theta \in \mathbb{R}.$$

   Hint: The function $x \mapsto (e^x - x - 1)/x^2$ is increasing.
2. Derive Bennett's inequality. For all $t \ge 0$,

$$\mathbb{P}\{X \ge t\} \le \exp\left(-\frac{\sigma^2}{B^2} h\left(\frac{Bt}{\sigma^2}\right)\right),$$

   where $h(u) := (1+u)\log(1+u) - u$ for all $u > -1$.
3. Compare and contrast Bennett's inequality with the upper Chernoff inequality (Problem 16.38).

Applications

Application 16.41 (Trace estimation). Let $\boldsymbol{A} \in \mathbb{R}^{d\times d}$ be a nonzero, (symmetric) positive-semidefinite (psd) matrix. Suppose that we only have access to the matrix via matrix–vector products: $\boldsymbol{u} \mapsto \boldsymbol{A}\boldsymbol{u}$. This situation can arise in statistics or in numerical analysis. Our goal is to estimate the trace, $\mathrm{tr}(\boldsymbol{A})$, using a small number of matrix–vector products with *random* vectors. The original application of this method was for cross-validation with smoothing splines (Lecture 0).

Draw a random vector $\boldsymbol{\epsilon} = (\varepsilon_1, \dots, \varepsilon_d)$ whose entries are i.i.d. UNIFORM$\{\pm 1\}$ random variables. Introduce the random variable

A UNIFORM$\{\pm 1\}$ random variable is often called a *Rademacher* random variable.

$$Y = \boldsymbol{\epsilon}^*(\boldsymbol{A}\boldsymbol{\epsilon}) = \sum\nolimits_{i,j=1}^{d} a_{ij}\varepsilon_i\varepsilon_j.$$

The *Hutchinson trace estimator* with n samples is the random variable

$$\bar{X}_n = \frac{1}{n}\sum\nolimits_{k=1}^{n} Y_k \quad \text{where } Y_k \sim Y \text{ i.i.d.}$$

1. Confirm that $\mathbb{E}\,Y = \mathrm{tr}(\boldsymbol{A}) > 0$. Show that

$$\mathrm{Var}[Y] = 2\sum\nolimits_{i\ne j} a_{ij}^2 < 2\,\mathrm{tr}(\boldsymbol{A})^2.$$

   The second inequality requires the assumption that $\boldsymbol{A}$ is psd.
2. Calculate $\mathbb{E}\,\bar{X}_n$ and $\mathrm{Var}[\bar{X}_n]$.
3. Using only the results from (a) and (b), bound the probability that the relative error is large:

$$\mathbb{P}\left\{|\bar{X}_n - \mathrm{tr}(\boldsymbol{A})|/\mathrm{tr}(\boldsymbol{A}) \ge \varepsilon\right\} \quad \text{for } \varepsilon \in [0,1].$$

4. How many samples n suffice so the relative error is less than ε with probability at least $1-\delta$? For concreteness, instantiate your bound for $\varepsilon = 10\%$ and $\delta = 10\%$.
5. **(*)** Implement the Hutchinson trace estimator. Consider (large) psd matrices $\boldsymbol{A}$ with various distributions of eigenvalues (say, flat, power-law decay, exponential decay). For each instance and for a range of values of n, run the trace estimator 100 times. For each instance, plot the average relative error with error bars as a function of n. Are the theoretical bounds accurate?
6. **(**)** Implement the smoothing splines described in Section 0.3. Use the randomized trace estimator to approximate the generalized cross-validation (gcv) functional. Fit some smoothing splines to some data, using the gcv functional to select the proper amount of smoothing.

7. **(**)** Develop a sharper error bound for the Hutchinson trace estimator via the Laplace transform method. This problem is much easier if you use Gaussian random variables instead of Rademacher variables in the trace estimator; see Problem 16.39 for the analysis. To deal with Rademachers, you will need the Hanson–Wright inequality (or elements from its proof).

Application 16.42 (The sample average estimator: Confidence intervals). In statistics, a basic task is to infer the mean of an observable from an (independent) sample. We can model this situation using an independent sum. Let Y be a real random variable that describes the distribution of an observable of the population (say, height or income), so $\mathbb{E}[Y]$ is the population mean of the observable. Draw an iid family $(Y_1, \dots, Y_n)$ of copies of Y, called a *sample*, and form the sample average $\bar{X} := n^{-1} \sum_{i=1}^{n} Y_i$. We regard $\bar{X}$ as an estimator for the population mean $\mathbb{E}[Y]$. For this problem, we will assume that $|Y - \mathbb{E}\, Y| \le B$. (Is this assumption reasonable? Is the independence assumption reasonable?)

1. We may want to ensure the sample average $\bar{X}$ is an accurate estimate for the mean $\mathbb{E}[Y]$:

$$\mathbb{P}\left\{|\bar{X} - \mathbb{E}\, Y| \ge \operatorname{stdev}(Y) \cdot \varepsilon\right\} \le \delta \quad \text{where } \varepsilon, \delta \in (0, 1).$$

   Using Bernstein's inequality (Theorem 16.30), check that it enough to take n samples where

$$n \ge \max\left\{4\varepsilon^{-2} \log(2/\delta),\ 4\varepsilon^{-1}(B/\operatorname{stdev}(Y)) \log(2/\delta)\right\}.$$

   When is the first term larger? The second? What are the implications if we try to make ε small? Look at a few particular values, say, $\varepsilon = 0.1, 0.01, 0.001$. Is it hard to achieve a small failure probability δ?
2. For $B = 1$ and large n, instantiate Bernstein's inequality to control

$$\mathbb{P}\left\{|\bar{X} - \mathbb{E}\, Y| \ge \operatorname{stdev}(Y) \cdot n^{-1/2} \cdot t\right\} \le \dots$$

   Give rough numerical bounds for the tail when $t = 1, 2, 3$.
3. **(*)** A *confidence interval* for the mean $\mathbb{E}[Y]$ is a (random) interval $[a, b]$, depending on observed data, that contains the mean $\mathbb{E}\, Y$ with probability $1 - \delta$. The probability is with respect to the randomness in the sample. Explain how to interpret this calculation as providing a confidence interval for the mean.

Application 16.43 (Median-of-means estimator). For natural numbers $k, n \in \mathbb{N}$, consider an independent family $(Y_1, \dots, Y_{kn})$ of copies of a square-integrable real random variable $Y \in \mathsf{L}_2$. From Example 16.3, we know that the sample average $\bar{X}_{kn}$ satisfies

$$\mathbb{P}\left\{\left|\bar{X}_{kn} - \mathbb{E}\, Y\right| > t\sqrt{\operatorname{Var}[Y]}\right\} \le \frac{1}{knt^2} \quad \text{for } t > 0.$$

In contrast, Example 16.25 yields much sharper concentration $\mathrm{e}^{-knt^2/2}$, provided that Y is bounded: $Y \in \mathsf{L}_\infty$. Is there a way to get the best of both worlds? That is, can we estimate the mean under weak integrability assumptions and still obtain sharp concentration?

One approach is called the *median-of-means* estimator. Maintain the square-integrability assumption: $Y \in \mathsf{L}_2$. We form k sample averages:

$$\bar{X}_n^{(i)} = \frac{1}{n} \sum\nolimits_{j=(i-1)n+1}^{in} Y_j \quad \text{for } i = 1, \dots, k.$$

Then, we find a median $M \in \mathbb{R}$ of the family of sample averages:

$$\#\{i : \bar{X}_n^{(i)} \le M\} \ge k/2 \quad \text{and} \quad \#\{i : \bar{X}_n^{(i)} \ge M\} \ge k/2.$$

We will argue that the median-of-means is a reliable estimator of the central tendency with extremely high probability.

1. Use Chebyshev's inequality (Proposition 15.2) to establish a concentration inequality for each $\bar{X}_n^{(i)}$. That is, provide a bound for the probability
$$\mathbb{P}\left\{\left|\bar{X}_n^{(i)} - \mathbb{E}\, Y\right| > t\right\}.$$
2. Observe that the family $\left(\bar{X}_n^{(i)} : i = 1, \dots, k\right)$ is statistically independent.
3. Use Chernoff's inequality (Problem 16.38) to bound the probability that
$$\mathbb{P}\left\{|M - \mathbb{E}\, Y| > t\right\}.$$
Hint: For each i, introduce the indicator of the event $\{|\bar{X}_n^{(i)} - \mathbb{E}\, Y| > t\}$. Use Chernoff's inequality to control the probability that more than $1/4$ of these events occur.
4. Compare the result with Examples 16.3 and 16.25.
5. **(*)** If we wish to achieve approximation error $\varepsilon\sqrt{\operatorname{Var}[Y]}$ with probability $1 - \delta$, how should we optimize (k, n) to minimize the total number of samples required?
6. **(*)** Show how to improve the analysis using the one-sided tail bound from Cantelli's inequality (Exercise 12.23).

Application 16.44 (*Johnson & Lindenstrauss). Randomized dimension reduction is a popular method in computational data science. Let $(\boldsymbol{a}_i : i = 1, \dots, N)$ be a family of distinct vectors in $\mathbb{R}^d$. Draw a matrix $\boldsymbol{G} \in \mathbb{R}^{m \times d}$ whose entries are i.i.d. $\textsc{normal}(0, m^{-1})$, and form the vectors $\boldsymbol{v}_i = \boldsymbol{G}\boldsymbol{a}_i \in \mathbb{R}^m$ for $i = 1, \dots, N$. We want to choose the embedding dimension m so that, with high probability,

$$\frac{1}{2}\|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2^2 \le \|\boldsymbol{v}_i - \boldsymbol{v}_j\|_2^2 \le \frac{3}{2}\|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2^2 \quad \text{for all } i, j. \tag{16.4}$$

That is, the low-dimensional vectors $(\boldsymbol{v}_i)$ have geometry comparable with the $(\boldsymbol{a}_i)$.

1. For a unit vector $\boldsymbol{u} \in \mathbb{R}^d$ and a standard normal random vector $\boldsymbol{g} \in \mathbb{R}^d$, explain why $|\langle \boldsymbol{g}, \boldsymbol{u} \rangle|^2 \sim X_1$. **Hint:** See Lecture 21.
2. Let $X_1 = g^2$ where $g \sim \textsc{normal}(0, 1)$. Show that the mgf takes the form $m_{X_1}(\theta) = (1 - 2\theta)^{-1/2}$ for $\theta < 1/2$. **Hint:** Unconscious statistician.
3. What is the mgf of $X_m = m^{-1} \sum_{i=1}^m g_i^2$ where the family (g_i) consists of i.i.d. standard normal random variables?
4. Deduce that
$$\mathbb{P}\{X_m - \mathbb{E}\, X_m \ge +t\} \le [(1 + t)\, \mathrm{e}^{-t}]^{m/2} \quad \text{for } t \ge 0;$$
$$\mathbb{P}\{X_m - \mathbb{E}\, X_m \le -t\} \le [(1 - t)\, \mathrm{e}^{+t}]^{m/2} \quad \text{for } t \in [0, 1].$$
(*) Combine and simplify: $\mathbb{P}\{|X_m - \mathbb{E}\, X_m| \ge t\} \le 2\exp(-mt^2/(4 + 4t))$ for $t \ge 0$.
5. Let $\boldsymbol{G} \in \mathbb{R}^{m \times d}$ be the random matrix defined as above. Conclude that
$$\mathbb{P}\left\{\left|\|\boldsymbol{G}\boldsymbol{u}\|_2^2 - 1\right| > t\right\} \le 2\exp\left(\frac{-mt^2/4}{1 + t}\right) \quad \text{for } t > 0.$$

6. Find a lower bound on the probability that $|\|\boldsymbol{G}\boldsymbol{u}_{ij}\|_2^2 - 1| \le t$ holds simultaneously for all the unit vectors $\boldsymbol{u}_{ij} = (\boldsymbol{a}_i - \boldsymbol{a}_j)/\|\boldsymbol{a}_i - \boldsymbol{a}_j\|_2$ with $i < j$.
7. To obtain failure probability $\delta \in (0, 1)$ in (16.4), how large should m be? Discuss.

Application 16.45 (*Random graphs). Recall that an *Erdős–Rényi graph* $\mathsf{G}(n, p)$ is an undirected, combinatorial graph (V, E) drawn at random from the following distribution. The vertex set V contains n vertices. For each choice $\{u, v\}$ of distinct vertices, the edge $e = \{u, v\}$ appears *independently* with probability p. The *degree* of a vertex v is the number of edges that are incident on the vertex:

$$\deg(v) := \sum\nolimits_{e \in \mathsf{E}} \mathbb{1}_{v \in e}.$$

We will explore the degrees of the vertices in a random graph with a large number n of vertices. We scale the probability p along with the number of vertices to preserve the expected degree.

1. For each vertex $v \in \mathsf{V}$, show that the expected degree $d := \mathbb{E}[\deg(v)] = (n-1)p$.
2. **Dense graphs, regularity:** Suppose that the expected degree $d \ge \text{Const} \cdot \log n$ for a sufficiently large constant. Show that

$$\mathbb{P}\{|\deg(v) - d| \le 0.1d \text{ for all } v \in \mathsf{V}\} \ge 0.9.$$

   In other words, every vertex in a dense random graph has about the same degree. The graph is *almost regular*. **Hint:** Use the Chernoff inequalities (Problem 16.38) along with a union bound.
3. ***Sparse graphs, irregularity:** Suppose that the expected degree $d = \text{Const} \cdot \log n$. Show that the graph has a vertex with degree no greater than d. Show that it is likely that the graph has a vertex with degree at least $10d$.
4. ****Very sparse graphs, irregularity:** Suppose that the expected degree d is a constant. Show that there is a vertex whose degree is constant. Show that it is likely that there is a vertex with degree at least $\text{const} \cdot (\log n)/(\log\log n)$.

Notes

Concentration inequalities for independent sums are the most basic example from a wide-ranging and powerful theory. Many other types of random variables satisfy strong concentration results. Michel Talagrand [Tal96] summarizes the basic principle:

> "A random variable that depends (in a 'smooth' way) on the influence of many independent random variables (but not too much on any of them) is essentially constant."

Concentration inequalities play a central role in modern statistics, mathematics of data science, theoretical algorithms, and other areas. You can learn more about this subject from [BLM13; VH16; Ver18; Tro21]. Some of the problems are drawn from these sources. For a discussion of large-deviation principles, see [DZ98].

Cumulants are very elegant, but they are not often covered in introductory classes. See the article of Speed [Spe83] for a brief introduction to the subject and some older references. A related class of cumulants plays a core role in the study of random matrices [Leh04; NS06].

Lecture bibliography

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. A nonasymptotic theory of independence, With a foreword by Michel Ledoux. Oxford University Press, Oxford, 2013. DOI: 10.1093/acprof:oso/9780199535255.001.0001.

[DZ98] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*. Second. Springer-Verlag, New York, 1998. DOI: 10.1007/978-1-4612-5320-4.

[Leho4] F. Lehner. "Cumulants in noncommutative probability theory. I. Noncommutative exchangeability systems". In: *Math. Z.* 248.1 (2004), pages 67–100. DOI: 10.1007/s00209-004-0653-0.

[NS06] A. Nica and R. Speicher. *Lectures on the combinatorics of free probability*. Cambridge University Press, Cambridge, 2006. DOI: 10.1017/CBO9780511735127.

[Spe83] T. P. Speed. "Cumulants and partition lattices". In: *Austral. J. Statist.* 25.2 (1983), pages 378–388.

[Tal96] M. Talagrand. "A new look at independence". In: *Ann. Probab.* 24.1 (1996), pages 1–34. DOI: 10.1214/aop/1042644705.

[Tro21] J. A. Tropp. *ACM 217: Probability in High Dimensions*. CMS Lecture Notes 2021-01. Caltech, 2021. DOI: 10.7907/mxr0-c422.

[VH16] R. Van Handel. "Probability in high dimension". APC 550 Lecture Notes, Princeton University. 2016. URL: https://web.math.princeton.edu/~rvan/APC550.pdf.

[Ver18] R. Vershynin. *High-dimensional probability*. An introduction with applications in data science, With a foreword by Sara van de Geer. Cambridge University Press, Cambridge, 2018. DOI: 10.1017/9781108231596.

17. Weak Convergence

Agenda:
1. Modes of convergence
2. Bounded Lipschitz functions
3. The BL distance
4. Weak convergence
5. Convergence of dfs
6. Integral probability metrics

"To rush into explanations is always a sign of weakness."

—*The Seven Dials Mystery*, Agatha Christie

Limit theorems in probability tell us when a sequence of random variables converges to a limiting random variable. There are many flavors of convergence, each suited for particular circumstances. So far, we have focused on notions of convergence that treat the random variables as *functions* on the sample space. In this lecture, we consider yet another approach that describes when the *distributions* of the random variables converge to a limiting distribution.

This change in perspective requires us to think about what it means for two probability distributions to be close to each other. Here is the key idea. When two distributions are similar, then many *moments* of the distributions are also similar. Conversely, if two distributions are very different, there is a moment that witnesses the discrepancy between them.

Using this approach, we introduce the *bounded Lipschitz* (BL) distance, a metric defined on the space of probability measures on the real line. The metric geometry leads to a notion of convergence for probability measures, called *weak convergence*. Weak convergence has a number of alternative formulations in terms of random variables, distribution functions, and characteristic functions. These results make weak convergence a core tool in probability theory.

We can generalize the bounded Lipschitz distance in several ways. First, it extends naturally to probability measures defined on $\mathbb{R}^n$. Second, we can introduce a whole family of distances, called *integral probability metrics*, that provide a useful framework for thinking about similarity of probability distributions and what it means for probability distributions to converge.

17.1 Modes of convergence

Suppose that $(X_i : i \in \mathbb{N})$ is a family of real random variables, defined on a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. What does it mean for the sequence of random variables to converge to a limiting random variable X, defined on the same sample space?

17.1.1 Convergence of functions

Recall that a real random variable $X : \Omega \to \mathbb{R}$ is a real-valued *function* on the sample space. This definition means that we can apply classic notions of convergence, defined for functions, to sequences of random variables. The modes of convergence we have encountered so far are all based on this idea.

Convergence pointwise, almost-sure, in probability

The simplest notion is pointwise convergence:

$$X_n \to X \text{ pointwise} \quad \text{when} \quad X_n(\omega) \to X(\omega) \text{ for all } \omega \in \Omega.$$

Almost-sure convergence is a relaxation of pointwise convergence:

$$X_n \to X \text{ a.s.} \quad \text{when} \quad \mathbb{P}\{\omega \in \Omega : X_n(\omega) \to X(\omega)\} = 1.$$

It is easy to see that pointwise convergence implies almost-sure convergence, but the converse does not hold.

Convergence in probability is even weaker:

$$X_n \to X \text{ in probability} \quad \text{when} \quad \sup_{t>0} \lim_{n\to\infty} \mathbb{P}\{\omega \in \Omega : |X_n(\omega) - X(\omega)| \geq t\} = 0.$$

Almost-sure convergence implies convergence in probability, but the converse is false.

Convergence in L_p

There is a separate notion of convergence defined for random variables in an L_p space ($p > 0$). If the members X_i of the sequence and the limit X all belong to L_p, we say that

$$X_n \to X \text{ in } \mathsf{L}_p \quad \text{when} \quad \int_\Omega |X_n(\omega) - X(\omega)|^p \, \mathbb{P}(\mathrm{d}\omega) \to 0.$$

This notion of convergence averages the distance between the two functions over the whole sample space, placing more emphasis on large discrepancies when p is large.

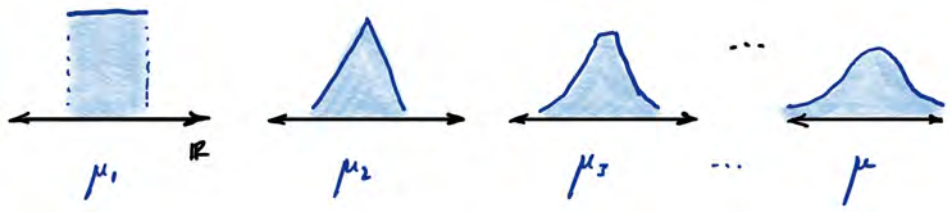
We can write the latter expression more compactly using homogeneous pth moments: $\|X_n - X\|_p \to 0$. By monotonicity (Theorem 11.4), convergence in L_p implies convergence in L_q for all $q \leq p$. On the other hand, convergence in L_p is incomparable with the concepts described in the last paragraph.

Observe that L_p convergence is determined by the pseudometric $\mathrm{dist}_p(X, Y) := \|X - Y\|_p$. In other words, the distance given by the L_p pseudonorm *metrizes* convergence in L_p.

17.1.2 Convergence of distributions

Recall that a real random variable Y induces a Borel probability measure μ_Y on the real line, called the distribution or the law of the random variable. The measure μ_Y depends on the probability space $(\Omega, \mathcal{F}, \mathbb{P})$, as well as the function Y. Nevertheless, once we have determined μ_Y, the probability space no longer plays a role. We can just think about μ_Y as a measure defined on the Borel sets of the real line.

As a consequence, it is natural to introduce the sequence $(\mu_i : i \in \mathbb{N})$ of distributions of the random variables X_i and the distribution μ of the random variable X. We can ask what it would mean for the sequence (μ_i) of distributions to converge to the limiting probability measure μ.



Note that this approach to convergence only depends on the *marginal* distributions μ_i of the random variables X_i. The interactions among the random variables do not play any role once we pass to the distributions. As a consequence, we anticipate that it is easier for the sequence of measures to converge than for the random variables themselves to converge.

17.1.3 Moments and similarity of measures

To develop a notion of convergence for distributions, we first discuss what it means for two distributions to be similar.

Recall that a *moment* of a probability measure μ on the real line is a linear functional of the measure. Each moment takes the form $\mu(h) = \int h \, d\mu$ for a measurable function $h : \mathbb{R} \to \mathbb{R}$.

We have seen that moments provide information about the distribution. For example, when the moment $\mu(|x|^p)$ is finite, then the tails of the distribution decay at least as fast as t^{-p}. In Problem 10.23, we developed a way to reconstruct a probability distribution that is supported on $\{0, 1, 2, \dots, n\}$ from a collection of n moments.

These observations motivate the idea that we can use moments to test whether two probability measures μ and ν are similar to each other. In other words, when $\mu \approx \nu$, we anticipate that the moments $\mu(h) \approx \nu(h)$ for many test functions h. The more moments that are close, the more evidence that the distributions are close.

Conversely, if two probability measures μ and ν are very dissimilar, we can search for a test function h for which the associated moments $\mu(h)$ and $\nu(h)$ are very different; this moment witnesses the discrepancy between the measures.

In this lecture, we implement techniques that use moments to measure the distance between distributions. We will begin with the most important special case, which leads to the concept of weak convergence of distributions. Afterward, we will generalize this construction and discuss several additional examples.

17.2 The bounded Lipschitz distance

In this section, we introduce a particular distance between probability distributions. To do so, we must describe a class of test functions that we will use to evaluate the similarity between distributions. Our choice of test functions may not seem obvious, but it is well motivated by the applications in probability theory.

17.2.1 Bounded, Lipschitz functions

First, we present some classical ways to measure regularity of a function.

Definition 17.1 (Bounded function). For a function $h : \mathbb{R} \to \mathbb{R}$, the *supremum norm* is

$$\|h\|_{\sup} := \sup_{a \in \mathbb{R}} |h(a)|.$$

If $\|h\|_{\sup} < +\infty$, then we say that h is *bounded*.

Exercise 17.2 (Supremum norm). Show that $\|\cdot\|_{\sup}$ is a norm on functions $h : \mathbb{R} \to \mathbb{R}$.

The supremum norm is always larger than the $\mathsf{L}_\infty(\mu)$ norm, which ignores the values of the function on a μ-negligible set.

Definition 17.3 (Lipschitz function). For a function $h : \mathbb{R} \to \mathbb{R}$, the *Lipschitz*

pseudonorm is

$$\|h\|_{\text{Lip}} := \inf\{L > 0 : |h(b) - h(a)| \leq L \cdot |b - a| \text{ for all } a, b \in \mathbb{R}\}.$$

If $\|h\|_{\text{Lip}} \leq L$ for a finite value L, then we say that h is *L-Lipschitz* or just *Lipschitz*. The number L is referred to as the *Lipschitz constant* of the function h.

In other words, an L-Lipschitz function changes by at most L units over each unit interval. You should confirm the following properties of Lipschitz functions.

Exercise 17.4 (Lipschitz constant). Show that $\|\cdot\|_{\text{Lip}}$ is a pseudonorm on functions $h : \mathbb{R} \to \mathbb{R}$.

Exercise 17.5 (Lipschitz function: Continuity). Show that every Lipschitz function is continuous. By example, show that there are continuous functions that are not Lipschitz.

Exercise 17.6 (Lipschitz function: Derivatives). Suppose that $h : \mathbb{R} \to \mathbb{R}$ is differentiable. Show that $\|h\|_{\text{Lip}} = \|h'\|_{\text{sup}}$. In particular, every differentiable function is Lipschitz on compact sets. By example, show that there are Lipschitz functions that are not differentiable.

A function can have any combination of boundedness and Lipschitz properties, or lack thereof. It is also convenient to introduce another norm that encapsulates both properties.

Definition 17.7 (Bounded, Lipschitz norm). For a function $h : \mathbb{R} \to \mathbb{R}$, the *bounded, Lipschitz (BL) norm* is

$$\|h\|_{\text{BL}} := \max\{\|h\|_{\text{sup}}, \ \|h\|_{\text{Lip}}\}.$$

If $\|h\|_{\text{BL}} < +\infty$, then h is both bounded and Lipschitz.

17.2.2 Bounded, Lipschitz functions separate probability measures

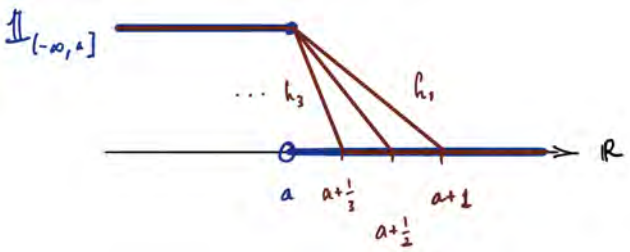
We can determine whether two probability measures are equal by checking whether the moments induced by bounded, Lipschitz functions are equal.

Proposition 17.8 (BL functions separate measures). Two Borel probability measures μ, ν on the real line are equal ($\mu = \nu$) if and only if $\mu(h) = \nu(h)$ for all bounded, Lipschitz functions $h : \mathbb{R} \to \mathbb{R}$.

Proof. The forward direction is easy. Let $h : \mathbb{R} \to \mathbb{R}$ be bounded and Lipschitz. A Lipschitz function is continuous, so it is also measurable. Each bounded, measurable function is integrable with respect to any probability measure. Therefore, if $\mu = \nu$, then the integrals $\mu(h) = \nu(h)$ are defined and must be equal for every function h that is bounded and Lipschitz.

For the reverse direction, assume that $\mu \neq \nu$. Since Borel probability measures are determined uniquely by their (cumulative) distribution functions (Theorem 3.26), we may just as well assume that the distribution functions $F_\mu(a) := \mu(-\infty, a]$ and $F_\nu(a) := \nu(-\infty, a]$ are not equal for some particular $a \in \mathbb{R}$.

We can approximate the indicator function $(-\infty, a]$ by a sequence of bounded, Lipschitz functions. Consider the piecewise linear function $h_n : \mathbb{R} \to \mathbb{R}$ with $h_n(x) = 1$ for $x \leq a$ and $h_n(x) = 0$ for $x \geq a + 1/n$.



By construction, $h_n \to \mathbb{1}_{(-\infty,a]}$ pointwise as $n \to \infty$. The bounded convergence theorem (Corollary 9.13) readily implies that

$$\mu(h_n) \to \mu(\mathbb{1}_{(-\infty,a]}) = F_\mu(a) \quad \text{and} \quad \nu(h_n) \to \nu(\mathbb{1}_{(-\infty,a]}) = F_\nu(a).$$

Since $F_\mu(a) \neq F_\nu(a)$, there must be some index $n \in \mathbb{N}$ for which $\mu(h_n) \neq \nu(h_n)$. In other words, the bounded, Lipschitz function h_n witnesses the fact that the two measures μ and ν are not equal. ∎

17.2.3 The bounded Lipschitz distance

Proposition 17.8 states that bounded, Lipschitz functions allow us to separate probability measures. Therefore, we can use them to define a metric.

Definition 17.9 (BL metric: Probability measures). Let μ, ν be Borel probability measures on the real line. Define the bounded Lipschitz (BL) probability metric:

$$\operatorname{dist}_{\mathrm{BL}}(\mu, \nu) := \sup \{|\mu(h) - \nu(h)| : \|h\|_{\mathrm{BL}} \leq 1\}.$$

We can also define the BL metric directly for random variables or distribution functions using the correspondences between them. Hence,

$$\operatorname{dist}_{\mathrm{BL}}(X, Y) := \operatorname{dist}_{\mathrm{BL}}(F_X, F_Y) := \operatorname{dist}_{\mathrm{BL}}(\mu_X, \mu_Y),$$

where F_X, F_Y are the distribution functions and μ_X, μ_Y are the laws of X, Y.

It is now quite easy to see that $\operatorname{dist}_{\mathrm{BL}}$ defines a metric on the class of Borel probability measures on the real line. In other words, it gives a rigorous way to quantify the similarity between two probability measures.

Exercise 17.10 (BL distance is a metric). Show that the BL distance is a metric. For all Borel probability measures μ, ν, ϱ on the real line,

1. **Positive definiteness:** $\operatorname{dist}_{\mathrm{BL}}(\mu, \nu) \geq 0$, with equality if and only if $\mu = \nu$.
2. **Symmetry:** $\operatorname{dist}_{\mathrm{BL}}(\mu, \nu) = \operatorname{dist}_{\mathrm{BL}}(\nu, \mu)$.
3. **Triangle inequality:** $\operatorname{dist}_{\mathrm{BL}}(\mu, \nu) \leq \operatorname{dist}_{\mathrm{BL}}(\mu, \varrho) + \operatorname{dist}_{\mathrm{BL}}(\varrho, \nu)$.

Hint: Proposition 17.8 takes care of the only tricky detail.

17.2.4 Convergence

Like every metric, the BL metric induces a notion of convergence. As we will learn, this type of convergence plays a central role in classical probability theory.

Definition 17.11 (BL metric: Convergence). Consider a sequence $(\mu_i : i \in \mathbb{N})$ of Borel probability measures on the real line. For another Borel probability measure μ on

the real line, suppose that

$$\text{dist}_{\text{BL}}(\mu_n, \mu) \to 0 \quad \text{as } n \to \infty.$$

Then we say that the measures $\mu_n \to \mu$ with respect to the BL metric.

In detail, convergence with respect to the BL metric means that

$$\text{dist}_{\text{BL}}(\mu_n, \mu) = \sup\{|\mu_n(h) - \mu(h)| : \|h\|_{\text{BL}} \leq 1\} \to 0.$$

As a particular consequence, convergence in the BL metric implies convergence for any particular bounded, Lipschitz test function:

$$\mu_n(h) \to \mu(h) \quad \text{for each bounded, Lipschitz function } h : \mathbb{R} \to \mathbb{R}.$$

In fact, these two statements are equivalent with each other. Convergence in the BL metric is the same as convergence for each BL test function.

Theorem 17.12 (BL metric: Convergence). Consider a sequence $(\mu_n : n \in \mathbb{N})$ of Borel probability measures on the real line, and let μ be another Borel probability measure on the real line. The following statements are equivalent:

1. $\text{dist}_{\text{BL}}(\mu_n, \mu) \to 0$ as $n \to \infty$.
2. $\mu_n(h) \to \mu(h)$ as $n \to \infty$ for each bounded, Lipschitz $h : \mathbb{R} \to \mathbb{R}$.

We postpone the proof of Theorem 17.12 to Section 17.7 because it requires a dose of functional analysis.

In the next section, we will glorify this type of convergence with its own name, and we will develop a number of equivalent conditions. Afterward, we will explore some generalizations of the BL probability metric.

17.3 Weak convergence

Convergence with respect to the BL distance plays a central role in classical probability theory because it is equivalent to several other types of convergence. In this section, we will develop some of these connections in detail.

17.3.1 Weak convergence of probability measures

Let us assign a name to the notion of convergence with respect to the BL distance. Motivated by Theorem 17.12, we pose the definition in terms of convergence for individual test functions.

Definition 17.13 (Weak convergence: Probability measures). Consider a sequence $(\mu_n : n \in \mathbb{N})$ of Borel probability measures on $\mathbb{R}$, and let μ be another Borel probability measure on $\mathbb{R}$. We say that μ_n *converges weakly* to μ when

$$\mu_n(h) \to \mu(h) \quad \text{for each bounded, Lipschitz function } h : \mathbb{R} \to \mathbb{R}.$$

Common notations for weak convergence include $\mu_n \rightsquigarrow \mu$ and $\mu_n \xrightarrow{w} \mu$.

Weak convergence is defined by the pointwise convergence of a large family of moments. Theorem 17.12 states that weak convergence is the same as convergence with respect to the BL metric. You may find one definition or the other more intuitive, but they give identical results. We will gain further intuition in the upcoming sections.

Our first observation is that weak limits are unique.

Exercise 17.14 (Weak limits: Uniqueness). Suppose that $\mu_n \rightsquigarrow \mu$ and $\mu_n \rightsquigarrow \nu$. Show that $\mu = \nu$. In other words, weak limits are unique. **Hint:** Just use Proposition 17.8.

The next exercise shows that we can also define weak convergence using another natural class of test functions. The alternative definition can sometimes be more convenient to use.

Exercise 17.15 (Weak convergence: Bounded, continuous functions). Show that $\mu_n \rightsquigarrow \mu$ if and only if $\mu_n(h) \to \mu(h)$ for each bounded, *continuous* function $h : \mathbb{R} \to \mathbb{R}$. **Hint:** Lipschitz functions are dense in the space of bounded, continuous functions, equipped with the supremum norm.

Warning 17.16 (Weak convergence: Bad notation). Some authors write $\mu_n \Rightarrow \mu$ for weak convergence. This notation is deprecated because of the potential for confusion with logical implication. ▪

Warning 17.17 (Weak convergence: Functional analysis). Weak convergence in probability theory is not the same as weak convergence in functional analysis. ▪

17.3.2 Weak convergence of random variables

We can also define weak convergence directly for random variables.

Definition 17.18 (Weak convergence: Random variables). Consider a sequence $(X_n : n \in \mathbb{N})$ of real random variables, and let X be another real random variable. We say that X_n *converges weakly* to X when

$$\mathbb{E}\, h(X_n) \to \mathbb{E}\, h(X) \quad \text{for each bounded, Lipschitz function } h : \mathbb{R} \to \mathbb{R}.$$

We write $X_n \rightsquigarrow X$ or $X_n \xrightarrow{w} X$.

Exercise 17.19 (Weak convergence: Equivalence). Verify that $X_n \rightsquigarrow X$ if and only if the laws of the random variables converge weakly: $\mu_{X_n} \rightsquigarrow \mu_X$.

This exercise emphasizes that the weak convergence is insensitive to relationships among the random variables X_n. Indeed, it only reflects the marginal laws of the random variables, regardless of any dependency or independency between them. The random variables X_n do not even need to be defined on the same probability space.

Exercise 17.20 (Almost-sure convergence implies weak convergence). Consider a sequence $(X_n : n \in \mathbb{N})$ of real random variables defined on the *same probability space*. Prove that $X_n \to X$ almost surely implies that $X_n \rightsquigarrow X$. In Section 17.8.5, we will see that this statement has a partial converse.

17.3.3 Weak convergence of distribution functions

For probability measures on the real line, weak convergence can also be formulated in terms of (cumulative) distribution functions. This is part of the reason that weak convergence plays a large role in classical probability theory.

Definition 17.21 (Weak convergence: Distribution functions). A sequence $(F_n : n \in \mathbb{N})$ of distribution functions on the real line *converges weakly* to a distribution function

F if and only if the associated laws converge weakly: $\mu_n \rightsquigarrow \mu$. We write $F_n \rightsquigarrow F$ to denote weak convergence of distribution functions.

There is an alternative characterization of weak convergence of distribution functions that may clarify these concepts.

Theorem 17.22 (Weak convergence: Distribution functions). Suppose that a sequence $(F_n : n \in \mathbb{N})$ of distribution functions on the real line, and let F be another distribution function on the real line. Then $F_n \rightsquigarrow F$ if and only if

$$F_n(a) \to F(a) \quad \text{for each } a \in \mathbb{R} \text{ at which } F \text{ is continuous}.$$

We will prove Theorem 17.22 in Section 17.8, along with some related results.

Under the *assumption* that the limit F is continuous, $F_n \rightsquigarrow F$ is the same as $F_n \to F$ pointwise. The most common situation where the limit F is continuous is when it is the distribution function of a continuous random variable. (That is, a random variable whose law is absolutely continuous with respect to Lebesgue measure.)

When the limit F is discontinuous, $F_n \rightsquigarrow F$ provides no guarantees at the points of discontinuity of F. The most common situation where the limit F is not continuous is when it is the distribution function of a discrete random variable. In this case, other (stronger) notions of convergence may be more appropriate so that we can control what happens at discontinuities.

Note that the convergence $F_n(a) \to F(a)$ may not be uniform, even restricted to points of continuity of F. That is, the distribution functions may converge at different rates at different points. Uniform convergence of distribution functions is a stricter requirement than weak convergence; see Section 17.6.1.

Owing to Theorem 17.22, weak convergence of measures is often called *convergence in distribution*. This terminology, however, may be confusing because there are many other modes of convergence for measures.

Exercise 17.23 (Weak convergence: Atoms). What is the weak limit of the sequence $(\delta_{1/n} : n \in \mathbb{N})$ of Dirac measures as $n \to \infty$? Do the distribution functions converge pointwise?

17.4 *Weak convergence and functional analysis

The probabilists' weak convergence can be understood in a functional analytic sense. We can equip the bounded, continuous functions on $\overline{\mathbb{R}}$ with the supremum norm to form a Banach space, denoted $\mathsf{C}(\overline{\mathbb{R}})$. The dual $\mathsf{C}(\overline{\mathbb{R}})^*$ of this Banach space consists of all finite, signed Borel measures on the extended real line, equipped with the norm $\|\mu\|_{\mathrm{TV}} := \sup\{\mu(h) : \|h\|_{\sup} \le 1\}$. By Exercise 17.15, weak convergence $\mu_n \rightsquigarrow \mu$ of probability measures is the same as weak-$*$ convergence of the probability measures, viewed as elements of the dual space.

By the Banach–Alaoglu theorem, the unit ball in $\mathsf{C}(\overline{\mathbb{R}})^*$ is weak-$*$ compact. Among other consequences, every sequence $(\mu_n : n \in \mathbb{N})$ of probability measures on $\overline{\mathbb{R}}$ has a weak-$*$ convergent subsequence. The weak-$*$ limit of the subsequence is a probability measure on $\overline{\mathbb{R}}$. Unfortunately, even if the μ_n are probability measures on $\mathbb{R}$, some of the mass can migrate to $\{\pm\infty\}$ and the limiting measure μ may place strictly less than one unit of mass on $\mathbb{R}$.

To ensure that the weak-$*$ limit of probability measures on $\mathbb{R}$ remains a probability measure on $\mathbb{R}$, the sequence $(\mu_n : n \in \mathbb{N})$ needs to have an additional property, called *tightness*. This condition ensures that mass does not "leak out" at infinity.

Definition 17.24 (Tightness). Consider a sequence $(\mu_n : n \in \mathbb{N})$ of Borel probability measures on the real line. For each $\varepsilon > 0$, suppose that there is a compact set $\mathsf{K} \subset \mathbb{R}$ with the property that

$$\mu_n(\mathsf{K}) > 1 - \varepsilon \quad \text{for all } n \in \mathbb{N}.$$

Then the sequence (μ_n) is said to be *tight*.

Theorem 17.25 (Prokhorov). Let $(\mu_n : n \in \mathbb{N})$ be a sequence of Borel probability measures on the real line.

1. If $\mu_n \rightsquigarrow \mu$ where μ is a *probability* measure, then the sequence (μ_n) is tight.
2. If the sequence (μ_n) is tight, then it admits a *subsequence* $(\mu_{n_k} : k \in \mathbb{N})$ that converges weakly to a *probability* measure μ. That is, $\mu_{n_k} \rightsquigarrow \mu$.

Proof. **(1).** Assume that $\mu_n \rightsquigarrow \mu$ where μ is a probability measure. We must argue that the sequence is tight.

Consider a sequence $(h_i : i \in \mathbb{N})$ of bounded Lipschitz functions where h_i approximates the indicator of $[-i, +i]$ from below. More precisely, set $h_i(t) = 1$ for all $|t| \le i - 1$ and $h_i(t) = 0$ for $|t| \ge i$. Since μ is a probability measure,

$$1 = \mu(\mathbb{1}_{\mathbb{R}}) = \lim_{i\to\infty} \mu(h_i) = \lim_{i\to\infty} \lim_{n\to\infty} \mu_n(h_i) \le \lim_{i\to\infty} \lim_{n\to\infty} \mu_n([-i, +i]).$$

We have used the fact that $h_i \uparrow 1$ pointwise to apply the monotone convergence theorem (Theorem 9.10). The weak convergence $\mu_n \rightsquigarrow \mu$ as $n \to \infty$ justifies the second limit. Last, we bound the Lipschitz function h_i above by the indicator of $[-i, +i]$. In fact, the right-hand side cannot exceed 1 because each μ_n is a probability measure.

Now, fix a parameter $\varepsilon > 0$. There exists an interval $[-i, +i]$ where

$$\lim_{n\to\infty} \mu_n([-i, +i]) > 1 - \varepsilon.$$

Next, we select a number N where $\mu_n([-i, +i]) > 1 - 2\varepsilon$ for all $n \ge N$. Since the μ_n are probability measures, we can also find a compact set K where $\mu_n(\mathsf{K}) > 1 - 2\varepsilon$ for each $n < N$. As a consequence,

$$\mu_n([-i, +i] \cup \mathsf{K}) > 1 - 2\varepsilon \quad \text{for all } n \in \mathbb{N}.$$

We conclude that the sequence (μ_n) is tight.

(2). Suppose that (μ_n) is a tight sequence of probability measures on the real line. By weak-$*$ compactness of $\mathsf{C}(\overline{\mathbb{R}})^*$, there is a subsequence $(\mu_{n_k} : k \in \mathbb{N})$ that converges to a measure μ on $\mathbb{R}$ in the sense that

$$\mu_{n_k}(h) \to \mu(h) \quad \text{for all bounded, continuous } h : \mathbb{R} \to \mathbb{R}.$$

In particular, the limit holds for all bounded, Lipschitz test functions. It remains to confirm that the limit μ is a *probability* measure.

For each $\varepsilon > 0$, there exists an interval $[-i, +i]$ for which $\mu_{n_k}([-i, +i]) > 1 - \varepsilon$ for all $k \in \mathbb{N}$ because the sequence is tight. Using the bounded Lipschitz functions h_i constructed above, we find that

$$1 - \varepsilon \le \lim_{k\to\infty} \mu_{n_k}([-i, +i]) \le \lim_{k\to\infty} \mu_{n_k}(h_{i+1}) = \mu(h_{i+1}) \le \mu(\mathbb{R}).$$

Since ε is arbitrary, we deduce that $\mu(\mathbb{R}) = 1$. ∎

17.5 *Weak convergence: Higher dimensions

Definition 17.13 describes weak convergence for distributions on the real line. The same ideas apply to distributions on $\mathbb{R}^n$. Let us summarize the approach, without proof.

Definition 17.26 (Bounded, Lipschitz function on $\mathbb{R}^n$). A function $h : \mathbb{R}^n \to \mathbb{R}$ is *bounded* when

$$\|h\|_{\sup} := \sup\{|h(\boldsymbol{a})| : \boldsymbol{a} \in \mathbb{R}^n\} < +\infty.$$

The function is *Lipschitz* when its Lipschitz pseudonorm is finite:

$$\|h\|_{\mathrm{Lip}} := \inf\{L \geq 0 : |h(\boldsymbol{a}) - h(\boldsymbol{b})| \leq L \cdot \|\boldsymbol{a} - \boldsymbol{b}\|_2 \text{ for all } \boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^n\}.$$

Here, $\|\cdot\|_2$ denotes the Euclidean norm.

The bounded, Lipschitz norm is now defined for functions on $\mathbb{R}^n$:

$$\|h\|_{\mathrm{BL}} := \max\{\|h\|_{\sup}, \|h\|_{\mathrm{Lip}}\}.$$

Definition 17.27 (Weak convergence on $\mathbb{R}^n$). Consider a sequence $(\mu_n : n \in \mathbb{R})$ of Borel probability measure on $\mathbb{R}^n$, and let μ be another Borel probability measure on $\mathbb{R}^n$. We say that μ_n *converges weakly* to μ when

$$\mu_n(h) \to \mu(h) \quad \text{for all bounded, Lipschitz } h : \mathbb{R}^n \to \mathbb{R}.$$

We write $\mu_n \rightsquigarrow \mu$. We use a similar notation for weak convergence of random variables taking values in $\mathbb{R}^n$.

As before, we can define a metric on distributions on $\mathbb{R}^n$ that is induced by the bounded, Lipschitz functions.

Definition 17.28 (BL metric: Probability measures on $\mathbb{R}^n$). Let μ, ν be Borel probability measures on $\mathbb{R}6n$. Define the bounded, Lipschitz (BL) probability metric:

$$\mathrm{dist}_{\mathrm{BL}}(\mu, \nu) := \sup\{|\mu(h) - \nu(h)| : \|h\|_{\mathrm{BL}} \leq 1\}.$$

We have similar notation for random variables taking values in $\mathbb{R}^n$.

Theorem 17.29 (BL metrizes weak convergence on $\mathbb{R}^n$). Consider a sequence $(\mu_n : n \in \mathbb{N})$ of Borel probability measures on $\mathbb{R}$, and let μ be another Borel probability measure on $\mathbb{R}$. Then

$$\mu_n \rightsquigarrow \mu \quad \text{if and only if} \quad \mathrm{dist}_{\mathrm{BL}}(\mu_n, \mu) \to 0.$$

We omit the proof. It is based on an argument similar to Theorem 17.12, but it requires more technical effort.

Aside: This machinery extends to a much more general setting. Indeed, we can use exactly the same ideas to define weak convergence of Borel probability measures on a complete, separable metric space. This generalization plays a major role in statistics and the study of empirical processes. See [Dudo2] for an introduction to this circle of ideas.

17.6 Integral probability metrics

We can generalize the idea behind the construction of the bounded Lipschitz distance to obtain other kinds of distances between probability distributions. These distances arise in many applications of probability theory, and they share some properties with the BL distance.

We would like to implement the general idea that two probability measures are close when a large collection of moments are similar. Here is the approach.

Definition 17.30 (Integral probability metric). Let H be a collection of measurable functions on $\mathbb{R}$. For two Borel probability measures μ, ν on $\mathbb{R}$, we define

$$\mathrm{dist}_{\mathsf{H}}(\mu, \nu) := \sup\{|\mu(h) - \nu(h)| : h \in \mathsf{H}\}.$$

This is called the *integral probability (pseudo)metric* induced by the set H of test functions.

It is easy to see that Definition 17.30 includes the bounded, Lipschitz distance as a special case when $\mathsf{H} = \{h : \mathbb{R} \to \mathbb{R} : \|h\|_{\mathrm{BL}} \leq 1\}$.

The terminology "integral probability metric" is partially justified by the fact that $\mu(h)$ is the integral of the function h against the measure μ. The rest of the justification appears in the next two exercises.

Exercise 17.31 (IPM is a pseudometric). Show that $\mathrm{dist}_{\mathsf{H}}$ is a pseudometric. That is, the function $\mathrm{dist}_{\mathsf{H}}$ is positive, symmetric, and satisfies the triangle inequality.

Exercise 17.32 (When is an IPM a metric?). Show that $\mathrm{dist}_{\mathsf{H}}$ is a metric if and only if the collection H separates points. That is, $\mu \neq \nu$ implies that $\mu(h) \neq \nu(h)$ for some function $h \in \mathsf{H}$.

As we discussed, the more moments that are similar, the more the probability distributions are similar. This intuition is captured by the next exercise.

Exercise 17.33 (IPM: Monotonicity). Verify that $\mathsf{H} \subseteq \mathsf{H}'$ implies that $\mathrm{dist}_{\mathsf{H}} \leq \mathrm{dist}_{\mathsf{H}'}$ for all arguments.

Note, however, that two IPMs, say $\mathrm{dist}_{\mathsf{H}}$ and $\mathrm{dist}_{\mathsf{H}'}$, may be incomparable as metrics when there is no containment between the classes H and H' of test functions.

17.6.1 Examples of IPMs

There are many important examples of IPMs. Here are a few.

Example 17.34 (Kolmogorov distance). The collection

$$\mathsf{H} = \{\mathbb{1}_{(-\infty, a]} : a \in \mathbb{R}\}$$

generates the *Kolmogorov distance* on probability measures. It is not hard to check that convergence in Kolmogorov distance is the same as uniform, pointwise convergence of distribution functions. Therefore, the Kolmogorov distance is stronger than the BL distance; in particular, it is a metric on probability measures. ▪

Example 17.35 (Kantorovich-1 distance). The collection

$$\mathsf{H} = \{h : \mathbb{R} \to \mathbb{R} : \|h\|_{\mathrm{Lip}} \leq 1\}$$

generates the *Kantorovich-1 distance* on probability measures. It is easy to see that H strictly contains the functions with BL norm less than one. Therefore, this distance is larger than the BL distance; in particular, it is a metric.

Convergence $\text{dist}_{\mathsf{H}}(\mu_n, \mu) \to 0$ is the same as weak convergence, plus convergence of first moments $\mu_n(|x|) \to \mu(|x|)$.

The Kantorovich-1 distance arises in theory of optimal transport, where it is often called the *Wasserstein-1* distance. ▪

Example 17.36 (Total variation distance). The collection

$$\mathsf{H} = \{h : \mathbb{R} \to \mathbb{R} : \|h\|_{\text{sup}} \le 1 \text{ and } h \text{ continuous}\}$$

generates the *total variation (TV) distance* on probability measures. Equivalently, we can take $\mathsf{H} = \{\mathbb{1}_{\mathsf{B}} : \mathsf{B} \in \mathscr{B}(\mathbb{R})\}$. It is clear that H strictly contains the functions with BL norm less than one. Therefore, the TV distance is also at as large as the BL distance, so is a metric.

Convergence in total variation $\text{dist}_{\mathsf{H}}(\mu_n, \mu) \to 0$ requires that $\mu_n(\mathsf{B}) \to \mu(\mathsf{B})$ for every Borel set $\mathsf{B} \in \mathscr{B}(\mathbb{R})$. You can see that this is a very strict requirement. As a consequence, total variation distance is rarely used for (absolutely) continuous distributions, but it may be appropriate for convergence of discrete distributions (e.g., supported on the integers). ▪

There is a lot more to say about each of these particular metrics, and they serve as appropriate tools for particular applications. Because IPMs derive from the same conceptual framework, there is a logical economy to these notions that makes them easier to understand and to compare with each other than other modes of convergence.

17.7 *BL distance metrizes weak convergence

In this section, we prove Theorem 17.12, which states that convergence in the BL metric is exactly the same as weak convergence. This argument requires a dose of functional analysis.

Proof of Theorem 17.12. Consider a sequence $(\mu_n : n \in \mathbb{N})$ of Borel probability measures on the real line, and let μ be another Borel probability measure on the real line. We must prove that

$$\mu_n \rightsquigarrow \mu \quad \text{if and only if} \quad \text{dist}_{\text{BL}}(\mu_n, \mu) \to 0.$$

We begin with the easier part.

Reverse direction

Assume that $\text{dist}_{\text{BL}}(\mu_n, \mu) \to 0$. Explicitly,

$$\text{dist}_{\text{BL}}(\mu_n, \mu) = \sup\{|\mu_n(h) - \mu(h)| : \|h\|_{\text{BL}} \le 1\} \to 0.$$

Therefore, for each (nonzero) bounded, Lipschitz function $h : \mathbb{R} \to \mathbb{R}$, we can calculate that

$$\begin{aligned} |\mu_n(h) - \mu(h)| &= \|h\|_{\text{BL}} \cdot |\mu_n(h/\|h\|_{\text{BL}}) - \mu(h/\|h\|_{\text{BL}})| \\ &\le \|h\|_{\text{BL}} \cdot \text{dist}_{\text{BL}}(\mu_n, \mu) \to 0. \end{aligned}$$

Thus, $\mu_n(h) \to \mu(h)$ for every bounded, Lipschitz function h.

Forward direction

Assume that $\mu_n \rightsquigarrow \mu$. Explicitly,

$$\mu_n(h) \to \mu(h) \quad \text{for each bounded, Lipschitz function } h : \mathbb{R} \to \mathbb{R}.$$

The challenge is to obtain uniform convergence over all functions with $\|h\|_{\mathrm{BL}} \le 1$.

Fix a parameter $\varepsilon > 0$. Observe that there is a *compact* interval $\mathsf{I} \subset \mathbb{R}$ with the property that $\mu(\text{closure}(\mathsf{I}^{\mathsf{c}})) \le \varepsilon$. This is an easy consequence of the increasing limit property of a measure (Proposition 2.30). This is a good time to note that

$$\limsup_{n\to\infty} \mu_n(\mathsf{I}^{\mathsf{c}}) \le \mu(\text{closure}(\mathsf{I}^{\mathsf{c}})) \le \varepsilon \quad \text{as } n \to \infty. \tag{17.1}$$

Indeed, we can approximate the indicator $\mathbb{1}_{\text{closure}(\mathsf{I}^{\mathsf{c}})}$ above by a Lipschitz function. At a small cost, we can neglect what happens outside the interval I.

Consider the Banach space $\mathsf{C}(\mathsf{I})$ of bounded, continuous, real-valued functions on the compact interval I, equipped with the supremum norm $\|\cdot\|_{\sup(\mathsf{I})}$. Introduce the family of bounded Lipschitz functions on the interval:

$$\mathsf{K} := \{h : \mathsf{I} \to \mathbb{R} : \|h\|_{\mathrm{BL}(\mathsf{I})} \le 1\}.$$

We have restricted the BL norm to real-valued functions on I in the obvious way. It is a standard consequence of the Arzelà–Ascoli theorem that K is a (relatively) compact subset of $\mathsf{C}(\mathsf{I})$. In particular, we can produce a finite ε-covering of the set K. More precisely, there exist a finite number J of functions $g_1, \dots, g_J \in \mathsf{K}$ such that

$$\min_j \|h - g_j\|_{\sup(\mathsf{I})} \le \varepsilon \quad \text{for each } h \in \mathsf{K}.$$

You can easily construct a finite ε-covering of K with your bare hands. Try it! In more general settings, it is less obvious how to do so.

It is convenient to treat each function $g_j : \mathsf{I} \to \mathbb{R}$ as the restriction to I of a bounded, Lipschitz function $g_j : \mathbb{R} \to \mathbb{R}$ on the real line. (For example, we can just make the constant extension of the value of g_j at the endpoints of I.)

This construction allows us to replace the supremum over bounded, Lipschitz functions by the maximum over the J functions in the covering. For each function $h : \mathbb{R} \to \mathbb{R}$ with $\|h\|_{\mathrm{BL}} \le 1$. Define g_h to be the function in $\{g_1, \dots, g_J\}$ that is closest to h in the supremum norm on I, breaking ties lexicographically. Thus, $\|h - g_h\|_{\sup(\mathsf{I})} \le \varepsilon$.

To bound the BL distance between μ_n and μ, we proceed as follows. Fix $h : \mathbb{R} \to \mathbb{R}$ with $\|h\|_{\mathrm{BL}} \le 1$. First, we replace h with its approximation g_h:

$$\begin{aligned} |\mu_n(h) - \mu(h)| &\le |\mu_n(g_h) - \mu(g_h)| + |\mu_n(h - g_h)| + |\mu(h - g_h)| \\ &\le \Big(\sum\nolimits_{j=1}^{J} |\mu_n(g_j) - \mu(g_j)|\Big) + |\mu_n(h - g_h)| + |\mu(h - g_h)|. \end{aligned}$$

We have bounded the first term by the sum over all possible choices for g_h, so it no longer depends on h. Notice that the limit of the sum is zero because $\mu_n \rightsquigarrow \mu$ and each g_j is a bounded, Lipschitz function.

Next, we obtain bounds for the error terms. For example,

$$\begin{aligned} |\mu_n(h - g_h)| &\le |\mu_n(h - g_h; \mathsf{I})| + |\mu_n(h - g_h; \mathsf{I}^{\mathsf{c}})| \\ &\le \|h - g_h\|_{\sup(\mathsf{I})} \cdot \mu_n(\mathsf{I}) + \|h - g_h\|_{\sup} \cdot \mu_n(\mathsf{I}^{\mathsf{c}}) \\ &\le \varepsilon + 2\mu_n(\mathsf{I}^{\mathsf{c}}). \end{aligned}$$

We used the triangle inequality to control the supremum of h and g_h on I^{c}. Using (17.1), we take the limit superior as $n \to \infty$ to see that

$$\limsup_{n\to\infty} \Big[\sup\{|\mu_n(h - g_h)| : \|h\|_{\mathrm{BL}} \le 1\} \Big] \le 3\varepsilon.$$

A similar argument implies that $\sup\{|\mu(h - g_h)| : \|h\|_{\mathrm{BL}} \le 1\} \le 3\varepsilon$.

Altogether, we see that the BL distance satisfies

$$\begin{aligned}
&\limsup_{n\to\infty} \mathrm{dist}_{\mathrm{BL}}(\mu_n, \mu) \\
&\qquad = \limsup_{n\to\infty} \left[\sup\{|\mu_n(h) - \mu(h)| : \|h\|_{\mathrm{BL}} \le 1\} \right] \\
&\qquad \le \limsup_{n\to\infty} \left[\left(\sum_{j=1}^{J} |\mu_n(g_j) - \mu(g_j)| \right) + \sup_{\|h\|_{\mathrm{BL}}\le 1} \left(|\mu_n(h - g_h)| + |\mu(h - g_h)| \right) \right] \\
&\qquad \le 6\varepsilon.
\end{aligned}$$

Since the parameter ε is arbitrary, we determine that

$$\mathrm{dist}_{\mathrm{BL}}(\mu_n, \mu) \to 0 \quad \text{as } n \to \infty.$$

In summary, the BL distance metrizes weak convergence. ∎

17.8 *Weak convergence of distribution functions

In this section, we present a proof of Theorem 17.22, which gives the alternative characterization of weak convergence in terms of distribution functions. We begin with the forward direction, which is an easy application of the definition of weak convergence. To prove the reverse direction, we also need a result, called the Skorokhod theorem, that clarifies the relationship between weak convergence and almost-sure convergence.

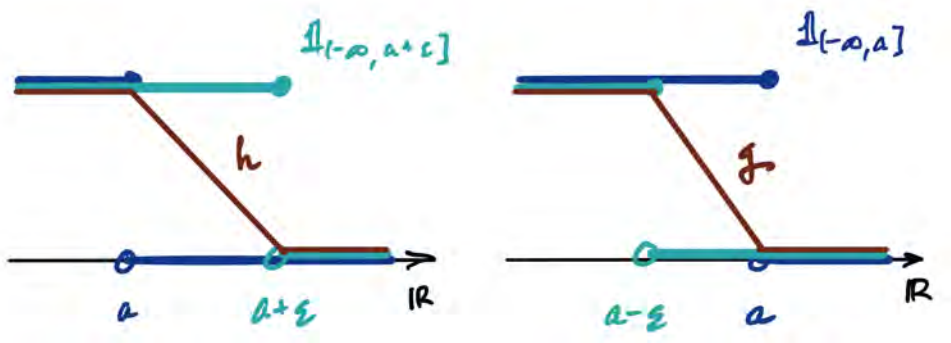
17.8.1 Forward direction

In this section, we prove that weak convergence implies that distribution functions converge pointwise at points of continuity of the limit.

Proposition 17.37 (Theorem 17.22: Forward direction). Consider a weakly convergent sequence of distribution functions $F_n \rightsquigarrow F$ on the real line, per Definition 17.21. Then $F_n(a) \to F(a)$ at each point $a \in \mathbb{R}$ where F is continuous.

Proof. For each $n \in \mathbb{N}$, let μ_n be the Borel probability measure with distribution function F_n, and let μ be the Borel probability measure with distribution function F.

Choose any point $a \in \mathbb{R}$ where F is continuous, and fix $\varepsilon > 0$. We will approximate indicator functions above and below by Lipschitz functions:



Let h be the piecewise linear function with $h(y) = 1$ for $y \le a$ and $h(y) = 0$ for $y \ge a + \varepsilon$. Since h is bounded and Lipschitz, weak convergence yields

$$F_n(a) = \mu_n(-\infty, a] \le \mu_n(h) \to \mu(h) \le \mu(-\infty, a + \varepsilon] = F(a + \varepsilon).$$

Therefore, $\limsup_{n\to\infty} F_n(a) \le F(a + \varepsilon)$.

Next, consider the piecewise linear function with $g(y) = 1$ for $y \le a - \varepsilon$ and $g(y) = 0$ for $y \ge a$. Since g is bounded and Lipschitz, weak convergence yields

$$F_n(a) = \mu_n(-\infty, a] \ge \mu_n(g) \to \mu(g) \ge \mu(-\infty, a - \varepsilon] = F(a - \varepsilon).$$

Therefore, $\liminf_{n\to\infty} F_n(a) \ge F(a - \varepsilon)$.

The distribution function F is continuous at a, and the parameter ε is arbitrary. Therefore, we may take $\varepsilon \downarrow 0$ to deduce that

$$F(a) \le \liminf_{n\to\infty} F_n(a) \le \limsup_{n\to\infty} F_n(a) \le F(a).$$

In other words, $\lim_{n\to\infty} F_n(a) = F(a)$. ∎

17.8.2 Skorokhod representation of a random variable

Weak convergence of probability measures does not necessarily involve random variables. Nevertheless, we might like to understand whether it is possible to model weak convergence of measures using some type of convergence for random variables. To do so, we need a way to produce a random variable with a specified distribution function.

Recall that the universal probability space is the triple $([0, 1], \mathscr{B}[0, 1], \lambda)$. The next result gives an explicit construction of a random variable on the universal probability space that has a specified distribution function.

Proposition 17.38 (Skorokhod representation). Consider any distribution function F on the real line. The universal probability space supports a real random variable with distribution function F. Here are two explicit expressions for such a random variable. For each $\omega \in [0, 1]$,

$$X^-(\omega) := \inf\{a \in \mathbb{R} : \omega \le F(a)\};$$
$$X^+(\omega) := \inf\{a \in \mathbb{R} : \omega < F(a)\}.$$

Furthermore, $X^- = X^+$ almost surely.

The random variables constructed here are essentially inverses of the distribution function. For $\omega \in [0, 1]$ the fact that the distribution function F is right-continuous guarantees that

$$X^-(\omega) \le a \quad \text{if and only if} \quad \omega \le F(a). \tag{17.2}$$

Meanwhile,

$$X^+(\omega) < a \quad \text{implies} \quad \omega < F(a); \tag{17.3}$$
$$\omega \le F(a) \quad \text{implies} \quad X^+(\omega) \le a. \tag{17.4}$$

You should verify these relations, which will be used heavily.

Proof. First, let us confirm that X^- has distribution function F. The probability measure is the Lebesgue measure on $[0, 1]$. For each $a \in \mathbb{R}$, the equivalence (17.2) implies that

$$\mathbb{P}\{X^- \le a\} = \lambda\{\omega \in [0, 1] : X^{-1}(\omega) \le a\} = \lambda\{\omega \in [0, 1] : \omega \le F(a)\} = F(a).$$

In words, the distribution function of X^- coincides with F.

Next, we demonstrate that $X^- = X^+$ almost surely, which further implies that X^+ has distribution function F. Since $X^- \le X^+$ pointwise, the event

$$\{X^- \neq X^+\} = \bigcup\nolimits_{q \in \mathbb{Q}} \{X^- \le q < X^+\}.$$

For each (real) $q \in \mathbb{R}$, we can control the probability of each member of the union:

$$\begin{aligned}\mathbb{P}\left\{X^- \le q < X^+\right\} &= \mathbb{P}\big(\{X^- \le q\} \setminus \{X^+ \le q\}\big) \\ &= \mathbb{P}\{X^- \le q\} - \mathbb{P}\{X^+ \le q\} \le F(q) - F(q) = 0.\end{aligned}$$

Indeed, since $X^- \le X^+$, we have the inclusion $\{X^+ \le q\} \subseteq \{X^- \le q\}$. The last relation $F(q) \le \mathbb{P}\{X^+ \le q\}$ is a consequence of (17.3). The result follows from countable additivity. ∎

17.8.3 Skorokhod's theorem

The next result shows that a weakly convergent sequence of probability measures can be modeled by a sequence of random variables that converges almost surely. To do so, we simply consider the Skorokhod representation of each probability measure, given by Proposition 17.38.

Theorem 17.39 (Skorokhod). Consider a sequence $(F_n : n \in \mathbb{N})$ of distribution functions that converges pointwise to a distribution function F at each point where F is continuous. That is, $F_n(a) \to F(a)$ whenever F is continuous at $a \in \mathbb{R}$.

Then the universal probability space supports a sequence $(X_n : n \in \mathbb{N})$ of random variables for which $X_n \to X$ almost surely, and each X_n has distribution function F_n, and X has distribution function F.

Proof. Using Proposition 17.38, for each $n \in \mathbb{N}$, construct random variables X_n^- and X_n^+ with the distribution function F_n and random variables X^- and X^+ with the distribution function F. We will show that $X_n^- \to X^-$ almost surely.

Fix a sample point $\omega \in [0, 1]$. Choose any point $a \in \mathbb{R}$ where F is continuous and where $X^+(\omega) < a$. According to (17.3), $\omega < F(a)$. For all large n, we must have $\omega < F_n(a)$, and (17.4) guarantees that $X_n^+(\omega) \le a$. Therefore, $\limsup_{n\to\infty} X_n^+(\omega) \le a$. Since F is an increasing function, it has at most a countable number of discontinuities. Therefore, we can decrease a sequentially to $X^+(\omega)$ to deduce that

$$\limsup\nolimits_{n\to\infty} X_n^+(\omega) \le X^+(\omega).$$

Analogous arguments show that

$$\liminf\nolimits_{n\to\infty} X_n^-(\omega) \ge X^-(\omega).$$

Since $X^- \le X^+$ pointwise and $X^- = X^+$ almost surely, we determine that $X_n^- \to X^-$ almost surely. ∎

17.8.4 Reverse direction

Finally, we may establish the reverse direction of Theorem 17.22.

Proposition 17.40 (Theorem 17.22: Reverse direction). Consider a sequence of distribution functions that satisfies $F_n(a) \to F(a)$ for all $a \in \mathbb{R}$ where F is continuous. Then the distribution functions converge weakly: $F_n \rightsquigarrow F$.

Proof. Skorokhod's result, Theorem 17.39, furnishes a sequence of random variables that satisfy $X_n \to X$ almost surely, where X_n has distribution function F_n, and where X has distribution function F.

For each $n \in \mathbb{N}$, let μ_n be the probability measure with distribution function F_n, and let μ be the probability measure with distribution function F. For any bounded, Lipschitz function $h : \mathbb{R} \to \mathbb{R}$, the bounded convergence theorem (Corollary 9.13) ensures that

$$\mathbb{E}\, h(X_n) \to \mathbb{E}\, h(X).$$

Equivalently, the associated probability measures satisfy $\mu_n(h) \to \mu(h)$. We conclude that $F_n \rightsquigarrow F$. ∎

17.8.5 Almost-sure convergence can model weak convergence

In Exercise 17.20, we saw that an almost-surely convergent sequence of random variables must also converge weakly. We are now prepared to establish a partial converse.

Corollary 17.41 (Almost-sure convergence can model weak convergence). Consider a weakly convergent sequence of distribution functions: $F_n \rightsquigarrow F$. Then the universal probability space supports a sequence of random variables where $X_n \to X$ almost surely, where X_n has distribution function F_n, and where X has distribution function F.

Proof. Combine Theorem 17.39 with Proposition 17.40. ∎

Problems

Problem 17.42 (Comparison of distances). Let $Z \sim \text{NORMAL}(0, 1)$. For a real random variable Y, recall the definitions of the Kolmogorov (Kol) and Kantorovich-1 (W_1) distances between Y and Z:

$$d_{\text{Kol}}(Y, Z) := \sup_{a \in \mathbb{R}} |\mathbb{P}\{Y \le a\} - \mathbb{P}\{Z \le a\}|;$$
$$d_{W_1}(Y, Z) := \sup_{\|f\|_{\text{Lip}} \le 1} |\mathbb{E}\, f(Y) - \mathbb{E}\, f(Z)|.$$

The purpose of this problem is to establish the comparison

$$d_{\text{Kol}}(Y, Z) \le \sqrt{2\mathrm{C}\, d_{W_1}(Y, Z)} \quad \text{for } \mathrm{C} = 1/\sqrt{2\pi}.$$

Thus, convergence in Kantorovich-1 distance implies convergence in Kolmogorov distance.

1. Explain why convergence in Kolmogorov distance implies convergence in distribution.
2. Fix $\varepsilon > 0$. For each $a \in \mathbb{R}$, define the piecewise linear function

$$h_a(x) = \begin{cases} 1, & x \le a \\ 1 - (x - a)/\varepsilon, & a < x \le a + \varepsilon \\ 0, & a + \varepsilon \le x. \end{cases}$$

   Observe that $\mathbb{1}\{x \le a\} \le h_a(x) \le \mathbb{1}\{x \le a + \varepsilon\}$ for all $x \in \mathbb{R}$. Compute $\|h_a\|_{\text{Lip}}$.
3. By adding and subtracting $\mathbb{E}\, h_a(Z)$, demonstrate that

$$\mathbb{E}\, \mathbb{1}\{Y \le a\} - \mathbb{E}\, \mathbb{1}\{Z \le a\} \le d_{W_1}(Y, Z)/\varepsilon + \mathrm{C}\varepsilon/2.$$

   Hint: C is the maximum value of the standard normal probability density function.

4. Develop the same inequality with Y and Z switched. Combine the two inequalities, select the optimal ε, and take the supremum over a to complete the proof.

Problem 17.43 (TV nation). For random variables X, Y that take values in $\mathbb{Z}_+$, the total variation metric $\mathrm{dist}_{\mathrm{TV}}$ is

$$\mathrm{dist}_{\mathrm{TV}}(X, Y) := \sup_{\mathsf{A} \subseteq \mathbb{Z}_+} |\mathbb{P}\{X \in \mathsf{A}\} - \mathbb{P}\{Y \in \mathsf{A}\}|.$$

1. Why does $\mathrm{dist}_{\mathrm{TV}}$ induce a metric on probability measures on $\mathbb{Z}_+$?
2. For random variables $(X_n : n \in \mathbb{N})$ taking values in $\mathbb{Z}_+$, explain why the limit $\mathrm{dist}_{\mathrm{TV}}(X_n, X) \to 0$ as $n \to \infty$ implies that $X_n \rightsquigarrow X$ weakly.
3. Establish the alternative formulation

$$\mathrm{dist}_{\mathrm{TV}}(X, Y) = \frac{1}{2} \sum_{k=0}^{\infty} |\mathbb{P}\{X = k\} - \mathbb{P}\{Y = k\}|.$$

   Hint: Consider the sets

$$\mathsf{A}_+ = \{k \in \mathbb{Z}_+ : \mathbb{P}\{X = k\} > \mathbb{P}\{Y = k\}\};$$
$$\mathsf{A}_- = \{k \in \mathbb{Z}_+ : \mathbb{P}\{X = k\} < \mathbb{P}\{Y = k\}\}.$$

4. Compute the TV distance between a $\textsc{geometric}(p)$ and a $\textsc{geometric}(q)$ distribution.
5. Prove that $\mathrm{dist}_{\mathrm{TV}}(X, Y) \le \mathbb{P}\{X \neq Y\}$. **Hint:** For each set A, the conditions $X \in \mathsf{A}$ and $Y \in \mathsf{A}^{\mathsf{c}}$ together imply that $X \neq Y$.
6. **(*)** Prove that there is a joint distribution of (X, Y) where equality holds in (5).
7. **(*)** We can define the TV metric for all random variables: $\mathrm{dist}_{\mathrm{TV}}(X, Y) := \sup_{\mathsf{B} \in \mathscr{B}(\mathbb{R})} |\mathbb{P}\{X \in \mathsf{B}\} - \mathbb{P}\{Y \in \mathsf{B}\}|$. Find a general representation of the TV metric similar with the result in (3).
8. **(**)** Given an upper bound for the TV distance between the $\textsc{normal}(m_1, v_1)$ and $\textsc{normal}(m_2, v_2)$ distributions.

Applications

Application 17.44 (Glivenko–Cantelli). In this application, we establish another classic result from statistics. We can uniformly approximate the distribution function of a real random variable from a sample. Let X be a real random variable with law μ_X and distribution function F_X. For simplicity, assume that F_X is continuous.

Let $(X_1, X_2, X_3, \dots)$ be an i.i.d. sample from μ_X. For $n \in \mathbb{N}$, define the (random) empirical measure and its (random) empirical distribution function

$$\mu_n = \frac{1}{n} \sum_{i=1}^{n} \delta_{X_i} \quad \text{and} \quad F_n(a) = \mu_n(-\infty, a] \quad \text{for } a \in R.$$

We can think of μ_n as an approximation of the measure μ_X from observed data. Meanwhile, F_n approximates the distribution function F_X.

1. Show that $F_n(a) = n^{-1} \sum_{i=1}^{n} \mathbb{1}\{X_i \le a\}$ is the distribution function of μ_n. For fixed $a \in \mathbb{R}$, explain why $F_n(a)$ is a real random variable. For a particular μ_X, sketch what μ_n and F_n might look like.
2. For a fixed $a \in \mathbb{R}$, confirm that $F_n(a) \to F_X(a)$ almost surely as $n \to \infty$.
3. For fixed $n \in \mathbb{N}$, show that there exist points $-\infty = t_0 < t_1 \le t_2 \le \cdots \le t_{n-1} < t_n = +\infty$ with the property that

$$F_X(t_j) = j/n \quad \text{for } j = 0, \dots, n.$$

4. For each fixed $a \in \mathbb{R}$, establish the bound

$$|F_n(a) - F_X(a)| \le \frac{1}{n} + \max\{|F_n(t_j) - F_X(t_j)| : j = 1, \dots, n-1\}.$$

   Hint: Isolate the index j where $t_{j-1} \le a < t_j$.
5. Conclude that

$$\sup_{a \in \mathbb{R}} |F_n(a) - F_X(a)| \to 0 \quad \text{almost surely as } n \to \infty.$$

6. **(*)** Show that the same conclusions are valid without the assumption that F_X is continuous. **Hint:** Find points t_j where $\mathbb{P}\{X < t_j\} \le j/n \le \mathbb{P}\{X \le t_j\}$ for each $j = 1, \dots, n$.
7. For probability measures μ, ν on the real line, define the *Kolmogorov metric*

$$\mathrm{dist}_{\mathrm{Kol}}(\mu, \nu) := \sup_{a \in \mathbb{R}} |\mu(-\infty, a] - \nu(-\infty, a]|.$$

   Why is $\mathrm{dist}_{\mathrm{Kol}}$ a metric? Why does convergence with respect to $\mathrm{dist}_{\mathrm{Kol}}$ imply weak convergence? Write the conclusion from (5) using the Kolmogorov metric.
8. **(**)** With notation as above, establish the convergence rate

$$\mathbb{P}\left\{\mathrm{dist}_{\mathrm{Kol}}(\mu_n, \mu) \ge t \cdot n^{-1/2}\right\} \le Cn \cdot \mathrm{e}^{-ct^2}$$

   for positive, absolute constants c, C. **Hint:** To get started, write $F_X(a) = \mathbb{E}\, F_n(a)$ and express the expectation in terms of an independent copy of the sample. Several other ideas are required, including symmetrization, a simple combinatorial estimate, and a concentration inequality.

Notes

The discussion of weak convergence is adapted from the books of Williams [Wil91], Dudley [Dud02], and Pollard [Pol02]. For more information about weak convergence, see Billingsley [Bil99] or van der Vaart & Wellner [VW23].

Lecture bibliography

[Bil99] P. Billingsley. *Convergence of probability measures*. Second. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1999. DOI: 10.1002/9780470316962.

[Dud02] R. M. Dudley. *Real analysis and probability*. Revised reprint of the 1989 original. Cambridge University Press, 2002. DOI: 10.1017/CBO9780511755347.

[Pol02] D. Pollard. *A user's guide to measure theoretic probability*. Cambridge University Press, 2002.

[VW23] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes—with applications to statistics*. Second. Springer, Cham, [2023] ©2023. DOI: 10.1007/978-3-031-29040-4.

[Wil91] D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

18. The Central Limit Theorem

Agenda:
1. Standardization
2. The central limit theorem
3. The Berry–Esséen theorem
4. Lindeberg universality
5. A quantitative CLT

"It is the supreme law of unreason."

—Sir Francis Galton

We have been studying the long-run behavior of the running averages of an independent and identically distributed sequence of real random variables. The law of large numbers states that the running average tends to the expectation almost surely. In Lecture 14, we saw evidence that the fluctuations of the running average around the expectation may converge to a limiting distribution. Today, we will prove that, with an appropriate scaling, the running average converges to a fixed distribution, and we will identify the limit as a normal random variable. This celebrated result is called the *central limit theorem*.

18.1 Standardization

Before we can prove a distributional limit theorem for the running averages, we need to agree on the right way to scale the averages so that they have a nontrivial limit.

As usual, let $Y \in \mathsf{L}_2$ be a fixed random variable, and draw a sequence $(Y_i : i \in \mathbb{N})$ of independent copies of Y. Form the running average $\bar{X}_n = n^{-1} \sum_{i=1}^{n} Y_i$ for each $n \in \mathbb{N}$.

Is there a sense in which $\bar{X}_n$ tends to a limiting distribution? Kolmogorov's SLLN states that

$$\bar{X}_n \to \mathbb{E}\, Y \quad \text{almost surely.}$$

The constant limit in the SLLN reflects the fact that the variance $\operatorname{Var}[\bar{X}_n]$ of the running average tends to zero with $n \to \infty$.

This observation suggests that we should rescale the running average so that its variance remains constant. To that end, we introduce the random variables

$$T_n := \sqrt{\frac{n}{\operatorname{Var}[Y]}} \cdot (\bar{X}_n - \mathbb{E}\, Y) \quad \text{for } n \in \mathbb{N}.$$

It is an easy exercise to check that

$$\mathbb{E}\, T_n = 0 \quad \text{and} \quad \operatorname{Var}[T_n] = 1 \quad \text{for each } n \in \mathbb{N}.$$

A random variable with these two properties is called *standardized*, and we call T_n the nth *standardized sum*.

The standardized sums have the same expectation and variance, so they are all comparable with each other. We will investigate whether there is a sense in which the standardized sums converge to a limit.

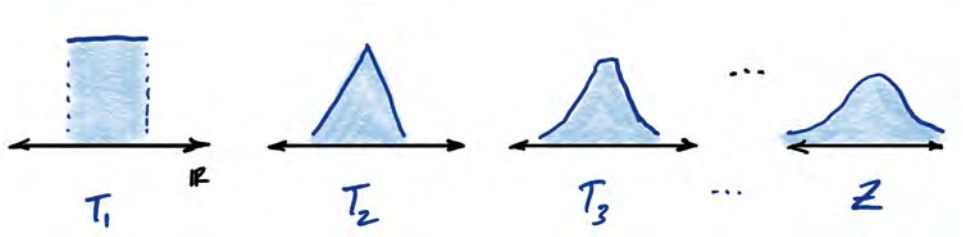


Figure 18.1 (CLT). Distributional convergence of the standardized sums to a normal distribution.

18.2 The distributional limit of standardized sums

It is, perhaps, too much to ask that the standardized sums converge as functions. Instead, we will ask whether the *distributions* of the standardized sums converge to a limiting distribution.

18.2.1 The central limit theorem

The first result states that the standardized sums converge weakly to a limiting distribution. Moreover, it identifies the weak limit as a normal distribution.

Theorem 18.1 (Central limit theorem (CLT)). Let $Y \in \mathsf{L}_2$ be a real random variable with expectation $m := \mathbb{E}\, Y$ and variance $\sigma^2 := \mathrm{Var}[Y] > 0$. Consider a sequence $(Y_i : i \in \mathbb{N})$ of i.i.d. copies of Y, and introduce the standardized sums

$$T_n := \frac{1}{\sqrt{n}} \sum_{i=1}^{n} \frac{Y_i - m}{\sigma} = \sqrt{n} \cdot \frac{\bar{X}_n - m}{\sigma}.$$

Let $Z \sim \textsc{normal}(0, 1)$ be a real standard normal random variable. Then

$$T_n \rightsquigarrow Z \quad \text{as } n \to \infty.$$

See Figure 18.1 for a schematic.

The main object of this lecture is to establish the CLT under slightly more generous assumptions.

To understand the result, we first recall that a standard normal random variable Z has the law

$$\gamma(\mathsf{B}) := \frac{1}{\sqrt{2\pi}} \int_{\mathsf{B}} \mathrm{e}^{-t^2/2}\, \lambda(\mathrm{d}t) \quad \text{for } \mathsf{B} \in \mathscr{B}(\mathbb{R}).$$

In other words, the standard normal distribution has density $(2\pi)^{-1/2}\mathrm{e}^{-t^2/2}$ with respect to the Lebesgue measure. The (cumulative) distribution function of the standard normal variable has its own notation:

$$\Phi(a) := \gamma(-\infty, a] = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} \mathrm{e}^{-t^2/2}\, \lambda(\mathrm{d}t).$$

By direct calculation, the expectation and variance of a standard normal variable are

$$\mathbb{E}\, Z = 0 \quad \text{and} \quad \mathrm{Var}[Z] = 1.$$

In other words, the standard normal variable Z is standardized. See Figure 18.2 for an illustration.

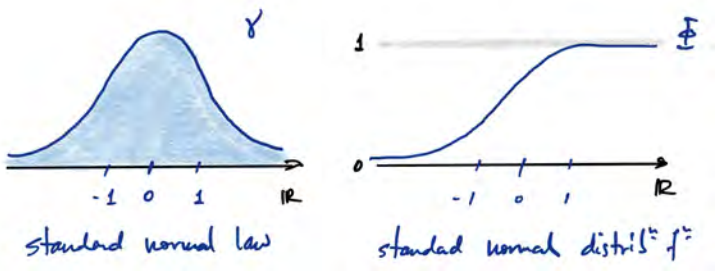


Figure 18.2 (Normal distribution). The law and distribution function of a standard normal random variable.

It is productive to rephrase what weak convergence means. Explicitly, weak convergence states that

$$\mathbb{E}\, h(T_n) \to \mathbb{E}\, h(Z) \quad \text{for each bounded, Lipschitz } h : \mathbb{R} \to \mathbb{R}.$$

In other words, every BL moment of T_n converges to the corresponding moment of a standard normal random variable Z. According to Theorem 17.22, we can rephrase weak convergence in terms of distribution functions:

$$\mathbb{P}\{T_n \le a\} = F_{T_n}(a) \to \Phi(a) = \mathbb{P}\{Z \le a\} \quad \text{for each } a \in \mathbb{R}.$$

Indeed, Φ is continuous on the real line.

Historically, the CLT played an important role in statistics to justify (asymptotic) normal confidence intervals for the sample average. This result has less importance now because we can use simulation methods (like the bootstrap) to obtain more accurate confidence intervals.

18.2.2 The Berry–Esséen theorem

You may recall that weak convergence is metrized by the BL distance (Theorem 17.12). As a consequence, we can also express the conclusion of the CLT as

$$\mathrm{dist}_{\mathrm{BL}}(T_n, Z) := \sup\{|\mathbb{E}\, h(T_n) - \mathbb{E}\, h(Z)| : \|h\|_{\mathrm{BL}} \le 1\} \to 0 \quad \text{as } n \to \infty.$$

The BL distance allows us to quantify how far the standardized sum T_n lies from the standard normal limit Z. Therefore, we can use the BL distance to quantify the rate of convergence in the CLT.

The next result provides a quantitative rate of convergence in the Kolmogorov metric, which is stronger than the BL metric.

Theorem 18.2 (Berry–Esséen). Let $Y \in \mathsf{L}_3$ be a real random variable with variance $\sigma^2 := \mathrm{Var}[Y] > 0$ and third central moment $M_3 := \mathbb{E}\,|Y - \mathbb{E}\, Y|^3$. Consider a sequence $(Y_i : i \in \mathbb{N})$ of i.i.d. copies of Y, and let T_n be the nth standardized sum, as in Theorem 18.1. Then, for each $n \in \mathbb{N}$,

$$\mathrm{dist}_{\mathrm{Kol}}(T_n, Z) := \sup_{a \in \mathbb{R}} |\mathbb{P}\{T_n \le a\} - \mathbb{P}\{Z \le a\}| \le \frac{M_3}{\sigma^3 \sqrt{n}}$$

In particular, the distribution functions of the standardized sums T_n converge

uniformly to the standard normal distribution function Φ.

Since the Kolmogorov metric is stronger than the BL metric, Theorem 18.2 implies that the standardized sums converge weakly to the standard normal distribution: $T_n \rightsquigarrow Z$ as $n \to \infty$.

Note that the Berry–Esséen theorem has slightly stronger hypotheses than the CLT. Indeed, the CLT holds when the underlying random variable $Y \in \mathsf{L}_2$, while the Berry–Esséen theorem requires that $Y \in \mathsf{L}_3$. In exchange, we obtain an explicit estimate for the rate at which the distribution functions of the standardized sums converge to the standard normal distribution function. The rate of convergence $n^{-1/2}$ specified in Theorem 18.2 is optimal.

We will prove a version of the Berry–Esséen theorem under the same hypotheses ($Y \in \mathsf{L}_3$). Our approach controls the BL metric rather than the Komogorov metric, and it only yields the suboptimal rate $n^{-1/6}$. Nevertheless, this argument is more than enough to deduce weak convergence. The proof here is based on a beautiful method due to Lindeberg (1922), which has found many new applications in the last 15 years.

Aside: The usual approach to the CLT and to the Berry–Esséen theorem is to show that the characteristic functions (that is, the Fourier transforms) of the distributions converge. This method is elegant, but it requires a significant amount of extra work to connect convergence in distribution to convergence of characteristic functions. See Lecture 21 for more information about this approach.

18.3 Lindeberg's universality principle

The key idea behind Lindeberg's proof of the central limit theorem is that we can estimate moments associated with smooth test functions by means of Taylor's theorem. That is,

$$\mathbb{E}\, h(T_n) \approx \mathbb{E}\, h(Z) \quad \text{when } h \text{ is sufficiently smooth.}$$

Lindeberg uses this observation to show that only the first and second moments of a distribution affect moments defined by smooth test functions. The fine structure of the distribution does not play a role. We refer to this fact as a *universality phenomenon*.

18.3.1 Universality for univariate functions

We begin with an elementary lemma, which shows how to control the discrepancy between the moments of two random variables with the same mean and variance.

Lemma 18.3 (Universality: Univariate case). Consider real random variables $Y, Z \in \mathsf{L}_3$ that have the same expectation and variance:

$$\mathbb{E}\, Y = \mathbb{E}\, Z \quad \text{and} \quad \operatorname{Var}[Y] = \operatorname{Var}[Z].$$

In this section, random variables Z and Z_i are not necessarily standard normal. Nevertheless, we have chosen this notation because Z will eventually play the role of the normal variable in the CLT.

For each function $h : \mathbb{R} \to \mathbb{R}$ with three bounded derivatives, we have the moment bound

$$|\mathbb{E}\, h(Y) - \mathbb{E}\, h(Z)| \le \tfrac{1}{6} \|h'''\|_{\sup} \big(\mathbb{E}\, |Y|^3 + \mathbb{E}\, |Z|^3 \big).$$

This is a basic universality result, which shows that moments with respect to a smooth test function depend primarily on the expectation and variance. Other features of the random variables are irrelevant.

Proof. Let $h : \mathbb{R} \to \mathbb{R}$ be a function with three bounded derivatives. By Taylor's theorem with remainder,

$$\begin{aligned}\left|h(Y) - h(0) - h'(0) \cdot Y - \tfrac{1}{2}h''(0) \cdot Y^2\right| &\le \tfrac{1}{6}\|h'''\|_{\sup} \cdot |Y|^3; \\ \left|h(Z) - h(0) - h'(0) \cdot Z - \tfrac{1}{2}h''(0) \cdot Z^2\right| &\le \tfrac{1}{6}\|h'''\|_{\sup} \cdot |Z|^3.\end{aligned}$$

Combining these two expressions, we can bound the discrepancy between the moment of Y and the moment of Z:

$$\begin{aligned}|\mathbb{E}[h(Y) - h(Z)]| \le \left|\mathbb{E}\left[h'(0) \cdot (Y - Z) + \tfrac{1}{2}h''(0) \cdot (Y^2 - Z^2)\right]\right| \\ + \tfrac{1}{6}\|h'''\|_{\sup} \cdot \left(\mathbb{E}\,|Y|^3 + \mathbb{E}\,|Z|^3\right).\end{aligned}$$

The contribution in the first line on the right-hand side vanishes because our assumption ensures that $\mathbb{E}\,Y = \mathbb{E}\,Z$ and $\mathbb{E}\,Y^2 = \mathbb{E}\,Z^2$. This is the stated result. ∎

18.3.2 Universality for multivariate functions

The next step in the proof is to extend Lemma 18.3 to the case of a multivariate function of independent random variables.

Theorem 18.4 (Lindeberg universality). Let $(Y_1, \dots, Y_n)$ and $(Z_1, \dots, Z_n)$ be mutually independent random variables that belong to L_3. Assume that

$$\mathbb{E}\,Y_i = \mathbb{E}\,Z_i \quad \text{and} \quad \operatorname{Var}[Y_i] = \operatorname{Var}[Z_i] \quad \text{for each } i = 1, \dots, n.$$

For each function $f : \mathbb{R}^n \to \mathbb{R}$ that satisfies $\|\partial_{iii} f\|_{\sup} < +\infty$ for each index i, we have the bound

$$|\mathbb{E}[f(Y_1, \dots, Y_n) - f(Z_1, \dots, Z_n)]| \le \frac{1}{6}\sum\nolimits_{i=1}^{n} \|\partial_{iii} f\|_{\sup}\left(\mathbb{E}\,|Y_i|^3 + \mathbb{E}\,|Z_i|^3\right).$$

As usual, ∂_{iii} denotes the third partial derivative with respect to the ith coordinate.

Proof. The proof is based on the *Lindeberg exchange argument*. We can interpolate between the values $f(Y_1, \dots, Y_n)$ and $f(Z_1, \dots, Z_n)$ by swapping one coordinate at a time. Lemma 18.3 allows us to control the error we incur at each step.

First, it is convenient to abbreviate lists of random variables using a vector notation:

$$\boldsymbol{y} := (Y_1, \dots, Y_n) \quad \text{and} \quad \boldsymbol{z} := (Z_1, \dots, Z_n).$$

Introduce the interpolating vectors

$$\boldsymbol{w}^i := (Y_1, \dots, Y_i, Z_{i+1}, \dots, Z_n) \quad \text{for each } i = 0, 1, 2, \dots, n.$$

Observe that $\boldsymbol{w}^0 = \boldsymbol{z}$ and $\boldsymbol{w}^n = \boldsymbol{y}$.

To compute the difference between the expectations, we use a telescoping sum:

$$\begin{aligned}\mathbb{E}\left[f(\boldsymbol{y}) - f(\boldsymbol{z})\right] &= \sum\nolimits_{i=1}^{n} \mathbb{E}\left[f(\boldsymbol{w}^i) - f(\boldsymbol{w}^{i-1})\right] \\ &= \sum\nolimits_{i=1}^{n} \mathbb{E}\,\mathbb{E}_i\left[f(\boldsymbol{w}^i) - f(\boldsymbol{w}^{i-1})\right].\end{aligned}$$

We temporarily use the notation $\mathbb{E}_i$ to denote the expectation with respect to Y_i and Z_i only. Since all the random variables are independent, Fubini's theorem (Theorem 6.23)

allows us to factor the expectation. We may bound the absolute value using Jensen's inequality (Theorem 9.26):

$$\left|\mathbb{E}\left[f(\boldsymbol{y}) - f(\boldsymbol{z})\right]\right| \le \sum\nolimits_{i=1}^{n} \mathbb{E}\left|\mathbb{E}_i\left[f(\boldsymbol{w}^i) - f(\boldsymbol{w}^{i-1})\right]\right|.$$

It remains to control each summand by an application of the lemma.

Explicitly, for each index i, we introduce the univariate function

$$h_i(t) := f(Y_1, \dots, Y_{i-1}, t, Z_{i+1}, \dots, Z_n) \quad \text{for } t \in \mathbb{R}.$$

Then Lemma 18.3 implies that

$$\begin{aligned}\left|\mathbb{E}_i\left[f(\boldsymbol{w}^i) - f(\boldsymbol{w}^{i-1})\right]\right| &= \left|\mathbb{E}_i\left[h_i(Y_i) - h_i(Z_i)\right]\right| \\ &\le \tfrac{1}{6}\|h_i'''\|_{\sup} \cdot \left(\mathbb{E}_i|Y_i|^3 + \mathbb{E}_i|Z_i|^3\right).\end{aligned}$$

Finally, note that $h_i''' = \partial_{iii} f$, and combine the results. ∎

18.4 A quantitative CLT

In this section, we show how the Lindeberg universality principle applies to standardized sums. From here, it is a short step to a quantitative version of the central limit theorem, expressed in term of an unusual probability metric. Afterward, we show that this result leads to suboptimal versions of the Berry–Esséen theorem and the CLT.

18.4.1 Universality for standardized sums

Lindeberg's universality principle can be specialized to the case of a standardized sum, as follows.

Corollary 18.5 (Universality: Standardized sums). Consider real random variables $Y, Z \in \mathsf{L}_3$ with

$$\mathbb{E}\,Y = \mathbb{E}\,Z = 0 \quad \text{and} \quad \operatorname{Var}[Y] = \operatorname{Var}[Z] = 1.$$

Introduce a bound for the third moments: $M := (\mathbb{E}|Y|^3) \vee (\mathbb{E}|Z|^3)$. Let $(Y_1, \dots, Y_n)$ be i.i.d. copies of Y, and let $(Z_1, \dots, Z_n)$ be i.i.d. copies of Z. For any function $h : \mathbb{R} \to \mathbb{R}$ with three bounded derivatives,

$$\left|\mathbb{E}\, h\Big(\tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n Y_i\Big) - \mathbb{E}\, h\Big(\tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n Z_i\Big)\right| \le \tfrac{1}{3} M \|h'''\|_{\sup} \cdot n^{-1/2}.$$

Proof. Define the multivariate function

$$f(x_1, \dots, x_n) := h\Big(\tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n x_i\Big) \quad \text{for } (x_1, \dots, x_n) \in \mathbb{R}^n.$$

By the chain rule, $\|\partial_{iii} f\|_{\sup} = n^{-3/2}\|h'''\|_{\sup}$. Thus, Theorem 18.4 implies that

$$\begin{aligned}\left|\mathbb{E}\left[f(Y_1, \dots, Y_n) - f(Z_1, \dots, Z_n)\right]\right| &\le \frac{1}{6} \sum\nolimits_{i=1}^{n} n^{-3/2}\|h'''\|_{\sup} \cdot (\mathbb{E}|Y_i|^3 + \mathbb{E}|Z_i|^3) \\ &\le \frac{1}{3} M \|h'''\|_{\sup} \cdot n^{-1/2}.\end{aligned}$$

We have used the fact that the Y_i are i.i.d. and the Z_i are i.i.d. to bound the parenthesis by $2M$. Note that the factor $n^{-3/2}$ emerges when we take the third derivative of the standardized sum, and this quantity is small enough to counteract the number n of terms in the sum. ∎

Alternatively, we can express the conclusion of Corollary 18.5 in terms of an integral probability metric. This is just a matter of reinterpretation.

Corollary 18.6 (Universality: Standardized sums). Instate the hypotheses of Corollary 18.5. Let $\mathsf{H} := \{h : \mathbb{R} \to \mathbb{R} : \|h'''\|_{\text{sup}} \leq 1\}$. Then

$$\text{dist}_{\mathsf{H}}\left(\tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n Y_i,\ \tfrac{1}{\sqrt{n}} \textstyle\sum_{i=1}^n Z_i\right) \leq \tfrac{1}{3} M \cdot n^{-1/2}.$$

What does Corollary 18.6 mean? It tells us that the distribution of a standardized sum depends primarily on the first two moments. Indeed, for any distributions Y and Z, standardizing the sums has the effect of shifting the distributions to have mean zero and scaling them to have variance one. Thus, the standardized sum of n i.i.d. copies of Y is very close the standardized sum of n i.i.d. copies of Z. We can express this result as saying that the distribution of a standardized sum is universal, given the first two moments.

For this proof, it is exceedingly natural to use the integral probability metric based on the family of test functions $\mathsf{H} = \{h : \|h'''\|_{\text{sup}} \leq 1\}$. On the other hand, this distance is somewhat unusual, and it is not obvious what it signifies. In Section 18.4.5, we will see how to pass from dist_{H} to the bounded, Lipschitz distance dist_{BL}.

18.4.2 The normal limit

Corollary 18.6 tells us that the standardized sum of i.i.d. random variables depends primarily on the first two moments. In other words, all the standardized sums are close to each other. But are they close to some particular distribution? The answer is a resounding "yes."

To find the particular distribution, we want to look for a random variable Z that generates standardized sums that we can treat analytically. Ideally, we can find a random variable Z where the standardized sums all have the same distribution. Let us confirm that the standard normal distribution fits the bill.

Exercise 18.7 (Standard normal variable: Standardized sums). Let $Z_1, \dots, Z_n$ be independent standard normal random variables. Show that the standardized sum $n^{-1/2} \sum_{i=1}^n Z_i$ follows the standard normal distribution. **Hint:** Generalize the calculation in Exercise 14.5(2), and use induction.

18.4.3 Our first CLT

With these results in place, we can easily derive a quantitative central limit theorem from Corollary 18.6.

Theorem 18.8 (Quantitative CLT). Instate the notation and hypotheses of Theorem 18.2. For the test functions $\mathsf{H} := \{h : \mathbb{R} \to \mathbb{R} : \|h'''\|_{\text{sup}} \leq 1\}$, we have

$$\text{dist}_{\mathsf{H}}(T_n, Z) \leq \frac{M_3}{\sigma^3 \sqrt{n}} \quad \text{for each } n \in \mathbb{N}.$$

Proof. We may assume that Y is centered and scaled so that $\sigma^2 = \text{Var}[Y] = 1$. By monotonicity of moments (Theorem 11.4), note that $M_3 := \mathbb{E}\,|Y|^3 \geq 1$. Let Z be a standard normal random variable.

Instantiate Corollary 18.6. According to Exercise 18.7, the standardized sum $n^{-1/2} \sum_{i=1}^n Z_i$ follows the standard normal distribution Z. By direct calculation, $\mathbb{E}\,|Z|^3 < 2$. We see that the third moment bound

$$M := (\mathbb{E}\,|Y|^3) \vee (\mathbb{E}\,|Z|^3) \leq M_3 \vee 2 \leq 2M_3.$$

The result follows once we correct for the scaling of Y. ∎

18.4.4 *From Lipschitz functions to smooth functions

To derive the CLT from Theorem 18.8, we still have to take another step. The remaining issue is that we established a rate of convergence with respect to a distance $\mathrm{dist}_{\mathsf{H}}$ that is not obviously comparable with $\mathrm{dist}_{\mathrm{BL}}$, which is the metric associated with weak convergence. Therefore, we need to change metrics.

This project requires us to approximate a bounded, Lipschitz function by a function with three bounded derivatives. To do so, we average the local values of the function by convolving it with a Gaussian distribution. This is a very useful technique in probability theory and analysis, so we include the details.

Lemma 18.9 (Gaussian smoothing). Let $h : \mathbb{R} \to \mathbb{R}$ be a bounded, Lipschitz function. For a parameter $\sigma > 0$, define the smoothed function

$$h_\sigma(a) := \mathbb{E}\big[h(a + \sigma Z)\big] \quad \text{for each } a \in \mathbb{R},$$

where $Z \sim \textsc{normal}(0, 1)$ is a standard normal random variable. Then h_σ approximates h pointwise:

$$\|h_\sigma - h\|_{\sup} < \sigma \cdot \|h\|_{\mathrm{Lip}}.$$

Furthermore, the approximation h_σ and all its derivatives are bounded:

$$\|h_\sigma\|_{\sup} \le \|h\|_{\sup} \quad \text{and} \quad \|\mathrm{D}^k h_\sigma\| \le \sigma^{-(k-1)} \cdot \|h\|_{\mathrm{Lip}} \quad \text{for } k \in \mathbb{N}.$$

Proof. To see that h_σ approximates h, we use the fact that h is Lipschitz:

$$\begin{aligned} |h_\sigma(a) - h(a)| &\le \mathbb{E}\,|h(a + \sigma Z) - h(a)| \\ &\le \mathbb{E}\big[|\sigma Z| \cdot \|h\|_{\mathrm{Lip}}\big] = \sqrt{\frac{2}{\pi}} \cdot \sigma \cdot \|h\|_{\mathrm{Lip}}. \end{aligned}$$

Indeed, $\mathbb{E}\,|Z| = \sqrt{2/\pi} < 1$.

Next, we will confirm that h_σ is differentiable. Fix a point $a \in \mathbb{R}$. By the law of the unconscious statistician,

$$h_\sigma(a) = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} h(a + \sigma z) \cdot \mathrm{e}^{-z^2/2}\,\mathrm{d}z = \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} h(u) \cdot \mathrm{e}^{-(u-a)^2/(2\sigma^2)}\,\frac{\mathrm{d}u}{\sigma}. \tag{18.1}$$

We have made the change of variables $a + \sigma z \mapsto u$. By dominated convergence, we can compute h_σ' by passing the derivative through the integral sign:

$$\begin{aligned} h_\sigma'(a) &= \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} h(u) \cdot \frac{-(u-a)}{\sigma^2}\,\mathrm{e}^{-(u-a)^2/(2\sigma^2)}\,\frac{\mathrm{d}u}{\sigma} \\ &= \frac{1}{\sigma\sqrt{2\pi}} \int_{\mathbb{R}} h(a + \sigma z) \cdot (-z)\mathrm{e}^{-z^2/2}\,\mathrm{d}z \\ &= \frac{1}{\sigma} \cdot \mathbb{E}\big[h(a + \sigma Z) \cdot (-Z)\big]. \end{aligned}$$

We have reversed the change of variables: $u \mapsto a + \sigma z$. Since h is bounded, it is easily seen that $h_\sigma'(a)$ is finite on the entire real line.

Since h_σ is differentiable, Exercise 17.6 ensures that $\|h_\sigma'\|_{\sup} = \|h\|_{\mathrm{Lip}}$. To bound the remaining derivatives, we exploit the calculations in the last paragraph. For each

$k \in \mathbb{N}$, assume that the smoothed function h_σ has $k-1$ bounded derivatives. Then we may calculate that

We write D for the derivative, and we write D_a for the (partial) derivative with respect to the variable a for emphasis.

$$(\mathrm{D}^k h_\sigma)(a) = \mathrm{D}_a^k \, \mathbb{E}\left[h(a+\sigma Z)\right] = \mathrm{D}_a \, \mathbb{E}\left[(\mathrm{D}^{k-1}h)(a+\sigma Z)\right].$$

Following the same calculation as before,

$$\mathrm{D}_a \, \mathbb{E}[(\mathrm{D}^{k-1}h)(a+\sigma Z)] = \sigma^{-1} \, \mathbb{E}[(\mathrm{D}^{k-1}h)(a+\sigma Z) \cdot (-Z)].$$

Therefore,

$$\|\mathrm{D}^k h_\sigma\|_{\sup} \le \sqrt{\frac{2}{\pi}} \cdot \sigma^{-1} \cdot \|\mathrm{D}^{k-1} h_\sigma\|_{\sup}.$$

Using induction, we determine that

$$\|\mathrm{D}^k h_\sigma\|_{\sup} \le (2/\pi)^{(k-1)/2} \cdot \sigma^{-(k-1)} \cdot \|h\|_{\mathrm{Lip}}.$$

This is the required result. ∎

18.4.5 Berry–Esséen with suboptimal rate

Finally, we can establish a version of the Berry–Esséen theorem.

Theorem 18.10 (Berry–Esséen: Suboptimal version). Instate the notation and hypotheses of Theorem 18.2. Then

$$\mathrm{dist}_{\mathrm{BL}}(T_n, Z) \le \frac{3M_3^{1/3}}{\sigma n^{1/6}} \quad \text{for each } n \in \mathbb{N}.$$

In particular, $T_n \rightsquigarrow Z$ as $n \to \infty$.

Proof. We need to obtain a uniform bound for moments determined by bounded Lipschitz functions. Choose a function $h : \mathbb{R} \to \mathbb{R}$ with $\|h\|_{\mathrm{BL}} \le 1$. Fix a parameter $\sigma > 0$, to be determined later. Then Lemma 18.9 promises us a function h_σ that satisfies

$$\|h_\sigma - h\|_{\sup} \le \sigma \quad \text{and} \quad \|h_\sigma'''\|_{\sup} \le \sigma^{-2}.$$

We can bound moments associated with h by passing to the function h_σ.

Compare h with h_σ using the uniform approximation bound:

$$|\mathbb{E}[h(T_n) - h(Z)]| \le |\mathbb{E}[h_\sigma(T_n) - h_\sigma(Z)]| + 2\sigma \le \frac{1}{\sigma^2}\mathrm{dist}_{\mathsf{H}}(T_n, Z) + 2\sigma.$$

We have used the fact that the function $\sigma^2 h_\sigma \in \mathsf{H}$. Now, the right-hand side is minimized when $\sigma^3 = \mathrm{dist}_{\mathsf{H}}(T_n, Z)$. Therefore,

$$\mathrm{dist}_{\mathrm{BL}}(T_n, Z) = \sup\{|\mathbb{E}[h(T_n) - h(Z)]| : \|h\|_{\mathrm{BL}} \le 1\} \le 3\,\mathrm{dist}_{\mathsf{H}}(T_n, Z)^{1/3}.$$

Apply Theorem 18.8 to complete the argument. ∎

Problem 18.11 (Berry–Esséen: Non-identical summands). The proof of Theorem 18.10 does not depend strongly on the assumption that the summands are i.i.d. Formulate and prove a version of this result for an arbitrary independent sum of real random variables.

Problems

Problem 18.12 (The Law of Small Numbers).** The central limit theorem applies, in particular, to a binomial random variable $X_n \sim \textsc{binomial}(p, n)$. It is well known that

$$\mathbb{E}\, X_n = np \quad \text{and} \quad \operatorname{Var}[X_n] = np(1-p).$$

For fixed $p \in (0,1)$, the central limit theorem implies that $n^{-1/2}(X_n - np) \rightsquigarrow \textsc{normal}(0, p(1-p))$.

We can also consider another limit of binomial random variables, where the probability p_n varies with n in a way that $np_n = \lambda$, a constant. In this setting, the central limit theorem is not valid. Instead, $\textsc{binomial}(p_n, n)$ converges to a Poisson random variable with expectation λ. This result is called the *law of small numbers*.

In this problem, we outline the steps in a proof of a strengthened form of the law of small numbers. As it happens, we can simply modify the Lindeberg method.

1. Define the forward difference operator $(\Delta_+ h)(k) := h(k+1) - h(k)$ on functions $h : \mathbb{Z}_+ \to \mathbb{R}$. For $i \in \mathbb{Z}_+$, show that

$$|h(k) - h(0) - (\Delta_+ h)(0) \cdot q - \tfrac{1}{2}(\Delta_+^2 h)(0) \cdot k(k-1)| \\ \leq \tfrac{1}{6}\|\Delta_+^3 h\|_{\sup} k(k-1)(k-2).$$

2. Suppose that Y, Q are random variables that take values in $\mathbb{Z}_+$, with $\mathbb{E}\, Y = \mathbb{E}\, Q$ and $|\operatorname{Var}[Y] - \operatorname{Var}[Q]| \leq \varepsilon$. Show that

$$|\mathbb{E}[h(Y) - h(Q)]| \leq \tfrac{1}{2}\|\Delta_+^2 h\|_{\sup} \cdot \varepsilon + \tfrac{1}{6}\|\Delta_+^3 h\|_{\sup}\big(\mathbb{E}\, Y^3 + \mathbb{E}\, Q^3\big).$$

3. Suppose that Y_i, Q_i are independent random variables with $\mathbb{E}\, Y_i = \mathbb{E}\, Q_i$ and $|\operatorname{Var}[Y_i] - \operatorname{Var}[Q_i]| \leq \varepsilon_i$. For each function $f : \mathbb{Z}_+^n \to \mathbb{R}$, show that

$$|\mathbb{E}[f(Y_1, \dots, Y_n) - f(Q_1, \dots, Q_n)]| \\ \leq \sum\nolimits_{i=1}^n \left[\tfrac{1}{2}\|\Delta_{i+}^2 f\|_{\sup} \cdot \varepsilon_i + \tfrac{1}{3}\|\Delta_{i+}^3 f\|_{\sup} \cdot M_i\right],$$

   where Δ_{i+} is the forward difference in the ith coordinate, and the third moment $M_i := (\mathbb{E}\, Y_i^3) \vee (\mathbb{E}\, Q_i^3)$.
4. Consider independent Poisson random variables $Q_i \sim \textsc{poisson}(p_i)$. Confirm that $\sum_{i=1}^n Q_i \sim \textsc{poisson}(\sum_{i=1}^n p_i)$.
5. Consider the case where $Y_i \sim \textsc{bernoulli}(p_i)$ and $Q_i \sim \textsc{poisson}(p_i)$ for each index i. Then apply the result from the (3) to the function $f(x_1, \dots, x_n) = \sum_{i=1}^n x_i$.
6. Deduce a quantitative version of the law of small numbers for an independent sum of Bernoulli random variables with arbitrary means p_i. Express the result in terms of the probability metric generated by the function class

$$\mathsf{H} := \{h : \mathbb{Z}_+ \to \mathbb{R} : \|\Delta_+^2 h\|_{\sup} \leq 1 \text{ and } \|\Delta_+^3 h\|_{\sup} \leq 1\}.$$

   Under what assumptions on the probabilities p_i is the Bernoulli sum approximated by a Poisson random variable?
7. We can smooth a bounded function on the integers to obtain a nearby function with bounded forwarded differences. Let $h : \mathbb{Z}_+ \to \mathbb{R}$ be uniformly bounded. For a parameter $\lambda > 0$, define the smoothed function

$$h_\lambda(k) = \mathbb{E}[h(k + Q_\lambda)] \quad \text{where} \quad Q_\lambda \sim \textsc{poisson}(\lambda).$$

   Show that h_λ is bounded. Obtain uniform bounds on the forward differences $\|\Delta_+^j h\|_{\sup}$ for each $j \in \mathbb{N}$.

8. The total variation distance on probability measures μ, ν supported on $\mathbb{Z}_+$ can be defined as

$$\operatorname{dist}_{\mathrm{TV}}(\mu, \nu) := \sup\{|\mu(h) - \nu(h)| : \|h\|_{\sup} \leq 1\}.$$

Show how to extend your law of small numbers from (6) to the TV distance. Deduce a limit theorem.

Applications

Application 18.13 (Normal confidence intervals). The central limit theorem provides a heuristic method for using data to estimate an interval that contains the expectation of a random variable. Consider an i.i.d. sequence $(Y_i : i \in \mathbb{N})$ of independent observations of $Y \in \mathsf{L}_2$. For $n \in \mathbb{N}$, we can calculate the sample average $\bar{X}_n := n^{-1} \sum_{i=1}^n Y_i$ and the sample variance $\hat{\sigma}_n^2 := (n-1)^{-1} \sum_{i=1}^n (Y_i - \bar{X}_n)^2$ from the observations.

1. Informally, explain why $\bar{X}_n \approx \mathbb{E}\, Y$ and $\hat{\sigma}_n^2 \approx \operatorname{Var}[Y]$ for large n. **Hint:** Add and subtract $\mathbb{E}\, Y$ in the definition of $\hat{\sigma}_n^2$, and expand the square. What is $\operatorname{Cov}(Y_i, \bar{X}_n)$?
2. Let $Z \sim \textsc{normal}(0, 1)$. For fixed $t > 0$ and large n, informally justify the approximation

$$\mathbb{P}\left\{\bar{X}_n - t \cdot \hat{\sigma}_n \cdot n^{-1/2} \leq \mathbb{E}\, Y \leq \bar{X}_n + t \cdot \hat{\sigma}_n \cdot n^{-1/2}\right\} \approx \mathbb{P}\{-t \leq Z \leq +t\}.$$

In words, explain what this formula means.
3. To find an interval containing $\mathbb{E}\, Y$ with probability $\approx 95\%$, how do we pick t? What about $\approx 99\%$?
4. For $Y \in \mathsf{L}_3$, per Berry–Esséen, how large does n need to be for these approximations to be reasonable?

Application 18.14 (*Bootstrap). As computers became more powerful and more widely available in the 1960s and 1970s, statisticians began to develop inference procedures based on computer simulation. In this application problem, we explore the simplest example of the bootstrap methodology for producing data-driven confidence intervals. The bootstrap was invented by Brad Efron (a Caltech alumnus!).

Let $Y \in \mathsf{L}_3$ be a real random variable with law μ_Y and distribution function F_Y. It is convenient to write $m = \mathbb{E}\, Y$ and $v = \operatorname{Var}[Y]$ and $s = \mathbb{E}\,|Y - m|^3$, the third central moment. Suppose that we acquire an i.i.d. sample $(Y_1, \dots, Y_n)$ from μ_Y. We would like to estimate the mean m and provide a data-driven confidence interval for m. The bootstrap can be used for many other estimation problems, but the justification becomes (even) harder.

1. Define the sample average estimator $\bar{X}_n = n^{-1} \sum_{i=1}^n Y_i$. This is a random variable. Use the Berry–Esséen theorem (Lecture 18) to explain why the sampling distribution of $\bar{X}_n$ satisfies

$$\operatorname{dist}_{\mathrm{Kol}}\left(\sqrt{n}\,(\bar{X}_n - m), \sqrt{v}\, Z\right) \leq \frac{s}{v^{3/2}\sqrt{n}} \quad \text{for each } n \in \mathbb{N}.$$

As usual, Z is a standard normal random variable.
2. Fix the sample $\boldsymbol{Y} = (Y_1, \dots, Y_n)$. Define the (nonrandom) empirical measure $\mu_n = n^{-1} \sum_{i=1}^n \delta_{Y_i}$. Let m_n^* be the mean and v_n^* the variance and s_n^* the third central moment of the empirical measure μ_n. Write down simple formulas to calculate these quantities in terms of the fixed sample $\boldsymbol{Y}$.

3. Now, suppose that we draw a new random sample $(Y_1^*, \dots, Y_n^*)$ i.i.d. from the empirical measure μ_n. This is called a *bootstrap sample*. We can form the sample average of the bootstrap sample: $\bar{X}_n^* = n^{-1} \sum_{i=1}^n Y_i^*$. This is a random variable. Explain why the (conditional) sampling distribution of $\bar{X}_n^* \,|\, \boldsymbol{Y}$ satisfies

$$\operatorname{dist}_{\mathrm{Kol}}\big(\sqrt{n}\,(\bar{X}_n^* - m_n^*) \,|\, \boldsymbol{Y}, \sqrt{v_n^*}\,Z \,|\, \boldsymbol{Y}\big) \le \frac{s_n^*}{(v_n^*)^{3/2}\sqrt{n}} \quad \text{for each } n \in \mathbb{N}.$$

   In this problem, we are not using conditioning in any substantive way. The notation is just intended to remind you that the sample $\boldsymbol{Y}$ was drawn at random. But you should think about $\boldsymbol{Y}$ as frozen, so you can treat it as a nonrandom quantity.
4. Using the triangle inequality, deduce that the bootstrap sample average distribution serves as a proxy for the sample average distribution:

$$\operatorname{dist}_{\mathrm{Kol}}\big(\sqrt{n}\,(\bar{X}_n^* - m_n^*) \,|\, \boldsymbol{Y}, \sqrt{n}\,(\bar{X}_n - m)\big) \le \frac{2s}{v^{3/2}\sqrt{n}} + \operatorname{error}(\boldsymbol{Y}).$$

   Find a formula or bound for $\operatorname{error}(\boldsymbol{Y})$ to show that it is a function of the fixed sample $\boldsymbol{Y}$.
5. **(*)** With respect to the randomness in the sample $\boldsymbol{Y} = (Y_1, \dots, Y_n)$, argue that $\operatorname{error}(\boldsymbol{Y}) \to 0$ almost surely as $n \to \infty$. **Hint:** Use a Taylor expansion to compare two normal cdfs, and invoke the SLLN. **(**)** A fortiori, prove that $\mathbb{P}\left\{|\operatorname{error}(\boldsymbol{Y})| \ge \text{Const} \cdot n^{-1/2}\right\} \to 0$ as $n \to \infty$.
6. Let F^* be the exact distribution function of the random variable $\sqrt{n}\,(\bar{X}_n^* - m_n^*)$. For a fixed confidence level $\alpha \in (0, 0.5)$, identify points where $F^*(t_\alpha) \approx \alpha$ and $F^*(t_{1-\alpha}) \approx 1 - \alpha$. Then we can define the bootstrap confidence interval for the population mean:

$$I^* = [m_n^* - n^{-1/2}\, t_{1-\alpha},\, m_n^* - n^{-1/2}\, t_\alpha].$$

   Use (b) and (d) to argue informally that the coverage probability $\mathbb{P}\{m \in I^*\} \approx 1 - 2\alpha$. The interval I^* is a prototype for our data-driven confidence interval.
7. Unfortunately, we do not have direct access to the bootstrap distribution function F^*. Instead, we must resort to simulation. For a large parameter B, let $W_1, \dots, W_B$ be independent copies of the random variable $\sqrt{n}\,(\bar{X}_n^* - m_n^*)$. These random variables are called *bootstrap replicates*, and they are accessible to us. Let F_B^* be the empirical distribution function of $W_1, \dots, W_B$. Use Application 17.44 to explain why F_B^* serves in place of F^* as $B \to \infty$. Explain how to use F_B^* to construct a confidence interval I_B^* where $\mathbb{P}\left\{m \in I_B^*\right\} \approx 1 - 2\alpha$.

Notes

Lindeberg's proof of the central limit theorem dates to 1922. Our presentation is adapted from Pollard's book [Polo2, Sec. 7.2]. Although this approach is both simple and classic, it has not achieved the same visibility as other proof strategies. In particular, most books use characteristic functions to establish the CLT; see Lecture 21 for additional information.

In recent decades, the Lindeberg principle has been revived because it has many other elegant applications. In particular, see the papers of Chatterjee [Chao8; Chao6], Korada & Montanari [KM11], and Tao [Tao19]. We have also borrowed some ideas from these treatments.

The bootstrap application of the Berry–Esséen theorem is adapted from Wasserman's book [Waso4].

The quotation that opens the chapter is attributed to Sir Francis Galton (1822–1911). Galton made major contributions to probability theory, statistics, genetics, psychology, meteorology, and other fields. He also has a much darker legacy as the leading proponent of "scientific eugenics". The eugenics movement inspired some of the greatest evils of the 20th century.

Lecture bibliography

[Chao8] S. Chatterjee. "A simple invariance theeorem". Available at https://arxiv.org/abs/math/0508213. 2008.

[Chao6] S. Chatterjee. "A generalization of the Lindeberg principle". In: *Ann. Probab.* 34.6 (2006), pages 2061–2076. DOI: 10.1214/009117906000000575.

[KM11] S. B. Korada and A. Montanari. "Applications of the Lindeberg principle in communications and statistical learning". In: *IEEE Trans. Inform. Theory* 57.4 (2011), pages 2440–2450. DOI: 10.1109/TIT.2011.2112231.

[Polo2] D. Pollard. *A user's guide to measure theoretic probability*. Cambridge University Press, 2002.

[Tao19] T. Tao. "Least singular value, circular law, and Lindeberg exchange". In: *Random matrices*. Volume 26. IAS/Park City Math. Ser. Amer. Math. Soc., Providence, RI, 2019, pages 461–498.

[Waso4] L. Wasserman. *All of statistics*. A concise course in statistical inference. Springer-Verlag, New York, 2004. DOI: 10.1007/978-0-387-21736-9.

IV. conditioning

19. Conditional Expectation in L_2

Agenda:
1. Expectation and least-squares
2. Conditional expectation in L_2
3. Conditioning on a σ-algebra
4. Properties

> "Celui qui désespère des événements est un lâche, mais celui qui espère en la condition humaine est un fou."
>
> "He who despairs at events is a coward, but he who has hope for the human condition is a fool."
>
> —Albert Camus

So far, we have focused our attention on independent random variables and their properties. In this lecture, we expand our scope to include dependent random variables. Roughly speaking, when random variables are dependent, they contain information about each other. By observing one random variable, we may gain additional knowledge about the probable values of the other. In particular, we can update our best guess about the expected value of one random variable, given the value of the other.

For example, consider the experiment where we flip two fair coins, and we are interested in the total number X of heads. *A priori*, $\mathbb{E}\, X = 1$. Suppose that we flip the first coin, and we observe that its value is heads. The value of X is not yet determined, but we now anticipate that the expected number of heads will be 1.5. Similarly, if the first coin turns up tails, then the expected number of heads will be 0.5. Once we flip both coins, the number X of heads is no longer random; its expected value has been determined by the outcomes of the two flips.

We aim to develop this notion of conditional expectation; that is, the expected value of a random variable that reflects the information that we have acquired. The discussion in the last paragraph yields several insights. First, the conditional expectation is our "best guess" for the expectation, given some data. Second, the conditional expectation is a function of the observed data. Therefore, if we regard the observations as random, the conditional expectation is a random variable. Third, the more data we acquire, the more accurately we can refine our prediction of the expectation.

In this lecture, we will use these intuitions to define conditional expectation for square-integrable random variables. In this case, conditional expectation can be viewed as an orthogonal projection, and many of its properties follow immediately from the corresponding properties of orthogonal projection. In the next lecture, we will extend the definition to integrable random variables and present a more complete slate of facts about the conditional expectation.

19.1 Least squares and conditional expectations

First, we develop a connection between the expectation of a random variable and a least-squares problem. This perspective allows us to reinterpret the ordinary expectation as an orthogonal projection, which will serve as a model for the general conditional expectation.

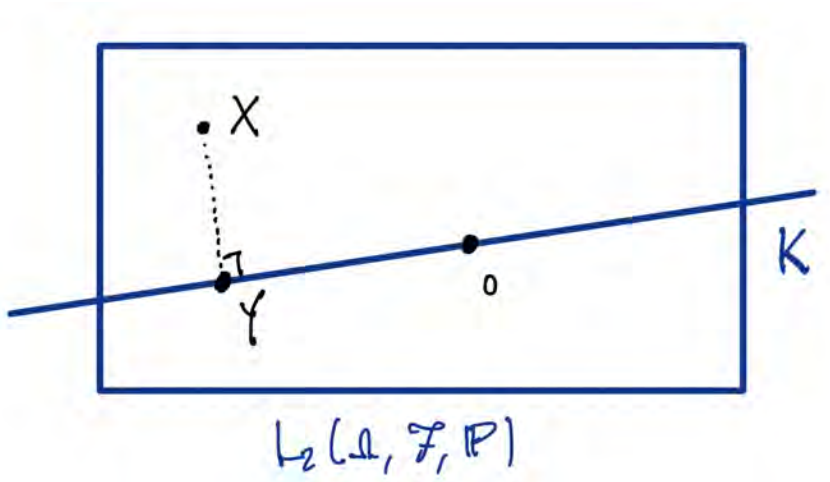


Figure 19.1 (Conditional expectation as an orthogonal projection). The subspace K contains random variables that are determined by the information given. The conditional expectation of X is a random variable $Y \in \mathsf{K}$ that best approximates X with respect to the L_2 pseudonorm. If we have no prior information, then K is simply the set of constant random variables, and the constant random variable $Y = \mathbb{E}\,X$ is our best guess for X.

19.1.1 Expectation

Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $X \in \mathsf{L}_2(\Omega, \mathcal{F}, \mathbb{P})$ be a square-integrable random variable. From Exercise 12.15, recall that the variance of X has a variational formulation:

$$\mathrm{Var}[X] = \|X - \mathbb{E}\,X\|_2^2 = \inf_{a \in \mathbb{R}} \|X - a\|_2^2. \tag{19.1}$$

This result leads to a variational interpretation of the expectation: $\mathbb{E}\,X$ is the *constant* that best approximates the random variable X with respect to the L_2 norm. It is our best guess for X in the absence of any information.

We can just as well treat a constant as a constant random variable. Consider the linear subspace of constant random variables:

$$\mathsf{K}_0 := \{Y \in \mathsf{L}_2 : Y(\omega) = a \text{ for all } \omega \in \Omega \text{ for some } a \in \mathbb{R}\}.$$

You should confirm that K_0 is a complete linear subspace of L_2. The constant random variable $Y : \omega \mapsto \mathbb{E}\,X$ obviously belongs to K_0, and the formula (19.1) ensures that Y is the (unique) orthogonal projection of X onto K_0. See Figure 19.1 for a reminder about the geometry. This innocent observation opens up a world of possibilities.

19.1.2 Linear least squares

Now, suppose that we observe a random variable $Z \in \mathsf{L}_2$, and we would like to use the value of Z to update our best guess about the expected value of X. Let us start with a simple approach that is widely used in practice. Suppose that we want to find the best approximation of X as a *linear* function of Z. This optimization problem takes the form

$$\text{minimize} \quad \|X - (a + bZ)\|_2^2 \quad \text{subject to} \quad a, b \in \mathbb{R}. \tag{19.2}$$

We can easily find the solution by setting the derivatives of the objective with respect to a, b to zero.

Exercise 19.1 (Linear least squares). Find the solution to (19.2). Express the result in terms of the expectations and covariances of X and Z.

Once again, we can interpret the variational problem (19.2) as an orthogonal projection. Introduce the linear subspace

$$\mathsf{K}_1 := \{a + bZ : a, b \in \mathbb{R}\}. \tag{19.3}$$

You should confirm that K_1 is a complete linear subspace of L_2. We can reframe the least-squares optimization problem (19.2) as

$$\text{minimize} \quad \|X - Y\|_2^2 \quad \text{subject to} \quad Y \in \mathsf{K}_1.$$

In other words, the solution Y is the (unique) orthogonal projection of X onto K_1. The geometry is similar with Figure 19.1.

Exercise 19.2 (Information never hurts). Let $X, Z \in \mathsf{L}_2$ be real random variables, and let Y be the orthogonal projection of X onto K_1, defined in (19.3). Show that

$$\|X - Y\|_2 \le \|X - \mathbb{E}\, X\|_2.$$

Find necessary and sufficient conditions for the inequality to be strict.

> **Aside:** Given observations $\{(X_i, Z_i) : i = 1, \dots, n\}$ that are distributed i.i.d. as (X, Z), we can formulate an empirical version of the optimization problem (19.2). That is,
>
> $$\text{minimize} \quad \frac{1}{n} \sum_{i=1}^{n} \big(X_i - (a + bZ_i)\big)^2 \quad \text{subject to} \quad a, b \in \mathbb{R}$$
>
> By solving this finite-dimensional optimization problem, we can obtain an empirical estimate for the coefficients $(a_\star, b_\star)$ that solve the problem (19.2). In statistical learning theory, a basic problem is to quantify how many samples n we need to obtain an accurate estimate of the true coefficients $(a_\star, b_\star)$.

19.1.3 Nonlinear least squares

In general, a real random variable Z contains more information about X than we can extract from the class K_1 of simple linear models. To make the best prediction we can, we need to consider *nonlinear* functions of the random variable Z. This leads to the optimization problem

$$\text{minimize} \quad \|X - h(Z)\|_2^2 \quad \text{subject to} \quad h(Z) \in \mathsf{L}_2. \tag{19.4}$$

In this expression, $h : \mathbb{R} \to \mathbb{R}$ ranges over Borel measurable functions for which $h(Z) \in \mathsf{L}_2$. In other words, we search over all square-integrable random variables that are completely determined by Z to find one that best approximates X in the L_2 sense.

As before, we can reframe (19.4) as an orthogonal projection. Introduce the linear subspace

$$\mathsf{K}_Z := \{h(Z) \in \mathsf{L}_2 : h : \mathbb{R} \to \mathbb{R} \text{ is Borel measurable}\}.$$

We arrive at the optimization problem

$$\text{minimize} \quad \|X - Y\|_2^2 \quad \text{subject to} \quad Y \in \mathsf{K}_Z. \tag{19.5}$$

Assuming that an orthogonal projection of X onto K_Z exists, it serves as a best L_2 estimate of X, given the random variable Z.

To see why an orthogonal projection is defined, we need an alternative expression for the linear subspace. As we have seen (Problem 13.20), a random variable Y is $\sigma(Z)$-measurable if and only if $Y = h(Z)$ for a Borel measurable function $h : \mathbb{R} \to \mathbb{R}$. Therefore, we can identify

$$\begin{aligned}\mathsf{K}_Z &= \{Y \in \mathsf{L}_2(\Omega, \mathcal{F}, \mathbb{P}) : Y \text{ is } \sigma(Z)\text{-measurable}\} \\ &= \mathsf{L}_2(\Omega, \sigma(Z), \mathbb{P}|_{\sigma(Z)}).\end{aligned}$$

The notation $\mathbb{P}|_{\mathcal{G}}$ means the restriction of the probability measure $\mathbb{P}$ to the events in a σ-algebra $\mathcal{G}$.

The square-integrable random variables on a probability space compose a complete linear subspace (Theorem 11.30). Therefore, Theorem 12.21 ensures that the orthogonal projection exists and is unique, up to its values on negligible sets.

We may now introduce the conditional expectation of X given Z. Indeed, we define the conditional expectation $\mathbb{E}[X \mid Z]$ to be any version of the orthogonal projection of X onto the subspace K_Z. That is, the conditional expectation is any one of the solutions to (19.5). As before, the geometry agrees with Figure 19.1.

Since the conditional expectation is a function $h(Z)$ of the random variable Z, the conditional expectation is itself a random variable. Moreover, the conditional expectation is determined up to its values on negligible sets because the orthogonal projection is determined up to its values on negligible sets.

> **Aside:** In spite of the terminology, the problem (19.4) is actually a linear least-squares problem posed on a linear space of random variables. In general, we cannot hope to find an explicit solution to (19.4). It is also challenging to solve the optimization problem numerically given the law of (X, Z) or to approximate the solution given an empirical sample from the law. Beyond that, there is no reason that an optimal estimator of X given Z is computationally tractable to implement. Computational statistics and statistical learning theory address these challenges.

19.2 Conditioning on a σ-algebra

The discussion in the last section culminated in the definition of the conditional expectation of a random variable X, given the value of another random variable Z. Indeed, we defined the conditional expectation $\mathbb{E}[X \mid Z]$ as the orthogonal projection of X onto a subspace of random variables determined by Z. In this section, we generalize this preliminary definition to allow for conditioning with respect to more general types of information.

19.2.1 Sigma-algebras carry information

We have the intuition that a σ-algebra $\mathcal{G}$ carries information about the state of the world. It is often helpful to think about $\mathcal{G}$ as a collection of events that have already been determined. In other words, we know whether each event has occurred or not. (That is, we can decide whether the distinguished sample point $\omega_0 \in \mathsf{G}$ or $\omega_0 \notin \mathsf{G}$ for each $\mathsf{G} \in \mathcal{G}$.)

For example, consider the experiment where we flip two fair coins. The natural sample space $\Omega = \{\mathrm{HH}, \mathrm{HT}, \mathrm{TH}, \mathrm{TT}\}$. Consider the sub-$\sigma$-algebra of events generated by the outcome of the first coin:

$$\{\emptyset, \{\mathrm{HH}, \mathrm{HT}\}, \{\mathrm{TH}, \mathrm{TT}\}, \Omega\}.$$

If we know how the first coin turns up, then we can decide whether each of these events has or has not occurred. For instance, if the first coin is heads, then the events $\{\mathrm{HH}, \mathrm{HT}\}$ and Ω occur, and the other events do not.

19.2.2 Conditional expectation with respect to a σ-algebra

In Lecture 13, we saw that independence can be formulated in terms of σ-algebras. Independence of events and random variables are just special cases of this definition. In the same way, we would like to define conditional expectation with respect to a σ-algebra.

In the last section, we saw that conditional expectation of $X \in \mathsf{L}_2$ with respect to a random variable Z is just the orthogonal projection onto the subspace of L_2 consisting of $\sigma(Z)$-measurable random variables. This connection suggests how we can define the conditional expectation with respect to a σ-algebra.

Definition 19.3 (Conditional expectation: L_2 case). Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra on Ω. For a square-integrable random variable $X \in \mathsf{L}_2(\Omega, \mathcal{F}, \mathbb{P})$, consider any solution Y to the least-squares problem

$$\text{minimize} \quad \|X - Y\|_2^2 \quad \text{subject to} \quad Y \in \mathsf{L}_2(\Omega, \mathcal{G}, \mathbb{P}|_{\mathcal{G}}).$$

We say that Y is a *version of the conditional expectation* $\mathbb{E}[X \mid \mathcal{G}]$.

Since there is rarely any ambiguity, we usually write $\mathbb{P} = \mathbb{P}|_{\mathcal{G}}$ in the sequel.

Definition 19.3 states that conditional expectation is simply an orthogonal projection, just as in Figure 19.1. The following result is fully justified by Theorem 12.21.

Theorem 19.4 (Conditional expectation: L_2 case). In the setting of Definition 19.3, there exists a version Y of the conditional expectation $\mathbb{E}[X \mid \mathcal{G}]$. Furthermore, if Y' is another version of the conditional expectation, then $Y = Y'$ almost surely. That is, $\mathbb{P}\{Y \neq Y'\} = 0$.

Let us emphasize that the conditional expectation $\mathbb{E}[X \mid \mathcal{G}]$ is a *random variable* defined on the sample space Ω. Its value is only determined once we specify which events in $\mathcal{G}$ have occurred and which have not. This fact has perplexed generations of probability students, so it merits reflection.

Exercise 19.5 (Conditioning: Trivial σ-algebra). Assume $X \in \mathsf{L}_2$, and let $\mathcal{G} = \{\emptyset, \Omega\}$. Show that $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X]$. What is the intuition for this statement?

Exercise 19.6 (Conditioning: Master σ-algebra). Assume $X \in \mathsf{L}_2$, and let $\mathcal{G} = \mathcal{F}$. Show that $\mathbb{E}[X \mid \mathcal{F}] = X$. What is the intuition for this statement?

Warning 19.7 (Conditional expectation: Almost sure). Sometimes, it is possible to make a canonical choice of the conditional expectation (e.g., when the sample space is finite). But, in general, the conditional expectation is only determined, modulo its values on negligible sets. As a consequence, identities involving conditional expectations hold almost surely. Initially, we will scrupulously remind you about this point, but later we will typically omit the qualification, as it is understood. ▪

19.2.3 Conditioning on a random variable

Definition 19.3 contains some familiar types of conditional expectation as special cases. First, each real random variable Z generates a σ-algebra $\sigma(Z)$. In this case, we define

$$\mathbb{E}[X \mid Z] := \mathbb{E}[X \mid \sigma(Z)].$$

This construction coincides with the preliminary definition of conditional expectation (Section 19.1.3) as the solution to a nonlinear least-squares problem.

More generally, we can define the conditional expectation with respect to a family of real random variables:

$$\mathbb{E}[X \mid Z_1, \dots, Z_n] := \mathbb{E}[X \mid \sigma(Z_1, \dots, Z_n)].$$

Recall that the σ-algebra $\sigma(Z_1, \dots, Z_n)$ is generated by all events of the form $Z_i^{-1}(\mathsf{B})$ for a Borel set $\mathsf{B} \in \mathcal{B}(\mathbb{R})$ and an index $i = 1, \dots, n$. The interpretation of this conditional expectation is similar to the univariate case, but we now search for approximations over the class of multivariate functions $h(Z_1, \dots, Z_n) \in \mathsf{L}_2$.

19.2.4 Conditioning on an event

We can also condition on an event, which leads to significant and informative simplifications. Each event $\mathsf{E} \in \mathcal{F}$ generates a σ-algebra $\sigma(\{\mathsf{E}\}) = \{\emptyset, \mathsf{E}, \mathsf{E}^c, \Omega\}$. We define the conditional expectation with respect to the event as

We often make abbreviations like $\sigma(\mathsf{E}) := \sigma(\{\mathsf{E}\})$ for legibility.

$$\mathbb{E}[X \mid \mathsf{E}] := \mathbb{E}[X \mid \sigma(\mathsf{E})] = \mathbb{E}[X \mid \mathbb{1}_{\mathsf{E}}].$$

We can develop an explicit expression for this conditional expectation that sheds further light on the concept.

Warning: In elementary probability, $\mathbb{E}[X \mid \mathsf{E}]$ is a number: the conditional expectation of X, given that E occurs. Here, it denotes a random variable whose value depends on whether or not E occurs. ∎

To compute the conditional expectation, we need to approximate X by a random variable Y that is measurable with respect to $\sigma(\{\mathsf{E}\})$. The measurability property forces the random variable Y to be constant on E and constant on E^c. Explicitly,

$$Y = a\mathbb{1}_{\mathsf{E}} + b\mathbb{1}_{\mathsf{E}^c} \quad \text{for some } a, b \in \mathbb{R}.$$

Thus, the conditional expectation $Y = \mathbb{E}[X \mid \mathsf{E}]$ is obtained by solving

$$\text{minimize} \quad \|X - (a\mathbb{1}_{\mathsf{E}} + b\mathbb{1}_{\mathsf{E}^c})\|_2^2 \quad \text{subject to} \quad a, b \in \mathbb{R}.$$

To do so, we use the orthogonality of $\mathbb{1}_{\mathsf{E}}$ and $\mathbb{1}_{\mathsf{E}^c}$ to split the squared norm:

$$\|X - (a\mathbb{1}_{\mathsf{E}} + b\mathbb{1}_{\mathsf{E}^c})\|_2^2 = \|(X - a)\mathbb{1}_{\mathsf{E}}\|_2^2 + \|(X - b)\mathbb{1}_{\mathsf{E}^c}\|_2^2.$$

We can now minimize independently over $a, b \in \mathbb{R}$. In light of (19.1), the solution is

$$a = \frac{\mathbb{E}[X\mathbb{1}_{\mathsf{E}}]}{\mathbb{P}(\mathsf{E})} \quad \text{and} \quad b = \frac{\mathbb{E}[X\mathbb{1}_{\mathsf{E}^c}]}{\mathbb{P}(\mathsf{E}^c)}.$$

If $\mathbb{P}(\mathsf{E}) = 0$, then we can take $a = 0$ because the numerator and denominator are both zero. Similarly, if $\mathbb{P}(\mathsf{E}^c) = 0$, then we can take $b = 0$. Therefore, the conditional expectation satisfies

$$Y(\omega) = \mathbb{E}[X \mid \mathsf{E}](\omega) = \begin{cases} \mathbb{E}[X\mathbb{1}_{\mathsf{E}}]/\mathbb{P}(\mathsf{E}), & \omega \in \mathsf{E}; \\ \mathbb{E}[X\mathbb{1}_{\mathsf{E}^c}]/\mathbb{P}(\mathsf{E}^c), & \omega \notin \mathsf{E}. \end{cases}$$

Thus, the conditional expectation admits a canonical definition in case $0 < \mathbb{P}(\mathsf{E}) < 1$. Otherwise, the conditional expectation can take an arbitrary value on the event E or E^c that is negligible.

In summary, the conditional expectation depends on whether the event E occurs. If E occurs, then the conditional expectation Y equals the average value of X on E. If E does not occur, then the conditional expectation Y equals the average value of X on E^c. This is precisely the elementary definition of conditional expectation. By direct calculation, you may also confirm that $\mathbb{E}\, Y = \mathbb{E}\, X$. See Figure 19.2 for an illustration.

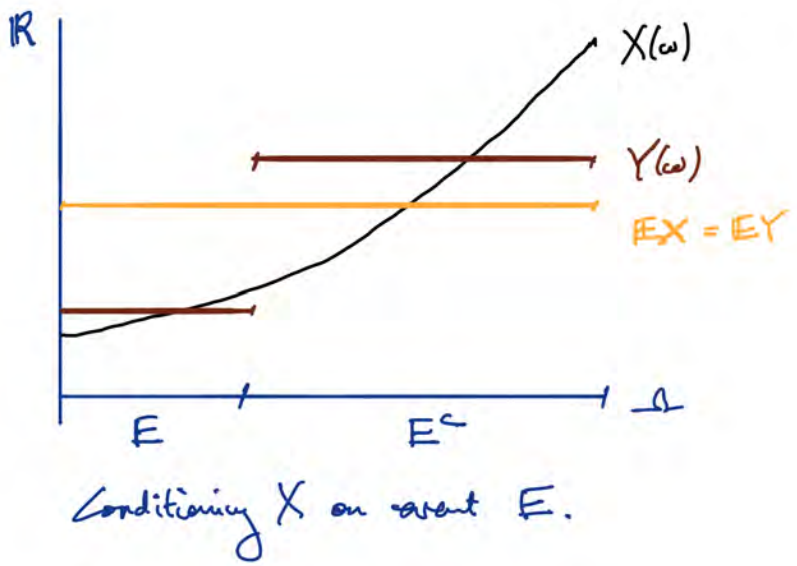


Figure 19.2 (Conditioning on an event). For a random variable $X \in \mathsf{L}_2$ and an event $\mathsf{E} \in \mathcal{F}$, we can easily compute $Y = \mathbb{E}[X \mid \mathsf{E}]$. The conditional expectation Y is constant on E and constant on E^c by measurability. By consistency, the constant value of Y on E (resp. E^c) agrees with the average value of X on E (resp. E^c). Moreover, the total expectations match: $\mathbb{E}\, Y = \mathbb{E}\, X$.

More generally, we can condition on a family of events:

$$\mathbb{E}[X \mid \mathsf{E}_1, \ldots, \mathsf{E}_n] := \mathbb{E}[X \mid \sigma(\mathsf{E}_1, \ldots, \mathsf{E}_n)].$$

In this case, the conditional expectation is constant over the "minimal" events in the σ-algebra, which take the form

$$\mathsf{A} = \mathsf{A}_1 \cap \cdots \cap \mathsf{A}_n \quad \text{where} \quad \mathsf{A}_i \in \{\mathsf{E}_i, \mathsf{E}_i^c\}.$$

For all $\omega \in \mathsf{A}$, the conditional expectation $\mathbb{E}[X \mid \mathsf{E}_1, \ldots, \mathsf{E}_n](\omega) = \mathbb{E}[X \mathbb{1}_{\mathsf{A}}]/\mathbb{P}(\mathsf{A})$.

19.3 Conditional expectation: Properties

For a square-integrable random variable, the conditional expectation is simply an orthogonal projection. In this setting, we can immediately derive a number of fundamental properties of the conditional expectation from facts about orthogonal projection. Lecture 20 contains generalizations and additional properties.

19.3.1 Measurability

We begin with two properties that characterize the conditional expectation. The first is a measurability property of the conditional expectation.

Proposition 19.8 (L_2 conditional expectation: Measurability). Let $X \in \mathsf{L}_2(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra contained in $\mathcal{F}$. If Y is a version of the conditional expectation $\mathbb{E}[X \mid \mathcal{G}]$, then Y is $\mathcal{G}$-measurable.

In detail, this result states that

$$Y^{-1}(\mathsf{B}) \in \mathcal{G} \quad \text{for each Borel set } \mathsf{B} \in \mathcal{B}(\mathbb{R}).$$

Therefore, if we know whether each event $\mathsf{G} \in \mathcal{G}$ occurs or not, then we can determine whether $Y \in \mathsf{B}$ for each Borel set B. The value of Y is completely determined by the events in $\mathcal{G}$.

Proof. By definition, a conditional expectation $Y = \mathbb{E}[X \,|\, \mathcal{G}]$ belongs to the space $\mathsf{L}_2(\Omega, \mathcal{G}, \mathbb{P})$. In particular, Y must be a $\mathcal{G}$-measurable random variable. ∎

In case $\mathcal{G} = \sigma(Z_1, \dots, Z_n)$ for a family $(Z_1, \dots, Z_n)$ of real random variables, this statement means that the conditional expectation $Y = h(Z_1, \dots, Z_n)$ for a Borel measurable function.

In case $\mathcal{G} = \sigma(\{\mathsf{E}_1, \dots, \mathsf{E}_n\})$ for a family $(\mathsf{E}_1, \dots, \mathsf{E}_n)$ of events, this statement means that the conditional expectation $Y = h(\varepsilon_1, \dots, \varepsilon_n)$ where $\varepsilon_i \in \{0, 1\}$ indicates whether event E_i occurs. In particular, Y is constant on "minimal" events. Figure 19.2 illustrates the simplest case.

19.3.2 Consistency

The second characteristic property of the conditional expectation states that certain averages of a random variable and its conditional expectation coincide.

Proposition 19.9 (L_2 conditional expectation: Consistency). Let $X \in \mathsf{L}_2(\Omega, \mathcal{F}, \mathbb{P})$, and let $Y = \mathbb{E}[X \,|\, \mathcal{G}]$ be a conditional expectation with respect to a σ-algebra $\mathcal{G} \subseteq \mathcal{F}$. Then

$$\mathbb{E}[X \mathbb{1}_{\mathsf{G}}] = \mathbb{E}[Y \mathbb{1}_{\mathsf{G}}] \quad \text{for each } \mathsf{G} \in \mathcal{G}. \tag{19.6}$$

This is a consistency or coarse-graining property. If $\mathsf{G} \in \mathcal{G}$ is one of the events on which we condition, the average value of the conditional expectation Y on G is the same as the average of X on G. In particular, if we consider a "minimal" event on which Y is constant, the constant equals the average value of X on the event. Figure 19.2 illustrates the simplest case.

Why do we restrict attention to these events? In light of Proposition 19.8, these are the ones where Y is naturally defined.

Proof. To check this claim, we use the dual characterization of orthogonal projection from Theorem 12.21:

$$\langle X - Y,\ W \rangle = 0 \quad \text{for each } W \in \mathsf{L}_2(\Omega, \mathcal{G}, \mathbb{P}). \tag{19.7}$$

Writing the inner product as an expectation, we see that

$$\mathbb{E}[XW] = \mathbb{E}[YW] \quad \text{for each } W \in \mathsf{L}_2(\Omega, \mathcal{G}, \mathbb{P}).$$

In particular, for any event $\mathsf{G} \in \mathcal{G}$, we can take $W = \mathbb{1}_{\mathsf{G}}$ because this random variable W is $\mathcal{G}$-measurable and square-integrable. ∎

19.3.3 The pull-through property

We continue with some special properties of conditional expectation. The first one describes a situation where we can draw a random variable through the conditional expectation.

Proposition 19.10 **(L_2 conditional expectation: Pull-through).** Suppose that $X \in \mathsf{L}_2(\Omega, \mathcal{F}, \mathbb{P})$, and let $Z \in \mathsf{L}_\infty(\Omega, \mathcal{G}, \mathbb{P})$ be an (essentially) *bounded, $\mathcal{G}$-measurable* random variable. Then

$$\mathbb{E}[XZ \,|\, \mathcal{G}] = \mathbb{E}[X \,|\, \mathcal{G}] \cdot Z.$$

In other words, given the information in the σ-algebra $\mathcal{G}$, the random variable Z is completely determined. Therefore, we may pull Z out of the expectation, as if it were a constant.

Proof. The pull-through property follows from the dual characterization of the orthogonal projection. Let $Y = \mathbb{E}[X \,|\, \mathcal{G}]$. For all $W \in \mathsf{L}_2(\Omega, \mathcal{G}, \mathbb{P})$,

$$\langle XZ - YZ,\ W \rangle = \mathbb{E}[(X - Y)(ZW)] = \langle X - Y,\ ZW \rangle = 0.$$

Indeed, since Z is bounded and $\mathcal{G}$-measurable, the product ZW is square-integrable and $\mathcal{G}$-measurable. The characterization (19.7) of the conditional expectation Y ensures that the inner-product is zero. Therefore, YZ is a version of the conditional expectation of XZ, given $\mathcal{G}$. ∎

Problem 19.11 **(L_2 conditional expectation: Pull-through).** For $X \in \mathsf{L}_2$, the pull-through property can be extended to the case where Z is *square-integrable* and $\mathcal{G}$-measurable. Prove it.

19.3.4 The tower property

There is a further orthogonality property that arises when we condition with respect to nested σ-algebras.

Proposition 19.12 **(L_2 conditional expectation: Tower).** Suppose that $\mathcal{G} \subseteq \mathcal{H} \subseteq \mathcal{F}$ are σ-algebras. Let $X \in \mathsf{L}_2(\Omega, \mathcal{F}, \mathbb{P})$. Then

$$\mathbb{E}[\mathbb{E}[X \,|\, \mathcal{H}] \,|\, \mathcal{G}] = \mathbb{E}[X \,|\, \mathcal{G}] \quad \text{almost surely.}$$

Heuristically, if we compute the expectation given a lot of information and then forget some of what we know, then the conditional expectation is the same as if we had less information to begin with. This result is called the *tower property*.

Proof. To establish this claim, note that

$$\mathsf{L}_2(\Omega, \mathcal{G}, \mathbb{P}) \subseteq \mathsf{L}_2(\Omega, \mathcal{H}, \mathbb{P}).$$

Problem 12.28(Nesting) shows that we can compute the orthogonal projection onto the smaller subspace ($\mathcal{G}$) in two stages. First, we compute the orthogonal projection onto the larger subspace ($\mathcal{H}$). Second, we compute its orthogonal projection onto the smaller subspace ($\mathcal{G}$). Reinterpreting this statement in terms of conditional expectation, we arrive at the tower property. ∎

19.3.5 Conditional expectation mimics an expectation

We motivated the construction of conditional expectation using the variational interpretation of the ordinary expectation. By what token does it make sense to call the conditional expectation an expectation? We may do so because it shares the four core properties of an expectation.

Proposition 19.13 **(L_2 conditional expectation: Expectation properties).** Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra on Ω. The conditional expectation with respect to $\mathcal{G}$ is...

1. **Unital:** $\mathbb{E}[1 \mid \mathcal{G}] = 1$ almost surely.
2. **Positive:** If X is square-integrable and $X \geq 0$ almost surely, then $\mathbb{E}[X \mid \mathcal{G}] \geq 0$ almost surely.
3. **Linear:** If $X, Y \in \mathsf{L}_2(\Omega, \mathcal{F}, \mathbb{P})$ and $\alpha, \beta \in \mathbb{R}$, then
$$\mathbb{E}[\alpha X + \beta Y \mid \mathcal{G}] = \alpha\, \mathbb{E}[X \mid \mathcal{G}] + \beta\, \mathbb{E}[Y \mid \mathcal{G}] \quad \text{almost surely.}$$
4. **Monotone:** In particular, for square-integrable random variables that satisfy $X \leq Y$, we have $\mathbb{E}[X \mid \mathcal{G}] \leq \mathbb{E}[Y \mid \mathcal{G}]$ almost surely.

Proof. **Unital:** Note that constant random variables are $\mathcal{G}$-measurable with respect to every σ-algebra. Therefore, $1 \in \mathsf{L}_2(\Omega, \mathcal{G}, \mathbb{P})$, and so 1 is a version of the orthogonal projection of 1 onto this linear subspace.

Positive: We may assume that $X \geq 0$ everywhere. (Why?) Let $Y = \mathbb{E}[X \mid \mathcal{G}]$ be a conditional expectation. The numerical inequality $(a - b_+)^2 \leq (a - b)^2$ for $a \geq 0$ and monotonicity of expectation together imply that

$$\|X - Y\|_2^2 = \mathbb{E}(X - Y)^2 \geq \mathbb{E}(X - Y_+)^2 = \|X - Y_+\|_2^2.$$

Note that the positive part $Y_+ = \max\{0, Y\} \in \mathsf{L}_2(\Omega, \mathcal{G}, \mathbb{P})$. The conditional expectation Y already minimizes the least-squares objective over this class, so Y_+ is also a minimizer. It follows that the conditional expectation has a version Y_+ that is positive.

Linear: The linearity property is a consequence of the fact that orthogonal projections are linear. See Problem 12.28. ∎

Exercise 19.14 (Conditional expectation: Monotonicity). Deduce the monotonicity property of conditional expectation from the properties we have already established.

Warning 19.15 (Conditional distribution). Although we might like to regard the conditional expectation as an ordinary expectation with respect to a (regular) conditional distribution, this interpretation is not valid in the more general settings. We will return to this point later. ▪

19.3.6 Special examples

There are a number of cases where the conditional expectation can be computed almost instantly from the definitions.

Proposition 19.16 (L_2 conditional expectation: Special cases). Let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra, and let $X \in \mathsf{L}_2(\Omega, \mathcal{F}, \mathbb{P})$ be a random variable.

1. **Expectations:** If $Y = \mathbb{E}[X \mid \mathcal{G}]$ almost surely, then $\mathbb{E}[Y] = \mathbb{E}[X]$.
2. **Full knowledge:** If X is $\mathcal{G}$-measurable, then $\mathbb{E}[X \mid \mathcal{G}] = X$ almost surely.
3. **Independence:** If $\sigma(X)$ is independent from $\mathcal{G}$, then $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X]$ almost surely.

Proof. **Expectations:** To see that the expectation of X and its conditional expectation $Y = \mathbb{E}[X \mid \mathcal{G}]$ coincide, we simply apply Proposition 19.9 with $\mathsf{G} = \Omega$. This is possible because every σ-algebra on Ω contains Ω.

Full knowledge: If X is $\mathcal{G}$-measurable, then $X \in \mathsf{L}_2(\Omega, \mathcal{G}, \mathbb{P})$. Therefore, X is its own orthogonal projection onto this space. Therefore, $\mathbb{E}[X \mid \mathcal{G}] = X$ almost surely. This result is natural because $\mathcal{G}$ already contains full information about X.

Independence: If $\sigma(X)$ is independent from $\mathcal{G}$, then you should confirm that

$$\mathbb{E}[(X - \mathbb{E}\,X)Y] = 0 \quad \text{for each } \mathcal{G}\text{-measurable } Y.$$

Now, if Y is $\mathcal{G}$-measurable, we may calculate that

$$\|X - Y\|_2^2 = \|(X - \mathbb{E}\,X) - (Y - \mathbb{E}\,X)\|_2^2 = \|X - \mathbb{E}\,X\|_2^2 + \|Y - \mathbb{E}\,X\|_2^2.$$

The last relation holds because $Y - \mathbb{E}\,X$ is $\mathcal{G}$-measurable. The right-hand side is minimized when $Y = \mathbb{E}\,X$. It follows that $Y = \mathbb{E}\,X$ is a version of the conditional expectation. ∎

Problems

Exercise 19.17 (Chain rule). Let $X \in \mathsf{L}_2$ be a real random variable. Let $\mathcal{G} \subseteq \mathcal{F}$ be a sub-σ-algebra of the master σ-algebra. Define the conditional variance:

$$\operatorname{Var}[X \,|\, \mathcal{G}] := \mathbb{E}[(X - \mathbb{E}[X \,|\, \mathcal{G}])^2 \,|\, \mathcal{G}].$$

1. Argue that the conditional variance is positive (almost surely).
2. Confirm the variational formulation:

$$\operatorname{Var}[X \,|\, \mathcal{G}](\omega) = \inf\{\mathbb{E}[(X - Y)^2 \,|\, \mathcal{G}](\omega) : Y \in \mathsf{L}_2(\Omega, \mathcal{G}, \mathbb{P})\} \quad \text{a.s.}$$

3. Establish the chain rule:

$$\operatorname{Var}[X] = \mathbb{E}[\operatorname{Var}[X \,|\, \mathcal{G}]] + \operatorname{Var}[\mathbb{E}[X \,|\, \mathcal{G}]].$$

Notes

Our approach to conditioning is modeled on Williams's book [Wil91].

Lecture bibliography

[Wil91] D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

20. Conditional Expectation in L_1

> "Do not become a mere recorder of facts, but try and penetrate the mystery of their origin."
>
> —Ivan Petrovich Pavlov

Agenda:

1. Conditional expectation
2. Convergence theorems
3. Further properties
4. Elementary conditional expectation

In the last lecture, we introduced the idea of conditional expectation. For a square-integrable random variable, we can update our best guess of its expected value in light of new information. In L_2, conditional expectation coincides with an orthogonal projection, so it has a geometric underpinning and an interpretation in terms of least squares. Furthermore, the properties of conditional expectation can be derived from facts about least-squares problems.

In this lecture, we will extend the conditional expectation to all integrable random variables. This result, due to Kolmogorov, is one of the most fundamental facts in modern probability. The generalization shares all of the properties outlined in the last lecture, but it lacks the pellucid geometry of least squares.

To begin, we will introduce the general definition of conditional expectation and prove that it is justified. Then we will develop the convergence theory for conditional expectation, which offers a quick path to obtaining the other properties of conditional expectation. Last, we will see that the abstract definition here captures all of the elementary notions of conditional expectation from introductory probability.

20.1 Conditional expectation, in general

In the *Grundbegriffe*, Kolmogorov introduced a general notion of conditional expectation. This definition is based on two characteristic properties of conditional expectation, outlined in Lecture 19. This section presents Kolmogorov's definition, along with the fundamental theorem of conditional expectation.

20.1.1 Characteristic properties

Let $X \in \mathsf{L}_2(\Omega, \mathcal{F}, \mathbb{P})$ be a *square-integrable* random variable, and let $Y \in \mathsf{L}_2(\Omega, \mathcal{G}, \mathbb{P})$ be a version of the conditional expectation $\mathbb{E}[X \,|\, \mathcal{G}]$. First, recall that the conditional expectation Y is a $\mathcal{G}$-measurable random variable, which means that Y is completely determined once we know which events in $\mathcal{G}$ occur (Proposition 19.8). Second, the expectation of Y over each event in $\mathcal{G}$ coincides with the expectation of X over the same event (Proposition 19.9).

Observe that neither of these properties requires the random variable X to be square-integrable. Therefore, we can use them as a template for extending the conditional expectation to random variables in L_1.

20.1.2 Kolmogorov's definition

Kolmogorov put forth the following definition of the conditional expectation. It requires further justification, which we will supply in a moment.

Definition 20.1 (Conditional expectation). Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra on Ω. Let $X \in \mathsf{L}_1(\Omega, \mathcal{F}, \mathbb{P})$ be an *integrable* real random variable. A real random variable Y on the probability space is called a *version of the conditional expectation* $\mathbb{E}[X \,|\, \mathcal{G}]$ if it satisfies three properties:

1. **Measurability:** Y is $\mathcal{G}$-measurable.
2. **Integrability:** Y is $\mathbb{P}$-integrable.
3. **Consistency:** For each event $\mathsf{G} \in \mathcal{G}$, we have $\mathbb{E}[Y \mathbb{1}_{\mathsf{G}}] = \mathbb{E}[X \mathbb{1}_{\mathsf{G}}]$.

The fundamental theorem of conditional expectation holds that the conditional expectation always exists, and it is essentially unique.

Theorem 20.2 (Conditional expectation: Fundamental theorem). In the setting of Definition 20.1, there exists a version Y of the conditional expectation $\mathbb{E}[X \,|\, \mathcal{G}]$. Furthermore, if Y' is another version of the conditional expectation, then $Y = Y'$ almost surely. That is, $\mathbb{P}\{Y \neq Y'\} = 0$.

In other words, the two properties (measurability, consistency) that we extracted from the definition of the conditional expectation in L_2 are sufficient to determine the conditional expectation in L_1. We will prove this result in the next subsection.

Exercise 20.3 (Conditional expectation: Positive case). Assuming that the random variable $X \geq 0$ is positive, but not necessarily integrable, adapt the proof of Theorem 20.2 to see that it has a positive conditional expectation $\mathbb{E}[Y \,|\, \mathcal{G}] \geq 0$ that is $\mathcal{G}$-measurable and satisfies the consistency property.

20.1.3 Fundamental theorem of conditional expectation: Proof

Let us prove Theorem 20.2. The approach is straightforward. We simply approximate an integrable random variable $X \in \mathsf{L}_1$ by a sequence of square-integrable random variables $(X_n : n \in \mathbb{N}) \subseteq \mathsf{L}_2$. The properties of the general conditional expectation are preserved in the limit.

Existence

Let $X \in \mathsf{L}_1(\Omega, \mathcal{F}, \mathbb{P})$. By decomposing the random variable into its positive and negative parts ($X = X_+ - X_-$ with $X_+ \geq 0$ and $X_- \geq 0$), we realize that it is sufficient to construct the conditional expectation of a positive, integrable random variable. (Why?)

Assume that $X \geq 0$. Define the random variables

$$X_n(\omega) := X(\omega) \wedge n \quad \text{for all } \omega \in \Omega \text{ and for all } n \in \mathbb{N}.$$

Evidently, X_n is positive, bounded, and $\mathcal{F}$-measurable. Furthermore, $X_n \uparrow X$ pointwise.

We may use Theorem 19.4 to construct a conditional expectation $Y_n = \mathbb{E}[X_n \,|\, \mathcal{G}]$ for each index $n \in \mathbb{N}$. Indeed, X_n is square-integrable because it is bounded. Since $X_n \geq 0$, Proposition 19.13 ensures that $Y_n \geq 0$ almost surely. The same proposition implies that $Y_{n+1} \geq Y_n$ almost surely for each $n \in \mathbb{N}$, so the conditional expectations are increasing. (Why?)

Now, we may define a candidate Y for the conditional expectation $\mathbb{E}[X \,|\, \mathcal{G}]$:

$$Y(\omega) := \limsup_{n \to \infty} Y_n(\omega) \quad \text{for each } \omega \in \Omega.$$

We must confirm the three properties of conditional expectation, stated in Definition 20.1.

First, we check that Y is $\mathcal{G}$-measurable. Each Y_n is an L_2 conditional expectation with respect to $\mathcal{G}$, so Y_n is $\mathcal{G}$-measurable (Proposition 19.8). Therefore, the limit superior Y of the sequence remains $\mathcal{G}$-measurable (Proposition 5.7).

Second, we confirm that Y is integrable. This is an easy consequence of monotone convergence (Theorem 9.10):

$$\mathbb{E}[Y] = \lim_{n\to\infty} \mathbb{E}[Y_n] = \lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X] < +\infty.$$

The first limit is valid because $(Y_n : n \in \mathbb{N})$ is an almost-surely increasing sequence of positive random variables. The second relation follows from the consistency property of the L_2 conditional expectation with $\mathsf{G} = \Omega$ (Proposition 19.9). The second limit follows from monotone convergence applied to the sequence $(X_n : n \in \mathbb{N})$. Finally, we use the fact that X is integrable.

Third, we deduce that Y satisfies the consistency property. This argument is almost the same as the last paragraph. Fix an event $\mathsf{G} \in \mathcal{G}$. Then

$$\mathbb{E}[Y\mathbb{1}_{\mathsf{G}}] = \lim_{n\to\infty} \mathbb{E}[Y_n\mathbb{1}_{\mathsf{G}}] = \lim_{n\to\infty} \mathbb{E}[X_n\mathbb{1}_{\mathsf{G}}] = \mathbb{E}[X\mathbb{1}_{\mathsf{G}}].$$

The first limit is a consequence of dominated convergence (Theorem 9.12), with the dominating random variable Y. The second relation follows from the consistency property of L_2 conditional expectation (Proposition 19.9). The last limit holds because of the monotone convergence theorem.

We conclude that $Y = \mathbb{E}[X \mid \mathcal{G}]$ is a version of the conditional expectation of the integrable random variable X with respect to the σ-algebra $\mathcal{G}$.

Uniqueness

Finally, we must verify that conditional expectations are determined up to their values on negligible sets. Suppose that Y and Y' are both versions of the conditional expectation $\mathbb{E}[X \mid \mathcal{G}]$. The consistency property requires that

$$\mathbb{E}[(Y - Y')\mathbb{1}_{\mathsf{G}}] = 0 \quad \text{for all events } \mathsf{G} \in \mathcal{G}. \tag{20.1}$$

For the sake of contradiction, assume that $\mathbb{P}\{Y \neq Y'\} > 0$.

Without loss of generality, we can arrange the two versions so that $\mathbb{P}\{Y > Y'\} > 0$. Define the events

$$\mathsf{E}_n := \{Y > Y' + n^{-1}\} \quad \text{for } n \in \mathbb{N}.$$

Since Y, Y' are both $\mathcal{G}$-measurable, each event $\mathsf{E}_n \in \mathcal{G}$. Beyond that, the sequence $(\mathsf{E}_n : n \in \mathbb{N})$ is increasing, and $\mathsf{E}_n \uparrow \{Y > Y'\}$. By the monotone limit property of a measure (Proposition 2.30),

$$\mathbb{P}(\mathsf{E}_n) \uparrow \mathbb{P}\{Y > Y'\} > 0.$$

For this limit to be valid, there must exist an index $n \in \mathbb{N}$ for which $\mathbb{P}(\mathsf{E}_n) > 0$.

We now approach the contradiction. For the distinguished index n, calculate that

$$\mathbb{E}[(Y - Y')\mathbb{1}_{\mathsf{E}_n}] \geq n^{-1} \cdot \mathbb{P}(\mathsf{E}_n) > 0.$$

Indeed, on the event E_n, the random variable $Y > Y' + n^{-1}$, and the probability of the event E_n is strictly positive. Unfortunately, the last display contradicts the consistency property (20.1) because $\mathsf{E}_n \in \mathcal{G}$. We must conclude that $\mathbb{P}\{Y \neq Y'\} = 0$.

20.2 Conditional expectation mimics an expectation

In this section, we will present a number of basic results which demonstrate that the conditional expectation behaves much like an ordinary expectation. In this section, we fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G} \subset \mathcal{F}$ be a σ-algebra on Ω.

20.2.1 Expectation properties

The conditional expectation shares four basic properties with the ordinary expectation. This result is the analog of Proposition 19.13.

Proposition 20.4 (Conditional expectation: Expectation properties). The conditional expectation is...

1. **Unital:** $\mathbb{E}[1 \mid \mathcal{G}] = 1$ almost surely.
2. **Positive:** If $X \geq 0$ almost surely, then $\mathbb{E}[X \mid \mathcal{G}] \geq 0$.
3. **Linear:** If X, Y are integrable and $\alpha, \beta \in \mathbb{R}$, then
$$\mathbb{E}[\alpha X + \beta Y \mid \mathcal{G}] = \alpha\, \mathbb{E}[X \mid \mathcal{G}] + \beta\, \mathbb{E}[Y \mid \mathcal{G}] \quad \text{almost surely.}$$
4. **Monotone:** In particular, for integrable random variables that satisfy $X \leq Y$ almost surely, it holds that $\mathbb{E}[X \mid \mathcal{G}] \leq \mathbb{E}[Y \mid \mathcal{G}]$ almost surely.

Exercise 20.5 (Conditional expectation: Expectation properties). Prove Proposition 20.4. In addition, show that the conditional expectation is positive linear: For positive random variables $X, Y \geq 0$ and $\alpha, \beta \geq 0$,
$$\mathbb{E}[\alpha X + \beta Y \mid \mathcal{G}] = \alpha\, \mathbb{E}[X \mid \mathcal{G}] + \beta\, \mathbb{E}[Y \mid \mathcal{G}] \quad \text{almost surely.}$$

20.2.2 Jensen's inequality

The conditional expectation also satisfies an analog of Jensen's inequality. This result is a very powerful tool.

Proposition 20.6 (Conditional expectation: Jensen). Let $\varphi : \mathsf{U} \to \mathbb{R}$ be a convex function on an open interval $\mathsf{U} \subseteq \mathbb{R}$. If X and $\varphi(X)$ are both integrable, then
$$\mathbb{E}[\varphi(X) \mid \mathcal{G}] \geq \varphi(\mathbb{E}[X \mid \mathcal{G}]) \quad \text{almost surely.}$$

As compared with Jensen's inequality for ordinary expectation, the proof is more involved. It is easy to prove the inequality for a single sample point, but we need to show that it holds on an almost-sure set of sample points. To do so, we have to piece together the result from a countable number of almost-sure inequalities. The dual representation of a convex function allows us to accomplish this task.

Proof. Corollary 9.21 furnishes a *countable* set $\mathsf{E} \subseteq \mathbb{R}^2$ for which
$$\varphi(y) = \sup\{\varphi(a) + g \cdot (y - a) : (a, g) \in \mathsf{E}\} \quad \text{for each } y \in \mathsf{U}.$$
For each pair $(a, g) \in \mathsf{E}$, we find that
$$\varphi(X) \geq \varphi(a) + g \cdot (X - a).$$
By monotonicity of conditional expectation (Proposition 20.4),
$$\mathbb{E}[\varphi(X) \mid \mathcal{G}] \geq \varphi(a) + g \cdot (\mathbb{E}[X \mid \mathcal{G}] - a) \quad \text{almost surely.}$$
We may take the countable supremum of the right-hand side without disturbing the almost sureness of the inequality:
$$\mathbb{E}[\varphi(X) \mid \mathcal{G}] \geq \sup\{\varphi(a) + g \cdot (\mathbb{E}[X \mid \mathcal{G}] - a) : (a, g) \in \mathsf{E}\} = \varphi(\mathbb{E}[X \mid \mathcal{G}]).$$
This is the required result. ∎

20.3 Convergence theorems

Like the ordinary expectation, the conditional expectation interacts well with limits. In this section, we present the main convergence theorems. These results allow us to deduce properties of the general conditional expectation from the corresponding properties of the L_2 conditional expectation. As usual, we fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a σ-algebra $\mathcal{G} \subseteq \mathcal{F}$ on Ω.

20.3.1 Monotone convergence and Fatou's lemma

For an increasing sequence of positive random variables, the conditional expectations also increase to an almost-sure limit.

Theorem 20.7 (Conditional monotone convergence). Consider an increasing sequence $(X_n : n \in \mathbb{N})$ of integrable, positive, real random variables with an integrable limit X. More precisely, $0 \le X_n \le X_{n+1}$ almost surely for each $n \in \mathbb{N}$, and $X_n \to X$ almost surely. Then

$$\mathbb{E}[X_n \,|\, \mathcal{G}] \uparrow \mathbb{E}[X \,|\, \mathcal{G}] \quad \text{almost surely.}$$

Proof. For each random variable X_n, we may extract a version $Y_n = \mathbb{E}[X_n \,|\, \mathcal{G}]$ of the conditional expectation. Define

$$Y(\omega) := \limsup_{n\to\infty} Y_n(\omega) \quad \text{for each } \omega \in \Omega.$$

Then Y is $\mathcal{G}$-measurable, and $Y_n \uparrow Y$ because of the monotonicity of conditional expectation (Proposition 20.4). For each event $\mathsf{G} \in \mathcal{G}$, two applications of the almost-sure monotone convergence theorem (Exercise 5.25) deliver

$$\mathbb{E}[Y \mathbb{1}_{\mathsf{G}}] = \lim_{n\to\infty} \mathbb{E}[Y_n \mathbb{1}_{\mathsf{G}}] = \lim_{n\to\infty} \mathbb{E}[X_n \mathbb{1}_{\mathsf{G}}] = \mathbb{E}[X \mathbb{1}_{\mathsf{G}}].$$

The second relation is the consistency property of the conditional expectation Y_n of the random variable X_n. ∎

Conditional expectations of positive random variables also satisfy a variant of Fatou's lemma. This result is useful because it does not require us to check convergence.

Theorem 20.8 (Conditional Fatou "lemma"). Consider a sequence $(X_n : n \in \mathbb{N})$ of integrable, almost surely *positive* real random variables. That is, $X_n \ge 0$ a.s. Then

$$\liminf_{n\to\infty} \mathbb{E}[X_n \,|\, \mathcal{G}] \ge \mathbb{E}[\liminf_{n\to\infty} X_n \,|\, \mathcal{G}] \quad \text{almost surely.}$$

Exercise 20.9 (Conditional Fatou). Prove Theorem 20.8.

20.3.2 Dominated and bounded convergence

As with ordinary expectation, the most widely used convergence theorem for conditional expectation is a dominated convergence result.

Theorem 20.10 (Conditional dominated convergence). Consider a sequence $(X_n : n \to \mathbb{N})$ of integrable, real-valued random variables that *converges* almost surely: $X_n \to X$. Assume that there is an *integrable* random variable Y for which

$$|X_n| \le |Y| \quad \text{almost surely for all } n \in \mathbb{N}.$$

Then the conditional expectations converge:

$$\mathbb{E}[X_n \mid \mathcal{G}] \to \mathbb{E}[X \mid \mathcal{G}] \quad \text{almost surely.}$$

Exercise 20.11 (Conditional dominated convergence). Prove Theorem 20.10.

Exercise 20.12 (Conditional bounded convergence). Consider a sequence $(X_n : n \in \mathbb{N})$ of real random variables that satisfy $|X_n| \leq R$ almost surely for each $n \in \mathbb{N}$, where $R \in \mathbb{R}$ is a number. Prove that $X_n \to X$ almost surely implies that $\mathbb{E}[X_n \mid \mathcal{G}] \to \mathbb{E}[X \mid \mathcal{G}]$ almost surely.

20.4 Conditional expectation: Special properties

The conditional expectation has a number of special properties that describe how particular types of random variables interact with conditioning. These results are all easy consequences of the definition of conditional expectation or the analogous statements for the L_2 conditional expectation.

Proposition 20.13 (Conditional expectation: Special properties). Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra on Ω. Let $X \in \mathsf{L}_1(\Omega, \mathcal{F}, \mathbb{P})$ be an integrable random variable.

1. **Expectations:** If $Y = \mathbb{E}[X \mid \mathcal{G}]$, then $\mathbb{E}[Y] = \mathbb{E}[X]$.
2. **Full knowledge:** If X is $\mathcal{G}$-measurable, then $\mathbb{E}[X \mid \mathcal{G}] = X$ almost surely.
3. **Independence:** If $\sigma(X)$ is independent from $\mathcal{G}$, then $\mathbb{E}[X \mid \mathcal{G}] = \mathbb{E}[X]$ almost surely.
4. **Pull-through:** If $Z \in \mathsf{L}_\infty(\Omega, \mathcal{G}, \mathbb{P})$ is a.s. *bounded* and $\mathcal{G}$-measurable, then
$$\mathbb{E}[XZ \mid \mathcal{G}] = \mathbb{E}[X \mid \mathcal{G}] \cdot Z \quad \text{almost surely.}$$
5. **Tower:** If $\mathcal{G} \subseteq \mathcal{H} \subseteq \mathcal{F}$, then
$$\mathbb{E}[\mathbb{E}[X \mid \mathcal{H}] \mid \mathcal{G}] = \mathbb{E}[X \mid \mathcal{G}] \quad \text{almost surely.}$$

Proof. **Expectations:** This is a consequence of the consistency property in Definition 20.1 for the certain event $\mathsf{G} = \Omega$.

Full knowledge: This is an immediate outcome of Definition 20.1. Indeed, X is assumed to be $\mathcal{G}$-measurable and integrable, and it is clear that the consistency property holds.

Independence: We can establish this result using approximation by square-integrable random variables. Let $X_n := (X \wedge n) \vee (-n)$ for each index $n \in \mathbb{N}$. Clearly, $\sigma(X_n)$ is still independent from $\mathcal{G}$, and X_n is square-integrable. Proposition 19.16 implies that $\mathbb{E}[X_n \mid \mathcal{G}] = \mathbb{E}[X_n]$ for each index n. We may apply conditional dominated convergence (Theorem 20.10) with X as the dominating random variable.

Pull-through: In the same way, the pull-through property follows from Proposition 19.10 using approximation. You should work through the steps of the argument.

Tower: This is just a matter of confirming that the left-hand side satisfies the properties that Definition 20.1 requires of a version of $\mathbb{E}[X \mid \mathcal{G}]$. You should write out the details. ∎

Exercise 20.14 (Pull-through law). The pull-through law holds under most reasonable integrability assumptions. For example, you can show that $\mathbb{E}[XZ \mid \mathcal{G}] = \mathbb{E}[X \mid \mathcal{G}] \cdot Z$ when $X, Z \in \mathsf{L}_2$ and Z is $\mathcal{G}$-measurable.

20.5 Conditional expectation in elementary probability

You are already familiar with several types of conditional expectation that arise in elementary probability. It is not immediately obvious how the elementary results are related to Kolmogorov's definition. In this section, we set this matter to rest by showing that everything is a special case of Definition 20.1. But the general construction also applies to many examples that fall outside the scope of elementary conditional expectation, such as conditioning a discrete random variable on a continuous random variable.

20.5.1 Conditioning on an event

Let $X \in \mathsf{L}_1(\Omega, \mathcal{F}, \mathbb{P})$ be an integrable random variable, and let $\mathsf{E} \in \mathcal{F}$ be an event. with $0 < \mathbb{P}(\mathsf{E}) < 1$. The conditional expectation of X, given that E has occurred, is defined as

$$\mathbb{E}[X \mid \mathsf{E} \text{ occurs}] := \frac{\mathbb{E}[X \mathbb{1}_{\mathsf{E}}]}{\mathbb{P}(\mathsf{E})}.$$

In other words, we compute the average value of X over the event E. The result is a *number*, not a random variable, because we assume that E occurs.

We have already seen a closely related result in Section 19.2.4. For a square-integrable random variable $X \in \mathsf{L}_2$, we calculated that

$$\mathbb{E}[X \mid \sigma(\mathsf{E})](\omega) = \begin{cases} \mathbb{E}[X \mathbb{1}_{\mathsf{E}}]/\mathbb{P}(\mathsf{E}), & \omega \in \mathsf{E}; \\ \mathbb{E}[X \mathbb{1}_{\mathsf{E}^c}]/\mathbb{P}(\mathsf{E}^c), & \omega \notin \mathsf{E}. \end{cases}$$

Using an approximation argument (or working directly from Definition 20.1), we see that the same formula is valid for an integrable random variable $X \in \mathsf{L}_1$.

The elementary notion of conditional expectation, given that the event E occurs, simply isolates the value of $\mathbb{E}[X \mid \sigma(\mathsf{E})](\omega)$ on the sample points $\omega \in \mathsf{E}$.

20.5.2 Conditioning on a discrete random variable

Next, we turn to the problem of conditioning a discrete random variable $X \in \mathsf{L}_1(\Omega, \mathcal{F}, \mathbb{P})$ on the value of another discrete random variable Z. The classical definition states that

$$\mathbb{E}[X \mid Z = z] = \frac{\mathbb{E}[X \mathbb{1}_{\{Z=z\}}]}{\mathbb{P}\{Z = z\}} \quad \text{for } z \in \operatorname{range}(Z).$$

Since Z is discrete, $\mathbb{P}\{Z = z\} > 0$ for each value $z \in \operatorname{range}(Z)$. This definition induces a random variable

$$\mathbb{E}[X \mid Z](\omega) := \mathbb{E}[X \mid Z = z_\omega] \quad \text{where } z_\omega := Z(\omega).$$

That is, the conditional expectation is a random variable whose value determined once the value $Z(\omega)$ is provided.

As with conditioning on an event, there is no difficulty in making the connection with the Kolmogorov definition. Consider the conditional expectation $Y = \mathbb{E}[X \mid Z] = \mathbb{E}[X \mid \sigma(Z)]$. The random variable Y is $\sigma(Z)$-measurable. Thus, Y must take a constant value, say $y(z)$, on each event $\{\omega : Z(\omega) = z\}$ where $z \in \operatorname{range}(Z)$. To determine the value $y(z)$, we use consistency:

$$\mathbb{E}[X \mathbb{1}_{\{Z=z\}}] = \mathbb{E}[Y \mathbb{1}_{\{Z=z\}}] = y(z) \cdot \mathbb{P}\{Z = z\}.$$

Since $\mathbb{P}\{Z = z\} > 0$ for each $z \in \operatorname{range}(Z)$, we have

$$Y(\omega) = \frac{\mathbb{E}[X \mathbb{1}_{\{Z=z\}}]}{\mathbb{P}\{Z = z\}} \quad \text{where } z = Z(\omega).$$

This is exactly the same as the classical definition.

20.5.3 Conditioning on a continuous random variable

Conditioning on a continuous random variable is more delicate because there is zero probability that a continuous random variable takes any particular value. In classical probability, this issue is addressed by passing to density functions. We will see that Kolmogorov's definition leads to the same result.

Suppose that (X, Z) is a jointly continuous pair of real random variables with joint density function $f_{X,Z}$. That is,

$$\mathbb{P}\{(X, Z) \in \mathsf{B}\} = \int_{\mathsf{B}} f_{X,Z}(x, z)\, \mathrm{d}x\, \mathrm{d}z \quad \text{for } \mathsf{B} \in \mathscr{B}(\mathbb{R}^2).$$

Keep in mind that an unadorned differential (such as $\mathrm{d}y$) means integration with respect to the Lebesgue measure (on the y-coordinate).

Introduce the marginal density f_Z of Z and the conditional density $f_{X \mid Z}$:

$$f_Z(z) := \int_{\mathbb{R}} f_{X,Z}(x, z)\, \mathrm{d}x \quad \text{and} \quad f_{X \mid Z}(x \mid z) := \frac{f_{X,Z}(x, z)}{f_Z(z)}.$$

Consider a Borel measurable function $h : \mathbb{R} \to \mathbb{R}$ where $h(X)$ is integrable. The classical definition of the conditional expectation is

$$\mathbb{E}[h(X) \mid Z = z] = \int_{\mathbb{R}} h(x)\, f_{X \mid Z}(x \mid z)\, \mathrm{d}z.$$

In other words, we simply compute the expectation with respect to the conditional distribution of X, given Z.

We need to verify that the Kolmogorov definition yields the same formulation of the conditional expectation. First, define the function

$$g(z) := \int_{\mathbb{R}} h(x)\, f_{X \mid Z}(x \mid z)\, \mathrm{d}x = \frac{1}{f_Z(z)} \int_{\mathbb{R}} h(x)\, f_{X,Z}(x, z)\, \mathrm{d}x \quad \text{for } z \in \mathbb{R}$$

In case $f_Z(z) = 0$, we set $g(z) = 0$. We *claim* that $g(Z)$ is a version of the conditional expectation $\mathbb{E}[h(X) \mid \sigma(Z)]$.

To check this statement, we first note that $g(Z)$ is measurable with respect to $\sigma(Z)$. To check consistency, observe that $\sigma(Z)$ is generated by events of the form $\{Z \in \mathsf{B}\}$ for a Borel set $\mathsf{B} \in \mathscr{B}(\mathbb{R})$. We may calculate that

$$\begin{aligned}
\mathbb{E}[h(X) \mathbb{1}_{\{Z \in \mathsf{B}\}}] &= \int_{\mathbb{R}^2} h(x) \mathbb{1}_{\mathsf{B}}(z)\, f_{X,Z}(x, z)\, \mathrm{d}x\, \mathrm{d}z \\
&= \int_{\mathbb{R}} \mathbb{1}_{\mathsf{B}}(z) \left(\int_{\mathbb{R}} h(x)\, f_{X,Z}(x, z)\, \mathrm{d}x \right) \mathrm{d}z \\
&= \int_{\mathbb{R}} \mathbb{1}_{\mathsf{B}}(z) g(z) \cdot f_Z(z)\, \mathrm{d}z = \mathbb{E}[g(Z) \mathbb{1}_{\{Z \in \mathsf{B}\}}].
\end{aligned}$$

We have used the (multivariate) law of the unconscious statistician (9.6) in the first and last relations. The second step is Fubini–Tonelli (Theorem 6.23, which is justified when $h(X)$ is integrable). To pass to the third step, we recognize the function $g(z)$.

Aside: The argument requires one more step, which is technical. So far, we have only checked consistency of $h(X)$ and $g(Z)$ on events of the form $Z^{-1}(B)$ where $B \in \mathcal{B}(\mathbb{R})$. The definition of conditional expectation requires us to check consistency on every event in $\sigma(Z)$. To do so, introduce the finite (signed) measures

$$\mu(E) := \int_E h(X)\,d\mathbb{P} \quad \text{and} \quad \nu(E) := \int_E g(Z)\,d\mathbb{P}.$$

These signed measures agree on all events of the form $E = Z^{-1}(B)$. This class of events is intersection stable. Therefore, the two measures must agree on the σ-algebra generated these events, namely $\sigma(Z)$. This point ultimately follows from Theorem E.4. You should work out the details.

20.6 *Regular conditional distributions

Intuitively, we might like to regard a conditional expectation as the integral with respect to a conditional probability measure. In many (but not all) cases of interest, this interpretation is possible. This section gives a short introduction to this extremely technical subject. As usual, we fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra on Ω.

20.6.1 Conditional probability

First, we remark that conditional expectation induces a notion of conditional probability if we restrict our attention to indicator functions.

Definition 20.15 (Conditional probability). The *conditional probability* $\mathbb{P}(E \mid \mathcal{G})$ of an event $E \in \mathcal{F}$, given the σ-algebra $\mathcal{G}$, is defined to be a version of the conditional expectation $\mathbb{E}[\mathbb{1}_E \mid \mathcal{G}]$.

The conditional probability is a random variable, not a number. This construction is consistent with the elementary notions of conditional probability if $\mathcal{G} = \sigma(A)$ for an event $A \in \mathcal{F}$ or if $\mathcal{G} = \sigma(Z)$ for a random variable Z.

In very simple cases (e.g., when Ω has finite cardinality), we can argue directly that the conditional probability $\mathbb{P}(\cdot \mid \mathcal{G})(\omega)$ defines a measure on the σ-algebra $\mathcal{F}$ of events for each sample point $\omega \in \Omega$. In more general settings, however, we cannot deduce that the conditional probability defines a probability measure. Indeed, it may not be clear how to stitch all of the conditional expectations together in a consistent way.

20.6.2 Conditional probability distributions

We are deeply interested in the case where conditioning results in a family of probability distributions on the master σ-algebra of events. The following definition captures the desired properties.

Definition 20.16 (Regular conditional distribution). We say that $\mathcal{G}$ induces a *regular conditional distribution* when there is a function

$$(E, \omega) \mapsto \mathbb{P}(E \mid \mathcal{G})(\omega) \quad \text{mapping} \quad \mathcal{F} \times \Omega \to [0, 1]$$

with two properties:

1. **Conditional:** For each $E \in \mathcal{F}$, the function $\omega \mapsto \mathbb{P}(E \mid \mathcal{G})(\omega)$ is a version of

the conditional probability $\mathbb{P}(\mathsf{E} \mid \mathcal{G})$.

2. **Distribution:** On a $\mathbb{P}|_{\mathcal{G}}$ almost-sure set of sample points $\omega \in \Omega$, the map $\mathsf{E} \mapsto \mathbb{P}(\mathsf{E} \mid \mathcal{G})(\omega)$ is a probability measure on $\mathcal{F}$.

We interpret the conditional distribution $\mathbb{P}(\cdot \mid \mathcal{G})$ as a new distribution of probability on events, updated based on the information encapsulated in the σ-algebra $\mathcal{G}$.

More rigorously, $\mathbb{P}(\cdot \mid \mathcal{G})$ can be regarded as a *random measure*. It is a function from the sample space Ω into the class of probability measures. When Tyche designates a sample point ω_0, it determines whether each event in $\mathcal{G}$ occurs. At this stage, the function $\mathsf{E} \mapsto \mathbb{P}(\mathsf{E} \mid \mathcal{G})$ is fixed for each event $\mathsf{E} \in \mathcal{F}$. There is a 100% chance that the result is a probability measure on $\mathcal{F}$.

Example 20.17 (Conditional distribution: Event). Consider the σ-algebra $\mathcal{G} = \sigma(\mathsf{A})$ generated by an event $\mathsf{A} \in \mathcal{F}$ with $0 < \mathbb{P}(\mathsf{A}) < 1$. It induces the conditional distribution

$$\mathbb{P}(\mathsf{E} \mid \sigma(\mathsf{A}))(\omega) = \begin{cases} \mathbb{P}(\mathsf{E} \mid \mathsf{A}), & \omega \in \mathsf{A}; \\ \mathbb{P}(\mathsf{E} \mid \mathsf{A}^{\mathsf{c}}), & \omega \notin \mathsf{A}. \end{cases}$$

This formula is valid for all $\mathsf{E} \in \mathcal{F}$ and all $\omega \in \Omega$. In words, the conditional distribution models the proportion of the probability of event E attributable to A occurring or not occurring. ∎

Example 20.18 (Conditional distribution: Discrete random variable). Consider the σ-algebra $\mathcal{G} = \sigma(Z)$ generated by a *discrete* random variable Z. For each event $\mathsf{E} \in \mathcal{F}$ and sample point $\omega \in \Omega$, the conditional distribution assigns the probability

$$\mathbb{P}(\mathsf{E} \mid \sigma(Z))(\omega) = \mathbb{P}(\mathsf{E} \mid Z = z_\omega) \quad \text{where } z_\omega = Z(\omega).$$

The conditional distribution models the proportion of the probability of event E attributable to each value of Z. ∎

When a conditional distribution exists, then we can calculate the conditional expectation of an integrable random variable X via the expression

$$\mathbb{E}[X \mid \mathcal{G}](\omega) = \int_\Omega X(s)\, \mathbb{P}(\mathrm{d}s \mid \mathcal{G})(\omega) \quad \text{almost surely.}$$

That is, for almost all sample points ω, the conditional expectation of X is determined as the expectation of X with respect to the probability measure $\mathbb{P}(\cdot \mid \mathcal{G})(\omega)$.

20.6.3 Conditional law of a random variable

In the same way that a random variable has a law, we can investigate when it has a conditional law.

Definition 20.19 (Regular conditional distribution: Random variable). Let X be a real random variable on the probability space. A *regular conditional distribution* of X, given $\mathcal{G}$, is a function

$$\mu_{X \mid \mathcal{G}} : \mathcal{B}(\mathbb{R}) \times \Omega \to [0, 1]$$

with two properties:

1. **Conditional:** For each Borel set $\mathsf{B} \in \mathcal{B}(\mathbb{R})$, the function $\omega \mapsto \mu_{X \mid \mathcal{G}}(\mathsf{B} \mid \omega)$ is a version of the conditional distribution $\mathbb{P}(X^{-1}(\mathsf{B}) \mid \mathcal{G})$.

2. **Distribution:** For a $\mathbb{P}|_{\mathcal{G}}$ almost-sure set of sample points $\omega \in \Omega$, the function $\mathsf{B} \mapsto \mu_{X\,|\,\mathcal{G}}(\mathsf{B}\,|\,\omega)$ is a probability measure on the Borel sets in $\mathbb{R}$.

We interpret the conditional law $\mu_{X\,|\,\mathcal{G}}$ as a new distribution over the outcomes of the random variable X, updated based on the information encapsulated in the σ-algebra $\mathcal{G}$.

That is, $\mu_{X\,|\,\mathcal{G}}$ can be regarded as a *random Borel measure* on the real line. Once a sample point $\omega_0 \in \Omega$ is chosen, the function $\mathsf{B} \mapsto \mu_{X\,|\,\mathcal{G}}(\mathsf{B})$ is determined for each Borel set $\mathsf{B} \in \mathcal{B}(\mathbb{R})$. This function is almost surely a Borel probability measure on $\mathbb{R}$, and it agrees with $\mathbb{P}\{X \in \mathsf{B}\,|\,\mathcal{G}\}$ almost surely.

Example 20.20 (Conditional law: Event). Consider the σ-algebra $\mathcal{G} = \sigma(\mathsf{A})$ generated by an event $\mathsf{A} \in \mathcal{F}$ with probability $0 < \mathbb{P}(\mathsf{A}) < 1$. The real random variable X has the conditional law

$$\mu_{X\,|\,\sigma(\mathsf{A})}(\mathsf{B}\,|\,\omega) = \begin{cases} \mathbb{P}(\{X \in \mathsf{B}\}\,|\,\mathsf{A}), & \omega \in \mathsf{A}; \\ \mathbb{P}(\{X \in \mathsf{B}\}\,|\,\mathsf{A}^{\mathsf{c}}), & \omega \notin \mathsf{A}. \end{cases}$$

This formula is valid for each Borel set $\mathsf{B} \in \mathcal{B}(\mathbb{R})$ and each $\omega \in \Omega$. Once we know whether or not A occurs, the conditional law of X is determined. For example, the probability that $\mathsf{E} := \{X \in \mathsf{B}\} \in \mathcal{F}$ occurs given that A occurs is the proportion of the probability in E that is attributable to A. ▪

Example 20.21 (Conditional law: Discrete random variable). Consider the σ-algebra $\mathcal{G} = \sigma(Z)$ generated by a *discrete* random variable Z. The real random variable X has the conditional law

$$\mu_{X\,|\,\sigma(Z)}(\mathsf{B}\,|\,\omega) = \mathbb{P}(\{X \in \mathsf{B}\}\,|\,\{Z = z_\omega\}) \quad \text{where } z_\omega = Z(\omega).$$

This formula is valid for each Borel set $\mathsf{B} \in \mathcal{B}(\mathbb{R})$ and each $\omega \in \Omega$. Once we know the value of Z, the conditional law of X is determined. The probability that $\mathsf{E} = \{X \in \mathsf{B}\} \in \mathcal{F}$ occurs given $\mathsf{A} = \{Z = z_\omega\}$ occurs is the proportion of the probability in E that is attributable to A. ▪

When a regular conditional law exists, we have a conditional version of the law of the unconscious statistician:

$$\mathbb{E}[h(X)\,|\,\mathcal{G}](\omega) = \int_{\mathbb{R}} h(x)\,\mu_{X\,|\,\mathcal{G}}(\mathrm{d}x\,|\,\omega) \quad \text{almost surely.}$$

This formula is valid when the function h is integrable with respect to the conditional law for all sample points $\omega \in \Omega$. For instance, it suffices that $h : \mathbb{R} \to \mathbb{R}$ is bounded and measurable.

Example 20.22 (Conditional law: Jointly continuous random variables). We can reinterpret the discussion from Section 20.5.3 to identify conditional laws in case (X, Z) is *jointly continuous* with a strictly positive density $f_{XZ} > 0$. In this case, we obtain the conditional law of X given Z by integrating the conditional density $f_{X\,|\,Z}$ over Borel sets. Indeed,

$$\mu_{X\,|\,\sigma(Z)}(\mathsf{B}\,|\,\omega) = \int_{\mathsf{B}} f_{X\,|\,Z}(x\,|\,Z(\omega))\,\mathrm{d}x.$$

This formula is valid for each Borel set $\mathsf{B} \in \mathcal{B}(\mathbb{R})$ and each sample point $\omega \in \Omega$.

More generally, the conditional law of the unconscious statistician reads

$$\mathbb{E}[h(X)\,|\,Z](\omega) = \int_{\mathbb{R}} h(x)\,f_{X\,|\,Z}(x\,|\,Z(\omega))\,\mathrm{d}x.$$

This formula holds for all bounded, measurable $h : \mathbb{R} \to \mathbb{R}$. At least for jointly continuous distributions, it is a valid intuition that conditional expectation derives from an integral. ∎

20.6.4 Disintegration

When a regular conditional distribution exists, we can also factorize the total expectation into a product of the marginal distribution and the conditional distribution. Moreover, we have a conditional variant of the Fubini–Tonelli theorem. Let us state following result without proof, even though it is not especially hard.

Theorem 20.23 (Disintegration). Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra on Ω. Suppose that the real random variable X admits a regular conditional law $\mu_{X \mid \mathcal{G}}$, and let Z be a $\mathcal{G}$-*measurable* real random variable. For each function $h : \mathbb{R}^2 \to \mathbb{R}$ with $\mathbb{E}\,|h(X,Z)| < +\infty$,

$$\mathbb{E}[h(X,Z) \mid \mathcal{G}] = \int_{\mathbb{R}} h(x,Z)\, \mu_{X \mid \mathcal{G}}(\mathrm{d}x) \quad \text{almost surely.}$$

In particular, the right-hand integral produces a $\mathcal{G}$-measurable random variable.

We can also take the expectation of the disintegration formula to obtain the promised extension of Fubini–Tonelli:

$$\mathbb{E}[h(X,Z)] = \mathbb{E}\left[\int_{\mathbb{R}} h(x,Z)\, \mu_{X \mid \mathcal{G}}(\mathrm{d}x)\right]. \tag{20.2}$$

In other words, we may compute the total expectation in two steps by first integrating with respect to the conditional law and then averaging over the residual randomness in the $\mathcal{G}$-measurable variable Z. The formula (20.2) is often called the *law of total expectation*. It generalizes several elementary results of the same name.

Consider the special case where $\mathcal{G} = \sigma(Z)$. It can be shown that the conditional law is given by a *probability kernel*:

$$\mu_{X \mid Z} = K(\cdot, Z) \quad \text{almost surely.}$$

In other words, the conditional law is a probability measure $K(\cdot, Z)$ that is essentially determined by the value of Z. Then we can simplify the disintegration formula to read

$$\mathbb{E}[h(X,Z) \mid Z] = \int_{\mathbb{R}} h(x,Z)\, K(\mathrm{d}x, Z) \quad \text{almost surely.}$$

Taking the expectation,

$$\mathbb{E}[h(X,Z)] = \int_{\mathbb{R}} \left[\int_{\mathbb{R}} h(x,z)\, K(\mathrm{d}x, z)\right] \mu_Z(\mathrm{d}z),$$

where μ_Z is the marginal law of Z. As above, we average with respect to the conditional law, and then we average over the residual randomness in Z. These rules are frequently used in the study of Markov chains.

20.6.5 Existence and uniqueness

The existence and uniqueness of regular conditional probability distributions is a delicate matter. Unlike many results in measure theory, it requires topological assumptions on the range of the random variable. Fortunately, for real random variables, no problems arise. The following theorem (also presented without proof) addresses this case.

Theorem 20.24 (Conditional distributions). Consider a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra on Ω. Let $X : \Omega \to \mathbb{R}$ be a real random variable. Then a regular conditional distribution $\mu_{X \mid \mathcal{G}}$ exists. If ν is another version of the conditional distribution of $X \mid \mathcal{G}$, then $\nu(\cdot \mid \omega) = \mu_{X \mid \mathcal{G}}(\cdot \mid \omega)$ on a $\mathbb{P}|_{\mathcal{G}}$ almost-sure set of sample points.

Problems

Exercise 20.25 (Pavlov). In this exercise, we will explore some basic examples of conditional expectation. Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Assume that $X, Z \in \mathsf{L}_1$.

1. Let $\mathcal{G} = \{\emptyset, \Omega\}$. What is $\mathbb{E}[X \mid \mathcal{G}]$?
2. Let $\mathcal{G} = \mathcal{F}$. What is $\mathbb{E}[X \mid \mathcal{G}]$?
3. Let $\mathsf{A} \in \mathcal{F}$, and set $\mathcal{G} = \sigma(\{\mathsf{A}\})$. What is $\mathbb{E}[X \mid \mathcal{G}]$?
4. Suppose that X, Z are independent. What is $\mathbb{E}[X \mid Z]$?
5. Suppose that $X = g(Z) \in \mathsf{L}_1$ for Borel measurable $g : \mathbb{R} \to \mathbb{R}$. What is $\mathbb{E}[X \mid Z]$?
6. Assume that X is a discrete random variable that takes values in $\mathbb{Z}$. Consider the random variable $Y = |X|$. Compute the conditional expectation $\mathbb{E}[X \mid Y]$.
7. Repeat the last part, now assuming that X is a continuous random variable taking values in $\mathbb{R}$.

Exercise 20.26 (Coarse graining). Consider the universal probability space $((0, 1], \mathcal{B}(0, 1], \lambda)$. For $n \in \mathbb{Z}_+$, let $\mathcal{G}_n := \sigma(\{(i2^{-n}, (i + 1)2^{-n}] : 0 \le i < 2^n \text{ and } i \in \mathbb{Z}\})$ be the σ-algebras generated by dyadic half-open intervals.

1. For an integrable random variable $X : (0, 1] \to \mathbb{R}$, compute $Y_n = \mathbb{E}[X \mid \mathcal{G}_n]$ for each $n \in \mathbb{N}$.
2. Relate Y_n and Y_{n+1}. **Hint:** Use the tower rule.
3. For a particular random variable, say $X(\omega) = \sin(2\pi\omega)$, illustrate X and Y_0, Y_1, Y_2, Y_3.

Exercise 20.27 (Memories). We say that a positive random variable X is *memoryless* when

$$\mathbb{P}\{X > t + r \mid X > t\} = \mathbb{P}\{X > r\} \quad \text{for all } t, r \ge 0.$$

To interpret this expression, we interpret X as the lifetime (say, of an electronic component), and we introduce the survival function $S(t) := \mathbb{P}\{X > t\}$ for $t \ge 0$. The memoryless property states that the probability of surviving for r seconds more does not depend on how long the component has already lasted.

1. Reinterpret the memoryless property as a functional equation for the survival function:
$$S(t + r) = S(t) \cdot S(r) \quad \text{for all } t, r \ge 0.$$
2. Assume that X is a discrete random variable, taking values in $\mathbb{N}$. What are the possible distributions for X? **Hint:** Consider $r = 1$.
3. Assume that X is an (absolutely) continuous random variable, taking values in $\mathbb{R}_+$. What are the possible distributions for X? **Hint:** Derive an ODE by taking $r \downarrow 0$.

Applications

Application 20.28 (Bayesian estimation). Bayesian statistics is based on the idea that we should encode our beliefs about the world using a probability distribution, which we should update as we acquire more information. In this application, we explore some basic facts about Bayes's rule and estimation.

1. Let E, F be events where $\mathbb{P}(\mathsf{F}) > 0$. Use elementary conditioning to confirm Bayes's rule:
$$\mathbb{P}(\mathsf{E} \mid \mathsf{F}) = \frac{\mathbb{P}(\mathsf{F} \mid \mathsf{E}) \cdot \mathbb{P}(\mathsf{E})}{\mathbb{P}(\mathsf{F})}.$$
2. Suppose that (X, Y) is a jointly continuous pair of real random variables whose joint density f_{XY} is strictly positive. Let f_X and f_Y be the marginal densities. Let $f_{X \mid Y}$ and $f_{Y \mid X}$ be the conditional densities. From the definition of a conditional density, establish Bayes's rule:
$$f_{X \mid Y}(x \mid y) = \frac{f_{Y \mid X}(y \mid x) f_X(x)}{f_Y(y)} \quad \text{for all } x, y \in \mathbb{R}.$$
Similar rules are valid in more general settings, e.g., when X is continuous and Y is discrete.
3. Suppose that (X, Y, Z) is a jointly continuous triple of real random variables whose joint density $f_{XYZ} > 0$. Using the tower law, show that conditioning is recursive:
$$f_{(X \mid Y) \mid Z}(x, y, z) = f_{X \mid (Y,Z)}(x, y, z) \quad \text{for all } x, y, z \in \mathbb{R}.$$
Similar rules are valid in more general settings.
4. **(Succession).** Imagine that we can perform repeated trials of a binary experiment (success/failure), resulting in outcomes $(X_1, X_2, X_3, \dots)$. Assume the outcomes (X_i) are independent $\textsc{bernoulli}(P)$ random variables, conditional on the unknown success probability P. Lacking further information, we might frame the "prior" model $P \sim \textsc{uniform}[0, 1]$. Now, suppose that we witness s successes and $n - s$ failures in the first n trials. Use Bayes's rule to show that the "posterior" probability distribution for P is
$$(P \mid X_1 + \cdots + X_n = s) \sim \textsc{beta}(s + 1, n - s + 1).$$
Confirm that the expected probability of success in the $(n + 1)$th trial is
$$\mathbb{E}[P \mid X_1 + \cdots + X_n = s] = \mathbb{P}\{X_{n+1} = 1 \mid X_1 + \cdots + X_n = s\} = \frac{s + 1}{n + 2}.$$
Hint: Given P, what is the distribution of $X_1 + \cdots + X_n$?
5. **(Sunrise).** As an impractical application, the mathematical astronomer Laplace observed that the sun has risen every morning for all of biblical history, which amounts to 5784 years at the time of writing. Compute the posterior probability that the sun will rise again tomorrow, assuming exactly 5784 years of 365.25 days. **(*)** Identify and explain the fallacy in this computation.
6. **(Counts).** Imagine that we can perform repeated trials of an experiment that results in counts $(N_1, N_2, N_3, \dots)$. Assume the outcomes (N_i) are independent $\textsc{poisson}(B)$ random variables, conditional on the unknown mean B. Lacking

further information, we model $B \sim \textsc{exponential}(b)$ for a (fixed) prior mean $\mathbb{E}[B] = b > 0$. Confirm that the posterior expectation of B satisfies

$$\mathbb{E}[B \mid N_1 + N_2 + \cdots + N_k = m] = \frac{m+1}{k+1/b} = \frac{b(m+1)}{bk+1}.$$

This is our best guess for the mean of the counts, given the observations. **Hint:** Given B, what is the distribution of $N_1 + \cdots + N_k$? Recall that Poisson random variables are stable.

7. **(Prussians).** It is dangerous to be a Prussian cavalryman. For example, you might get kicked to death by your horse. Contemplating this possibility, one might imagine that 1 soldier per corp per year suffers this tragic fate. For a period of years, beginning in 1875, the actual total number of deaths in 14 full corps were

   [1875] 3, 5, 7, 9, 10, 18, 6, 14, 11, 9, 5, 11, 15, 6, 11, 17, 12, 15, 8, 4 [1894].

   Use the Bayesian count estimator to obtain a sequence of posterior estimates for the mean number of deaths per corp per year (for the observations prior to each year). Discuss.

Notes

The treatment of conditional expectation in L_1 is adapted from Williams [Wil91, Chap. 9]. For more information about regular conditional distributions, refer to Pollard [Pol02, Chap. 5] or Dudley [Dud02, Chap. 10]. See Kallenberg [Kal02, pp. 107–108] for the proofs of the disintegration theorem (Theorem 20.23) and the existence of conditional laws for real random variables (Theorem 20.24).

Lecture bibliography

[Dud02] R. M. Dudley. *Real analysis and probability*. Revised reprint of the 1989 original. Cambridge University Press, 2002. DOI: 10.1017/CBO9780511755347.

[Kal02] O. Kallenberg. *Foundations of modern probability*. Second. Springer-Verlag, New York, 2002. DOI: 10.1007/978-1-4757-4015-8.

[Pol02] D. Pollard. *A user's guide to measure theoretic probability*. Cambridge University Press, 2002.

[Wil91] D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

21. Gaussians and Conditioning

"*Pauca sed matura.* Few but ripe."

—Motto of Karl Friedrich Gauss

Agenda:
1. Gaussian random variables
2. Characteristic functions
3. Characterization of distributions
4. Independence and conditioning

In this lecture, we make a deeper investigation of the properties of univariate and multivariate normal distributions. These random variables are the most fundamental ones of all, and they enjoy some truly remarkable properties. Our primary goal in this lecture is to show that, for normal random variables, conditional expectation reduces to linear least-squares. This incredible fact motivates many of the applications of normal random variables in statistics, machine learning, and applied mathematics. Along the way, we will introduce the idea of a characteristic function, which is an elegant tool for working with multivariate distributions.

21.1 Normal random variables

To begin, we recall the definitions of univariate and multivariate normal random variables, as well as some basic results about them.

21.1.1 Univariate normal distributions

In the central limit theorem, we encountered the standard normal distribution as the limiting distribution of a standardized sum of i.i.d. random variables in L_2. This is the single most important distribution in probability theory. In this section, we introduce the class of all normal distributions on the real line and discuss some of its properties.

Definition 21.1 (Standard normal distribution). The *standard normal distribution* γ is a Borel probability measure on the real line defined by

$$\gamma(\mathsf{B}) := \int_{\mathsf{B}} \frac{\mathrm{e}^{-z^2/2}}{\sqrt{2\pi}} \lambda(\mathrm{d}z) \quad \text{for all Borel } \mathsf{B} \in \mathscr{B}(\mathbb{R}).$$

Recall that λ is the Lebesgue measure on $\mathbb{R}$.

A real random variable Z with law γ is called a *standard normal* random variable. The distribution function has special notation that is widely used:

$$\Phi(a) := F_Z(a) = \gamma(-\infty, a] = \int_{-\infty}^{a} \frac{\mathrm{e}^{-z^2/2}}{\sqrt{2\pi}} \lambda(\mathrm{d}z) \quad \text{for all } a \in \mathbb{R}.$$

We sometimes write $\varphi(z) := (2\pi)^{-1/2}\mathrm{e}^{-z^2/2}$ for $z \in \mathbb{R}$ to denote the density of a standard normal random variable; this notation is not universal.

We obtain the class of univariate standard normal random variables by affine transformations of a real standard normal random variable.

Definition 21.2 (Normal random variable). Let Z be a standard normal random variable. For $m, \sigma \in \mathbb{R}$, we say that the random variable $X = m + \sigma Z$ follows a normal distribution with mean m and variance σ^2, and we write $X \sim \textsc{normal}(m, \sigma^2)$.

It is very common to refer to normal distributions as *Gaussian distributions*.

You should confirm that the mean and variance in Definition 21.2 are correct. By a change of variables argument, we quickly identify the law of a normal random variable.

Exercise 21.3 (Normal distribution). Let $X \sim \textsc{normal}(m, \sigma^2)$ where $\sigma > 0$. Verify that the law μ_X of X takes the form

$$\mu_X(\mathsf{B}) = \int_{\mathsf{B}} \frac{\mathrm{e}^{-(x-m)^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}} \, \lambda(\mathrm{d}x).$$

It is important to allow the case $\sigma = 0$. What is the distribution then?

Observe that there is a one-to-one correspondence between univariate normal distributions and pairs (m, σ) with $m \in \mathbb{R}$ and $\sigma \geq 0$.

21.1.2 Univariate normal distribution: Properties

This section contains several exercises that describe fundamental properties of normal random variables on the real line.

Problem 21.4 (Normal distribution: Density). Confirm that the standard normal law γ is a probability density by showing that $\gamma(\mathbb{R}) = 1$. **Hint:** You can argue that $(\gamma \times \gamma)(\mathbb{R}^2) = 1$ by computing the integral in polar coordinates.

Exercise 21.5 (Normal distribution: Approximate identity). Consider a real standard normal random variable Z. For a parameter $\sigma \geq 0$, show that

$$\mathbb{E}[h(a + \sigma Z)] = \int_{\mathbb{R}} h(a+z) \cdot \frac{\mathrm{e}^{-z^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}} \, \lambda(\mathrm{d}z) \to h(a) \quad \text{as } \sigma \to 0,$$

for all bounded, continuous $h : \mathbb{R} \to \mathbb{R}$ and for each $a \in \mathbb{R}$. Reinterpret this statement as a weak limit $\sigma Z \rightsquigarrow 0$ when $\sigma \to 0$. Express the result as a limit of probability measures.

Exercise 21.6 (Gaussian integration by parts). Let Z be a real standard normal variable variable. Suppose that $h : \mathbb{R} \to \mathbb{R}$ is a continuously differentiable function whose derivative is bounded. Verify that

$$\mathbb{E}[Zh(Z)] = \mathbb{E}[h'(Z)]. \tag{21.1}$$

Hint: This is just the integration by parts rule from calculus. **(*)** Conversely, show that a random variable Z that satisfies (21.1) must follow a standard normal distribution.

Exercise 21.7 (Normal distribution: Moments). Let Z be a standard normal random variable. For each natural number $p \in \mathbb{N}$, prove that

$$\mathbb{E}[Z^{2p-1}] = 0 \quad \text{and} \quad \mathbb{E}[Z^{2p}] = (2p-1)!!. \tag{21.2}$$

Recall that $(2p-1)!! := (2p-1)(2p-3)(2p-5)\cdots 3 \cdot 1$. **Hint:** For the odd moments, notice that the normal density is an odd function. For the even moments, use Gaussian integration by parts (Exercise 21.6) recursively.

Exercise 21.8 (Normal distribution: Tail bounds). Let Z be a real standard normal random variable.

1. For $t \geq 0$, show that $\mathbb{P}\{Z \geq t\} \leq \frac{1}{2}e^{-t^2/2}$.
2. For $t > 0$, show that $\mathbb{P}\{Z \geq t\} \leq (2\pi t^2)^{-1/2} e^{-t^2/2}$.

Hint: First the first part, find the maximum difference between left- and right-hand sides by global optimization of the resulting differentiable function. For the second part, introduce an extra factor into the integral that defines the tail probability.

21.1.3 Multivariate normal distributions

Now, we expand our scope to include multivariate normal distributions on $\mathbb{R}^n$. As in the univariate case, we begin with the fundamental example: an independent family of independent, real standard normal random variables.

Exercise 21.9 (Multivariate standard normal distribution). Consider an independent family $(Z_1, \dots, Z_n)$ of real, standard normal random variables. Show that the joint distribution has law

$$\gamma^n(\mathsf{B}) = \int_{\mathsf{B}} \frac{e^{-\|\boldsymbol{z}\|_2^2/2}}{(2\pi)^{n/2}} \lambda^n(d\boldsymbol{z}) \quad \text{for all Borel } \mathsf{B} \in \mathscr{B}(\mathbb{R}^n).$$

In the integral, the variable $\boldsymbol{z} \in \mathbb{R}^n$, and $\|\cdot\|_2$ is the Euclidean norm on $\mathbb{R}^n$, and λ^n is the Lebesgue measure on $\mathbb{R}^n$.

This law is called the *standard normal distribution* on $\mathbb{R}^n$. **Hint:** This point follows from the definition of a product measure and Fubini–Tonelli.

In much the same way that we defined the univariate normal distribution as an affine function of a standard normal distribution, we can define multivariate normal distributions.

Definition 21.10 (Multivariate normal random variable). Consider an independent family $(Z_1, \dots, Z_n)$ of real, standard normal random variables. For a vector $\boldsymbol{m} \in \mathbb{R}^n$ and a matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$, we can construct a family of random variables

$$X_j := m_j + \sum_{k=1}^n \sigma_{jk} Z_k \quad \text{for each } j = 1, \dots, n. \tag{21.3}$$

We say that a random vector $(X_1, \dots, X_n) \in \mathbb{R}^n$ follows a *multivariate normal* distribution if and only if it can be written in the form (21.3).

We write m_j for the jth entry of the vector $\boldsymbol{m}$ and σ_{jk} for the (j,k) entry of the matrix $\boldsymbol{\Sigma}$.

Exercise 21.11 (Multivariate normal: Mean and covariance). Construct the random vector $(X_1, \dots, X_n)$ as in (21.3). Introduce the positive-semidefinite matrix $\boldsymbol{C} = \boldsymbol{\Sigma}\boldsymbol{\Sigma}^* \in \mathbb{R}^n$. Show that

$$\mathbb{E}[X_j] = m_j \quad \text{and} \quad \operatorname{Cov}(X_j, X_k) = c_{jk} \quad \text{for all } j, k = 1, \dots, n.$$

A matrix $\boldsymbol{C} \in \mathbb{R}^{n \times n}$ is *positive semidefinite* if it is symmetric and $\boldsymbol{u}^* \boldsymbol{C} \boldsymbol{u} \geq 0$ for each vector $\boldsymbol{u} \in \mathbb{R}^n$. The symbol * denotes the (conjugate) transpose of a vector or matrix.

To be clear, we have written c_{jk} for the (j,k) entry of the matrix $\boldsymbol{C}$. We call $\boldsymbol{C}$ the *covariance matrix* of the vector $(X_1, \dots, X_n)$.

For every vector $\boldsymbol{m} \in \mathbb{R}^n$ and positive-semidefinite matrix $\boldsymbol{C} \in \mathbb{R}^{n \times n}$, confirm that there is a normal random vector $(X_1, \dots, X_n)$ with mean vector $\boldsymbol{m}$ and covariance matrix $\boldsymbol{C}$. **Hint:** Consider the positive-semidefinite square root of the covariance matrix.

Exercise 21.12 (Multivariate normal: Affine transformations). Let $\boldsymbol{X}$ be a multivariate normal distribution on $\mathbb{R}^n$. For a vector $\boldsymbol{m} \in \mathbb{R}^n$ and a matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$, show that $\boldsymbol{Y} = \boldsymbol{m} + \boldsymbol{\Sigma}\boldsymbol{X}$ follows a multivariate normal distribution. What are the mean and covariance?

We have seen that there exists a normal distribution with a given mean and positive-semidefinite covariance. Our primary task is to establish that the mean and covariance uniquely determine the normal distribution. Anticipating this result, we give a formal definition of the multivariate normal distribution.

Definition 21.13 (Multivariate normal distribution). Fix a vector $\boldsymbol{m} \in \mathbb{R}^n$ and a positive-semidefinite matrix $\boldsymbol{C} \in \mathbb{R}^{n \times n}$. We say that a multivariate normal random variable with mean $\boldsymbol{m}$ and covariance $\boldsymbol{C}$ follows the $\textsc{normal}(\boldsymbol{m}, \boldsymbol{C})$ distribution.

Right now, this definition remains suspect because the parameterization $(\boldsymbol{m}, \boldsymbol{\Sigma})$ in Definition 21.10 could result in distributions that are different, even though they have the same mean and covariance matrix. One approach to resolving this matter is to explicitly compute the form of the multivariate normal distribution. In case the matrix $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ in the definition (21.3) is invertible, then the multivariate normal distribution has a density with respect to the Lebesgue measure on $\mathbb{R}^n$.

Exercise 21.14 (Multivariate normal: Density). Suppose that $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ is an *invertible* matrix, and construct the random vector $(X_1, \dots, X_n)$ as in (21.3). Show that the joint distribution of the random vector satisfies

$$\mathbb{P}\{(X_1, \dots, X_n) \in \mathsf{B}\} = \int_{\mathsf{B}} \frac{\mathrm{e}^{-(\boldsymbol{x}-\boldsymbol{m})^* \boldsymbol{C}^{-1} (\boldsymbol{x}-\boldsymbol{m})}}{(\det(2\pi \boldsymbol{C}))^{1/2}} \, \lambda^n(\mathrm{d}\boldsymbol{x}) \quad \text{for all Borel } \mathsf{B} \in \mathscr{B}(\mathbb{R}^n).$$

We have written $\boldsymbol{C}^{-1}$ for the inverse of the matrix $\boldsymbol{C}$, and det denotes the determinant.

Hint: This result follows from the change of variables formula for a multivariate integral in terms of the Jacobian of the transformation. **(*)** If $\boldsymbol{\Sigma}$ is singular, what happens?

Exercise 21.14 settles the matter of uniqueness when the covariance matrix $\boldsymbol{C}$ is invertible. When the matrix $\boldsymbol{\Sigma}$ in (21.3) is singular, it is also possible to produce an explicit (but somewhat fussy) representation for the distribution. This representation implies that the distribution is also uniquely determined when $\boldsymbol{C}$ is singular. Instead of pursuing this argument, we will develop a more elegant approach for studying normal distributions, via characteristic functions.

21.1.4 *Multivariate normal distribution: Properties

We proceed with several exercises in analogy with the results in Section 21.1.2.

Exercise 21.15 (Multivariate normal: Approximate identity). Consider a multivariate standard normal random variable $\boldsymbol{Z}$ on $\mathbb{R}^n$. For a parameter $\sigma \geq 0$, show that

$$\mathbb{E}[h(\boldsymbol{a} + \sigma \boldsymbol{Z})] = \int_{\mathbb{R}^n} h(\boldsymbol{a} + \boldsymbol{z}) \cdot \frac{\mathrm{e}^{-\|\boldsymbol{z}\|^2/(2\sigma^2)}}{(2\pi\sigma^2)^{n/2}} \, \lambda^n(\mathrm{d}\boldsymbol{z}) \to h(\boldsymbol{a}) \quad \text{as } \sigma \to 0,$$

for each $\boldsymbol{a} \in \mathbb{R}^n$ and for all bounded, continuous $h : \mathbb{R}^n \to \mathbb{R}$. Reinterpret this statement as the weak limit $\sigma \boldsymbol{Z} \rightsquigarrow \mathbf{0}$ when $\sigma \to 0$.

Exercise 21.16 (Gaussian integration by parts). Let $\boldsymbol{Z}$ be a multivariate standard normal variable variable on $\mathbb{R}^n$. Suppose that $h : \mathbb{R}^n \to \mathbb{R}$ is a twice continuously differentiable function whose second derivative is bounded. Verify that

$$\mathbb{E}[\langle \boldsymbol{Z}, \, \nabla h(\boldsymbol{Z}) \rangle] = \mathbb{E}[\operatorname{tr} \nabla^2 h(\boldsymbol{Z})].$$

Here, $\langle \cdot, \, \cdot \rangle$ is the standard Euclidean inner product, and tr is the trace. As usual, ∇ denotes the gradient and ∇^2 is the Hessian.

Hint: Once again, this is just integration by parts. **(*)** More generally, show that a multivariate normal variable $\boldsymbol{X}$ on $\mathbb{R}^n$ with mean zero and covariance $\boldsymbol{C} \in \mathbb{R}^{n \times n}$ satisfies

$$\mathbb{E}[\langle \boldsymbol{X}, \, \nabla h(\boldsymbol{X}) \rangle] = \mathbb{E}[\operatorname{tr}[\boldsymbol{C} \nabla^2 h(\boldsymbol{X})]].$$

Although this result may seem esoteric, it is truly fundamental.

21.2 Characteristic functions

In this section, we introduce the characteristic function of a real random variable, and then we extend the definition to a vector of real random variables. We will prove that this function completely determines the distribution of a vector of real random variables, so it is an incredibly useful tool for reasoning about when two distributions coincide or differ from each other.

21.2.1 Real random variables

We begin with the basic definition of the characteristic function of a real random variable.

This definition is the reason that probabilists call $\mathbb{1}_E$ an *indicator* function, not a characteristic function (as some other mathematicians do).

We reserve the symbol $\mathrm{i} := \sqrt{-1}$ for the imaginary unit.

Definition 21.17 (Characteristic function). Let X be a *real* random variable. The *characteristic function* $\chi_X : \mathbb{R} \to \mathbb{C}$ is a complex-valued function defined as

$$\chi_X(\theta) := \mathbb{E}[\mathrm{e}^{\mathrm{i}\theta X}] := \mathbb{E}[\cos(\theta X)] + \mathrm{i} \cdot \mathbb{E}[\sin(\theta X)] \quad \text{for all } \theta \in \mathbb{R}.$$

Exercise 21.18 (Characteristic function: Bounded and continuous). Let X be a real random variable. Explain why $\chi_X(\theta)$ is defined and finite for all $\theta \in \mathbb{R}$. Argue that the characteristic function is uniformly bounded in magnitude by one and continuous.

Exercise 21.19 (*Characteristic function: Smoothness). Suppose that $X \in \mathsf{L}_1$ is a real random variable. Show that the derivative $\chi'_X(\theta)$ exists and is a continuous function. **(*)** Under what conditions does χ_X have n continuous derivatives?

Exercise 21.20 (Characteristic function: Affine transformations). Let X be a real random variable. Let $a, b \in \mathbb{R}$ be real numbers. Check that

$$\chi_{a+bX}(\theta) = \mathrm{e}^{\mathrm{i}a\theta} \cdot \chi_X(b\theta) \quad \text{for all } \theta \in \mathbb{R}.$$

Exercise 21.21 (Characteristic function: Multiplicativity). Let X, Y be *independent*, real random variables. Prove that

$$\chi_{X+Y}(\theta) = \chi_X(t) \cdot \chi_Y(\theta) \quad \text{for all } \theta \in \mathbb{R}.$$

Extend this identity to a finite sum of independent, real random variables.

21.2.2 Examples

Let us continue with computations of the characteristic functions for the most important real random variables.

Example 21.22 (Binomial distribution: Characteristic function). First, observe that a Bernoulli random variable $Y \sim \textsc{bernoulli}(p)$ has characteristic function

$$\chi_Y(\theta) = \mathbb{E}[\mathrm{e}^{\mathrm{i}\theta Y}] = 1 + p \cdot (\mathrm{e}^{\mathrm{i}\theta} - 1) \quad \text{for all } \theta \in \mathbb{R}.$$

Since the binomial random variable $X \sim \textsc{binomial}(n, p)$ can be written as the sum of n i.i.d. copies of Y, we immediately recognize that

$$\chi_X(\theta) = \left[1 + p \cdot (\mathrm{e}^{\mathrm{i}\theta} - 1)\right]^n \quad \text{for all } \theta \in \mathbb{R}.$$

Indeed, Exercise 21.21 states that the characteristic function is multiplicative. ∎

Example 21.23 (Univariate normal distribution: Characteristic function). Consider a univariate normal random variable $X \sim \text{NORMAL}(m, \sigma^2)$ with mean $m \in \mathbb{R}$ and variance $\sigma^2 \in \mathbb{R}_+$. We will calculate that

$$\chi_X(\theta) = \mathrm{e}^{\mathrm{i}m\theta + \sigma^2\theta^2/2} \quad \text{for all } \theta \in R.$$

The case $\sigma = 0$ is trivial, so we may assume that $\sigma > 0$. To simplify matters, it is enough to standardize so that $m = 0$ and $\sigma^2 = 1$. Indeed, we may write $X = m + \sigma \cdot [(Z - m)/\sigma]$ and apply the affine transformation rule (Exercise 21.20).

Let $Z \sim \text{NORMAL}(0, 1)$ be a real standard normal variable. The argument is based on the fact that the Taylor series for the exponential function converges everywhere in the complex plane. By dominated convergence (Theorem 9.12),

$$\chi_Z(\theta) = \mathbb{E}\left[1 + \sum\nolimits_{p=1}^{\infty} \frac{(\mathrm{i}\theta)^p}{p!} Z^p\right] = 1 + \sum\nolimits_{p=1}^{\infty} \frac{(\mathrm{i}\theta)^p}{p!} \mathbb{E}[Z^p].$$

(Why?) Exercise 21.7 shows us how to compute the moments of a standard normal random variable. Therefore, the series simplifies to

$$\chi_Z(t) = 1 + \sum\nolimits_{p=1}^{\infty} \frac{(-1)^p \theta^{2p}}{(2p)!}(2p-1)!! = 1 + \sum\nolimits_{p=1}^{\infty} \frac{(-\theta^2)^p}{2^p p!} = \mathrm{e}^{-\theta^2/2}.$$

This is the required result. ∎

Exercise 21.24 (Normal random variable: Inversion). Let Z be a standard normal random variable. Show that the standard normal density φ admits the inversion formula

$$\varphi(z) := \frac{\mathrm{e}^{-z^2/2}}{\sqrt{2\pi}} = \frac{1}{2\pi} \int_{\mathbb{R}} \chi_Z(\theta) \cdot \mathrm{e}^{-\mathrm{i}\theta z} \, \lambda(\mathrm{d}\theta).$$

Hint: Use the computation from Example 21.23, the definition of the characteristic function, and the law of the unconscious statistician.

Exercise 21.25 (Poisson distribution: Characteristic function). Let $Q \sim \text{POISSON}(\beta)$ be a Poisson random variable with mean $\beta \in \mathbb{R}_+$. Prove that

$$\chi_Q(\theta) = \exp\left(\beta(\mathrm{e}^{\mathrm{i}\theta} - 1)\right) \quad \text{for all } \theta \in \mathbb{R}.$$

21.2.3 Multivariate characteristic functions

Characteristic functions are especially useful for studying multivariate distributions. For this purpose, we require a small generalization.

Definition 21.26 (Characteristic function: Multivariate distribution). Let $\boldsymbol{X} := (X_1, \dots, X_n)$ be an arbitrary family of real random variables. The *characteristic function* $\chi_{\boldsymbol{X}} : \mathbb{R}^n \to \mathbb{C}$ is the multivariate complex-valued function

$$\chi_{\boldsymbol{X}}(\boldsymbol{\theta}) := \mathbb{E}[\mathrm{e}^{\mathrm{i}\langle \boldsymbol{\theta}, \boldsymbol{X} \rangle}] \quad \text{for all } \boldsymbol{\theta} \in \mathbb{R}^n.$$

We write $\langle \cdot, \cdot \rangle$ for the standard Euclidean inner product on $\mathbb{R}^n$.

Multivariate characteristic functions share many of the basic properties of univariate characteristic functions.

Exercise 21.27 (Multivariate characteristic function: Affine transformations). Let $\boldsymbol{X}$ be a random vector taking values in $\mathbb{R}^n$. Let $\boldsymbol{m} \in \mathbb{R}^n$ be a vector, and let $\boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}$ be a matrix. Form the random vector $\boldsymbol{Y} := \boldsymbol{m} + \boldsymbol{\Sigma}\boldsymbol{X}$. Then

$$\chi_{\boldsymbol{Y}}(\boldsymbol{\theta}) = \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\theta}, \boldsymbol{m} \rangle} \cdot \chi_{\boldsymbol{X}}(\boldsymbol{\Sigma}^* \boldsymbol{\theta}) \quad \text{for all } \boldsymbol{\theta} \in \mathbb{R}^n.$$

Exercise 21.28 (Multivariate characteristic function: Multiplicativity). Let $\boldsymbol{X}$ and $\boldsymbol{Y}$ be *independent* random vector taking values in $\mathbb{R}^n$. Verify that

$$\chi_{\boldsymbol{X}+\boldsymbol{Y}}(\boldsymbol{\theta}) = \chi_{\boldsymbol{X}}(\boldsymbol{\theta}) \cdot \chi_{\boldsymbol{Y}}(\boldsymbol{\theta}) \quad \text{for all } \boldsymbol{\theta} \in \mathbb{R}^n.$$

Exercise 21.29 (Multivariate normal distribution: Characteristic function). Consider a vector $\boldsymbol{Z} = (Z_1, \dots, Z_n)$ of *independent* standard normal random variables. Prove that

$$\chi_{\boldsymbol{Z}}(\boldsymbol{\theta}) = \mathrm{e}^{-\|\boldsymbol{\theta}\|_2^2/2} \quad \text{for all } \boldsymbol{\theta} \in \mathbb{R}^n.$$

Now, consider the multivariate normal distribution $\boldsymbol{X} = \boldsymbol{m} + \boldsymbol{\Sigma}\boldsymbol{Z}$. Show that

$$\chi_{\boldsymbol{X}}(\boldsymbol{\theta}) = \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\theta}, \boldsymbol{m}\rangle} \cdot \mathrm{e}^{-\boldsymbol{\theta}^* \boldsymbol{C} \boldsymbol{\theta}/2} \quad \text{for all } \boldsymbol{\theta} \in \mathbb{R}^n,$$

where $\boldsymbol{C} = \boldsymbol{\Sigma}\boldsymbol{\Sigma}^*$ is the covariance matrix.

The result of Exercise 21.29 is intriguing because it shows that the mean $\boldsymbol{m}$ and covariance $\boldsymbol{C}$ are the only aspects of the multivariate normal distribution that appear in the characteristic function. Ultimately, this fact will allow us to deduce that the multivariate normal distribution is completely determined by its mean and covariance.

21.3 Characterization of distributions

What is the source of the terminology "characteristic function"? It comes from the next group of results, which state that the characteristic function of a distribution determines the distribution completely.

Theorem 21.30 (Characteristic functions). Suppose that X and Y are real random variables with distributions μ_X and μ_Y. Then the distributions are identical ($\mu_X = \mu_Y$) if and only if the characteristic functions are identical ($\chi_X = \chi_Y$).

The key idea in the proof is to smooth the distributions of X and Y so that we can find an explicit representation of the densities in terms of the characteristic functions. Before continuing on to the proof, let us state some important facts that follow from closely related arguments.

Problem 21.31 (Multivariate characteristic functions). Suppose that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are random vectors taking values in $\mathbb{R}^n$. Then the distributions of $\boldsymbol{X}$ and $\boldsymbol{Y}$ are equal if and only if the multivariate characteristic functions are equal: $\chi_{\boldsymbol{X}} = \chi_{\boldsymbol{Y}}$. Prove this claim by generalizing the proof of Theorem 21.30.

Hint: The argument is structurally identical to the result for real random variables. The main difference is that we need to smooth the distribution using a multivariate standard normal distribution.

Exercise 21.32 (Linear marginals). Suppose that $\boldsymbol{X}$ and $\boldsymbol{Y}$ are random vectors taking values in $\mathbb{R}^n$. Show that $\boldsymbol{X}$ and $\boldsymbol{Y}$ share the same distribution if and only if $\langle \boldsymbol{a}, \boldsymbol{X}\rangle \sim \langle \boldsymbol{a}, \boldsymbol{Y}\rangle$ for every vector $\boldsymbol{a} \in \mathbb{R}^n$. In other words, a multivariate distribution is completely determined by its linear marginals. **Hint:** This result is an easy consequence of Problem 21.31.

21.3.1 Smoothing

Let X be a real random variable. Let Z be a real standard normal random variable, independent from X. For each $\sigma > 0$, we construct the smoothed random variable $X_\sigma := X + \sigma Z$.

Exercise 21.33 (Smoothed random variable: Approximation). Show that $X_\sigma \rightsquigarrow X$ weakly as $\sigma \to 0$. **Hint:** This is a consequence of the observation that $X_\sigma \to X$ pointwise.

When X and Y have the same characteristic function, we will prove that the smoothed random variables X_σ and Y_σ share the same distribution for each $\sigma > 0$. Since weak limits are unique (Exercise 17.14), the facts that $X_\sigma \rightsquigarrow X$ and $Y_\sigma \rightsquigarrow Y$ force X and Y to share the same distribution.

21.3.2 Inversion

The critical step in the argument is to write the distribution of the smoothed random variable X_σ in terms of its characteristic function. This result can be viewed as a Fourier inversion formula for the density of a particular type of random variable.

Proposition 21.34 (Smoothed random variable: Inversion). Let X_σ be the smoothed random variable defined above. The density f_{X_σ} of the random variable X_σ can be written as

$$f_{X_\sigma}(x) = \int_{\mathbb{R}} \chi_X(\theta) \cdot \chi_{\sigma Z}(\theta) \cdot \mathrm{e}^{-\mathrm{i}\theta x} \, \lambda(\mathrm{d}\theta) \quad \text{for all } x \in \mathbb{R}.$$

Proof. The basic result underlying the proof is a similar inversion formula for a standard normal random variable. According to Exercise 21.24, the standard normal density can be represented in terms of a characteristic function:

$$\frac{\mathrm{e}^{-z^2/2}}{\sqrt{2\pi}} = \frac{1}{2\pi} \int_{\mathbb{R}} \chi_Z(\theta) \cdot \mathrm{e}^{-\mathrm{i}\theta z} \, \lambda(\mathrm{d}\theta) \quad \text{for all } z \in \mathbb{R}.$$

We can obtain a similar result for the density φ_σ of the random variable σZ. To do so, we make the change of variables $z \mapsto z/\sigma$ in the last display. This step yields

$$\begin{aligned} \varphi_\sigma(z) := \frac{\mathrm{e}^{-z^2/(2\sigma^2)}}{\sqrt{2\pi\sigma^2}} &= \frac{1}{2\pi\sigma} \int_{\mathbb{R}} \chi_Z(\theta) \cdot \mathrm{e}^{-\mathrm{i}\theta(z/\sigma)} \, \lambda(\mathrm{d}\theta) \\ &= \frac{1}{2\pi} \int_{\mathbb{R}} \chi_{\sigma Z}(\theta) \cdot \mathrm{e}^{-\mathrm{i}\theta z} \, \lambda(\mathrm{d}\theta). \end{aligned}$$

The last identity follows from the change of variables $\theta \mapsto \sigma\theta$ and the affine transformation rule for cgfs (Exercise 21.20).

According to Exercise 14.4, the independent sum $X_\sigma = X + \sigma Z$ is an (absolutely) continuous random variable with density

$$f_{X_\sigma}(x) = \int_{\mathbb{R}} \varphi_\sigma(x - y) \, \mu_X(\mathrm{d}y) \quad \text{for all } x \in \mathbb{R}.$$

Our plan is to express φ_σ in terms of the characteristic function of σZ. After reorganizing the integrals, this computation will deliver the result. Combine the last two displays to obtain

$$\begin{aligned} f_{X_\sigma}(y) &= \int_{\mathbb{R}} \left[\frac{1}{2\pi} \int_{\mathbb{R}} \chi_{\sigma Z}(\theta) \cdot \mathrm{e}^{-\mathrm{i}\theta(x-y)} \, \lambda(\mathrm{d}\theta) \right] \mu_X(\mathrm{d}y) \\ &= \frac{1}{2\pi} \int_{\mathbb{R}} \left[\int_{\mathbb{R}} \mathrm{e}^{\mathrm{i}\theta y} \, \mu_X(\mathrm{d}y) \right] \cdot \chi_{\sigma Z}(\theta) \cdot \mathrm{e}^{-\mathrm{i}\theta x} \, \lambda(\mathrm{d}\theta) \\ &= \frac{1}{2\pi} \int_{\mathbb{R}} \chi_X(\theta) \cdot \chi_{\sigma Z}(\theta) \cdot \mathrm{e}^{-\mathrm{i}\theta x} \, \lambda(\mathrm{d}\theta). \end{aligned}$$

Since the integrand is bounded, we may apply Fubini–Tonelli (Theorem 6.23) to rearrange the integrals in the first line. We used the law of the unconscious statistician (Proposition 9.4) to identify the characteristic function χ_X of the random variable X. This is what we needed to show. ∎

21.3.3 Proof of Theorem 21.30

We are now prepared to establish Theorem 21.30, which states that the distribution of a real random variable is completely determined by its characteristic function.

It is clear that the equality $\mu_X = \mu_Y$ of the distributions implies the equality $\chi_X = \chi_Y$ of the characteristic functions. Indeed, the characteristic function just packs up a collection of moments of the (common) distribution.

For the reverse direction, assume that the characteristic functions coincide:

$$\chi_X(\theta) = \chi_Y(\theta) \quad \text{for all } \theta \in \mathbb{R}.$$

Let Z be a real standard normal random variable. For $\sigma > 0$, define the (absolutely) continuous random variables $X_\sigma := X + \sigma Z$ and $Y_\sigma := Y + \sigma Z$. Proposition 21.34 ensures that their respective densities f_{X_σ} and f_{Y_σ} also coincide:

$$\begin{aligned} f_{X_\sigma}(x) &= \int_{\mathbb{R}} \chi_X(\theta) \cdot \chi_{\sigma Z}(\theta) \cdot \mathrm{e}^{-\mathrm{i}\theta x} \, \lambda(\mathrm{d}x) \\ &= \int_{\mathbb{R}} \chi_Y(\theta) \cdot \chi_{\sigma Z}(\theta) \cdot \mathrm{e}^{-\mathrm{i}\theta x} \, \lambda(\mathrm{d}x) = f_{Y_\sigma}(x) \quad \text{for all } x \in \mathbb{R}. \end{aligned}$$

The density of a continuous random variable determines its distribution. We realize that X_σ and Y_σ have a common distribution, say μ_σ.

Now, Exercise 21.33 states that the distributions of the smoothed random variables converge weakly to the original distributions: $\mu_{X_\sigma} \rightsquigarrow \mu_X$ and $\mu_{Y_\sigma} \rightsquigarrow \mu_Y$ as $\sigma \to 0$. We have seen that the distributions of the smoothed variables both coincide with μ_σ, so $\mu_\sigma \rightsquigarrow \mu_X$ and $\mu_\sigma \rightsquigarrow \mu_Y$. But weak limits are unique (Exercise 17.14), so we must conclude that $\mu_X = \mu_Y$. The random variables X and Y share the same distribution.

21.3.4 *Characteristic functions and weak convergence

Characteristic functions have historically played a significant role in studying weak convergence because of the following equivalence.

Theorem 21.35 (Characteristic functions and weak convergence). Consider a sequence $(W_n : n \in \mathbb{N})$ of real random variables, and let W be a real random variable. Then

$$W_n \rightsquigarrow W \quad \text{if and only if} \quad \chi_{W_n}(\theta) \to \chi_W(\theta) \quad \text{pointwise.}$$

Note that it is necessary to assume that the limit is a random variable.

Using bounded convergence, it is quite easy to check that weak convergence implies pointwise convergence of the characteristic functions. The reverse direction requires some techniques from Fourier analysis, and it is somewhat more involved. We omit the proof, which you may find in most probability books.

Theorem 21.35 and its relatives lead to short proofs of distributional limit theorems for independent sums.

Problem 21.36 (CLT). Assume that $Y \in \mathsf{L}_2$ is a real random variable. Let T_n be the standardized sum of n i.i.d. copies of Y. Show that $\chi_{T_n} \to \chi_Z$ pointwise, where

$Z \sim \text{NORMAL}(0,1)$ is a standard normal random variable. Deduce the central limit theorem.

Problem 21.37 (Multivariate CLT). Assume that $\boldsymbol{Y}$ is a random vector taking values in $\mathbb{R}^k$, with square-integrable coordinates. Let $\boldsymbol{m}$ be the mean vector of $\boldsymbol{Y}$, and let $\boldsymbol{C}$ be the covariance matrix. Let $\boldsymbol{T}_n = n^{-1/2} \sum_{i=1}^n (\boldsymbol{Y}_i - \boldsymbol{m})$ be the normalized sum of n independent copies of the vector $\boldsymbol{Y}$. Prove that $\boldsymbol{T}_n$ converges weakly to a $\text{NORMAL}(\boldsymbol{0}, \boldsymbol{C})$ distribution.

Other types of limit theorems follow from similar considerations.

Problem 21.38 (Poisson limit of the binomial). Let $\beta \geq 0$. For all sufficiently large $n \in \mathbb{N}$, consider the random variables $Q_n \sim \text{BINOMIAL}(n, \beta/n)$. Prove that $Q_n \rightsquigarrow Q$, where Q follows the Poisson distribution with mean β.

21.4 Gaussians, independence, and conditioning

Now that we understand how to identify distributions using the (multivariate) characteristic function, we can undertake a deeper study of the multivariate normal distribution. Our objective in this section is to determine the conditional expectation of a normal random variable, given a family of normal random variables.

21.4.1 Existence and uniqueness

In this section, we use characteristic functions to argue the there is only one multivariate normal distribution with a given mean and covariance. This fact has some significant implications.

> **Theorem 21.39 (Multivariate normal: Existence and uniqueness).** Select a vector $\boldsymbol{m} \in \mathbb{R}^n$ and a positive-semidefinite matrix $\boldsymbol{C} \in \mathbb{R}^{n \times n}$. There is a unique normal distribution, $\text{NORMAL}(\boldsymbol{m}, \boldsymbol{C})$, on $\mathbb{R}^n$ with mean $\boldsymbol{m}$ and covariance matrix $\boldsymbol{C}$.

Proof. By definition, a multivariate normal random vector $\boldsymbol{X}$ on $\mathbb{R}^n$ is the affine image of a multivariate standard normal vector $\boldsymbol{Z} \in \mathbb{R}^n$. That is,

$$\boldsymbol{X} = \boldsymbol{m} + \boldsymbol{\Sigma} \boldsymbol{Z} \quad \text{where } \boldsymbol{m} \in \mathbb{R}^n \text{ and } \boldsymbol{\Sigma} \in \mathbb{R}^{n \times n}.$$

Define the positive-definite covariance matrix $\boldsymbol{C} = \boldsymbol{\Sigma} \boldsymbol{\Sigma}^*$.

According to Exercise 21.29, the characteristic function of $\boldsymbol{X}$ takes the form

$$\chi_{\boldsymbol{X}}(\boldsymbol{\theta}) = \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\theta}, \boldsymbol{m} \rangle} \mathrm{e}^{-\boldsymbol{\theta}^* \boldsymbol{C} \boldsymbol{\theta}/2} \quad \text{for all } \boldsymbol{\theta} \in \mathbb{R}.$$

Invoking Problem 21.31, we realize that the mean vector $\boldsymbol{m}$ and the covariance matrix $\boldsymbol{C}$ completely determine the distribution of the multivariate normal random vector $\boldsymbol{X}$.

It remains to confirm that there is a multivariate normal distribution with an arbitrary mean vector $\boldsymbol{m}$ and positive-semidefinite covariance matrix $\boldsymbol{C}$. Of course, the mean presents no difficulty. We can obtain the desired covariance by choosing $\boldsymbol{\Sigma} = \boldsymbol{C}^{1/2}$, the unique positive-semidefinite square-root of $\boldsymbol{C}$. ∎

21.4.2 Independence and rotational invariance

Theorem 21.39 has striking consequences.

> **Corollary 21.40 (Multivariate normal: Independence).** Consider a multivariate normal random vector $(X_1, \dots, X_n) \sim \text{NORMAL}(\boldsymbol{m}, \boldsymbol{C})$ on $\mathbb{R}^n$. Then the random variables $(X_1, \dots, X_n)$ compose an independent family if and only if $\boldsymbol{C}$ is diagonal. In other words, $(X_1, \dots, X_n)$ is independent if and only if $\text{Cov}(X_i, X_j) = 0$ for all $i \neq j$.

Proof. It is well-known that an independent pair (X, Y) of random variables necessarily satisfies $\operatorname{Cov}(X, Y) = 0$. The content of the result is the reverse direction.

Let $\boldsymbol{m} \in \mathbb{R}^n$, and suppose that $\boldsymbol{C} = \operatorname{diag}(c_{11}, \dots, c_{nn})$ is a diagonal, positive-semidefinite matrix. According to Theorem 21.39, the distribution $\textsc{normal}(\boldsymbol{m}, \boldsymbol{C})$ is uniquely determined. Therefore, it suffices to produce a multivariate normal vector $\boldsymbol{X} = (X_1, \dots, X_n)$ with independent coordinates and the specified mean and covariance. But this is trivial. For a standard normal vector $\boldsymbol{Z} = (Z_1, \dots, Z_n)$, we simply choose

$$X_j = m_j + \sqrt{c_{jj}}\, Z_j \quad \text{for } j = 1, \dots, n.$$

The coordinates $(X_1, \dots, X_n)$ compose an independent family because $(Z_1, \dots, Z_n)$ is independent. ∎

Exercise 21.41 (Multivariate standard normal: Rotational invariance). Let $\boldsymbol{Z}$ be a standard normal vector on $\mathbb{R}^n$. Let $\boldsymbol{U} \in \mathbb{R}^{n \times n}$ be an orthogonal matrix. Show that the random vector $\boldsymbol{UZ}$ remains standard normal. **Hint:** Compute the covariance matrix.

21.4.3 Linear marginals

Another important property of the multivariate normal distribution is that each linear marginal follows a normal distribution on the real line. Conversely, a distribution whose linear marginals are all normal must be a multivariate normal distribution.

Theorem 21.42 (Multivariate normal: Linear marginals). Let $\boldsymbol{X}$ be a multivariate normal random distribution on $\mathbb{R}^n$. For every vector $\boldsymbol{a} \in \mathbb{R}^n$, the linear marginal $\langle \boldsymbol{a}, \boldsymbol{X} \rangle$ follows a normal distribution on $\mathbb{R}$.

Conversely, suppose that $\boldsymbol{X}$ is a distribution taking values on $\mathbb{R}^n$. Suppose that $\langle \boldsymbol{a}, \boldsymbol{X} \rangle$ follows a normal distribution for every vector $\boldsymbol{a} \in \mathbb{R}^n$. Then $\boldsymbol{X}$ must be a multivariate normal distribution.

Proof. Suppose that $\boldsymbol{X} \sim \textsc{normal}(\boldsymbol{m}, \boldsymbol{C})$ on $\mathbb{R}^n$. Exercise 21.29 states that the characteristic function is

$$\chi_{\boldsymbol{X}}(\boldsymbol{\theta}) = \mathbb{E}[\mathrm{e}^{\mathrm{i}\langle \boldsymbol{\theta}, \boldsymbol{X} \rangle}] = \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\theta}, \boldsymbol{m} \rangle} \mathrm{e}^{-\boldsymbol{\theta}^* \boldsymbol{C} \boldsymbol{\theta}/2} \quad \text{for all } \boldsymbol{\theta} \in \mathbb{R}^n.$$

Let $Y = \langle \boldsymbol{a}, \boldsymbol{X} \rangle$ for some vector $\boldsymbol{a} \in \mathbb{R}^n$. Reading from the last display, we see that the characteristic function of Y must satisfy

$$\chi_Y(\theta) = \mathbb{E}[\mathrm{e}^{\mathrm{i}\theta Y}] = \mathrm{e}^{\mathrm{i}\theta m_Y} \mathrm{e}^{-\theta^2 v_Y/2},$$

where $m_Y = \langle \boldsymbol{a}, \boldsymbol{m} \rangle$ and $v_Y = \boldsymbol{a}^* \boldsymbol{C} \boldsymbol{a}$. From Example 21.23, we recognize the characteristic function of a real $\textsc{normal}(m_Y, v_Y)$ random variables. According to Theorem 21.30, we must conclude that Y follows a normal distribution on $\mathbb{R}$.

Conversely, suppose that $\boldsymbol{X}$ is a distribution on $\mathbb{R}^n$ whose linear marginals each follow a standard normal distribution. Introduce the mean $\boldsymbol{m} = \mathbb{E}\,\boldsymbol{X}$ and covariance matrix $\boldsymbol{C} = \mathbb{E}[(\boldsymbol{X} - \boldsymbol{m})(\boldsymbol{X} - \boldsymbol{m})^*]$ of the distribution. Now, for a vector $\boldsymbol{a} \in \mathbb{R}^n$, consider the real normal random variable $Y = \langle \boldsymbol{a}, \boldsymbol{X} \rangle$. A short calculation shows that the mean and variance of Y must satisfy

$$m_Y = \mathbb{E}[Y] = \langle \boldsymbol{a}, \boldsymbol{m} \rangle \quad \text{and} \quad v_Y = \operatorname{Var}[Y] = \boldsymbol{a}^* \boldsymbol{C} \boldsymbol{a}.$$

Since the mean and variance determine a univariate normal distribution, the characteristic function of Y satisfies

$$\chi_Y(\theta) = \mathbb{E}[\mathrm{e}^{\mathrm{i}\theta Y}] = \mathrm{e}^{\mathrm{i}m_Y \theta} \mathrm{e}^{-v_Y^2 \theta^2/2}.$$

Changing variables $\theta \boldsymbol{a} \mapsto \boldsymbol{\theta}$ and introducing the computed mean and variance, we find that

$$\mathbb{E}[\mathrm{e}^{\mathrm{i}\langle \boldsymbol{\theta},\, \boldsymbol{X}\rangle}] = \mathrm{e}^{\mathrm{i}\langle \boldsymbol{\theta},\, \boldsymbol{m}\rangle} \mathrm{e}^{-\boldsymbol{\theta}^* \boldsymbol{C} \boldsymbol{\theta}/2} \quad \text{for all } \boldsymbol{\theta} \in \mathbb{R}^n.$$

Exercise 21.29 shows that this is the characteristic function of a multivariate normal distribution. Problem 21.31 now compels us to conclude that $\boldsymbol{X}$ itself is multivariate normal. ∎

Theorem 21.42 allow us to work with multivariate normal distributions by focusing on the marginals. This is a very useful technique, and it suggests an alternative definition of a multivariate normal distribution.

Definition 21.43 (Jointly Gaussian distribution). A random vector $\boldsymbol{X}$ taking values in $\mathbb{R}^n$ is called *jointly Gaussian* when the linear marginal $\langle \boldsymbol{a},\, \boldsymbol{X}\rangle$ follows a univariate normal distribution for every vector $\boldsymbol{a} \in \mathbb{R}^n$.

Warning 21.44 (Jointly Gaussian). To satisfy Definition 21.43, every single linear marginal of the random vector $\boldsymbol{X}$ must follow a normal distribution. It is possible for a random vector to have some Gaussian marginals without being jointly Gaussian. In particular, the sole conditions $X \sim \textsc{normal}(0, 1)$ and $Y \sim \textsc{normal}(0, 1)$ do not imply that (X, Y) is jointly Gaussian! ∎

Aside: Definition 21.43 provides the right framework for extending the concept of a Gaussian distribution to an infinite-dimensional linear space.

21.4.4 Gaussian least squares

Next, we study the solution to least-squares problems involving multivariate normal distributions. The simple structure of the distribution allows for clean statements of the result.

Example 21.45 (Multivariate normal distribution: Least squares). Consider a jointly Gaussian family $(X, \boldsymbol{Y})$ of real random variables where $\boldsymbol{Y} = (Y_1, \dots, Y_n)$. For simplicity, we will assume that these random variables all have mean zero. Now, frame the least-squares problem:

$$\operatorname{minimize}_{\boldsymbol{a} \in \mathbb{R}^n} \quad \|X - \langle \boldsymbol{a},\, \boldsymbol{Y}\rangle\|_2^2.$$

In other words, we seek the best L_2 approximation of X as a linear function of the vector $\boldsymbol{Y}$.

We can solve this problem by differential calculus. Let $c_{XX} = \operatorname{Var}[X]$, and define the covariance vector and matrix:

$$\boldsymbol{c}_{\boldsymbol{XY}} = (\mathbb{E}[XY_i] : i = 1, \dots, n) \in \mathbb{R}^n;$$
$$\boldsymbol{C}_{\boldsymbol{YY}} = (\mathbb{E}[Y_i Y_j] : i, j = 1, \dots, n) \in \mathbb{R}^{n \times n}.$$

Writing the norm as an expectation and expanding the square, we detect that

$$\begin{aligned} \|X - \langle \boldsymbol{a},\, \boldsymbol{Y}\rangle\|_2^2 &= \operatorname{Var}[X] - 2\,\mathbb{E}[X\langle \boldsymbol{a},\, \boldsymbol{Y}\rangle] + \operatorname{Var}[\langle \boldsymbol{a},\, \boldsymbol{Y}\rangle] \\ &= c_{XX} - 2\langle \boldsymbol{a},\, \boldsymbol{c}_{\boldsymbol{XY}}\rangle + \boldsymbol{a}^* \boldsymbol{C}_{\boldsymbol{YY}} \boldsymbol{a}. \end{aligned}$$

Set the derivative with respect to $\boldsymbol{a}$ to zero, and rearrange to deduce that there is a minimizer of the form $\boldsymbol{a} = \boldsymbol{C}_{\boldsymbol{YY}}^{\dagger} \boldsymbol{c}_{\boldsymbol{XY}}$. It follows that the best approximation $\hat{X}$ of X as a linear function of $\boldsymbol{Y}$ is the random variable

$$\hat{X} = \boldsymbol{c}_{\boldsymbol{XY}}^* \boldsymbol{C}_{\boldsymbol{YY}}^{\dagger} \boldsymbol{Y}.$$

We write $\boldsymbol{C}^{\dagger}$ for the *pseudoinverse* of the matrix $\boldsymbol{C}$. It reduces to the ordinary inverse when $\boldsymbol{C}$ is invertible.

In particular, the norm of the residual satisfies

$$\|X - \hat{X}\|_2^2 = c_{XX} - \boldsymbol{c}_{X\boldsymbol{Y}}^* \boldsymbol{C}_{\boldsymbol{YY}}^\dagger \boldsymbol{c}_{X\boldsymbol{Y}}.$$

Observe that the residual norm is a Schur complement of the covariance matrix. This is not an accident.

Let us emphasize that the coefficients in the formula for $\hat{X}$ are simply numbers that depend on the joint statistics of $(X, \boldsymbol{Y})$.

So far, we have not used the assumption that the random variables are Gaussian. From the orthogonal projection theorem (Theorem 12.21), recall that the residual $X - \hat{X}$ in the least-squares problem must be orthogonal to the subspace of random variables that can be written as linear combinations of the components of $(Y_1, \dots, Y_n)$. That is,

$$\mathbb{E}[(X - \hat{X})\langle \boldsymbol{u}, \boldsymbol{Y}\rangle] = 0 \quad \text{for all } \boldsymbol{u} \in \mathbb{R}^n.$$

But $(X - \hat{X})$ and $\langle \boldsymbol{u}, \boldsymbol{Y}\rangle$ are jointly Gaussian (because they are fixed linear combinations of jointly Gaussian random variables). From Corollary 21.40, we deduce that the residual $(X - \hat{X})$ is *independent* from $\langle \boldsymbol{u}, \boldsymbol{Y}\rangle$ for each $\boldsymbol{u} \in \mathbb{R}^n$. *A fortiori*, the residual $(X - \hat{X})$ is independent from the Gaussian vector $\boldsymbol{Y}$. We will exploit this observation in the next section. ∎

Exercise 21.46 (Multivariate normal: Least squares). In Example 21.45, consider the more general case where the random variables may have nonzero expectations. Solve the affine least-squares problem:

$$\text{minimize}_{\boldsymbol{a} \in \mathbb{R}^n, b \in \mathbb{R}} \quad \|X - (b + \langle \boldsymbol{a}, \boldsymbol{Y}\rangle)\|_2^2.$$

Find representations for the optimal coefficients $\boldsymbol{a}$ and b. What can we deduce about the relationship between the residual and the random vector $\boldsymbol{Y}$?

21.4.5 Gaussian conditioning

Finally, we are in a position to compute conditional expectations in the setting of a jointly Gaussian distribution. This situation allows for a dramatic simplification over the general case.

Theorem 21.47 (Multivariate normal distribution: Conditional expectation). Consider a jointly Gaussian family $(X, \boldsymbol{Y})$ of real random variables where $\boldsymbol{Y} = (Y_1, \dots, Y_n)$. For simplicity, we will assume that each of these random variables has mean zero. Then, almost surely, conditional expectation of X given $\boldsymbol{Y}$ takes the form

$$\mathbb{E}[X \,|\, \boldsymbol{Y}] = \boldsymbol{c}_{X\boldsymbol{Y}}^* \boldsymbol{C}_{\boldsymbol{YY}}^{-1} \boldsymbol{Y},$$

where $\boldsymbol{c}_{X\boldsymbol{Y}} \in \mathbb{R}^n$ is the covariance vector of X and $\boldsymbol{Y}$ and $\boldsymbol{C}_{\boldsymbol{YY}} \in \mathbb{R}^{n \times n}$ is the covariance matrix of $\boldsymbol{Y}$.

Theorem 21.47 is a truly remarkable result. In a general setting, the conditional expectation $\mathbb{E}[X \,|\, \boldsymbol{Y}]$ is a *measurable function* of $\boldsymbol{Y}$. Yet, in the Gaussian setting, the conditional expectation is a *linear function* of $\boldsymbol{Y}$. This is a dramatic conceptual simplification, and it supports the use of linear least-squares procedures in statistical estimation.

Proof. This result follows from the computation in Example 21.45 and the properties of conditional expectation. Let $\hat{X} = \boldsymbol{c}_{X\boldsymbol{Y}}^* \boldsymbol{C}_{\boldsymbol{YY}}^\dagger \boldsymbol{Y}$ be the best approximation of X as a linear function of $\boldsymbol{Y}$. Recall that the residual $(X - \hat{X})$ is *independent* from $\boldsymbol{Y}$. Therefore, we

may calculate that

$$\begin{aligned}\mathbb{E}[X \mid \boldsymbol{Y}] &= \mathbb{E}[(X - \hat{X}) + \hat{X} \mid \boldsymbol{Y}] \\ &= \mathbb{E}[(X - \hat{X}) \mid \boldsymbol{Y}] + \mathbb{E}[\hat{X} \mid \boldsymbol{Y}] \\ &= \mathbb{E}[X - \hat{X}] + \hat{X} = \hat{X}.\end{aligned}$$

To reach the second line, we apply the linearity of conditional expectation. Since $(X - \hat{X})$ is independent from $\boldsymbol{Y}$, its conditional expectation coincides with its complete expectation. But X and $\hat{X}$ are both mean zero, so this term vanishes by linearity. As for the second term, the estimate $\hat{X}$ is a function of $\boldsymbol{Y}$, so it is $\sigma(\boldsymbol{Y})$-measurable. Thus, its conditional expectation equals itself. ∎

Exercise 21.48 (Multivariate normal: Conditional expectation). Extend the computation in Theorem 21.47 to the case where $(X, \boldsymbol{Y})$ may have nonzero expectations.

Problems

Exercise 21.49 (Stability). Characteristic functions provide an easy way to check stability properties of probability distributions.

1. Let (X, Y) be an *independent* pair of Poisson random variables, perhaps with different means. Show that $X + Y$ is Poisson by computing characteristic functions and invoking Theorem 21.30.
2. Let (X, Y) be an *independent* pair of Cauchy random variables, perhaps with different means and scale parameters. Show that $X + Y$ is Cauchy.

Exercise 21.50 (Combinations of Gaussians). We can check that two distributions are equal by checking that their characteristic functions are equal. Alternatively, we can check that all of the linear marginals have the same distribution. There are many situations where these are the most efficient paths to identifying a distribution. Recall that a pair (X, Y) of random variables is *jointly Gaussian* if $aX + bY$ follows a normal distribution for all $a, b \in \mathbb{R}$. Jointly Gaussian variables are special.

1. Suppose that a real random variable X has the characteristic function $\chi_X(\theta) = \exp(\mathrm{i}m\theta - v\theta^2/2)$ for all $\theta \in \mathbb{R}$. What is the distribution of X?
2. Let $X \sim \textsc{normal}(m_X, v_X)$ and $Y \sim \textsc{normal}(m_Y, v_Y)$ be *independent*. By calculating the characteristic function, show that the sum $X + Y \sim \textsc{normal}(m_X + m_Y, v_X + v_Y)$. Explain why (X, Y) is jointly Gaussian.
3. Assume that (X, Y) is jointly Gaussian. Show that $\mathrm{Cov}(X, Y) = 0$ if and only if X and Y are independent. **Hint:** Look at the linear marginals.
4. Let (X, Y) be jointly Gaussian. Compute $\mathbb{E}[X \mid Y]$ in terms of the mean and covariance of (X, Y). **Hint:** Prove that the conditional expectation is an affine function of Y.

Applications

Application 21.51 (Kalman). The Kalman filter applies the Bayesian framework to track a random state variable X_t that is evolving with time. The input to the Kalman filter is a sequence (Y_t) of noisy observations that are (indirectly) related to the state X_t. The Kalman filter is used by self-driving vehicles to track their positions and to map their environments. Nonlinear extensions of the Kalman filter arise in a multitude of other domains, such as weather predictions in geophysics and tracking virus populations in

epidemiology. In this problem, we will analyze the stability and error properties of the Kalman filter.

Assume that the initial state follows a normal distribution: $X_0 \sim \textsc{normal}(m_0, c_0)$ with known mean and variance. Consider a discrete-time dynamical system that links the (scalar) state variable X_t and the scalar observations Y_t. For $t \in \mathbb{N}$,

$$\begin{aligned} X_t &= \lambda X_{t-1} + \sigma_X \varepsilon_t \\ Y_t &= X_t + \sigma_Y \eta_t, \end{aligned}$$

where $\lambda \in \mathbb{R}$ is a known stability parameter and $\sigma_X, \sigma_Y > 0$ are fixed standard deviations, assumed known. The forecast and observational noise variables follow $\varepsilon_t \sim \textsc{normal}(0, 1)$ and $\eta_t \sim \textsc{normal}(0, 1)$. The noise variables are drawn independently from X_t for all t, and they are treated as unknown.

The Kalman filter tracks the conditional distribution of X_t given the measurements $Y_1, \dots, Y_t$. We denote the conditional distribution as $\mathcal{L}_t := \mathcal{L}(X_t \mid Y_1, \dots, Y_t)$. The filter computes the distribution $\mathcal{L}_t$ recursively in terms of the previous distribution $\mathcal{L}_{t-1}$ and the new observation Y_t. The update is performed in two stages. First, the *forecast step* describes $\mathcal{L}(X_t \mid Y_1, \dots, Y_{t-1})$ using $\mathcal{L}_{t-1}$ and the model dynamics. Second, the *analysis step* uses Bayesian inference to incorporate the latest measurement Y_t.

1. First, a warmup. Assume the law $\mathcal{L}(X) = \textsc{normal}(m, u)$ and the law $\mathcal{L}(Y \mid X) = \textsc{normal}(X, w)$, where the variances $u, w > 0$. Use Bayes's rule to verify that

$$\mathcal{L}(X \mid Y) = \textsc{normal}(\hat{X}, v) \quad \text{where} \quad \frac{1}{v} = \frac{1}{u} + \frac{1}{w} \quad \text{and} \quad \frac{\hat{X}}{v} = \frac{m}{u} + \frac{Y}{w}.$$

2. The next task is to show by induction that $\mathcal{L}_t \sim \textsc{normal}(m_t, c_t)$ for each $t \in \mathbb{N}$. To do so, write out the distribution $\mathcal{L}(X_t \mid Y_1, \dots, Y_{t-1})$ after the forecast step. Then use Bayes's rule to calculate the distribution $\mathcal{L}_t$ after the analysis step.
3. Give explicit formulas for the evolution $f : m_t \mapsto m_{t+1}$ of the filtering mean and the evolution $g : c_t \mapsto c_{t+1}$ of the filtering variance.
4. Deduce that $c_t \le \sigma_Y^2$ for all t, regardless of the observations (Y_t).

For the rest of the problem, we will assume that the dynamics are deterministic ($\sigma_X = 0$), while the observations are contaminated with noise with a constant standard deviation ($\sigma_Y > 0$).

5. Derive the form of the asymptotic filtering variance $c_\infty := \lim_{t\to\infty} c_t$ by computing a fixed point of the evolution equation. Describe what happens when the dynamics are stable ($|\lambda| \le 1$) and when they are unstable ($|\lambda| > 1$).
6. Regardless of whether the dynamical system is stable or unstable, show that the asymptotic filtering variance computed in (d) is globally asymptotically stable. That is, $|g'(c_\infty)| < 1$.
7. Last, we consider the evolution of the filtering mean for a fixed state sequence $X^*_{t+1} = \lambda X^*_t$ where X^*_0 is deterministic. Define the error $E_t := m_t - X^*_t$ for each $t \in \mathbb{N}$. Using the previous results, show that the expected error in the filtering mean with respect to the noise in the observations satisfies

$$\lim_{t\to\infty} \left| \frac{\mathbb{E}[E_{t+1}]}{\mathbb{E}[E_t]} \right| < 1.$$

Therefore, the expected error in the filtering mean converges to zero. Altogether, these results show that we can reliably track the evolution of an (unstable) dynamical system from noisy observations.

Notes

The material on characteristic functions can be found in most books on probability theory. Application 21.51 was written by Dr. Ricardo Baptista.

22. *Densities

> "**Dr. Peter Venkman:** Ray, pretend for a moment that I don't know anything about metallurgy, engineering, or physics, and just tell me what the hell is going on."
>
> —*Ghostbusters*, 1984

Agenda:
1. Densities
2. Absolute continuity
3. Radon–Nikodým theorem
4. Conditional expectation as a density
5. Lebesgue decomposition theorem

In Lecture 19, we introduced the concept of conditional expectation by considering the best least-squares approximation of a random variable given some additional data. This construction shows that the conditional expectation of a square-integrable random variable has two properties (measurability and consistency) that characterize it.

In Lecture 20, we defined the conditional expectation of an integrable random variable by means of the two characteristic properties. Then, we constructed the general conditional expectation by approximating an integrable random variable by a square-integrable random variable and taking limits.

In this lecture, we will present another perspective on conditional expectation based on the notion of densities and justified by the Radon–Nikodým theorem from measure theory. This is the classical approach to conditional expectation. It provides yet another way to think about what conditional expectation means. Furthermore, it is valuable to understand the concept of the relative density of a pair of measures, which arises in other applications of probability theory.

22.1 The relative density of two measures

To motivate the idea of a relative density, we begin with the elementary conditional expectation of a random variable given an event. This definition suggests that we might be able to view conditional expectation as a ratio of measures. With this idea in place, we recall the notion of a density, and we present the Radon–Nikodým theorem.

22.1.1 Elementary conditional expectation

Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. Let $X \in \mathsf{L}_1(\Omega, \mathcal{F}, \mathbb{P})$ be an integrable real random variable. For conceptual simplicity, we assume that the random variable is positive: $X \geq 0$. Consider an event $\mathsf{G} \in \mathcal{F}$ with strictly positive probability: $\mathbb{P}(\mathsf{G}) > 0$.

The elementary conditional expectation of the random variable X given the event G is given by the formula

$$\mathbb{E}[X \,|\, \mathsf{G} \text{ occurs}] := \frac{\mathbb{E}[X \mathbb{1}_{\mathsf{G}}]}{\mathbb{P}(\mathsf{G})}.$$

Note that this conditional expectation is a *number*, not a random variable, because it reflects the concrete situation where the event G has occurred.

We can interpret the conditional expectation as a ratio of the expectation of X on the event G against the probability of the event G. Since X is positive, this looks like

a ratio of two measures applied to the event G. This point suggests the prospect of constructing the conditional expectation as a relative density of two measures.

What is the connection with Kolmogorov's definition of conditional expectation? Suppose that Y is a version of the conditional expectation $\mathbb{E}[X \mid \sigma(\mathsf{G})]$. Then

$$\mathbb{E}[X \mid \mathsf{G} \text{ occurs}] = \frac{\mathbb{E}[X \mathbb{1}_{\mathsf{G}}]}{\mathbb{P}(\mathsf{G})} = \frac{\mathbb{E}[Y \mathbb{1}_{\mathsf{G}}]}{\mathbb{P}(\mathsf{G})}.$$

The second relation is the consistency property of the conditional expectation. In other words, elementary conditional expectation given an event is the average of the abstract conditional expectation over the event. Note that we must normalize by the probability of the event to make this formulation correct. One of the insights behind Kolmogorov's construction is that the normalization prevents us from defining the conditional expectation over negligible sets, even though we may need to do so (e.g., when conditioning on a continuous random variable).

22.1.2 Densities

It is easiest to discuss densities using measures, rather than limiting our attention to probabilities. Let us recall what it means for one measure to have a density with respect to another, along with the basic facts about this construction.

Suppose that μ is a (finite) measure on the measurable space $(\Omega, \mathcal{F})$. Consider a positive, μ-integrable function $f \in \mathsf{L}_1(\mu)$ with $f \geq 0$. We can construct another (finite) measure ν on the same space using an integral:

$$\nu(\mathsf{E}) := \int_{\mathsf{E}} f \, \mathrm{d}\mu \quad \text{for } \mathsf{E} \in \mathcal{F}. \tag{22.1}$$

It is an easy matter to check that ν is a measure by means of Tonelli's theorem for sums (Exercise 5.39). When (22.1) holds, we say that the function f is a *density* of the measure ν with respect to the measure μ.

Integrals with respect to the measure ν satisfy the formula

$$\int g \, \mathrm{d}\nu = \int g f \, \mathrm{d}\mu \quad \text{for each } g \in \mathsf{L}_1(\nu).$$

This point follows from an application of Fubini–Tonelli (Theorem 6.23). In view of this formula, it is natural to use a differential notation for the density:

$$f = \frac{\mathrm{d}\nu}{\mathrm{d}\mu}.$$

This notation is compatible with the intuition that we are changing measures from μ to ν by "cancellation." But it is worth emphasis that *the density is a function* (defined on points of the domain Ω), and not a measure.

Densities are essentially unique. If the function h is another density of ν with respect to μ, then $f = h$ μ-almost everywhere. This point requires a standard, but nontrivial, measure theory argument. It is similar to the proof that integrals are almost positive (Theorem 5.14).

The measures μ and ν can distribute mass in rather disparate ways. Nevertheless, for a measurable set $\mathsf{E} \in \mathcal{F}$, it must be the case that $\mu(\mathsf{E}) = 0$ implies that $\nu(\mathsf{E}) = 0$. This is a consequence of the definition of the integral. As we will see, this property characterizes when ν has a density with respect to μ.

22.1.3 Densities: Examples

The most common situations where we encounter densities are in elementary discrete and continuous probability.

Example 22.1 (Counting measure: Densities). Consider the measurable space $(\mathbb{Z}_+, \mathscr{P}(\mathbb{Z}_+))$. Recall that the counting measure $\#\mathsf{E}$ reports the cardinality of a set $\mathsf{E} \subseteq \mathbb{Z}_+$. Let $(p_n : n \in \mathbb{N})$ be a sequence of positive numbers, and consider the measure

$$\nu(\mathsf{E}) = \sum\nolimits_{n \in \mathsf{E}} p_n \quad \text{for } \mathsf{E} \subseteq \mathbb{Z}_+.$$

The measure ν has density $\boldsymbol{p} = (p_n : n \in \mathbb{Z}_+)$ with respect to the counting measure.

Familiar cases include the standard discrete probability distributions. For example, $p_n = q^{-n}/(1-q)$ is the density of the $\textsc{geometric}(q)$ distribution. Meanwhile, $p_n = \mathrm{e}^{-\beta}\beta^n/n!$ is the density of the $\textsc{poisson}(\beta)$ distribution. ▪

Example 22.2 (Lebesgue measure: Densities). Consider the measurable space $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. Recall that the Lebesgue measure λ reports the total length of a Borel set. Let $f \geq 0$ be a positive, Borel measurable function, and consider the measure

$$\nu(\mathsf{B}) = \int_{\mathsf{B}} f \, \mathrm{d}\lambda \quad \text{for } \mathsf{B} \in \mathscr{B}(\mathbb{R}).$$

The measure ν has density f with respect to the Lebesgue measure.

Familiar examples include "continuous" probability distributions. For example, $f = \mathbb{1}_{[0,1]}$ is the density of the $\textsc{uniform}[0,1]$ distribution. Meanwhile, the function $f(x) = (2\pi)^{-1/2}\mathrm{e}^{-x^2/2}$ is the density of the standard normal distribution. ▪

22.1.4 The Radon–Nikodým theorem

In this section, we present an important theorem that describes when one measure has a density with respect to another. First, we give a definition.

Definition 22.3 (Absolutely continuous). Let μ and ν be two measures on the same measurable space. Suppose that, for any measurable set E, the condition $\mu(\mathsf{E}) = 0$ implies that $\nu(\mathsf{E}) = 0$. Then we say that ν is *absolutely continuous* with respect to μ, and we write $\nu \ll \mu$.

As we have seen, when ν has a density with respect to μ, then ν is absolutely continuous with respect to μ. The converse of this statement is also valid.

Theorem 22.4 (Radon–Nikodým). Suppose that μ, ν are σ-finite measures on the same measurable space. If ν is absolutely continuous with respect to μ (that is, $\nu \ll \mu$), then

$$\nu(\mathsf{E}) = \int_{\mathsf{E}} f \, \mathrm{d}\mu \quad \text{for all measurable } \mathsf{E},$$

where f is a positive, real-valued, measurable function. The function f is determined μ-almost everywhere.

We will establish this theorem in Section 22.3.

As before, we can use the differential notation to express the density $f = \mathrm{d}\nu/\mathrm{d}\mu$ that is promised by Theorem 22.4. The density $f = \mathrm{d}\nu/\mathrm{d}\mu$ is often called the *Radon–Nikodým derivative* of ν with respect to μ.

Even though $\mathrm{d}\nu/\mathrm{d}\mu$ is a function and not a measure, the notation hints at the role of absolute continuity. It is natural to insist that $\mu(\mathsf{E}) = 0$ whenever $\nu(\mathsf{E}) = 0$, or else the ratio would formally be infinite for points in E.

The Radon–Nikodým theorem arises in contexts where we change measures or compare two measures. Our purpose in introducing the result is to give an alternative construction of the conditional expectation.

22.1.5 Absolute continuity: Examples

To get a feel for what absolute continuity means, we can consider some simple examples.

Example 22.5 (Absolute continuity: Discrete case). Once again, consider the measurable space $(\mathbb{Z}_+, \mathscr{P}(\mathbb{Z}_+))$. Generically, any two measures on this space can be written as

$$\mu(\mathsf{E}) = \sum\nolimits_{n\in\mathsf{E}} p_n \quad \text{and} \quad \nu(\mathsf{E}) = \sum\nolimits_{n\in\mathsf{E}} q_n \quad \text{for } \mathsf{E} \subseteq \mathbb{Z}_+,$$

where $(p_n : n \in \mathbb{Z}_+)$ and $(q_n : n \in \mathbb{Z}_+)$ are sequences that take values in $\overline{\mathbb{R}}_+$.

The measure ν is absolutely continuous with respect to μ if and only if $p_n = 0$ implies that $q_n = 0$ for every index n. In other words, $\nu \ll \mu$ if and only if $\operatorname{supp}(\boldsymbol{q}) \subseteq \operatorname{supp}(\boldsymbol{p})$. ∎

The *support* of a sequence $\boldsymbol{a} = (a_n)$ is $\operatorname{supp}(\boldsymbol{a}) := \{n : a_n \neq 0\}$.

Example 22.6 (Absolute continuity: Lebesgue case). Now, consider the measurable space $(\mathbb{R}, \mathscr{B}(\mathbb{R}))$. We introduce two measures that have densities with respect to the Lebesgue measure:

$$\mu(\mathsf{B}) = \int_{\mathsf{B}} f \,\mathrm{d}\lambda \quad \text{and} \quad \nu(\mathsf{B}) = \int_{\mathsf{B}} g \,\mathrm{d}\lambda \quad \text{for } \mathsf{B} \in \mathscr{B}(\mathbb{R}).$$

The densities $f, g : \mathbb{R} \to \overline{\mathbb{R}}_+$ are Borel measurable.

For the measure ν to be absolutely continuous with respect to μ, it is *sufficient* that $f(a) = 0$ implies $g(a) = 0$ for all $a \in \mathbb{R}$. That is, $\operatorname{supp}(g) \subseteq \operatorname{supp}(f)$ implies that $\nu \ll \mu$. ∎

The *support* of a real-valued function f is $\operatorname{supp}(f) := \{a : f(a) \neq 0\}$.

Exercise 22.7 (Absolute continuity: Lebesgue case). Find a necessary and sufficient condition for $\nu \ll \mu$ in Example 22.6.

Warning 22.8 (Continuous versus absolutely continuous). Recall that a real random variable X is continuous when it has a density with respect to the Lebesgue measure. Equivalently, the law μ_X is absolutely continuous with respect to the Lebesgue measure: $\mu_X \ll \lambda$.

Note, however, that $X : \Omega \to \mathbb{R}$ might not be a continuous function. (Indeed, the sample space Ω need not have a topology).

The distribution function of a continuous random variable is always a continuous function. Nevertheless, there are random variables that have continuous distribution functions that do not have a density with respect to the Lebesgue measure. ∎

22.2 Conditional expectation: Construction via densities

Let us return to the matter of constructing the conditional expectation by means of a density. We can accomplish this goal by applying the Radon–Nikodým theorem to the ratio of an expectation against a probability that arises from the elementary definition of conditional expectation.

Theorem 22.9 (Conditional expectation: Fundamental theorem, redux). Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$, and let $\mathcal{G} \subseteq \mathcal{F}$ be a σ-algebra on the sample space Ω. Let $X \in \mathsf{L}_1(\Omega, \mathcal{F}, \mathbb{P})$ be a real random variable.

As in Definition 20.1, there exists a version Y of the conditional expectation $\mathbb{E}[X \,|\, \mathcal{G}]$. Furthermore, if Y' is another version of the conditional expectation, then $Y = Y'$ almost surely. That is, $\mathbb{P}\{Y \neq Y'\} = 0$.

The uniqueness claim in Theorem 22.9 follows from a direct argument, just as it did in Theorem 20.2. Let us present an independent proof of the existence result.

Proof. *On the measurable space* $(\Omega, \mathcal{G})$, we construct two finite measures

$$\begin{aligned} \nu_+(\mathsf{G}) &:= \int_{\mathsf{G}} X_+ \,\mathrm{d}\mathbb{P} = \mathbb{E}[X_+ \mathbb{1}_{\mathsf{G}}]; \\ \nu_-(\mathsf{G}) &:= \int_{\mathsf{G}} X_- \,\mathrm{d}\mathbb{P} = \mathbb{E}[X_- \mathbb{1}_{\mathsf{G}}] \end{aligned} \qquad \text{for all } \mathsf{G} \in \mathcal{G}. \tag{22.2}$$

Let us emphasize that the events G here belong to $\mathcal{G}$, the σ-algebra on which we are conditioning. As usual, $X_\pm$ are the positive and negative parts of the random variable X, which are both $\mathcal{F}$-measurable (but typically not $\mathcal{G}$-measurable).

We can verify that both ν_+ and ν_- are finite measures using Tonelli's theorem (Exercise 5.39). Furthermore, by definition of the Lebesgue integral, the condition $\mathbb{P}(\mathsf{G}) = 0$ implies that $\nu_+(\mathsf{G}) = 0$ and $\nu_-(\mathsf{G}) = 0$. Therefore, each of the two measures is absolutely continuous with respect to $\mathbb{P}$ on the measurable space $(\Omega, \mathcal{G})$.

An application of Theorem 22.4 delivers two positive, $\mathcal{G}$-measurable functions Y_+ and Y_- for which

$$\begin{aligned} \nu_+(\mathsf{G}) &= \int_{\mathsf{G}} Y_+ \,\mathrm{d}\mathbb{P} = \mathbb{E}[Y_+ \mathbb{1}_{\mathsf{G}}]; \\ \nu_-(\mathsf{G}) &= \int_{\mathsf{G}} Y_- \,\mathrm{d}\mathbb{P} = \mathbb{E}[Y_- \mathbb{1}_{\mathsf{G}}] \end{aligned} \qquad \text{for all } \mathsf{G} \in \mathcal{G}. \tag{22.3}$$

We may now construct the random variable $Y := Y_+ - Y_-$, which will serve as a version of the conditional expectation $\mathbb{E}[X \,|\, \mathcal{G}]$.

Let us confirm that Y has the properties required by Definition 20.1. First, Y is $\mathcal{G}$-measurable, because it is the difference of two $\mathcal{G}$-measurable random variables. Second, Y is integrable because $\nu_\pm(\Omega) = \mathbb{E}[X_\pm] < +\infty$. Last, Y is consistent with X in the sense that

$$\mathbb{E}[Y \mathbb{1}_{\mathsf{G}}] = \mathbb{E}[X \mathbb{1}_{\mathsf{G}}] \quad \text{for all } \mathsf{G} \in \mathcal{G}.$$

This point follows when we subtract the relations (22.3) and invoke the definitions (22.2) of $\nu_\pm$. Thus, Y is a conditional expectation of X, given $\mathcal{G}$. ∎

The proof of Theorem 22.9 indicates that the conditional expectation can be viewed as the density of the (signed) measure $\nu(\mathsf{G}) := \mathbb{E}[X \mathbb{1}_{\mathsf{G}}]$ with respect to the probability measure $\mathsf{G} \mapsto \mathbb{P}(\mathsf{G})$ on the measurable space $(\Omega, \mathcal{G})$. That is,

$$Y = \frac{\mathrm{d}\nu}{\mathrm{d}\mathbb{P}} \quad \text{on } (\Omega, \mathcal{G}).$$

This corresponds to our intuition that we are comparing the mass that X places on G with the probability mass of G for events $\mathsf{G} \in \mathcal{G}$.

A subtle feature of this argument is that we apply the Radon–Nikodým theorem in the measurable space $(\Omega, \mathcal{G})$ instead of the original measurable space $(\Omega, \mathcal{F})$. This slight change of perspective forces the conditional expectation Y to be $\mathcal{G}$-measurable, which means that Y is a coarse-grained approximation of X, quantized to events in $\mathcal{G}$.

22.3 The Lebesgue decomposition theorem

We will derive the Radon–Nikodým theorem as a consequence of a more general decomposition theorem for measures, due to Lebesgue.

Theorem 22.10 (Lebesgue decomposition). Let μ, ν be σ-finite measures on the same measurable space. Then there is a measurable set N with $\mu(\mathsf{N}) = 0$ and a positive, measurable function f for which

$$\nu(\mathsf{E}) = \nu(\mathsf{E} \cap \mathsf{N}) + \int_{\mathsf{E}} f \, d\mu \quad \text{for all measurable } \mathsf{E}.$$

This decomposition is essentially unique, in that the set N and the function f are determined μ-almost everywhere.

A discussion and proof of Theorem 22.10 occupy the rest of this section. Observe that the Radon–Nikodým theorem is an immediate consequence of Theorem 22.10.

Corollary 22.11 (Radon–Nikodým). If $\nu \ll \mu$, then ν has a density f with respect to μ.

Proof. Since $\mu(\mathsf{N}) = 0$, we have $\nu(\mathsf{N}) = 0$. Therefore, in Theorem 22.10, the term $\nu(\mathsf{E} \cap \mathsf{N}) = 0$ for every measurable set E. ∎

22.3.1 Lebesgue decomposition

Let us discuss what Theorem 22.10 means. We begin with a definition.

Definition 22.12 (Mutually singular). Two measures μ, ν on the same measurable space are called *mutually singular* if there is a measurable set N for which $\mu(\mathsf{N}) = 0$ and $\nu(\mathsf{N}^c) = 0$. We write $\mu \perp \nu$.

In other words, two measures are mutually singular when each one is concentrated on a negligible set of the other. In this sense, the two measures have nothing to do with each other.

Theorem 22.10 treats μ as reference measure. This reference measure allows us to split the measure ν into two parts:

$$\nu = \nu_\perp + \nu_{\ll} \quad \text{where} \quad \nu_\perp \perp \nu_{\ll}. \tag{22.4}$$

The measure $\nu_\perp$ and the reference measure μ are mutually singular, while the measure $\nu_{\ll}$ is absolutely continuous with respect to the reference μ.

Exercise 22.13 (Mutually singular). Show that the set N that witnesses that $\mu \perp \nu$ is determined $(\mu + \nu)$-almost everywhere. More precisely, suppose there is another measurable set M for which $\mu(\mathsf{M}) = 0$ and $\nu(\mathsf{M}^c) = 0$. Then $\mu(\mathsf{N} \triangle \mathsf{M}) = \nu(\mathsf{N} \triangle \mathsf{M}) = 0$.

Exercise 22.14 (Mutually singular: Discrete and continuous). Suppose that X is a discrete random variable, and Y is a continuous random variable. Show that the laws μ_X and μ_Y are mutually singular.

Exercise 22.15 (Lebesgue decomposition). Show that the measures $\nu_\perp$ and $\nu_{\ll}$ in the Lebesgue decomposition (22.4) are completely determined by μ, ν.

Problem 22.16 (*Singular continuous). Recall that the Cantor distribution has a continuous distribution function, but it does not have a density with respect to the Lebesgue measure. Let X be a random variable that is either discrete or continuous. Show that the law μ_X and the Cantor distribution are mutually singular.

Prove the following extension of the Lebesgue decomposition. Every Borel measure on the real line is the sum of three mutually singular distributions: a purely discrete distribution (concentrated on a countable set), a distribution that has a density with respect to Lebesgue measure, and a singular continuous distribution (having a continuous distribution function but no density with respect to Lebesgue measure).

22.3.2 Background

We will prove Theorem 22.10 using a functional-analytic argument due to John von Neumann. This approach involves a small dose of theory for inner-product spaces.

Consider a σ-finite measure ρ on a measurable space $(\Omega, \mathcal{F})$. As usual, define the L_2 pseudonorm:

$$\|g\|_2 := \left(\int_\Omega |g|^2 \, d\rho \right)^{1/2} \quad \text{for measurable } g : \Omega \to \mathbb{R}.$$

We introduce the space of square-integrable functions:

$$\mathsf{L}_2(\rho) := \{g : \|g\|_2 < +\infty\}.$$

As in Lecture 11 or 12, this is a complete pseudonormed space. Let us single out an important class of functions, defined on $\mathsf{L}_2(\rho)$.

Definition 22.17 (Bounded linear functional). A *linear functional* is a linear map $\varphi : \mathsf{L}_2(\rho) \to \mathbb{R}$. That is,

$$\varphi(\alpha f + \beta g) = \alpha \varphi(f) + \beta \varphi(g) \quad \text{for all } \alpha, \beta \in \mathbb{R} \text{ and } f, g \in \mathsf{L}_2(\rho).$$

The linear functional φ is *bounded* if there exists a constant $C > 0$ such that

$$|\varphi(g)| \le C \cdot \|g\|_2 \quad \text{for all } g \in \mathsf{L}_2(\rho).$$

Every bounded linear functional on $\mathsf{L}_2(\rho)$ can be expressed as an integral. This famous result is called the Riesz representation theorem.

Theorem 22.18 (Riesz representation). Let ρ be a σ-finite measure on a measurable space. Consider the space $\mathsf{L}_2(\rho)$ of square-integrable functions. If φ is a bounded linear functional on $\mathsf{L}_2(\rho)$, then there is a function $h \in \mathsf{L}_2(\rho)$ for which

$$\varphi(g) = \int_\Omega g h \, d\rho \quad \text{for all } g \in \mathsf{L}_2(\rho).$$

In the special case of a probability measure ρ, Problem 12.29 contains an outline of the proof. The argument is no different in the general setting here. The key idea is to show that every function in $\mathsf{L}_2(\rho)$ induces another function that is orthogonal to the null space of φ. This orthogonality relation yields the Riesz representation.

22.3.3 Lebesgue decomposition: Proof

We establish the theorem under the additional assumption that both μ and ν are *finite* measures. We will work in the space $\mathsf{L}_2(\mu + \nu)$. Recall that

$$(\mu + \nu)(\mathsf{E}) = \mu(\mathsf{E}) + \nu(\mathsf{E}) \quad \text{for measurable } \mathsf{E}.$$

Similarly,

$$\int_\Omega g \, d(\mu + \nu) = \int_\Omega g \, d\mu + \int_\Omega g \, d\nu \quad \text{for } g \in \mathsf{L}_2(\mu + \nu).$$

The sum of two measures behaves exactly as you imagine it should.

A bounded linear functional

On $\mathsf{L}_2(\mu + \nu)$, we consider the bounded linear functional

$$\varphi(g) := \int_\Omega g \, \mathrm{d}\nu \quad \text{for all } g \in \mathsf{L}_2(\mu + \nu).$$

Linearity follows from the linearity of the integral. To see that φ is bounded, we calculate that

$$\begin{aligned} |\varphi(g)| \le \int_\Omega |g| \, \mathrm{d}\nu &\le \int_\Omega |g| \, \mathrm{d}(\mu + \nu) \\ &\le ((\mu + \nu)(\Omega))^{1/2} \left(\int_\Omega |g|^2 \, \mathrm{d}(\mu + \nu) \right)^{1/2} =: C \cdot \|g\|_2. \end{aligned}$$

The first inequality is the absolute value inequality for integrals. The second inequality holds because μ is a positive measure, so $\int_\Omega |g| \, \mathrm{d}\mu \ge 0$. The third inequality is Cauchy–Schwarz. Finally, we invoke the fact that μ, ν are finite measures and the definition of the L_2 pseudonorm.

Riesz representation

Since φ is a bounded linear functional on $\mathsf{L}_2(\mu + \nu)$, Theorem 22.18 furnishes a representation as an integral. There is a fixed function $h \in \mathsf{L}_2(\mu + \nu)$ with the property that

$$\varphi(g) = \int_\Omega g \, \mathrm{d}\nu = \int_\Omega g h \, \mathrm{d}(\mu + \nu) \quad \text{for all } g \in \mathsf{L}_2(\mu + \nu). \tag{22.5}$$

The relation (22.5) is the key to the proof. We can extract everything we need by inserting particular functions into the identity and making deductions.

The representing function is bounded: $0 \le h \le 1$

First, let us establish that $0 \le h \le 1$ ν-almost everywhere. Heuristically, $h \le 1/2$ in places where μ is bigger than ν, while $h \ge 1/2$ in places where ν is bigger than μ.

Let us start with the upper bound. We will present this argument in some detail because variants arise several times in the proof. For a fixed $n \in \mathbb{N}$, we consider the bounded function $g(\omega) = \mathbb{1}\{\omega \in \Omega : h(\omega) > 1 + n^{-1}\}$. Since bounded functions are square-integrable, the relation (22.5) shows that

$$\int_\Omega \mathbb{1}\{h > 1 + n^{-1}\} \, \mathrm{d}\nu = \int_\Omega \mathbb{1}\{h > 1 + n^{-1}\} \cdot h \, \mathrm{d}(\mu + \nu).$$

On the left-hand side, we recognize the measure of a set, and we bound the right-hand side below:

$$\begin{aligned} \nu\{h > 1 + n^{-1}\} &\ge (1 + n^{-1}) \cdot (\mu + \nu)\{h > 1 + n^{-1}\} \\ &\ge (1 + n^{-1}) \cdot \nu\{h > 1 + n^{-1}\}. \end{aligned}$$

The first inequality holds by monotonicity of the integral since $h > 1 + n^{-1}$ on the designated event. The last inequality depends on the fact that μ is a positive measure. This relation implies that

$$\nu\{h > 1 + n^{-1}\} = 0 \quad \text{for each } n \in \mathbb{N}.$$

Using the decreasing limit property of a finite measure (Proposition 2.30), we conclude that $\nu\{h > 1\} = 0$.

The proof of the lower bound is essentially the same. This time, we consider functions of the form $g = \mathbb{1}\{h < -1/n\}$ for $n \in \mathbb{N}$

Constructing the singular set

We introduce the measurable set $\mathsf{N} := \{\omega \in \Omega : h(\omega) = 1\}$. This is the singular set promised in the theorem.

To verify that $\mu(\mathsf{N}) = 0$, we simply insert the function $g = \mathbb{1}\{h = 1\}$ into the relation (22.5).

Constructing the density

We are now prepared to construct the density that certifies that the remaining part of the measure ν is absolutely continuous with respect to μ.

Let us rearrange the relation (22.5) to obtain

$$\int_\Omega g(1-h)\,\mathrm{d}\nu = \int_\Omega gh\,\mathrm{d}\mu. \quad \text{for all } g \in \mathsf{L}_2(\mu+\nu).$$

Fix a measurable set E. By the change of variables $g \mapsto g\mathbb{1}_\mathsf{E}$, we can restrict the integrals to the set E. That is,

$$\int_\mathsf{E} g(1-h)\,\mathrm{d}\nu = \int_\mathsf{E} gh\,\mathrm{d}\mu. \quad \text{for all } g \in \mathsf{L}_2(\mu+\nu).$$

To obtain the formula for the density f in the theorem statement, we need to choose a function g that cancels out the factor $(1-h)$ on the left-hand side.

For $n \in \mathbb{N}$, we activate (22.5) with the bounded function

$$g := \frac{1}{1-h} \cdot \mathbb{1}\{h < 1 - n^{-1}\}.$$

Using monotone convergence (Theorem 5.18), we obtain

$$\int_\mathsf{E} \mathbb{1}\{h<1\}\,\mathrm{d}\nu = \int_\mathsf{E} \frac{h}{1-h} \cdot \mathbb{1}\{h<1\}\,\mathrm{d}\mu.$$

Since $\mathsf{N}^\mathsf{c} = \{h < 1\}$, we can reinterpret the last display to reach the relation

$$\nu(\mathsf{E} \cap \mathsf{N}^\mathsf{c}) = \int_\mathsf{E} f\,\mathrm{d}\mu \quad \text{where} \quad f = \frac{h}{1-h}\mathbb{1}_{h<1}.$$

The function f is the density in the part of ν that is absolutely continuous with respect to μ.

Endgame

Finally, we note that

$$\nu(\mathsf{E}) = \nu(\mathsf{E} \cap \mathsf{N}) + \nu(\mathsf{E} \cap \mathsf{N}^\mathsf{c}) = \nu(\mathsf{E} \cap \mathsf{N}) + \int_\mathsf{E} f\,\mathrm{d}\nu.$$

This is the statement of Theorem 22.10.

Exercise 22.19 (Lebesgue decomposition: σ-finite case). Extend the Lebesgue decomposition result to the case where μ, ν are σ-finite measures.

Exercise 22.20 (Lebesgue decomposition: Uniqueness). Show that the set N and the density f are essentially unique.

V. martingales

23. Martingales

Agenda:
1. Filtrations
2. Martingales
3. Examples

> "I took all the gold I found, and playing the martingale, and doubling my stakes continuously, I won every day during the remainder of the carnival. I was fortunate enough never to lose the sixth card, and, if I had lost it, I should have been without money to play, for I had two thousand sequins on that card. I congratulated myself upon having increased the treasure of my dear mistress...
>
> "I still played on the martingale, but with such bad luck that I was soon left without a sequin. As I shared my property with M. M. I was obliged to tell her of my losses, and it was at her request that I sold all her diamonds, losing what I got for them; she had now only five hundred sequins by her. There was no more talk of her escaping from the convent, for we had nothing to live on! I still gamed, but for small stakes, waiting for the slow return of good luck."
>
> —*Histoire de ma via*, Giacomo Casanova (1822), transl. Arthur Machen (1902)

Having completed our study of conditional expectation, we turn back to the theory of stochastic processes. We have already attended on the basic facts about the partial sums of a sequence of independent random variables. Although these partial sums are not independent, they have a rather trivial type of dependency structure. Now that we understand conditioning, we can discuss stochastic processes with more interesting dependencies.

In this lecture, we introduce several important classes of stochastic processes: martingales and their relatives. These processes consist of random variables that are indexed by time, and each element can depend in an almost arbitrary way on the past. The random variables are linked to each other by means of assumptions about their conditional expectations. At each time, the conditional expectation of the next random variable in the sequence is related to the current value of the sequence.

Among other things, a martingale models a sequence of repeated games, where the player's strategy may depend on the past. For conceptual simplicity, we will use gambling analogies to motivate the definition of a martingale and related concepts. For computational mathematicians, the relevant applications of martingales include prediction, filtering, adaptive learning and decision making, and the study of randomized algorithms.

23.1 Filtrations and adapted processes

We have the intuition that a σ-algebra captures knowledge about the world. In particular, at a given time, we can collect all the events that have been determined so far. Therefore, a small σ-algebra contains less information than a larger σ-algebra that contains it. In this section, building on this idea, we develop a mathematical formalization of the process of accumulating information about the state of affairs.

23.1.1 Filtrations

A filtration is simply an increasing sequence of σ-algebras. Each algebra collects the events that have been determined at a given time.

Definition 23.1 (Filtration). Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$. A (discrete-time) *filtration* is an increasing sequence of σ-algebras on Ω:

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}_\infty \subseteq \mathcal{F}.$$

We may abbreviate the filtration as a sequence $(\mathcal{F}_k : k \in \overline{\mathbb{Z}}_+)$.

You should think about $k = 0, 1, 2, \ldots$ as a discrete time index. The σ-algebra $\mathcal{F}_k$ contains all events that have been determined up to and including time instant k. That is, for each event $\mathsf{E} \in \mathcal{F}_k$, at the time k, we know whether the distinguished sample point $\omega_0 \in \mathsf{E}$ or $\omega_0 \notin \mathsf{E}$.

The fact that the σ-algebras are increasing reflects the assumption that we accumulate knowledge with time, and we never forget what we know. With age comes wisdom.

It is common (but not necessary) for the initial element of the filtration to be the trivial σ-algebra: $\mathcal{F}_0 := \{\emptyset, \Omega\}$. In this case, no events are determined at the outset; we are born ignorant.

If the last element $\mathcal{F}_\infty$ of the filtration is not specified, we define $\mathcal{F}_\infty := \sigma(\bigcup_{k=1}^\infty \mathcal{F}_k)$. This is the collection of all events that are discoverable during our exploration. We do not assume that $\mathcal{F}_\infty$ coincides with the master σ-algebra $\mathcal{F}$, so there may be many events that remain outside the scope of our experience.

Note that $\bigcup_{k=1}^\infty \mathcal{F}_k$ need not be a σ-algebra; we may need to generate additional events.

It is also very common to consider a finite filtration, which takes the form

$$\mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \cdots \subseteq \mathcal{F}_N \subseteq \mathcal{F} \quad \text{for a fixed } N \in \mathbb{Z}_+.$$

These filtrations can be used to model random processes that have a fixed horizon.

23.1.2 Filtrations generated by events and by random variables

We can generate a filtration from a sequence of events.

Example 23.2 (Filtration: Events). Consider an arbitrary sequence $(\mathsf{E}_i : i \in \mathbb{N})$ of events that belong to the master σ-algebra $\mathcal{F}$. Define the σ-algebras

$$\mathcal{F}_k := \sigma(\mathsf{E}_1, \ldots, \mathsf{E}_k) \quad \text{for } k \in \mathbb{N}.$$

Let $\mathcal{F}_0 := \{\emptyset, \Omega\}$ and $\mathcal{F}_\infty := \sigma(\bigcup_{k=1}^\infty \mathcal{F}_k)$.

Then $(\mathcal{F}_k : k \in \overline{\mathbb{Z}}_+)$ is a filtration. At each time k, we know whether the events $\mathsf{E}_1, \ldots, \mathsf{E}_k$ and all of their set-theoretic combinations have occurred.

This filtration models a sequence of simple experiments (not necessarily independent) where E_i is the event that the ith experiment succeeds. At time k, we know the outcomes of the first k experiments. ▪

Most commonly, filtrations arise from a sequence of random variables.

Example 23.3 (Filtration: Random variables). Consider an arbitrary sequence $(Z_i : i \in \mathbb{Z}_+)$ of real random variables on Ω that are $\mathcal{F}$-measurable. Define the σ-algebras

$$\mathcal{F}_k := \sigma(Z_0, Z_1, Z_2, \ldots, Z_k) \quad \text{for } k \in \mathbb{Z}_+.$$

Let $\mathcal{F}_\infty := \sigma(\bigcup_{k=1}^\infty \mathcal{F}_k)$.

As before, $(\mathcal{F}_k : k \in \overline{\mathbb{Z}}_+)$ is a filtration. At each time k, we know the values of the random variables $Z_0, Z_1, \dots, Z_k$. Recall that a random variable Y is $\mathcal{F}_k$-measurable if and only if $Y = g(Z_0, Z_1, \dots, Z_k)$ for a measurable function $g : \mathbb{R}^k \to \mathbb{R}$.

In many cases, the initial random variable Z_0 takes a constant value (so is not random). This assumption is equivalent with $\mathcal{F}_0 = \{\emptyset, \Omega\}$.

This filtration models a sequence $Z_0, Z_1, Z_2, \dots$ of numerical observations (not necessarily independent), although there is no need for the measurements to be alike or independent from each other. ▪

The last example is fairly general, and it describes a large class of situations. For example, we can think about the random variables Z_k as the random outcome of the kth game in a sequence. If we are repeatedly rolling a fair die, then Z_k are i.i.d. copies of a UNIFORM$\{1, \dots, 6\}$ random variable. At time k, the outcomes of the first k rolls are completely described by $\mathcal{F}_k$.

23.1.3 Adapted processes

We can now introduce a new kind of stochastic process, consisting of a sequence of random variables whose values are determined as time passes.

Definition 23.4 (Adapted process). Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a filtration $(\mathcal{F}_k : k \in \overline{\mathbb{Z}}_+)$. We say that a sequence $(X_k : k \in \mathbb{Z}_+)$ of real random variables is *adapted* to the filtration if X_k is $\mathcal{F}_k$-measurable for each index $k \in \mathbb{Z}_+$.

In other words, the kth random variable X_k in the adapted sequence is completely determined by the information we have available at time k.

If you think about the filtration $(\mathcal{F}_k)$ as describing the random outcomes (Z_k) of a sequence of games, you might think about (X_k) as describing the sequence of accumulated winnings after k games. The total winnings X_k after k games depends on the player's winnings X_{k-1} from the first $k-1$ games, the player's bet on the kth game (which may depend on the entire history of the game), and the random outcome Z_k of the kth game. Although these effects may interact in a complicated way, the sequence (X_k) is still adapted.

23.2 Martingales and friends

An adapted process provides a very flexible model for describing random outcomes that evolve with time in a causal fashion (i.e., the future does not influence the past). Nevertheless, at this level of generality, it is hard to say very much about the behavior of an adapted process because there is no relationship among its constituents. If we want to be able to understand the trajectory of the process, we need a way to link the random variables.

23.2.1 Martingales

Martingales (and their relatives) are adapted processes that are based on a minimal assumption about how the random variables evolve. Although the value X_k of an adapted process is determined at time k, the future values of the process remain random. A martingale is a random process that is indifferent about the future. Given where things stand now, our best prediction is that the future will be the same.

Definition 23.5 (Martingale sequence). Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a filtration

$(\mathscr{F}_k : k \in \overline{\mathbb{Z}}_+)$. A sequence $(X_k : k \in \mathbb{Z}_+)$ of real random variables is called a (discrete-time) *martingale* with respect to the filtration when it satisfies three properties:

1. **Adaptivity:** (X_k) is adapted to the filtration for each $k \in \mathbb{Z}_+$.
2. **Integrability:** $\mathbb{E}\,|X_k| < +\infty$ for each $k \in \mathbb{Z}_+$.
3. **Status quo:** $\mathbb{E}[X_{k+1} \,|\, \mathscr{F}_k] = X_k$ almost surely for each $k \in \mathbb{Z}_+$.

The integrability property does *not* require that the sequence $\|X_k\|_{\mathsf{L}_1}$ is uniformly bounded.

The adaptivity property means that the value X_k of the martingale sequence at time k is determined by the information $\mathscr{F}_k$ that we have available at time k. The distinctive assumption is the status quo property, which states that the next value X_{k+1} of the sequence is, on average, the same as the current value. In other words, we expect tomorrow to be the same as today. The integrability requirement is needed so that we can compute the conditional expectation.

Example 23.6 (Gambling). In the kth game, I flip a fair coin, and I pay you \$1 if the coin comes up heads, whereas you pay me \$1 if the coin comes up tails. Let X_k denote your total winnings after k games are complete.

We can model your total winnings as a martingale sequence. For each $k = 1, 2, 3, \dots$ define the random variable $Z_k = +1$ if the kth coin comes up heads and $Z_k = -1$ if the kth coin comes up tails. Introduce the σ-algebras $\mathscr{F}_0 = \{\emptyset, \Omega\}$ and $\mathscr{F}_k = \sigma(Z_1, \dots, Z_k)$. The initial winnings $X_0 = 0$, so it is $\mathscr{F}_0$-measurable. Your total winnings after the kth game satisfy

$$X_k = X_{k-1} + Z_k \quad \text{for } k \in \mathbb{N}.$$

By induction, we can see that X_k is $\mathscr{F}_k$-measurable for each k. Therefore, (X_k) is adapted to the filtration. Second, induction shows that X_k is integrable because L_1 is a linear space. Third, we compute the conditional expectation:

$$\mathbb{E}[X_{k+1} \,|\, \mathscr{F}_k] = \mathbb{E}[X_k + Z_{k+1} \,|\, \mathscr{F}_k] = X_k + \mathbb{E}[Z_{k+1} \,|\, \mathscr{F}_k] = X_k.$$

Therefore, the accumulated winnings from a fair game form a martingale sequence. ▪

The last example is probably the source of the terminology "martingale". Indeed, there is a famous betting strategy called a martingale where you double your bet each time you play a fair game. Paradoxically, it may seem that this strategy is guaranteed to yield a profit. If your first win occurs on the kth trial, then your total winnings after the k trial satisfy

This is also called the St. Petersburg game.

$$-1 - 2 - 4 - \cdots - 2^{k-1} + 2^k = 1.$$

Unfortunately, this strategy potentially requires an infinite amount of capital and infinite gameplay, so it is not entirely practical.

Exercise 23.7 (Martingale: Future expectation). Assuming that $n \geq k$, show that

$$\mathbb{E}[X_n \,|\, \mathscr{F}_k] = X_k \quad \text{almost surely.}$$

In particular, $\mathbb{E}[X_n] = \mathbb{E}[X_0]$ for each $n \in \mathbb{Z}_+$.

23.2.2 Supermartingales and submartingales

As we have mentioned, martingales can be used to model fair games. It is natural to consider two related processes that describe unfair games.

Definition 23.8 (Supermartingale). Maintain the notation of Definition 23.5. We say that the sequence $(X_k : k \in \mathbb{Z}_+)$ is a *supermartingale* if the status quo property (3) is replaced by the condition

3↓. **Decreasing expectations:** $\mathbb{E}[X_{k+1} \,|\, \mathcal{F}_k] \le X_k$ almost surely for each $k \in \mathbb{Z}_+$.

The decreasing expectation property in Definition 23.8 reflects a pessimistic view of the world. On average, tomorrow is worse than today. A supermartingale can be used to model a player's total winnings in a game that is unfair to him (e.g., casino games from the player's point of view).

" 'That Accounts for a Good Deal,' said Eeyore. 'How Like Them,' he added, after a long silence."
—*Winnie-The-Pooh*, A. A. Milne

Definition 23.9 (Submartingale). Maintain the notation of Definition 23.5. We say that the sequence $(X_k : k \in \mathbb{Z}_+)$ is a *submartingale* if the status quo property (3) is replaced by the condition

3↑. **Increasing expectations:** $\mathbb{E}[X_{k+1} \,|\, \mathcal{F}_k] \ge X_k$ almost surely for each $k \in \mathbb{Z}_+$.

The increasing expectation property in Definition 23.9 reflects an optimistic view of the world. On average, tomorrow is better than today. A submartingale can be used to model a player's total winnings in a game that is unfair to her opponent (e.g., casino games from the casino's point of view).

"Pangloss enseignait la métaphysico-théologo-cosmolonigologie. Il prouvait admirablement qu'il n'y a point d'effet sans cause, et que, dans ce meilleur des mondes possibles, le château de monseigneur le baron était le plus beau des châteaux, et madame la meilleure des baronnes possibles."
—*Candide*, Voltaire

Exercise 23.10 (Supermartingale and submartingale). Show that $(X_k : k \in \mathbb{Z}_+)$ is a supermartingale if and only if $(-X_k : k \in \mathbb{Z}_+)$ is a submartingale.

Exercise 23.11 (Supermartingale and submartingale: Future expectations). What is the correct extension of Exercise 23.7 to a supermartingale? For a submartingale?

Warning 23.12 (Super?). The terms "supermartingale" and "submartingale" are the opposite of what you might anticipate. Supermartingales have decreasing expectations, while submartingales have increasing expectations. Be careful!

In fact, there is good reason for this terminology, connected to potential theory. A superharmonic function on a Markov chain induces a supermartingale, while a subharmonic function induces a submartingale. ∎

23.2.3 *Positive martingales

For a sequence of positive random variables, we can define a martingale sequence without the integrability assumption.

Definition 23.13 (Positive martingale sequence). Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a filtration $(\mathcal{F}_k : k \in \overline{\mathbb{Z}}_+)$. Consider a sequence $(X_k : k \in \mathbb{Z}_+)$ of *positive* random variables taking extended real values. The sequence is called a *positive martingale* with respect to the filtration when it satisfies two properties:

1. **Adaptivity:** (X_k) is adapted to the filtration for each $k \in \mathbb{Z}_+$.
2. **Status quo:** $\mathbb{E}[X_{k+1} \,|\, \mathcal{F}_k] = X_k$ almost surely for each $k \in \mathbb{Z}_+$.

We admit the possibility that the random variables and the conditional expectations take the value $+\infty$.

We can define positive supermartingales and positive submartingales in a similar fashion.

23.3 Examples

Martingales arise in a wide range of situations. In this section, we describe some mathematical models that lead to martingales, along with some applications of these models.

23.3.1 Independent sums

First, let us demonstrate that the independent sums we studied before can provide an example of a martingale sequence. In particular, we may gain new insights on independent sums using martingale methods.

Consider an independent sequence $(Z_i : i \in \mathbb{N})$ of *centered* random variables; that is, $\mathbb{E}\, Z_i = 0$ for each index i. In particular, the random variables Z_i are integrable. Consider the partial sums

$$X_0 = 0 \quad \text{and} \quad X_k = \sum_{i=1}^{k} Z_i \quad \text{for } k \in \mathbb{N}.$$

We will check that these partial sums compose a martingale.

Construct the filtration $\mathcal{F}_k = \sigma(Z_1, \dots, Z_k)$ for $k \in \mathbb{N}$, where $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Since $X_0 = 0$ is $\mathcal{F}_0$-measurable, a short inductive argument demonstrates that X_k is $\mathcal{F}_k$-measurable. The integrability of X_k follows from the triangle inequality and the integrability of the Z_k. Last, to check the status quo property, we calculate that

$$\mathbb{E}[X_{k+1} \,|\, \mathcal{F}_k] = \mathbb{E}[X_k + Z_{k+1} \,|\, \mathcal{F}_k] = X_k + \mathbb{E}[Z_{k+1} \,|\, \mathcal{F}_k] = X_k \quad \text{almost surely.}$$

We have used linearity of conditional expectation, the fact that X_k is $\mathcal{F}_k$-measurable, and the fact that Z_{k+1} is independent from $\mathcal{F}_k$.

As we discussed in Lecture 14, independent sums arise in many contexts. They model random walks and renewal processes (e.g., queues). They describe the total number of successes in a sequence of independent experiments and the sample average of a sequence of i.i.d. measurements. They also describe the behavior of Monte Carlo integration procedures. In some of these instances, we may need to center the independent sum to treat it using martingale methods.

Exercise 23.14 (Independent sum: Uncentered case). Suppose instead that the independent summands Z_i are uncentered, but $\mathbb{E}\, Z_i \geq 0$ for each index i. Show that the partial sums (X_k) compose a submartingale.

23.3.2 Independent products

Next, we consider products of independent random variables. This example is closely related to independent sums, but it falls outside the scope of our previous analyses.

Consider an independent sequence $(Z_i : i \in \mathbb{N})$ of positive random variables with expectation $\mathbb{E}\, Z_i = 1$ for each index i. In particular, the random variables Z_i are integrable. Consider the partial products

$$X_0 = 1 \quad \text{and} \quad X_k = \prod_{i=1}^{k} Z_i \quad \text{for } k \in \mathbb{N}.$$

We claim that this sequence (X_k) is a martingale.

As before, construct the filtration $\mathcal{F}_k = \sigma(Z_1, \dots, Z_k)$ for $k \in \mathbb{N}$, where $\mathcal{F}_0 = \{\emptyset, \Omega\}$. Since $X_0 = 1$ is $\mathcal{F}_0$-measurable, induction confirms that X_k is $\mathcal{F}_k$-measurable. The integrability of X_k follows from the independence and the integrability of the Z_k. Last, to check the status quo property, we calculate that

$$\mathbb{E}[X_{k+1} \,|\, \mathcal{F}_k] = \mathbb{E}[X_k \cdot Z_{k+1} \,|\, \mathcal{F}_k] = X_k \cdot \mathbb{E}[Z_{k+1} \,|\, \mathcal{F}_k] = X_k \quad \text{almost surely.}$$

We have used the pull-through property of conditional expectation, the fact that X_k is $\mathcal{F}_k$-measurable, and the fact that Z_{k+1} is independent from $\mathcal{F}_k$.

By assuming that $Z_i \geq 0$, so that $X_k \geq 0$, we can justify the pull-through property without integrability assumptions.

It may not be as clear to you how independent products appear in practice. First, they arise in the study of branching processes, which describe the growth or decay of a population (Exercise 23.18). Branching processes model things like actual populations (in ecology or bacteriology), as well as more abstract things like family names (in genealogy) or chain reactions (in nuclear physics). Independent products also have statistical applications, including the analysis of likelihood ratio tests (Application 25.32). Furthermore, independent products can describe the trajectory of a (stochastic) dynamical system.

It is also easy to see how independent products can arise from the mgf of an independent sum. Indeed, if the family (Y_i) is independent, then

$$\mathrm{e}^{\theta(Y_1+\cdots+Y_n)} = \mathrm{e}^{\theta Y_1} \cdots \mathrm{e}^{\theta Y_n}.$$

The right-hand side is an independent product. If $\mathbb{E}\, Y_i = 0$, then we can only be sure that $\mathbb{E}\, \mathrm{e}^{\theta Y_i} \geq 1$, so the product is a submartingale (rather than a martingale). We will return to this example in Lecture 26.

Exercise 23.15 (Independent product: Out of equilibrium). Suppose instead that the independent factors Z_i satisfy $\mathbb{E}\, Z_i \geq 1$ for each index i. Confirm that the partial products $X_k = Z_k \cdots Z_1 Z_0$ compose a submartingale.

23.3.3 The Lévy–Doob martingale

There is another fundamental construction of a martingale, variously attributed to Lévy or to Doob. Suppose that $(\mathcal{F}_k : k \in \overline{\mathbb{Z}}_+)$ is an arbitrary filtration of the probability space. Let Z be any integrable random variable. Then we may define

$$X_k = \mathbb{E}[Z \,|\, \mathcal{F}_k] \quad \text{for } k \in \overline{\mathbb{Z}}_+.$$

More precisely, X_k is a version of the conditional expectation $\mathbb{E}[Z \,|\, \mathcal{F}_k]$. In other words, X_k is our best prediction of Z, given the information that we have acquired at time k. This sequence (X_k) of random variables also composes a martingale.

By definition of the conditional expectation, X_k is integrable, and it is $\mathcal{F}_k$-measurable (hence adapted). The status quo property follows instantly from the tower law:

$$\mathbb{E}[X_{k+1} \,|\, \mathcal{F}_k] = \mathbb{E}[\mathbb{E}[Z \,|\, \mathcal{F}_{k+1}] \,|\, \mathcal{F}_k] = \mathbb{E}[Z \,|\, \mathcal{F}_k] = X_k.$$

Indeed, $\mathcal{F}_k \subseteq \mathcal{F}_{k+1}$, so the tower law is valid.

Lévy–Doob martingales arise whenever we make a sequence of predictions based on accumulated information. For example, they appear in the problem of predicting the value of a random variable from a sequence of noisy observations. They arise in filtering problems, where our goal is to track the evolution of a random sequence given ongoing observations. They also come up in combinatorics and in the analysis of randomized algorithms.

23.3.4 Further applications

More generally, martingale arise in any setting where we have a sequence of random variables that evolves in time and where we our predictions of the future have a clear relationship with the present value. Some problems where martingales are central include...

- Gambling and game theory;

- Pricing financial assets;
- Diffusion processes;
- Statistical estimation;
- Adaptive learning and decision making;
- Prediction, filtering, and control;
- Stochastic algorithms.

As a consequence, understanding the behavior of martingale processes can pay significant dividends for the modern computational mathematician.

Problems

Exercise 23.16 (The drifters). There are several ways to construct martingales related to random walks on the real line. Consider an i.i.d. family $(Z_k : k \in \mathbb{N})$ of copies of a real random variable Z. For an initial point $a \in \mathbb{R}$, we can construct a random walk with increment distribution Y via

$$S_0 := a \quad \text{and} \quad S_{k+1} := S_k + Z_{k+1} \quad \text{for } k \in \mathbb{Z}_+.$$

Observe that these random variables S_k do not generally compose a martingale sequence. Nevertheless, we can use martingale methods to analyze the random walk by passing to another sequence.

1. For a simple random walk on the integers, the increment distribution Z takes the form
$$\mathbb{P}\{Z = +1\} = p \quad \text{and} \quad \mathbb{P}\{Z = -1\} = 1 - p =: q,$$
where $p \in (0, 1)$. Confirm that the sequence $M_k := (q/p)^{S_k}$ for $k \in \mathbb{Z}_+$ composes a martingale, called *De Moivre's martingale*.
2. If $Z \in \mathsf{L}_1$ with $m = \mathbb{E}[Z]$, show that the random variables $X_k := S_k - mk$ for $k \in \mathbb{Z}_+$ compose a martingale sequence.
3. If $Z \in \mathsf{L}_2$ with mean $m = \mathbb{E}[Z]$ and variance $v = \mathrm{Var}[Z]$, check that the random variables $X_k := (S_k - mk)^2 - vk$ for $k \in \mathbb{Z}_+$ also compose a martingale.
4. Assume that $Z \in \mathsf{L}_\infty$ with mgf m_Z. For each fixed real number $\theta \in \mathbb{R}$, confirm that the following sequence composes a martingale:
$$M_k(\theta) := \frac{\mathrm{e}^{\theta S_k}}{m_Z(\theta)^k} \quad \text{for } k \in \mathbb{Z}_+.$$

Exercise 23.17 (Le rouge et le noir). Urn models can be used to describe the evolution of discrete populations, including the spread of epidemics. Here is the most elementary example. Initially, a candy bowl contains one red and one black M&M. At each time instant $k = 1, 2, 3, \ldots$, a uniformly random M&M is extracted from the bowl, and we return this M&M to the bowl along with a new M&M of the same color. Let R_k and B_k denote the number of red and black M&Ms in the bowl after k steps. Check that the random variables $M_k := R_k/(R_k + B_k)$ compose a martingale. (Note: Black M&Ms are available at Halloween, so this problem statement is not vacuous.)

Exercise 23.18 (Branch water). Branching processes are often used to model the growth or decay of a population (e.g., family names, bunnies, bacteria, free neutrons). A branching process is a random sequence $(Z_k : k \geq 0)$ constructed in the following manner.

i. The initial population Z_0 is a positive, integer-valued random variable with finite mean.
ii. For each time k and each $i = 1, 2, 3, \ldots$, the family sizes $Y_k^{(i)}$ are positive, integer-valued, i.i.d. random variables
iii. The size Z_k of the population evolves as

$$Z_{k+1} = \sum_{i=1}^{Z_k} Y_k^{(i)}.$$

Check that the random variables $X_k = Z_k / s^k$ for $k \in \mathbb{Z}_+$ compose a martingale.

Notes

I learned of the delicious Casanova quotation about martingales from Grimett & Stirzaker [GS01], and many of the problems and exercise are adapted from their book. Our overall presentation of martingales is based on Williams's book [Wil91].

Lecture bibliography

[GS01] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. 3rd ed. Oxford University Press, 2001.

[Wil91] D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

24. Stopping Times

"Stop! In the name of love
Before you break my heart.
Think it over."

—*Stop! In the name of love*, The Supremes, 1965

Agenda:

1. Martingale transforms
2. Stopping times
3. Stopped processes
4. Optional stopping
5. Monotone stopping

In Lecture 23, we introduced the concept of a martingale process, a sequence of random variables with the status quo property: on average, tomorrow is the same as today. We also defined supermartingales (whose conditional expectations are decreasing) and submartingales (whose conditional expectations are increasing). We saw that these models arise in a wide range of circumstances, with many possible applications in computational mathematics.

Among other things, martingales model the accumulated winnings from a sequence of fair games. This application suggests some questions. First, we may ask if there is a betting strategy that allows us to profit, on average, from a fair game. Second, we may ask if there is a strategy for stopping play that allows us to profit on average. I am sorry to report that both answers are essentially negative.

To settle these questions, we need to introduce some additional tools for understanding the behavior of martingales. First, we define the martingale transform of a sequence, which converts a sequence of bets and a sequence of game outcomes into a sequence of total winnings. Second, we introduce the important notion of a stopping time, which is a random time at which we may elect to quit playing a game. Although these methods are motivated by gambling analogies, they are also quite valuable for proving mathematical facts about martingale sequences.

24.1 The martingale transform

As discussed, martingales are closely connected to games of chance. In this section, we develop a mathematical formalization for betting strategies. Then we define the martingale transform, which describes the total winnings from a sequence of games. We prove the important fact that the sequence of accumulated winnings from a fair game compose a martingale sequence. This settles the matter of whether we can outfox a fair game by a sequence of clever bets.

24.1.1 Previsible processes

If we are playing a game of chance, we must place a bet before the random outcome of the game is revealed to us. At the same time, our betting strategy may depend in a complicated way on the trajectory of the game up to the present time. A previsible process encapsulates these requirements.

Definition 24.1 (Previsible process). Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a filtration $(\mathcal{F}_k : k \in \overline{\mathbb{Z}}_+)$. A sequence $(C_1, C_2, C_3, \dots)$ of real random variables is called *previsible* with respect to the filtration when C_k is $\mathcal{F}_{k-1}$-measurable for each $k \in \mathbb{N}$. Previsible processes are also called *predictable* processes.

A previsible process starts with index $k = 1$!

You can think about C_k as modeling the value of a bet on the kth game in a sequence. The requirement that C_k is $\mathcal{F}_{k-1}$-measurable means that the bet may depend on everything that has happened in the first $k-1$ games, but it must be determined before the outcome of the kth game is revealed.

24.1.2 The martingale transform

We can model a fair game using a martingale sequence, and we can model bets on these outcomes using a previsible process. The martingale transform then describes the accumulated winnings.

Definition 24.2 (Martingale transform). Fix a probability space and a filtration. Suppose that $(X_k : k \in \mathbb{Z}_+)$ is a martingale and $(C_k : k \in \mathbb{N})$ is a previsible process. Assume that each random variable C_k is almost surely bounded. Then the *martingale transform* of the previsible process by the martingale is the sequence

$$(\boldsymbol{C} \cdot \boldsymbol{X})_k := \sum_{i=1}^{k} C_i \cdot (X_i - X_{i-1}) \quad \text{for each } k \in \mathbb{N},$$

with the understanding that $(\boldsymbol{C} \cdot \boldsymbol{X})_0 := 0$.

The idea here is that you are betting on the *change* in the martingale sequence. Each *increment* $\Delta_i := X_i - X_{i-1}$ of the martingale describes the outcome of the ith game. The value C_i is the bet you place on the ith game, while $C_i \Delta_i$ is the amount you win from the ith game. Then the martingale transform $(\boldsymbol{C} \cdot \boldsymbol{X})_k$ is the total amount that you have won after k games.

Exercise 24.3 (Martingale increments). Let $(X_k : k \in \mathbb{Z}_+)$ be a martingale with respect to a filtration $\mathcal{F}_k$. Check that the increments of the margingale are conditionally zero mean: $\mathbb{E}[X_{k+1} - X_k \,|\, \mathcal{F}_k] = 0$ for each $k \in \mathbb{Z}_+$.

> **Aside:** The martingale transform is a discrete version of a stochastic integral. The theory of stochastic integration is closely connected with continuous stochastic processes. It is a fundamental tool for studying diffusion processes and stochastic PDEs. It is also plays a role in pricing financial assets.

24.1.3 The martingale transform is a martingale

The key fact about the martingale transform is that it always results in a martingale sequence.

Proposition 24.4 (Martingale transform). Fix a probability space and a filtration. Suppose that $(X_k : k \in \mathbb{Z}_+)$ is a martingale. Assume that $(C_k : k \in \mathbb{N})$ is a previsible process that is bounded in the sense that $\|C_k\|_{\mathsf{L}_\infty} < +\infty$ for each $k \in \mathbb{N}$. Then the martingale transform $((\boldsymbol{C} \cdot \boldsymbol{X})_k : k \in \mathbb{Z}_+)$ is a martingale with initial value zero.

The sequence $\|C_k\|_{\mathsf{L}_\infty}$ does *not* need to be uniformly bounded, so the bets can increase without bound provided that each one is a.s. finite.

Proof. First, since $(\boldsymbol{C} \cdot \boldsymbol{X})_0 = 0$, it is $\mathcal{F}_0$-measurable. A short inductive argument, using Definition 24.2, shows that $(\boldsymbol{C} \cdot \boldsymbol{X})_k$ is $\mathcal{F}_k$-measurable. Integrability of $(\boldsymbol{C} \cdot \boldsymbol{X})_k$ follows

from the triangle inequality, the boundedness of $C_1, \dots, C_k$, and the integrability of $X_1, \dots, X_k$.

The status quo property is an easy consequence of the pull-through law for conditional expectation (Proposition 20.13). Indeed,

$$\begin{aligned}\mathbb{E}[(\boldsymbol{C} \cdot \boldsymbol{X})_{k+1} - (\boldsymbol{C} \cdot \boldsymbol{X})_k \,|\, \mathcal{F}_k] &= \mathbb{E}[C_{k+1}(X_{k+1} - X_k) \,|\, \mathcal{F}_k] \\ &= C_{k+1} \cdot \mathbb{E}[X_{k+1} - X_k \,|\, \mathcal{F}_k] = 0.\end{aligned}$$

We are justified in drawing the previsible multiplier C_{k+1} out from the conditional expectation because it is bounded and $\mathcal{F}_k$-measurable. The last relation holds because (X_k) is a martingale.

Last, using linearity of the conditional expectation and the fact that $(\boldsymbol{C} \cdot \boldsymbol{X})_k$ is $\mathcal{F}_k$-measurable, we quickly determine that the random variables $(\boldsymbol{C} \cdot \boldsymbol{X})_k$ compose a martingale with initial value zero. ∎

We can interpret Proposition 24.4 in terms of gambling. On average, we can never gain an advantage in a sequence of fair games by a clever betting strategy.

There are many natural variants on Proposition 24.4 that follow from straightforward technical modifications.

Exercise 24.5 (Martingale transform). Suppose that (X_k) is a martingale and (C_k) is a previsible process, and assume that both take values in L_2. Show that the martingale transform $((\boldsymbol{C} \cdot \boldsymbol{X})_k)$ is a martingale.

Exercise 24.6 (Supermartingale transform). Suppose that (X_k) is a *super*martingale. Assume that (C_k) is a *positive*, previsible process that takes values in L_∞. We can define the transform $((\boldsymbol{C} \cdot \boldsymbol{X})_k)$ of the supermartingale, just as in Definition 24.2. Prove that $((\boldsymbol{C} \cdot \boldsymbol{X})_k)$ is a *super*martingale.

Exercise 24.7 (Submartingale transform). Suppose that (X_k) is a *sub*martingale. Assume that (C_k) is a *positive*, previsible process that takes values in L_∞. Prove that the transform $((\boldsymbol{C} \cdot \boldsymbol{X})_k)$ is a *sub*martingale.

Exercise 24.6 means that an unfair sequence of games remains unfair, no matter how we choose to make (positive) bets. Exercise 24.7 means that a favorable sequence of games remains favorable, no matter how we choose to make (positive) bets.

24.2 Stopping times and stopped processes

We have seen that it is not possible to gain an advantage in a fair sequence of games by calibration of our betting strategy. Next, we may ask whether it is possible to gain an advantage by leaving the game at a favorable moment. To answer this question, we need to formalize strategies for stopping play. We will prove that, if we exit a fair game at a random time, the accumulated winnings also compose a martingale sequence. As we will see, this point is somewhat delicate, and it requires careful interpretation.

24.2.1 Stopping times

A strategy for stopping play may involve complicated considerations based on the trajectory of the game, but it cannot involve foreknowledge of the future. The next definition encapsulates this idea.

Definition 24.8 (Stopping time). Fix a probability space and a filtration. A *stopping time* $\tau : \Omega \to \overline{\mathbb{Z}}_+$ is a random variable that may take the value $+\infty$ and with the

property that the event

$$\{\tau \leq k\} \in \mathcal{F}_k \quad \text{for each } k \in \overline{\mathbb{Z}}_+.$$

At each moment k, we can decide whether the stopping time τ has already arrived, given the information we have at hand.

Exercise 24.9 (Stopping time: Equivalence). In the discrete-time setting, we can also define a stopping-time as a (positive, extended value) random variable that satisfies

$$\{\tau = k\} \in \mathcal{F}_k \quad \text{for each } k \in \overline{\mathbb{Z}}_+.$$

Prove that this definition is equivalent with Definition 24.8.

Exercise 24.9 has an intuitive meaning. After playing the kth game, we have the information to decide whether to quit at that moment (before playing the next game).

Exercise 24.10 (Compound stopping times). Let τ_1 and τ_2 be two stopping times with respect to the same filtration. Show that $\min\{\tau_1, \tau_2\}$ and $\max\{\tau_1, \tau_2\}$ are stopping times. Show that $\tau_1 + \tau_2$ is a stopping time.

24.2.2 Examples

We can dramatize the notion of a stopping time by considering the problem of when to liquidate a stock portfolio.

Example 24.11 (Stopping times: Investing). A stopping time describes a rule for deciding when to convert our stock portfolio into cash.

- We can stop investing at the close of market on a date that is fixed in advance, such as the date of your 70th birthday or the Monday before the upcoming presidential election. This is a nonrandom time, so it is not a very interesting choice.
- We can stop investing at the close of market on the first day that the Dow surpasses 40,000 points. This is a random time, but we can certainly ascertain whether it has arrived from available data. It is also possible that this day will never arrive, in which case the stopping time is infinite.
- We can stop investing the first day that the value of our portfolio exceeds \$100,000, the amount required to purchase an aspirational good, such as a very nice fountain pen. As before, this is a random time that we can distinguish when it arrives, although it may never occur.
- Here is an example of an "investing rule" that is not a stopping time. Suppose that we plan to stop investing on the day before the next stock market crash. This is a random time, and it is almost surely finite (I claim!). On the other hand, we cannot evaluate whether the market will crash before it actually does so, because this would require clairvoyance.

Likewise, stopping "in the name of love before you break my heart" is not a stopping time. Think it over.

Indeed, any strategy that you can actually implement will be a stopping time. ▪

Here is an important mathematical example of a stopping time.

Example 24.12 (Hitting times). Let (X_k) be a martingale sequence. Fix a Borel set $\mathsf{B} \in \mathcal{B}(\mathbb{R})$. The *first hitting time* is defined as

$$\tau := \inf\{k \in \mathbb{Z}_+ : X_k \in \mathsf{B}\}.$$

In other words, τ is the first time that the martingale takes a value in the distinguished Borel set. Since X_k is $\mathcal{F}_k$-measurable, it is certainly possible to evaluate whether $X_k \in \mathsf{B}$ given the information we have at time k. It is also possible that the martingale never enters the set B, in which case τ is infinite.

For a related example of a random variable that is *not a stopping time*, consider $T := \sup\{k \in \mathbb{Z}_+ : X_k \in \mathsf{B}\}$, the last time that the martingale takes a value in B. In a general setting, we cannot be sure that the martingale will not return to the set later, so the event $\{T \le k\}$ is certainly not $\mathcal{F}_k$-measurable. ∎

24.2.3 Stopped processes

Given an adapted sequence and a stopping time, we can construct a stopped process whose value is frozen after the stopping time.

Definition 24.13 (Stopped process). Fix a probability space and a filtration. Consider an adapted process $(X_k : k \in \mathbb{Z}_+)$, and let τ be a stopping time. The *stopped process* is the adapted process $(X_{k\wedge\tau} : k \in \mathbb{Z}_+)$.

Recall that $\wedge$ is the infix minimum.

The idea here is that the original process X_k evolves until the random stopping time τ arrives. For future times $k \ge \tau$, the stopped process persists with the same value X_τ.

You can think about the stopped processes as a model for your accumulated winnings from a repeated sequence of games that you may choose to stop playing. The game may go on, but your winnings do not change after you leave the casino.

Exercise 24.14 (Stopped process). Confirm that the stopped process $(X_{k\wedge\tau})$ is indeed an adapted process.

24.2.4 Stopped martingales are martingales

The key fact about a stopped process is that it inherits conditional expectation properties from the original process. For example, a stopped martingale is a martingale.

Theorem 24.15 (Stopped martingale). Fix a probability space and a filtration. Consider a martingale $(X_k : k \in \mathbb{Z}_+)$ and a stopping time τ. Then the stopped process $(X_{k\wedge\tau} : k \in \mathbb{Z}_+)$ is a martingale. In particular, $\mathbb{E}\, X_{k\wedge\tau} = \mathbb{E}\, X_0 = \mathbb{E}\, X_k$ for each $k \in \mathbb{Z}_+$.

Theorem 24.15 means that, on average, at any fixed time k, you cannot gain an advantage in a fair game by any implementable strategy for quitting. But, as we will discuss in the next section, the interpretation of this result does require further thought.

Proof. We will prove this result by showing that the stopped process can be represented using a martingale transform. Construct the previsible process

$$C_k := \begin{cases} 1, & k \le \tau; \\ 0, & k > \tau \end{cases} \qquad \text{for all } k \in \mathbb{N}.$$

To confirm that (C_k) is previsible, note that

$$\{C_k = 0\} = \{\tau \le k - 1\} \in \mathcal{F}_{k-1}.$$

Therefore, the complementary event $\{C_k = 1\} \in \mathcal{F}_{k-1}$ as well. We conclude that C_k is $\mathcal{F}_{k-1}$ measurable for each $k \in \mathbb{N}$.

Now, since (C_k) is bounded, Proposition 24.4 ensures that the martingale transform $((\boldsymbol{C} \cdot \boldsymbol{X})_k)$ is a martingale with initial value zero. Let us reinterpret what this means:

$$(\boldsymbol{C} \cdot \boldsymbol{X})_k = \sum_{i=1}^{k} C_i(X_i - X_{i-1}) = \sum_{i=1}^{k\wedge\tau} (X_i - X_{i-1}) = X_{k\wedge\tau} - X_0.$$

It now follows that $X_{k\wedge\tau}$ is a martingale with initial value X_0. This is what we needed to show. ∎

Variants of Theorem 24.15 hold for supermartingales and submartingales.

Exercise 24.16 (Stopped supermartingales). Suppose that (X_k) is a supermartingale and τ is a stopping time. Show that the stopped process $(X_{k\wedge\tau})$ is a supermartingale. In particular, $\mathbb{E}[X_{k\wedge\tau}] \le \mathbb{E}[X_0]$.

Exercise 24.17 (Stopped submartingales). Suppose that (X_k) is a submartingale and τ is a stopping time. Show that the stopped process $(X_{k\wedge\tau})$ is a submartingale. In particular, $\mathbb{E}[X_{k\wedge\tau}] \ge \mathbb{E}[X_0]$.

24.3 Optional stopping

You can always quit while you are ahead. Suppose that we stop playing a game at the first time that our accumulated winnings exceed \$1. Therefore, at the stopping time, our total winnings must exceed \$1.

More formally, there are martingales (X_k) and a.s. finite stopping times τ for which $\mathbb{E}\, X_\tau \neq \mathbb{E}\, X_0$. This situation can occur for a number of different reasons, including when the stopping time τ is unbounded or the martingale lacks suitable integrability properties. Nevertheless, there are simple conditions under which we reach conclusions about the expected value of a martingale at a stopping time.

The next result describes some conditions where the expected value of a martingale at a stopping time coincides with the expectation of its initial value. Among other things, this kind of result is useful for studying hitting times.

Theorem 24.18 (Optional stopping). Fix a probability space and a filtration. Let (X_k) be a martingale, and let τ be a stopping time that is almost surely finite. Fix a (nonrandom) number $B \ge 0$. Then $\mathbb{E}\, X_\tau = \mathbb{E}\, X_0$ under *any one of* the following assumptions:

This list of conditions is not exhaustive!

1. The stopping time τ is bounded. That is, $\tau \le B$.
2. The martingale (X_k) is uniformly bounded and the stopping time τ is almost surely finite. That is, $|X_k(\omega)| \le B$ uniformly for each time k and each sample point ω.
3. The *increments* of the martingale are uniformly bounded and the stopping time τ is integrable. That is, $\mathbb{E}\,\tau < +\infty$ and $|X_{k+1}(\omega) - X_k(\omega)| \le B$ for each time k and sample point ω.

Proof. This result follows when we take appropriate limits of the fact that $\mathbb{E}\, X_{k\wedge\tau} = \mathbb{E}\, X_0$ for all k.

First, suppose that the stopping time τ is bounded; say, $\tau \le B$ for a natural number B. Then we may take $k = B$ to obtain

$$\mathbb{E}\, X_0 = \mathbb{E}\, X_{B\wedge\tau} = \mathbb{E}\, X_\tau.$$

This is the first result.

Second, suppose that the martingale is uniformly bounded and the stopping time is finite, almost surely. Then

$$X_{k\wedge\tau} = X_{k\wedge\tau}\mathbb{1}_{\tau<+\infty} \to X_\tau \quad \text{almost surely as } k \to \infty.$$

The second result now follows by taking expectations and invoking the bounded convergence theorem.

Third, suppose that the stopping time is integrable and the martingale increments are uniformly bounded, say, by B. Then

$$|X_{k\wedge\tau} - X_0| = \left|\sum\nolimits_{i=1}^{k\wedge\tau} (X_i - X_{i-1})\right| \le B\tau.$$

Since τ is integrable, we can take the limit of $0 = \mathbb{E}[X_{k\wedge\tau} - X_0]$ as $k \to \infty$ using dominated convergence. ∎

Exercise 24.19 (Optional stopping: Submartingales). Fix a probability space and a filtration. Consider a *sub*martingale (X_k) and a stopping time τ.

1. Show that $\mathbb{E}[X_\tau] \ge \mathbb{E}[X_0]$ under each one of the conditions in Theorem 24.18.
2. If (X_k) is positive and τ is a.s. finite, show that $\mathbb{E}[X_\tau] \ge \mathbb{E}[X_0]$.
3. Assume that $\tau \le B$ almost surely for a fixed integer $B \in \mathbb{Z}_+$. Deduce that $\mathbb{E}[X_\tau] \le \mathbb{E}[X_B]$. **Hint:** Invoke Proposition 24.4 to analyze the decomposition

$$X_B - X_\tau = \sum\nolimits_{i=1}^{B} \mathbb{1}\{\tau < i\} \cdot (X_i - X_{i-1}).$$

4. Consider stopping times that satisfy $\tau \le \tau' \le B$ almost surely for a fixed integer $B \in \mathbb{Z}_+$. Deduce that $\mathbb{E}[X_\tau] \le \mathbb{E}[X_{\tau'}]$.
5. Develop parallel results for supermartingales.

Problem 24.20 (Optional stopping: Weaker conditions). Consider a martingale sequence (X_k) and a stopping time τ that satisfy

i. $\mathbb{P}\{\tau < +\infty\} = 1$;
ii. $\mathbb{E}|X_\tau| < +\infty$; and
iii. $\mathbb{E}[X_k \mathbb{1}\{\tau > k\}] \to 0$ as $k \to \infty$.

Prove that these conditions imply $\mathbb{E}[X_\tau] = \mathbb{E}[X_0]$. **Hint:** Work with the decomposition $X_\tau = X_{\tau\wedge k} + (X_\tau - X_k)\mathbb{1}\{k > \tau\}$.

Problems

Exercise 24.21 (Martingale increments in L_2). Consider a martingale $(X_k : k \in \mathbb{Z}_+)$ taking values in L_2, and define the difference sequence $\Delta_k := X_k - X_{k-1}$ for each $k \in \mathbb{Z}_+$.

1. Prove that the difference sequence is orthogonal: $\mathbb{E}[\Delta_j \Delta_k] = 0$ when $j \ne k$. **Hint:** Use the tower law.
2. Deduce that $\mathbb{E}[X_k^2] = \sum_{i=1}^{k} \mathbb{E}[\Delta_i^2]$ for each index $k \in \mathbb{Z}_+$.

Problem 24.22 (Doob decomposition). Let $(X_k : k \in \mathbb{Z}_+)$ be a *submartingale* sequence with respect to a filtration $(\mathcal{F}_k : k \in \mathbb{Z}_+)$.

1. Prove that we can decompose

$$X_k = M_k + C_k \quad \text{for } k \in \mathbb{Z}_+$$

where $(M_k : k \in \mathbb{N})$ is a martingale sequence and $(C_k : k \in \mathbb{N})$ is an *increasing* previsible process with $C_0 = 0$, called a *compensator*. **Hint:** The compensator process has increments $C_{k+1} - C_k := \mathbb{E}[X_{k+1} \,|\, \mathcal{F}_k] - X_k$.
2. Show that the Doob decomposition is uniquely determined (up to values on negligible sets).
3. Apply the Doob decomposition to give an alternative account of the results in Exercise 24.19.

Problem 24.23 (Time to ruin: Fair game). Consider the simple symmetric random walk: $X_0 = 0$ and $X_{k+1} = X_k + Z_k$, where the increments $Z_k \sim \text{UNIFORM}\{\pm 1\}$ are independent. For $a, b \in \mathbb{N}$, we would like to understand whether the random walk first visits the number $-a$ or the number $+b$.

This is an idealization of a fair game where you win or lose one unit of capital at each play, and (X_k) tabulates your total winnings (or losses). Suppose that you have a units of capital to start, and you are hoping to win b more units so that you can purchase an aspirational good, such as a Taylor Swift NFT. If you visit $-a$ first, then you have depleted all of your capital. Conversely, if you visit b first, your total fortune is augmented to $a + b$, and you can acquire the cryptotoken.

1. Check that the sequence (X_k) is a martingale with respect to $\mathcal{F}_k = \sigma(Z_1, \dots, Z_k)$.
2. Let $\tau := \inf\{k \in \mathbb{Z}_+ : X_k = -a \text{ or } X_k = +b\}$. Confirm that τ is a stopping time.
3. Prove that the stopping time is integrable: $\mathbb{E}[\tau] < +\infty$. In particular, τ is a.s. finite. **Hint:** Observe that $\mathbb{P}\{\tau > k\} \le \mathbb{P}\{|X_k| \le a \wedge b\}$. Bound the probability below using the Paley–Zygmund inequality (Exercise 12.23).
4. Verify that the martingale (X_k) and stopping time τ satisfy one the conditions of the optional stopping theorem (Theorem 24.18), so that $\mathbb{E}[X_\tau] = \mathbb{E}[X_0] = 0$. Reinterpret this statement to see that the probability of ruin satisfies

$$\mathbb{P}\{X_\tau = -a\} = \frac{b}{a+b}.$$

Discuss what this result means in terms of the game.
5. Check that the random variables $M_k := X_k^2 - k$ for $k \in \mathbb{Z}_+$ compose a martingale with respect to the same filtration.
6. Do the martingale (M_k) and the stopping time τ satisfy any one of the hypotheses of Theorem 24.18? What about Problem 24.20?
7. Compute $\mathbb{E}[M_\tau]$ two different ways to prove that $\mathbb{E}[\tau] = ab$. Discuss what this result means.

Exercise 24.24 (Probability of ruin: Unfair games). Consider the simple random walk: $S_0 = a \in \mathbb{N}$ and $S_{k+1} = S_k + Z_k$, where the independent increments $Z_k = +1$ with probability p and $Z_k = -1$ with probability $1 - p =: q$. For a number $N \in \mathbb{N}$, we would like to understand whether the sequence S_k first visits 0 or N. We assume that $\varrho := q/p \neq 1$, so this model is an idealization of an unfair game.

1. Confirm that the random variables $X_k = \varrho^{S_k}$ compose a martingale with respect to $\mathcal{F}_k = \sigma(Z_1, \dots, Z_k)$.
2. Let $\tau := \inf\{k \in \mathbb{Z}_+ : S_k = 0 \text{ or } S_k = N\}$. Confirm that τ is a stopping time.

3. Verify that the martingale (X_k) and stopping time τ satisfy one the conditions of the optional stopping theorem (Theorem 24.18), so that $\mathbb{E}[X_\tau] = \mathbb{E}[X_0] = 1$. Reinterpret this statement to see that

$$\mathbb{P}\{X_\tau = 0\} = \frac{\varrho^k - \varrho^N}{1 - \varrho^N}.$$

Discuss what this result means in terms of the game.

Problem 24.25 (Wald identities). In this problem, we consider the sum $S_k = \sum_{i=1}^k Z_i$ of i.i.d. copies of a random variable Z. Let τ be a stopping time with respect to the filtration $\mathcal{F}_k = \sigma(Z_1, \dots, Z_k)$ with the property that $\mathbb{E}[\tau] < +\infty$. We can use the optional stopping theorem (Theorem 24.18) to analyze the randomly stopped sum X_τ.

1. Assuming that $Z \in \mathsf{L}_1$, prove that

$$\mathbb{E}[X_\tau] = \mathbb{E}[\tau] \cdot \mathbb{E}[Y].$$

Hint: Consider $X_k := S_k - k\,\mathbb{E}[Y]$.

2. Assuming that $Y \in \mathsf{L}_2$, prove that

$$\operatorname{Var}[X_\tau] = \mathbb{E}[\tau] \cdot \operatorname{Var}[Y].$$

Hint: Consider the quadratic martingale described in Exercise 23.16. First consider the bounded stopping time $\tau' := \tau \wedge N$. Then take a limit as $N \to \infty$.

3. Assume that $Y \in \mathsf{L}_\infty$ with mgf $m_Y(\theta) := \mathbb{E}[\mathrm{e}^{\theta Y}]$. Consider the sequence with $S_0(\theta) = 1$ and

$$S_k(\theta) := \frac{\mathrm{e}^{\theta S_k}}{m_Y(\theta)^k} \quad \text{for fixed } \theta \in \mathbb{R}.$$

Prove that (S_k) is a martingale sequence. Confirm that

$$\mathbb{E}[S_\tau(\theta)] = 1 \quad \text{provided that } m_Y(\theta) \geq 1.$$

This result is often used to compute the distribution of first passage times, but the details of the argument are subtle. **Hint:** Use the third condition of the optional stopping theorem.

Notes

This lecture is based on Williams's book [Wil91]. Many problems are adapted from Grimmett & Stirzaker [GS01].

Lecture bibliography

[GS01] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. 3rd ed. Oxford University Press, 2001.

[Wil91] D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

25. Martingale Convergence

> "Remain true to yourself, but move ever upward toward greater consciousness and greater love! At the summit you will find yourselves united with all those who, from every direction, have made the same ascent. For everything that rises must converge."
>
> —*The Omega Point*, Pierre Teilhard de Chardin

Agenda:
1. Doob's convergence theorem
2. Convergence and crossings
3. Upcrossing inequality
4. Proof of Doob's theorem
5. *Uniform integrability

Martingales are much-loved by probabilists because they converge to a limiting random variable under minimal conditions. Supermartingale and submartingales also enjoy strong convergence properties.

There are many random processes that have an obvious martingale structure (e.g., prediction of a random variable from noisy observations), and the martingale convergence theorem ensures that these processes converge to an equilibrium. There are other random processes where there is a "hidden" martingale structure that can be exploited to understand the limiting behavior.

In this lecture, we will state and prove the most basic martingale convergence theorem, due to Joseph Doob. The argument reframes the question about convergence as a question about the number of times the martingale traverses a range of values. This proof strategy is very powerful, and it extends almost verbatim to continuous-time martingales.

25.1 Doob's convergence theorem

The main result of this lecture states that a martingale converges almost surely under a mild integrability condition.

Theorem 25.1 (Martingale: Convergence). Fix a probability space $(\Omega, \mathcal{F}, \mathbb{P})$ and a filtration $(\mathcal{F}_k : k \in \overline{\mathbb{Z}}_+)$. Suppose that $(X_k : k \in \mathbb{Z}_+)$ is a *uniformly bounded* martingale. That is, for a fixed number R,

$$\mathbb{E}\,|X_k| \le R \quad \text{for all } k \in \mathbb{Z}_+.$$

Then $X_k \to X_\infty$ almost surely, where X_∞ is an integrable random variable. In particular, X_∞ is almost surely finite.

The proof of Theorem 25.1 occupies the rest of the lecture.

In words, a uniformly bounded martingale converges almost surely to a finite limit. We may assume that the limit X_∞ is a random variable because

$$X_\infty(\omega) = \liminf_{k\to\infty} X_k(\omega) \quad \text{almost surely.}$$

We can view this random variable X_∞ as the equilibrium of the martingale process. To identify the limiting random variable, we must conduct a separate and specialized investigation.

Warning 25.2 (Convergence in L_1). The martingale convergence provided by Theorem 25.1 does not imply that (X_k) converges in L_1. In particular, $\mathbb{E}\, X_k$ may not converge to $\mathbb{E}\, X_\infty$. This conclusion requires further assumptions or a separate argument. See Section 25.5 for a discussion of uniformly integrable martingales, where the situation is simpler.

Martingales that are uniformly bounded in L_2 also enjoy a more satisfactory convergence theory; see Problem 25.31. ▪

25.1.1 Extensions

For clarity of argument, we will prove Theorem 25.1 for a martingale, but the same proof applies to supermartingales and submartingales with essentially no change.

Exercise 25.3 (Supermartingale: Convergence). Suppose that $(X_k : k \in \mathbb{Z}_+)$ is a *uniformly bounded supermartingale*. Then $X_k \to X_\infty$ almost surely, where X_∞ is an integrable random variable. Prove this result by surgical modifications to the proof of Theorem 25.1.

Exercise 25.4 (Submartingale: Convergence). Use Exercise 25.3 to deduce that a *uniformly bounded submartingale* converges almost surely to an integrable random variable.

In certain circumstances, we can even avoid placing a separate boundedness assumption on the random process.

Exercise 25.5 (Positive supermartingale: Convergence). Use Exercise 25.3 to deduce that a *positive supermartingale* converges almost surely to an integrable, positive random variable.

25.1.2 Example: Urn models

As a first example, let us investigate how Doob's theorem applies to the urn model described in Exercise 23.17.

Let us recall the setting. Initially, a candy bowl contains one red and one black Skittle. At each time instant $k = 1, 2, 3, \dots$, a uniformly random Skittle is extracted from the bowl, and we return this Skittle to the bowl along with a new Skittle of the same color. Let R_k and B_k denote the number of red and black Skittles in the bowl after k steps.

It is not hard to confirm that the (random) proportion $M_k := R_k/(R_k + B_k)$ of red Skittles at time k composes a martingale sequence. This is the content of Exercise 23.17. It is also clear that $\mathbb{E}\,|M_k| \le 1$ for all k, so the martingale is uniformly bounded in L_1. Therefore, Doob's martingale convergence theorem (Theorem 25.1) furnishes a limiting random variable M_∞ with the property that $M_k \to M_\infty$ almost surely. As a consequence, $M_k \rightsquigarrow M_\infty$ in distribution.

Figure 25.1 illustrates the a.s. convergence of the sample paths of the martingale (M_k). You can also see that the density of sample paths settles down to an equilibrium as time passes. What is the limiting distribution? As it happens, $M_\infty \sim \textsc{uniform}[0, 1]$.

Doob's theorem does *not* assert that $\mathbb{E}\, M_k \to \mathbb{E}\, M_\infty$. Nevertheless, the sequence (M_k) also converges in L_1 because it is uniformly integrable; see Theorem 25.28.

Exercise 25.6 (Skittles: Limiting distribution). For the martingale $(M_k : k \in \mathbb{N})$ defined in this section, prove that the limiting random variable M_∞ follows the $\textsc{uniform}[0, 1]$

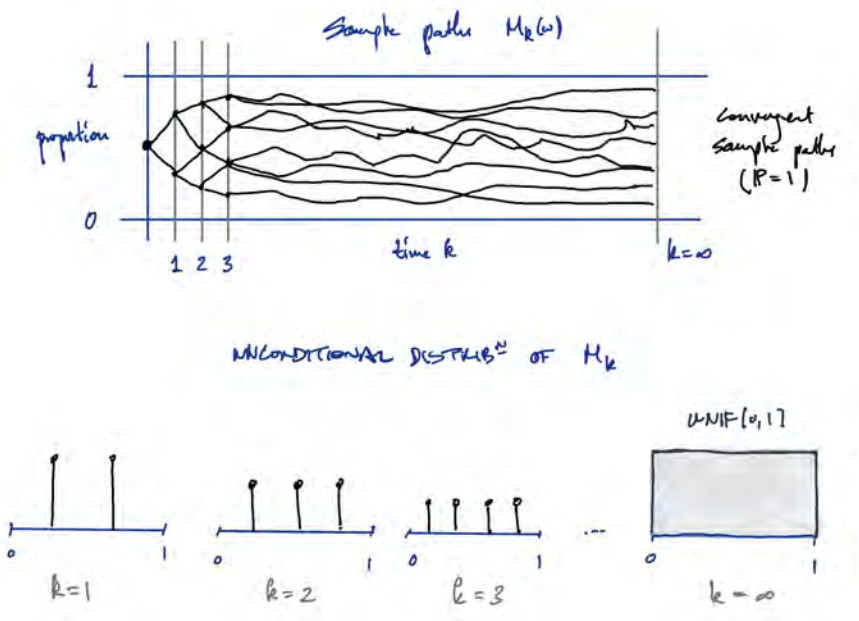


Figure 25.1 (Urn model: Convergence). The top chart illustrates the almost sure convergence of the sample paths of the proportion M_k of red Skittles at time k in the candy bowl to a limiting distribution. The panels below show the unconditional distribution of M_k at several specific instants $k \in \{1, 2, 3, \infty\}$. You can see that the proportion of red Skittles tends to a uniform limit.

distribution. **Hint:** During the first n draws, what is the probability that we select a red Skittle in each of the first k draws? What is the probability that we select exactly k red Skittle during the first n draws?

Problem 25.7 (Skittles: Initial conditions and limits). Consider the more general setting where the bowl initially contains r red Skittles and b black Skittles. As before, we draw a uniformly random Skittle and we return this Skittle to the bowl along with another Skittle of the same color. Confirm that the proportion $M_k := R_k/(R_k + B_k)$ of red Skittles is still a martingale that is uniformly bounded in L_1. Prove that the limiting distribution $M_\infty \sim \textsc{beta}(r, b)$.

25.1.3 Example: Branching processes

As a second example, let us investigate how Doob's theorem applies to the branching process described in Exercise 23.18.

Let Y be a positive, integer-valued random variable that models the size of a family. We require that $s := \mathbb{E}\, Y$ be finite. For simplicity, assume that the initial population size $Z_0 = 1$, and the population size evolves as

$$Z_{k+1} = \sum_{i=1}^{Z_k} Y_i^{(k)} \quad \text{for } k \in \mathbb{Z}_+,$$

where the $Y_i^{(k)}$ are i.i.d. copies of Y. Define the relative population size

$$M_k := Z_k/s^k \quad \text{for } k \in \mathbb{Z}_+.$$

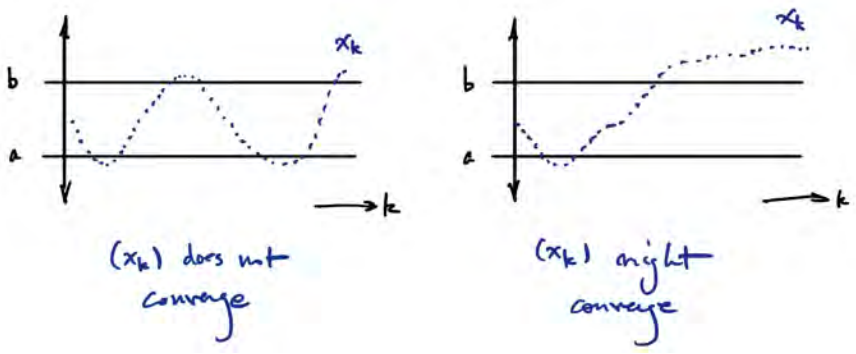


Figure 25.2 (Crossings and convergence). A sequence cannot converge if it crosses an interval an infinite number of times.

The scaling s^k is important to make sure that the sequence M_k does not trivially collapse or blow up.

It is not hard to check that $(M_k : k \in \mathbb{Z}_+)$ composes a positive martingale sequence. This is the content of Exercise 23.18. As a consequence, the sequence must be uniformly bounded in L_1. Doob's theorem (Theorem 25.1) yields the almost sure limit $M_k \to M_\infty$, where M_∞ is a finite, integrable random variable. This result can be interpreted as a statement that the random population size has a distribution relative to the baseline population growth s^k, which is exponentially increasing or decreasing, depending on the value s of the typical family size.

What is the nature of the limiting distribution? When the family size $s < 1$, the population dies out, and $M_\infty = 0$ almost surely. Observe that $0 = \mathbb{E}\, M_\infty \neq \mathbb{E}\, M_k = 1$, so this is a situation where the martingale fails to converge in L_1!

When the family size $s > 1$ and $\mathbb{E}[Y \log Y] < +\infty$, then the random variable M_∞ has a nontrivial distribution. Unfortunately, it is not possible to obtain formulas for the limit, except in very special cases.

25.2 Convergence and crossings

Before we turn to the probabilistic parts of the proof, we need to develop some conditions under which a real-valued sequence converges to a limit.

25.2.1 Interval sandwiches

The key observation is that a convergent sequence cannot oscillate across an interval. See Figure 25.2 for an illustration.

Lemma 25.8 (Interval sandwich). A (nonrandom) real-valued sequence $(x_k : k \in \mathbb{Z}_+)$ *fails to converge* to a limit in $\overline{\mathbb{R}}$ if and only if there are *rational* numbers $a < b$ with $a, b \in \mathbb{Q}$ for which

$$\liminf_{k\to\infty} x_k < a < b < \limsup_{k\to\infty} x_k.$$

Exercise 25.9 (Interval sandwich). Prove Lemma 25.8.

In other words, we can witness the failure of a sequence to converge by identifying an interval $[a, b]$ with rational endpoints where the sequence takes an infinite number of values below a and an infinite number of values above b.

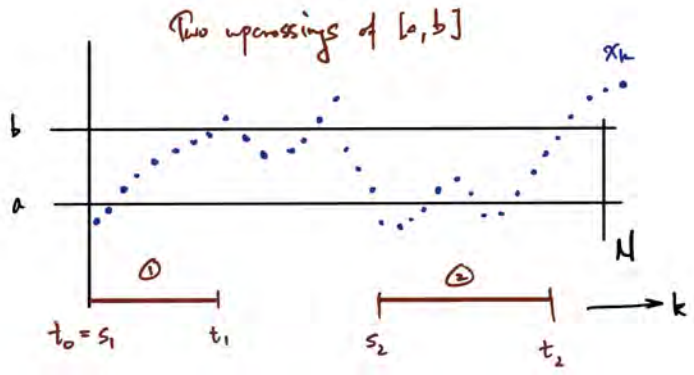


Figure 25.3 (Completed upcrossings). This diagram shows a sequence that completes two upcrossings of the interval $[a, b]$ before the index N. The first upcrossing (s_1, t_1) occurs right at the beginning of the sequence. Afterward, the sequence wanders for a while before dropping below the level a. The second upcrossing (s_2, t_2) occurs at this point, including a section that rises above the level a and falls below a before continuing upward past the level b.

25.2.2 Upcrossings

This observation suggests that we count how many times the sequence passes upward through the interval.

Definition 25.10 (Upcrossing). Consider a nonrandom real sequence $(x_k : k \in \mathbb{Z}_+)$, and fix real numbers $a, b \in \mathbb{R}$. An *upcrossing* of the interval $[a, b]$ is a pair (s, t) of indices for which $x_s < a < b < x_t$.

Definition 25.11 (Completed upcrossings). Consider a nonrandom real sequence $(x_k : k \in \mathbb{Z}_+)$, real numbers $a, b \in \mathbb{R}$, and an index $N \in \mathbb{Z}_+$. The number $u_N[a, b]$ denotes the (maximum) number of upcrossings that are completed by time N. The total number $u_\infty[a, b]$ of upcrossings is given by

$$u_\infty[a, b] := \lim_{N\to\infty} u_N[a, b] \in \overline{\mathbb{Z}}_+.$$

The monotone limit is always defined, but it may take the value $+\infty$.

More formally, we define pairs of upcrossings recursively. Let $t_0 = 0$, and construct

$$\begin{aligned} s_1 &:= \inf\{k \geq t_0 : x_k < a\} \quad \text{and} \quad t_1 := \inf\{k > s_1 : x_k > b\}; \\ s_2 &:= \inf\{k > t_1 : x_k < a\} \quad \text{and} \quad t_2 := \inf\{k > s_2 : x_k > b\}; \quad \dots. \end{aligned}$$

Then $u_N[a, b] = \sup\{m \in \mathbb{Z}_+ : t_m \leq N\}$. See Figure 25.3 for an illustration.

25.2.3 Convergence from upcrossings

We can now express convergence properties of a real sequence in terms of the number of times that it crosses an interval.

Lemma 25.12 (Finite upcrossings). Consider a real sequence $(x_k : k \in \mathbb{Z}_+)$. If $u_\infty[a, b] <$

$+\infty$ for *all rational* numbers $a < b$, then (x_k) converges to a limit in $\overline{\mathbb{R}}$.

Proof. Let us establish the contrapositive: If (x_k) *fails* to converge to a limit in $\overline{\mathbb{R}}$, then $u_\infty[a, b] = +\infty$ for some rational numbers $a < b$.

According to Lemma 25.8, the premise is equivalent to the condition that

$$\liminf_{k\to\infty} x_k < a < b < \limsup_{k\to\infty} x_k \quad \text{for some rational } a < b.$$

Consider the sequences $(s_i : i \in \overline{\mathbb{Z}}_+)$ and $(t_i : i \in \overline{\mathbb{Z}}_+)$ from Definition 25.11. Since the limit inferior of (x_k) is smaller than a, the times s_i are all finite. Since the limit superior of (x_k) is bigger than b, the times t_i are all finite. Furthermore, each pair (s_i, t_i) is a distinct upcrossing of the interval $[a, b]$. We conclude that $u_\infty[a, b] = +\infty$. ∎

25.3 Upcrossing inequalities

Our next goal is to count the number of times that a martingale sequence makes an upcrossing of a fixed interval. To do so, we recall that martingale increments model the outcomes of a fair game. If the martingale values cross an interval repeatedly, we could prescribe a betting strategy that exploits this event to make money. Unfortunately, we have already seen that it is impossible to profit on a fair game on average, so we must conclude that the martingale only traverses the interval a finite number of times.

25.3.1 Betting on upcrossings

Consider a martingale sequence $(X_k : k \in \mathbb{Z}_+)$. The martingale increments $\Delta_i := X_i - X_{i-1}$ model the outcomes of a sequence of fair games. We can make (linear) bets on the values of the increments in the hope of turning a profit.

We will develop a betting strategy based on the surmise that the martingale repeatedly traverses a particular interval $[a, b]$ upward, where $a < b$ are real numbers. The basic idea is that we start betting \$1 per game as soon as the martingale dips below the level a, and we continue betting \$1 per game until the martingale exceeds the level b. Then we bet \$0 per game until the value of the martingale falls below a again. See Figure 25.4.

The endpoints of the interval are parameters in the argument. We will see that there is no benefit from picking any particular set of endpoints.

Formally, we define the sequence $(C_k : k \in \mathbb{N})$ of bets:

$$\begin{aligned} C_1 &:= \mathbb{1}\{X_0 < a\} \\ C_k &:= \mathbb{1}\{C_{k-1} = 1, X_{k-1} \le b\} + \mathbb{1}\{C_{k-1} = 0, X_{k-1} < a\} \quad \text{for } k \ge 2. \end{aligned} \tag{25.1}$$

You should convince yourself that this sequence implements exactly the strategy described in the previous paragraph.

Exercise 25.13 (Upcrossing: Bets). Show that the sequence $(C_k : k \in \mathbb{N})$ is positive, bounded, and previsible.

The martingale transform of (X_k) by the sequence (C_k) models your accumulated winnings from this betting strategy:

$$Y_k := (\boldsymbol{C} \cdot \boldsymbol{X})_k = \sum\nolimits_{i=1}^{k} C_i(X_i - X_{i-1}) \quad \text{for } k \in \mathbb{Z}_+,$$

with the understanding that $Y_0 = 0$. Proposition 24.4 shows that (Y_k) is also a martingale. In particular, your expected winnings $\mathbb{E}\, Y_k = \mathbb{E}\, Y_0 = 0$ at each time $k \in \mathbb{Z}_+$.

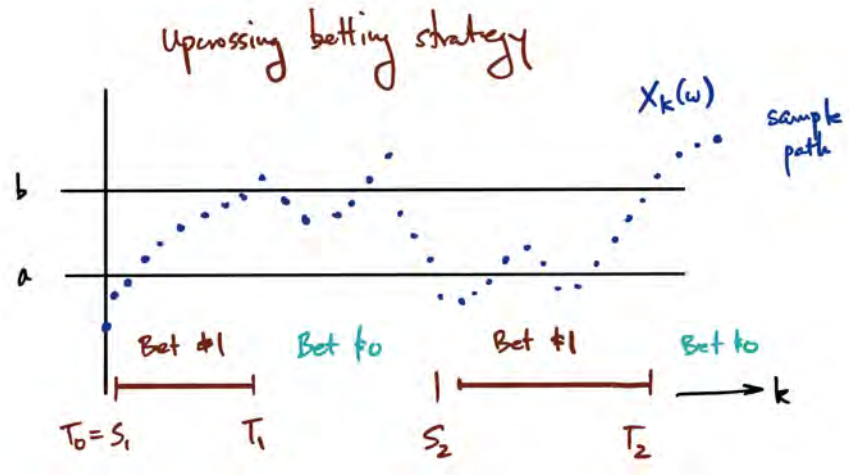


Figure 25.4 (Betting on upcrossings). This diagram shows the upcrossing betting strategy for a particular sample path $X_k(\omega)$ and a particular interval $[a, b]$.

25.3.2 Snell's inequality

We can develop a coarse lower bound on the accumulated winnings from the upcrossing betting strategy. This bound allows us to control the total number of upcrossings of a sample path of the martingale (X_k).

For each fixed time $N \in \mathbb{Z}_+$ and each sample point $\omega \in \Omega$, define the number $U_N[a, b](\omega)$ of upcrossings of $[a, b]$ that the sample path $(X_k(\omega) : k \in \mathbb{Z}_+)$ completes by time N. Similarly, we introduce the total number $U_\infty[a, b](\omega)$ of upcrossings of $[a, b]$ achieved by the entire sample path $X_k(\omega)$.

Exercise 25.14 (Upcrossing random variables). Confirm that $U_N[a, b]$ and $U_\infty[a, b]$ are positive random variables.

The next result bounds the expected number of completed upcrossings at time N in terms of the expected value of the martingale at time N.

Proposition 25.15 (Snell's inequality). Let (X_k) be a martingale. For all real numbers $a < b$ and integers $N \in \mathbb{Z}_+$,

$$(b - a) \cdot \mathbb{E}\, U_N[a, b] \le \mathbb{E}(X_N - a)_-.$$

Recall that

$$(x)_- := \max\{-x, 0\} \ge 0.$$

There are two phenomena that affect the expected number of completed upcrossings. The wider the interval $[a, b]$, the fewer upcrossings we can anticipate. The integrability properties of the martingale also play a role.

Proof. We consider the martingale transform $Y_k = (\boldsymbol{C} \cdot \boldsymbol{X})_k$ of the martingale (X_k) applied to the betting strategy (C_k), defined in (25.1).

Using the number $U_N[a, b](\omega)$ of upcrossings, we can develop a lower bound on our total winnings $Y_N(\omega)$ up to time N:

$$Y_N(\omega) \ge (b - a) \cdot U_N[a, b](\omega) - (X_N(\omega) - a)_-.$$

Indeed, we win at least $(b - a)$ dollars each time the sample path $(X_k(\omega))$ completes an upcrossing of the interval $[a, b]$. After the last upcrossing, completed before time N, there is a final interval of play. If the martingale passes below a during this interval,

then we have started betting actively. The amount that we lose during this interval of play is bounded above by $(X_N(\omega) - a)_-$, the amount that the martingale drops below the level a.

Since (Y_k) is a martingale, at the fixed time N, we have

$$0 = \mathbb{E}\, Y_0 = \mathbb{E}\, Y_N \geq (b-a) \cdot \mathbb{E}\, U_N[a,b] - \mathbb{E}(X_N - a)_-.$$

The second inequality follows from the display in the last paragraph. Rearrange to complete the proof. ∎

25.3.3 Infinite upcrossings are negligible

The key implication of Snell's inequality is that a uniformly bounded martingale has zero probability of traversing any given interval $[a, b]$ an infinite number of times.

Corollary 25.16 (Infinite upcrossings are negligible). Assume (X_k) is a martingale that is uniformly bounded in L_1. Fix real numbers $a < b$. Then $\mathbb{P}\{U_\infty[a,b] = +\infty\} = 0$.

Proof. Suppose that $\mathbb{E}\,|X_k| \leq R$ uniformly for all k. According to Snell's inequality (Proposition 25.15),

$$(b-a) \cdot \mathbb{E}\, U_N[a,b] \leq \mathbb{E}(X_N - a)_- \leq \mathbb{E}\,|X_N| + |a| \leq R + |a|.$$

Note that $U_N[a,b] \uparrow U_\infty[a,b]$ as $N \to \infty$. Therefore, we can invoke the bounded convergence theorem to deduce that

$$(b-a) \cdot \mathbb{E}\, U_\infty[a,b] \leq R + |a|.$$

Since $U_\infty[a,b]$ is a positive random variable, its expectation is finite only if it takes exclusively finite values, almost surely: $\mathbb{P}\{U_\infty[a,b] = +\infty\} = 0$. ∎

25.3.4 *Dubins's inequality

For a *positive* martingale $(X_k : k \in \mathbb{N})$, it is possible to prove a much sharper bound on the number of upcrossings. Fix positive real numbers $a < b$. As in Definition 25.11, we may introduce (random) times $T_0 = 0$ and

$$\begin{aligned} S_1 &:= \inf\{k \geq T_0 : x_k < a\} \quad \text{and} \quad T_1 := \inf\{k > S_1 : x_k > b\}; \\ S_2 &:= \inf\{k > T_1 : x_k < a\} \quad \text{and} \quad T_2 := \inf\{k > S_2 : x_k > b\}; \quad \dots. \end{aligned}$$

Then $U_N[a,b] = \sup\{m \in \mathbb{Z}_+ : T_m \leq N\}$.

Exercise 25.17 (Upcrossing: Stopping times). Show that each T_k is a stopping time, and each S_k is a stopping time.

Exercise 25.18 (Upcrossing: Bets). Express the sequence (C_k) of bets from (25.1) more succinctly using the stopping times (S_i) and (T_i).

Problem 25.19 (Dubins's upcrossing inequality). Let $(X_k : k \in \mathbb{Z}_+)$ be a *positive* martingale, and fix positive real numbers $a < b$. For each index $k \in \mathbb{N}$, let T_k be the stopping time described above. Then

$$\mathbb{P}\{T_k < \infty\} \leq (a/b) \cdot \mathbb{P}\{T_{k-1} < \infty\}.$$

In particular, $\mathbb{P}\{T_k < +\infty\} \leq (a/b)^k$. Deduce that $\mathbb{P}\{U_\infty[a,b] = +\infty\} = 0$.

Hint: For any fixed index N, show that

$$\begin{aligned} b \cdot \mathbb{P}\{T_k \leq N\} + \mathbb{E}[X_N \mathbb{1}\{T_k > N\}] &\leq \mathbb{E}\, X_{T_k \wedge N} \\ &= \mathbb{E}\, X_{S_k \wedge N} \leq a \cdot \mathbb{P}\{S_k \leq N\} + \mathbb{E}[X_N \mathbb{1}\{S_k > N\}]. \end{aligned}$$

Exercise 25.20 (Dubins: Supermartingale case). Show that Dubins's inequality also holds for a *positive supermartingale*.

25.4 Doob's martingale convergence theorem: Proof

We are now prepared to prove Doob's convergence theorem (Theorem 25.1). For concreteness, define the limiting random variable

$$X_\infty(\omega) := \liminf_{k\to\infty} X_k(\omega) \quad \text{for all } \omega \in \Omega.$$

Consider the event

$$\mathsf{E} := \{\omega : X_k(\omega) \text{ does } \textit{not} \text{ have a limit in } \overline{\mathbb{R}} \text{ as } k \to \infty\}.$$

Using Lemma 25.8 and Lemma 25.12, we can express this event as

$$\begin{aligned}\mathsf{E} &= \bigcup_{\substack{a<b\\ a,b\in\mathbb{Q}}} \{\omega : \liminf_{k\to\infty} X_k(\omega) < a < b < \limsup_{k\to\infty} X_k(\omega)\} \\ &\subseteq \bigcup_{\substack{a<b\\ a,b\in\mathbb{Q}}} \{\omega : U_\infty[a,b](\omega) = +\infty\}.\end{aligned}$$

Corollary 25.16 states that each event in the latter union has probability zero. Since the union is indexed by a countable set (pairs of rational numbers), we deduce that E is an event with $\mathbb{P}(\mathsf{E}) = 0$. As a consequence,

$$X_\infty(\omega) = \lim_{k\to\infty} X_k(\omega) \quad \text{almost surely.}$$

Indeed, the limit and limit inferior coincide whenever the limit exists.

To confirm that X_∞ is almost surely finite, we bound its expectation using Fatou's lemma (Theorem 9.11):

$$\mathbb{E}\,|X_\infty| = \mathbb{E}[\liminf_{k\to\infty} |X_k|] \le \liminf_{k\to\infty} \mathbb{E}\,|X_k| \le R,$$

where R is the uniform L_1 bound on the martingale (X_k). In other words, X_∞ is integrable. In particular, X_∞ must be finite almost surely.

25.5 *Uniformly integrable martingales

A disappointing feature of Doob's convergence theorem is that it does not imply the L_1 convergence of a martingale (X_k). A natural question is when we can assert that $\mathbb{E}\,X_k \to \mathbb{E}\,X_\infty$. To address this problem, we give a brief tour of the theory of uniformly integrable martingales.

25.5.1 Uniform integrability

The notion of uniform integrability arises when we attempt to connect L_1 convergence with notions of pointwise convergence. As we have discussed, these two concepts are incomparable with each other. Uniform integrability offers a bridge.

First, we observe that the tails $\mathbb{P}\{|X| \ge t\}$ of an integrable random variable X must decay at least as fast as t^{-1} as $t \to \infty$. The next result gives an alternative perspective on this fact.

Exercise 25.21 (Integrable random variable: Tails). Suppose that X is an integrable random variable. Show that

$$\mathbb{E}\left[|X| \cdot \mathbb{1}\{|X| > R\}\right] \to 0 \quad \text{as } R \to \infty.$$

Motivated by this exercise, we carve out a family of integrable random variables whose tails all decay uniformly.

Definition 25.22 (Uniform integrability). A family $(X_t : t \in \mathsf{T})$ of random variables is *uniformly integrable* (UI) if

$$\sup_{t \in \mathsf{T}} \mathbb{E}\left[|X_t| \cdot \mathbb{1}\{|X_t| > R\}\right] \to 0 \quad \text{as } R \to \infty.$$

Uniform integrability is a kind of compactness property. Among other things, a UI family is uniformly bounded in L_1. In particular, UI random variables are integrable.

Exercise 25.23 (Uniform integrability: L_1 boundedness). Show that a UI family $(X_t : t \in \mathsf{T})$ is uniformly bounded in L_1. That is, $\sup_t \|X_t\|_{\mathsf{L}_1} \le R$ for a finite number R.

Uniform integrability is the missing ingredient that we need to deduce L_1 convergence from almost-sure convergence.

Proposition 25.24 (Convergence in L_1: Sufficient condition). Consider a sequence $(X_k : k \in \mathbb{Z}_+)$ of integrable random variables. If the sequence is uniformly integrable and $X_k \to X_\infty$ almost surely, then $\|X_k - X_\infty\|_{\mathsf{L}_1} \to 0$.

Proof. First, we verify that the limit X_∞ is integrable. Using Fatou's lemma (Theorem 9.11),

$$\mathbb{E}\,|X_\infty| = \mathbb{E}[\liminf_{k\to\infty} |X_k|] \le \liminf_{k\to\infty} \mathbb{E}\,|X_k| < +\infty.$$

The last relation holds because a UI sequence is uniformly bounded (Exercise 25.23).

The key idea is to approximate the random variables in the sequence $(X_k : k \in \overline{\mathbb{Z}}_+)$ by thresholding them all at the same level. For a parameter $R \ge 0$, introduce the bounded, continuous function

$$\varphi_R(x) := \operatorname{sgn}(x) \cdot (|x| \wedge R).$$

By choosing R sufficiently large, we can ensure that

$$\mathbb{E}\,|\varphi_R(X_\infty) - X_\infty| \le \varepsilon \quad \text{and} \quad \mathbb{E}\,|\varphi_R(X_k) - X_k| \le \varepsilon \quad \text{for all } k \in \mathbb{Z}_+.$$

The first statement follows from Exercise 25.21; the second statement is the definition of uniform integrability. Since φ_R is bounded and continuous, the bounded convergence theorem implies that

$$\mathbb{E}\,|\varphi_R(X_k) - \varphi_R(X_\infty)| \to 0 \quad \text{as } k \to \infty.$$

Therefore, if we choose k large enough, the triangle inequality implies that

$$\mathbb{E}\,|X_k - X_\infty| \le \mathbb{E}\left[|X_k - \varphi_R(X_k)| + |\varphi_R(X_k) - \varphi_R(X_\infty)| + |\varphi_R(X_\infty) - X_\infty|\right] \le 3\varepsilon.$$

Take the limit as $k \to \infty$ and then as $\varepsilon \to 0$ to deduce that $X_k \to X_\infty$ in L_1. ∎

We can strengthen the result of Proposition 25.24 to obtain a necessary and sufficient condition for L_1 convergence.

Problem 25.25 (*Convergence in L_1: Characterization). Consider a sequence $(X_k : k \in \mathbb{Z}_+)$ of integrable random variables. Prove that the following statements are equivalent.

1. **Convergence in L_1:** The sequence (X_k) converges to an integrable random variable X_∞ with respect to the L_1 pseudonorm: $\|X_k - X_\infty\|_{\mathsf{L}_1} \to 0$.
2. **Convergence in probability + UI:** The sequence (X_k) is uniformly integrable, and it converges in probability to an integrable random variable X_∞.

Hint: The direction (2) ⇒ (1) is similar to the proof of Proposition 25.24. The direction (1) ⇒ (2) requires uniform boundedness of (X_k) in L_1 and Markov's inequality.

25.5.2 Example: Lévy–Doob martingales

The key example of a UI sequence is the Lévy–Doob martingale associated with a random variable and a filtration. Recall that these martingales arise when we improve our prediction of a random variable by progressively acquiring more information.

Proposition 25.26 (Lévy–Doob martingale: Uniform integrability). Let $(\mathscr{F}_k : k \in \overline{\mathbb{Z}}_+)$ be a filtration, and let Z be an integrable random variable. Consider the martingale sequence

$$X_k := \mathbb{E}[Z \,|\, \mathscr{F}_k] \quad \text{for } k \in \overline{\mathbb{Z}}_+.$$

Then $(X_k : k \in \overline{\mathbb{Z}}_+)$ is a UI family.

Proof. Jensen's inequality for conditional expectation (Proposition 20.6) yields

$$|X_k| = |\mathbb{E}[Z \,|\, \mathscr{F}_k]| \le \mathbb{E}[|Z| \,|\, \mathscr{F}_k].$$

Therefore, without loss, we may assume that $Z \ge 0$. In particular, the sequence (X_k) is positive as well.

Fix an index $k \in \mathbb{Z}_+$. We will obtain a bound on the tail expectation that is independent of k. For a parameter $R \ge 0$, we calculate that

$$\begin{aligned}\mathbb{E}[X_k \cdot \mathbb{1}\{X_k > R^2\}] &= \mathbb{E}[\mathbb{E}[Z \,|\, \mathscr{F}_k] \cdot \mathbb{1}\{X_k > R^2\}] \\ &= \mathbb{E}[\mathbb{E}[Z \cdot \mathbb{1}\{X_k > R^2\} \,|\, \mathscr{F}_k]] \\ &= \mathbb{E}[Z \cdot \mathbb{1}\{X_k > R^2\}].\end{aligned}$$

The second relation holds because $\{X_k > R^2\}$ is $\mathscr{F}_k$-measurable, so we can draw it inside the conditional expectation. Then we invoke the tower law to drop the conditioning. To continue, we decompose $Z = Z \cdot \mathbb{1}\{Z \le R\} + Z \cdot \mathbb{1}\{Z > R\}$ to make the bounds

$$\mathbb{E}[Z \cdot \mathbb{1}\{X_k > R^2\}] \le R \cdot \mathbb{P}\left\{X_k > R^2\right\} + \mathbb{E}[Z \cdot \mathbb{1}\{Z > R\}].$$

We can control the first term using Markov's inequality:

$$\mathbb{P}\left\{X_k > R^2\right\} \le (\mathbb{E}\, X_k)/R^2 = (\mathbb{E}\, Z)/R^2.$$

Sequencing the last three displays and taking the supremum over k,

$$\sup_k \mathbb{E}[X_k \cdot \mathbb{1}\{X_k > R^2\}] \le (\mathbb{E}\, Z)/R + \mathbb{E}[Z \cdot \mathbb{1}\{Z > R\}].$$

The right-hand side tends to zero as $R \to \infty$ because of Exercise 25.21. ∎

Exercise 25.27 (Conditional expectations: Uniform integrability). Establish that the family containing *all* conditional expectations of an integrable random variable Z is UI:

$$\{\mathbb{E}[Z \,|\, \mathscr{G}] : \mathscr{G} \subseteq \mathscr{F} \text{ and } \mathscr{G} \text{ a } \sigma\text{-algebra on } \Omega\}.$$

Hint: The proof is no different from Proposition 25.26.

25.5.3 UI martingales converge

Consider a martingale sequence that is uniformly integrable. This extra condition guarantees almost-sure convergence and L_1 convergence of the martingale. More strikingly, we can deduce that every UI martingale is a Lévy–Doob martingale.

Theorem 25.28 (UI martingale: Convergence). Consider a martingale sequence $(X_k : k \in \mathbb{Z}_+)$ that is uniformly integrable. Then $X_k \to X_\infty$ almost surely and in L_1 where X_∞ is an a.s. finite random variable. Furthermore,

$$X_k = \mathbb{E}[X_\infty \,|\, \mathcal{F}_k] \quad \text{almost surely for each } k \in \mathbb{Z}_+.$$

Proof. According to Exercise 25.23, the UI martingale $(X_k : k \in \mathbb{Z}_+)$ is uniformly bounded in L_1. Doob's theorem (Theorem 25.1) now implies that $X_k \to X_\infty$ almost surely. Proposition 25.24 allows us to upgrade the convergence to L_1.

Last, we must argue that the martingale has the Lévy–Doob form. Since (X_k) is a martingale,

$$\mathbb{E}[X_n \mathbb{1}_\mathsf{E}] = \mathbb{E}[X_k \mathbb{1}_\mathsf{E}] \quad \text{for all } n \geq k \text{ and all } \mathsf{E} \in \mathcal{F}_k.$$

By the conditional Jensen inequality (Proposition 20.6) and convergence $X_n \to X_\infty$ in L_1,

$$|\mathbb{E}[X_n \mathbb{1}_\mathsf{E}] - \mathbb{E}[X_\infty \mathbb{1}_\mathsf{E}]| \leq \mathbb{E}[|X_n - X_\infty| \cdot \mathbb{1}_\mathsf{E}] \to 0 \quad \text{as } n \to \infty.$$

Therefore, $\mathbb{E}[X_\infty \mathbb{1}_\mathsf{E}] = \mathbb{E}[X_k \mathbb{1}_\mathsf{E}]$ for all $\mathsf{E} \in \mathcal{F}_k$. We conclude that $X_k = \mathbb{E}[X_\infty \,|\, \mathcal{F}_k]$ by Definition 20.1 of conditional expectation. ∎

25.5.4 Lévy's upward convergence theorem

We can apply Theorem 25.28 to a Lévy–Doob martingale, but one issue remains hanging. The Lévy–Doob martingale converges, but does its limit coincide with its terminal value? The next result clarifies that the answer is positive, so UI martingales are essentially the same thing as Lévy–Doob martingales.

Corollary 25.29 (Lévy–Doob martingale: Convergence). Consider a filtration $(\mathcal{F}_k : k \in \overline{\mathbb{Z}}_+)$ where $\mathcal{F}_\infty = \sigma(\bigcup_{k=1}^\infty \mathcal{F}_k)$. For any integrable random variable Z, the Lévy–Doob martingale associated with the filtration converges:

$$X_k := \mathbb{E}[Z \,|\, \mathcal{F}_k] \to \mathbb{E}[Z \,|\, \mathcal{F}_\infty] =: X_\infty \quad \text{almost surely and in } \mathsf{L}_1.$$

Proof. Proposition 25.26 ensures that the Lévy–Doob martingale (X_k) is UI. Theorem 25.28 states that it converges: $X_k \to X$ almost surely and in L_1. Furthermore, each element of the martingale is a conditional expectation of its limit: $X_k = \mathbb{E}[X \,|\, \mathcal{F}_k]$ almost surely.

We must confirm that almost-sure equality at the finite indices implies almost-sure equality at the terminal index $k = +\infty$. We know that

$$\mathbb{E}[X \,|\, \mathcal{F}_k] = \mathbb{E}[Z \,|\, \mathcal{F}_k] \quad \text{almost surely for all } k \in \mathbb{Z}_+.$$

Let us see how this point follows from a short uniqueness of measure argument.

By splitting X and Z into positive and negative parts, we may assume that both random variables are positive. For any event $\mathsf{E} \in \mathcal{F}_\infty$, we define measures

$$\mu(\mathsf{E}) := \mathbb{E}[X \mathbb{1}_\mathsf{E}] \quad \text{and} \quad \nu(\mathsf{E}) := \mathbb{E}[Z \mathbb{1}_\mathsf{E}].$$

Suppose that $\mathsf{F} \in \mathcal{F}_k$ for some index k. Then the tower and pull-through laws guarantee that

$$\mu(\mathsf{F}) = \mathbb{E}[\mathbb{E}[X \mathbb{1}_\mathsf{F} \,|\, \mathcal{F}_k]] = \mathbb{E}[\mathbb{E}[X \,|\, \mathcal{F}_k] \cdot \mathbb{1}_\mathsf{F}] = \mathbb{E}[X_k \mathbb{1}_\mathsf{F}] = \nu(\mathsf{F}).$$

Therefore, μ and ν coincide on the intersection-stable system $\bigcup_{k=1}^\infty \mathcal{F}_k$, which generates the σ-algebra $\mathcal{F}_\infty$. Theorem E.4, on uniqueness of measures, implies that μ and ν coincide on $\mathcal{F}_\infty$.

Finally, we need to pass from events to random variables. Since $X = \limsup_{k\to\infty} X_k$ almost surely, both X and X_∞ are $\mathcal{F}_\infty$-measurable. In particular, we realize that the event $\mathsf{E} := \{X > X_\infty\} \in \mathcal{F}_\infty$, so

$$\mathbb{E}[(X - X_\infty)\mathbb{1}_{\mathsf{E}}] = \mu(\mathsf{E}) - \nu(\mathsf{E}) = 0.$$

Since expectations are almost positive operators, we deduce that $\mathbb{P}\{X > X_\infty\} = 0$. Likewise, $\mathbb{P}\{X < X_\infty\} = 0$. We conclude that $X = X_\infty$ almost surely. ∎

Problems

Exercise 25.30 (Time to ruin: Limits). Consider the simple random walk described in Exercise 24.24. Assuming $q/p \neq 1$, show that the martingale (X_k) converges almost surely to a finite limit: $X_k \to 0$ almost surely. What are the implications for the random walk S_k? Check that $\mathbb{E}[X_k] \not\to 0$.

Problem 25.31 (Martingales in L_2). Let $(X_k : k \in \mathbb{Z}_+)$ be a martingale in L_2 with difference sequence $\Delta_k := X_k - X_{k-1}$ for $k \in \mathbb{N}$. The theory of L_2 martingales is more elementary than the case of L_1 martingales.

1. Prove that the difference sequence is orthogonal: $\mathbb{E}[\Delta_j \Delta_k] = 0$ when $j \neq k$.
2. Use the Pythagorean theorem to establish that the martingale (X_k) is uniformly bounded in L_2 if and only if $\sum_{k\geq 1} \mathbb{E}[\Delta_k^2] < \infty$.
3. From now on, assume (X_k) is uniformly bounded in L_2. Deduce that (X_k) is a Cauchy sequence in L_2, so it converges in the L_2 sense to a random variable $M_\infty \in \mathsf{L}_2$.
4. Apply Doob's convergence theorem to conclude that (X_k) also converges almost surely to a random variable X_∞.
5. In particular, consider $X_k = \sum_{i=1}^k \Delta_i$, where Δ_i are centered, independent random variables in L_2. Argue that (X_k) converges almost surely and in L_2, provided that $\sum_{i=1}^\infty \operatorname{Var}[\Delta_i] < \infty$.
6. Let $\varepsilon_i \sim \textsc{unif}\{\pm 1\}$ be i.i.d. Does the random series $\sum_{i=1}^\infty \varepsilon_i / i$ converge? In what sense? **(*)** For context, does the series $\sum_{i=1}^\infty i^{-1}$ converge? What about $\sum_{i=1}^\infty (-1)^i / i$?
7. **(*)** Establish Kronecker's lemma: Let (a_k) and (b_k) be real sequences for which $0 < b_k \uparrow \infty$ and $\sum_{i=1}^\infty a_i / b_i$ converges to a finite limit. Then $b_k^{-1} \sum_{i=1}^k a_i \to 0$ as $k \to \infty$. **Hint:** Summation by parts.
8. **(*)** Consider an independent sequence (Z_k) of real random variables whose variances satisfy the bound $\sum_{i=1}^\infty i^{-2} \operatorname{Var}[Z_i] < \infty$. Use Kronecker's lemma to deduce a variant of the SLLN:
$$\frac{1}{k} \sum_{i=1}^k (Z_i - \mathbb{E}\, Z_i) \to 0 \quad \text{almost surely as } k \to \infty.$$
This result, due to Kolmogorov, is the standard SLLN for non-identically distributed sums.
9. **(**)** For an i.i.d. sum, show how to extend the argument to the case where the summands are in L_1 but not necessarily in L_2. **Hint:** Truncate the summands. The slickest way to handle the technical details is to invoke Toeplitz's lemma, a relative of Kronecker's lemma.

Applications

Application 25.32 (Likelihood ratio tests). Let $(Z_k : k \in \mathbb{N})$ of i.i.d. copies of a continuous random variable Z with strictly positive density with respect to Lebesgue measure. We

pose competing hypotheses:

$$\begin{cases} \mathsf{H}_0: & X \text{ has density } p : \mathbb{R} \to \mathbb{R}_{++}; \\ \mathsf{H}_1: & X \text{ has density } q : \mathbb{R} \to \mathbb{R}_{++}. \end{cases}$$

Assume $\lambda\{p \neq q\} > 0$. The likelihood ratio test forms the sequence of random variables

$$X_0 = 1 \quad \text{and} \quad X_{k+1} = \frac{q(Z_{k+1})}{p(Z_{k+1})} \cdot X_k \quad \text{for } k \in \mathbb{Z}_+.$$

For a level $a > 0$, we reject the null hypothesis H_0 if we observe that $X_k \geq a$ at any time k.

1. Assuming the null hypothesis H_0 is valid, show that $(X_k : k \in \mathbb{Z}_+)$ is a positive martingale.
2. Explain why the martingale converges almost surely.
3. Under the null hypothesis H_0, prove that $X_k \to 0$ almost surely. **Hint:** Take the logarithm and use the SLLN. You will need the strict case of Jensen (Exercise 9.30) as well.
4. **(*)** Using Ville's maximal inequality (Theorem 26.12), bound the probability that the test commits a Type I error, that is, the event that the test mistakenly rejects the null when it is true.

Application 25.33 (Randomized Kaczmarz). Consider a matrix $\boldsymbol{A} \in \mathbb{R}^{d \times r}$ with full column-rank. Let $\boldsymbol{a}_i^*$ denote the ith row of $\boldsymbol{A}$, and assume that $\|\boldsymbol{a}_i\|_{\ell_2} = 1$ for each index i. Let $\boldsymbol{b} \in \mathbb{R}^d$ be a vector. We wish to find the (unique) solution $\boldsymbol{x}_\star \in \mathbb{R}^d$ to the overdetermined, consistent linear system $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$.

Here, $\|\cdot\|_{\ell_2}$ is the ordinary Euclidean norm on $\mathbb{R}^r$, and * is the transpose of a vector or matrix.

Here is a simple randomized algorithm. Fix an arbitrary point $\boldsymbol{x}_0 \in \mathbb{R}^r$. For each iteration $k = 1, 2, 3, \ldots$, choose a random index $T(k) \sim \textsc{uniform}\{1, \ldots, d\}$, independent of everything. Update the current iterate $\boldsymbol{x}_k \in \mathbb{R}^r$ by enforcing equation $T(k)$ exactly:

$$\boldsymbol{x}_k = \boldsymbol{x}_{k-1} - \left(\boldsymbol{a}_{T(k)}{}^* \boldsymbol{x}_{k-1} - b_{T(k)}\right) \boldsymbol{a}_{T(k)}.$$

Define the *squared* error $E_k := \|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_{\ell_2}^2$ for each $k \in \mathbb{Z}_+$.

1. Consider the random rank-one orthogonal projector $\boldsymbol{P}_k := \boldsymbol{a}_{T(k)} \boldsymbol{a}_{T(k)}{}^*$ for $k \in \mathbb{N}$. Show that the error vectors satisfy the relation

$$\boldsymbol{x}_k - \boldsymbol{x}_\star = (\mathbf{I} - \boldsymbol{P}_k)(\boldsymbol{x}_{k-1} - \boldsymbol{x}_\star).$$

2. Deduce that the squared errors are decreasing: $E_{k+1} \leq E_k$ for each $k \in \mathbb{Z}_+$.
3. Observe that $\mathbb{E}\, \boldsymbol{P}_k = d^{-1} \boldsymbol{A}^* \boldsymbol{A}$ for each $k \in \mathbb{N}$.
4. Deduce that

$$\mathbb{E}\left[E_{k+1} \,|\, T(1), \ldots, T(k)\right] \leq (1 - \varrho)\, E_k \quad \text{for each } k \in \mathbb{Z}_+,$$

   where $0 < \varrho \leq 1$. **Hint:** The minimum singular value of $\boldsymbol{A}^* \boldsymbol{A}$ is strictly positive.
5. Show that (E_k) converges almost surely.
6. Show that (E_k) also converges in L_1. **Hint:** Use part (2).
7. Conclude that $E_k \to 0$ almost surely and in L_1.
8. Give a lower bound on the probability that $E_k \leq 10(1 - \varrho)^k E_0$ for all $k \in \mathbb{Z}_+$.
9. **(*)** Prove that the sequence $(E_k / (1-\varrho)^k)$ converges almost surely to an integrable random variable.

10. **(**)** Show that $(E_k/(1-\varrho)^k)$ converges in L_p for all $0 < p < 1$.

Application 25.34 (Stochastic gradient descent). Martingale methods provide a natural tool for studying the behavior of randomized iterative algorithms. In this problem, we will develop some convergence properties of the simplest stochastic gradient algorithm, a very important optimization method.

Let $f : \mathbb{R}^d \to \mathbb{R}$ be a differentiable convex function. The set $\mathsf{F}_\star := \arg\min\{f(\boldsymbol{x}) : \boldsymbol{x} \in \mathbb{R}^n\}$ of minimizers is always closed and convex, and we will assume that it is nonempty.

The stochastic gradient algorithm attempts to minimize f. It begins with a (random) initial point $\boldsymbol{x}_1 \in \mathbb{R}^d$. At each step $k \in \mathbb{N}$, we construct a *random* unbiased estimator $\boldsymbol{g}_k \in \mathbb{R}^d$ for the gradient at the current iterate:

$$\mathbb{E}\,\boldsymbol{g}_k = \nabla f(\boldsymbol{x}_k) \quad \text{for each } k \in \mathbb{N}.$$

Then we update the iterate using the rule

$$\boldsymbol{x}_{k+1} = \boldsymbol{x}_k - \eta_k \boldsymbol{g}_k,$$

where $\eta_k \geq 0$ is a (nonrandom) step size parameter. For simplicity, we also assume that $\|\boldsymbol{g}_k\|_2^2 \leq L$ uniformly.

The random gradient approximation $\boldsymbol{g}_k$ can depend in an arbitrary way on the observed trajectory $\boldsymbol{x}_1, \dots, \boldsymbol{x}_k$ and on the previous gradient approximations $\boldsymbol{g}_1, \dots, \boldsymbol{g}_{k-1}$ and on some auxiliary randomness. But it cannot anticipate the future trajectory.

1. Let $\boldsymbol{y} \in \mathbb{R}^d$ be an arbitrary point. For each $k \in \mathbb{N}$, prove that

$$\mathbb{E}\left[\|\boldsymbol{x}_{k+1} - \boldsymbol{y}\|_2^2 \,|\, \mathcal{F}_k\right] \leq \|\boldsymbol{x}_k - \boldsymbol{y}\|_2^2 - 2\eta_k(f(\boldsymbol{x}_k) - f(\boldsymbol{y})) + \eta_k^2 L.$$

   Hint: The gradient ∇f of the convex function f satisfies the inequality

$$f(\boldsymbol{y}) - f(\boldsymbol{x}) \geq \langle \nabla f(\boldsymbol{x}),\ \boldsymbol{y} - \boldsymbol{x} \rangle \quad \text{for all } \boldsymbol{y} \in \mathbb{R}^d.$$

2. Fix an optimal point $\boldsymbol{x}_\star \in \mathsf{F}_\star$. Define the random error in the current iterate relative to the optimal point:

$$E_k(\boldsymbol{x}_\star) := \|\boldsymbol{x}_k - \boldsymbol{x}_\star\|_2^2.$$

   Instantiate the result from (1) with $\boldsymbol{y} = \boldsymbol{x}_\star$ to relate E_{k+1} and E_k.

The result in (2) shows that the error sequence is almost—but not quite—a positive supermartingale. To establish that it converges, we need to develop a convergence theorem that addresses our situation.

3. Let (S_k) and (Y_k) be *positive* adapted sequences of random variables that satisfy

$$\mathbb{E}[S_{k+1} \,|\, \mathcal{F}_k] \leq S_k - Y_k + c_k, \quad \text{where} \quad c_i \geq 0 \text{ and } \sum\nolimits_{i=1}^{\infty} c_i < \infty.$$

   Prove that (S_k) converges almost surely to a (finite) positive random variable. Conclude that $\sum_{i=1}^{\infty} Y_i < \infty$ almost surely. **Hint:** Consider the positive supermartingale

$$T_k = S_k + \sum\nolimits_{i<k} (Y_i - c_i) + \sum\nolimits_{i=1}^{\infty} c_i.$$

4. With this result in view, identify a sequence of step size parameters that satisfy

$$\sum_{k=1}^{\infty} \eta_k = +\infty \quad \text{and} \quad \sum_{k=1}^{\infty} \eta_k^2 < +\infty.$$

5. Invoke (3) to argue that $\liminf_{k\to\infty} f(\boldsymbol{x}_k) = f(\boldsymbol{x}_\star)$ almost surely.
6. Deduce that $(\boldsymbol{x}_k)$ a.s. has a limit point in $\mathsf{F}_\star$. **Hint:** The sequence $E_k(\boldsymbol{x}_\star)$ is bounded a.s.
7. Assume that the optimal set $\mathsf{F}_\star = \{\boldsymbol{x}_\star\}$ is a singleton. In this case, demonstrate that $\boldsymbol{x}_k \to \boldsymbol{x}_\star$ a.s. As a consequence, $f(\boldsymbol{x}_k) \to f(\boldsymbol{x}_\star)$ a.s. **Hint:** Prove that $E_k(\boldsymbol{x}_\star) \to 0$ a.s.
8. **(*)** Without assuming that $\mathsf{F}_\star$ is a singleton, prove that $(\boldsymbol{x}_k)$ converges a.s. to a point in $\mathsf{F}_\star$. Conclude that $f(\boldsymbol{x}_k) \to f(\boldsymbol{x}_\star)$ a.s. **Hint:** Amplify the last argument by noting that $E_k(\boldsymbol{y})$ converges a.s. for every $\boldsymbol{y}$ in a countable dense subset of $\mathsf{F}_\star$.
9. Stochastic gradient is often applied when the function f can be written as a sum of many terms: $f(\boldsymbol{x}) = n^{-1} \sum_{i=1}^{n} f_i(\boldsymbol{x})$. Consider a random vector $\boldsymbol{g}$ with the distribution

$$\mathbb{P}\{\boldsymbol{g} = \nabla f_i(\boldsymbol{x})\} = 1/n \quad \text{for } i = 1, \dots, n.$$

   Confirm that $\boldsymbol{g}$ is an unbiased estimator for $\nabla f(\boldsymbol{x})$.
10. Explain how to apply the stochastic gradient method to solve the least-squares problem:

$$\operatorname{minimize}_{\boldsymbol{x}} \quad \frac{1}{n} \sum_{i=1}^{n} (\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle - b_i)^2.$$

11. **(*)** Explain how to apply stochastic gradient to solve the logistic regression problem:

$$\operatorname{maximize}_{\boldsymbol{x}} \quad \frac{1}{n} \sum_{i=1}^{n} \left[b_i \langle \boldsymbol{a}_i, \boldsymbol{x} \rangle - \log\left(1 + \mathrm{e}^{\langle \boldsymbol{a}_i, \boldsymbol{x} \rangle}\right) \right].$$

12. **(*)** Solve some least-squares and logistic regression problems using stochastic gradient. Plot the convergence of the objective value and the iterates.

Notes

The quotation heading this chapter is due to a Jesuit theologian named Pierre Teilhard de Chardin. In contrast, the proof of the Doob martingale convergence theorem might be summarized as "everything that rises repeatedly doesn't converge". The Teilhard de Chardin quotation has been immortalized by Flannery O'Connor in her short story "Everything That Rises Must Converge". O'Connor's collected work [O'C71] received the National Book Award in 1972.

See the introduction of Williams [Wil91] for an overview of branching processes and the connection with martingales. Grimmett & Stirzaker [GS01] also contains a treatment of branching processes.

The proof of the martingale convergence theorem and the discussion of uniformly integrable martingales are both adapted from Williams [Wil91]. Some of the problems are adapted from Grimmett & Stirzaker [GS01].

Lecture bibliography

[GS01] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. 3rd ed. Oxford University Press, 2001.

[O'C71] F. O'Connor. *The Complete Stories*. Farrar, Straus, and Giroux, 1971.

[Wil91] D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

26. Maximal Inequalities

> "La loi, dans un grand souci d'égalité, interdit aux riches comme aux pauvres de coucher sous les ponts, de mendier dans les rues et de voler du pain."
>
> "The law, in its great concern for equality, forbids the rich and the poor alike from sleeping under bridges, from begging in the streets, and from stealing bread."
>
> —Anatole France

Agenda:

1. Doob's maximal inequality
2. Submartigales and convexity
3. Hoeffding–Azuma maximal inequality
4. Uniform confidence intervals
5. *Ville's maximal inequality
6. *Adapted Hoeffding
7. *Empirical Bernstein

A concentration inequality provides bounds on the probability that a random variable takes a value far from its expectation. We have already encountered some powerful concentration inequalities for independent sums (Chebyshev, Hoeffding, Chernoff, Bernstein, and others). These results show that the probability of a large deviation has a profile similar to the probability that a normal random variable exhibits a large deviation. They stand among the most widely used probabilistic tools in computational mathematics and statistics.

A martingale is a stochastic process whose expectation is stable. We may now ask whether it is possible to control the probability that a martingale takes a value far from its expectation (at a finite time). In other words, we want to investigate concentration inequalities for martingales.

A remarkable feature of the martingale setting is that we can establish stronger results than we were able to achieve for independent sums. In particular, we can show that the *entire trajectory* of the martingale (up to a fixed time) is unlikely to deviate from its expectation. As a consequence, we can even enhance our results for independent sums.

This type of result is called a *maximal inequality*, because it controls the maximum value of the martingale over a portion of its trajectory. To establish maximal inequalities, we must exploit our toolkit of martingale methods, including martingale transforms and stopping times.

26.1 Doob's maximal inequality

We begin with the simplest maximal inequality, which is the analog of Markov's inequality (Theorem 10.13) for a submartingale. As we will see, the submartingale assumption is critical for extensions of this inequality.

Theorem 26.1 (Doob's maximal inequality). Fix a probability space and a filtration. Consider a *positive submartingale* $(X_k : k \in \mathbb{Z}_+)$. For each index $N \in \mathbb{Z}_+$,

$$\mathbb{P}\{\max_{0 \le k \le N} X_k \ge t\} \le \frac{\mathbb{E}\, X_N}{t} \quad \text{for all } t > 0.$$

If the submartingale converges to a limit X_∞ almost surely and in L_1, then

$$\mathbb{P}\left\{\sup_{k\in\mathbb{Z}_+} X_k > t\right\} \le \frac{\mathbb{E}\, X_\infty}{t} \quad \text{for all } t > 0.$$

The proof of Theorem 26.1 appears in the next subsection. First, let us take a moment to compare the result with Markov's inequality.

There are two obvious ways that we might invoke Markov's inequality to control the supremum of a positive random process. First, we can apply it to the random variable X_N to obtain

$$\mathbb{P}\{X_N \ge t\} \le \frac{\mathbb{E}\, X_N}{t} \quad \text{for all } t > 0.$$

In this case, the right-hand side of the bound matches Theorem 26.1, but we only control the tail of X_N, a single element of the process. Instead, we might apply Markov's inequality to $\sup_{k\le N} X_k$, which yields

$$\mathbb{P}\left\{\sup_{k\le N} X_k \ge t\right\} \le \frac{\mathbb{E}\sup_{k\le N} X_k}{t} \quad \text{for all } t > 0.$$

Now, the left-hand side of the bound matches Theorem 26.1, but the right-hand side involves the *expected supremum*. The latter expectation may not be easy to compute and, in the worst case, it might be as large as $N \cdot \max_{k\le N} \mathbb{E}\, X_k$.

For submartingale processes, Doob's inequality asserts that we can do better. Since the random variables in a submartingale are linked, we can control the entire trajectory (up to time N) by means of the expectation $\mathbb{E}\, X_N$ of the submartingale at the end of the time horizon. In particular, Doob's maximal inequality applies to martingale sequences, in which case $\mathbb{E}\, X_N = \mathbb{E}\, X_0$.

As a first application of Doob's inequality, we can obtain bounds for the polynomial moments of the maximum of a submartingale in terms of the moments of the submartingale. These estimates improve with the order of the moment (why?).

Problem 26.2 (Doob's maximal inequality: Moments). Fix a power $p > 1$. Consider a positive submartingale $(X_k : k \in \mathbb{Z}_+)$ that takes values in L_p. Define the sequence of partial maxima: $S_k := \max_{i\le k} X_i$. For each time horizon $N \in \mathbb{N}$, prove that

$$(\mathbb{E}\, X_N^p)^{1/p} \le (\mathbb{E}\, S_N^p)^{1/p} \le \frac{p}{p-1} \cdot (\mathbb{E}\, X_N^p)^{1/p}.$$

Hint: Use integration by parts: $\mathbb{E}\, S_N^p = \int_0^\infty \mathbb{P}\{S_N > t\} \cdot pt^{p-1}\, \mathrm{d}t$.

26.1.1 Doob's inequality: Proof

We establish Theorem 26.1 under the assumption that $N < +\infty$; you are invited to establish the limit case in Exercise 26.3. The argument involves a powerful stopping time argument that has many other applications in probability theory.

Define the extended random variable

$$\tau := \inf\{k \le N : X_k \ge t\} \in \mathbb{Z}_+ \cup \{+\infty\}.$$

In other words, τ is the first index k where the submartingale X_k surpasses the level t. If this event does not occur by the fixed time N, then we set $\tau = +\infty$. You should confirm that τ is indeed a *stopping time*.

Consider the event that the submartingale surmounts the level t on or before the time N. That is,

$$\mathsf{E} := \{\max_{k \le N} X_k \ge t\} = \{\tau \le N\}.$$

Indeed, the maximum exceeds t if and only if $X_k \ge t$ at some particular time $k \le N$. Equivalently, the stopping time is at most N.

Our task is to compute the probability of the excursion event E. We proceed backward so that we can exploit the submartingale property. Invoke the optional stopping theorem for submartingales (Exercise 24.19) to compare the expectation $\mathbb{E}[X_N]$ with the expectation $\mathbb{E}[X_{N\wedge\tau}]$ of the stopped process:

$$\mathbb{E}\, X_N \ge \mathbb{E}\, X_{N\wedge\tau} \ge \mathbb{E}[X_{N\wedge\tau} \cdot \mathbb{1}_{\mathsf{E}}] = \mathbb{E}[X_\tau \cdot \mathbb{1}_{\mathsf{E}}] \ge \mathbb{E}[t \cdot \mathbb{1}_{\mathsf{E}}] = t \cdot \mathbb{P}(\mathsf{E}).$$

Since the (stopped) process is positive, the expectation only gets smaller if we introduce an indicator random variable. On the event E, the stopping time $\tau \le N$. At the stopping time, $X_\tau \ge t$ by construction. The last relations follow from familiar properties of the expectation. This is the required result.

Problem 26.3 (Doob maximal inequality: Limiting case). Assuming that $X_k \to X_\infty$ almost surely and in L_1, extend the statement of Theorem 26.1 to the limiting case. **Hint:** This result follows from the inequality we have already established.

26.2 Submartingales and convexity

Markov's inequality is a powerful tool for obtaining other concentration inequalities because we can transform a random variable before invoking the inequality. For example, Chebyshev's inequality follows when we apply Markov's inequality to the squared deviation, and the Laplace transform method involves an application to the exponential. Doob's maximal inequality plays a similar role in the theory of martingales because convex transformations interact well with (sub)martingales.

26.2.1 Convex transformations

The next two results are very simple, and yet they have significant ramifications.

Proposition 26.4 (Martingale: Convex transformation). Consider a *martingale* sequence $(X_k : k \in \mathbb{Z}_+)$. For any convex function $\varphi : \mathbb{R} \to \mathbb{R}$, the transformed process $(\varphi(X_k) : k \in \mathbb{Z}_+)$ is a *submartingale*, provided that each $\varphi(X_k)$ is integrable.

Proof. Granted integrability, we just need to confirm the submartingale property. For any index $k \in \mathbb{Z}_+$, we find that

$$\mathbb{E}[\varphi(X_{k+1}) \,|\, \mathcal{F}_k] \ge \varphi(\mathbb{E}[X_{k+1} \,|\, \mathcal{F}_k]) = \varphi(X_k) \quad \text{almost surely.}$$

This is an immediate consequence of the conditional Jensen inequality (Proposition 20.6). Therefore, $(\varphi(X_k))$ is a submartingale. ∎

Similarly, increasing convex transformations preserve the submartingale property.

Exercise 26.5 (Submartingale: Convex transformation). Consider a *submartingale* sequence $(X_k : k \in \mathbb{Z}_+)$. For any *increasing* convex function $\varphi : \mathbb{R} \to \mathbb{R}$, show that the transformed process $(\varphi(X_k) : k \in \mathbb{Z}_+)$ is a *submartingale*, provided that each $\varphi(X_k)$ remains integrable. Show by example that the conclusion can fail if φ is not increasing.

26.2.2 Kolmogorov's inequality

Just as we obtained Chebyshev's inequality from Markov's inequality, Doob's inequality yields a maximal inequality for squared deviations.

Exercise 26.6 (Kolmogorov's maximal inequality). Consider a *martingale* sequence $(X_k : k \in \mathbb{Z}_+)$ in L_2. For each index $N \in \mathbb{Z}_+$, prove that

$$\mathbb{P}\left\{\max_{k \le N} (X_k - \mathbb{E}[X_k])^2 \ge t^2\right\} \le \frac{\operatorname{Var}[X_N]}{t^2} \quad \text{for all } t > 0.$$

Confirm that $\operatorname{Var}[X_N] = \sum_{k=1}^{N} \operatorname{Var}[\Delta_k]$, where $\Delta_k := X_k - X_{k-1}$ is the kth martingale difference. **Hint:** See Exercise 24.21.

26.2.3 The Laplace transform method

Next, we present a maximal inequality version of the Laplace transform method.

Proposition 26.7 (Exponential maximal inequality). Let $(X_k : k \in \mathbb{Z}_+)$ in L_∞ be a *bounded submartingale* sequence. For each index $N \in \mathbb{Z}_+$,

$$\mathbb{P}\{\max_{k \le N} X_k \ge t\} \le \exp\big(-\sup_{\theta > 0}(\theta t - \xi_{X_N}(\theta))\big).$$

As usual, $\xi_X(\theta) := \log \mathbb{E}\, \mathrm{e}^{\theta X}$ denotes the cgf.

This result takes some more thought to use because we need to find a way to control the cgf of an adapted process in terms of its increments. In the next subsection, we give a simple example of how this can work.

Proof. Let $\theta > 0$, since the exponential function $x \mapsto \mathrm{e}^{\theta x}$ is convex and strictly increasing,

$$\mathbb{P}\{\max_{k \le N} X_k \ge t\} = \mathbb{P}\left\{\max_{k \le N} \mathrm{e}^{\theta X_k} \ge \mathrm{e}^{\theta t}\right\} \le \mathrm{e}^{-\theta t} \cdot \mathbb{E}\, \mathrm{e}^{\theta X_N}.$$

We have invoked Exercise 26.5 to see that $(\mathrm{e}^{\theta X_k} : k \in \mathbb{Z}_+)$ is a (positive) submartingale. The inequality follows from Doob's maximal inequality (Theorem 26.1). Finally, rewrite the right-hand side in terms of the cgf, and optimize over $\theta > 0$. ∎

Exercise 26.8 (Exponential maximal inequality). Explain how we can weaken the boundedness assumption in Proposition 26.7, as we did in Theorem 16.16.

26.2.4 The Hoeffding–Azuma maximal inequality

As a particular example of the exponential maximal inequality, let us derive a very useful concentration inequality for submartingales with bounded increments. In the next section, we will explore some applications in probability and statistics; see the Problems section for some additional examples.

Theorem 26.9 (Hoeffding–Azuma maximal inequality). Consider a martingale sequence $(X_k : k \in \mathbb{Z}_+)$ whose difference sequence is bounded in the sense that

$$|\Delta_k| := |X_k - X_{k-1}| \le a_k \quad \text{almost surely for } k \in \mathbb{N}.$$

For $N \in \mathbb{N}$, define the variance proxy

$$v_N := \sum_{k=1}^{N} a_k^2.$$

Then, for all $t > 0$,

$$\mathbb{P}\{\max_{k \le N} |X_k - X_0| \ge t\} \le 2\,\mathrm{e}^{-t^2/(2v_N)}.$$

Theorem 26.9 is the maximal inequality that corresponds with Hoeffding's inequality (Theorem 16.23). The cornerstone of the argument is a cgf bound for a martingale with a bounded difference sequence, which we obtain by applying the Hoeffding cgf bound (Lemma 16.27) conditionally.

Lemma 26.10 (Hoeffding–Azuma: Cgf bound). Under the assumptions of Theorem 26.9,

$$\xi_{X_N - X_0}(\theta) \le \theta^2 v_N/2.$$

Proof. We express $X_N - X_0$ as a telescoping sum of the difference sequence:

$$m(\theta) := \mathbb{E}\,\mathrm{e}^{\theta(X_N - X_0)} = \mathbb{E}\left[\mathrm{e}^{\theta\Delta_N}\mathrm{e}^{\theta\Delta_{N-1}} \cdot \mathrm{e}^{\theta\Delta_1}\right].$$

To bound the expectation, we repeatedly condition using the tower rule:

$$\begin{aligned} m(\theta) &= \mathbb{E}\left[\,\mathbb{E}[\mathrm{e}^{\theta\Delta_N} \,|\, \mathcal{F}_{N-1}] \cdot \mathrm{e}^{\theta\Delta_{N-1}} \cdots \mathrm{e}^{\theta\Delta_1}\right] \\ &\le \mathrm{e}^{\theta^2 a_N^2/2} \cdot \mathbb{E}[\mathrm{e}^{\theta\Delta_{N-2}} \cdots \mathrm{e}^{\theta\Delta_1}] \\ &\le \cdots \\ &\le \mathrm{e}^{\theta^2(a_N^2 + a_{N-1}^2 + \cdots + a_1^2)/2} = \mathrm{e}^{\theta^2 v_N^2/2}. \end{aligned}$$

The first step follows from the pull-through rule and the fact that $X_{N-1}, \dots, X_0$ are bounded and $\mathcal{F}_{N-1}$-measurable. At each step $k = N, N-1, \dots, 1$, we have invoked the Hoeffding cgf bound (Lemma 16.27), conditional on $\mathcal{F}_{k-1}$. This action is legal because $\mathbb{E}[\Delta_k \,|\, \mathcal{F}_{k-1}] = 0$ by the martingale property and $|\Delta_k| \le a_k$ by assumption. Take the logarithm to complete the proof. ∎

We are now prepared to complete the proof of the Hoeffding–Azuma maximal inequality.

Proof of Theorem 26.9. According to Proposition 26.7 and Lemma 26.10,

$$\mathbb{P}\{\max_{k \le N}(X_N - X_0) \ge t\} \le \exp\left(-\sup_{\theta > 0}(\theta t - \theta^2 v_N/2)\right) = \mathrm{e}^{-t^2/(2v_N)}.$$

The same argument, applied to the martingale $(-X_k)$, yields a bound on the lower tail. Use the union bound to combine the two inequalities. ∎

26.3 Uniform concentration: Applications

In this section, we develop some applications of the Hoeffding–Azuma maximal inequality (Theorem 26.9) in statistics and probability. These results also support modern techniques for stochastic decision problems, such as A/B testing and stochastic bandits.

26.3.1 Uniform confidence intervals

Suppose that we observe an i.i.d sequence of *bounded* random variables, and we compute the sequence of sample average estimators:

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} Y_i \quad \text{where} \quad Y_i \text{ are i.i.d. copies of } Y \in \mathsf{L}_\infty.$$

Of course, the SLLN tells us that the sample average estimators converge to the common expectation of the input sequence: $\bar{X}_n \to \mathbb{E}\, Y$ almost surely. Can we be confident that *every individual* sample average $\bar{X}_n$ provides a reliable estimate of the expectation? More precisely, can we guarantee that

$$\mathbb{P}\left\{|\bar{X}_n - \mathbb{E}\, Y| \le c_n \text{ for all } n \in \mathbb{N}\right\} \ge 1 - \alpha,$$

where $(c_n : n \in \mathbb{N})$ is a coverage sequence and $1 - \alpha$ is the confidence level?

We can indeed achieve this goal, provided that we select the right coverage sequence. This is not an immediate consequence of Theorem 26.9 because the time horizon is infinite, while the variance proxies $v_N \to \infty$. To achieve this goal, we stitch together bounds that are valid up to dyadically increasing horizons.

Theorem 26.11 (Sample average: Uniform confidence interval). Let $Y \in \mathsf{L}_\infty$ with $\mathbb{E}\, Y = 0$ and $|Y| \le 1$. Consider a sequence $(Y_i : i \in \mathbb{N})$ of i.i.d. copies of Y, and form the sample averages $\bar{X}_k = k^{-1} \sum_{i=1}^k Y_i$. Then, for $\alpha \le 1/16$,

$$\mathbb{P}\left\{\exists k : |\bar{X}_k| \ge \sqrt{8k^{-1}(\log(1 + \log_2 k) + \log(1/\alpha))}\right\} \le \alpha.$$

Proof. Consider the partial sum process $X_k = \sum_{i=1}^k Y_i$ for $k \in \mathbb{N}$. These random variables compose a martingale with $\mathbb{E}[X_k] = 0$. The difference sequence $|X_k - X_{k-1}| = |Y_k| \le 1$ for each k, so the variance proxy v_N defined in Theorem 26.9 satisfies $v_N \le N$ for each time horizon N. It follows immediately from the Hoeffding–Azuma maximal inequality that

$$\mathbb{P}\left\{\sup_{1 \le k \le N} |X_k| \ge t\right\} \le 2\, \mathrm{e}^{-t^2/2N}.$$

We deploy this bound on each interval $2^i \le k < 2^{i+1}$ separately with an appropriate choice of t to control the overall trajectory of the partial sum process. Since the upper and lower limits of the interval are comparable, we can rescale by k to derive a uniform confidence interval for the sample average.

For $i \in \mathbb{Z}_+$, consider the dyadic interval $2^i \le k < 2^{i+1} = N_i$. By the following choice of the level t, we obtain

$$\mathbb{P}\left\{\sup_{k \le N_i} |X_k| \ge \sqrt{2N_i \log[(\log_2^2 N_i)/\alpha]}\right\} \le \frac{2\alpha}{\log_2^2(N_i)} = \frac{2\alpha}{(i+1)^2}.$$

To pass to the sample average, we restrict the supremum to the range $N_i/2 \le k < N_i$ and observe that

$$\mathbb{P}\left\{|X_k| \ge \sqrt{2N_i \log[(\log_2^2 N_i)/\alpha]} \text{ when } N_i/2 \le k < N_i\right\} \le \frac{2\alpha}{(i+1)^2}.$$

Dividing the inequality in the event through by k and noting that $N_i/k^2 \le 2/k$ and $N_i \le 2k$ on this dyadic interval,

$$\mathbb{P}\left\{|\bar{X}_k| \ge \sqrt{4k^{-1} \log[(\log_2^2(2k))/\alpha]} \text{ when } 2^i \le k < 2^{i+1}\right\} \le \frac{2\alpha}{(i+1)^2}.$$

Summing these inequalities over $i \in \mathbb{Z}_+$, we obtain a uniform bound over all choices of $k \in \mathbb{N}$.

$$\mathbb{P}\left\{|\bar{X}_k| \ge \sqrt{4k^{-1} \log[(\log_2^2(2k))/\alpha]} \text{ for some } k \in \mathbb{N}\right\} \le \sum\nolimits_{i=0}^{\infty} \frac{2\alpha}{(i+1)^2} = \frac{\pi^2 \alpha}{3}.$$

This formula implies the one stated in the theorem. ∎

26.3.2 *The law of the iterated logarithm

The calculation in Theorem 26.11 is closely related to a classic problem in probability theory. What are the extreme limits achieved by a random walk with i.i.d. increments?

Consider the partial sum $X_k := \sum_{i=1}^k Y_i$ where the Y_i are i.i.d. copies of a bounded random variable $Y \in \mathsf{L}_\infty$ with $\mathbb{E}\, Y = 0$ and $\operatorname{Var}[Y] =: \sigma^2$. Khintchine's *law of the iterated logarithm* states that

$$\begin{aligned} \liminf_{k\to\infty} \frac{X_k}{\sqrt{2\sigma^2 k \log\log k}} &= -1; \\ \limsup_{k\to\infty} \frac{X_k}{\sqrt{2\sigma^2 k \log\log k}} &= +1 \end{aligned} \qquad \text{almost surely.}$$

This is the blue envelope depicted in Figure 14.3. In other words, with probability one, a sample path $X_k(\omega)$ approaches the upper limit and lower limit an infinite number of times, but it escapes these limits only a finite number of times.

Under the assumption $|Y| \le 1$, the argument in Theorem 26.11 can be tuned and combined with the Borel–Cantelli lemma (Proposition 15.14) to establish that

$$\limsup_{k\to\infty} \frac{|X_k|}{\sqrt{2k \log\log k}} \le 1 \quad \text{almost surely.}$$

We need to use a fancier maximal inequality (analogous to the Bernstein inequality) if we wish to obtain scaling that is proportional to the standard deviation, $\operatorname{stdev}(Y) = 1$, rather than the uniform bound, $\|Y\|_{\mathsf{L}_\infty} = 1$. The complementary lower bounds (showing that the limit is achieved) require a separate argument.

26.4 Maximal inequalities for supermartingales

As we have seen, Doob's maximal inequality has many elegant corollaries and a wide range of applications. On the other hand, it does not allow us to control the trajectory of a process using information that we acquire along the way. To accomplish this goal, we need to develop maximal inequalities for *supermartingales* and learn how to apply them. This is a vast and tricky subject, so we will introduce some basic principles and establish a couple representative results.

26.4.1 Ville's maximal inequality

As usual, the fundamental result is an analog of Markov's inequality for supermartingales.

Theorem 26.12 (Ville's maximal inequality). Fix a probability space and a filtration. Consider a *positive supermartingale* $(X_k : k \in \mathbb{Z}_+)$. Then

$$\mathbb{P}\left\{\sup\nolimits_{k\in\mathbb{Z}_+} X_k > t\right\} \le \frac{\mathbb{E}\, X_0}{t} \quad \text{for all } t > 0.$$

Let us emphasize that Ville's maximal inequality controls the *entire* infinite trajectory of the supermartingale in terms of its *initial* value $\mathbb{E}\, X_0$.

Proof sketch. The argument is similar to the proof of Doob's maximal inequality (Theorem 26.1). Introduce a stopping time

$$\tau := \inf\{k \in \mathbb{Z}_+ : X_k > t\}.$$

Define the event $\mathsf{E} := \{\tau < +\infty\}$. Use the fact that a stopped supermartingale is a supermartingale (Exercise 24.16) to determine that

$$\begin{aligned}\mathbb{E}[X_0] &\geq \liminf_{k\to\infty} \mathbb{E}[X_{k\wedge\tau}] \geq \liminf_{k\to\infty} \mathbb{E}[X_{k\wedge\tau}\mathbb{1}_{\mathsf{E}}] \\ &\geq \mathbb{E}\left[\liminf_{k\to\infty} X_{k\wedge\tau}\mathbb{1}_{\mathsf{E}}\right] = \mathbb{E}[X_\tau \mathbb{1}_{\mathsf{E}}] \geq t \cdot \mathbb{P}(\mathsf{E}).\end{aligned}$$

The third inequality is Fatou's lemma (Theorem 9.11). Reinterpret this bound to obtain Ville's maximal inequality. ∎

26.4.2 *Where do we get supermartingales?

Doob's inequality is easy to deploy because convex transformations of (sub)martingale sequences produce submartingales. It takes more ingenuity to identify positive supermartingales that allow us to extract the full power of Ville's inequality.

Here is one natural construction. The most powerful concentration results are obtained by applying Markov's inequality to an exponential. We would like to mimic this approach, but the exponential of a martingale is a *sub*martingale. To fix this problem, we need to modify the exponent with a term that forces the random process to decline.

Exercise 26.13 (Cgf identity). Let $X \in \mathsf{L}_\infty$. For all $\theta \in \mathbb{R}$, confirm that

$$\mathbb{E}\exp\left(\theta X - \log \mathbb{E}\, \mathrm{e}^{\theta X}\right) = 1.$$

We can apply the cgf identity conditionally to construct a martingale sequence.

Proposition 26.14 (Martingale: Conditional cgfs). Consider an adapted sequence $(X_k : k \in \mathbb{Z}_+)$ with $X_0 = 0$ and with bounded differences $\Delta_k := X_k - X_{k-1} \in \mathsf{L}_\infty$. Fix a parameter $\theta \in \mathbb{R}$, and define the partial sum of conditional cgfs:

$$V_k(\theta) := \sum\nolimits_{i=1}^{k} \log \mathbb{E}[\mathrm{e}^{\theta\Delta_k} \,|\, \mathcal{F}_{k-1}] \quad \text{for } k \in \mathbb{N}.$$

Set $S_0(\theta) := 1$, and construct the sequence

$$S_k(\theta) := \exp\left(\theta X_k - V_k(\theta)\right) \quad \text{for } k \in \mathbb{N}.$$

Then $(S_k(\theta) : k \in \mathbb{Z}_+)$ is a positive martingale.

Proof. Positivity is clear. Note that $S_k(\theta)$ is $\mathcal{F}_k$-measurable for each $k \in \mathbb{Z}_+$. We simply need to check the status-quo condition:

$$\begin{aligned}\mathbb{E}[S_{k+1}(\theta) \,|\, \mathcal{F}_k] &= \mathbb{E}\left[\exp\left(\theta X_{k+1} - V_{k+1}(\theta)\right) \,|\, \mathcal{F}_k\right] \\ &= \mathbb{E}\left[\exp\left(\theta X_k - V_k(\theta) + \theta\Delta_{k+1} - \log \mathbb{E}[\mathrm{e}^{\theta\Delta_{k+1}} \,|\, \mathcal{F}_k]\right) \,|\, \mathcal{F}_k\right] \\ &= \mathbb{E}\left[S_k(\theta) \cdot \exp\left(\theta\Delta_{k+1} - \log \mathbb{E}[\mathrm{e}^{\theta\Delta_{k+1}} \,|\, \mathcal{F}_k]\right) \,|\, \mathcal{F}_k\right] \\ &= S_k(\theta) \cdot \mathbb{E}\left[\exp\left(\theta\Delta_{k+1} - \log \mathbb{E}[\mathrm{e}^{\theta\Delta_{k+1}} \,|\, \mathcal{F}_k]\right) \,|\, \mathcal{F}_k\right] \\ &= S_k(\theta).\end{aligned}$$

The first three relations follow from the definitions of $S_{k+1}(\theta)$, the differences Δ_{k+1}, and the conditional cgf process $V_{k+1}(\theta)$. Last, we invoke the pull-through property of the conditional expectation and the cgf identity (Exercise 26.13). ∎

In many situations, it is too much to ask that we know the exact value of the cgf of the random increments. If we replace each conditional cgf by an adapted upper bound, then we obtain a supermartingale instead of a martingale.

Exercise 26.15 (Supermartingale: Conditional cgf bounds). With the assumptions of Proposition 26.14, suppose that $(W_k(\theta) : k \in \mathbb{N})$ is an adapted process where

$$\log \mathbb{E}[\mathrm{e}^{\theta \Delta_k} \,|\, \mathcal{F}_{k-1}] \le W_k(\theta) \quad \text{for all } \theta \in \mathbb{R}.$$

Define $V_k(\theta) = \sum_{i=1}^{k} W_k(\theta)$ for $k \in \mathbb{N}$. Construct the sequence $S_0(\theta) = 0$ and

$$S_k(\theta) = \exp\big(\theta X_k - V_k(\theta)\big) \quad \text{for } k \in \mathbb{N}.$$

Check that $(S_k(\theta) : k \in \mathbb{Z}_+)$ is a positive *supermartingale*.

One approach to using the supermartingale from Exercise 26.15 is to introduce $S_k(\theta)$ into Ville's maximal inequality (Theorem 26.12) and to choose the parameter θ and the level t cleverly. As in the proof of Theorem 26.11, we can also stitch together a sequence of bounds to obtain a uniform envelope.

In contrast with the situation for independent sums and submartingales, the processes (X_k) and (V_k) are both random. As a consequence, we may not be able to optimize the parameter θ to obtain the best probability bound. One alternative is to draw the parameter θ from a suitable probability distribution and *average*.

Exercise 26.16 (Supermartingale: Pseudomaximization). For each $\theta \in \mathbb{R}$, assume that $(S_k(\theta) : k \in \mathbb{Z}_+)$ is a positive supermartingale with initial value $S_0(\theta) = 1$. For a probability distribution μ on the real line, construct the sequence

$$S_k := \int S_k(\theta)\, \mu(\mathrm{d}\theta) \quad \text{for } k \in \mathbb{Z}_+.$$

Show that $(S_k : k \in \mathbb{Z}_+)$ is a positive supermartingale with initial value $S_0 = 1$.

Yet another approach is to choose a separate value θ_k of the parameter to control each increment of the supermartingale. This leads to results with a more complicated structure.

26.4.3 *Example: Random Hoeffding–Azuma bounds

At this stage, it may not be clear that the machinery in this section leads to fruitful results. In this section, we present a simple example, which shows that we can produce uniform concentration inequalities for martingales with bounded differences, where the bounds are *random* and may depend on the trajectory of the process.

Theorem 26.17 (Hoeffding–Azuma: Adapted bounds). Consider a martingale $(X_k : k \in \mathbb{Z}_+)$ whose differences admit *random* uniform bounds A_k that are *previsible* ($\mathcal{F}_{k-1}$-measurable):

$$|\Delta_k| := |X_k - X_{k-1}| \le A_k \quad \text{almost surely for } k \in \mathbb{N}.$$

Define the random variance proxies $V_k := \sum_{i=1}^{k} A_k^2$. Then, for all $t > 0$,

$$\mathbb{P}\left\{\exists k : |X_k - X_0| > \sqrt{(1 + V_k)(\log(1 + V_k) + t^2)}\right\} \le \mathrm{e}^{-t^2/2}.$$

Theorem 26.17 offers an intriguing refinement over the Hoeffding–Azuma maximal inequality (Theorem 26.9). Instead of a sequence of uniform bounds that are fixed *a priori*, we can control the probability of a large deviation in terms of an evolving sequence of (previsible) uniform bounds. There is a modest cost for this flexibility, which is a larger logarithmic factor.

As a simple example, imagine that the martingale models your capital after playing a sequence of fair games where the winnings in each round are bounded. Both results control the probability that your capital ever differs very much from your initial capital. But Theorem 26.9 requires an *a priori* sequence of bounds on the winnings from each game. So, for example, it only applies when all your bets are bounded in advance of playing the game. In contrast, Theorem 26.17 allows random bounds that are chosen in advance of each round. For example, it reflects the prospect that you may win a significant amount and choose to make a large bet using your good fortune.

Proof. The proof of this result requires several lemmas, which we frame as exercises. Without loss of generality, we may assume that $X_0 = 0$.

Exercise 26.18 (Adapted Hoeffding: Supermartingale). Instate the notation of Theorem 26.17. For a parameter $\theta \in \mathbb{R}$, show that the sequence

$$S_k(\theta) = \exp\left(\theta X_k - (\theta^2/2)V_k\right) \quad \text{for } k \in \mathbb{Z}_+$$

composes a positive supermartingale. **Hint:** Use Exercise 26.15 and the Hoeffding cgf bound (Lemma 16.27).

Exercise 26.19 (Gaussian pseudomaximization). Consider a family of supermartingales of the form

$$S_k(\theta) = \exp\left(\theta X_k - (\theta^2/2)V_k\right) \quad \text{for } k \in \mathbb{Z}_+ \text{ and } \theta \in \mathbb{R}.$$

Check that the supermartingale (S_k) obtained by averaging with respect to the distribution $\theta \sim \textsc{normal}(0, 1)$ takes the form

$$S_k := \frac{1}{\sqrt{2\pi}} \int_{\mathbb{R}} S_k(\theta)\, \mathrm{e}^{-\theta^2/2}\, \mathrm{d}\theta = \frac{1}{\sqrt{1+V_k}} \exp\left(\frac{X_k^2}{2(1+V_k)}\right).$$

Hint: Recognize the mgf of a normal random variable with an appropriate mean and variance.

We are now prepared to establish Theorem 26.17. Apply Ville's maximal inequality to the supermartingale

$$S_k = \frac{1}{\sqrt{1+V_k}} \exp\left(\frac{X_k^2}{2(1+V_k)}\right) \quad \text{for } k \in \mathbb{N}.$$

Since $S_0 = 1$, this step yields

$$\mathbb{P}\left\{\sup\nolimits_{k\geq 0} \frac{1}{\sqrt{1+V_k}} \exp\left(\frac{X_k^2}{2(1+V_k)}\right) > \mathrm{e}^t\right\} \leq \mathrm{e}^{-t}.$$

Rearrange the event to obtain

$$\mathbb{P}\left\{\exists k : |X_k| > \sqrt{2(1+V_k)(t + \log\sqrt{1+V_k})}\right\} \leq \mathrm{e}^{-t}.$$

This is equivalent to the stated result. ∎

In comparison with the uniform confidence intervals described in Theorem 26.11, Theorem 26.17 has a logarithm instead of an iterated logarithm. If you invest some thought, you will see that it is possible to obtain the iterated logarithm bound by replacing Gaussian pseudomaximization with stitching. The cost is a more involved argument and worse constants.

Problem 26.20 (Adapted Hoeffding: Iterated logarithm). Instate the assumptions of Theorem 26.17. Show that

$$\mathbb{P}\left\{\exists k : |X_k - X_0| > \sqrt{(1+V_k)(2\log(1+\log_2(1+V_k)) + t^2)}\right\} \le 2\,\mathrm{e}^{-t^2/2}.$$

Hint: Partition the *values* of the variance proxy V_k into dyadic intervals: $2^i \le 1 + V_k < 2^{i+1}$ for $i \in \mathbb{Z}_+$.

26.4.4 *Example: Empirical martingale inequalities

The supermartingale approach to concentration also allows us to obtain concentration inequalities that depend primarily on empirical information. This kind of bound is useful because it allows us to produce uniform confidence intervals under minimal assumptions about the process generating the data. Here is one example of this phenomenon.

Theorem 26.21 (Empirical Bernstein). Consider a martingale $(X_k : k \in \mathbb{Z}_+)$ whose differences $\Delta_k := X_k - X_{k-1}$ are uniformly bounded: $|\Delta_k| \le B$ where $B > 0$ is a known constant. Introduce the quadratic variation process:

$$V_k := \sum\nolimits_{i=1}^{k} \Delta_i^2.$$

Then, for all $\alpha \le 1/9$,

$$\mathbb{P}\left\{\exists k : \frac{|X_k - X_0|}{2(B + \sqrt{V_k})} > 2\log_2(1 + \log(1 + B^{-1}\sqrt{V_k})) + \log(1/\alpha)\right\} \le 9\alpha.$$

This result measures the fluctuations $|X_k - X_0|$ of the martingale with respect to the increasing scale parameter $B + \sqrt{V_k}$. This is a natural scale for measuring the fluctuations of the martingale because it incorporates both the observed variability (V_k) and the maximum possible value (B) of the martingale differences. Bernstein's inequality (Theorem 16.30) tells us that both of these quantities are relevant to the concentration of an independent sum, so they should also arise when studying martingales.

In contrast with previous results, this bound only requires a weak piece of prior information: a uniform upper bound on the martingale differences. The scale parameter only reflects observed values of the martingale, so we can instantiate it using only the information that we acquire along the trajectory. The price we pay for this improvement is a very slowly growing function (log log) of the scale for fluctuations.

Proof. In spirit, the proof of Theorem 26.21 is similar to the argument suggested in Problem 26.20. First, we may assume that $X_0 = 0$ without loss of generality. Making the transformation $X_k \mapsto X_k/B$, we may also assume that $|X_k| \le 1$ uniformly for all k.

Exercise 26.22 (Empirical Bernstein: Cgf bound). Establish the numerical inequality

$$\exp\left(a - \frac{a^2/2}{1-|a|}\right) \le 1 + a \quad \text{for } |a| < 1.$$

Let Z be a random variable with $\mathbb{E}\,Z = 0$ and $|Z| < 1$. Deduce that

$$\mathbb{E}\exp\left(\theta Z - \frac{\theta^2 Z^2/2}{1-|\theta|}\right) \le 1 \quad \text{when } |\theta| < 1.$$

Exercise 26.23 (Empirical Bernstein: Supermartingale). Instate the notation of Theorem 26.21, and place the assumptions $X_0 = 0$ and $|X_k| \le 1$. For $|\theta| < 1$, verify the supermartingale property for the sequence

$$S_k(\theta) := \exp\left(\theta X_k - \frac{\theta^2 V_k/2}{1-|\theta|}\right) \quad \text{for } k \in \mathbb{Z}_+.$$

Hint: Modify the proof of Proposition 26.14.

For each $t > 0$, Ville's maximal inequality (Theorem 26.12) yields the uniform bound

$$\mathbb{P}\left\{\exists k : \theta X_k > \frac{\theta^2 V_k/2}{1-|\theta|} + t\right\} \le \mathrm{e}^{-t}. \tag{26.1}$$

To use this result, we need to break the trajectory into segments based on the value of the (random) intrinsic time parameter

$$T_k := 1 + \sqrt{V_k}.$$

For future reference, note that $V_k/T_k^2 < 1$.

Next, we partition the time indices into disjoint sets. For each natural number $j \in \mathbb{Z}_+$, define

$$\mathsf{I}_j^+ := \{k : 2^j \le T_k < 2^{j+1} \text{ and } X_k \ge 0\};$$
$$\mathsf{I}_j^- := \{k : 2^j \le T_k < 2^{j+1} \text{ and } X_k < 0\}.$$

We will develop a separate probability bound for each of these sets of time indices.

Fix an index $j \in \mathbb{Z}_+$. For this index, we will select a value of the parameters θ and t in (26.1) to analyze the fluctuations of the martingale on the (random) indices $k \in \mathsf{I}_j^+$. For fixed $\alpha \in (0,1)$, select

$$\theta := 2^{-(j+1)} \le 1/2 \quad \text{and} \quad t := \log((j+1)^2/\alpha).$$

From (26.1), we obtain the bound

$$\mathbb{P}\left\{\exists k : \frac{X_k}{2^{j+1}} > 2^{-2(j+1)} V_k + \log((j+1)^2/\alpha)\right\} \le \frac{\alpha}{(j+1)^2}.$$

This event addresses all time indices k, so the event certainly includes the (random) indices $k \in \mathsf{I}_j^+$. On these indices, we know that $X_k \ge 0$ and that $2^{j+1} \le 2T_k$ and that $2^{-2(j+1)} V_k \le V_k/T_k^2 < 1$. It follows that

$$\mathbb{P}\left\{\exists k \in \mathsf{I}_j^+ : \frac{|X_k|}{2T_k} > 1 + \log((\log_2(T_k)+1)^2/\alpha)\right\} \le \frac{\alpha}{(j+1)^2}.$$

A parallel argument gives the identical bound for indices $k \in \mathsf{I}_j^-$.

Finally, recall that the sets $\mathsf{I}_j^\pm$ partition $\mathbb{Z}_+$, so we may conclude that

$$\mathbb{P}\left\{\exists k : \frac{|X_k|}{2T_k} > 2\log(1+\log_2(T_k)) + \log(\mathrm{e}/\alpha)\right\} \le \sum\nolimits_{j=0}^{\infty} \frac{2\alpha}{(j+1)^2}.$$

The series on the right-hand side equals $\alpha\pi^2/3 \le 4\alpha$. Adjust constants to arrive at the stated result. ∎

Problems

Exercise 26.24 (Uniform Chebyshev–Cantelli). Using martingale methods, we can extend many classic concentration inequalities to martingales and simultaneously strengthen them to obtain uniform bounds over a portion of the trajectory.

Let $(X_k : k \in \mathbb{Z}_+)$ be a martingale with $\mathbb{E}\, X_0 = 0$ and taking values in L_2. Prove the uniform extension of the Chebyshev–Cantelli inequality (Exercise 12.23):

$$\mathbb{P}\left\{\max_{k \le n} X_k \ge t\right\} \le \frac{\mathbb{E}[X_n^2]}{\mathbb{E}[X_n^2] + t^2} \quad \text{for } t \ge 0.$$

Problem 26.25 (Freedman's inequality). We can also develop martingale inequalities that depend on the typical variability we expect to encounter at the next step. Consider a martingale sequence $(X_k : k \in \mathbb{Z}_+)$ whose differences $\Delta_k := X_k - X_{k-1}$ are uniformly bounded: $|\Delta_k| \le R$. Introduce the *predictable quadratic variation* of the martingale:

$$W_k := \sum\nolimits_{i=1}^{k} \mathbb{E}[\Delta_i^2 \,|\, \mathcal{F}_{i-1}].$$

For all $t > 0$, we will prove Freedman's inequality:

$$\mathbb{P}\left\{\exists k : |X_k| \ge t \text{ and } W_k \le \sigma^2\right\} \le \exp\left(\frac{-t^2/2}{\sigma^2 + Rt/3}\right). \tag{26.2}$$

In words, it is unlikely that the martingale achieves a large value when its predicted fluctuations are not very large.

1. Define the function
$$g(\theta) := \frac{\theta^2/2}{1 - R|\theta|} \quad \text{for } \theta \in \mathbb{R}.$$
For fixed $\theta \in \mathbb{R}$, introduce the sequence
$$S_0(\theta) := 1 \quad \text{and} \quad S_k(\theta) := \exp(\theta X_k - g(\theta) W_k) \quad \text{for } k \in \mathbb{Z}_+.$$
Prove that $S_k(\theta)$ composes a supermartingale. **Hint:** Use the Bernstein cgf bound (Lemma 16.32).
2. Introduce the stopping time
$$\tau := \inf\{k \ge 0 : |X_k| \ge t \text{ and } W_k \le \sigma^2\}.$$
As usual, we understand that $\tau = +\infty$ if there is no finite k where the conditions hold. Confirm that
$$1 \ge \mathbb{E}[S_\tau \mathbb{1}\{\tau < +\infty\}] \ge \mathrm{e}^{\theta t - g(\theta)\sigma^2} \cdot \mathbb{P}\{\tau < +\infty\}.$$
Hint: Modify the proof of Ville's maximal inequality (Theorem 26.12).
3. Derive Freedman's inequality (26.2) by choosing an appropriate value of θ.

Problem 26.26 (Bounded differences). Martingale methods lead to many useful new probability inequalities. The following result is a classic tool with a number of applications to combinatorics and algorithms.

1. For a set $\mathsf{S} \subset \mathbb{R}^n$, let $f : \mathsf{S} \to \mathbb{R}$ be a function that satisfies the bounded difference property:
$$\sup_{x_1, \dots, x_n, x_k'} |f(x_1, \dots, x_k, \dots x_n) - f(x_1, \dots, x_k', \dots x_n)| \le a_k \quad \text{for each } k.$$
This type of function often arises in combinatorial problems. Explain what the bounded difference property means in words.

2. Give an example of a function that has bounded differences.
3. Let f be a function with bounded differences, and define the variance proxy

$$v_n := \sum_{i=1}^n a_i^2.$$

Consider an independent family $(X_1, \dots, X_n)$ of real random variables taking values in S. Prove that

$$\mathbb{P}\{|f(X_1, \dots, X_n) - \mathbb{E} f(X_1, \dots, X_n)| \geq t\} \leq 2\,\mathrm{e}^{-t^2/(2v_n)}.$$

Hint: Introduce the Lévy–Doob martingale

$$M_k = \mathbb{E}[f(X_1, \dots, X_n) \,|\, \mathcal{F}_k] \quad \text{where} \quad \mathcal{F}_k = \sigma(X_1, \dots, X_k).$$

Exercise 26.27 (Bin space). Suppose that n pieces of luggage have weights $w_1, \dots, w_n \in [0, 1]$. A flight attendant wishes to place the luggage into overhead bins, but each bin holds at most one unit of weight. Let $B(w_1, \dots, w_n)$ be the minimum number of bins that suffice.

1. For random weights $W_1, \dots, W_n$, use the bounded differences inequality to control

$$\mathbb{P}\{|B(W_1, \dots, W_n) - \mathbb{E} B(W_1, \dots, W_n)| \geq t\}.$$

Explain what this result means for Thanksgiving air travel.
2. **(**)** Prove that $\mathbb{E} B(W_1, \dots, W_n)/n \to \text{const}$ as $n \to \infty$. Can we determine the constant? **Hint:** The sequence $n \mapsto B(W_1, \dots, W_n)$ is subadditive.

Exercise 26.28 (Graph coloring). A *coloring* of the graph is an assignment of colors (or labels) to the vertices so that no pair of connected vertices shares the same color. The *chromatic number* $\chi(\mathsf{G})$ of a graph G is the minimal number of colors required for a coloring.

We construct a random graph on n vertices $v_1, \dots, v_n$. For each pair $\{v_i, v_j\}$ of distinct vertices, we randomly place an edge with probability $p \in (0, 1)$. This is called an *Erdős–Rényi (ER)* graph.

1. For an ER graph G_n on n vertices, use the bounded differences inequality to control

$$\mathbb{P}\{|\chi(\mathsf{G}) - \mathbb{E} \chi(\mathsf{G})| > t\}.$$

2. **(**)** Prove that $\mathbb{E} \chi(\mathsf{G}_n)/n \to \text{const}$ as $n \to \infty$. Can we determine the constant?

Applications

Application 26.29 (A/B testing). A basic problem in applied statistics is to determine which of two alternative procedures leads to the best outcomes on average. For example, which of two experimental toothpastes instills the most personal charisma in a typical patient? Which of two website designs leads users to buy more widgets? We can administer each of the alternatives and track the outcomes to see which emerges as the favorite.

To model these problems, let $(X_i : i \in \mathbb{N})$ be i.i.d. copies of a random variable with mean m_X, taking values in $[0, 1]$. Let $(Y_i : i \in \mathbb{N})$ be i.i.d. copies of a random variable with mean m_Y, taking values in $[0, 1]$. By drawing as few samples as possible, our goal is to assess which of the means is greater. Define the gap between the means: $G := |m_X - m_Y| > 0$. For concreteness, we assume that $m_X > m_Y$, but

this information is not available to the experimenter. Fix a confidence parameter $\delta \in (0, 1/2)$.

First, suppose that the gap size G is known. In this case, we can obtain an *a priori* bound on the number $n = n(G, \delta)$ samples that suffice to determine which sequence has the larger mean, up to the specified confidence.

1. Let $\bar{X}_t$ and $\bar{Y}_t$ be the sample averages of the first t draws from each sequence. Use the (sharp) Hoeffding inequality to confirm that

$$\mathbb{P}\left\{|\bar{X}_t - m_X| > \sqrt{\log(2/\delta)/(2t)}\right\} \geq 1 - \delta.$$

   A similar bound holds for $|\bar{Y}_t - m_Y|$. Find a lower bound for n which ensures that $\bar{X}_n > \bar{Y}_n$ with probability at least $1 - 2\delta$?

In practice, we do not know the gap size G, so we cannot predict the number n of samples that are required in advance. Instead, we will maintain a uniform confidence interval for each sample mean. For example,

$$\mathbb{P}\left\{\forall t \in \mathbb{N} : |\bar{X}_t - m_X| \leq U(t, \delta)\right\} \geq 1 - \delta.$$

We can take $U(t, \delta) = \sqrt{8t^{-1} \log(\delta^{-1} \log t)}$ when $\delta \leq 1/16$.

2. Find a condition involving t and $\bar{X}_t$ and $\bar{Y}_t$ which guarantees that $m_X > m_Y$ with probability at least $1 - 2\delta$. Use this condition to obtain a lower bound on the number $n = n(G, \delta)$ of samples that suffice for your condition to hold.

In practice, for each trial, we can only draw a sample from one of the two distributions. (For example, each user only sees one of the two websites.) We want to achieve the best payoff we can, as we try to narrow our uncertainty. The key idea is to select the distribution that is most likely to have the larger mean, given our uncertainty. This approach provides a way to balance "exploration versus exploitation".

To implement this approach, we introduce a clock k that counts the number of trials. We track $T_X(k)$ and $T_Y(k)$, the number of times we have drawn a sample from X or Y in trial k. After drawing one sample from each distribution in the first step,

$$\text{Sample } X \text{ when } \quad \bar{X}_{T_X(k)} + U(T_X(k), \delta) \geq \bar{Y}_{T_Y(k)} + U(T_Y(k), \delta).$$

Otherwise, sample from Y. That is, the larger of our upper bounds signals the preferred distribution. This approach is called an "upper confidence bound" method.

3. **(*)** With high probability, show that we have identified the correct distribution when $T_X(k) \geq \text{Const} \cdot T_Y(k)$ for a sufficiently large constant, independent of everything. **Hint:** Keep in mind that a uniform concentration inequality is valid for all t, including random values T_X or T_Y. Under what circumstances do we choose to sample from the distribution Y with the smaller mean?
4. **(*)** Implement the A/B testing procedure for various choices of the gap G. Plot the evolution of $T_X(t)$ and $T_Y(t)$ as a function of t. Plot the evolution of the upper confidence bounds $\bar{X}_t + U(T_X(t), \delta)$ and $\bar{Y}_t + U(T_Y(t), \delta)$.
5. **(**)** Explain how to extend these ideas to the case where there are more than two alternatives with means $m_1 > m_2 \geq \cdots \geq m_k$. Express the required number n of samples in terms of $\Delta_k := m_1 - m_k$. Can you obtain a procedure where the number n of samples depends only on the gaps and not directly on the number k of alternatives?

Notes

Kolmogorov's maximal inequality (Exercise 26.6) is one of the earliest maximal inequalities for martingales. Many proofs are based on an explicit decomposition of the event that the martingale differs from its mean. This argument is effectively a stopping time proof, even though it is not couched in these terms. The proof of Doob's maximal inequality (Theorem 26.1) crystallizes this stopping time argument.

See Boucheron et al. [BLM13] or Alon & Spencer [AS16] for many applications of the bounded differences inequality and its relatives. The application of Doob's maximal inequality to the A/B testing problem is adapted from a paper [Jam+14] of Jamieson and coauthors.

Ville's maximal inequality (Theorem 26.12) is older than Doob's maximal inequality, as it was derived in Ville's 1939 thesis along with many other foundational properties of martingale sequences. Ville was one of the first researchers to appreciate the power and beauty of martingales, but his contributions have often been overshadowed by Doob's work.

Ville's inequality plays a central role in recent research [How+20; How+21] on uniform concentration inequalities. We have adapted the discussion of concentration for supermartingales from this body of work.

Lecture bibliography

[AS16] N. Alon and J. H. Spencer. *The probabilistic method*. Fourth. John Wiley & Sons, 2016.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. A nonasymptotic theory of independence, With a foreword by Michel Ledoux. Oxford University Press, Oxford, 2013. DOI: 10.1093/acprof:oso/9780199535255.001.0001.

[How+20] S. R. Howard et al. "Time-uniform Chernoff bounds via nonnegative supermartingales". In: *Probab. Surv.* 17 (2020), pages 257–317. DOI: 10.1214/18-PS321.

[How+21] S. R. Howard et al. "Time-uniform, nonparametric, nonasymptotic confidence sequences". In: *Ann. Statist.* 49.2 (2021), pages 1055–1080. DOI: 10.1214/20-aos1991.

[Jam+14] K. Jamieson et al. "lil' UCB : An Optimal Exploration Algorithm for Multi-Armed Bandits". In: *Proceedings of Machine Learning Research*. Volume 35. PMLR, 2014, pages 423–439. URL: http://proceedings.mlr.press/v35/jamieson14.html.

VI. appendices

A. Extension of Measures

Agenda:
1. Extension theorems
2. Set algebras
3. Hahn–Kolmogorov theorem
4. Construction of Lebesgue measure
5. Distribution functions
6. Proof of Hahn–Kolmogorov

We now return to the question that was posed at the end of Lecture 2. What information do we need to determine a measure?

For example, to specify a measure on the real line, we might like to provide only the measure of the open intervals. Is this data consistent with the values of a measure on the Borel sets? Does it determine a unique measure on Borel sets?

The abstract setting of a measurable space provides the appropriate context for discussing this problem. Here, our goal is to construct a measure by specifying its value for a small family of "elementary sets". Having done so, we need to extend this partial definition to the entire family of measurable sets. We also need to understand when the extension is unique.

This appendix discusses and proves one of the major results on measure extension, the Hahn–Kolmogorov theorem. It also shows how this theorem can be used to produce the Lebesgue measure. These tools also play a role in the construction of product measures (Lecture 6), and they provide the scaffolding that supports most of our probability models (Lecture 7).

A.1 Set algebras

As we have said, we hope to define measures on a "small family" of subsets and then extend to all measurable sets. To formalize what properties we need the "small family" to enjoy, we recall the concept of a set algebra. This definition should look familiar.

Definition A.1 (Algebra of sets). Let X be a domain. A family $\mathcal{A} \subseteq \mathcal{P}(X)$ of subsets of X is called an *algebra* on the domain X if it satisfies three properties.

1. **Nothing and everything:** The empty set $\emptyset$ and the domain X belong to $\mathcal{A}$.
2. **Complements:** If a set $A \in \mathcal{A}$, then its complement $A^c := X \setminus A$ belongs to $\mathcal{A}$.
3. **Unions and intersections:** If two sets $A, B \in \mathcal{A}$, then their union and intersection belong to $\mathcal{A}$:
$$A \cup B \in \mathcal{A} \quad \text{and} \quad A \cap B \in \mathcal{A}.$$

As before, some of the requirements in Definition A.1 are redundant (which ones?), but we have included them for clarity.

Exercise A.2 (Algebra: Stability for finite unions and intersections). Let $\mathcal{A}$ be a set algebra. Show that $\mathcal{A}$ is stable under finite unions and under finite intersections. That is,
$$A_1, \dots, A_n \in \mathcal{A} \quad \text{implies that} \quad \bigcup_{i=1}^n A_i \in \mathcal{A} \quad \text{and} \quad \bigcap_{i=1}^n A_i \in \mathcal{A}.$$

Exercise A.3 (Algebra: Stability for set differences). Let $\mathcal{A}$ be a set algebra, and let $A, B \in \mathcal{A}$. Verify that the differences $A \setminus B$ and $B \setminus A$ and $A \triangle B$ all belong to the algebra $\mathcal{A}$.

Exercise A.4 (Algebra: Intersection). Show that the intersection of an arbitrary family of algebras on X remains an algebra on X.

Obviously, every σ-algebra is also a set algebra, but the converse is not true. Indeed, although the set algebra is closed under finite unions and intersections, it need not be closed under countable unions and intersections. Here is an example.

Example A.5 (Co-finite algebra). For an (infinite) set X, consider the family

$$\mathscr{A} := \{\mathsf{A} \subseteq \mathsf{X} : \mathsf{A} \text{ is finite or } \mathsf{A}^c \text{ is finite}\}.$$

The collection $\mathscr{A}$ is an algebra on X. Since the domain X has an infinite cardinality, $\mathscr{A}$ is not a σ-algebra (why?). This algebra often arises as a counterexample to refute "intuitively obvious" statements. ∎

A.1.1 Generation of algebras

Given a collection of subsets of the domain, we can always construct a minimal algebra that contains these sets.

Definition A.6 (Set algebra: Generation). Let $\mathcal{S} \subseteq \mathcal{P}(\mathsf{X})$ be a collection of subsets of X. The family $\mathcal{S}$ generates a unique minimal algebra:

$$\operatorname{algebra}(\mathcal{S}; \mathsf{X}) := \big\{\mathsf{A} \subseteq \mathsf{X} : \mathsf{A} \text{ belongs to every algebra } \mathscr{A} \text{ on } \mathsf{X} \text{ with } \mathcal{S} \subseteq \mathscr{A}\big\}.$$

We may omit the domain X from the notation if it is clear.

In other words, $\operatorname{algebra}(\mathcal{S})$ is the intersection of all set algebras on X that contain $\mathcal{S}$. The intersection is nonempty because $\mathcal{S}$ is contained in the complete algebra $\mathcal{P}(\mathsf{X})$. The intersection remains an algebra because of Exercise A.4.

In many cases, the algebra generated by a set is small enough that we can find an explicit expression for a generic element of the algebra. In contrast, the σ-algebra generated by the set has to be stable under countably many set operations, so we may not be able to write down a formula for an element of the σ-algebra.

Example A.7 (Interval algebra). Consider the family $\mathcal{S} = \{(a, b] : a, b \in \mathbb{R}\}$ of half-open intervals of the real line. It generates $\operatorname{algebra}(\mathcal{S}; \mathbb{R})$, the *algebra of half-open intervals* of $\mathbb{R}$, consisting of all finite unions of half-open intervals and their complements. ∎

Exercise A.8 (Interval algebra). Does the algebra of half-open intervals of $\mathbb{R}$ contain closed intervals $[a, b]$ for $a, b \in \mathbb{R}$? What about open intervals and semi-infinite open intervals, such as (a, b) and $(a, +\infty)$? Describe a generic element of the algebra.

Exercise A.8 suggests that the algebra of open intervals is, somehow, very small. This is both a curse and a blessing. On the one hand, the algebra is missing a lot of sets that we might want to measure. On the other hand, it is relatively easy to describe all of the sets in the algebra and to assign them a measure. This observation is very useful for constructing measures.

A.2 The Hahn–Kolmogorov theorem

The basic idea behind measure extension is to construct an algebra of "elementary sets" that we understand well. Then we define a function, called a *premeasure*, that assigns mass to each elementary set. Afterward, we lift the premeasure to a measure defined on the full σ-algebra generated by the elementary sets. The lifting is accomplished by means of a construction called an *outer measure*.

It should be clear that the premeasure cannot be an arbitrary function on elementary sets. Its structure must be consistent with the structure of a measure. The next definition bakes in everything that we need.

Definition A.9 (Premeasure). Let X be a domain equipped with an algebra $\mathscr{A}$ of *elementary sets*. A *premeasure* is a positive function $\mu_0 : \mathscr{A} \to [0, +\infty]$ on the elementary sets that satisfies three properties.

1. **Empty set:** $\mu_0(\emptyset) = 0$;
2. **Pre-countable additivity:** For each sequence $(\mathsf{A}_i : i \in \mathbb{N})$ of *disjoint, elementary* sets $\mathsf{A}_i \in \mathscr{A}$ whose *union is an elementary set*, we have
$$\mu_0\big(\dot{\bigcup}_{i=1}^{\infty} \mathsf{A}_i\big) = \sum_{i=1}^{\infty} \mu_0(\mathsf{A}_i) \quad \text{when} \quad \dot{\bigcup}_{i=1}^{\infty} \mathsf{A}_i \in \mathscr{A}.$$
3. **Strong σ-finiteness:** The domain has a countable cover by *elementary* sets $\mathsf{A}_i \in \mathscr{A}$, each with finite premeasure:
$$\mathsf{X} = \bigcup_{i=1}^{\infty} \mathsf{A}_i \quad \text{where} \quad \mu_0(\mathsf{A}_i) < +\infty \quad \text{for each } i \in \mathbb{N}.$$

Now, we can state precisely what we mean when we say that a measure extends a premeasure.

Definition A.10 (Extension of premeasures). Let μ_0 be a premeasure defined on an algebra $\mathscr{A}$ of elementary sets. Let $\mu : \sigma(\mathscr{A}) \to [0, +\infty]$ be a measure on the σ-algebra generated by the elementary sets. We say that the measure μ *extends* the premeasure μ_0 when they coincide on elementary sets:
$$\mu(\mathsf{A}) = \mu_0(\mathsf{A}) \quad \text{for each } \mathsf{A} \in \mathscr{A}.$$

More precisely, we have defined the *minimal* extension of the premeasure.

To find a candidate for the extension of the measure, we will use the outer measure associated with the premeasure. Heuristically, the outer measure "shrink-wraps" each set by means of elementary sets.

Definition A.11 (Outer measure). Let X be a domain equipped with a set algebra $\mathscr{A}$, and let $\mu_0 : \mathscr{A} \to [0, +\infty]$ be a premeasure defined on the algebra. For each set $\mathsf{E} \in \mathscr{P}(\mathsf{X})$, the *outer measure* is defined as
$$\mu^*(\mathsf{E}) := \inf \Big\{ \sum_{i=1}^{\infty} \mu_0(\mathsf{A}_i) : \mathsf{E} \subseteq \bigcup_{i=1}^{\infty} \mathsf{A}_i \text{ and } \mathsf{A}_i \in \mathscr{A} \Big\}.$$
In other words, we cover the set E by a countable number elementary sets, compute the total premeasure of the cover, and minimize over all such covers.

With these definitions at hand, our main result on measure extension can be written succinctly.

Theorem A.12 (Hahn–Kolmogorov). Each premeasure μ_0 on an algebra $\mathscr{A}$ extends to a unique measure μ on the generated σ-algebra $\sigma(\mathscr{A})$. On this σ-algebra, the measure μ agrees with the associated outer measure μ^*.

We will give almost the entire proof of this theorem, leaving a few small pieces for the avid reader. See Section A.6.

To invoke the Hahn–Kolmogorov theorem, we typically proceed in four steps:

1. Construct an algebra $\mathscr{A}$ of elementary sets.
2. Identify the σ-algebra generated by the elementary sets, which will be the domain of the extension.
3. Construct a finitely additive function $\mu_0 : \mathscr{A} \to [0, +\infty]$ on the elementary sets.

4. Verify that μ_0 is a premeasure.

In most applications, the main challenge is to check that the function μ_0 is pre-countably additive.

> **Aside:** There are other kinds of measure extension theorems. In analysis, the Carathéodory extension theorem is prominent; it allows for the construction of Lebesgue sets. In probability, it is common to use Dynkin's theorem on intersection-stable systems, which gives uniqueness of measures with a minimum of effort. We will prove Dynkin's theorem in Appendix E.

A.3 The Lebesgue measure

Our first application of the Hahn–Kolmogorov theorem is to construct the Lebesgue measure on the real line.

The algebra of half-open intervals

Our goal is to define the Lebesgue measure on the class $\mathscr{B}(\mathbb{R})$ of Borel sets in the real line. Therefore, we begin with an algebra that generates the Borel class and where the length of a set is easy to determine. Construct the algebra of half-open intervals:

$$\mathscr{A} = \text{algebra}\big\{(a,b] : a < b \text{ and } a, b \in \mathbb{R}\big\}. \tag{A.1}$$

Referring back to Exercise 3.6, we may confirm that $\sigma(\mathscr{A})$ coincides with $\mathscr{B}(\mathbb{R})$, the family of Borel sets. A generic nonempty element $\mathsf{A} \in \mathscr{A}$ takes the form

$$\begin{aligned} \mathsf{A} &= (a_1, b_1] \cup (a_2, b_2] \cup \cdots \cup (a_n, b_n] \quad \text{or} \\ \mathsf{A} &= (a_1, b_1] \cup (a_2, b_2] \cup \cdots \cup (a_n, +\infty). \end{aligned} \tag{A.2}$$

The endpoints can be arranged so that $a_1 < b_1 < a_2 < \cdots < a_n < b_n$. We allow the possibility that $a_1 = -\infty$, but the rest of the endpoints must be finite. In this context, we understand $(-\infty, b]$ and $(a, +\infty)$ and $(-\infty, +\infty)$ to be half-open intervals for $a, b \in \mathbb{R}$.

The Lebesgue premeasure

We define a function $\lambda_0 : \mathscr{A} \to [0, +\infty]$ by specifying its value on each elementary set. If $\mathsf{A} \in \mathscr{A}$ is a bounded, then we can define

$$\lambda_0(\mathsf{A}) := \textstyle\sum_{i=1}^n |b_i - a_i| \quad \text{when} \quad \mathsf{A} = \dot{\bigcup}_{i=1}^n (a_i, b_i]. \tag{A.3}$$

Of course, we demand that $\lambda_0(\emptyset) = 0$ and that $\lambda_0(\mathsf{A}) = +\infty$ when A is unbounded. The function λ_0 is called the *Lebesgue premeasure*, in anticipation that we will verify the required properties.

Exercise A.13 (Lebesgue premeasure: Well-definition). Confirm that the definition (A.3) does not depend on the representation of the set A as a disjoint union of half-open intervals. **Hint:** Consider a representation of A using the minimum number of half-open intervals. Show that it is unique. For any other representation, argue that you can reduce the number of intervals without changing the value of λ_0.

Exercise A.14 (Lebesgue premeasure: Finite additivity and monotonicity). Show that λ_0 is *finitely* additive on *disjoint* unions of elementary sets:

$$\lambda_0\big(\dot{\bigcup}_{i=1}^n \mathsf{A}_i\big) = \textstyle\sum_{i=1}^n \lambda_0(\mathsf{A}_i) \quad \text{when } \mathsf{A}_i \in \mathscr{A} \text{ are disjoint.}$$

Deduce that λ_0 is finitely subadditive, without assuming disjointness:

$$\lambda_0\big(\bigcup_{i=1}^n \mathsf{A}_i\big) \le \sum_{i=1}^n \lambda_0(\mathsf{A}_i) \quad \text{when } \mathsf{A}_i \in \mathscr{A}.$$

Conclude that λ_0 is monotone: $\mathsf{A} \subseteq \mathsf{B}$ implies $\lambda_0(\mathsf{A}) \le \lambda_0(\mathsf{B})$ for sets $\mathsf{A}, \mathsf{B} \in \mathscr{A}$.

Exercise A.15 (Lebesgue premeasure: Strong σ-finiteness). Check that the function λ_0 is strongly σ-finite.

Pre-countable additivity

The only real difficulty is to prove that the function λ_0 is pre-countably additive on the algebra. This argument requires us to exploit the topology of the real line.

Our duty is to establish that

$$\lambda_0(\mathsf{A}) = \sum_{i=1}^\infty \lambda_0(\mathsf{A}_i) \quad \text{whenever} \quad \mathsf{A} = \dot{\bigcup}_{i=1}^\infty \mathsf{A}_i \quad \text{and } \mathsf{A}, \mathsf{A}_i \in \mathscr{A}. \tag{A.4}$$

It is easy to check that λ_0 is pre-countably *superadditive*. By monotonicity and finite additivity of λ_0 on the algebra $\mathscr{A}$,

$$\lambda_0(\mathsf{A}) = \lambda_0\big(\dot{\bigcup}_{i=1}^\infty \mathsf{A}_i\big) \ge \lambda_0\big(\dot{\bigcup}_{i=1}^k \mathsf{A}_i\big) = \sum_{i=1}^k \lambda_0(\mathsf{A}_i) \quad \text{for each } k \in \mathbb{N}.$$

Taking the (increasing) limit as $k \to \infty$, we find that

$$\lambda_0(\mathsf{A}) \ge \sum_{i=1}^\infty \lambda_0(\mathsf{A}_i). \tag{A.5}$$

If any one of the A_i is unbounded, then A is unbounded and both sides of (A.5) are infinite.

Therefore, it suffices to reverse the inequality (A.5) when the sets A_i are bounded. We need to show that

$$\lambda_0(\mathsf{A}) \le \sum_{i=1}^\infty \lambda_0(\mathsf{A}_i) \quad \text{when} \quad \mathsf{A} = \dot{\bigcup}_{i=1}^\infty \mathsf{A}_i \quad \text{and } \mathsf{A}, \mathsf{A}_i \in \mathscr{A} \text{ with } \mathsf{A}_i \text{ bounded}.$$

To simplify matters further, we can limit our attention to the situation where each set A and A_i is a half-open interval. Indeed, we may invoke the representation (A.2) of members of the algebra $\mathscr{A}$ to slice the set $\mathsf{A} = \dot{\bigcup}_{j=1}^n \mathsf{B}_j$ into a finite number of half-open intervals B_j, perhaps unbounded. By finite additivity of λ_0, it is enough to show that $\lambda_0(\mathsf{B}_j) \le \sum_{i=1}^\infty \lambda_0(\mathsf{A}_i \cap \mathsf{B}_j)$ for each index j. We can use finite additivity of λ_0 again to break down each elementary set $\mathsf{A}_i \cap \mathsf{B}_j$ into a finite intersection of bounded, half-open intervals. You may wish to write out the details.

In summary, we must obtain pre-countable subadditivity for a half-open interval that is represented as a disjoint union of finite half-open intervals. This amounts to the following pair of claims:

$$(a, b] = \dot{\bigcup}_{i=1}^\infty (a_i, b_i] \qquad \text{implies} \qquad b - a = \lambda_0(a, b] \le \sum_{i=1}^\infty \lambda_0(a_i, b_i]. \tag{A.6}$$

$$(a, +\infty) = \dot{\bigcup}_{i=1}^\infty (a_i, b_i] \qquad \text{implies} \qquad +\infty = \sum_{i=1}^\infty \lambda_0(a_i, b_i]. \tag{A.7}$$

In these expressions, we allow the left-hand endpoint $a = -\infty$, but the remaining endpoints are all finite: $b, a_i, b_i \in \mathbb{R}$.

The main challenge inheres in proving (A.6) for a finite interval ($a, b \in \mathbb{R}$). Elect a positive parameter $\varepsilon > 0$. Adjusting the endpoints of the intervals, we obtain the inclusions

Partitioning ε into dyadically decreasing parts is a very useful trick for proving things about measures.

$$[a + \varepsilon, b] \subseteq (a, b] = \dot{\bigcup}_{i=1}^\infty (a_i, b_i] \subseteq \bigcup_{i=1}^\infty (a_i, b_i + 2^{-i}\varepsilon).$$

We have constructed an open cover of the compact interval $[a+\varepsilon, b]$. By the Heine–Borel theorem, the open cover has a finite subcover. Thus, for a *finite* set $\mathsf{I} \subseteq \mathbb{N}$,

$$(a+\varepsilon, b] \subseteq [a+\varepsilon, b] \subseteq \bigcup_{i\in\mathsf{I}} (a_i, b_i + 2^{-i}\varepsilon) \subseteq \bigcup_{i\in\mathsf{I}} (a_i, b_i + 2^{-i}\varepsilon].$$

Apply the premeasure λ_0, using monotonicity and finite subadditivity:

$$\begin{aligned} b - (a+\varepsilon) = \lambda_0(a+\varepsilon, b] &\leq \lambda_0\big(\textstyle\bigcup_{i\in\mathsf{I}} (a_i, b_i + 2^{-i}\varepsilon]\big) \\ &\leq \textstyle\sum_{i\in\mathsf{I}} \big((b_i + 2^{-i}\varepsilon) - a_i\big) \leq \varepsilon + \sum_{i=1}^{\infty} (b_i - a_i) = \varepsilon + \sum_{i=1}^{\infty} \lambda_0(a_i, b_i]. \end{aligned}$$

Since ε is arbitrary, we arrive at the statement (A.6) for finite $a, b \in \mathbb{R}$.

The infinite cases follow from the finite case by exhausting the infinite interval with a sequence of finite subintervals:

- **Semi-infinite case in (A.6):** For $a = -\infty$ and $b \in \mathbb{R}$, we apply the finite case of (A.6) to the sequence $((-n, b] : n \in \mathbb{N})$ of finite half-open intervals.
- **Semi-infinite case in (A.7):** For $a \in \mathbb{R}$, we apply (A.6) to the sequence $((a, n] : n \in \mathbb{N})$ of finite half-open intervals.
- **Infinite case in (A.7):** For $a = -\infty$, we apply (A.6) to the sequence $((-n, n] : n \in \mathbb{N})$ of finite half-open intervals.

Together, these arguments establish (A.6) and (A.7) for all valid choices of a, b. We conclude that λ_0 is indeed a premeasure on the algebra $\mathcal{A}$ of half-open intervals.

The Lebesgue measure

An incantation with the Hahn–Kolmogorov theorem now yields the existence of a unique measure λ that extends the Lebesgue premeasure λ_0. Since the algebra of half-open intervals generates the Borel σ-algebra $\mathcal{B}(\mathbb{R})$, the measure $\lambda : \mathcal{B}(\mathbb{R}) \to [0, +\infty]$ assigns a well-defined value to every Borel set. On the Borel sets, we can obtain a formula for the Lebesgue measure by means of the outer measure (cf. Definition 3.15 and Definition A.11). We can interpret $\lambda(\mathsf{B})$ as the length of a Borel set B. We call λ the Lebesgue measure.

Exercise A.16 (Lebesgue measure: Properties). You may now confirm the remaining properties of the Lebesgue measure stated in Theorem 3.16.

A.4 Distributions on the real line

The strategy we used to construct the Lebesgue measure works in a more general setting. From Lecture 3, recall that a distribution function $F : \mathbb{R} \to \mathbb{R}_+$ is increasing, has asymptotic limits, and is right-continuous. Theorem 3.26 states that every distribution function induces a finite measure. We are now in a position to establish this result.

Problem A.17 (Measures from distribution functions). Prove Theorem 3.26. Let $F : \mathbb{R} \to \mathbb{R}_+$ be a distribution function; that is, F is increasing, has limits at $\pm\infty$, and is right-continuous (see Proposition 3.24). Define

$$\mu_0(a, b] := F(b) - F(a) \quad \text{for all } a, b \in \mathbb{R} \text{ with } a < b.$$

Use finite additivity and appropriate limits to extend μ_0 to the algebra $\mathcal{A}$ of half-open intervals in $\mathbb{R}$. Prove that μ_0 is a premeasure on $\mathcal{A}$. Deduce that μ_0 extends to a unique (finite) measure on the Borel sets $\mathcal{B}(\mathbb{R})$. **Hint:** The proof hews closely to the construction of the Lebesgue measure. What are the differences? How do the properties of a distribution function enter into the argument?

There is a related, but somewhat more general, construction of a measure from an increasing function.

Problem A.18 **(*Lebesgue–Stieltjes measure).** Let $F : \mathbb{R} \to \mathbb{R}$ be an increasing, right-continuous function with $F(0) = 0$. Prove that there exists a unique locally finite Borel measure μ that satisfies

$$\mu(a, b] = F(b) - F(a) \quad \text{for all } a, b \in \mathbb{R} \text{ with } a < b.$$

The converse is also true, so locally finite Borel measures are in one-to-one correspondence with the class of increasing functions that fix zero.

A measure on a topological space is *locally finite* when $\mu(\mathsf{K}) < +\infty$ for every compact set K.

> **Aside:** The Lebesgue measure is both σ-finite and locally finite. In spite of the similar terminology, locally finite measures and σ-finite measures are incomparable concepts. Neither one of these conditions on a measure implies the other in a general setting.

A.5 Approximating Borel sets

The proof of the Hahn–Kolmogorov theorem (Theorem A.12) also lends us insight into the structure of a Borel set. After digesting the proof, you can attempt the next problem.

Problem A.19 **(*Borel sets: Approximation by intervals).** In this problem, we will argue that Borel sets are well-approximated by unions of half-open intervals. Define $\mathscr{A}$ to be the algebra (A.1) of half-open intervals in $\mathbb{R}$.

1. Choose a *bounded* Borel set $\mathsf{B} \in \mathscr{B}(\mathbb{R})$. In particular, the set has finite Lebesgue measure: $\lambda(\mathsf{B}) < +\infty$. For each $\varepsilon > 0$, show that B is ε-close to a *finite* union of (disjoint) half-open intervals, in the sense that

   $$\lambda(\mathsf{B} \triangle \mathsf{A}) < \varepsilon \quad \text{where} \quad \mathsf{A} = (a_1, b_1] \dot\cup \cdots \dot\cup (a_n, b_n].$$

   As usual, the endpoints $a_i, b_i \in \mathbb{R}$. **Hint:** Each element of the algebra $\mathscr{A}$ is a finite union of half-open intervals. The statement follows directly from the definition of the σ-algebra $\mathscr{C}$ in the proof of Theorem A.12. Why is there no semi-infinite interval in A?
2. By a stronger argument, we can refine the representation to obtain

   $$\mathsf{B} = (a_1, b_1] \dot\cup \cdots \dot\cup (a_n, b_n] \dot\cup \mathsf{E} \quad \text{where } \lambda(\mathsf{E}) < \varepsilon.$$

   In this expression, E is a Borel set, and the endpoints a_i, b_i may differ from the first representation. **Hint:** Given the approximation A from above, cover the set $\mathsf{A} \setminus \mathsf{B}$ using half-open intervals, and use the construction of the Lebesgue measure as an outer measure (see Definition 3.15 or Definition A.11).
3. Let $\mathsf{B} \in \mathscr{B}(\mathbb{R})$ be a Borel set, possibly unbounded, but with finite Lebesgue measure: $\lambda(\mathsf{B}) < +\infty$. Show that we can represent B as a *finite* union of (disjoint) half-open intervals, plus a set with arbitrarily small Lebesgue measure. **Hint:** Find a compact set that contains most of B.
4. Now, consider a general Borel set B. In particular, it might have infinite Lebesgue measure: $\lambda(\mathsf{B}) = +\infty$. In this case, the situation is a little more complicated. For each $\varepsilon > 0$, show that we can approximate the set B by a *countable* union of (disjoint) half-open intervals, united with a (disjoint) set that has Lebesgue measure smaller than ε. **Hint:** Cut the set B into a countable number of bounded pieces, and apply the previous results to each piece.

5. In this context, the Lebesgue measure does not play any special role. Show that similar results hold if we replace λ by any locally finite Borel measure. See Problem A.18.

In the last problem, we saw that Borel sets can be approximated by simpler sets. In a related vein, we may ask when the measure of a set can be approximated from the inside or from the outside. For a (nice) Borel measure, both types of approximation are always possible.

Problem A.20 (*Borel measures: Regularity). The Lebesgue measure can be computed by approximating Borel sets from the inside or the outside.

1. Prove that the Lebesgue measure λ is *outer regular*. That is,
$$\lambda(\mathsf{B}) = \inf\{\lambda(\mathsf{G}) : \mathsf{B} \subseteq \mathsf{G} \text{ where } \mathsf{G} \subseteq \mathbb{R} \text{ is open}\}.$$
Hint: Use the construction of the Lebesgue measure as an outer measure (Definition 3.15), and adjust the sets in the cover so they are open.
2. Prove that the Lebesgue measure λ is *inner regular*. That is,
$$\lambda(\mathsf{B}) = \sup\{\lambda(\mathsf{K}) : \mathsf{K} \subseteq \mathsf{B} \text{ where } \mathsf{K} \subseteq \mathbb{R} \text{ is compact}\}.$$
Hint: The results in Problem A.19 can be used to this effect.
3. Show that a locally finite Borel measure $\mu : \mathscr{B}(\mathbb{R}) \to [0, +\infty]$ on the real line is both inner regular and outer regular. **Hint:** See Problem A.18 and Problem A.19.

In other words, every (locally finite) Borel measure on the real line is both inner regular and outer regular.

A.6 Hahn–Kolmogorov theorem: Proof

In this section, we establish Theorem A.12. The argument is based on the intuition that the sets in a σ-algebra should be "limits" of elementary sets.

We will prove the theorem for the special case of a finite premeasure. The extension to strongly σ-finite premeasures is left for the reader (Exercise A.22).

A.6.1 Properties of a premeasure

To begin, let us collect some basic properties of a premeasure that parallel the properties of a measure.

Exercise A.21 (Premeasure: Properties). Prove that every premeasure $\mu_0 : \mathscr{A} \to [0, +\infty]$ enjoys the following properties.

1. **Monotonicity:** $\mathsf{A} \subseteq \mathsf{B}$ implies $\mu_0(\mathsf{A}) \leq \mu_0(\mathsf{B})$ for elementary sets $\mathsf{A}, \mathsf{B} \in \mathscr{A}$.
2. **Finite additivity:** For a *finite, disjoint* family $(\mathsf{A}_1, \ldots, \mathsf{A}_n)$ of elementary sets in $\mathscr{A}$,
$$\mu_0\big(\dot{\bigcup}_{i=1}^{n} \mathsf{A}_i\big) = \textstyle\sum_{i=1}^{n} \mu_0(\mathsf{A}_i).$$
3. **Pre-countable subadditivity:** Let $\mathsf{A} \in \mathscr{A}$ be an elementary set. For a sequence $(\mathsf{A}_i : i \in \mathbb{N})$ of elementary sets in $\mathscr{A}$ whose union covers A,
$$\mu_0(\mathsf{A}) \leq \textstyle\sum_{i=1}^{\infty} \mu_0(\mathsf{A}_i) \quad \text{when} \quad \mathsf{A} \subseteq \bigcup_{i=1}^{\infty} \mathsf{A}_i \quad \text{and } \mathsf{A}, \mathsf{A}_i \in \mathscr{A}.$$
Hint: The argument is related to the proof that a measure is countably subadditive. By considering the sets $\mathsf{A}_n \setminus \bigcup_{i=1}^{n-1} \mathsf{A}_i$, we may use monotonicity to reduce to the case where A is covered by a union of *disjoint* elementary sets. Then intersect each set in the disjoint cover with A to pass to the case where A equals a disjoint union of elementary sets.

A.6.2 The outer measure

Let $\mathscr{A}$ be a set algebra on a domain X. Recall that the sets in the algebra $\mathscr{A}$ are referred to as *elementary sets*. Consider a *finite* premeasure $\mu_0 : \mathscr{A} \to [0, +\infty)$ on the algebra. That is, μ_0 assigns zero mass to the empty set, is pre-countably additive, and $\mu(\mathsf{X}) < +\infty$.

To extend a premeasure, we need to find a way to assign mass to sets that do not belong to the algebra. The approach is based on the concept of an *outer measure* (Definition A.11). For a set $\mathsf{E} \in \mathscr{P}(\mathsf{X})$, define

The word "outer" indicates that we are approximating a set by supersets.

$$\mu^*(\mathsf{E}) := \inf \left\{ \sum_{i=1}^{\infty} \mu_0(\mathsf{A}_i) : \mathsf{E} \subseteq \bigcup_{i=1}^{\infty} \mathsf{A}_i \text{ and } \mathsf{A}_i \in \mathscr{A} \right\}. \tag{A.8}$$

We emphasize that this definition applies to every subset E of the domain. You can think about this construction as "shrink wrapping" the set E. We cover E with a countable union of elementary sets, and we seek to minimize the total premeasure of the sets in the cover.

For an elementary set $\mathsf{A} \in \mathscr{A}$, the outer measure and premeasure coincide: $\mu^*(\mathsf{A}) = \mu_0(\mathsf{A})$. Indeed, an elementary set covers itself, so $\mu^*(\mathsf{A}) \le \mu_0(\mathsf{A})$. At the same time, by pre-countable subadditivity of the premeasure μ_0,

$$\mathsf{A} \subseteq \bigcup_{i=1}^{\infty} \mathsf{A}_i \text{ for } \mathsf{A}_i \in \mathscr{A} \quad \text{implies} \quad \mu_0(\mathsf{A}) \le \sum_{i=1}^{\infty} \mu_0(\mathsf{A}_i).$$

Taking the infimum of the right-hand side, $\mu_0(\mathsf{A}) \le \mu^*(\mathsf{A})$.

We remark that the outer measure is uniformly bounded: $\mu^*(\mathsf{E}) \le \mu_0(\mathsf{X}) < +\infty$ for all $\mathsf{E} \subseteq \mathsf{X}$. The construction (A.8) also ensures that the outer measure μ^* is monotone for arbitrary sets:

$$\mathsf{E} \subseteq \mathsf{F} \subseteq \mathsf{X} \quad \text{implies} \quad \mu^*(\mathsf{E}) \le \mu^*(\mathsf{F}) \tag{A.9}$$

because any countable cover of F by elementary sets is also a cover of E. Similarly, the outer measure is countably subadditive for arbitrary sets:

$$\mu^*\Big(\bigcup_{i=1}^{\infty} \mathsf{E}_i\Big) \le \sum_{i=1}^{\infty} \mu^*(\mathsf{E}_i) \quad \text{for all } \mathsf{E}_i \subseteq \mathsf{X}. \tag{A.10}$$

Indeed, given a countable cover of E_i for each $i \in \mathbb{N}$, we can combine them to obtain a countable cover for $\bigcup_{i=1}^{\infty} \mathsf{E}_i$. This cover witnesses the inequality.

Assuming the axiom of choice, the countable union of countable sets remains a countable set.

A.6.3 A pseudometric space

Next, we use the outer measure μ^* to equip the subsets of the domain with a pseudometric structure. Define a (pseudo)distance

The pseudodistance between two distinct sets can equal zero. For brevity, we usually just refer to the pseudodistance as a distance.

$$\operatorname{dist}(\mathsf{E}, \mathsf{F}) := \mu^*(\mathsf{E} \triangle \mathsf{F}) \quad \text{for } \mathsf{E}, \mathsf{F} \in \mathscr{P}(\mathsf{X}).$$

The distance is clearly nonnegative and symmetric. To verify the triangle inequality, first note that

$$\mathsf{E} \triangle \mathsf{G} \subseteq (\mathsf{E} \triangle \mathsf{F}) \cup (\mathsf{F} \triangle \mathsf{G}) \quad \text{for all } \mathsf{E}, \mathsf{F}, \mathsf{G} \in \mathscr{P}(\mathsf{X}).$$

You may want to draw a Venn diagram to persuade yourself of the set inclusion. As a consequence,

$$\begin{aligned} \operatorname{dist}(\mathsf{E}, \mathsf{G}) = \mu^*(\mathsf{E} \triangle \mathsf{G}) &\le \mu^*\big((\mathsf{E} \triangle \mathsf{F}) \cup (\mathsf{F} \triangle \mathsf{G})\big) \\ &\le \mu^*(\mathsf{E} \triangle \mathsf{F}) + \mu^*(\mathsf{F} \triangle \mathsf{G}) = \operatorname{dist}(\mathsf{E}, \mathsf{F}) + \operatorname{dist}(\mathsf{F}, \mathsf{G}). \end{aligned}$$

We have used the facts that the outer measure is monotone (A.9) and subadditive (A.10).

A.6.4 A bigger algebra

Let $\mathscr{C} := \text{closure}(\mathscr{A})$ be the closure of the algebra of elementary sets in the pseudometric space. Roughly, we will use continuity to extend the premeasure μ^* from the algebra $\mathscr{A}$ to its closure $\mathscr{C}$.

To do so, we will argue that the closure $\mathscr{C}$ remains a set algebra. Furthermore, we will establish that the outer measure μ^* is finitely additive on this algebra. These two claims are the content of this subsection.

Complements

Let us show that the family $\mathscr{C}$ is stable under complements. That is, for each set $\mathsf{E} \in \mathscr{C}$, we verify that $\mathsf{E}^c := \mathsf{X} \setminus \mathsf{E} \in \mathscr{C}$.

For each $\varepsilon > 0$, we can find $\mathsf{A} \in \mathscr{A}$ with $\operatorname{dist}(\mathsf{A}, \mathsf{E}) < \varepsilon$ because E is in the closure of $\mathscr{A}$ with respect to the pseudometric. Now,

$$\operatorname{dist}(\mathsf{A}^c, \mathsf{E}^c) = \operatorname{dist}(\mathsf{A}, \mathsf{E}) < \varepsilon.$$

Recall the set identity $\mathsf{A}^c \triangle \mathsf{E}^c = \mathsf{A} \triangle \mathsf{E}$.

The display states that we can approximate E^c by a set in the algebra $\mathscr{A}$, namely A^c. Since $\varepsilon > 0$ is arbitrary, we conclude that $\mathsf{E}^c \in \mathscr{C}$.

Unions and intersections

Next, we show that the family $\mathscr{C}$ is stable under unions. Since we have shown that $\mathscr{C}$ is also stable under complements, De Morgan's identity ensures that $\mathscr{C}$ is also stable under intersections.

Suppose that $\mathsf{E}, \mathsf{F} \in \mathscr{C}$. We must confirm that $\mathsf{E} \cup \mathsf{F} \in \mathscr{C}$. This argument depends on the set identities

$$\begin{aligned} (\mathsf{A} \cup \mathsf{B}) \triangle (\mathsf{E} \cup \mathsf{F}) &\subseteq (\mathsf{A} \triangle \mathsf{E}) \cup (\mathsf{B} \triangle \mathsf{F}) \\ (\mathsf{A} \cap \mathsf{B}) \triangle (\mathsf{E} \cap \mathsf{F}) &\subseteq (\mathsf{A} \triangle \mathsf{E}) \cup (\mathsf{B} \triangle \mathsf{F}) \end{aligned} \quad \text{for all } \mathsf{A}, \mathsf{B}, \mathsf{E}, \mathsf{F} \subseteq \mathsf{X}.$$

As always, make Venn diagrams as needed to check set relations. Alternatively, see Exercise 2.47.

For each $\varepsilon > 0$, we can find elementary sets $\mathsf{A}, \mathsf{B} \in \mathscr{A}$ for which $\operatorname{dist}(\mathsf{A}, \mathsf{E}) < \varepsilon$ and $\operatorname{dist}(\mathsf{B}, \mathsf{F}) < \varepsilon$. The outer measure μ^* is monotone (A.9) and subadditive (A.10), so

$$\begin{aligned} \operatorname{dist}(\mathsf{A} \cup \mathsf{B}, \mathsf{E} \cup \mathsf{F}) &= \mu^*((\mathsf{A} \cup \mathsf{B}) \triangle (\mathsf{E} \cup \mathsf{F})) \le \mu^*((\mathsf{A} \triangle \mathsf{E}) \cup (\mathsf{B} \triangle \mathsf{F})) \\ &\le \mu^*(\mathsf{A} \triangle \mathsf{E}) + \mu^*(\mathsf{B} \triangle \mathsf{F}) = \operatorname{dist}(\mathsf{A}, \mathsf{E}) + \operatorname{dist}(\mathsf{B}, \mathsf{F}) < 2\varepsilon. \end{aligned} \tag{A.11}$$

Since ε is arbitrary, we deduce that $\mathsf{E} \cup \mathsf{F} \in \mathscr{C}$.

Finite additivity

These arguments demonstrate that $\mathscr{C}$ is a set algebra. We are now prepared to show that the outer measure is finitely additive on the algebra $\mathscr{C}$. In other words,

$$\mu^*(\mathsf{E} \dot\cup \mathsf{F}) = \mu^*(\mathsf{E}) + \mu^*(\mathsf{F}) \quad \text{for disjoint } \mathsf{E}, \mathsf{F} \in \mathscr{C}. \tag{A.12}$$

The outer measure is subadditive (A.10), so we already know that $\mu^*(\mathsf{E} \cup \mathsf{F}) \le \mu^*(\mathsf{E}) + \mu^*(\mathsf{F})$. Our job is to prove the reverse inequality.

Choose disjoint sets $\mathsf{E}, \mathsf{F} \in \mathscr{C}$. Fix $\varepsilon > 0$, and select $\mathsf{A}, \mathsf{B} \in \mathscr{A}$ with $\operatorname{dist}(\mathsf{A}, \mathsf{E}) < \varepsilon$ and $\operatorname{dist}(\mathsf{B}, \mathsf{F}) < \varepsilon$. Using (A.11) and the triangle inequality,

$$2\varepsilon > \operatorname{dist}(\mathsf{A} \cup \mathsf{B}, \mathsf{E} \cup \mathsf{F}) \ge \operatorname{dist}(\mathsf{A} \cup \mathsf{B}, \emptyset) - \operatorname{dist}(\mathsf{E} \cup \mathsf{F}, \emptyset) = \mu_0(\mathsf{A} \cup \mathsf{B}) - \mu^*(\mathsf{E} \cup \mathsf{F}).$$

We have also applied the fact that the outer measure μ^* coincides with the premeasure μ_0 on elementary sets. By a similar argument,

$$2\varepsilon > \operatorname{dist}(\mathsf{A} \cap \mathsf{B}, \mathsf{E} \cap \mathsf{F}) \ge \operatorname{dist}(\mathsf{A} \cap \mathsf{B}, \emptyset) - \operatorname{dist}(\mathsf{E} \cap \mathsf{F}, \emptyset) = \mu_0(\mathsf{A} \cap \mathsf{B}).$$

We have used the disjointness of E and F at the last step.

Now, combine these inequalities:

$$\begin{aligned}\mu^*(\mathsf{E} \cup \mathsf{F}) &\geq \mu_0(\mathsf{A} \cup \mathsf{B}) - 2\varepsilon \\ &= \mu_0(\mathsf{A}) + \mu_0(\mathsf{B}) - \mu_0(\mathsf{A} \cap \mathsf{B}) - 2\varepsilon \geq \mu^*(\mathsf{E}) + \mu^*(\mathsf{F}) - 6\varepsilon.\end{aligned}$$

To reach the second line, we invoked the finite additivity of the premeasure μ_0. To reach the last member, we used the assumption that E, F are ε-distant from A, B. Since ε was arbitrary, we conclude that

$$\mu^*(\mathsf{E} \cup \mathsf{F}) \geq \mu^*(\mathsf{E}) + \mu^*(\mathsf{F}).$$

This is what we needed to show.

A.6.5 A σ-algebra and an extension

In this section, we complete the proof that the premeasure has at least one extension. To do so, we first demonstrate that the outer measure μ^* is countably additive on $\mathscr{C}$. Then we deduce that $\mathscr{C} = \text{closure}(\mathscr{A})$ is actually a σ-algebra.

Hahn–Kolmogorov: Existence

Granted the statements in the last paragraph, we can easily establish the existence claim in the Hahn–Kolmogorov theorem. Indeed, since $\sigma(\mathscr{A})$ is the *smallest* σ-algebra that contains the algebra $\mathscr{A}$, it must be the case that the σ-algebra $\mathscr{C}$ contains $\sigma(\mathscr{A})$. By construction, the outer measure μ^* agrees with the premeasure μ_0 on $\mathscr{A}$. Moreover, we have verified that μ^* is countably additive on $\sigma(\mathscr{A})$. Therefore, μ^* is a *measure* that extends μ_0.

Countable additivity

Let us turn to proving the two claims. Consider a disjoint sequence $(\mathsf{E}_i : i \in \mathbb{N})$ of sets in the closure $\mathscr{C}$. We intend to argue that

$$\mu^*\big(\dot{\bigcup}_{i=1}^{\infty} \mathsf{E}_i\big) = \sum_{i=1}^{\infty} \mu^*(\mathsf{E}_i) \quad \text{for disjoint } \mathsf{E}_i \in \mathscr{C}. \tag{A.13}$$

By countable subadditivity (A.10) of the outer measure, $\mu^*\big(\bigcup_{i=1}^{\infty} \mathsf{E}_i\big) \leq \sum_{i=1}^{\infty} \mu^*(\mathsf{E}_i)$. We must develop the reverse inequality.

To do so, we simply calculate that

$$\mu^*\big(\bigcup_{i=1}^{\infty} \mathsf{E}_i\big) \geq \mu^*\big(\bigcup_{i=1}^{k} \mathsf{E}_i\big) = \sum_{i=1}^{k} \mu^*(\mathsf{E}_i) \quad \text{for each } k \in \mathbb{N}.$$

Indeed, the first inequality follows from monotonocity (A.9). Since $\mathscr{C}$ is stable under finite unions, $\bigcup_{i=1}^{k} \mathsf{E}_i \in \mathscr{C}$. Therefore, it is legal to deploy the finite additivity of μ^* on $\mathscr{C}$. Taking the increasing limit as $k \to \infty$, we discover that $\mu^*(\bigcup_{i=1}^{\infty} \mathsf{E}_i) \geq \sum_{i=1}^{\infty} \mu^*(\mathsf{E})$. This is the required result.

Countable unions

Last, we must show that the algebra $\mathscr{C}$ is a σ-algebra. To accomplish this task, we only need to confirm that $\mathscr{C}$ is stable under countable unions. It suffices to consider a disjoint sequence $(\mathsf{E}_i : i \in \mathbb{N})$ of sets in $\mathscr{C}$ and its union $\mathsf{E} = \dot{\bigcup}_{i=1}^{\infty} \mathsf{E}_i$. Let us verify that the union $\mathsf{E} \in \mathscr{C}$.

Since the outer measure μ^* is countably additive (A.13) on $\mathscr{C}$ and we assumed that μ^* is finite,

$$\sum_{i=1}^{\infty} \mu^*(\mathsf{E}_i) = \mu^*(\mathsf{E}) < +\infty.$$

Next, calculate that

$$\operatorname{dist}\bigl(\dot{\bigcup}_{i=1}^{k} \mathsf{E}_i, \mathsf{E}\bigr) = \mu^*\bigl(\dot{\bigcup}_{i=k+1}^{\infty} \mathsf{E}_i\bigr) = \textstyle\sum_{i=k+1}^{\infty} \mu^*(\mathsf{E}_i) \to 0 \quad \text{as } k \to \infty.$$

Again, we rely on the countable additivity (A.13) of the outer measure μ^* on $\mathscr{C}$. The right-hand side tends to zero because it is the tail of a summable sequence. Now, for each index k, the finite union $\dot{\bigcup}_{i=1}^{k} \mathsf{E}_i$ belongs to the algebra $\mathscr{C}$. Therefore, we can approximate E by elements of the set $\mathscr{C}$. But the pseudometric space $\mathscr{C}$ is closed with respect to the pseudodistance. It follows that $\mathsf{E} \in \mathscr{C}$.

A.6.6 Uniqueness of the extension

Finally, we demonstrate that the premeasure μ_0 on the algebra $\mathscr{A}$ admits a unique extension. The key idea is to show that every measure that extends μ_0 is continuous with respect to the pseudometric.

Let μ be a measure on $\sigma(\mathscr{A})$ that agrees with μ_0 on the algebra $\mathscr{A}$. Choose a set $\mathsf{E} \in \sigma(\mathscr{A})$ in the σ-algebra generated by $\mathscr{A}$. We must prove that $\mu(\mathsf{E}) = \mu^*(\mathsf{E})$. That is, the (outer) measure μ^* is the only measure that extends μ_0 to $\sigma(\mathscr{A})$.

We have already shown that $\sigma(\mathscr{A})$ is a subset of $\mathscr{C} = \text{closure}(\mathscr{A})$. As a consequence, $\operatorname{dist}(\mathsf{A}_i, \mathsf{E}) \to 0$ for a sequence $(\mathsf{A}_i : i \in \mathbb{N})$ of elementary sets in $\mathscr{A}$. Since the measure μ and the outer measure μ^* both agree with μ_0 on elementary sets,

$$\mu(\mathsf{A}_i) = \mu_0(\mathsf{A}_i) = \mu^*(\mathsf{A}_i) \to \mu^*(\mathsf{E}) \quad \text{as } i \to \infty.$$

The last relation follows because $\operatorname{dist}(\mathsf{A}_i, \mathsf{E}) = \mu^*(\mathsf{A}_i \mathbin{\triangle} \mathsf{E}) \to 0$. Therefore, $\mu(\mathsf{A}_i)$ has an unambiguous limit.

Now, let us prove that $\mu(\mathsf{A}_i) \to \mu(\mathsf{E})$, which implies that $\mu(\mathsf{E}) = \mu^*(\mathsf{E})$. By definition (A.8) of the outer measure, for each i, we can cover the set $\mathsf{A}_i \mathbin{\triangle} \mathsf{E}$ by a countable family of elementary sets $(\mathsf{B}_{i,j} : j \in \mathbb{N})$ in $\mathscr{A}$ whose total premeasure μ_0 is arbitrarily close to the outer measure $\mu^*(\mathsf{A}_i \mathbin{\triangle} \mathsf{E})$, which tends to zero as $i \to \infty$. Whence

$$\mathsf{A}_i \mathbin{\triangle} \mathsf{E} \subseteq \textstyle\bigcup_{j=1}^{\infty} \mathsf{B}_{i,j} \quad \text{and} \quad \sum_{j=1}^{\infty} \mu_0(\mathsf{B}_{i,j}) \to 0 \quad \text{as } i \to \infty.$$

By monotonicity and countable subadditivity of the measure μ,

$$\mu(\mathsf{A}_i \mathbin{\triangle} \mathsf{E}) \le \mu\bigl(\textstyle\bigcup_{j=1}^{\infty} \mathsf{B}_{i,j}\bigr) \le \sum_{j=1}^{\infty} \mu(\mathsf{B}_{i,j}) = \sum_{j=1}^{\infty} \mu_0(\mathsf{B}_{i,j}) \to 0.$$

In the last step, we have used the assumption that the measure μ agrees with the premeasure μ_0 on elementary sets. It now follows that $\mu(\mathsf{A}_i) \to \mu(\mathsf{E})$.

A.6.7 Strongly σ-finite measures

We have established the Hahn–Kolmogorov theorem for finite premeasures. The result holds equally for strongly σ-finite premeasures.

Exercise A.22 (Extension of σ-finite premeasures). Prove the Hahn–Kolmogorov theorem for a premeasure μ_0 that is strongly σ-finite. **Hint:** Reduce to the finite case by cutting the domain X into a disjoint union of elementary sets, each with finite premeasure.

Notes

Our approach to the Hahn–Kolmogorov theorem is attributed to Marshall Stone [Sto48], with further contributions by Dorothy Maharam [Mah87]. Parts of the presentation are drawn from Terry Tao's online lecture notes on measure theory [Taoa].

The earliest proofs of measure extension results were based on lengthy set-theoretic arguments. These approaches remain important, and you will find them in many books on measure theory, real analysis, and probability.

Lecture bibliography

[Mah87] D. Maharam. “From finite to countable additivity”. In: *Portugal. Math.* 44.3 (1987), pages 265–282.

[Sto48] M. H. Stone. “Notes on integration. II”. In: *Proc. Nat. Acad. Sci. U.S.A.* 34 (1948), pages 447–455. DOI: 10.1073/pnas.34.9.447.

[Taoa] T. Tao. *An alternate approach to the Carathéodory extension theorem*. URL: https://terrytao.wordpress.com/2009/01/03/254a-notes-0a-an-alternate-approach-to-the-caratheodory-extension-theorem/.

B. Unmeasurable Sets

We spent a lot of energy motivating the definition of Borel sets and Borel measures. We have also defined the Lebesgue measure, which assigns a length to each Borel set. At the same time, we asserted that is impossible to assign a length to every subset of the real line, which is our prime reason for introducing the Borel sets in particular and measurable sets in general. In this appendix, we give substance to this claim by describing sets that cannot possibly have a well-defined length. This material will not be needed later in the course.

Agenda:

1. Lebesgue sets
2. Lebesgue vs. Borel
3. Vitali sets
4. Banach–Tarski paradox

B.1 Lebesgue sets

If you have taken a course on measure theory, you may recall that the Lebesgue measure can be extended from the Borel sets $\mathscr{B}(\mathbb{R})$ to a larger family, called the Lebesgue sets $\mathscr{L}(\mathbb{R})$. In this section, we state a characterization of the Lebesgue sets, due to Carathéodory, and we start to see why we cannot define the length of every set.

The *exterior length* of a general subset of the real line is defined as

$$\lambda^*(\mathsf{A}) := \inf\left\{ \textstyle\sum_{i=1}^{\infty} |b_i - a_i| : \mathsf{A} \subseteq \dot{\bigcup}_{i=1}^{\infty} (a_i, b_i] \right\} \quad \text{for } \mathsf{A} \subseteq \mathbb{R}. \tag{B.1}$$

In other words, the exterior length of the set is the minimum length of a cover by half-open intervals. It is easy to check that the exterior length is translation invariant, so it is a reasonable candidate for the length.

We have defined the Lebesgue measure λ as the restriction of the exterior length λ^* to the Borel sets. It is natural to ask whether we can extend the exterior length to more sets, while retaining the countable additivity property. The answer is yes.

A subset $\mathsf{E} \subseteq \mathbb{R}$ of the real line is called a *Lebesgue set* if

$$\lambda^*(\mathsf{A}) = \lambda^*(\mathsf{A} \cap \mathsf{E}) + \lambda^*(\mathsf{A} \cap \mathsf{E}^c) \quad \text{for } \textit{all} \text{ subsets } \mathsf{A} \subseteq \mathbb{R}. \tag{B.2}$$

That is, if we slice a general set A in two by a Lebesgue set E and its complement E^c, the exterior length of the set A equals the sum of the exterior lengths of the slices. This is a kind of additivity property, although it is not widely regarded as intuitive.

The family $\mathscr{L}(\mathbb{R})$ of Lebesgue sets contains all of the Borel sets from $\mathscr{B}(\mathbb{R})$, and $\mathscr{L}(\mathbb{R})$ is stable under complements and countable unions. Furthermore, the restriction of the exterior length to the Lebesgue sets is countably additive:

$$\lambda(\mathsf{E}) := \lambda^*(\mathsf{E}) \quad \text{for } \mathsf{E} \in \mathscr{L}(\mathbb{R}) \text{ is countably additive on } \mathscr{L}(\mathbb{R}).$$

In other words, we can extend the measure λ from Borel sets to a measure on Lebesgue sets, which is why we retain the same symbol.

In a moment, we will see that there are subsets of the real line that are not Lebesgue sets. For these sets, the Carathéodory criterion (B.2) fails. As a consequence, there are pairs of subsets of the real line for which

$$\lambda^*(\mathsf{A} \dot{\cup} \mathsf{B}) < \lambda^*(\mathsf{A}) + \lambda^*(\mathsf{B}) \quad \text{where } \mathsf{A}, \mathsf{B} \subseteq \mathbb{R}.$$

(The reverse inequality is not possible because the exterior length is a subadditive function.) That is, the exterior length is not even *finitely* additive on all subsets of the real line. One may wonder if we just need a more clever definition of length, but we will rule out this possibility as well.

Exercise B.1 (Exterior length). Establish the following facts about the exterior length λ^* defined in (B.1).

1. **Translation invariance:** $\lambda^*(\mathsf{A} + t) = \lambda^*(\mathsf{A})$ for each subset $\mathsf{A} \subseteq \mathbb{R}$ and each shift $t \in \mathbb{R}$.
2. **Monotonicity:** $\mathsf{A} \subseteq \mathsf{B}$ implies $\lambda^*(\mathsf{A}) \leq \lambda^*(\mathsf{B})$.
3. **Subadditivity:** $\lambda^*(\mathsf{A} \cup \mathsf{B}) \leq \lambda^*(\mathsf{A}) + \lambda^*(\mathsf{B})$.
4. **Additivity fails:** Assume that there exists a subset $\mathsf{E} \subseteq \mathbb{R}$ for which the Carathédory criterion (B.2) fails. Show that the exterior length is not finitely additive.

B.2 Lebesgue versus Borel

What is the relationship between Borel sets and Lebesgue sets? In this section, we clarify this picture by giving an alternative definition of the Lebesgue sets.

Let $\mathsf{Z} \in \mathscr{B}(\mathbb{R})$ be a Borel set that is negligible for the Lebesgue measure: $\lambda(\mathsf{Z}) = 0$. It is not always the case that a subset $\mathsf{N} \subseteq \mathsf{Z}$ is a Borel set. Nevertheless, it is reasonable to desire that the subset N be measurable, in which case monotonicity demands that $\lambda(\mathsf{N}) = 0$.

The Lebesgue sets repair this deficiency of the Borel sets. Indeed, we can define the Lebesgue sets as the *smallest* σ-algebra that contains all the Borel sets and *all subsets* of λ-negligible Borel sets. In a word, the Lebesgue sets form the *completion* of the Borel sets with respect to the Lebesgue measure.

In summary, the completion process guarantees that every subset of a negligible set is both measurable and negligible, while securing the countable additivity of the Lebesgue measure. It can be shown that the completion definition of Lebesgue sets gives the same result as the earlier definition (B.2), via the Carathéodory criterion.

There are way more Lebesgue sets than Borel sets. The family $\mathscr{B}(\mathbb{R})$ of Borel sets has the same cardinality as the real line $\mathbb{R}$. The family $\mathscr{L}(\mathbb{R})$ of Lebesgue sets has the same cardinality as the power set $\mathscr{P}(\mathbb{R})$ of the real line.

For real analysis and for certain applications of probability theory, the completeness of the Lebesgue sets plays an important role. For our purposes, the plethora of Lebesgue sets actually causes more trouble than it is worth because continuous functions interact better with Borel sets than with Lebesgue sets.

"**El Guapo:** I would not like to think that someone would tell someone else he has a plethora, and then find out that that person has no idea what it means to have a plethora.
"**Jefe:** El Guapo, I know that I, Jefe, do not have your superior intellect and education, but could it be that once again, you are angry at something else, and are looking to take it out on me?"

—*¡Three Amigos!*, 1986

Problem B.2 (Lebesgue = Carathéodory). Prove that the two definitions of Lebesgue sets are equivalent. **Hint:** Use the definition of the Lebesgue measure as the restriction of the exterior length and the property that the Lebesgue sets are complete.

B.3 Vitali sets

In 1905, several years after Lebesgue's thesis, Vitali demonstrated that there are subsets of the real line that are not Lebesgue measurable. In fact, Vitali's argument establishes a stronger statement.

Theorem B.3 (Vitali). Assume the axiom of choice. There is no (nontrivial) translation invariant, countably additive, positive function defined on all subsets of the real

A nontrivial set function cannot equal $+\infty$ on all sets.

line. In particular, not all subsets of the real line are Lebesgue measurable.

Proof. Suppose that $\mu : \mathcal{P}(\mathbb{R}) \to [0, +\infty]$ is a translation invariant, countably additive, positive function defined on all subsets of the real line. The function μ must be monotone increasing with respect to set inclusion because it is countably additive. Without loss of generality, we may impose the normalization $\mu(0, 1] = 1$.

Consider the abelian group $(\mathbb{R}, +)$ of real numbers with ordinary addition. The subset $\mathbb{Q}$ of rational numbers forms a normal subgroup. Therefore, we can construct the quotient group $\mathbb{R}/\mathbb{Q}$, which consists of the cosets

$$[r] := \{r + q : q \in \mathbb{Q}\} \quad \text{for } r \in \mathbb{R}.$$

By standard considerations, the cosets compose a partition of $\mathbb{R}$. There are uncountably many cosets, and each coset corresponds with a copy of the rational numbers. A *Vitali set* V is a subset of $(0, 1]$ that contains exactly one representative $r + q(r)$ from each coset $[r]$. We need the axiom of choice to make this selection.

Choose and fix a Vitali set V. We can use this Vitali set to lodge a contradiction against our assumptions about the function μ. Enumerate the rational numbers $q_1, q_2, q_3, \dots$ belonging to the interval $[-1, +1]$. Introduce the sets $\mathsf{V}_i := \mathsf{V} + q_i$ for $i \in \mathbb{N}$. The family $(\mathsf{V}_i : i \in \mathbb{N})$ must be disjoint because the Vitali set contains only one element of each coset. Moreover,

$$(0, 1] \subseteq \bigcup_{i=1}^{\infty} \mathsf{V}_i \subseteq (-1, 2]. \tag{B.3}$$

The upper inclusion is straightforward because $\mathsf{V} \subseteq (0, 1]$ and $|q_i| \leq 1$. For the lower inclusion, fix a real number $r \in (0, 1]$. The Vitali set $\mathsf{V} \subseteq (0, 1]$ contains exactly one representative $r + q(r)$ from the coset $[r]$, where $q(r) \in [-1, +1] \cap \mathbb{Q}$. Since $q(r) = -q_i$ for some index i, depending on r, the number $r \in \mathsf{V}_i$.

Apply the function μ to the countable union of the shifts V_i using countable additivity and then translation invariance:

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

Owing to the normalization and monotonicity of μ, the inclusions (B.3) imply that

$$1 \leq \sum_{i=1}^{\infty} \mu(\mathsf{V}) \leq 3.$$

If $\mu(\mathsf{V}) = 0$, then the central member is zero. If $\mu(\mathsf{V}) > 0$, then the central member is infinite. Either way, we have a problem. ∎

Aside: In 1970, Solovay showed that there are models of set theory, without the axiom of choice, where every subset of the real line is Lebesgue measurable. Keep in mind, however, that there remain explicit subsets of $\mathbb{R}$ that are not Borel measurable.

B.4 The Banach–Tarski "paradox"

The horrors arising from unmeasurable sets multiply when we move to settings more general than the real line. We may agree that every reasonable notion of volume in three dimensions should assign the unit ball a positive volume, it should be invariant under proper rigid motions, and it should be finitely additive. Granted this principle, we must conclude that there are subsets of $\mathbb{R}^3$ that cannot be assigned a volume.

A proper rigid motion is a translation followed by a rotation.

Introduce the three-dimensional Euclidean unit ball:

$$\mathsf{B}_3 := \{\boldsymbol{x} \in \mathbb{R}^3 : \|\boldsymbol{x}\|_{\ell_2} \leq 1\}.$$

The Banach–Tarski theorem states that the ball B_3 can be cut into a finite number of disjoint pieces (say, five) that can be reassembled via rigid motions to obtain the ball $2\mathsf{B}_3$ with twice the radius:

$$\mathsf{B}_3 = \dot{\bigcup}_{i=1}^{n} \mathsf{E}_i \quad \text{and} \quad 2\mathsf{B}_3 = \bigcup_{i=1}^{n} g_i \mathsf{E}_i,$$

where the $g_i : \mathbb{R}^3 \to \mathbb{R}^3$ are proper rigid motions. Assuming that every subset of $\mathbb{R}^3$ has a volume, we can use finite (sub)additivity and motion invariance to compute

$$\operatorname{vol}(\mathsf{B}_3) = \sum_{i=1}^{n} \operatorname{vol}(\mathsf{E}_i) = \sum_{i=1}^{n} \operatorname{vol}(g_i \mathsf{E}_i) \geq \operatorname{vol}(2\mathsf{B}_3) = 8 \operatorname{vol}(\mathsf{B}_3)$$

This is impossible because the unit ball has strictly positive volume.

By iterating the Banach–Tarski theorem, we could repeatedly dissect and reassemble a pea until it were larger than the sun. Because of this apparent absurdity, people often refer to the Banach–Tarski "paradox." But there is no paradox: we have simply established that there are regions in three-dimensional space where volume is meaningless.

"To see a World in a Grain of Sand
And a Heaven in a Wild Flower
Hold Infinity in the palm of your hand
And Eternity in an hour..."
—William Blake

> **Aside:** The Banach–Tarski theorem depends on the fact that the special Euclidean group $\mathsf{SE}(n)$ for $n \geq 3$ contains a free group with two generators. The same argument, however, does not work in lower dimensions. If you are interested in understanding these results more deeply, there are simpler settings where you can encounter similar phenomena. In particular, you may want to read about the Banach–Tarski "paradox" for the free group with two generators.

C. The Riemann–Darboux Integral

The integral of a function computes the (signed) area lying between the function and the horizontal axis. This area can be approximated by subdividing the region under the curve into very thin vertical rectangles and adding up the area of these rectangles; see Figure 4.1. This geometric intuition was already present in the work of Archimedes. Integrals found wide application in mathematics and physics after the late 17th century work of Newton and Leibniz on the calculus. Only in the 19th century were mathematicians able to place integration on a solid footing, by exploiting new-fangled concepts from analysis, such as limits.

This section describes a classic construction of an integral, commonly known as the *Riemann integral*, that was introduced in the mid-1800s. We rely on this material to develop the Lebesgue integral, so we will treat it in some detail.

Agenda:

1. Riemann vs. Darboux
2. Lower and upper sums
3. The Darboux integral
4. Properties of the integral
5. Calculus rules
6. Improper integrals
7. Monotone convergence
8. Riemann vs. Lebesgue

In Latin, the word *calculus* means "pebble." Pebbles were used for reckoning, as in an abacus. The etymological relative "chalk" is also an important part of mathematics.

C.1 Riemann versus Darboux

In the 1850s, Bernhard Riemann developed the first rigorous definition of an integral. He did so by formalizing the concept of partitioning the interval of integration into small pieces. His approximation sums the areas of rectangles that touch the function at the knots of the partition. Then one takes the limit as the partition becomes increasingly fine. The resulting object is called a *Riemann integral*.

As in most modern treatments, we present a variant of Riemann's approach, properly called a *Darboux integral*. Darboux constructed lower and upper approximations to the area under the curve by describing rectangles that rise to the minimum and maximum function value within each piece of the partition. If the lower and upper sums share a limit as the partition becomes increasingly fine, we declare the integral to equal their common value.

The Riemann and Darboux integrals are defined for the same functions, and they give identical results. Darboux's approach, however, leads to somewhat easier proofs. In the notes, we just refer to the Riemann integral, even though the term is inaccurate.

C.2 Partitions

Fix a compact subinterval $[a, b]$ of the real line. A *partition* $\boldsymbol{p} = (x_0, \dots, x_n)$ is a finite, ordered list of points in the interval:

$$a = x_0 \le x_1 \le x_2 \le \cdots \le x_n = b.$$

We allow any natural number $n \in \mathbb{N}$ here. The *width* $\|\boldsymbol{p}\|$ of the partition is the maximum separation between adjacent points:

$$\|\boldsymbol{p}\| := \max\{x_i - x_{i-1} : 1 \le i \le n\}.$$

The collection $\mathsf{P}_{a,b}$ contains all partitions of the interval $[a, b]$.

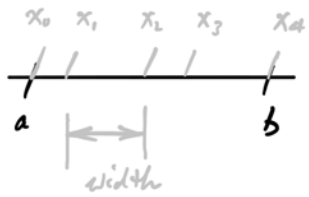


Figure C.1 (A partition).

Exercise C.1 (Narrow partitions). For each $\delta > 0$, describe a partition $\boldsymbol{p} \in \mathsf{P}_{a,b}$ whose width satisfies $\|\boldsymbol{p}\| \le \delta$.

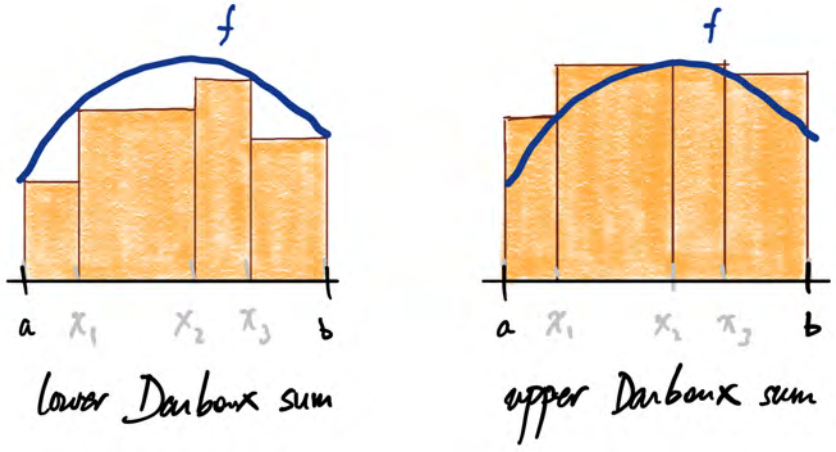


Figure C.2 (Darboux sums). In the lower (resp. upper) Darboux sum, the height of each rectangle is the infimal (resp. supremal) value of the function on each piece of the partition.

C.3 Lower and upper sums

Let $f : [a, b] \to \mathbb{R}$ be a bounded function defined on the interval $[a, b]$. Let $\boldsymbol{p}$ be a partition of $[a, b]$. The *lower* and *upper Darboux sums* are computed as

$$\begin{aligned} L(f, \boldsymbol{p}) &:= \sum_{i=1}^{n} \inf\{f(x) : x_{i-1} \le x \le x_i\} \cdot (x_i - x_{i-1}); \\ U(f, \boldsymbol{p}) &:= \sum_{i=1}^{n} \sup\{f(x) : x_{i-1} \le x \le x_i\} \cdot (x_i - x_{i-1}). \end{aligned} \tag{C.1}$$

Since f is bounded, each infimum and supremum takes a finite value, so both the sums are finite. See Figure C.2 for an illustration.

Obviously, the lower sum is smaller than the upper sum: $L(f, \boldsymbol{p}) \le U(f, \boldsymbol{p})$. We can bound the difference between the upper and lower sums:

$$\begin{aligned} &U(f, \boldsymbol{p}) - L(f, \boldsymbol{p}) \\ &\qquad \le \sum_{i=1}^{n} \sup\{f(x) - f(y) : x_{i-1} \le x, y \le x_i\} \cdot (x_i - x_{i-1}). \end{aligned} \tag{C.2}$$

This expression suggests the heuristic that the upper and lower sums are close for functions with "bounded total variation." We will not justify this statement, but an appropriate formulation is valid.

C.4 The Darboux integral

The lower and upper sums allow us to obtain a family of approximations to the signed area between the function f and the horizontal axis.

$$\begin{aligned} L_a^b(f) &:= \sup\{L(f, \boldsymbol{p}) : \boldsymbol{p} \in \mathsf{P}_{a,b}\}; \\ U_a^b(f) &:= \inf\{U(f, \boldsymbol{p}) : \boldsymbol{p} \in \mathsf{P}_{a,b}\}. \end{aligned}$$

These quantities are called *lower* and *upper Darboux integrals*. It is clear that $L_a^b(f) \le U_a^b(f)$, but they may not coincide.

The *Darboux integral* is defined when the lower and upper integrals match. Under the *assumption* that $L_a^b(f) = U_a^b(f)$,

$$\int_a^b f(x)\,\mathrm{d}x := L_a^b(f) = U_a^b(f).$$

In this case, we say that f is *Darboux integrable* on $[a, b]$.

Here is a necessary and sufficient condition for the Darboux integrability of f on $[a, b]$. For each $\varepsilon > 0$, there exists a partition $\boldsymbol{p}_\varepsilon \in \mathsf{P}_{a,b}$ for which

$$U(f, \boldsymbol{p}_\varepsilon) - L(f, \boldsymbol{p}_\varepsilon) \leq \varepsilon.$$

The bound (C.2) aids us checking this criterion.

C.5 Darboux-integrable functions

Let us establish two theorems that describe important collections of functions that are Darboux integrable. The first result has central importance for us.

Theorem C.2 (Darboux integral: Monotone function). Let $f : [a, b] \to \mathbb{R}$ be a bounded, monotone function. Then f is Darboux integrable on $[a, b]$.

Proof. Assume that f is increasing. For any partition $\boldsymbol{p}$ of $[a, b]$, the definition (C.1) of the Darboux sums implies that

$$\begin{aligned} U(f, \boldsymbol{p}) - L(f, \boldsymbol{p}) &= \sum\nolimits_{i=1}^{n} (f(x_i) - f(x_{i-1})) \cdot (x_i - x_{i-1}) \\ &\leq \left[\sum\nolimits_{i=1}^{n} (f(x_i) - f(x_{i-1}))\right] \cdot \|\boldsymbol{p}\| = [f(b) - f(a)] \cdot \|\boldsymbol{p}\|. \end{aligned}$$

Thus, we can make the difference $U(f, \boldsymbol{p}) - L(f, \boldsymbol{p})$ as small as we like by selecting a partition $\boldsymbol{p}$ of sufficiently small width. Thus f is integrable. The proof for decreasing f is similar. ∎

Theorem C.3 (Darboux integral: Continuous function). Let $f : [a, b] \to \mathbb{R}$ be continuous. Then f is Darboux integrable on $[a, b]$.

Proof. Let $\varepsilon > 0$, and define $\eta := \varepsilon/(b-a)$. On the compact set $[a, b]$, the continuous function f is uniformly continuous, so there exists $\delta > 0$ for which

$$|x - y| < \delta \quad \text{implies} \quad |f(x) - f(y)| < \eta.$$

Choose a partition $\boldsymbol{p}$ of the interval $[a, b]$ with width $\|\boldsymbol{p}\| \leq \delta$. Then

$$U(f, \boldsymbol{p}) - L(f, \boldsymbol{p}) \leq \sum\nolimits_{i=1}^{n} \eta \cdot (x_i - x_{i-1}) = (b - a) \cdot \eta = \varepsilon.$$

Therefore, f is Darboux integrable. ∎

C.6 Properties of the Darboux integral

The Darboux integral has a number of important structural properties. In particular, the integral of a positive function is a positive number, and the integral of a linear combination is the linear combination of integrals. The Darboux integral also satisfies the usual computational rules, such as the Fundamental Theorem of Calculus and the change of variables formula.

Theorem C.4 (Darboux integral: Properties). Let $f, g : [a,b] \to \mathbb{R}$ be bounded, Darboux integrable functions.

1. **Monotonicity:** If $f \leq g$, then
$$\int_a^b f(x)\,\mathrm{d}x \leq \int_a^b g(x)\,\mathrm{d}x.$$
In particular, $f \geq 0$ implies that the integral of f is positive.
2. **Linearity:** For $\alpha, \beta \in \mathbb{R}$,
$$\int_a^b (\alpha f + \beta g)(x)\,\mathrm{d}x = \alpha \int_a^b f(x)\,\mathrm{d}x + \beta \int_a^b g(x)\,\mathrm{d}x.$$
3. **Domain decomposition:** If $c \in [a,b]$, then
$$\int_a^b f(x)\,\mathrm{d}x = \int_a^c f(x)\,\mathrm{d}x + \int_c^b f(x)\,\mathrm{d}x.$$

Proof. **Monotonicity:** Suppose that $f \leq g$ pointwise. Then the definition (C.1) of the Darboux sums immediately implies that $L(f, \boldsymbol{p}) \leq L(g, \boldsymbol{p})$ for any partition $\boldsymbol{p}$. As an immediate consequence, $L_a^b(f) \leq L_a^b(g)$. Both f and g are Darboux integrable, so
$$\int_a^b f(x)\,\mathrm{d}x = L_a^b(f) \leq L_a^b(g) = \int_a^b g(x)\,\mathrm{d}x.$$
This is the required result.

Homogeneity: To verify the linearity property, we first check that the integral is homogeneous. It is easy to see that the lower and upper Darboux integrals are positively homogeneous, but negative scaling reverses the lower and upper integrals:
$$\begin{aligned} L_a^b(\alpha f) &= \alpha L_a^b(f) \quad \text{and} \quad U_a^b(\alpha f) = \alpha U_a^b(f) \quad \text{for } \alpha \geq 0; \\ L_a^b(\alpha f) &= \alpha U_a^b(f) \quad \text{and} \quad U_a^b(\alpha f) = \alpha L_a^b(f) \quad \text{for } \alpha < 0. \end{aligned}$$
When f is Darboux integrable, we may deduce that
$$\int_a^b (\alpha f)(x)\,\mathrm{d}x = \alpha \int_a^b f(x)\,\mathrm{d}x \quad \text{for } \alpha \in \mathbb{R}.$$
Therefore, the integral is homogeneous.

Additivity: Next, we show that the integral is additive. The key observation is that
$$\begin{aligned} L(f, \boldsymbol{p}) + L(g, \boldsymbol{p}) &\leq L(f+g, \boldsymbol{p}) \\ &\leq U(f+g, \boldsymbol{p}) \leq U(f, \boldsymbol{p}) + U(g, \boldsymbol{p}) \end{aligned}$$
for any partition $\boldsymbol{p} \in \mathsf{P}_{a,b}$. (Why?)

To verify that the sum $f + g$ is Darboux integrable, fix $\varepsilon > 0$. There exist partitions $\boldsymbol{p}_f$ and $\boldsymbol{p}_g$ for which
$$U(f, \boldsymbol{p}_f) - L(f, \boldsymbol{p}_f) \leq \varepsilon \quad \text{and} \quad U(g, \boldsymbol{p}_g) - L(g, \boldsymbol{p}_g) \leq \varepsilon.$$
For both functions f, g, we can reduce the error bound by passing to the common refinement $\boldsymbol{p}_{fg}$ of the partitions $\boldsymbol{p}_f$ and $\boldsymbol{p}_g$. (Formalize and check this claim!) As a

consequence,

$$U(f+g, \boldsymbol{p}_{fg}) - L(f+g, \boldsymbol{p}_{fg}) \\ \leq [U(f, \boldsymbol{p}_{fg}) - L(f, \boldsymbol{p}_{fg})] + [U(f, \boldsymbol{p}_{fg}) - L(f, \boldsymbol{p}_{fg})] \leq 2\varepsilon.$$

We determine that $f + g$ is Darboux integrable.

To compute the integral of the sum, we first bound it *above*:

$$\int_a^b (f+g)(x)\,\mathrm{d}x = L_a^b(f+g) \\ \leq U_a^b(f) + U_a^b(g) = \int_a^b f(x)\,\mathrm{d}x + \int_a^b g(x)\,\mathrm{d}x.$$

We have used the fact that all three functions $f + g, f, g$ are Darboux integrable. A similar argument shows that the integral of the sum $f + g$ is bounded *below* by the sum of the integrals of f and g. Altogether,

$$\int_a^b (f+g)(x)\,\mathrm{d}x = \int_a^b f(x)\,\mathrm{d}x + \int_a^b g(x)\,\mathrm{d}x.$$

We conclude that the integral is additive. Therefore, it is linear.

Domain decomposition: We leave the proof for the reader. You just need to consider partitions that contain the intermediate point $c \in [a, b]$. ∎

C.7 Calculus rules

In this framework, it is easy to establish some basic operational rules from calculus.

Theorem C.5 (Darboux integral: Fundamental theorem of calculus). Suppose that $f : [a, b] \to \mathbb{R}$ is Darboux integrable. Let F be an antiderivative of f; that is, $F' = f$ on (a, b). Then

$$\int_a^b f(x)\,\mathrm{d}x = F(b) - F(a).$$

Proof. Let $\boldsymbol{p} = (x_0, \dots, x_n)$ be a partition of $[a, b]$. By the mean value theorem for derivatives, we can find a point $\xi_i \in [x_{i-1}, x_i]$ in each subinterval that satisfies

$$f(\xi_i)(x_i - x_{i-1}) = F(x_i) - F(x_{i-1}) \quad \text{for } i = 1, \dots, n.$$

By definition of the Darboux sums,

$$L(f; \boldsymbol{p}) \leq \sum\nolimits_{i=1}^{n} f(\xi_i)(x_i - x_{i-1}) \leq U(f; \boldsymbol{p}).$$

But the center term equals $F(b) - F(a)$, regardless of the choice of partition. Since f is Darboux integrable, the integral must equal $F(b) - F(a)$. ∎

As a corollary, we obtain a partial result on change of variables.

Corollary C.6 (Darboux integral: Change of variables). Suppose that u is a strictly increasing, continuously differentiable function that maps $[A, B]$ onto $[a, b]$. Suppose that $f : [a, b] \to \mathbb{R}$ is continuously differentiable. Then

$$\int_A^B f'(u(x))\,u'(x)\,\mathrm{d}x = \int_a^b f'(x)\,\mathrm{d}x.$$

Proof. This is just an application of the chain rule for derivatives and the previous result. Let $F(x) = f(u(x))$, so that $F'(x) = f'(u(x))u'(x)$. Then

$$\int_A^B f'(u(x))\, u'(x)\, \mathrm{d}x = \int_A^B F'(x)\, \mathrm{d}x = F(A) - F(B)$$
$$= f(a) - f(b) = \int_a^b f'(x)\, \mathrm{d}x.$$

We have used the fact that $u(A) = a$ and $u(B) = b$. ∎

C.8 Improper integrals

It is convenient to be able to integrate functions over infinite intervals. A simple approach to this problem is to take limits, which results in an object called an *improper integral*. For our purposes, it is enough to construct the improper integral of a decreasing, positive function.

Theorem C.7 (Darboux integral: Decreasing, positive function). Let $h : \mathbb{R}_+ \to \overline{\mathbb{R}}_+$ be a decreasing, positive function. Introduce the *improper integral*

$$I(h) := \int_0^\infty h(x)\, \mathrm{d}x := \lim_{N\to\infty} \int_{1/N}^N h(x)\, \mathrm{d}x. \tag{C.3}$$

We can define this limit unambiguously, although it may take the value $+\infty$.

The resulting improper integral remains monotone. For a decreasing, positive function $g : \mathbb{R}_+ \to \overline{\mathbb{R}}_+$,

$$0 \le g \le h \quad \text{implies} \quad 0 \le I(g) \le I(h).$$

Proof. If $h(x) = +\infty$ for some $x > 0$, then we may declare the improper integral $I(h) = +\infty$.

We may suppose that $h(x) < +\infty$ for all $x > 0$. The function h is bounded and monotone decreasing on the compact interval $[1/N, N]$, where $N \in \mathbb{N}$. Therefore, h is Darboux integrable on $[1/N, N]$.

Now, by the positivity and domain decomposition properties of the Darboux integral, the function $N \mapsto \int_{1/N}^N h(x)\, \mathrm{d}x$ is increasing for $N \in \mathbb{N}$. A monotone increasing sequence has a limit, which may equal $+\infty$.

By a similar argument, the monotonicity property of the improper integral is a consequence of the monotonicity of the Darboux integral. ∎

C.9 Doubly monotone convergence

For the most part, Darboux integrals do not interact smoothly with limits. Nevertheless, there is a special situation where we can prove a satisfactory limit theorem. For an increasing sequence of monotone decreasing functions, the limit of the integrals is the integral of the limit. This result depends crucially on everything being monotone.

Theorem C.8 (Darboux integral: Doubly monotone convergence). For each $j \in \mathbb{N}$, let $h_j : \mathbb{R}_+ \to \overline{\mathbb{R}}_+$ be a decreasing, positive function. Assume that the sequence is

pointwise increasing (and thus has a limit). That is,

$$h_j(x) \uparrow h(x) \quad \text{for each } x \in \mathbb{R}_+.$$

Then the sequence of improper integrals (C.3) increases to its limiting value:

$$I(h_j) \uparrow I(h).$$

Proof. The limit h must also be decreasing and positive (why?), so its improper integral $I(h)$ is defined. The improper integral is monotone, so the sequence $I(h_j)$ is increasing and has a limit. We must show that the limit coincides with $I(h)$.

Everything is monotone increasing, so we can rewrite limits as suprema and interchange them:

$$\lim_{j\to\infty} I(h_j) = \sup_{j\in\mathbb{N}} \sup_{N\in\mathbb{N}} \int_{1/N}^{N} h_j(x)\,\mathrm{d}x = \sup_{N\in\mathbb{N}} \sup_{j\in\mathbb{N}} \int_{1/N}^{N} h_j(x)\,\mathrm{d}x.$$

We *claim* that

$$\sup_{j\in\mathbb{N}} \int_{1/N}^{N} h_j(x)\,\mathrm{d}x = \int_{1/N}^{N} h(x)\,\mathrm{d}x. \tag{C.4}$$

Owing to the definition (C.3) of the improper integral, we may then conclude that $\lim_{j\to\infty} I(h_j) = I(h)$.

To prove the claim, fix $N \in \mathbb{N}$ and a partition $\boldsymbol{p}$ of the interval $[1/N, N]$. In view of the facts that each h_j is decreasing and that $h = \sup_j h_j$,

$$\begin{aligned}\sup_j L(h_j, \boldsymbol{p}) &= \sup_j \sum\nolimits_{i=1}^{n} h_j(x_i)(x_i - x_{i-1}) \\ &= \sum\nolimits_{i=1}^{n} h(x_i)(x_i - x_{i-1}) = L(h, \boldsymbol{p}).\end{aligned}$$

Therefore, the lower Darboux integral satisfies

$$\sup_j L_{1/N}^{N}(h_j) = \sup_j \sup_{\boldsymbol{p}} L(h_j, \boldsymbol{p}) = \sup_{\boldsymbol{p}} \sup_j L(h_j, \boldsymbol{p}) = L_{1/N}^{N}(h).$$

The suprema over $\boldsymbol{p}$ range over $\mathsf{P}_{1/N,N}$. The functions h_j and h are both Darboux integrable on $[1/N, N]$, so the lower integral coincides with the integral. Therefore, we reach the stronger conclusion (C.4). This is what we had to show. ∎

C.10 Riemann implies Lebesgue

In this section, we sketch the proof of Proposition 4.37, which states that a bounded Riemann integrable function $f : [a, b] \to \mathbb{R}$ is Lebesgue integrable with respect to Lebesgue measure and that the two integrals coincide.

By adding a constant, we may assume that $f : [a, b] \to \mathbb{R}_+$ is positive. A *step function* is a function that is constant in the interior of each interval of a partition of $[a, b]$. From Section 5.2.3, recall that a *positive simple function* is a positive linear combination of indicator functions of (Borel) measurable sets. Every positive step function is a positive simple function because intervals are Borel measurable.

Beginning with the lower Riemann sum, we have a chain of inequalities:

$$\begin{aligned}
L_a^b(f) &= \sup\left\{\int_a^b s(x)\,\mathrm{d}x : s \le f \text{ for a positive step function } s\right\} \\
&= \sup\left\{\int_{[a,b]} s(x)\,\lambda(\mathrm{d}x) : s \le f \text{ for a positive step function } s\right\} \\
&\le \sup\left\{\int_{[a,b]} s(x)\,\lambda(\mathrm{d}x) : s \le f \text{ for a positive simple function } s\right\} \\
&\le \inf\left\{\int_{[a,b]} s(x)\,\lambda(\mathrm{d}x) : s \ge f \text{ for a positive simple function } s\right\} \\
&= \inf\left\{\int_{[a,b]} s(x)\,\lambda(\mathrm{d}x) : s \ge f \text{ for a positive step function } s\right\} \\
&= \inf\left\{\int_a^b s(x)\,\mathrm{d}x : s \ge f \text{ for a positive step function } s\right\} \\
&= U_a^b(f).
\end{aligned}$$

Since f is Riemann integrable, the lower Riemann sum $L_a^b(f)$ and the upper Riemann sum $U_a^b(f)$ both equal the Riemann integral of f. Meanwhile, by formula (5.5) and Proposition 5.35, the Lebesgue integral of f is the supremum of the integrals of simple functions that satisfy $s \le f$. We conclude that

$$\int_a^b f(x)\,\mathrm{d}x = \int_{[a,b]} f(x)\,\lambda(\mathrm{d}x). \tag{C.5}$$

This is what we needed to show.

C.11 Integration by parts

Finally, we establish Proposition 4.39. This result states that we can replace the improper Darboux integral on the right-hand side of Definition 4.20 (of the Lebesgue integral) with the corresponding Lebesgue integral.

We can just as easily establish the result in a general measure space $(\mathsf{X}, \mathscr{F}, \mu)$. Let $f : \mathsf{X} \to \overline{\mathbb{R}}_+$ be a positive, measurable function. Define

$$h_\mu(t) := \mu\{x \in \mathsf{X} : f(x) > t\} \quad \text{for } t \ge 0.$$

We must establish the identity

$$\int_0^\infty h_\mu(t)\,\mathrm{d}t = \int_{\mathbb{R}_+} h_\mu(t)\,\lambda(\mathrm{d}t)$$

The integral on the left-hand side is the improper Darboux integral of h_μ, while the right-hand side is the Lebesgue integral of h_μ with respect to the Lebesgue measure λ. Without loss of generality, we may assume that $h_\mu(t) < +\infty$ for $t \ge 0$, or else both integrals are infinite.

At this stage, the computation is straightforward. Since h_μ is positive and decreasing, it is improperly Darboux integrable (Theorem C.7). Moreover, using the definition of

the improper Darboux integral (C.3), we find that

$$\begin{aligned}\int_0^\infty h_\mu(t)\,\mathrm{d}t &= \lim_{N\to\infty}\int_{1/N}^{N} h_\mu(t)\,\mathrm{d}t \\ &= \lim_{N\to\infty}\int_{[1/N,N]} h_\mu(t)\,\lambda(\mathrm{d}t) = \int_{(0,+\infty)} h_\mu(t)\,\lambda(\mathrm{d}t).\end{aligned}$$

The second relation is (C.5), and the last relation is monotone convergence (Theorem 5.18) for the Lebesgue integral. Finally, to pass from the domain $(0,+\infty)$ to $\mathbb{R}_+ = [0,+\infty)$, we invoke Theorem 5.14 to take advantage of the fact that the Lebesgue integral is insensitive to negligible sets (in this case $\{0\}$).

Notes

This material may be found in any introductory book on real analysis. We have adapted this presentation from [Rud76].

D. Product Measures

In this appendix, we establish the basic results about the existence and uniqueness of product measures. Then we prove Kolmogorov's extension theorem, which asserts that there is a probability space that supports an independent family of random variables with specified marginal laws. Last, we prove the Fubini–Tonelli theorem on the interchange of integrals.

Agenda:

1. Existence of product measures
2. Monotone class theorem
3. Proof of Fubini–Tonelli

D.1 Construction of product measures

In this section, we prove Theorem 6.14 by constructing the product measure. For reference, we restate the result.

Theorem D.1 (Product measure: Existence and uniqueness). Let $(\mathsf{X}_i, \mathscr{F}_i, \mu_i)$ be *σ-finite* measure spaces for $i = 1, 2$. The product $(\mathsf{X}, \mathscr{F}) := (\mathsf{X}_1, \mathscr{F}_1) \times (\mathsf{X}_2, \mathscr{F}_2)$ carries a unique measure $\mu := \mu_1 \times \mu_2$, called the *product measure*, that satisfies

$$\mu(\mathsf{E} \times \mathsf{F}) = \mu_1(\mathsf{E}) \cdot \mu_2(\mathsf{F}) \quad \text{for all } \mathsf{E} \in \mathscr{F}_1 \text{ and } \mathsf{F} \in \mathscr{F}_2. \tag{D.1}$$

The triple $(\mathsf{X}, \mathscr{F}, \mu)$ is called the *product* of the measure spaces.

Warning: The construction of product measures fails without σ-finiteness! ∎

To establish this theorem, we activate the only machine we have for this purpose: Hahn–Kolmogorov (Theorem A.12).

The algebra of rectangles

First, construct the algebra of measurable rectangles:

$$\mathscr{A} := \operatorname{algebra}\{\mathsf{E} \times \mathsf{F} : \mathsf{E} \in \mathscr{F}_1 \text{ and } \mathsf{F} \in \mathscr{F}_2\} \subseteq \mathscr{F}.$$

A generic element $\mathsf{A} \in \mathscr{A}$ can be written as a finite union of disjoint rectangles:

$$\mathsf{A} = \dot{\bigcup}_{i=1}^{n} (\mathsf{E}_i \times \mathsf{F}_i) \quad \text{with } \mathsf{E}_i \in \mathscr{F}_1 \text{ and } \mathsf{F}_i \in \mathscr{F}_2. \tag{D.2}$$

By Exercise 6.4, this algebra generates the product σ-algebra: $\sigma(\mathscr{A}) = \mathscr{F}$.

A candidate premeasure

Define a candidate premeasure $\mu_0 : \mathscr{A} \to [0, +\infty]$ by the rule

$$\mu_0\big(\dot{\bigcup}_{i=1}^{n} (\mathsf{E}_i \times \mathsf{F}_i)\big) = \sum_{i=1}^{n} \mu_1(\mathsf{E}_i) \cdot \mu_2(\mathsf{F}_i). \tag{D.3}$$

One may check that the function μ_0 is well-defined; it does not depend on the particular representation of the union. In addition, μ_0 is finitely additive. Every measure on the product that satisfies the product condition (D.1) must also satisfy (D.3) by finite additivity, so this is the natural definition of a product premeasure.

Pre-countable additivity

As usual, the difficult step is to prove that μ_0 is pre-countably additive. That is,

$$\mu_0(\mathsf{A}) = \textstyle\sum_{i=1}^{\infty} \mu_0(\mathsf{A}_i) \quad \text{when} \quad \mathsf{A} = \dot{\bigcup}_{i=1}^{\infty} \mathsf{A}_i \in \mathscr{A} \quad \text{for sets } \mathsf{A}_i \in \mathscr{A}.$$

By the representation (D.2) and finite additivity of μ_0, we may assume that each member of the union is a measurable rectangle. Similarly, we may assume that the union itself is a measurable rectangle. In other words, it is enough to show that

$$\mu_0(\mathsf{E} \times \mathsf{F}) = \textstyle\sum_{i=1}^{\infty} \mu_0(\mathsf{E}_i \times \mathsf{F}_i) \quad \text{when} \quad \mathsf{E} \times \mathsf{F} = \dot{\bigcup}_{i=1}^{\infty} \mathsf{E}_i \times \mathsf{F}_i.$$

As usual $\mathsf{E}, \mathsf{E}_i \in \mathscr{F}_1$ and $\mathsf{F}, \mathsf{F}_i \in \mathscr{F}_2$. You should convince yourself of these claims, by pictures or by algebra.

In this situation, our task is considerably simplified because we can exploit integration theory for the component measures μ_1 and μ_2. Passing to indicator functions, the condition on the sets can be written in the form

$$\mathbb{1}_{\mathsf{E}}(x) \cdot \mathbb{1}_{\mathsf{F}}(y) = \sum\nolimits_{i=1}^{\infty} \mathbb{1}_{\mathsf{E}_i}(x) \cdot \mathbb{1}_{\mathsf{F}_i}(y) \quad \text{for all } x \in \mathsf{X}_1 \text{ and } y \in \mathsf{X}_2.$$

Integrate both sides with respect to μ_2 and then with respect to μ_1. The left-hand side equals

$$\begin{aligned} &\int_{\mathsf{X}_1} \mu_1(\mathrm{d}x) \left(\int_{\mathsf{X}_2} \mu_2(\mathrm{d}y)\, \mathbb{1}_{\mathsf{E}}(x) \cdot \mathbb{1}_{\mathsf{F}}(y) \right) \\ &\qquad = \int_{\mathsf{X}_1} \mu_1(\mathrm{d}x)\, \mathbb{1}_{\mathsf{E}}(x) \left(\int_{\mathsf{X}_2} \mu_2(\mathrm{d}y)\, \mathbb{1}_{\mathsf{F}}(y) \right) \\ &\qquad = \int_{\mathsf{X}_1} \mu_1(\mathrm{d}x)\, \mathbb{1}_{\mathsf{E}}(x) \cdot \mu_2(\mathsf{F}) \\ &\qquad = \mu_1(\mathsf{E}) \cdot \mu_2(\mathsf{F}). \end{aligned} \tag{D.4}$$

We have used positive linearity to draw positive constants out of both integrals, and there is no question about existence because everything is positive. Meanwhile, the right-hand side equals

$$\begin{aligned} &\int_{\mathsf{X}_1} \mu_1(\mathrm{d}x) \left(\int_{\mathsf{X}_2} \mu_2(\mathrm{d}y) \sum\nolimits_{i=1}^{\infty} \mathbb{1}_{\mathsf{E}_i}(x) \cdot \mathbb{1}_{\mathsf{F}_i}(y) \right) \\ &\qquad = \int_{\mathsf{X}_1} \mu_1(\mathrm{d}x) \left(\sum\nolimits_{i=1}^{\infty} \int_{\mathsf{X}_2} \mu_2(\mathrm{d}y)\, \mathbb{1}_{\mathsf{E}_i}(x) \cdot \mathbb{1}_{\mathsf{F}_i}(y) \right) \\ &\qquad = \sum\nolimits_{i=1}^{\infty} \int_{\mathsf{X}_1} \mu_1(\mathrm{d}x) \left(\int_{\mathsf{X}_2} \mu_2(\mathrm{d}y)\, \mathbb{1}_{\mathsf{E}_i}(x) \cdot \mathbb{1}_{\mathsf{F}_i}(y) \right) \\ &\qquad = \sum\nolimits_{i=1}^{\infty} \mu_1(\mathsf{E}_i) \cdot \mu_2(\mathsf{F}_i). \end{aligned}$$

We have used monotone convergence (Theorem 5.18) twice, once to draw the sum out of the integral with respect to μ_2 and then to draw the sum out of the integral with respect to μ_1. The last identity follows from (D.4). And everything depends on positivity.

Since the last two displays are equal, we have shown that the premeasure μ_0 is pre-countably additive on the algebra of measurable rectangles.

Strong σ-finiteness

Finally, since μ_1 and μ_2 are σ-finite measures, there are countable covers of X_1 and X_2 where

$$\mathsf{X}_i \subseteq \textstyle\bigcup_{j=1}^{\infty} \mathsf{E}_j^i \quad \text{and} \quad \mu_i(\mathsf{E}_j^i) < +\infty \quad \text{for } i = 1, 2.$$

The products $(\mathsf{E}_j^1 \times \mathsf{E}_j^2 : j \in \mathbb{N}) \subseteq \mathscr{A}$ form a countable cover of the product space:

$$\mathsf{X}_1 \times \mathsf{X}_2 \subseteq \bigcup_{j=1}^{\infty} (\mathsf{E}_j^1 \times \mathsf{E}_j^2) \quad \text{and} \quad \mu_0(\mathsf{E}_j^1 \times \mathsf{E}_j^2) = \mu_1(\mathsf{E}_j^1) \cdot \mu_2(\mathsf{E}_j^2) < +\infty.$$

In other words, μ_0 is strongly σ-finite.

Hahn–Kolmogorov

We deduce that μ_0 is a premeasure on $\mathscr{A}$. Theorem A.12 implies that μ_0 extends to a unique measure μ on the product σ-algebra $\mathscr{F} = \sigma(\mathscr{A})$. This measure μ satisfies (D.3), so it is the product measure.

D.2 Kolmogorov's extension theorem

Let us recall the statement of the Kolmogorov extension theorem, which allows us to construct a sequence of independent random variables. This section offers a proof of the result by means of the Hahn–Kolmogorov theorem.

Theorem D.2 (Kolmogorov extension). Let $(\mu_1, \mu_2, \mu_3, \dots)$ be a sequence of probability measures defined on the Borel sets of the real line. There exists a probability space $(\Omega, \mathscr{F}, \mathbb{P})$ on which we can define a sequence $(X_1, X_2, X_3, \dots)$ of *independent* random variables where the law of X_i is μ_i for each index $i \in \mathbb{N}$. That is,

$$\mathbb{P}\left\{(X_{i_1}, \dots, X_{i_n}) \in \mathsf{B}_{i_1} \times \cdots \times \mathsf{B}_{i_n}\right\} = \prod\nolimits_{j=1}^{n} \mathbb{P}\left\{X_{i_j} \in \mathsf{B}_{i_j}\right\} = \prod\nolimits_{j=1}^{n} \mu_{i_j}(\mathsf{B}_{i_j})$$

for all $n \in \mathbb{N}$, and distinct indices $i_1 < \cdots < i_n \in \mathbb{N}$, and Borel sets $\mathsf{B}_{i_j} \in \mathscr{B}(\mathbb{R})$ for $j = 1, \dots, n$.

Proof. We mimic the construction behind Theorem D.1 on the existence of product measures (Section D.1). First, we define the sample space to be the family of all real-valued sequences, which is the countable product of copies of the real line:

$$\Omega := \mathbb{R}^{\mathbb{N}} := \{\boldsymbol{\omega} = (\omega_1, \omega_2, \omega_3, \dots) : \omega_i \in \mathbb{R} \text{ for each } i \in \mathbb{N}\}.$$

Second, we introduce the coordinate functions

$$X_i(\boldsymbol{\omega}) = \pi_i(\boldsymbol{\omega}) = \omega_i \quad \text{for each } \boldsymbol{\omega} \in \Omega.$$

Third, we equip Ω with the σ-algebra $\mathscr{F}$ generated by all Borel cylinders:

$$\mathscr{F} := \sigma(X_i : i \in \mathbb{N}) := \sigma\{X_i^{-1}(\mathsf{B}) : \mathsf{B} \in \mathscr{B}(\mathbb{R}) \text{ and } i \in \mathbb{N}\}.$$

In particular, the σ-algebra contains all Borel rectangles of the form $\mathsf{B}_1 \times \mathsf{B}_2 \times \mathsf{B}_3 \times \cdots$ where $\mathsf{B}_i \in \mathscr{B}(\mathbb{R})$. Fourth, we define the probability measure on Borel rectangles in a finite number of coordinates, as in the statement of the theorem:

$$\mathbb{P}\left\{(X_{i_1}, \dots, X_{i_n}) \in \mathsf{B}_{i_1} \times \cdots \times \mathsf{B}_{i_n}\right\} = \prod\nolimits_{j=1}^{n} \mathbb{P}\left\{X_{i_j} \in \mathsf{B}_{i_j}\right\} = \prod\nolimits_{j=1}^{n} \mu_{i_j}(\mathsf{B}_{i_j}).$$

This is the natural definition of the product measure for a sequence.

The proof that $\mathbb{P}$ extends to a unique probability measure on the σ-algebra $\mathscr{F}$ is the same as the proof of Theorem 6.14. We use the same method to check that the definition of $\mathbb{P}$ yields a (finite) pre-measure on the algebra generated by Borel rectangles in a finite number of coordinates. Hahn–Kolmogorov (Theorem A.12) then yields the extension to $\mathscr{F}$. ∎

D.3 The monotone class theorem

To prove the Fubini–Tonelli theorem, we need a result from set theory that gives an alternative method for generating σ-algebras.

Definition D.3 (Monotone class). A system $\mathscr{M}$ of sets is called a *monotone class* if it is stable under increasing countable unions and decreasing countable intersections:

1. **Increase:** If $\mathsf{A}_1 \subseteq \mathsf{A}_2 \subseteq \mathsf{A}_3 \subseteq \cdots$ and $\mathsf{A}_i \in \mathscr{M}$, then $\bigcup_{i=1}^{\infty} \mathsf{A}_i \in \mathscr{M}$.
2. **Decrease:** If $\mathsf{A}_1 \supseteq \mathsf{A}_2 \supseteq \mathsf{A}_3 \supseteq \cdots$ and $\mathsf{A}_i \in \mathscr{M}$, then $\bigcap_{i=1}^{\infty} \mathsf{A}_i \in \mathscr{M}$.

Every σ-algebra is a monotone class because a σ-algebra contains all countable unions and intersections of its members. In general, a monotone class $\mathscr{M}$ may not even contain enough sets to form an algebra. Nevertheless, if $\mathscr{M}$ happens to contain an algebra $\mathscr{A}$, then it even contains $\sigma(\mathscr{A})$. This statement is the content of the monotone class theorem, which we are about to prove.

This result is useful because we may be able to extend an algebra to a monotone class with a minimum of effort. In contrast, it may take a lot of work to construct the generated σ-algebra directly.

Theorem D.4 (Monotone class). Let $\mathscr{A}$ be a set algebra. The σ-algebra $\sigma(\mathscr{A})$ equals the *smallest* monotone class that contains $\mathscr{A}$.

D.3.1 Monotone class theorem: Proof

The proof of Theorem D.4 involves set gymnastics and close reasoning.

Setup

Fix an algebra $\mathscr{A}$ on the domain X. Let $\mathscr{M}$ be the intersection of every monotone class that contains $\mathscr{A}$. You may check that the family $\mathscr{M}$ is itself a monotone class that contains $\mathscr{A}$. We interpret $\mathscr{M}$ as the *smallest* monotone class that contains the algebra $\mathscr{A}$. In particular, $\mathscr{M} \subseteq \sigma(\mathscr{A})$, because a σ-algebra is a monotone class.

We need to prove the reverse inclusion $\sigma(\mathscr{A}) \subseteq \mathscr{M}$. Since $\sigma(\mathscr{A})$ is the smallest σ-algebra containing $\mathscr{A}$, it is enough to show that $\mathscr{M}$ is itself a σ-algebra. What steps must we take? Since $\mathscr{M}$ contains the algebra $\mathscr{A}$, it obviously contains the empty set $\emptyset$ and the domain X. It remains to demonstrate that $\mathscr{M}$ is stable under complements and under countable unions.

Complements

First, we check that $\mathscr{M}$ is stable under complements. Introduce the family $\mathscr{C}$ of complemented sets in $\mathscr{M}$:

$$\mathscr{C} := \{\mathsf{M} \in \mathscr{M} : \mathsf{M}^{\mathsf{c}} \in \mathscr{M}\} \subseteq \mathsf{M}.$$

Since the algebra $\mathscr{A}$ is contained in $\mathscr{M}$ and the algebra is stable under complements, we see that the algebra $\mathscr{A}$ belongs to $\mathscr{C}$. Now, consider an increasing sequence $(\mathsf{B}_i : i \in \mathbb{N}) \subseteq \mathscr{C}$. By De Morgan's law,

$$\Big(\bigcup_{i=1}^{\infty} \mathsf{B}_i\Big)^{\mathsf{c}} = \bigcap_{i=1}^{\infty} \mathsf{B}_i^{\mathsf{c}} \in \mathscr{M}$$

because $(\mathsf{B}_i^{\mathsf{c}} : i \in \mathbb{N}) \subseteq \mathscr{M}$ is a decreasing sequence and $\mathscr{M}$ is a monotone class. It follows that $\mathscr{C}$ is stable under increasing unions. By a parallel argument, the family $\mathscr{C}$ is stable under decreasing intersections. Therefore, $\mathscr{C}$ is a monotone class that contains the algebra $\mathscr{A}$. By minimality of the monotone class, $\mathscr{M} \subseteq \mathscr{C}$, so that $\mathscr{C} = \mathscr{M}$. Therefore, the family $\mathscr{M}$ is stable under complementation.

Finite unions

Second, let us show that $\mathscr{M}$ is stable under finite unions. Fix a set $\mathsf{A} \in \mathscr{A}$ in the *algebra*, and define the family $\mathscr{U}(\mathsf{A})$ of sets in $\mathscr{M}$ that stay in $\mathscr{M}$ after union with A:

$$\mathscr{U}(\mathsf{A}) := \{\mathsf{M} \in \mathscr{M} : \mathsf{A} \cup \mathsf{M} \in \mathscr{M}\} \subseteq \mathscr{M}.$$

Since the algebra $\mathscr{A}$ is contained in $\mathscr{M}$ and the algebra is stable under unions, the algebra $\mathscr{A}$ belongs to $\mathscr{U}(\mathsf{A})$. Now, if $(\mathsf{B}_i : i \in \mathbb{N}) \subseteq \mathscr{U}(\mathsf{A})$ is an increasing sequence,

$$\mathsf{A} \cup \bigcup_{i=1}^{\infty} \mathsf{B}_i = \bigcup_{i=1}^{\infty} (\mathsf{A} \cup \mathsf{B}_i) \in \mathscr{M}$$

because $(\mathsf{A} \cup \mathsf{B}_i : i \in \mathbb{N})$ is increasing and $\mathscr{M}$ is a monotone class. Therefore, the collection $\mathscr{U}(\mathsf{A})$ is stable under increasing unions. By a parallel argument, $\mathscr{U}(\mathsf{A})$ is stable under decreasing intersections. It follows that $\mathscr{U}(\mathsf{A})$ is itself a monotone class containing $\mathscr{A}$, and $\mathscr{U}(\mathsf{A}) = \mathscr{M}$ by minimality of $\mathscr{M}$. In other words, $\mathsf{A} \cup \mathsf{M} \in \mathscr{M}$ whenever $\mathsf{A} \in \mathscr{A}$ and $\mathsf{M} \in \mathscr{M}$.

Next, fix an *arbitrary* set $\mathsf{E} \in \mathscr{M}$, and make the same construction:

$$\mathscr{U}(\mathsf{E}) := \{\mathsf{M} \in \mathscr{M} : \mathsf{E} \cup \mathsf{M} \in \mathscr{M}\} \subseteq \mathscr{M}.$$

We have just shown that $\mathsf{A} \cup \mathsf{E} \in \mathscr{M}$ for every $\mathsf{A} \in \mathscr{A}$, so the algebra $\mathscr{A} \subseteq \mathscr{U}(\mathsf{E})$. Repeating the argument from the last paragraph, we learn that $\mathscr{U}(\mathsf{E})$ is a monotone class containing $\mathscr{A}$. Therefore, $\mathscr{U}(\mathsf{E}) = \mathscr{M}$. We confirm that $\mathsf{E} \cup \mathsf{M} \in \mathscr{M}$ for all $\mathsf{E}, \mathsf{M} \in \mathscr{M}$.

Countable unions

Last, we prove that the monotone class $\mathscr{M}$ is stable under countable unions, increasing or not. Elect a sequence of sets from the class: $(\mathsf{M}_i : i \in \mathbb{N}) \subseteq \mathscr{M}$. Consider the increasing sequence of partial unions: $(\bigcup_{i=1}^{k} \mathsf{M}_i : k \in \mathbb{N}) \in \mathscr{M}$. Indeed, the partial unions belong to $\mathscr{M}$ because it is stable under finite unions. Since $\mathscr{M}$ is a monotone class, it contains the increasing union of the partial unions. That is, $\bigcup_{i=1}^{\infty} \mathsf{M}_i \in \mathscr{M}$.

D.4 Fubini–Tonelli theorem: Proof

Consider measure spaces $(\mathsf{X}_i, \mathscr{F}_i, \mu_i)$ for $i = 1, 2$ with product $(\mathsf{X}, \mathscr{F}, \mu)$. The proof of Theorem 6.23 follows a standard strategy. First, we treat the case where the measures μ_i are finite. For the finite case, we begin with indicators, proceed to positive simple functions, positive measurable functions, and then integrable functions. Finally, we pass to the σ-finite case by partitioning the domain.

Indicators

Until further notice, assume that μ_1 and μ_2 are *finite* measures, so that μ is also a finite measure. Introduce the family $\mathscr{M}$ that contains every product-measurable set $\mathsf{A} \in \mathscr{F}$ for which

$$\int_{\mathsf{X}} \mathrm{d}\mu \, \mathbb{1}_{\mathsf{A}} = \int_{\mathsf{X}_2} \mathrm{d}\mu_2 \left(\int_{\mathsf{X}_1} \mathrm{d}\mu_1 \, \mathbb{1}_{\mathsf{A}} \right) = \int_{\mathsf{X}_1} \mathrm{d}\mu_1 \left(\int_{\mathsf{X}_2} \mathrm{d}\mu_2 \, \mathbb{1}_{\mathsf{A}} \right). \tag{D.5}$$

The validity of the expression (D.5) depends on the fact that the indicator function $\mathbb{1}_{\mathsf{A}}$ is product-measurable, so that its sections are measurable functions on the factor spaces (Exercise 6.22).

Recalling the calculation (D.4), we realize that the family $\mathscr{M}$ contains every measurable rectangle $\mathsf{E} \times \mathsf{F}$ where $\mathsf{E} \in \mathscr{F}_1$ and $\mathsf{E} \in \mathscr{F}_2$. Since $\mathscr{F}$ is a σ-algebra, it is

stable under increasing unions. Therefore, by the monotone convergence theorem, $\mathscr{M}$ is also stable under increasing unions. In detail, if $\mathsf{B}_i \uparrow \mathsf{B}$ and the formula (D.5) is valid for each $\mathsf{A} = \mathsf{B}_i \in \mathscr{M}$, then (D.5) is valid when $\mathsf{A} = \mathsf{B}$. Likewise, $\mathscr{M}$ is stable under decreasing intersections of sets in $\mathscr{F}$. This claim requires *downward* monotone convergence (Exercise 5.38), which is only valid for finite measures. In other words, $\mathscr{M}$ is a monotone class that contains all rectangles.

Theorem D.4 now ensures that $\mathscr{M}$ is a σ-algebra that contains all rectangles. By definition, $\mathscr{F}$ is the smallest such σ-algebra. Therefore, $\mathscr{M} = \mathscr{F}$. We conclude that the interchange (D.5) of integrals is valid for the indicator of every product-measurable set.

Positive simple functions

The rest of the argument follows a prescribed route. A positive simple function $f : \mathsf{X} \to \mathbb{R}_+$ takes the form

$$f = \sum\nolimits_{i=1}^{n} \alpha_i \mathbb{1}_{\mathsf{A}_i} \quad \text{for } \alpha_i \geq 0 \text{ and } \mathsf{A}_i \in \mathscr{F}.$$

By positive linearity of the Lebesgue integral,

$$\int_{\mathsf{X}} \mathrm{d}\mu\, f = \int_{\mathsf{X}_2} \mathrm{d}\mu_2 \left(\int_{\mathsf{X}_1} \mathrm{d}\mu_1\, f \right) = \int_{\mathsf{X}_1} \mathrm{d}\mu_1 \left(\int_{\mathsf{X}_2} \mathrm{d}\mu_2\, f \right). \tag{D.6}$$

We have also used the fact that linear combinations of indicators are measurable (Exercise 4.15), and so their sections are measurable (Exercise 6.22).

Positive functions

Now, every positive, product-measurable function $f : \mathsf{X} \to \overline{\mathbb{R}}_+$ is an increasing limit of positive simple functions (Exercise 5.10). The sections of the approximations are all measurable and converge pointwise to the sections of the limit f. Therefore, monotone convergence allows us to extend (D.6) to the positive function f.

Integrable functions

Last, if $f : \mathsf{X} \to \mathbb{R}$ is integrable, then its positive part f_+ and negative part f_- and their sections are integrable. We can extend (D.6) to f by applying the result to the positive and negative parts and subtracting. Integrability ensures that there are no competing infinities.

The σ-finite case

To complete the argument, we must address the case where μ_1 and μ_2 are σ-finite. Consider countable covers by disjoint sets of finite measure:

$$\mathsf{X}_1 = \dot{\bigcup}_{i=1}^{\infty} \mathsf{E}_i \quad \text{and} \quad \mathsf{X}_2 = \dot{\bigcup}_{j=1}^{\infty} \mathsf{F}_j \quad \text{whence} \quad \mathsf{X} = \dot{\bigcup}_{i,j=1}^{\infty} \mathsf{E}_i \times \mathsf{F}_j.$$

Here, $\mu_1(\mathsf{E}_i) < +\infty$ and $\mu_2(\mathsf{F}_j) < +\infty$ for all $i, j \in \mathbb{N}$. Let $f : \mathsf{X} \to \overline{\mathbb{R}}_+$ be a positive, measurable function. Using domain decomposition (Exercise 5.40), the Fubini–Tonelli identities (D.6), and the Tonelli theorem for sums (Exercise 5.39) repeatedly,

$$\begin{aligned} \int_{\mathsf{X}} \mathrm{d}\mu\, f &= \sum\nolimits_{i,j=1}^{\infty} \int_{\mathsf{E}_i \times \mathsf{F}_j} \mathrm{d}\mu\, f = \sum\nolimits_{i,j=1}^{\infty} \int_{\mathsf{E}_i} \mathrm{d}\mu_1 \left(\int_{\mathsf{F}_j} \mathrm{d}\mu_2\, f \right) \\ &= \sum\nolimits_{i=1}^{\infty} \int_{\mathsf{E}_i} \mathrm{d}\mu_1 \left(\sum\nolimits_{j=1}^{\infty} \int_{\mathsf{F}_j} \mathrm{d}\mu_2\, f \right) = \int_{\mathsf{X}_1} \mathrm{d}\mu_1 \left(\int_{\mathsf{X}_2} \mathrm{d}\mu_2\, f \right). \end{aligned}$$

A similar calculation yields the same result with the integrals over μ_1 and μ_2 exchanged. Finally, we extend this result to integrable functions by considering the positive and negative parts.

E. Uniqueness of Measures

This appendix covers some more topics around the uniqueness of measures. In particular, we describe some additional tools for checking independence of σ-algebras and uniqueness of measures. Last, as a more sophisticated application of these ideas, we prove Kolmogorov's 0–1 law.

Agenda:

1. Intersection-stable systems
2. Uniqueness of measure
3. Dynkin's lemma
4. Kolmogorov's 0–1 law

E.1 Intersection-stable systems

It can be challenging to give a direct proof that two σ-algebras are independent. The basic reason is that the sets in a σ-algebra are not very explicit. Instead, it is often convenient to work with smaller families of sets that we can describe completely. This approach is widely used in probability theory.

E.1.1 Intersection-stable systems

We begin with the definition of a new type of set system.

Definition E.1 (Intersection-stable system). A family $\mathcal{S} \subseteq \mathcal{P}(\Omega)$ of subsets of Ω is called an *intersection-stable system* when $\Omega \in \mathcal{S}$ and

$$A, B \in \mathcal{S} \quad \text{implies} \quad A \cap B \in \mathcal{S}.$$

Intersection-stable systems are also called *multiplicative systems* or *π-systems*.

Obviously, a σ-algebra on Ω is an intersection-stable system. We will be interested in the case where an intersection-stable system $\mathcal{S}$ *generates* a σ-algebra $\mathcal{F}$. That is, $\mathcal{F} = \sigma(\mathcal{S})$. Even in this setting, the intersection-stable system may be far smaller than the σ-algebra.

Example E.2 (Semi-infinite intervals). Consider the family $\mathcal{S} = \{(-\infty, a] : a \in \mathbb{R}\}$ of semi-infinite subintervals of the real line. Then $\mathcal{S}$ is an intersection-stable system in $\mathbb{R}$. It generates the Borel sets of the real line: $\sigma(\mathcal{S}) = \mathcal{B}(\mathbb{R})$. ∎

By convention, the class $\mathcal{S}$ includes the real line $\mathbb{R}$.

Example E.3 (Semi-infinite rectangles). The family $\mathcal{S} = \{(-\infty, a] \times (-\infty, b] : a, b \in \mathbb{R}\}$ is an intersection-stable system in $\mathbb{R}^2$. It generates the Borel sets of the plane: $\sigma(\mathcal{S}) = \mathcal{B}(\mathbb{R}^2)$. ∎

By convention, the class $\mathcal{S}$ includes the real plane $\mathbb{R}^2$.

E.1.2 Uniqueness of measure

The key fact about intersection-stable systems is that they are already large enough to determine a measure completely. Let us emphasize that this result does not guarantee the existence of a measure—just the uniqueness.

Theorem E.4 (Uniqueness of measure). Let $(\Omega, \mathcal{F})$ be a measurable space. Suppose that $\mathcal{S}$ is an intersection-stable system with $\mathcal{F} = \sigma(\mathcal{S})$. Let μ_1 and μ_2 be finite

measures on $\mathcal{F}$. If μ_1 and μ_2 agree on $\mathcal{S}$, then they agree on $\mathcal{F}$. That is,

$$\begin{aligned} &\mu_1(\mathsf{A}) = \mu_2(\mathsf{A}) \text{ for all } \mathsf{A} \in \mathcal{S} \\ \text{implies} \quad &\mu_1(\mathsf{E}) = \mu_2(\mathsf{E}) \text{ for all } \mathsf{E} \in \mathcal{F}. \end{aligned}$$

We will prove Theorem E.4 below in Section E.1.4. For now, we outline the consequences for distribution functions and independence of σ-algebras.

Corollary E.5 (Distribution functions: Uniqueness). Suppose that $F_X : \mathbb{R} \to [0,1]$ is a distribution function; that is, an increasing, right-continuous function with left asymptote zero and right asymptote one. Then there is *at most* one probability measure μ_X with the property that

$$\mu_X(a,b] = F_X(b) - F_X(a) \quad \text{for all real numbers } a < b.$$

Warning: By itself, this result does not guarantee that μ_X extends to a Borel measure on $\mathbb{R}$. See Problem A.17. ■

Proof. The class $\mathcal{S} = \{(a,b] : a < b\}$ of semi-open intervals (including $\mathbb{R}$) forms an intersection-stable system in $\mathbb{R}$. By Exercise 3.6, the class $\mathcal{S}$ generates the Borel σ-algebra $\mathcal{B}(\mathbb{R})$.

According to the increasing limit property of a measure, the total mass of the measure μ_X must satisfy

$$\mu_X(\mathbb{R}) = \lim_{b\uparrow+\infty} F_X(b) - \lim_{a\downarrow-\infty} F_X(a) = 1.$$

As a consequence, Theorem E.4 guarantees that there is at most one probability measure μ_X that takes the specified values on the semi-open intervals. ■

Corollary E.6 (Independence: Intersection-stable systems). Let $(\Omega, \mathcal{F}, \mathbb{P})$ be a probability space. Suppose that $\mathcal{S}_i \subseteq \mathcal{F}$ are intersection-stable systems that generate σ-algebras $\mathcal{G}_i = \sigma(\mathcal{S}_i)$ for $i = 1, 2$. Assume that

$$\mathbb{P}(\mathsf{A} \cap \mathsf{B}) = \mathbb{P}(\mathsf{A}) \cdot \mathbb{P}(\mathsf{B}) \quad \text{for all } \mathsf{A} \in \mathcal{S}_1 \text{ and } \mathsf{B} \in \mathcal{S}_2.$$

Then $\mathcal{G}_1$ and $\mathcal{G}_2$ are independent.

Proof. Fix an event $\mathsf{B} \in \mathcal{S}_2$. Define measures

$$\mu_1(\mathsf{A}) = \mathbb{P}(\mathsf{A} \cap \mathsf{B}) \quad \text{and} \quad \mu_2(\mathsf{A}) = \mathbb{P}(\mathsf{A}) \cdot \mathbb{P}(\mathsf{B}) \quad \text{for } \mathsf{A} \in \mathcal{F}.$$

Clearly, $\mu_1(\Omega) = \mu_2(\Omega)$. By hypothesis, μ_1 and μ_2 agree on events $\mathsf{A} \in \mathcal{S}_1$. Theorem E.4 implies that they agree on events $\mathsf{E} \in \mathcal{G}_1 = \sigma(\mathcal{S}_1)$.

We can now fix an event $\mathsf{E} \in \mathcal{G}_1$. Define measures

$$\nu_1(\mathsf{B}) = \mathbb{P}(\mathsf{E} \cap \mathsf{B}) \quad \text{and} \quad \nu_2(\mathsf{B}) = \mathbb{P}(\mathsf{E}) \cdot \mathbb{P}(\mathsf{B}) \quad \text{for } \mathsf{B} \in \mathcal{F}.$$

Taking $\mathsf{B} = \Omega$, we see that these measures have the same total mass. By the argument in the last paragraph, these measures agree on events $\mathsf{B} \in \mathcal{S}_2$. Theorem E.4 now implies that the measures agree on all events $\mathsf{F} \in \mathcal{G}_2 = \sigma(\mathcal{S}_2)$. That is,

$$\mathbb{P}(\mathsf{E} \cap \mathsf{F}) = \mathbb{P}(\mathsf{E}) \cdot \mathbb{P}(\mathsf{F}) \quad \text{for all } \mathsf{E} \in \mathcal{G}_1 \text{ and } \mathsf{F} \in \mathcal{G}_2.$$

The σ-algebras $\mathcal{G}_1$ and $\mathcal{G}_2$ are independent. ■

Example E.7 (Independence and distribution functions). Suppose that X, Y are real random variables that satisfy

$$\mathbb{P}\{X \le a, Y \le b\} = \mathbb{P}\{X \le a\} \cdot \mathbb{P}\{Y \le b\} \quad \text{for all } a, b \in \mathbb{R}.$$

Then X and Y are independent. Indeed, $\mathcal{S} = \{(-\infty, a] : a \in \mathbb{R}\}$ is an intersection-stable system that generates the Borel σ-algebra on $\mathbb{R}$. Therefore, Corollary E.6 implies that $\sigma(X)$ and $\sigma(Y)$ are independent σ-algebras. That is, X and Y are independent random variables. ∎

Example E.8 (Independence of families of σ-algebras). Let $\{\mathcal{G}_i : i \in \mathbb{N}\}$ be independent σ-algebras. We can prove that, for any $\mathsf{I} \subseteq \mathbb{N}$, the σ-algebra $\sigma(\mathcal{G}_i : i \in \mathsf{I})$ and $\sigma(\mathcal{G}_j : j \notin \mathsf{I})$ are independent.

Consider the intersection-stable system $\mathcal{S}_1$ consisting of all finite intersections of sets from $\bigcup_{i\in\mathsf{I}} \mathcal{G}_i$, which generates $\sigma(\mathcal{G}_i : i \in \mathsf{I})$. Consider the intersection-stable system $\mathcal{S}_2$ consisting of all finite intersections of sets from $\bigcup_{j\notin\mathsf{I}} \mathcal{G}_j$, which generates $\sigma(\mathcal{G}_j : j \notin \mathsf{I})$. By the definition of independence,

$$\mathbb{P}(\mathsf{E} \cap \mathsf{F}) = \mathbb{P}(\mathsf{E}) \cdot \mathbb{P}(\mathsf{F}) \quad \text{for all } \mathsf{E} \in \mathcal{S}_1 \text{ and } \mathsf{F} \in \mathcal{S}_2.$$

Corollary E.6 implies the result. ∎

E.1.3 Limit-stable systems

To establish Theorem E.4, we need to introduce another type of set system that is simpler than a σ-algebra and has properties complementary to an intersection-stable system.

Definition E.9 (Limit-stable system). A collection $\mathcal{D} \subseteq \mathcal{P}(\Omega)$ of subsets of Ω is called a *limit-stable system* when

1. **Everything:** $\Omega \in \mathcal{D}$.
2. **Set difference:** For nested $\mathsf{D}_1 \subseteq \mathsf{D}_2$ with $\mathsf{D}_1, \mathsf{D}_2 \in \mathcal{D}$, we have $\mathsf{D}_2 \setminus \mathsf{D}_1 \in \mathcal{D}$.
3. **Increasing limits:** For an increasing sequence $\mathsf{D}_1 \subseteq \mathsf{D}_2 \subseteq \mathsf{D}_3 \subseteq \cdots$ of sets that belong to $\mathcal{D}$, the limit $\bigcup_{i=1}^{\infty} \mathsf{D}_i \in \mathcal{D}$.

Limit-stable systems are often called *Dynkin systems* or *λ-systems*.

It is immediate that a limit-stable system $\mathcal{D}$ is a σ-algebra if and only if it is stable under (finite) intersections.

Exercise E.10 (Intersection-stable + limit-stable). Show that a collection $\mathcal{F}$ of subsets of Ω is a σ-algebra if and only if $\mathcal{F}$ is both an intersection-stable system and a limit-stable system.

In fact, to obtain a σ-algebra, it is sufficient that a limit-stable system contains an intersection-stable system. Our next goal is to establish this claim.

Definition E.11 (Generated limit-closed system). Suppose that $\mathcal{S}$ is a collection of subsets of Ω. Define the collection $d(\mathcal{S})$ to be the intersection of all limit-stable systems $\mathcal{D}$ on Ω with $\mathcal{S} \subseteq \mathcal{D}$.

Exercise E.12 (Generated limit-closed system). Check that $d(\mathcal{S})$ is a limit-stable system.

Proposition E.13 (Dynkin's lemma). If $\mathcal{S}$ is an intersection-stable system, then $d(\mathcal{S}) = \sigma(\mathcal{S})$. In particular, a limit-stable system that contains an intersection-stable system contains the σ-algebra generated by the intersection-stable system.

Proof. In view of Exercise E.10, it suffices to check that $d(\mathcal{S})$ is an intersection-stable system. This argument is reminiscent of the proof of the monotone class theorem (Theorem D.4).

Introduce the part of the limit-stable system $d(\mathcal{S})$ that is stable under intersection with all sets in $\mathcal{S}$:

$$\mathcal{I} := \{\mathsf{D} \in d(\mathcal{S}) : \mathsf{S} \cap \mathsf{D} \in d(\mathcal{S}) \text{ for all } \mathsf{S} \in \mathcal{S}\}.$$

Since $\mathcal{S}$ is stable under intersections, $\mathcal{S} \subseteq \mathcal{I}$. We *claim* that $\mathcal{I}$ is a limit-stable system. As a consequence, $\mathcal{I} = d(\mathcal{S})$ because $d(\mathcal{S})$ is the smallest limit-stable system containing $\mathcal{S}$.

Second, we introduce the part of the limit-stable system $d(\mathcal{S})$ that is stable under intersection with all sets in $d(\mathcal{S})$:

$$\mathcal{J} := \{\mathsf{D} \in d(\mathcal{S}) : \mathsf{S} \cap \mathsf{D} \in d(\mathcal{S}) \text{ for all } \mathsf{S} \in d(\mathcal{S})\}.$$

We have already shown that $\mathcal{I} \subseteq \mathcal{J}$. For the same reasons (below) that $\mathcal{I}$ is a limit-stable system, $\mathcal{J}$ is a limit-stable system. Therefore, $\mathcal{J} = d(\mathcal{S})$. But $\mathcal{J}$ is intersection-stable by construction. This is what we needed to prove.

To verify the outstanding claim, we need to check that $\mathcal{I}$ satisfies the three properties of a limit-stable system. First, $\Omega \in \mathcal{I}$ because $\mathsf{S} \cap \Omega = \mathsf{S} \in \mathcal{I}$. Second, choose nested sets $\mathsf{D}_1 \subseteq \mathsf{D}_2$ from $\mathcal{I}$. For each $\mathsf{S} \in \mathcal{S}$,

$$(\mathsf{D}_2 \setminus \mathsf{D}_1) \cap \mathsf{S} = (\mathsf{D}_2 \cap \mathsf{S}) \setminus (\mathsf{D}_1 \cap \mathsf{S}) \in \mathcal{I}.$$

Indeed, since $\mathsf{D}_n \in \mathcal{I}$ for $n = 1, 2$, we must have $\mathsf{D}_n \cap \mathsf{S} \in d(\mathcal{I})$, which is stable under set difference. Last, we select an increasing sequence $(\mathsf{D}_n \subseteq \mathcal{I} : n \in \mathbb{N})$ with $\mathsf{D}_n \uparrow \mathsf{E}$. Then

$$(\mathsf{D}_n \cap \mathsf{S}) \uparrow (\mathsf{E} \cap \mathsf{S}) \in \mathcal{I}.$$

Indeed, $\mathsf{D}_n \cap \mathsf{S} \in d(\mathsf{S})$ for each $n \in \mathbb{N}$, and $d(\mathsf{S})$ is stable under increasing limits.

The identical argument implies that $\mathcal{J}$ is a limit-stable system as well. ∎

E.1.4 Proof of uniqueness of measure theorem

We are now prepared to establish Theorem E.4 on the uniqueness of measures. Let μ_1 and μ_2 be two measures on $(\Omega, \mathcal{F})$. We assume that the measures agree on an intersection-stable system $\mathcal{S} \subseteq \mathcal{F}$ that generates $\mathcal{F}$. In particular, $\mu_1(\Omega) = \mu_2(\Omega)$, so the measures have the same total mass.

Introduce the class of events in $\mathcal{F}$ that have the same measure:

$$\mathcal{D} := \{\mathsf{E} \in \mathcal{F} : \mu_1(\mathsf{E}) = \mu_2(\mathsf{E})\}.$$

By hypothesis, $\mathcal{D}$ contains the intersection-stable system $\mathcal{S}$. We *claim* that $\mathcal{D}$ is a limit-stable system as well. Therefore, Dynkin's lemma (Proposition E.13) implies that $\mathcal{F} = \sigma(\mathcal{S}) = d(\mathcal{S}) \subseteq \mathcal{D}$. We determine that the measures agree on $\mathcal{F}$.

To prove the claim, we must check that $\mathcal{D}$ satisfies the three properties of a limit-stable system. First, $\Omega \in \mathcal{D}$ because the two measures have the same total mass. Second, choose nested sets $\mathsf{A} \subseteq \mathsf{B}$ from $\mathcal{D}$. Calculate that

$$\mu_1(\mathsf{B} \setminus \mathsf{A}) = \mu_1(\mathsf{B}) - \mu_1(\mathsf{A}) = \mu_2(\mathsf{B}) - \mu_2(\mathsf{A}) = \mu_2(\mathsf{B} \setminus \mathsf{A}).$$

Thus, $\mathsf{B} \setminus \mathsf{A} \in \mathcal{D}$. Finally, consider an increasing sequence $(\mathsf{D}_n : n \in \mathbb{N})$ with $\mathsf{D}_n \uparrow \mathsf{E}$. By the increasing limit property of a measure (Proposition 2.30),

$$\mu_1(\mathsf{E}) = \sup_{n \in \mathbb{N}} \mu_1(\mathsf{D}_n) = \sup_{n \in \mathbb{N}} \mu_2(\mathsf{D}_n) = \mu_2(\mathsf{E}).$$

We conclude that $\mathsf{E} \in \mathcal{D}$. Therefore, $\mathcal{D}$ is a limit-closed system.

E.2 *Kolmogorov's 0–1 law

In this section, we give an example of a situation where independence of σ-algebras plays a central role.

E.2.1 The tail σ-algebra

The Kolmogorov extension theorem (Theorem 13.24) asserts that it is possible to construct an independent sequence of random variables. When we try to study convergence properties associated with this sequence, we encounter the following σ-algebra.

Definition E.14 (Tail σ-algebra). Let $(X_1, X_2, X_3, \dots)$ be an independent sequence of real random variables. The *tail σ-algebra* $\mathcal{T}$ is defined as

$$\mathcal{T} := \bigcap\nolimits_{n=1}^{\infty} \sigma(X_n, X_{n+1}, X_{n+2}, \dots)$$

In other words, the tail σ-algebra includes only events that do not depend on any prefix of the sequence of random variables. Another way to say this is that the events in the tail σ-algebra do not depend on any finite subcollection of the random variables. Events in the tail σ-algebra include things like...

1. $\mathsf{E}_1 = \{\lim_{n\to\infty} X_n \text{ exists}\}$.
2. $\mathsf{E}_2 = \{\sum_{n=1}^{\infty} X_n \text{ converges}\}$.
3. $\mathsf{E}_3 = \{\limsup_{n\to\infty} X_n = +\infty\}$.

There are also random variables that are measurable with respect to the tail σ-algebra, such as

$$\xi = \limsup_{n\to\infty} \frac{1}{n} \sum\nolimits_{i=1}^{n} X_i.$$

This is the limiting upper bound on the running average of the random variables.

E.2.2 The tail σ-algebra is almost trivial

Kolmogorov proved a striking fact about the tail σ-algebra:

Theorem E.15 (Kolmogorov 0–1 law). Let $(X_1, X_2, X_3, \dots)$ be an independent sequence of real random variables. Every event E in the tail σ-algebra $\mathcal{T}$ has probability $\mathbb{P}(\mathsf{E}) = 0$ or $\mathbb{P}(\mathsf{E}) = 1$. In particular, every random variable that is measurable with respect to $\mathcal{T}$ is constant almost surely.

Although this theorem seems to conjure a strong conclusion from minimal assumptions, it can be quite hard to check whether an event $\mathsf{E} \in \mathcal{T}$ has probability zero or probability one. Similarly, it may take serious effort to determine the constant value of a random variable that is measurable with respect to $\mathcal{T}$.

Why is this theorem covered in this appendix? The key idea in the proof is to demonstrate that the tail σ-algebra $\mathcal{T}$ is independent from itself!

Proof. For each index $n \in \mathbb{N}$, we introduce two σ-algebras:

$$\mathcal{F}_n := \sigma(X_1, \dots, X_n) \quad \text{and} \quad \mathcal{T}_n := \sigma(X_{n+1}, X_{n+2}, \dots).$$

First, observe that $\mathcal{F}_n$ is independent from $\mathcal{T}_n$ because of Example E.8. It follows that $\mathcal{F}_n$ is independent from $\mathcal{T}$ because $\mathcal{T} \subseteq \mathcal{T}_n$ for each $n \in \mathbb{N}$.

A fortiori, the σ-algebra $\mathcal{F}_\infty = \sigma(X_n : n \in \mathbb{N})$ is independent from $\mathcal{T}$. To check this point, note that $\mathcal{S} = \bigcup_{n=1}^{\infty} \mathcal{F}_n$ is an intersection-stable system that generates $\mathcal{F}_\infty$. Since $\mathcal{S}$ is independent from $\mathcal{T}$, Corollary E.6 ensures that $\mathcal{F}_\infty$ and $\mathcal{T}$ are independent.

In general, $\mathcal{S}$ is not a σ-algebra!

But $\mathcal{T} \subseteq \mathcal{F}_\infty$. Thus, $\mathcal{T}$ is *independent from* $\mathcal{T}$. In particular, for $\mathsf{E} \in \mathcal{T}$,

$$\mathbb{P}(\mathsf{E}) = \mathbb{P}(\mathsf{E} \cap \mathsf{E}) = \mathbb{P}(\mathsf{E}) \cdot \mathbb{P}(\mathsf{E}).$$

We conclude that $\mathbb{P}(\mathsf{E}) = 0$ or $\mathbb{P}(\mathsf{E}) = 1$. ∎

Exercise E.16 (Tail random variables are almost constant). Prove that a real random variable that is measurable with respect to $\mathcal{T}$ is constant almost surely.

Notes

The discussion of intersection-stable and limit-closed systems is adapted from Williams' book [Wil91, Chap. 1, Chap. 4, App. A.1]. We have also drawn on insights from Pollard [Pol02, Sec. 2.10].

Lecture bibliography

[Pol02] D. Pollard. *A user's guide to measure theoretic probability*. Cambridge University Press, 2002.

[Wil91] D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

VII.

back matter

Bibliography

[AS16] N. Alon and J. H. Spencer. *The probabilistic method*. Fourth. John Wiley & Sons, 2016.

[Aro+18] R. Arora et al. "Understanding Deep Neural Networks with Rectified Linear Units". In: *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net, 2018. URL: https://openreview.net/forum?id=B1J_rgWRW.

[Bil99] P. Billingsley. *Convergence of probability measures*. Second. A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1999. DOI: 10.1002/9780470316962.

[Bil12] P. Billingsley. *Probability and measure*. Anniversary ed. John Wiley & Sons Inc., 2012.

[Bot98] L. Bottou. "Online Algorithms and Stochastic Approximations". In: *Online Learning and Neural Networks*. Revised, Oct. 2012. Cambridge University Press, 1998. URL: http://leon.bottou.org/papers/bottou-98x.

[BLM13] S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities*. A nonasymptotic theory of independence, With a foreword by Michel Ledoux. Oxford University Press, Oxford, 2013. DOI: 10.1093/acprof:oso/9780199535255.001.0001.

[BV04] S. Boyd and L. Vandenberghe. *Convex optimization*. Cambridge University Press, Cambridge, 2004. DOI: 10.1017/CBO9780511804441.

[But03] O. E. Butler. *Kindred*. Beacon, 2003.

[CB90] G. Casella and R. L. Berger. *Statistical inference*. Wadsworth & Brooks/Cole Advanced Books & Software, Pacific Grove, CA, 1990.

[Cha08] S. Chatterjee. "A simple invariance theeorem". Available at https://arxiv.org/abs/math/0508213. 2008.

[Cha06] S. Chatterjee. "A generalization of the Lindeberg principle". In: *Ann. Probab.* 34.6 (2006), pages 2061–2076. DOI: 10.1214/009117906000000575.

[Con] K. Conrad. "Differentiation under the integral sign". Available online. URL: https://kconrad.math.uconn.edu/blurbs/analysis/diffunderint.pdf.

[CT06] T. M. Cover and J. A. Thomas. *Elements of information theory*. Second. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, 2006.

[DZ98] A. Dembo and O. Zeitouni. *Large deviations techniques and applications*. Second. Springer-Verlag, New York, 1998. DOI: 10.1007/978-1-4612-5320-4.

[Dri12] B. K. Driver. "Analysis tools with examples". Available online. 2012. URL: https://mathweb.ucsd.edu/~bdriver/240A-C-2016-17/Lecture_Notes/2012%20Notes/240Lecture_Notes_Ver8.pdf.

[Dud02] R. M. Dudley. *Real analysis and probability*. Revised reprint of the 1989 original. Cambridge University Press, 2002. DOI: 10.1017/CBO9780511755347.

[Dur19] R. Durrett. *Probability—theory and examples*. 5th ed. Cambridge University Press, 2019. DOI: 10.1017/9781108591034.

[Fol99] G. B. Folland. *Real analysis*. Second. Modern techniques and their applications, A Wiley-Interscience Publication. John Wiley & Sons, Inc., New York, 1999.

[Gar07] D. J. H. Garling. *Inequalities: a journey into linear analysis*. Cambridge University Press, 2007. DOI: 10.1017/CBO9780511755217.

[GG84] S. Geman and D. Geman. "Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* PAMI-6.6 (1984), pages 721–741. DOI: 10.1109/TPAMI.1984.4767596.

[Gir89] D. A. Girard. "A fast "Monte Carlo cross-validation" procedure for large least squares problems with noisy data". In: *Numer. Math.* 56.1 (1989), pages 1–23. DOI: 10.1007/BF01395775.

[GS01] G. Grimmett and D. Stirzaker. *Probability and Random Processes*. 3rd ed. Oxford University Press, 2001.

[Gru07] P. M. Gruber. *Convex and discrete geometry*. Springer, Berlin, 2007.

[Has70] W. K. Hastings. "Monte Carlo sampling methods using Markov chains and their applications". In: *Biometrika* 57.1 (1970), pages 97–109. DOI: 10.1093/biomet/57.1.97.

[How+20] S. R. Howard et al. "Time-uniform Chernoff bounds via nonnegative supermartingales". In: *Probab. Surv.* 17 (2020), pages 257–317. DOI: 10.1214/18-PS321.

[How+21] S. R. Howard et al. "Time-uniform, nonparametric, nonasymptotic confidence sequences". In: *Ann. Statist.* 49.2 (2021), pages 1055–1080. DOI: 10.1214/20-aos1991.

[Hut90] M. F. Hutchinson. "A stochastic estimator of the trace of the influence matrix for Laplacian smoothing splines". In: *Comm. Statist. Simulation Comput.* 19.2 (1990), pages 433–450. DOI: 10.1080/03610919008812864.

[Jam+14] K. Jamieson et al. "lil' UCB : An Optimal Exploration Algorithm for Multi-Armed Bandits". In: *Proceedings of Machine Learning Research*. Volume 35. PMLR, 2014, pages 423–439. URL: http://proceedings.mlr.press/v35/jamieson14.html.

[Kal02] O. Kallenberg. *Foundations of modern probability*. Second. Springer-Verlag, New York, 2002. DOI: 10.1007/978-1-4757-4015-8.

[Kec95] A. S. Kechris. *Classical descriptive set theory*. Springer-Verlag, New York, 1995. DOI: 10.1007/978-1-4612-4190-4.

[KM11] S. B. Korada and A. Montanari. "Applications of the Lindeberg principle in communications and statistical learning". In: *IEEE Trans. Inform. Theory* 57.4 (2011), pages 2440–2450. DOI: 10.1109/TIT.2011.2112231.

[Lax02] P. D. Lax. *Functional analysis*. Wiley-Interscience, 2002.

[LC98] E. L. Lehmann and G. Casella. *Theory of point estimation*. Second. Springer-Verlag, New York, 1998.

[Leh04] F. Lehner. "Cumulants in noncommutative probability theory. I. Noncommutative exchangeability systems". In: *Math. Z.* 248.1 (2004), pages 67–100. DOI: 10.1007/s00209-004-0653-0.

[LP17] D. A. Levin and Y. Peres. *Markov chains and mixing times*. Second edition of [MR2466937], With contributions by Elizabeth L. Wilmer, With a chapter on "Coupling from the past" by James G. Propp and David B. Wilson. American Mathematical Society, Providence, RI, 2017. DOI: 10.1090/mbk/107.

[LL01] E. H. Lieb and M. Loss. *Analysis*. 2nd ed. American Mathematical Society, 2001. DOI: 10.1090/gsm/014.

[Mah87] D. Maharam. “From finite to countable additivity”. In: *Portugal. Math.* 44.3 (1987), pages 265–282.

[MT20] P.-G. Martinsson and J. A. Tropp. “Randomized numerical linear algebra: Foundations and algorithms”. In: *Acta Numerica* (2020).

[MR95] R. Motwani and P. Raghavan. *Randomized algorithms*. Cambridge University Press, Cambridge, 1995. DOI: 10.1017/CBO9780511814075.

[NWZ22] J. Nair, A. Wierman, and B. Zwart. *The Fundamentals of Heavy Tails: Properties, Emergence, and Estimation*. 53. Cambridge University Press, 2022.

[NS06] A. Nica and R. Speicher. *Lectures on the combinatorics of free probability*. Cambridge University Press, Cambridge, 2006. DOI: 10.1017/CBO9780511735127.

[O’C71] F. O’Connor. *The Complete Stories*. Farrar, Straus, and Giroux, 1971.

[Owe13] A. B. Owen. “Monte Carlo theory, methods and examples”. Available online. 2013. URL: https://artowen.su.domains/mc/.

[Pol02] D. Pollard. *A user’s guide to measure theoretic probability*. Cambridge University Press, 2002.

[RM51] H. Robbins and S. Monro. “A stochastic approximation method”. In: *Ann. Math. Statistics* 22 (1951), pages 400–407. DOI: 10.1214/aoms/1177729586.

[Roc70] R. T. Rockafellar. *Convex analysis*. Princeton University Press, Princeton, N.J., 1970.

[Roy88] H. L. Royden. *Real analysis*. 3rd ed. Macmillan Publishing Company, New York, 1988.

[Rud76] W. Rudin. *Principles of mathematical analysis*. 3rd ed. McGraw-Hill, 1976.

[Rud91] W. Rudin. *Functional analysis*. Second. McGraw-Hill, Inc., New York, 1991.

[Sch15] R. Schwartz. “The change of variables formula”. Available online. 2015. URL: https://www.math.brown.edu/reschwar/M114/notes8.pdf.

[Shi96] A. N. Shiryaev. *Probability*. Second. Translated from the first (1980) Russian edition by R. P. Boas. Springer-Verlag, New York, 1996. DOI: 10.1007/978-1-4757-2539-1.

[Sim15] B. Simon. *Real analysis*. With a 68 page companion booklet. American Mathematical Society, 2015. DOI: 10.1090/simon/001.

[Spe83] T. P. Speed. “Cumulants and partition lattices”. In: *Austral. J. Statist.* 25.2 (1983), pages 378–388.

[Ste04] J. M. Steele. *The Cauchy-Schwarz master class*. An introduction to the art of mathematical inequalities. Mathematical Association of America / Cambridge Univ. Press, 2004. DOI: 10.1017/CBO9780511817106.

[Sto48] M. H. Stone. “Notes on integration. II”. In: *Proc. Nat. Acad. Sci. U.S.A.* 34 (1948), pages 447–455. DOI: 10.1073/pnas.34.9.447.

[Stu10] A. M. Stuart. “Inverse problems: a Bayesian perspective”. In: *Acta Numer.* 19 (2010), pages 451–559. DOI: 10.1017/S0962492910000061.

[Tal96] M. Talagrand. “A new look at independence”. In: *Ann. Probab.* 24.1 (1996), pages 1–34. DOI: 10.1214/aop/1042644705.

[Taoa] T. Tao. *An alternate approach to the Carathéodory extension theorem*. URL: https://terrytao.wordpress.com/2009/01/03/254a-notes-0a-an-alternate-approach-to-the-caratheodory-extension-theorem/.

[Taob] T. Tao. *Second-moment and entropy methods*. URL: https://terrytao.wordpress.com/2019/11/12/254a-notes-9-second-moment-and-entropy-methods/.

[Tao11] T. Tao. *An introduction to measure theory*. American Mathematical Society, Providence, RI, 2011. DOI: 10.1090/gsm/126.

[Tao16] T. Tao. *Analysis I*. 3rd ed. Springer, 2016. DOI: 10.1007/978-981-10-1789-6.

[Tao19] T. Tao. "Least singular value, circular law, and Lindeberg exchange". In: *Random matrices*. Volume 26. IAS/Park City Math. Ser. Amer. Math. Soc., Providence, RI, 2019, pages 461–498.

[Tie94] L. Tierney. "Markov chains for exploring posterior distributions". In: *Ann. Statist.* 22.4 (1994). With discussion and a rejoinder by the author, pages 1701–1762. DOI: 10.1214/aos/1176325750.

[Tro21] J. A. Tropp. *ACM 217: Probability in High Dimensions*. CMS Lecture Notes 2021-01. Caltech, 2021. DOI: 10.7907/mxr0-c422.

[Tro22] J. A. Tropp. *ACM 204: Matrix Analysis*. CMS Lecture Notes 2022-01. Caltech, 2022. DOI: 10.7907/m421-yb89.

[VW23] A. W. van der Vaart and J. A. Wellner. *Weak convergence and empirical processes—with applications to statistics*. Second. Springer, Cham, [2023] ©2023. DOI: 10.1007/978-3-031-29040-4.

[VH16] R. Van Handel. "Probability in high dimension". APC 550 Lecture Notes, Princeton University. 2016. URL: https://web.math.princeton.edu/~rvan/APC550.pdf.

[Ver18] R. Vershynin. *High-dimensional probability*. An introduction with applications in data science, With a foreword by Sara van de Geer. Cambridge University Press, Cambridge, 2018. DOI: 10.1017/9781108231596.

[Was04] L. Wasserman. *All of statistics*. A concise course in statistical inference. Springer-Verlag, New York, 2004. DOI: 10.1007/978-0-387-21736-9.

[Wil91] D. Williams. *Probability with Martingales*. Cambridge University Press, 1991.

[YY87] A. M. Yaglom and I. M. Yaglom. *Challenging mathematical problems with elementary solutions. Vol. II*. Problems from various branches of mathematics, Translated from the Russian by James McCawley, Jr., Reprint of the 1967 edition. Dover Publications, Inc., New York, 1987.