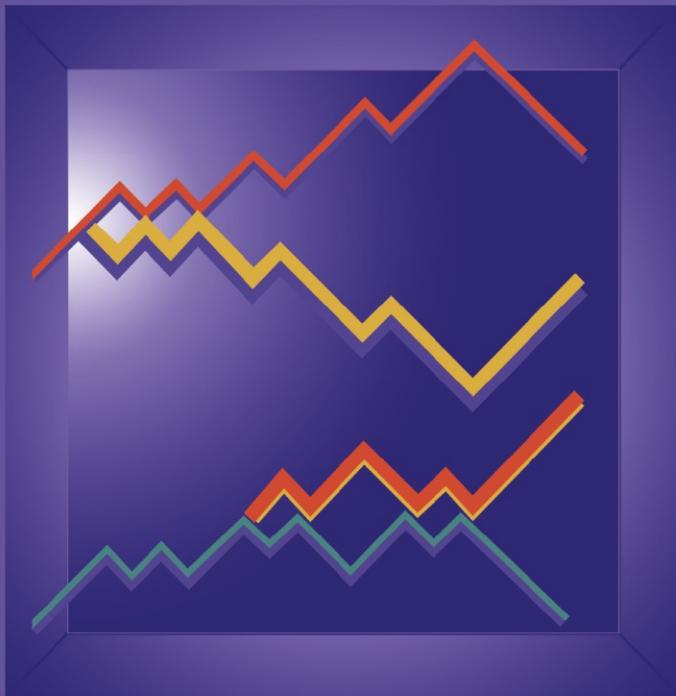




SPINGER TEXTS IN STATISTICS

Alan F. Karr

# Probability



Springer Science+Business Media, LLC

*Springer Texts in Statistics*

---

*Advisors:*

Stephen Fienberg   Ingram Olkin

## *Springer Texts in Statistics*

---

<i>Alfred</i>	Elements of Statistics for the Life and Social Sciences
<i>Berger</i>	An Introduction to Probability and Stochastic Processes
<i>Blom</i>	Probability and Statistics: Theory and Applications
<i>Chow and Teicher</i>	Probability Theory: Independence, Interchangeability, Martingales, Second Edition
<i>Christensen</i>	Plane Answers to Complex Questions: The Theory of Linear Models
<i>Christensen</i>	Linear Models for Multivariate, Time Series, and Spatial Data
<i>Christensen</i>	Log-Linear Models
<i>du Toit, Steyn and Stumpf</i>	Graphical Exploratory Data Analysis
<i>Finkelstein and Levin</i>	Statistics for Lawyers
<i>Jobson</i>	Applied Multivariate Data Analysis, Volume I: Regression and Experimental Design
<i>Jobson</i>	Applied Multivariate Data Analysis, Volume II: Categorical and Multivariate Methods
<i>Kalbfleisch</i>	Probability and Statistical Inference: Volume I: Probability, Second Edition
<i>Kalbfleisch</i>	Probability and Statistical Inference: Volume 2: Statistical Inference, Second Edition
<i>Karr</i>	Probability

*continued at end of book*

Alan F. Karr

# *Probability*



Springer Science+Business Media, LLC

Alan F. Karr  
National Institute of Statistical Sciences  
P.O. Box 14162  
Research Triangle Park, NC 27709-4162 USA

*Editorial Board*

Stephen Fienberg	Ingram Olkin
Office of the Vice President	Department of Statistics
York University	Stanford University
4700 Keele Street	Stanford, CA 94305 USA
Ontario M3J 1P3 Canada	

---

Mathematical Subject Classification (1992): 60-01

---

Cataloging-in-Publication Data is available from the Library of Congress

© 1993 Springer Science+Business Media New York  
Originally published by Springer-Verlag New York, Inc. in 1993  
Softcover reprint of the hardcover 1st edition 1993

All rights reserved. This work may not be translated or copied in whole or in part without the written permission of the publisher  
Springer Science+Business Media, LLC, except for brief excerpts in connection with reviews or scholarly analysis. Use in connection with any form of information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed is forbidden.

The use of general descriptive names, trade names, trademarks, etc., in this publication, even if the former are not especially identified, is not to be taken as a sign that such names, as understood by the Trade Marks and Merchandise Marks Act, may accordingly be used freely by anyone.

Production managed by Karen Phillips; manufacturing supervised by  
Vincent Scelta.

Photocomposed pages prepared using the author's L<sup>A</sup>T<sub>E</sub>X file.

9 8 7 6 5 4 3 2 1

ISBN 978-1-4612-6937-3 ISBN 978-1-4612-0891-4 (eBook)  
DOI 10.1007/978-1-4612-0891-4

# Preface

This book is a text at the introductory graduate level, for use in the one-semester or two-quarter probability course for first-year graduate students that seems ubiquitous in departments of statistics, biostatistics, mathematical sciences, applied mathematics and mathematics. While it is accessible to advanced (“mathematically mature”) undergraduates, it could also serve, with supplementation, for a course on measure-theoretic probability. Students who master this text should be able to read the “hard” books on probability with relative ease, and to proceed to further study in statistics or stochastic processes.

This is a book to teach from. It is not encyclopædic, and may not be suitable for all reference purposes.

Pascal once apologized to a correspondent for having written a long letter, saying that he hadn’t the time to write a short one. I have tried to write a short book, which is quite deliberately incomplete, globally and locally. Many topics, including at least one of everyone’s favorites, are omitted, among them, infinite divisibility, interchangeability, large deviations, ergodic theory and the Markov property. These can be supplied at the discretion and taste of instructors and students, or to suit particular interests.

The major issue in writing a book on probability at this level is what to do about measure theory. To assume that students know it ignores reality. After several years of experimentation, I have concluded that it is intellectually imperative and pedagogically sensible to introduce and use concepts and results from measure theory, but that it is not necessary (and perhaps not even desirable) to develop and prove them individually nor to treat measure theory as a subject in its own right. Not everyone, of course, will agree. Students should be made aware of the role of probability within the broader context of measure theory, but to dismiss probability as the special case of a “space with total measure one” is neither honest nor helpful.

Thus, in this book, classes of sets, monotone class theorems, probability measures, measurability of random variables, and similar concepts, are given proper definitions and used at full strength, but many proofs are downgraded or omitted altogether. Other concepts are treated in isolation,

even though they are instances of more general theories. For example, expectation is developed from scratch, with most of the main theorems rather proved carefully, without taking the point of view that it is a special case of Lebesgue integration.

Consequently, no background in real variables (at the graduate level) on students' part is assumed. However, facility with "real analysis" at the level of Apostol's *Mathematical Analysis* or Rudin's *Principles of Mathematical Analysis*, primarily in regard to convergence in  $\mathbb{R}$  and continuity and convergence of functions, is necessary. Students sometimes take such a course simultaneously with probability, and while this is disadvantageous, it is feasible. Perhaps the most important thing is to be aware that interchanges of limiting processes, among which are differentiation, integration and summation, require justification in every instance, even if it is often not provided in this book.

Another requirement is the ability to recognize, devise and write down coherent, concise proofs.

A prior course in probability (at the calculus level) is not necessary, although clearly it would help.

In order to accommodate a streamlined presentation without sacrificing all of the details, I have adopted a series of devices. Every chapter concludes with a section entitled "Complements," in which I have placed additional topics and results pertinent to the main material of the chapter, but not (in my opinion and at this level) absolutely central to it. Examples include Fubini's theorem, integration with respect to Lebesgue measure, conditional expectation given  $\sigma$ -algebras and reversed martingales. The book is usable if these sections are ignored, but then some of the richness of the subject is lost.

In the main text, and in the "Complements" sections as well, some of the more arcane details are relegated to "Technical Asides," which are generally short enough to be skimmed.

Educated in part as an engineer and a long-time collaborator with engineers, I have an affinity for concreteness, which has generally led me to choose illuminating specificity over generalization beyond recognition. For example, combinatorial probability is not treated in generality, but instead illustrated via rather detailed consideration of occupancy problems, allowing one to focus in a meaningful manner on computational techniques, construction of random variables and asymptotics. Similarly, Bernoulli and Poisson processes are used to epitomize construction of random variables from other random variables (and stochastic processes from other stochastic processes) and the identification of independence relationships among random variables. And yet again, convergence is exemplified by limit theorems for sums of Bernoulli-distributed random variables, even though these theorems are subsumed by subsequent results. Many examples and exercises are "elementary," even to the point that they would not be out of place in a calculus-level probability book, in order to partially bridge the gap between

honest theory and real application. It will be apparent, however, that the bridging is incomplete: there is no serious treatment of applications here, to my chagrin. Were the book longer, there would be some substantive applications, not more theory.

Throughout this book, I take the point of view that random variables are the most important objects in probability. Most students study probability in order to work with statistics or stochastic processes, both of which deal entirely with properties of random variables, and even in probability itself, random variables are arguably of paramount importance. Each chapter, other than the first, focuses on a specific aspect of random variables: closure, distributional and transformation properties, expectation, convergence, transforms, limit theorems, conditioning and the martingale property. In some cases, this point of view leads to topic orderings that may seem unnatural: for example, independent random variables are introduced before independent events. To some extent, this is a matter of taste, but I have never found sets (events) more intuitive or easier to manipulate than functions (random variables).

Perhaps the most distinctive aspect of the book is the treatment of conditioning. Conditional expectation is developed from the geometric/engineering point of view of minimum mean squared error prediction of unobservable random variables. This seems more effective than the obscure and cumbersome “equal expectations over sets in some  $\sigma$ -algebra” criterion ordinarily used. The prediction formulation has the further advantage of making many properties natural and intuitive. Major emphasis is placed upon conditioning given a finite family of random variables, which seems appropriate to the level of the book.

Along the same lines, martingales are defined relative to a base stochastic process rather than a filtration (although the latter are presented as a complement). This is adequate to encompass essentially all important examples, and maintains the focus on random variables.

Inevitably, mathematical rigor is sacrificed from time to time. This happens mainly in connection with sets of probability or Lebesgue measure zero, and in connection with the sense in which various integrals are taken. Integrals on the real line, for example, are often (either explicitly or by implication) taken as Riemann integrals, yet manipulated as Lebesgue integrals. Excessive pedanticism, however, seems a less desirable alternative. In any case, I have tried to be honest about this when it happens, so that readers are at least aware that there is an issue, even if it is left unresolved.

I have been somewhat more careful about “operation interchange” issues, especially insofar as they involve the monotone and dominated convergence theorems and Fubini’s theorem. These kinds of questions are the *raison d’être* for studying analysis in order to understand probability.

The exercises are an integral component of the book: one cannot learn mathematics without doing it. In a book intended to be concise, there is

great temptation to “sneak” additional concepts into the exercises. Only in relatively few cases have I acceded to this: for example, several exercises develop limit theorems for maxima of independent random variables. Not all of the exercises are easy, but relatively few, I hope, are impossible, and none is meant to be.

A lot of time has elapsed in the course of writing this book, whose genesis dates to 1980. At that point, I had taught the first-year graduate probability course at Johns Hopkins several times, and had searched each year, unsuccessfully, for a suitable text. In 1984 and 1985, a preliminary set of notes was developed and used once, but then lay fallow, to be resurrected in 1991, when I taught the probability course once again.

My students and former colleagues at Hopkins, especially Robert Serfling and John Wierman, have been unreservedly generous and extremely helpful throughout the process. Professor Wierman, along with Robert Smythe of George Washington University, taught from the penultimate version of the manuscript in the fall of 1992. Their and their students’ insightful comments and criticisms helped shape the final version.

Stamatis Cambanis and Gordon Simons, now my colleagues at the University of North Carolina at Chapel Hill, contributed a number of superb problems. Martin Gilchrist and others at Springer-Verlag made the editorial process a pleasure.

Finally, I wish to thank Senora DeCosta for her assistance and support during the latter stages of the project. She knows how much she helped.

*Alan F. Karr*

# Reader's Guide

The book is basically intended to be read linearly, as the following chart of chapter dependences indicates. The first five chapters are the irreducible core of the subject, and cannot be omitted. However, Chapter 8, on conditioning and the geometry of  $L^2$ , requires neither Chapter 6 nor Chapter 7. Chapter 9, on martingales, does not require Chapter 7 in a strict sense, but otherwise its impact is vitiated.

The prelude should be read, at least initially, more for the questions and ideas raised there, and for some sense of the kinds of tools with which they are addressed, than for the precise form of the results. Reading it again as a postlude may be equally enlightening.

An important concept raised in the prelude, and one deserving of more emphasis than it is commonly accorded, is *exploitation of special structure*. Too often, students (and others) seem to view specificity as a last resort, and prefer unthinking application of formulas or theorems to insight. The (to some, excessive) concreteness of the book is an attempt to counteract this tendency.

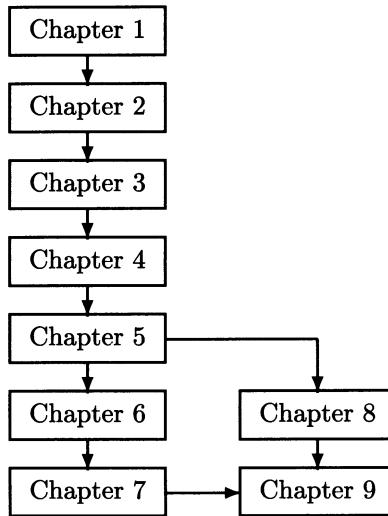
More advanced or arcane portions of the material can be skimmed, at least on first reading. These are comprised of

1. A series of Technical Asides, which either elaborate upon points in the main text or address “technical” issues such as measurability or the sense in which various integrals are to be taken.
2. A section entitled “Complements” that concludes each chapter. The material relegated to these sections tends to be more complicated, and more important, than that in the Technical Asides.

Readers, however, should not presume *carte blanche* to ignore the material they contain, much of which is interesting, important and beautiful.

Even those not obliged to work exercises as part of a course should look at them. Mathematics is better learned actively than passively. Like the complements and technical asides, the exercises convey the richness of the subject.

There are relatively few perverse notations, usages or terminologies in the book. Terms such as “positive” and “increasing” are used in the



weak sense (to mean, for example, “nonnegative” and “nondecreasing”), unless qualified with “strictly.” With no pun intended, this is a more positive and sensible approach than littering the book with “non-”s. We use  $a/bc$  to mean  $a/(bc)$ . For a function  $f$  and a subset  $B$  of its domain,  $\arg \max_{x \in B} f(x)$  denotes any maximizer  $x^*$  of  $f$  over  $B$ :  $f(x^*) \geq f(x)$  for all  $x \in B$ .

I have attempted to stem the proliferation of equation numbers by referring insofar as reasonable to results rather than equations. This is especially effective for terms and results with names attached to them; as an aid, Appendix B is an index of these. Along the same lines, Appendix A is a guide to the notation.

Finally, ■ is used to indicate the ends of proofs, and □ to mark the ends of definitions, examples, technical asides and unproved results.

# Contents

<b>Preface</b>	v
<b>Reader's Guide</b>	ix
<b>List of Figures</b>	xix
<b>List of Tables</b>	xxi
<b>Prelude: Random Walks</b>	1
The Model . . . . .	1
Random variables . . . . .	1
Probability . . . . .	1
First calculations . . . . .	3
Issues and Approaches . . . . .	4
Issues . . . . .	4
Approaches . . . . .	4
Tools . . . . .	5
Functionals of the Random Walk . . . . .	6
Times of returns to the origin . . . . .	7
Numbers of returns to the origin . . . . .	9
First passage times . . . . .	9
Maxima . . . . .	11
Time spent positive . . . . .	11
Limit Theorems . . . . .	12
Summary . . . . .	14
<b>1 Probability</b>	15
1.1 Random Experiments and Sample Spaces . . . . .	16
1.1.1 Random experiments . . . . .	16
1.1.2 Sample spaces . . . . .	16
1.2 Events and Classes of Sets . . . . .	17
1.2.1 Events . . . . .	18

1.2.2	Basic set operations . . . . .	18
1.2.3	Indicator functions . . . . .	18
1.2.4	Operations on sequences of sets . . . . .	20
1.2.5	Classes of sets closed under set operations . . . . .	21
1.2.6	Generated classes . . . . .	22
1.2.7	The monotone class theorem . . . . .	23
1.2.8	Events, bis . . . . .	23
1.3	Probabilities and Probability Spaces . . . . .	23
1.3.1	Probability . . . . .	23
1.3.2	Elementary properties . . . . .	25
1.3.3	More advanced properties . . . . .	26
1.3.4	Almost sure and null events . . . . .	28
1.3.5	Uniqueness . . . . .	28
1.4	Probabilities on $\mathbb{R}$ . . . . .	29
1.4.1	Distribution functions . . . . .	29
1.4.2	Discrete probabilities . . . . .	32
1.4.3	Absolutely continuous probabilities . . . . .	33
1.4.4	Mixed distributions . . . . .	34
1.5	Conditional Probability Given a Set . . . . .	35
1.6	Complements . . . . .	36
1.6.1	The extended real numbers . . . . .	36
1.6.2	Measures . . . . .	37
1.6.3	Lebesgue measure . . . . .	38
1.6.4	Singular probabilities on $\mathbb{R}$ . . . . .	38
1.6.5	Representation of probabilities on $\mathbb{R}$ . . . . .	39
1.7	Exercises . . . . .	40
<b>2</b>	<b>Random Variables</b> . . . . .	<b>43</b>
2.1	Fundamentals . . . . .	44
2.1.1	Random variables . . . . .	44
2.1.2	Random vectors . . . . .	45
2.1.3	Stochastic processes . . . . .	45
2.1.4	Complex-valued random variables . . . . .	45
2.1.5	The $\sigma$ -algebra generated by a random variable . . . . .	46
2.1.6	Simplified criteria . . . . .	47
2.2	Combining Random Variables . . . . .	47
2.2.1	Algebraic operations . . . . .	47
2.2.2	Limiting operations . . . . .	49
2.2.3	Transformations . . . . .	50
2.2.4	Approximation of positive random variables . . . . .	50
2.2.5	Monotone class theorems . . . . .	51
2.3	Distributions and Distribution Functions . . . . .	51
2.3.1	Random variables . . . . .	52
2.3.2	Random vectors . . . . .	53
2.4	Key Random Variables and Distributions . . . . .	54

<b>CONTENTS</b>	<b>xiii</b>
-----------------	-------------

2.4.1 Discrete random variables . . . . .	54
2.4.2 Absolutely continuous random variables . . . . .	56
2.4.3 Random vectors . . . . .	59
2.5 Transformation Theory . . . . .	60
2.5.1 Random variables . . . . .	60
2.5.2 Random vectors . . . . .	62
2.6 Random Variables with Prescribed Distributions . . . . .	63
2.6.1 Individual random variables . . . . .	63
2.6.2 Random vectors . . . . .	65
2.6.3 Sequences of random variables . . . . .	65
2.7 Complements . . . . .	65
2.7.1 Measurability with respect to sub- $\sigma$ -algebras . . . . .	65
2.7.2 Borel measurable functions . . . . .	66
2.8 Exercises . . . . .	67
<b>3 Independence</b>	<b>71</b>
3.1 Independent Random Variables . . . . .	71
3.1.1 Fundamentals . . . . .	71
3.1.2 Criteria for independence . . . . .	71
3.1.3 Examples . . . . .	75
3.2 Functions of Independent Random Variables . . . . .	76
3.2.1 Transformation properties . . . . .	76
3.2.2 Sums of independent random variables . . . . .	77
3.3 Constructing Independent Random Variables . . . . .	79
3.3.1 Finite families . . . . .	79
3.3.2 Sequences . . . . .	79
3.4 Independent Events . . . . .	80
3.5 Occupancy Models . . . . .	82
3.5.1 Four occupancy models . . . . .	83
3.5.2 Occupancy numbers . . . . .	84
3.5.3 Asymptotics . . . . .	86
3.6 Bernoulli and Poisson Processes . . . . .	88
3.6.1 Bernoulli processes . . . . .	88
3.6.2 Poisson processes . . . . .	91
3.7 Complements . . . . .	94
3.7.1 Independent $\sigma$ -algebras . . . . .	94
3.7.2 Products of probability spaces . . . . .	94
3.8 Exercises . . . . .	95
<b>4 Expectation</b>	<b>101</b>
4.1 Definition and Fundamental Properties . . . . .	102
4.1.1 Simple random variables . . . . .	102
4.1.2 Positive random variables . . . . .	103
4.1.3 Integrable random variables . . . . .	107
4.1.4 Complex-valued random variables . . . . .	109

4.2	Integrals with respect to Distribution Functions . . . . .	110
4.2.1	Generalities . . . . .	110
4.2.2	Discrete distribution functions . . . . .	111
4.2.3	Absolutely continuous distribution functions . . . . .	112
4.2.4	Mixed distribution functions . . . . .	112
4.3	Computation of Expectations . . . . .	113
4.3.1	Positive random variables . . . . .	113
4.3.2	Integrable random variables . . . . .	114
4.3.3	Functions of random variables . . . . .	114
4.3.4	Functions of random vectors . . . . .	115
4.3.5	Functions of independent random variables . . . . .	117
4.3.6	Sums of independent random variables . . . . .	118
4.4	$L^p$ Spaces and Inequalities . . . . .	119
4.4.1	$L^p$ spaces . . . . .	119
4.4.2	Key inequalities . . . . .	119
4.5	Moments . . . . .	123
4.5.1	Moments of random variables . . . . .	123
4.5.2	Variance and standard deviation . . . . .	124
4.5.3	Covariance and correlation . . . . .	125
4.5.4	Moments of random vectors . . . . .	126
4.5.5	Multivariate normal distributions . . . . .	126
4.6	Complements . . . . .	127
4.6.1	Integration with respect to Lebesgue measure . . . . .	127
4.6.2	Expectation for product probabilities . . . . .	128
4.7	Exercises . . . . .	130
<b>5</b>	<b>Convergence of Sequences of Random Variables</b>	<b>135</b>
5.1	Modes of Convergence . . . . .	135
5.1.1	Convergence of random variables as functions . . . . .	135
5.1.2	Convergence of distribution functions . . . . .	136
5.1.3	Alternative criteria . . . . .	137
5.2	Relationships Among the Modes . . . . .	140
5.2.1	Implications always valid . . . . .	140
5.2.2	Counterexamples . . . . .	141
5.2.3	Implications of restricted validity . . . . .	142
5.2.4	Implications involving subsequences . . . . .	144
5.3	Convergence under Transformations . . . . .	145
5.3.1	Algebraic operations . . . . .	145
5.3.2	Continuous mappings . . . . .	148
5.4	Convergence of Random Vectors . . . . .	149
5.4.1	Convergence of random vectors as functions . . . . .	149
5.4.2	Convergence in distribution . . . . .	149
5.4.3	Continuous mappings . . . . .	150
5.5	Limit Theorems for Bernoulli Summands . . . . .	150
5.5.1	Laws of large numbers . . . . .	151

5.5.2	Central limit theorems . . . . .	152
5.5.3	The Poisson limit theorem . . . . .	155
5.5.4	Approximation of continuous functions . . . . .	156
5.6	Complements . . . . .	157
5.6.1	$L^p$ Convergence of random variables . . . . .	157
5.7	Exercises . . . . .	158
<b>6</b>	<b>Characteristic Functions</b>	<b>163</b>
6.1	Definition and Basic Properties . . . . .	163
6.1.1	Fundamentals . . . . .	163
6.1.2	Elementary properties . . . . .	164
6.2	Inversion and Uniqueness Theorems . . . . .	166
6.2.1	The inversion theorem . . . . .	166
6.2.2	The uniqueness theorem . . . . .	167
6.2.3	Specialized inversion theorems . . . . .	167
6.3	Moments and Taylor Expansions . . . . .	169
6.3.1	Calculation of moments known to exist . . . . .	169
6.3.2	Establishing existence of moments . . . . .	170
6.3.3	Taylor expansions of characteristic functions . . . . .	171
6.4	Continuity Theorems and Applications . . . . .	171
6.4.1	Convergence in distribution . . . . .	171
6.4.2	The Lévy continuity theorem . . . . .	172
6.4.3	Application to classical limit theorems . . . . .	173
6.5	Other Transforms . . . . .	174
6.5.1	Characteristic functions of random vectors . . . . .	175
6.5.2	Laplace transforms . . . . .	176
6.5.3	Moment generating functions . . . . .	177
6.5.4	Generating functions . . . . .	177
6.6	Complements . . . . .	178
6.6.1	Helly's theorem . . . . .	178
6.7	Exercises . . . . .	179
<b>7</b>	<b>Classical Limit Theorems</b>	<b>183</b>
7.1	Series of Independent Random Variables . . . . .	183
7.1.1	Kolmogorov's inequality . . . . .	183
7.1.2	The three series theorem . . . . .	185
7.2	The Strong Law of Large Numbers . . . . .	187
7.3	The Central Limit Theorem . . . . .	190
7.3.1	The Lyapunov condition . . . . .	191
7.3.2	The Lindeberg condition . . . . .	192
7.4	The Law of the Iterated Logarithm . . . . .	196
7.4.1	Normally distributed summands . . . . .	197
7.4.2	More general versions . . . . .	200
7.5	Applications of the Limit Theorems . . . . .	200
7.5.1	Monte Carlo integration . . . . .	201

7.5.2	Maximum likelihood estimation . . . . .	201
7.5.3	Empirical distribution functions . . . . .	205
7.5.4	Random sums of independent random variables . . . . .	207
7.5.5	Renewal processes . . . . .	208
7.6	Complements . . . . .	210
7.6.1	The Berry-Esséen theorem . . . . .	210
7.7	Exercises . . . . .	212
<b>8</b>	<b>Prediction and Conditional Expectation</b>	<b>217</b>
8.1	Prediction in $L^2$ . . . . .	218
8.1.1	The inner product and norm . . . . .	218
8.1.2	$L^2$ as metric space . . . . .	219
8.1.3	Orthogonality and orthonormality . . . . .	221
8.1.4	The orthogonal decomposition theorem . . . . .	222
8.1.5	Computation of MMSE predictors . . . . .	224
8.1.6	Linear prediction . . . . .	224
8.2	Conditional Expectation Given a Finite Set of Random Variables . . . . .	225
8.2.1	Basics . . . . .	225
8.2.2	Examples . . . . .	226
8.2.3	Conditional probability . . . . .	227
8.3	Conditional Expectation for $X \in L^2$ . . . . .	227
8.3.1	Conditional expectation as MMSE prediction . . . . .	227
8.3.2	Properties of conditional expectation . . . . .	228
8.4	Positive and Integrable Random Variables . . . . .	230
8.5	Conditional Distributions . . . . .	233
8.5.1	Generalities . . . . .	233
8.5.2	Discrete random variables . . . . .	234
8.5.3	Absolutely continuous random variables . . . . .	235
8.6	Computational Techniques . . . . .	236
8.6.1	General results . . . . .	236
8.6.2	Special cases . . . . .	238
8.7	Complements . . . . .	238
8.7.1	Mixed conditional distributions . . . . .	238
8.7.2	Conditional expectation given a $\sigma$ -algebra . . . . .	239
8.8	Exercises . . . . .	239
<b>9</b>	<b>Martingales</b>	<b>243</b>
9.1	Fundamentals . . . . .	243
9.1.1	Definitions . . . . .	243
9.1.2	Examples . . . . .	245
9.1.3	Compositions and transformations . . . . .	247
9.2	Stopping Times . . . . .	248
9.3	Optional Sampling Theorems . . . . .	250
9.3.1	Optional sampling theorems for martingales . . . . .	251

9.3.2 Applications of optional sampling theorems . . . . .	253
9.4 Martingale Convergence Theorems . . . . .	255
9.4.1 Upcrossings and almost sure convergence . . . . .	255
9.4.2 Almost sure convergence of submartingales . . . . .	256
9.4.3 Almost sure convergence of martingales . . . . .	258
9.4.4 Uniformly integrable martingales . . . . .	259
9.5 Applications of Convergence Theorems . . . . .	260
9.5.1 The Radon-Nikodym theorem . . . . .	260
9.5.2 Zero-one laws . . . . .	262
9.5.3 Likelihood ratios . . . . .	264
9.6 Complements . . . . .	264
9.6.1 Conditioning on $Y_0, \dots, Y_T$ . . . . .	264
9.6.2 Martingales with respect to filtrations . . . . .	267
9.6.3 Reversed martingales . . . . .	267
9.7 Exercises . . . . .	268
<b>A Notation</b>	<b>271</b>
<b>B Named Objects</b>	<b>275</b>
<b>Bibliography</b>	<b>277</b>
<b>Index</b>	<b>279</b>

# List of Figures

0.1	Two Realizations of a Random Walk . . . . .	2
0.2	The First Reflection Principle . . . . .	8
0.3	The Second Reflection Principle . . . . .	10
1.1	The Basic Set Operations . . . . .	19
1.2	The Poisson Distribution with $\lambda = 1$ . . . . .	32
1.3	The Exponential Distribution with $\lambda = 1$ . . . . .	33
1.4	The Mixed Distribution of Example 1.42, with $\lambda = 1$ . . . . .	35
2.1	The Standard Normal Density Function . . . . .	57
2.2	Some Gamma Density Functions ( $\lambda = 1$ ) . . . . .	58
2.3	A Distribution Function and its Inverse . . . . .	64
4.1	The Proof of Young's Inequality . . . . .	119
5.1	Implications Among Forms of Convergence . . . . .	140
5.2	Bernstein Polynomials for $f(p) = e^{-p} \cos 3p $ . . . . .	156
8.1	Orthogonal Decompositions and Projections . . . . .	222
9.1	Upcrossings for a Sequence $x_0, \dots, x_{20}$ . . . . .	256

# List of Tables

1.1	Probabilities in Terms of Distribution Functions . . . . .	30
2.1	Discrete Distributions . . . . .	56
2.2	Absolutely Continuous Distributions . . . . .	59
2.3	Notation for Particular Distributions . . . . .	60
4.1	Computational Formulas for $E[g(X)]$ . . . . .	115
4.2	The Key Inequalities . . . . .	123
4.3	Moments of Key Distributions . . . . .	125
5.1	Definitions of Convergence for Random Variables . . . . .	137
6.1	Characteristic Functions of Key Distributions . . . . .	164

# Prelude: Random Walks

Consider the random experiment of observing a particle moving randomly on the set  $\mathbb{Z} = \{\dots, -1, 0, 1, \dots\}$  of integers. It begins at the origin at time 0 and at each time  $1, 2, \dots$  thereafter, moves, with equal likelihood, either one step up or one step down. This motion is termed a *random walk*, and might be interpreted, for example, as the path followed by an atom in a gas moving under the influence of collisions with other atoms.

How might such an experiment be modeled and analyzed? What issues are raised, and what tools are needed to resolve them?

## The Model

Let us first introduce terminology and notation. A *realization* of the random walk is an infinite sequence  $\omega = (\omega_1, \omega_2, \dots)$ , where  $\omega_n = \pm 1$  is the step at time  $n$ . Repetitions of the experiment lead to different realizations, which are envisioned to be drawn in a random manner from the set  $\Omega$  of all possible realizations, which is termed the *sample space* of the experiment. Two possible realizations of the first one hundred steps of the path are illustrated in Figure 0.1.

## Random variables

The steps are functions on  $\Omega$ , and we denote the  $n$ th step by  $Y_n(\omega) \stackrel{\text{def}}{=} \omega_n$ . In faded terminology, these are (dependent) variables, and since their values are random, we call them *random variables*. Thus, a random variable is a *function on the sample space*.

From these steps, we may define other random variables. For example, the position of the random walk after  $n$  steps is  $X_n(\omega) = \sum_{i=1}^n Y_i(\omega)$ . Still other random variables will be defined momentarily.

## Probability

The randomness that underlies the random walk is specified by *probability*, which measures the likelihoods of sets of outcomes, termed *events*. An event

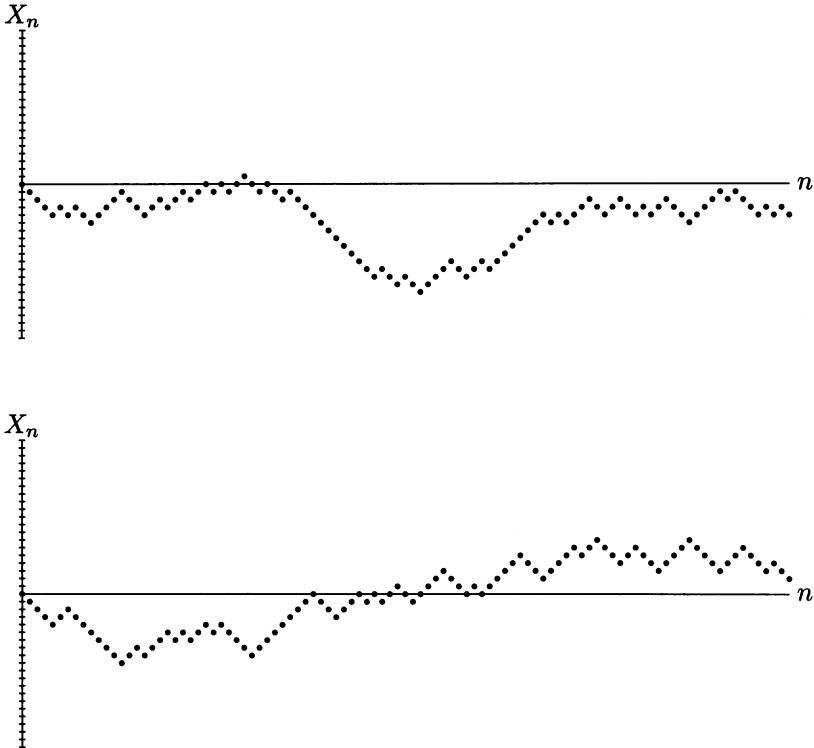


Figure 0.1. Two Realizations of a Random Walk

$A \subseteq \Omega$  has probability denoted by  $P(A)$ . The most important characteristic of probability is that it is *additive*: if the events  $A_1, \dots, A_k$  are pairwise disjoint ( $A_i \cap A_j = \emptyset$  whenever  $i \neq j$ ), then the probability of their union is the sum of their individual probabilities. In symbols,

$$P\left(\bigcup_{i=1}^k A_i\right) = \sum_{i=1}^k P(A_i). \quad (0.1)$$

The mathematics actually requires *countable additivity*, a stronger version of (0.1) for sequences of disjoint events: if  $A_1, A_2, \dots$  are disjoint, then

$$P\left(\bigcup_{i=1}^{\infty} A_i\right) = \sum_{i=1}^{\infty} P(A_i). \quad (0.2)$$

As an invocation of “randomness,” we stipulate that, for each  $n$ , all  $2^n$  possible values of the random variables  $Y_1, \dots, Y_n$  are equally likely, so that

for all choices  $y_1, \dots, y_n$  of  $\pm 1$ ,

$$P\{Y_1 = y_1, \dots, Y_n = y_n\} = \left(\frac{1}{2}\right)^n. \quad (0.3)$$

In particular,  $P\{Y_i = 1\} = P\{Y_i = -1\} = 1/2$ , so that each step is, indeed, equally likely to be up or down. Also, since these probabilities are the same for all  $i$ , the  $Y_i$  are *identically distributed*.

The steps have the crucial probabilistic property that they are *independent random variables*, in the sense that knowledge of the values of some of them is of no help in the predicting values of others. Mathematically, this takes the form of the multiplicative relationship

$$P\{Y_1 = y_1, \dots, Y_n = y_n\} = \prod_{j=1}^n P\{Y_j = y_j\}. \quad (0.4)$$

## First calculations

From the probabilistic structure of the steps, we infer that of the positions.

First, we compute  $P\{X_n = k\}$  for fixed  $n$  and each  $k$ . Since  $X_0 \equiv 0$  and steps are of size  $\pm 1$ ,  $|X_n|$  cannot exceed  $n$ , so we suppose that  $k \in \{-n, \dots, 0, \dots, n\}$ . Also, the position is even at even times and odd at odd times. Assuming, therefore, that  $(n+k)/2$  is an integer,  $X_n = k$  if  $(n+k)/2$  of the steps  $Y_1, \dots, Y_n$  are 1 and the remainder are -1, in other words, if there is a subset  $K$  of  $\{1, \dots, n\}$ , containing  $(n+k)/2$  elements, such that  $Y_j = 1$  if and only if  $j \in K$ . Thus, with  $|K|$  the number of elements in  $K$ ,

$$\begin{aligned} P\{X_n = k\} &= P\left(\bigcup_{|K|=(n+k)/2} \{Y_j = 1 \iff j \in K\}\right) \quad (0.5) \\ &= \sum_{|K|=(n+k)/2} P\{Y_j = 1 \iff j \in K\} \end{aligned}$$

[by (0.1): as  $K$  varies with  $|K| = (n+k)/2$ , the events  $\{Y_j = 1 \iff j \in K\}$  are disjoint]

$$\begin{aligned} &= \sum_{|K|=(n+k)/2} \left(\frac{1}{2}\right)^n \\ &= \binom{n}{(n+k)/2} \left(\frac{1}{2}\right)^n, \end{aligned}$$

where the third equality is by (0.3) and the fourth is by the property that there are  $\binom{n}{j} = n!/j!(n-j)!$  size- $j$  subsets of a set of size  $n$ .

More generally, for  $B$  a subset of  $\mathbb{R} = (-\infty, \infty)$ , by countable additivity of  $P$ , the event  $\{X_n \in B\} \stackrel{\text{def}}{=} \{\omega: X_n(\omega) \in B\}$  has probability

$$P\{X_n \in B\} = P(\bigcup_{k \in B \cap \mathbb{Z}} \{X_n = k\}) = \sum_{k \in B \cap \mathbb{Z}} P\{X_n = k\}. \quad (0.6)$$

## Issues and Approaches

Having specified the probabilistic structure of the random walk, we proceed to further aspects of its analysis. Before discussing details, we consider some general aspects, which apply to much of probability.

### Issues

Among the major issues regarding the random walk are:

1. **Computations.** Quantities whose computation is of interest are probabilities other than those specified by assumption in (0.3), some of which have been considered already, and expectations (average values of random variables, defined below). Not only exact results, but also approximations and inequalities, are important.
2. **Functionals.** From the step process  $(Y_i)$  or the position process  $(X_n)$ , one can define, as we do below, many other random variables with important physical interpretations. Their analysis is central to full understanding of the random walk.
3. **Structure.** Identification of probabilistic structure, particularly independence, is a key step in developing properties.
4. **Asymptotics.** A characteristic feature of probability is that random variation “in the small” is accompanied by regularity “in the large,” which is elucidated by means of limit theorems.

### Approaches

Addressing these issues depends, above all, on *exploitation of special structure*. For example, a crucial physical property of the random walk is that it is “continuous.” Because its jumps are all of size  $\pm 1$ , it cannot move from one level to another without passing through all values between.

Similarly, the key to computing probabilities for the random walk is the specific property that for a given value of  $n$ , all  $2^n$  length- $n$  paths are equally likely, so that the probability of an event determined by  $Y_1, \dots, Y_n$  is the number of paths in that event divided by  $2^n$ . In particular, two events containing the same number of paths have the same probability, which allows the probability of one event to be determined by showing

that outcomes belonging to it are in one-to-one correspondence with those of an event of known probability. In many cases, this correspondence is established geometrically, via reasoning known generically as the *reflection principle*, since the geometry involves reflections of paths.

In addition, the random walk is symmetric: for each  $n$  and  $k$ ,

$$P\{X_n = k\} = P\{X_n = -k\}.$$

Yet another key structural element is that a random walk is temporally and spatially homogeneous: for fixed  $k$ , the process  $\tilde{X}_n = X_{k+n} - X_k$  has the same probabilistic structure as  $(X_n)$  itself.

## Tools

Three concepts are used throughout probability.

### Independence

Independence can be defined for any family of random variables. By analogy with (0.4), random variables  $Z_1, \dots, Z_n$  are *independent* if

$$P\{Z_1 \in B_1, \dots, Z_n \in B_n\} = \prod_{i=1}^n P\{Z_i \in B_i\} \quad (0.7)$$

for all appropriate subsets  $B_1, \dots, B_n$  of  $\mathbb{R}$ . The interpretation of independence, as before, is *absence of probabilistic interaction*.

### Expectation

A random variable has associated to it an average value, its *expectation*, which is the sum (or, in other cases, integral) of its possible values, weighted by the probabilities with which they are assumed. For example, the expectation of an integer-valued random variable  $Z$  is

$$E[Z] = \sum_{k \in \mathbb{Z}} k P\{Z = k\}. \quad (0.8)$$

For each  $i$ ,

$$E[Y_i] = 1 \times P\{Y_1 = 1\} + (-1) \times P\{Y_i = -1\} = 0,$$

so that, consistent with symmetry, the average step is zero. Additivity of probability translates to linearity of expectation, so that for each  $n$ ,

$$E[X_n] = E\left[\sum_{i=1}^n Y_i\right] = \sum_{i=1}^n E[Y_i] = 0.$$

## Conditioning

In probability, one often deals with phenomena, such as random walks, that evolve sequentially, or admit other forms of partial observation. This leads to the issue of how to revise probabilities in light of the knowledge that some event has occurred, or on the basis of having observed some set of random variables. For the random walk, for example, we have seen that  $P\{X_{2n} = 0\} = \binom{2n}{n}/2^{2n}$ , whereas if we knew that  $X_{2n-1}$  were 1, then clearly the probability that  $X_{2n} = 0$  should be one-half.

The tool used to effect these revisions is conditional probability. For events  $A$  and  $B$ , with  $P(A) > 0$ , the *conditional probability of  $B$  given  $A$*  is

$$P(B|A) = \frac{P(B \cap A)}{P(A)}, \quad (0.9)$$

in other words, that “portion” of  $B$  also lying within  $A$ . We interpret  $P(B|A)$  as the probability of  $B$  given the knowledge that  $A$  has occurred. An alternative expression is

$$P(B \cap A) = P(B|A)P(A). \quad (0.10)$$

For the random walk, with  $A = \{X_{2n-1} = 1\}$  and  $B = \{X_{2n} = 0\}$ , (0.9) gives

$$\begin{aligned} P\{X_{2n} = 0 | X_{2n-1} = 1\} &= \frac{P\{X_{2n-1} = 1, X_{2n} = 0\}}{P\{X_{2n-1} = 1\}} \\ &= \frac{P\{X_{2n-1} = 1, Y_{2n} = -1\}}{P\{X_{2n-1} = 1\}} \\ &= \frac{P\{X_{2n-1} = 1\}P\{Y_{2n} = -1\}}{P\{X_{2n-1} = 1\}} \\ &= 1/2. \end{aligned}$$

The third equality uses the property that  $Y_{2n}$  and  $X_{2n-1}$ , which depends only on  $Y_1, \dots, Y_{2n-1}$ , are independent random variables.

Conditioning extends to *conditional expectations*, which are sums (more generally, integrals) of values of random variables weighted by conditional probabilities rather than ordinary probabilities, and to *conditioning given random variables*, instead of events. Conditional expectations given random variables are minimum error predictors of unobservable random variables on the basis of the values of observable random variables.

## Functionals of the Random Walk

We now examine several functionals of random walks.

## Times of returns to the origin

The random variable

$$\mathbf{1}(X_n = 0) \stackrel{\text{def}}{=} \begin{cases} 1 & \text{if } X_n = 0 \\ 0 & \text{if } X_n \neq 0, \end{cases}$$

which is one if  $X_n = 0$  and zero otherwise, is the *indicator* of the event  $\{X_n = 0\}$  that the random walk returns to the origin at time  $n$ . Then, the random variable

$$Z_n = \sum_{i=1}^n \mathbf{1}(X_i = 0)$$

is the number of times the random walk has returned to the origin in the first  $n$  steps.

One may also study returns according to the times at which they occur. The random time

$$T^0 = \min \{n \geq 1 : X_n = 0\}$$

is the time at which the random walk first returns to the origin, and one may also define times  $T_2^0, T_3^0, \dots$  of the second, third, ... returns. It is not even obvious that  $T^0$  must be finite, since there are realizations  $\omega$  for which  $X_n(\omega) \neq 0$  for all  $n \geq 1$ , in which case we put  $T^0(\omega) = \infty$ . It turns out, however, that

$$P\{T^0 < \infty\} = \sum_{n=1}^{\infty} P\{T^0 = n\} = 1 - \lim_{n \rightarrow \infty} P\{T^0 > n\} = 1, \quad (0.11)$$

but  $T^0$  assumes large values with probabilities large enough that

$$E[T^0] = \sum_{n=1}^{\infty} n P\{T^0 = n\} = \infty.$$

Thus, the average time to return to the origin is infinite.

We illustrate computation of probabilities by counting paths by showing that for each  $n$ ,

$$P\{T^0 > 2n\} = P\{X_1 \neq 0, \dots, X_{2n} \neq 0\} = \binom{2n}{n} \left(\frac{1}{2}\right)^{2n}, \quad (0.12)$$

which we write in this form since  $T^0$  must be even. Validity of (0.12) follows at once from the following geometric observation.

**Reflection Principle.** For each  $n$ , there are as many paths of length  $2n$  [originating at  $(0, 0)$ ] that do not return to the origin before or at time  $2n$  as there are from  $(0, 0)$  to  $(2n, 0)$ .  $\square$

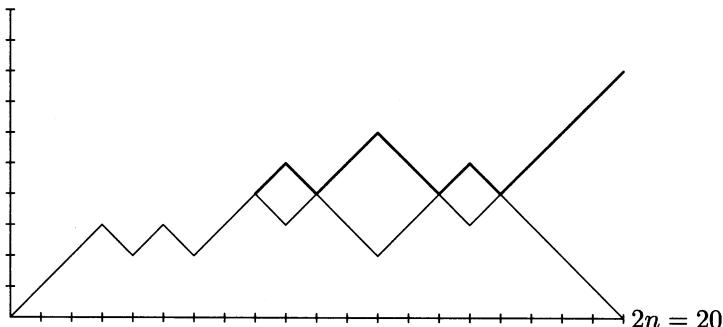
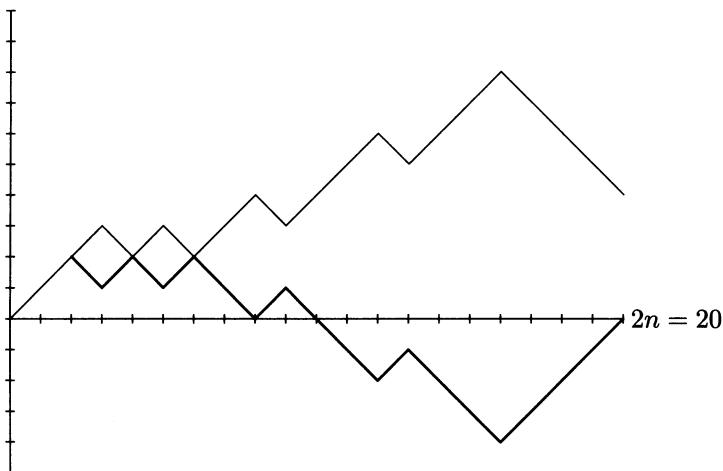


Figure 0.2. The First Reflection Principle

The one-to-one correspondence embodied the reflection principle is illustrated in Figure 0.2. In the first graph, the lighter line represents a path of length 20 that does not return to the origin. The final value of this path is  $2k = 4$ . The darker path is obtained by reflecting the lighter about the horizontal line  $y = k = 2$  beginning at the first time this level is hit, and is, as shown, at the origin at time 20.

Conversely, in the lower graph, the lighter path, likewise of length 20, is at the origin at  $n = 20$ . The maximum value of this path is 4, and is first attained at time 8. The darker path results from reflecting the lighter about the horizontal line  $y = 4$  from time 8 on, and does not return to the origin before time 20.

Further insight ensues from the behavior of  $P\{T^0 > 2n\}$  as  $n \rightarrow \infty$ .

From (0.12) and *Stirling's approximation* to  $n!$  for large values of  $n$ :

$$n! \cong \sqrt{2\pi} n^{n+1/2} e^{-n}, \quad (0.13)$$

it follows that

$$\begin{aligned} P\{T^0 > 2n\} &= \frac{(2n)!}{n! n!} \left(\frac{1}{2}\right)^{2n} \\ &\cong \frac{\sqrt{2\pi} (2n)^{2n+1/2} e^{-2n}}{[\sqrt{2\pi} n^{n+1/2} e^{-n}]^2} \left(\frac{1}{2}\right)^{2n} \\ &= 1/\sqrt{\pi n}. \end{aligned} \quad (0.14)$$

Consequently, (0.11) holds. Also, (0.12) and (0.14) show that as  $n \rightarrow \infty$ ,  $P\{X_{2n} = 0\} \cong 1/\sqrt{\pi n}$ .

Another implication of (0.12) is that

$$P\{T^0 = 2n\} = \frac{1}{2n-1} \binom{2n}{n} \left(\frac{1}{2}\right)^{2n},$$

in consequence of which  $E[T^0] = \infty$ .

Suppose now that  $T_j^0$  is the time of the  $j$ th return of the random walk to the origin. These times are all finite by (0.12). Moreover, since the steps are independent and identically distributed, once the random walk returns to the origin, its future behavior is independent of the past, and has the same probabilistic properties as the original process. Thus, the differences  $U_j^0 = T_j^0 - T_{j-1}^0$  are *independent and identically distributed* random variables. The distributions of the  $T_j^0$  are computed in (0.18), using results for first passage times.

## Numbers of returns to the origin

Recall that  $Z_n$  is the number of returns of the random walk to the origin in steps  $1, \dots, n$  (actually,  $2, \dots, n$ , since returns can occur only at even times). A reflection argument may be used to show that for  $k \leq n$ ,

$$P\{Z_{2n} = k\} = \binom{2n-k}{k} \left(\frac{1}{2}\right)^{2n-k}.$$

## First passage times

For each  $k$ , we define  $T^k$ , the *first passage time* to  $k$ , as the smallest value of  $n$  for which  $X_n = k$ :

$$T^k = \min \{n \geq 1 : X_n = k\}.$$

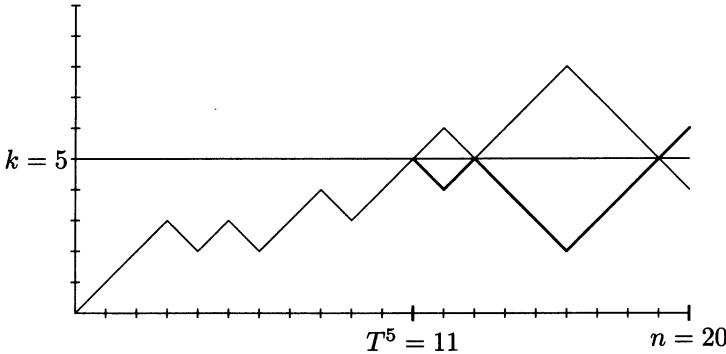


Figure 0.3. The Second Reflection Principle

For the same reason that it is not evident that (0.11) holds, we must address the issue of whether the  $T^k$  are finite. The key to computing the probabilities  $P\{T^k = n\}$  is another reflection principle.

**Reflection Principle, bis.** Suppose that  $k > 0$  and  $n > 0$ . Then, the number of paths from  $(0,0)$  to  $(n, k - 1)$  that hit or cross level  $k$  prior to time  $n$  is the same as the number of paths from  $(0,0)$  to  $(n, k + 1)$ .  $\square$

Figure 0.3 illustrates the second reflection principle. The heavier path hits  $k = 5$  prior to  $n = 20$  and has value  $k - 1 = 4$  at  $n = 20$ , while its lighter “twin” has value  $k + 1 = 6$  at  $n$ . The two coincide until  $T^5 = 11$ , and, thereafter, are mirror images of one another about the line  $y = 5$ .

In order to calculate  $P\{T^k = n\}$ , we must count the number of length- $n$  paths that hit the level  $k$  precisely at time  $n$ , which is  $\binom{n}{(n+k)/2}$ , the number of paths whose value at  $n$  is  $k$ , minus the number of paths from whose value at  $n$  is  $k$  and which have hit  $k$  (strictly) prior to  $n$ . Paths of the latter kind fall into two categories: those whose value at  $n - 1$  is  $k + 1$  and that have crossed  $k$  at some earlier time, of which there are

$$\binom{n-1}{[(n-1)+(k+1)]/2} = \binom{n-1}{(n+k)/2},$$

and those whose value at  $n - 1$  is  $k - 1$ , and which have reached  $k$  prior to  $n - 1$ . By the reflection principle, these number  $\binom{n-1}{(n+k)/2}$ . Consequently,

$$P\{T^k = n\} = \frac{k}{n} \binom{n}{(n+k)/2} \left(\frac{1}{2}\right)^n = \frac{k}{n} P\{X_n = k\}. \quad (0.15)$$

By computations that we omit, (0.15) implies that

$$P\{T^k < \infty\} = \sum_{n=1}^{\infty} P\{T^k = n\} = \sum_{n=1}^{\infty} \frac{k}{n} \binom{n}{(n+k)/2} \left(\frac{1}{2}\right)^n = 1, \quad (0.16)$$

yet also that

$$E[T^1] = \sum_{n=1}^{\infty} n P\{T^1 = n\} = \sum_{n=1}^{\infty} \binom{n}{(n+1)/2} \left(\frac{1}{2}\right)^n = \infty.$$

Since “continuity” of the random walk means that for  $k > 1$ ,  $T^k > T^1$ , we have  $E[T^k] \geq E[T^1] = \infty$  for all  $k$ . By symmetry,  $E[T^{-1}] = \infty$  as well, even though  $\min\{T^1, T^{-1}\} \equiv 1$ .

First passage times also yield information about times of returns to the origin. At time 1, the random walk is at  $\pm 1$ , from either of which, by spatial symmetry and homogeneity, its time to reach the origin has the same distribution as the time to reach 1 from 0. Therefore,  $T^0 \stackrel{d}{=} 1 + T^1$ , where  $\stackrel{d}{=}$  means “has the same distribution as.” More generally, also by spatial homogeneity,  $T^j$ , the time to get from 0 to  $j$ , has the distribution of the sum of  $j$  independent random variables, each with the same distribution as  $T^1$ :

$$T_j^0 \stackrel{d}{=} j + T^j. \quad (0.17)$$

As a consequence, for example, (0.15) and (0.17) imply that

$$P\{T_j^0 = n\} = P\{T^j = n - j\} = \frac{j}{n-j} P\{X_{n-j} = j\}. \quad (0.18)$$

## Maxima

For each  $n$ ,  $M_n^* = \max\{X_0, \dots, X_n\}$  is the maximum value attained by the random walk by time  $n$ . There is an important duality between maxima and first passage times: for  $k, n > 0$ ,  $T^k(\omega) > n$  if and only if  $M_n^*(\omega) < k$ .

The reflection principle implies that for each  $n$ , and for  $\ell \leq k$ ,

$$P\{M_n^* \geq k, X_n = \ell\} = P\{X_n = 2k - \ell\}.$$

Validity of this expression may be argued as follows: by reflection about the line  $y = \ell$  at the first time it is hit, there are as many paths from  $(0, 0)$  to  $(n, \ell)$  whose maximum value exceeds  $k$  as there are from  $(0, 0)$  to  $(n, 2k - \ell)$ . Consequently,

$$P\{M_n^* = k\} = P\{X_n = k\} + P\{X_n = k + 1\}. \quad (0.19)$$

Interestingly, (0.19) shows that  $M_n^*$  has the same magnitude as  $X_n$ .

## Time spent positive

For each  $n$ ,

$$W_n = \sum_{i=1}^n \mathbf{1}(X_i + X_{i-1} > 0)$$

is the number of times among  $\{0, \dots, n\}$  at which the random walk is positive. Being “positive” at time  $i$  requires that either  $X_i > 0$  or  $X_{i-1} > 0$  (or both).

The random variable  $W_n/n$ , the *fraction of time spent positive*, is often more informative. For ease in analysis, we restrict attention to even times, which we write as  $2n$ . Necessarily  $W_{2n}$  is even, and in fact,

$$P\{W_{2n} = 2k\} = \binom{2k}{2} \binom{2n-2k}{n-k} \left(\frac{1}{2}\right)^{2n} \quad (0.20)$$

This relationship typifies formulas in some areas of probability: albeit true, it is not very enlightening. Instead, below we consider asymptotics.

## Limit Theorems

The major limit theorems of probability pertain to asymptotic behavior of partial sums of independent random variables. The position process  $(X_n)$  is such a sequence, and satisfies two celebrated limit theorems:

1. **Law of large numbers.** As  $n \rightarrow \infty$ ,  $X_n/n$  converges to zero, the average value of the  $Y_i$ , with probability one:

$$P\{\lim_{n \rightarrow \infty} X_n/n = 0\} = 1.$$

Put differently, the “empirical averages”  $(1/n) \sum_{i=1}^n Y_i$  converge to the “theoretical average”  $E[Y_i] = 0$ .

2. **Central limit theorem.** As  $n \rightarrow \infty$ , the errors  $\sqrt{n}(X_n/n - E[Y_1])$ , albeit not convergent as functions on the sample space, have a particular, and very important, limit distribution:

$$\lim_{n \rightarrow \infty} P\{\sqrt{n}(X_n/n - E[Y_1]) \leq x\} = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy. \quad (0.21)$$

The central limit theorem is especially useful. For large  $n$ , the difficult-to-compute probability

$$P\{a < X_n \leq b\} = \sum_{a < k \leq b} \binom{n}{(n+k)/2} \left(\frac{1}{2}\right)^n$$

can be approximated, using (0.21):

$$P\{a < X_n \leq b\} = P\left\{\frac{a}{\sqrt{n}} < \frac{X_n}{\sqrt{n}} \leq \frac{b}{\sqrt{n}}\right\} \cong \int_{a/\sqrt{n}}^{b/\sqrt{n}} \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy.$$

A random variable  $Z$  with

$$P\{Z \leq x\} = \int_{-\infty}^x \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy, \quad x \in \mathbb{R},$$

is said to have a *standard normal distribution*. It differs qualitatively from integer-valued random variables such as the  $X_n$  and  $T^0$ . In particular,  $P\{Z = x\} = 0$  for every  $x \in \mathbb{R}$ . The integrand,

$$\phi(y) = \frac{1}{\sqrt{2\pi}} e^{-y^2/2},$$

the *standard normal density function*, is ubiquitous in probability and statistics, and is shown pictorially in Figure 2.1. The density function has the interpretation that for each  $y$ ,

$$P\{Z \in (y, y + dy]\} \cong \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy. \quad (0.22)$$

Also, the expectation of  $Z$  is

$$E[Z] = \int_{-\infty}^{\infty} y \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy = 0.$$

In view of (0.22), this expression, like (0.8), is the “sum” of the possible values  $y$  of  $Z$  weighted by their probabilities  $P\{Z \in (y, y + dy]\}$ .

Limit theorems are not valid only for sums of independent random variables. The arc sine law for random walks, which describes the asymptotic behavior of  $W_n/n$ , the fraction of time spent positive, is an example.

**Arc sine law.** For  $x \in [0, 1]$ ,

$$\lim_{n \rightarrow \infty} P\{W_n/n \leq x\} = \frac{2}{\pi} \arcsin \sqrt{x}. \quad \square$$

We sketch the derivation. As  $n \rightarrow \infty$ , by (0.20), for  $x \in [0, 1]$ ,

$$\begin{aligned} P\{W_{2n}/2n \leq x\} &= \sum_{k=0}^{xn} \binom{2k}{k} \binom{2n-2k}{n-k} \left(\frac{1}{2}\right)^{2n} \\ &\cong \sum_{k=0}^{xn} \frac{\sqrt{2\pi}(2k)^{2k+1/2} e^{-2k}}{[\sqrt{2\pi}k^{k+1/2}e^{-k}]^2} \frac{\sqrt{2\pi}(2n-2k)^{2n-2k+1/2} e^{-(2n-2k)}}{[\sqrt{2\pi}(n-k)^{(n-k+1/2)}e^{-(n-k)}]^2} \\ &= \sum_{k=0}^{xn} \frac{1}{\pi \sqrt{k(n-k)}} \\ &\cong \int_0^x \frac{1}{\pi} \frac{1}{\sqrt{y(1-y)}} dy \\ &= \frac{2}{\pi} \arcsin \sqrt{x}. \end{aligned}$$

Here, the first “ $\cong$ ” is by Stirling’s formula (0.13) and the second represents approximation of an integral by Riemann sums.

A random variable  $W$  with

$$P\{W \leq x\} = \int_0^x \frac{1}{\pi} \frac{1}{\sqrt{y(1-y)}} dy = \frac{2}{\pi} \arcsin \sqrt{x}, \quad x \geq 0,$$

has the *arc sine distribution*. The function  $f_W(y) = 1/\pi\sqrt{y(1-y)}$  is the *arc sine density* function. This function is U-shaped, so that  $W_n/n$  is more likely to be near 0 or 1 than near 1/2, which is in no sense apparent from (0.20).

## Summary

We conclude by reviewing the key concepts introduced so far. We began with a *random experiment*, to which is associated the *sample space*, the set of all possible *outcomes* of the experiment. *Events* are subsets of the sample space for which probability is defined. *Probability* is a set function measuring the likelihood of events. Its crucial property is countable additivity, in the form (0.2). *Random variables*, real-valued functions on the sample space, are the most important objects in probability. *Independence* is the absence of probabilistic interaction among random variables, manifested in (0.7). It engenders structure whose usefulness ranges from being a tool for computations to being the basis for limit theorems. *Expectation* defines an average value for random variables. *Limit theorems*, especially for sums of independent random variables, illuminate the “regularity in the large” that permeates probability. *Exploitation of particulars* is crucial in probability, and should be the first resort when attacking a problem, not the last.

# Chapter 1

# Probability

Probability is the mathematics of uncertainty. It has flourished under the stimulus of applications. Phenomena as diverse as games of chance, insurance, demography, waiting lines, astronomy, agricultural experiments, clinical trials, signal processing, traffic flow, reliability of complex systems, population growth, spread of infectious diseases, genetics, neurophysiology, microstructure of materials, precipitation, storage systems, telecommunications, statistical mechanics, earthquakes and medical imaging have furnished both difficult mathematical questions and genuine interest in the answers.

In probability, the main undefined concept is a *random experiment*, whose outcome cannot be determined in advance. The toss of a coin and an election poll are simple examples, but the outcome can be as complicated as the time-varying position of a molecule moving under the influence of collisions with other molecules or the locations, sizes and shapes of objects in an image.

Associated with a random experiment is a *sample space*  $\Omega$ , the set of all possible outcomes. “Possible” means “conceptually possible,” and often compromises between physical reality and mathematical convenience. For example, if the random experiment is to measure the time until a computer system fails, then negative outcomes are logically impossible, but even though a lifetime of  $10^{17}$  seconds is implausible and measurements of infinite precision are precluded, it may be more sensible to take  $\Omega$  to be  $\mathbb{R}_+ = [0, \infty)$  than to impose an arbitrary discretization or upper bound on the operating time.

*Probability* is a *set function*, defined for sets of outcomes, known as *events*. The probability of an event measures its “size,” in the sense of likelihood or frequency. Like other measures of size such as cardinality, length, area and mass, probability is *additive*: if  $A$  and  $B$  are disjoint events, then

$$P(A \cup B) = P(A) + P(B).$$

If outcomes are complicated, interest often focuses instead on numerical (real-valued) functions of them, known as *random variables*. Much of probability theory, and essentially all of statistics and stochastic processes, deals with random variables and their properties.

## 1.1 Random Experiments; Sample Spaces

In this section and the next, we introduce sample spaces and events.

### 1.1.1 Random experiments

A *random experiment* has associated with it the following objects:

1. **Sample space.** The set  $\Omega$  of all (conceptually) possible outcomes.
2. **Outcomes.** Elements  $\omega$  of the sample space, also referred to as *sample points* or *realizations*.
3. **Events.** Subsets of  $\Omega$  for which probability is defined.

### 1.1.2 Sample spaces

As previously intimated, the choice of the sample space for a random experiment balances fidelity to physical reality with mathematical convenience. In practice, most sample spaces fall into one of six categories.

**Example 1.1 (Finite set).** The simplest random experiment, for example, the toss of one coin, has only two outcomes. While one might take  $\Omega = \{H, T\}$ , representing heads and tails, a more flexible and mathematically efficient choice is  $\Omega = \{0, 1\}$ .

More generally, a random experiment with  $n$  possible outcomes may be modeled with a sample space consisting of  $n$  integers, usually either  $\{0, \dots, n - 1\}$  or  $\{1, \dots, n\}$ . Physical considerations may dictate, however, some other set of integers. For one step of a random walk, for example, the sensible sample space is the one used in the Prelude:  $\{-1, 1\}$ , so that the outcome is the size of the step.  $\square$

**Example 1.2 (Countable set).** The sample space for an experiment with countably many possible outcomes is ordinarily the set  $\mathbb{N} = \{0, 1, \dots\}$  of positive integers or the set  $\mathbb{Z} = \{\dots, -1, 0, 1, \dots\}$  of all integers.  $\square$

Whether a finite or countable sample space better describes a given phenomenon is a matter of judgment and compromise. Often, unless the number of outcomes is *logically finite* and a *known number* (as, for example, in the toss of a coin or the shuffle of a deck of cards), a countable sample space is chosen. Moreover, many useful and important probabilistic models

(such as those based on Poisson distributions) employ countable sample spaces.

**Example 1.3 (The real line  $\mathbb{R}$ ).** The most common sample space is the real line  $\mathbb{R} = (-\infty, \infty)$ , which is used for virtually all numerical phenomena that are not inherently integer-valued. For example, measurement errors in scientific observations, whose systematic study dates at least to Gauss in 1809, usually have  $\mathbb{R}$  as the sample space.

Intervals in  $\mathbb{R}$ , especially the “unit interval”  $[0, 1]$  and the positive half-line  $\mathbb{R}_+ = [0, \infty)$ , are also common sample spaces.  $\square$

Many random experiments are comprise replications of a basic experiment. In such cases, the sample space is a Cartesian product.

**Example 1.4 (Finitely many replications).** Consider the random experiment consisting of  $n$  replications of an underlying experiment with sample space  $\Omega_0$  (which might be  $\{0, 1\}$ ,  $\mathbb{N}$  or  $\mathbb{R}$ ). Then, the sample space for the repeated experiment is

$$\Omega = \Omega_0^n = \{\omega = (\omega_1, \dots, \omega_n) : \omega_i \in \Omega_0 \text{ for all } i\}.$$

An outcome for the “big” experiment is an  $n$ -tuple  $\omega = (\omega_1, \dots, \omega_n)$ , with  $\omega_i$  the outcome of the  $i$ th “sub”-experiment.  $\square$

**Example 1.5 (Infinitely many replications).** If a basic experiment is repeated infinitely many times, the sample space is the set  $\Omega = \Omega_0^{\mathbb{N}}$  of all infinite sequences  $\omega = (\omega_1, \omega_2, \dots)$  of elements of  $\Omega_0$ . This model includes the random walk, where  $\Omega_0 = \{-1, 1\}$ : an outcome is the entire sequence  $\omega$  of steps, each of which is  $\pm 1$ .  $\square$

Sample spaces can be as complicated as spaces of functions.

**Example 1.6 (Function spaces).** In some random experiments the outcome is the path or trajectory followed by a system over an interval of time. Outcomes are then functions. For example, if the system is observed over  $[0, 1]$  and its path is continuous, we could take  $\Omega = \mathbf{C}[0, 1]$ , the vector space of continuous, real-valued functions on  $[0, 1]$ . The probability models for such systems are known as *stochastic processes*.  $\square$

## 1.2 Events and Classes of Sets

Given a random experiment, we ultimately assign probabilities to various sets of outcomes, which are termed events. The family of events must be closed under appropriate set operations.

### 1.2.1 Events

The crucial concept is that *events are subsets of the sample space*. An event occurs if the outcome of the random experiment belongs to it.

**Provisional Definition.** Given a random experiment with sample space  $\Omega$ , an *event* is a subset of  $\Omega$  whose probability is defined.  $\square$

### 1.2.2 Basic set operations

As subsets of  $\Omega$ , events are manipulated using set operations. We begin by reviewing complementation, union and intersection, as well as the possibly less familiar operations of difference and symmetric difference. All sets below are subsets of a sample space  $\Omega$ . The concepts “is an element of,” “is a subset of” and the empty set  $\emptyset$  are presumed understood.

The *complement* of  $A$  is

$$A^c = \{\omega : \omega \notin A\}.$$

The *union* and *intersection* of  $A$  and  $B$  are

$$A \cup B = \{\omega : \omega \in A \text{ or } \omega \in B\}$$

and

$$A \cap B = \{\omega : \omega \in A \text{ and } \omega \in B\}.$$

Sets  $A$  and  $B$  are *disjoint* if  $A \cap B = \emptyset$ , that is, if they have no elements in common. As a mnemonic device, we use  $A + B$  to denote  $A \cup B$  when  $A$  and  $B$  are disjoint.

The *set difference* between  $B$  and  $A$  consists of those points in  $B$  but not in  $A$ :

$$B \setminus A \stackrel{\text{def}}{=} B \cap A^c.$$

The *symmetric difference* between  $A$  and  $B$  is the set of points in one but not both of  $A$  and  $B$ :

$$A \Delta B \stackrel{\text{def}}{=} A \setminus B + B \setminus A.$$

Set operations can be represented pictorially by *Venn diagrams*. Seemingly mundane, these are useful notwithstanding, and the five basic operations are depicted in Figure 1.1.

### 1.2.3 Indicator functions

Functions on the sample space (defined in Chapter 2 as random variables) are even more important than events, and we now discuss how a set is identified with a  $\{0, 1\}$ -valued function.

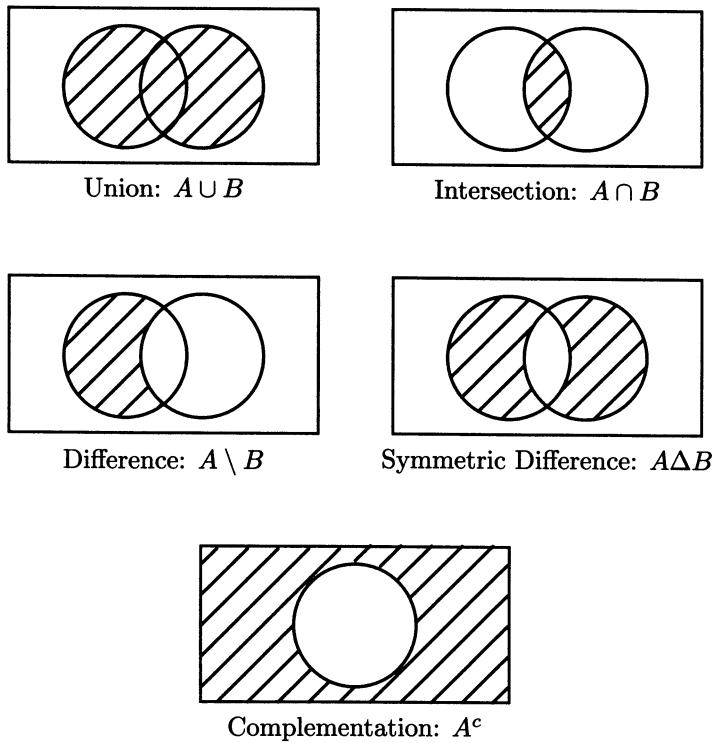


Figure 1.1. The Basic Set Operations

**Definition 1.7.** The *indicator function* of the set  $A \subseteq \Omega$  is the function on  $\Omega$  given by

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

Thus,  $\mathbf{1}_A$  “indicates” whether  $A$  occurs. Algebraic operations on indicator functions generalize set operations on events:

$$\mathbf{1}_{A \cup B} = \max \{\mathbf{1}_A, \mathbf{1}_B\} \tag{1.1}$$

$$\mathbf{1}_{A \cap B} = \min \{\mathbf{1}_A, \mathbf{1}_B\} = \mathbf{1}_A \mathbf{1}_B \tag{1.2}$$

$$\mathbf{1}_{A^c} = 1 - \mathbf{1}_A \tag{1.3}$$

$$\mathbf{1}_{A \Delta B} = |\mathbf{1}_A - \mathbf{1}_B|. \tag{1.4}$$

### 1.2.4 Operations on sequences of sets

Let  $(A_n)$  be a sequence of subsets of  $\Omega$ . The *union* of  $(A_n)$  is

$$\bigcup_{n=1}^{\infty} A_n = \{\omega : \omega \in A_n \text{ for some } n\},$$

and the *intersection* is

$$\bigcap_{n=1}^{\infty} A_n = \{\omega : \omega \in A_n \text{ for all } n\}.$$

The sequence  $(A_n)$  is *disjoint* if no two of its members have any elements in common:  $A_i \cap A_j = \emptyset$  whenever  $i \neq j$ . (Sometimes we use *pairwise disjoint* for emphasis.) The union of a disjoint sequence is denoted by  $\sum_{n=1}^{\infty} A_n$ .

There are also limits for sequences of sets.

**Definition 1.8.** Let  $A_1, A_2, \dots$  and  $A$  be subsets of  $\Omega$ .

- a) The  $\limsup$  of  $(A_n)$  is the set of  $\omega$  such that  $\omega \in A_n$  for *infinitely many* values of  $n$ :

$$\limsup_n A_n = \bigcap_{k=1}^{\infty} \bigcup_{n=k}^{\infty} A_n. \quad (1.5)$$

- b) The  $\liminf$  of  $(A_n)$  is the set of  $\omega$  belonging to  $A_n$  for *all but finitely many* values of  $n$ :

$$\liminf_n A_n = \bigcup_{k=1}^{\infty} \bigcap_{n=k}^{\infty} A_n. \quad (1.6)$$

- c) The sequence  $(A_n)$  *converges to*  $A$ , which we write as  $A = \lim_{n \rightarrow \infty} A_n$  or simply  $A_n \rightarrow A$ , if  $\liminf_n A_n = \limsup_n A_n = A$ .  $\square$

We also employ the mnemonic notations

$$\begin{aligned} \{A_n, \text{ i.o.}\} &\stackrel{\text{def}}{=} \limsup_n A_n \\ \{A_n, \text{ ult.}\} &\stackrel{\text{def}}{=} \liminf_n A_n, \end{aligned}$$

where “i.o.” and “ult.” abbreviate “infinitely often” and “ultimately.”

Monotone sequences of set converge.

**Proposition 1.9.** Let  $A_1, A_2, \dots$  be subsets of  $\Omega$ .

- a) If  $A_1 \subseteq A_2 \subseteq \dots$ , then  $A_n \rightarrow A = \bigcup_{n=1}^{\infty} A_n$ . (We denote this by  $A_n \uparrow A$ .)

- b) If  $A_1 \supseteq A_2 \supseteq \dots$ , then  $A_n \rightarrow A = \bigcap_{n=1}^{\infty} A_n$ . (This is written as  $A_n \downarrow A$ .)

**Proof:** We prove only a). Let  $A = \bigcup_{n=1}^{\infty} A_n$ . For each  $k$ ,  $\bigcup_{n=k}^{\infty} A_n = A$  as well, and, hence,  $\limsup_n A_n = A$ . On the other hand, for each  $k$ ,  $\bigcap_{n=k}^{\infty} A_n = A_k$ , so that  $\liminf_n A_n = A$  as well. ■

In terms of indicator functions,  $A_n \rightarrow A$  if and only if  $\mathbf{1}_{A_n}(\omega) \rightarrow \mathbf{1}_A(\omega)$  for every  $\omega \in \Omega$ . Indeed,

$$\mathbf{1}_{\liminf_n A_n} = \liminf_n \mathbf{1}_{A_n}$$

and

$$\mathbf{1}_{\limsup_n A_n} = \limsup_n \mathbf{1}_{A_n},$$

so that convergence of sets is the same as pointwise convergence of their indicator functions.

### 1.2.5 Classes of sets closed under set operations

We next define three classes of subsets of a sample space having prescribed closure properties under various set operations. Of these,  $\sigma$ -algebras are the most central, because the family of sets for which probability is defined must be a  $\sigma$ -algebra.

**Definition 1.10.** A  $\sigma$ -algebra on  $\Omega$  is a family of subsets of  $\Omega$  containing  $\Omega$  and closed under countable union, countable intersection and complementation. □

That is,  $\mathcal{F}$  is a  $\sigma$ -algebra if  $\Omega \in \mathcal{F}$ , if whenever  $A_1, A_2, \dots$  are sets such that  $A_i \in \mathcal{F}$  for each  $i$ , then  $\bigcup_{i=1}^{\infty} A_i \in \mathcal{F}$  and  $\bigcap_{i=1}^{\infty} A_i \in \mathcal{F}$ , and if whenever  $A \in \mathcal{F}$ , then  $A^c \in \mathcal{F}$  as well.

The two extreme  $\sigma$ -algebras on  $\Omega$  are the *trivial*  $\sigma$ -algebra  $\mathcal{F} = \{\emptyset, \Omega\}$  and the *power set*  $\mathcal{P}(\Omega)$ , consisting of all subsets of  $\Omega$ . In general, neither of these is especially interesting or useful.

Two other classes are also important, for establishing that families of interest are  $\sigma$ -algebras.

**Definition 1.11.** Let  $\Omega$  be a sample space.

- a) A *d-system* is a family of subsets containing  $\Omega$  and closed under proper difference (if  $A, B \in \mathcal{D}$  and  $A \subseteq B$ , then  $B \setminus A \in \mathcal{D}$ ) and countable increasing union.
- b) A  *$\pi$ -system* is a family of subsets closed under finite intersection. □

A  $\sigma$ -algebra is a *d*-system and a  *$\pi$* -system. In addition, a class that is both a  *$\pi$* -system and a *d*-system is a  $\sigma$ -algebra.

### 1.2.6 Generated classes

The mathematics of probability necessitates that we work with  $\sigma$ -algebras, but rarely are these given *a priori*. Rather, one desires a  $\sigma$ -algebra that contains some explicitly defined class of “elementary” events. Fortunately, there is a mechanistic procedure for achieving this: given any class of subsets of  $\Omega$ , there always exists a *minimal* family containing that class and closed under prescribed set operations. Minimality is important because unless  $\Omega$  is finite or countable, the power set  $\mathcal{P}(\Omega)$  is too large.

Here is the most important case. Versions hold for  $d$ -systems and  $\pi$ -systems as well.

**Theorem 1.12.** For every family  $\mathcal{G}$  of subsets of  $\Omega$  there exists a unique  $\sigma$ -algebra  $\sigma(\mathcal{G})$ , the  $\sigma$ -algebra *generated by*  $\mathcal{G}$ , such that

- a)  $\mathcal{G} \subseteq \sigma(\mathcal{G})$  ( $A \in \sigma(\mathcal{G})$  for all  $A \in \mathcal{G}$ ).
- b) For any  $\sigma$ -algebra  $\mathcal{H}$  with  $\mathcal{G} \subseteq \mathcal{H}$ ,  $\sigma(\mathcal{G}) \subseteq \mathcal{H}$ .  $\square$

We next present several key examples. The first describes a setting common in applications, which the second extends.

**Example 1.13 (Finite partition).** If  $\mathcal{G} = \{A_1, \dots, A_n\}$  is a finite partition of  $\Omega$  (that is,  $A_1, \dots, A_n$  are disjoint and  $\Omega = \sum_{i=1}^n A_i$ ), then  $\mathcal{S} = \{\emptyset, A_1, \dots, A_n\}$  is a  $\pi$ -system and

$$\sigma(\mathcal{G}) = \sigma(\mathcal{S}) = \left\{ \sum_{i \in I} A_i : I \subseteq \{1, \dots, n\} \right\}$$

consists of all unions of some of the  $A_i$ .  $\square$

**Example 1.14 (Countable partition).** If  $\mathcal{G} = \{A_1, A_2, \dots\}$  is a countable partition of  $\Omega$ , then  $\mathcal{S} = \{\emptyset, A_1, A_2, \dots\}$  is a  $\pi$ -system and

$$\sigma(\mathcal{G}) = \sigma(\mathcal{S}) = \left\{ \sum_{i \in I} A_i : I \subseteq \mathbb{N} \right\}. \quad \square$$

Next is the most important  $\sigma$ -algebra in probability, because of its role vis-à-vis random variables.

**Definition 1.15.** The *Borel  $\sigma$ -algebra* on  $\mathbb{R}$  is the  $\sigma$ -algebra  $\mathcal{B}(\mathbb{R})$  generated by the  $\pi$ -system  $\mathcal{J}$  of intervals  $(a, b]$ , where  $a < b$  in  $\mathbb{R}$ . (We also allow the possibility that  $a = -\infty$  or  $b = \infty$ .) Its elements are called *Borel sets*.

For  $A \in \mathcal{B}(\mathbb{R})$ , the  $\sigma$ -algebra

$$\mathcal{B}(A) = \{B \subseteq A : B \in \mathcal{B}(\mathbb{R})\}$$

of Borel subsets of  $A$  is termed the *Borel  $\sigma$ -algebra* on  $A$ .  $\square$

Every “reasonable” subset of  $\mathbb{R}$  — in particular, each interval, closed set, open set, finite set and countable set — is a Borel set. However,  $\mathcal{B}(\mathbb{R}) \neq \mathcal{P}(\mathbb{R})$ , although this latter property is deep and difficult to prove.

Borel sets are also defined in multi-dimensional Euclidean spaces.

**Definition 1.16.** The *Borel  $\sigma$ -algebra* on  $\mathbb{R}^d$  is the  $\sigma$ -algebra  $\mathcal{B}(\mathbb{R}^d)$  generated by the  $\pi$ -system  $\mathcal{J}$  of rectangles  $\prod_{i=1}^d (a_i, b_i]$  that are Cartesian products of intervals.  $\square$

The Borel  $\sigma$ -algebra  $\mathcal{B}(\mathbb{R}^d)$  is also generated by the  $\pi$ -system  $\mathcal{J}'$  of “rectangles”  $\prod_{i=1}^d B_i$ , where the  $B_i$  are one-dimensional Borel sets. As with  $\mathcal{B}(\mathbb{R})$ , all “reasonable” subsets of  $\mathbb{R}^d$  belong to  $\mathcal{B}(\mathbb{R}^d)$ .

### 1.2.7 The monotone class theorem

Often (for example, in Theorem 1.29), it is necessary to prove that some class of sets is a  $\sigma$ -algebra. For some classes, this is more easily done by showing instead that they are  $d$ -systems. Denote by  $d(\mathcal{G})$  the  $d$ -system generated by  $\mathcal{G}$ .

**Theorem 1.17 (Monotone class theorem).** Let  $\mathcal{S}$  be a  $\pi$ -system on  $\Omega$ . Then,  $\sigma(\mathcal{S}) = d(\mathcal{S})$ .  $\square$

To show, then, that some property holds for all sets in  $\sigma(\mathcal{S})$ , it is enough to establish that the class  $\mathcal{U}$  of sets for which it holds is a  $d$ -system, for then  $d(\mathcal{S}) \subseteq \mathcal{U}$  by the definition of  $d(\mathcal{S})$ , while  $\sigma(\mathcal{S}) = d(\mathcal{S})$  by Theorem 1.17, and, consequently,  $\sigma(\mathcal{S}) \subseteq \mathcal{U}$ . This pattern of reasoning is termed a *monotone class argument*.

### 1.2.8 Events, bis

Finally, we can give a full definition of events, not one-by-one, but in terms of the class of events.

**Definition 1.18.** The family of *events* associated with a random experiment with sample space  $\Omega$  is a  $\sigma$ -algebra  $\mathcal{F}$  on  $\Omega$ .  $\square$

## 1.3 Probabilities and Probability Spaces

### 1.3.1 Probability

Probability is a *set function*, defined for events, and is (*countably*) *additive*: the probability of the countable disjoint union of events is the sum of their individual probabilities.

Let  $(\Omega, \mathcal{F})$  be a sample space and a  $\sigma$ -algebra of events.

**Definition 1.19.** A *probability* on  $(\Omega, \mathcal{F})$  is a function  $P: \mathcal{F} \rightarrow \mathbb{R}$  such that

- a)  $P(A) \geq 0$  for all  $A \in \mathcal{F}$ .
- b)  $P(\Omega) = 1$ .
- c) Whenever  $A_1, A_2, \dots$  are (pairwise) disjoint sets in  $\mathcal{F}$ ,

$$P\left(\sum_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} P(A_n). \quad (1.7)$$

The triple  $(\Omega, \mathcal{F}, P)$  is a *probability space*.  $\square$

The terminology is as follows:  $P(A)$  is the *probability* of the event  $A$ . The relationship (1.7), the crucial part of the definition, is *countable additivity*. Probabilities are sometimes called *probability measures*, because they are special cases of set functions known as *measures*, defined in §6.

Although other interpretations are possible, perhaps the most plausible interpretation of  $P(A)$  is as the long-run frequency of occurrence of  $A$  under independent replications of the random experiment, as discussed following Theorem 5.31.

At this point only a few examples can be introduced. The simplest probabilities are concentrated on one outcome.

**Example 1.20 (Point mass).** The *point mass* at  $\omega \in \Omega$  is the probability

$$\varepsilon_{\omega}(A) = \begin{cases} 1 & \text{if } \omega \in A \\ 0 & \text{otherwise.} \end{cases} \quad (1.8)$$

Like many degenerate cases, this one is good for testing conjectures and constructing counterexamples. Note also that  $\varepsilon_{\omega}(A) = \mathbf{1}(\omega \in A)$ .  $\square$

Uniform distributions depict “equally likely” outcomes.

**Example 1.21 (Uniform distribution).** Suppose that  $\Omega$  is a finite set, and that  $\mathcal{F} = \mathcal{P}(\Omega)$ . If all outcomes are to be “equally likely,” then the probability of an event  $A$  is proportional to its cardinality  $|A|$ , the number of outcomes it contains. Thus, because  $P(\Omega) = 1$ ,

$$P(A) = |A|/|\Omega|.$$

The probability  $P$  is known as the *uniform distribution* on  $\Omega$ .

Even here we are not defining probabilities of individual outcomes. The outcome  $\omega$  and the event  $\{\omega\}$  are logically distinct objects.  $\square$

The next example shows how a probability on the  $\sigma$ -algebra generated by a partition is constructed from its values on the partition sets.

**Example 1.22 (Finite partition).** Let  $\mathcal{G} = \{A_1, \dots, A_n\}$  be a finite partition of  $\Omega$  and let  $\mathcal{F}$  be the  $\sigma$ -algebra generated by  $\mathcal{G}$  (see Example 1.13). Let  $p_1, \dots, p_n$  be positive numbers with  $\sum_{i=1}^n p_i = 1$ . Then,

$$P(\sum_{i \in I} A_i) = \sum_{i \in I} p_i, \quad I \subseteq \{1, \dots, n\}, \quad (1.9)$$

defines a probability on  $(\Omega, \mathcal{F})$ . Conversely, every probability  $P$  on  $(\Omega, \mathcal{F})$  satisfies (1.9), with  $p_i = P(A_i)$ .  $\square$

### 1.3.2 Elementary properties

Two things should be noted about these properties: they have to be proved, and how this is done.

**Theorem 1.23.** Let  $(\Omega, \mathcal{F}, P)$  be a probability space. Then,

- a)  $P(\emptyset) = 0$ .
- b)  $P$  is *finitely additive*: if  $A_1, \dots, A_n$  are (pairwise) disjoint, then

$$P(\sum_{i=1}^n A_i) = \sum_{i=1}^n P(A_i). \quad (1.10)$$

Consequently, for each  $A$ ,

$$P(A^c) = 1 - P(A). \quad (1.11)$$

- c) If  $A \subseteq B$ , then

$$P(B \setminus A) = P(B) - P(A). \quad (1.12)$$

Thus,  $P$  is *monotone*: if  $A \subseteq B$ , then  $P(A) \leq P(B)$ .

- d) For all  $A$  and  $B$  (disjoint or not),

$$P(A \cup B) + P(A \cap B) = P(A) + P(B). \quad (1.13)$$

Hence,  $P$  is *(finitely) subadditive*: for all  $A$  and  $B$ , disjoint or not,

$$P(A \cup B) \leq P(A) + P(B). \quad (1.14)$$

**Proof:** a) Taking  $A_1 = A_2 = \dots = \emptyset$  in (1.7) gives  $P(\emptyset) = \sum_{n=1}^{\infty} P(\emptyset)$ , which can hold only if  $P(\emptyset) = 0$ .

b) To deduce (1.10) from (1.7), take  $A_{n+1} = A_{n+2} = \dots = \emptyset$  and apply a). In particular, since  $A + A^c = \Omega$ ,  $P(A) + P(A^c) = 1$  for each  $A$ .

c) When  $A \subseteq B$ ,  $B = A + (B \setminus A)$ , so this follows from (1.10). Note that (1.12) need not hold when  $A$  is not a subset of  $B$ .

d) We have  $A \cup B = A + (B \setminus A \cap B)$ , so that by (1.10) and c),

$$\begin{aligned} P(A \cup B) &= P(A) + P(B \setminus A \cap B) \\ &= P(A) + P(B) - P(A \cap B). \quad \blacksquare \end{aligned}$$

### 1.3.3 More advanced properties

Boole's inequality, because of its simplicity, is one of the most ubiquitous results in probability. It establishes *countable subadditivity* of  $P$ , which generalizes the finite subadditivity in (1.14).

**Proposition 1.24 (Boole's inequality).** For events  $A_1, A_2, \dots$ ,

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

**Proof:** This argument illustrates a very important technique known colloquially as “disjointification.” For each  $n$ , let  $B_n = A_n \setminus \bigcup_{i=1}^{n-1} A_i$  be the set of outcomes in  $A_n$  but not in any of  $A_1, \dots, A_{n-1}$ . Then,  $B_1, B_2, \dots$  are disjoint,  $\bigcup_{n=1}^{\infty} A_n = \sum_{n=1}^{\infty} B_n$  and  $B_n \subseteq A_n$  for each  $n$ . Consequently,

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

by (1.7) and monotonicity of  $P$ . ■

The next result shows that countable additivity is equivalent to the confluence of finite additivity, which is reasonable physically, and (monotone) continuity, which is convenient and desirable mathematically.

**Theorem 1.25.** Let  $P$  be a positive, finitely additive set function on  $\mathcal{F}$  with  $P(\Omega) = 1$ . Then, the following are equivalent:

- a)  $P$  is countably additive (and, hence, a probability).
- b) Whenever  $A_n \uparrow A$  in  $\mathcal{F}$ ,  $P(A_n) \uparrow P(A)$ .
- c) Whenever  $A_n \downarrow A$  in  $\mathcal{F}$ ,  $P(A_n) \downarrow P(A)$ .
- d) Whenever  $A_n \downarrow \emptyset$  in  $\mathcal{F}$ ,  $P(A_n) \downarrow 0$ .

**Proof:** We show that  $a) \Rightarrow b) \Rightarrow c) \Rightarrow d) \Rightarrow a)$ . Two of these implications are easy: b) and c) are equivalent by complementation, while d) is a special case of c).

a)  $\Rightarrow$  b): Suppose that  $A_n \uparrow A$ . Since  $(A_n)$  is increasing, we can “disjointify”  $A$  as  $A = \sum_{i=1}^{\infty} (A_i \setminus A_{i-1})$  (with  $A_0 = \emptyset$ ), so that

$$P(A) = \sum_{i=1}^{\infty} P(A_i \setminus A_{i-1}) = \lim_{n \rightarrow \infty} \sum_{i=1}^n [P(A_i) - P(A_{i-1})] = \lim_{n \rightarrow \infty} P(A_n).$$

Here, the first equality is by countable additivity, while at each of the last two steps we used the property that the  $A_n$  are increasing.

d)  $\Rightarrow$  a): For disjoint events  $A_1, A_2, \dots, \sum_{i=n+1}^{\infty} A_i \downarrow \emptyset$  as  $n \rightarrow \infty$ , because any  $\omega$  belongs to at most one of the  $A_i$ , so for each  $n$ ,

$$\begin{aligned} P\left(\sum_{i=1}^{\infty} A_i\right) &= P\left(\sum_{i=1}^n A_i + \sum_{i=n+1}^{\infty} A_i\right) \\ &= \sum_{i=1}^n P(A_i) + P\left(\sum_{i=n+1}^{\infty} A_i\right) \\ &\rightarrow \sum_{i=1}^{\infty} P(A_i), \end{aligned}$$

using successively finite additivity of  $P$  and then d). ■

Probabilities are continuous, in the sense of Definition 1.8.

**Theorem 1.26.** If  $A_n \rightarrow A$ , then  $P(A_n) \rightarrow P(A)$ .

**Proof:** We show that

$$P\left(\liminf_n A_n\right) \leq \liminf_{n \rightarrow \infty} P(A_n); \quad (1.15)$$

$$\limsup_{n \rightarrow \infty} P(A_n) \leq P\left(\limsup_n A_n\right). \quad (1.16)$$

These inequalities, particularly (1.15), a special case of Fatou's lemma (Theorem 4.8), have independent interest. They yield the corollary as follows: if  $A_n \rightarrow A$ , then  $P(\liminf_n A_n) = P(\limsup_n A_n) = P(A)$ , and since we always have  $\liminf_n P(A_n) \leq \limsup_n P(A_n)$ , it must be that

$$\liminf_{n \rightarrow \infty} P(A_n) = \limsup_{n \rightarrow \infty} P(A_n) = P(A),$$

and, hence, that  $P(A_n) \rightarrow P(A)$ .

Since (1.15) and (1.16) are equivalent, by complementation, it suffices to establish the former. For each  $k$ ,  $\bigcap_{n \geq k} A_n \subseteq A_k$ , so that

$$P\left(\bigcap_{n \geq k} A_n\right) \leq P(A_k),$$

and since  $\bigcap_{n \geq k} A_n \uparrow \liminf_n A_n$ , taking limits as  $k \rightarrow \infty$  gives (1.15). ■

The following result is known as the (first) Borel-Cantelli lemma. It is crucial to the proofs of many convergence theorems.

**Theorem 1.27 (Borel-Cantelli lemma).** If  $\sum_{n=1}^{\infty} P(A_n) < \infty$ , then  $P\{A_n, \text{i.o.}\} = 0$ .

**Proof:** As  $m \rightarrow \infty$ ,  $\bigcup_{k=m}^{\infty} A_k \downarrow \{A_n, \text{i.o.}\}$ , so by Proposition 1.24,

$$P\{A_n, \text{i.o.}\} = \lim_{m \rightarrow \infty} P\left(\bigcup_{k=m}^{\infty} A_k\right) \leq \lim_{m \rightarrow \infty} \sum_{k=m}^{\infty} P(A_k) = 0. \quad \blacksquare$$

### 1.3.4 Almost sure and null events

Events whose probability is one carry “the whole truth” in probability. In hypotheses, it is unnecessary (or even impossible) to assume that something is true of every outcome, but rather only that it is true of outcomes belonging to an event of probability one. Correspondingly, in conclusions, that some property holds on an event of probability one is, ordinarily, all that one can establish.

**Definition 1.28.** An event  $A$  is *almost sure* if  $P(A) = 1$  and *null* if  $P(A) = 0$ . A property of outcomes holds *almost surely* if there is an almost sure event on which it is satisfied.  $\square$

### 1.3.5 Uniqueness

The following result, proved using a prototypical “monotone class argument,” states that a probability is determined by its values on a  $\pi$ -system generating the  $\sigma$ -algebra of events.

**Theorem 1.29.** Let  $\mathcal{F}$  be a  $\sigma$ -algebra on  $\Omega$  and let  $\mathcal{S}$  be a  $\pi$ -system with  $\sigma(\mathcal{S}) = \mathcal{F}$ . If  $P$  and  $P'$  are probabilities on  $\mathcal{F}$ , such that  $P = P'$  on  $\mathcal{S}$ , then  $P = P'$  on  $\mathcal{F}$ .

**Proof:** By assumption,  $\mathcal{S}$  is contained in  $\mathcal{G} = \{A \in \mathcal{F}: P(A) = P'(A)\}$ . If we show that  $\mathcal{G}$  is a  $d$ -system, then

$$\mathcal{G} \supseteq d(\mathcal{S}) = \sigma(\mathcal{S}) = \mathcal{F},$$

with the first equality by the monotone class theorem (Theorem 1.17). This suffices to complete the proof.

Since  $P(\Omega) = P'(\Omega) = 1$ ,  $\Omega$  belongs to  $\mathcal{G}$ . If  $A \subseteq B$  in  $\mathcal{G}$ , then by two applications of (1.12),

$$P(B \setminus A) = P(B) - P(A) = P'(B) - P'(A) = P'(B \setminus A),$$

confirming that  $\mathcal{G}$  is closed under proper differences. Finally, if  $A_n \in \mathcal{G}$  for each  $n$  and  $A_n \uparrow A$ , then by Theorem 1.26,

$$P(A) = \lim_{n \rightarrow \infty} P(A_n) = \lim_{n \rightarrow \infty} P'(A_n) = P'(A),$$

so that  $A \in \mathcal{G}$ , and, hence,  $\mathcal{G}$  is indeed a  $d$ -system. ■

The following illustration is especially important.

**Example 1.30 (Probabilities on  $\mathcal{B}(\mathbb{R})$ ).** Theorem 1.29 implies that a probability  $P$  on the Borel  $\sigma$ -algebra  $\mathcal{B}(\mathbb{R})$  is determined by its values  $P((a, b])$  for all intervals  $(a, b]$ .  $\square$

## 1.4 Probabilities on $\mathbb{R}$

Probabilities on the real line play a key role as distributions of random variables. To avoid pedanticism, we speak of probabilities on  $\mathbb{R}$  rather than on  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ .

### 1.4.1 Distribution functions

Each probability on  $\mathbb{R}$  is specified by its values on intervals  $(-\infty, t]$ , which are given by a function on  $\mathbb{R}$ .

**Definition 1.31.** The *distribution function* of  $P$  is the function  $F_P: \mathbb{R} \rightarrow [0, 1]$  defined by

$$F_P(t) = P((-\infty, t]). \quad \square$$

Our first result is a uniqueness theorem.

**Proposition 1.32.** If  $F_P = F_{P'}$ , then  $P = P'$ .

**Proof:** Let  $\mathcal{J}$  be the  $\pi$ -system of intervals  $(a, b]$ . Since  $\mathcal{J}$  generates  $\mathcal{B}(\mathbb{R})$ , it suffices by Theorem 1.29 to show that  $P$  and  $P'$  agree on  $\mathcal{J}$ . For  $a < b$ ,

$$(a, b] = (-\infty, b] \setminus (-\infty, a],$$

so that by (1.12),

$$\begin{aligned} P((a, b]) &= P((-\infty, b] \setminus (-\infty, a]) = P((-\infty, b]) - P((-\infty, a]) \\ &= F_P(b) - F_P(a) \\ &= F_{P'}(b) - F_{P'}(a) \\ &= P'((a, b]). \quad \blacksquare \end{aligned}$$

Here are key properties of distribution functions.

**Theorem 1.33.** Let  $F_P$  be the distribution function of  $P$ . Then,

- a)  $F_P$  is increasing. Hence, the left-hand limit

$$F_P(t-) \stackrel{\text{def}}{=} \lim_{s \uparrow t, s < t} F_P(s)$$

and the right-hand limit

$$F_P(t+) \stackrel{\text{def}}{=} \lim_{s \downarrow t, s > t} F_P(s)$$

exist for each  $t$ , and  $F_P(t-) \leq F_P(t) \leq F_P(t+)$ .

Interval $I$	$P(I)$	Interval $I$	$P(I)$
$(-\infty, t]$	$F_P(t)$	$(t, \infty)$	$1 - F_P(t)$
$(-\infty, t)$	$F_P(t-)$	$[t, \infty)$	$1 - F_P(t-)$
$(a, b]$	$F_P(b) - F_P(a)$	$[a, b)$	$F_P(b-) - F_P(a-)$
$[a, b]$	$F_P(b) - F_P(a-)$	$(a, b)$	$F_P(b-) - F_P(a)$
$\{t\}$	$F_P(t) - F_P(t-)$		

Table 1.1. Probabilities in Terms of Distribution Functions

- b)  $F_P$  is right-continuous:  $F_P(t+) = F_P(t)$  for each  $t$ .  
 c)  $\lim_{t \rightarrow -\infty} F_P(t) = 0$  and  $\lim_{t \rightarrow \infty} F_P(t) = 1$ .

**Proof:** a) Monotonicity of  $F_P$  mirrors that of  $P$ : for  $s < t$ ,  $(-\infty, s] \subseteq (-\infty, t]$ , and, hence,

$$F_P(s) = P((-\infty, s]) \leq P((-\infty, t]) = F_P(t).$$

Existence of left-hand and right-hand limits follows from monotonicity. (See Rudin, 1975.)

b) To prove right-continuity, note that if  $t_n \downarrow t$ , then  $(-\infty, t_n] \downarrow (-\infty, t]$  and, therefore, by Theorem 1.25,

$$F_P(t) = P((-\infty, t]) = \lim_{n \rightarrow \infty} P((-\infty, t_n]) = \lim_{n \rightarrow \infty} F_P(t_n).$$

c) As  $n \rightarrow \infty$ ,  $(-\infty, -n] \downarrow \emptyset$ , while  $(-\infty, n] \uparrow \mathbb{R}$ , so both statements are consequences of continuity of  $P$  and monotonicity of  $F_P$ . For example,

$$\lim_{t \rightarrow \infty} F_P(t) \geq \lim_{t \rightarrow \infty} F_P(\lfloor t \rfloor) = \lim_{n \rightarrow \infty} F_P(n) = 1,$$

where  $\lfloor t \rfloor$  is the integer part of  $t$ . ■

As  $t_n \uparrow t$  strictly ( $t_n < t$  for each  $n$ ), we have  $(-\infty, t_n] \uparrow (-\infty, t)$ , which means that

$$F_P(t-) = P((-\infty, t)).$$

Table 1.1 expresses probabilities of various intervals in terms of distribution functions.

For occasions when it is more convenient to work with probabilities

$$P((t, \infty)) = 1 - P((-\infty, t]),$$

we use the following terminology.

**Definition 1.34.** The *survivor function* of  $P$  is the function  $S_P(t) = 1 - F_P(t) = P((t, \infty))$ .  $\square$

Probabilities  $1 - F_P(t)$  are also termed *tail probabilities* of  $F_P$ .

The only distribution functions we are able to calculate at this juncture are those of point masses.

**Example 1.35 (Point mass).** Let  $P = \delta_s$ , the point mass at  $s$ . Then,

$$F_P(t) = \begin{cases} 0 & t < s \\ 1 & t \geq s. \end{cases}$$

This property that  $F_P$  is only 0 or 1 characterizes point masses.  $\square$

Every function on  $\mathbb{R}$  with the properties in Theorem 1.33 is the distribution function of a probability on  $\mathcal{B}(\mathbb{R})$ .

**Theorem 1.36.** Let  $F: \mathbb{R} \rightarrow \mathbb{R}$  be an increasing, right-continuous function with  $F(-\infty) = 0$  and  $F(\infty) = 1$ . Then, there is a unique probability  $P$  on  $\mathcal{B}(\mathbb{R})$  such that  $F_P = F$ .  $\square$

Using Theorem 1.36, we can construct more probabilities on  $\mathbb{R}$ . Further examples appear in §2.4, as distributions of random variables.

**Example 1.37 (Uniform distribution).** The experiment of selecting a point “at random” from the interval  $[0, 1]$  can be realized as follows. Heuristically, all outcomes are equally likely, so “at random” is interpreted to mean that the probability the point chosen belongs to an interval is proportional to its length. By Theorem 1.36 there exists a probability  $P$  on  $\mathbb{R}$ , known as the *uniform distribution* on  $[0, 1]$  and denoted by  $U[0, 1]$ , with

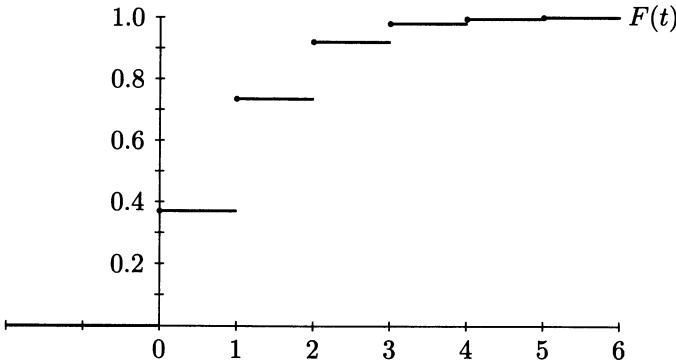
$$F_P(t) = \begin{cases} 0 & t < 0 \\ t & 0 \leq t \leq 1 \\ 1 & t > 1. \end{cases}$$

More generally, for  $a < b$ , the function

$$F(t) = \begin{cases} 0 & t < a \\ (t - a)/(b - a) & a \leq t \leq b \\ 1 & t > b, \end{cases}$$

is the distribution function of a probability  $P$  on  $\mathbb{R}$  (or on  $[a, b]$ ), called the *uniform distribution* on  $[a, b]$  and denoted by  $U[a, b]$ .  $\square$

Two classes of probabilities on  $\mathbb{R}$  are especially amenable to computational manipulation: the discrete and absolutely continuous probabilities, for which computations are effected via summations and integrals.

Figure 1.2. The Poisson Distribution with  $\lambda = 1$ 

### 1.4.2 Discrete probabilities

**Definition 1.38.** A probability  $P$  on  $\mathbb{R}$  is *discrete* if there exists a countable set  $C$  such that  $P(C) = 1$ .  $\square$

Figure 1.2 illustrates a key discrete distribution: the Poisson distribution with parameter 1 (see Example 2.34).

Discrete probabilities are finite or countable convex combinations of point masses. Their distribution functions increase only by means of jumps, rather than “smoothly.”

**Proposition 1.39.** The following are equivalent for each probability  $P$  on  $\mathbb{R}$ :

- a)  $P$  is discrete.
- b) There exist a real sequence  $(t_i)$  and numbers  $p_i$  with  $p_i > 0$  for each  $i$  and  $\sum_{i=1}^{\infty} p_i = 1$  such that  $P = \sum_i p_i \varepsilon_{t_i}$ .
- c) There exist a real sequence  $(t_i)$  and numbers  $p_i$  with  $p_i > 0$  for each  $i$  and  $\sum_{i=1}^{\infty} p_i = 1$  such that  $F_P(t) = \sum_i p_i \mathbf{1}(t_i \leq t)$  for all  $t \in \mathbb{R}$ .

**Proof:** a)  $\Rightarrow$  b): If  $C = \{t_i : i \in \mathbb{N}\}$  is a countable set with  $P(C) = 1$ , then for each Borel set  $B$ ,

$$P(B) = \sum_i P(B \cap \{t_i\}) = \sum_{t_i \in B} P(\{t_i\}) = \sum_i P(\{t_i\}) \varepsilon_{t_i}(B),$$

so that b) holds with  $p_i = P(\{t_i\})$ .

b)  $\Rightarrow$  c): For each  $t$ ,

$$F_P(t) = \sum_i p_i \varepsilon_{t_i} ((-\infty, t]) = \sum_i p_i \mathbf{1}(t_i \leq t).$$

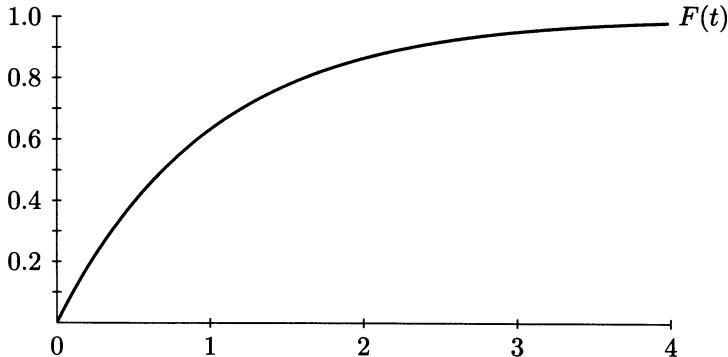


Figure 1.3. The Exponential Distribution with  $\lambda = 1$

c)  $\Rightarrow$  a): If  $F_P$  has the form stipulated in c), then  $P$  is discrete since

$$\begin{aligned} P(\{t_i : i \in \mathbb{N}\}) &= \sum_i P(\{t_i\}) = \sum_i [F_P(t_i) - F_P(t_{i-})] \\ &= \sum_i p_i = 1. \quad \blacksquare \end{aligned}$$

The distribution function of a discrete probability increases only by jumps, with  $p_i$  the size of the jump at  $t_i$ . One should not, however, conclude that the jumps are always isolated from each other: the set of jumps could be, for example, that of all rational numbers.

### 1.4.3 Absolutely continuous probabilities

Absolutely continuous probabilities are the antithesis of discrete, with distribution functions smooth enough to be indefinite integrals of their derivatives. Figure 1.3 shows one key example, the exponential distribution (see Example 2.36), which is related to Poisson distributions through Poisson processes (§3.6).

**Definition 1.40.** A probability  $P$  on  $\mathbb{R}$  is *absolutely continuous* if there exists a positive function  $f_P$  on  $\mathbb{R}$ , the *density function* of  $P$ , such that for every interval  $(a, b]$ ,

$$P((a, b]) = \int_a^b f_P(t) dt. \quad \square \tag{1.17}$$

**Technical Aside.** In the most general case, the integrals in (1.17) are Lebesgue integrals (see §4.6). However, in nearly all specific cases,  $f_P$  is piecewise continuous, and the integrals are Riemann integrals.

The term “the” density function is a misnomer, since in a technical sense it is not unique. However, any two functions satisfying (1.17) differ at most on a set of Lebesgue measure zero (Definition 1.48), and may be regarded as identical.  $\square$

Virtually by definition, absolutely continuous probabilities are also characterized via their distribution functions.

**Proposition 1.41.** The probability  $P$  is absolutely continuous if and only if there exists a positive function  $f$  on  $\mathbb{R}$  with  $\int_{-\infty}^{\infty} f(s) ds = 1$  and

$$F_P(t) = \int_{-\infty}^t f(s) ds, \quad t \in \mathbb{R}. \quad (1.18)$$

**Proof:** If  $P$  is absolutely continuous, then its density  $f_P$  satisfies these properties. Conversely, if  $\int_{-\infty}^{\infty} f(s) ds = 1$ , then (1.18) defines a distribution function, to which Theorem 1.36 applies. ■

Hence, any positive function  $f$  satisfying with  $\int_{-\infty}^{\infty} f(s) ds = 1$  may safely be termed a density function.

#### 1.4.4 Mixed distributions

A probability need not be discrete or absolutely continuous; it may have components of both types, or neither. The mixed case, in which both components are present, entails no new difficulties.

**Example 1.42 (Mixed distribution).** The probability  $P$  with distribution function

$$F_P(t) = \begin{cases} 0 & t < 0 \\ 1 - e^{-\lambda t} & 0 \leq t < s \\ 1 & t \geq s, \end{cases}$$

where  $s$  and  $\lambda$  are positive, which arises from backward recurrence times in Poisson processes (§3.6), has both discrete and absolutely continuous components. In fact,  $P$  is a convex combination of  $\varepsilon_s$ , the point mass at  $s$ , and  $E(\lambda)$ , the exponential distribution with parameter  $\lambda$ :

$$P = e^{-\lambda s} \varepsilon_s + (1 - e^{-\lambda s}) E(\lambda).$$

See Figure 1.4, where  $\lambda = 1$  and  $s = 2$ .  $\square$

A third (in fact, *the* third — see Theorem 1.52) class of probabilities on  $\mathbb{R}$ , the *singular* probabilities, is discussed in §6.

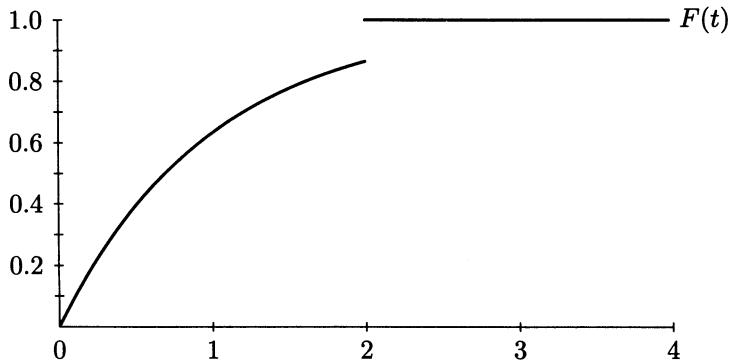


Figure 1.4. The Mixed Distribution of Example 1.42, with  $\lambda = 1$

## 1.5 Conditional Probability Given a Set

In this section, we show how to revise probabilities to account for the knowledge that an event has occurred, using a concept known as conditional probability. A more powerful version of conditioning, namely, conditional expectation, is presented in Chapter 8.

**Definition 1.43.** Let  $A$  and  $B$  be events. Provided that  $P(A) > 0$ , the *conditional probability of  $B$  given  $A$*  is

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

In case  $P(A) = 0$ , we make the convention that  $P(B|A) = P(B)$ .  $\square$

One interprets  $P(B|A)$  as the relative likelihood that  $B$  occurs given that  $A$  is known to have occurred. Observe that Definition 1.43 implies that

$$P(B \cap A) = P(B|A)P(A)$$

regardless of whether  $P(A) > 0$ .

We now present two extremely important computational formulas. The first expresses the probability of an event in terms of its conditional probabilities given elements of a partition of  $\Omega$ .

**Proposition 1.44 (Law of total probability).** Let  $\Omega = \sum_{i=1}^{\infty} A_i$  be a countable partition of  $\Omega$ . Then, for each event  $B$ ,

$$P(B) = \sum_{i=1}^{\infty} P(B|A_i)P(A_i).$$

**Proof:** By countable additivity of  $P$ , since  $B = \sum_{i=1}^{\infty} B \cap A_i$ ,

$$\begin{aligned} P(B) &= P\left(\sum_{i=1}^{\infty} B \cap A_i\right) = \sum_{i=1}^{\infty} P(B \cap A_i) \\ &= \sum_{i=1}^{\infty} P(B|A_i)P(A_i). \quad \blacksquare \end{aligned}$$

Traditionally (and probably incorrectly) attributed to the English cleric Thomas Bayes (1702–1761), the theorem that bears his name is used to compute conditional probabilities “the other way around.” One can interpret the  $A_j$  as unobservable states of nature, each of which has some *a priori* probability  $P(A_j)$  of occurring, together with a probability  $P(B|A_j)$  of “causing” an observable event  $B$  to occur. Then, given that  $B$  actually has occurred, (1.19) calculates the revised, *a posteriori* probabilities  $P(A_j|B)$  of the states  $A_j$ .

**Proposition 1.45 (Bayes’ theorem).** Suppose that  $\{A_1, A_2, \dots\}$  is a partition of  $\Omega$ . Then, for each event  $B$  with  $P(B) > 0$  and each  $j$ ,

$$P(A_j|B) = \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{\infty} P(B|A_i)P(A_i)}. \quad (1.19)$$

**Proof:** By two applications of Definition 1.43,

$$\begin{aligned} P(A_j|B) &= \frac{P(A_j \cap B)}{P(B)} = \frac{P(B|A_j)P(A_j)}{P(B)} \\ &= \frac{P(B|A_j)P(A_j)}{\sum_{i=1}^{\infty} P(B|A_i)P(A_i)}, \end{aligned}$$

where the last step is by Proposition 1.44. ■

## 1.6 Complements

### 1.6.1 The extended real numbers

The *extended real number system* is the set  $\overline{\mathbb{R}} = \mathbb{R} \cup \{-\infty\} \cup \{\infty\}$ , together with the following extended rules of arithmetic:

$$x + y = \begin{cases} \infty & \text{if } x \in \mathbb{R} \text{ and } y = \infty \text{ or vice versa} \\ \infty & \text{if } x = y = \infty \\ -\infty & \text{if } x \in \mathbb{R} \text{ and } y = -\infty \text{ or vice versa} \\ -\infty & \text{if } x = y = -\infty \end{cases}$$

$$xy = \begin{cases} \infty & \text{if } x > 0 \text{ and } y = \infty \text{ or vice versa} \\ \infty & \text{if } x < 0 \text{ and } y = -\infty \text{ or vice versa} \\ \infty & \text{if } x = y = \infty \text{ or } x = y = -\infty \\ -\infty & \text{if } x > 0 \text{ and } y = -\infty \text{ or vice versa} \\ -\infty & \text{if } x < 0 \text{ and } y = \infty \text{ or vice versa} \\ -\infty & \text{if } x = \infty \text{ and } y = -\infty \text{ or vice versa} \\ 0 & \text{if } x = 0 \text{ or } y = 0 \end{cases}$$

$$x/y = \begin{cases} \infty & \text{if } x > 0 \text{ and } y = 0 \\ -\infty & \text{if } x < 0 \text{ and } y = 0 \\ 0 & \text{if } x \in \mathbb{R} \text{ and } y = \pm\infty. \end{cases}$$

The operations  $\infty - \infty$  and  $\pm\infty / \pm\infty$  are *not defined*.

Convergence in  $\overline{\mathbb{R}}$  or  $\overline{\mathbb{R}}_+ = [0, \infty]$  is identical to that in  $\mathbb{R}$ , except that  $\pm\infty$  are allowable limits, provided that the divergence is unambiguous. Thus, for example, *every* monotone sequence has a limit in  $\overline{\mathbb{R}}$ .

## 1.6.2 Measures

Measures generalize probabilities: they are positive and countably additive, but the “measure” of the entire space need not be one. Indeed, it need not even be finite, although the measure of the empty must be zero.

**Definition 1.46.** Let  $(E, \mathcal{E})$  be a *measurable space*, consisting of a set  $E$  and a  $\sigma$ -algebra  $\mathcal{E}$  of subsets of it.

- a) A *measure* on  $(E, \mathcal{E})$  is a set function  $\mu: \mathcal{E} \rightarrow \overline{\mathbb{R}}_+$  such that  $\mu(\emptyset) = 0$  and  $\mu$  is countably additive: for disjoint sets  $A_1, A_2, \dots$  in  $\mathcal{E}$ ,

$$\mu\left(\sum_{n=1}^{\infty} A_n\right) = \sum_{n=1}^{\infty} \mu(A_n).$$

- b) A measure  $\mu$  is *finite* if  $\mu(E) < \infty$ .  
c) A measure  $\mu$  is  *$\sigma$ -finite* if there is a countable partition  $E = \sum_{n=1}^{\infty} A_n$  such that  $\mu(A_n) < \infty$  for each  $n$ .  
d) The triple  $(E, \mathcal{E}, \mu)$  is a *measure space*.  $\square$

In the countable additivity condition, we explicitly allow  $\infty = \infty$ . The sum there is infinite if  $\mu(A_n) = \infty$  for some  $n$  or if  $\sum_{n=1}^{\infty} \mu(A_n)$  diverges.

Finite measures do not differ dramatically from probabilities, since every finite measure  $\mu$  can be written as  $\mu(\cdot) = \mu(E)P(\cdot)$ , where  $P$  is a probability. The  $\sigma$ -finite measures, though, are truly more general than probabilities, yet still tractable mathematically.

The properties of measures mimic those of probabilities, with one important exception:  $A_n \downarrow \emptyset$  does not imply that  $\mu(A_n) \downarrow 0$  unless  $\mu(A_n) < \infty$  for some  $n$ .

### 1.6.3 Lebesgue measure

The most important measure of all is Lebesgue measure on  $\mathbb{R}$ , which defines a “length” for every Borel set. Given an interval  $I$  with endpoints  $a$  and  $b$ , the length of  $I$ , denoted by  $|I|$ , is  $b - a$ . This makes sense as well if  $a = -\infty$  or  $b = \infty$ , in which case  $|I| = \infty$ .

**Theorem 1.47.** There exists a unique  $\sigma$ -finite measure  $\lambda$  on  $(\mathbb{R}, \mathcal{B}(\mathbb{R}))$ , known as *Lebesgue measure*, such that  $\lambda(I) = |I|$  for every interval  $I$ .  $\square$

For  $A \in \mathcal{B}(\mathbb{R})$ ,  $\lambda(A)$  is termed the Lebesgue measure of  $A$ . Sets of Lebesgue measure zero are assigned a special name.

**Definition 1.48.** A set  $A \in \mathcal{B}(\mathbb{R})$  is a (Lebesgue) *null set* if  $\lambda(A) = 0$ . A property of real numbers holds *almost everywhere* if the set on which it is *not true* is a null set.  $\square$

Given a subset  $A$  of  $\mathbb{R}$  and  $y \in \mathbb{R}$ , the set  $A + y = \{x + y : y \in A\}$  is known as the *translate* of  $A$  by  $y$ . For example, if  $A = (a, b]$ , then  $A + y = (a + y, b + y]$ . Length is *translation invariant*:  $|I + x| = |I|$  for every interval  $I$  and  $x \in \mathbb{R}$ .

Lebesgue measure is characterized by  $\sigma$ -finiteness and translation invariance.

**Theorem 1.49.** Let  $\mu$  be a  $\sigma$ -finite, translation invariant measure on  $\mathcal{B}(\mathbb{R})$ . Then, there is a constant  $c \geq 0$  such that  $\mu = c\lambda$ .  $\square$

Lebesgue measure is defined in higher dimensions as well. In  $\mathbb{R}^n$ , Lebesgue measure  $\lambda$  is a  $\sigma$ -finite, translation invariant measure satisfying

$$\lambda\left(\prod_{i=1}^n (a_i, b_i]\right) = \prod_{i=1}^n (b_i - a_i)$$

for all choices of  $a_i \leq b_i$ ,  $i = 1, \dots, n$ . In particular, Lebesgue measure in two and three dimensions extends area and volume to all Borel sets.

### 1.6.4 Singular probabilities on $\mathbb{R}$

We introduce a third class of probabilities on  $\mathbb{R}$ .

**Definition 1.50.** A probability  $P$  on  $\mathbb{R}$  is *singular* if there exists a Lebesgue null set  $A$  such that  $F'_P(t)$  exists and is zero for all  $t \notin A$ .  $\square$

Every discrete probability is singular:  $F'_P$  is zero except at the countable (and, hence, null) set of jump points. However, a probability can be singular and yet have a continuous distribution function.

**Example 1.51 (Cantor distribution).** We first construct the Cantor subset  $C$  of  $[0, 1]$ . At the initial step, the open interval  $(1/3, 2/3)$  is removed, followed by the intervals  $(1/9, 2/9)$  and  $(7/9, 8/9)$ , then the intervals  $(1/27, 2/27)$ ,  $(7/27, 8/27)$ ,  $(19/27, 20/27)$  and  $(25/27, 26/27)$ , and so on. After  $n$  steps,  $2^n - 1$  intervals have been removed, and there remain  $2^n$  disjoint closed intervals, each of length  $3^{-n}$ . For each  $n$ , let  $J_{nk}$ ,  $k = 1, \dots, 2^n$ , denote the intervals removed in steps  $1, \dots, n$ , numbered from left to right. The *Cantor set*

$$C = \left( \bigcup_{n=1}^{\infty} \bigcup_{k=1}^{2^n-1} J_{nk} \right)^c, \quad (1.20)$$

consists of those points never removed. That  $C$  is uncountably infinite will not be proved here.

For  $1 \leq k \leq 2^n - 1$ , let  $c_{nk} = k/2^n$ . Since  $J_{n+1,2k} = J_{nk}$  and  $c_{n+1,2k} = c_{nk}$  we can define a function  $\tilde{F}: C^c \rightarrow [0, 1]$  by putting  $\tilde{F}(t) = c_{nk}$  for  $t \in J_{nk}$ . Moreover,  $\tilde{F}$  is increasing and uniformly continuous on  $C^c$ , and, hence (because  $C^c$  is dense in  $[0, 1]$ ), there is a uniformly continuous function  $F$  such that  $F = \tilde{F}$  on  $C^c$ . Since by construction  $F(0) = 0$  and  $F(1) = 1$ ,  $F$  is a distribution function, which we call the *Cantor distribution function*.

We now verify that  $F$  is singular. By construction,  $F' = 0$  on  $C^c$ , so it is enough to show that  $C$  is a null set. Also, (1.20) implies that for each  $n$ ,  $C$  is a subset of  $\left( \sum_{k=1}^{2^n-1} J_{nk} \right)^c$ , a set which consists of  $2^n$  disjoint intervals  $I_{nk}$ , each of length  $3^{-n}$ . Thus,  $\sum_{k=1}^{2^n} |I_{nk}| = (2/3)^n \rightarrow 0$ , which proves that  $C$  is a null set.  $\square$

### 1.6.5 Representation of probabilities on $\mathbb{R}$

Every probability on  $\mathbb{R}$  is the convex combination of three probabilities, one discrete, the second absolutely continuous and the third singular, but with a continuous distribution function.

**Theorem 1.52.** Every probability  $P$  on  $\mathbb{R}$  has a unique representation

$$P = \alpha_d P_d + \alpha_a P_a + \alpha_s P_s,$$

where  $\alpha_d \geq 0$ ,  $\alpha_a \geq 0$ ,  $\alpha_s \geq 0$ ,  $\alpha_d + \alpha_a + \alpha_s = 1$ ,  $P_d$  is discrete,  $P_a$  is absolutely continuous, and  $P_s$  is singular with continuous distribution function  $F_{P_s}$ .  $\square$

## 1.7 Exercises

**1.1.** Prove *de Morgan's laws*:

$$(A \cup B)^c = A^c \cap B^c$$

and

$$(A \cap B)^c = A^c \cup B^c,$$

both directly and using indicator functions.

**1.2.** Prove that for events  $A$  and  $B$ ,  $A\Delta B = A^c\Delta B^c$ .

**1.3.** Let  $B$  and  $C$  be events  $(A_n)$  and let  $A_n = B$  if  $n$  is odd and  $A_n = C$  if  $n$  is even. Calculate  $\limsup_n A_n$  and  $\liminf_n A_n$ .

**1.4.** Prove part b) of Proposition 1.9.

**1.5.** Prove that if  $t_n \downarrow t$ , then  $(-\infty, t_n] \downarrow (-\infty, t]$ , but that if  $t_n \uparrow t$  and  $t_n < t$  for each  $n$ , then  $(-\infty, t_n] \uparrow (-\infty, t)$ .

**1.6.** Show that the disjointification procedure used to prove Proposition 1.24 actually yields  $\bigcup_{n=1}^k A_n = \sum_{n=1}^k B_n$  for every  $k$ .

**1.7.** Prove (1.1) through (1.4).

**1.8.** Let  $A_1, A_2, \dots$  be events. Prove that  $A_n \rightarrow A$  if and only if  $\mathbf{1}_{A_n} \rightarrow \mathbf{1}_A$  as functions on  $\Omega$  (that is,  $\mathbf{1}_{A_n}(\omega) \rightarrow \mathbf{1}_A(\omega)$  for every  $\omega$ ).

**1.9.** Given subsets  $A$  and  $B$  of  $\Omega$ , identify all sets in  $\sigma(A, B)$ .

**1.10.** Prove that  $\{x\}$  is a Borel set for every  $x \in \mathbb{R}$ .

**1.11.** Let  $A$ ,  $B$  and  $C$  be disjoint events with  $P(A) = .6$ ,  $P(B) = .3$  and  $P(C) = .1$ . Calculate the probabilities of all events in the  $\sigma$ -algebra generated by  $\{A, B, C\}$ .

**1.12.** Verify that (1.8) defines a probability.

**1.13.** Confirm that (1.9) defines a probability.

**1.14.** Prove that for all events  $A$  and  $B$ ,

$$\begin{aligned} P(A\Delta B) &= P(A \cup B) - P(A \cap B) \\ &= P(A) + P(B) - 2P(A \cap B). \end{aligned}$$

**1.15.** Let  $\Omega$  be a finite set and let  $\mathcal{F} = \mathcal{P}(\Omega)$  be the  $\sigma$ -algebra of all subsets of  $\Omega$ . For  $A \subseteq \Omega$ , let  $|A|$  be the number of elements in  $A$ . Prove that the formula  $P(A) = |A|/|\Omega|$  defines a probability on  $(\Omega, \mathcal{F})$  (the uniform distribution on  $\Omega$ ).

- 1.16.** Prove that (1.15) and (1.16) are equivalent.
- 1.17.** Prove that there does not exist a uniform distribution on the set  $\mathbb{N} = \{0, 1, \dots\}$ .
- 1.18.** Suppose that  $P_1$  and  $P_2$  are probabilities on  $(\Omega, \mathcal{F})$  and that  $0 \leq \alpha \leq 1$ . Prove that the set function

$$P(A) = \alpha P_1(A) + (1 - \alpha)P_2(A)$$

is also a probability.

- 1.19.** Prove that if  $P(A_i) = 1$  for each  $i$ , then

$$P(\bigcap_{i=1}^{\infty} A_i) = 1.$$

- 1.20.** Show that for events  $A$  and  $B$ ,

$$P(A \cap B) \geq P(A) + P(B) - 1.$$

- 1.21.** Let  $P$  be a probability on  $\mathbb{R}$ . Prove that for every  $\varepsilon > 0$  there is a compact set  $K$  such that  $P(K) > 1 - \varepsilon$ .
- 1.22.** Prove that a distribution function on  $\mathbb{R}$  has at most countably many points of discontinuity.
- 1.23.** Prove that if  $P(A) > 0$ , then the set function

$$P_A(B) = P(B|A)$$

is a probability on  $(\Omega, \mathcal{F})$  satisfying  $P_A(A^c) = 0$ .

- 1.24.** Let  $P$  be the uniform distribution on a finite set  $\Omega$  and let  $A$  be a subset of  $\Omega$ . Prove that  $P(\cdot|A)$  is the uniform distribution on  $A$ .
- 1.25.** Let  $A_1, \dots, A_n$  be events, and for  $J \subseteq \{1, \dots, n\}$ , let  $B_J = \bigcap_{j \in J} A_j$ . For  $k \geq 1$ , let  $S_k = \sum_{|J|=k} P(B_J)$ , where the sum is over all subsets  $J$  of  $\{1, \dots, n\}$  with  $|J| = k$ .

- a) Prove the *inclusion-exclusion principle*:

$$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{k=1}^n (-1)^{k-1} S_k.$$

- b) Suppose that  $P(B_J)$  depends only on  $|J|$ , i.e., there are numbers  $q_0, \dots, q_n$  such that  $P(B_J) = q_k$  whenever  $|J| = k$ . (This is true, for example, in Exercise 1.15.) Prove that  $S_k = \binom{n}{k} q_k$ .

**1.26.** This is known as the *coupon collector's problem*. There are  $t$  different types of coupons available and the collector is seeking to collect one of each (for example, in order to win some premium). Show that if  $n$  coupons have been collected, then the probability  $p_n$  of having at least one of each type is

$$p_n = \sum_{k=1}^t (-1)^{k-1} \binom{t}{k} \left(1 - \frac{k}{t}\right)^n.$$

**1.27.** Consider an urn that initially contains  $r$  red balls and  $b$  black balls. At each trial one ball is drawn. It is replaced and  $c \geq 0$  balls of the same color added to the urn. Let  $A_j$  be the event that the  $j$ th ball drawn is black. Show that  $P(A_j) = b/(b+r)$  for every  $j$ . [This is the *Pólya urn scheme*, analyzed using martingales in Chapter 9.]

# Chapter 2

## Random Variables

Random variables, the essence of probability, comprise the subject matter of this chapter and much of the remainder of the book.

We begin with the concept of inverse images of sets. Let  $(\Omega, \mathcal{F}, P)$  be a probability space.

**Definition 2.1.** Let  $X$  be a function from  $\Omega$  to  $\mathbb{R}$ . The *inverse image under  $X$*  of  $B \in \mathcal{B}(\mathbb{R})$  is the subset of  $\Omega$  given by

$$X^{-1}(B) = \{\omega : X(\omega) \in B\},$$

which we abbreviate as  $\{X \in B\}$ .  $\square$

The inverse image mapping  $X^{-1}$  maps subsets of  $\mathbb{R}$  to subsets of  $\Omega$ . It preserves all set operations, as well as disjointness.

**Proposition 2.2.** Let  $B, B'$  and  $\{B_\alpha : \alpha \in I\}$  be Borel sets. Then,

- a) If  $B \subseteq B'$ , then  $X^{-1}(B) \subseteq X^{-1}(B')$ .
- b)  $X^{-1}(\bigcup_I B_\alpha) = \bigcup_I X^{-1}(B_\alpha)$  and  $X^{-1}(\bigcap_I B_\alpha) = \bigcap_I X^{-1}(B_\alpha)$ .
- c) If  $B$  and  $B'$  are disjoint, then so are  $X^{-1}(B)$  and  $X^{-1}(B')$ .
- d)  $X^{-1}(B^c) = [X^{-1}(B)]^c$ .

**Proof:** We prove the first part of b). If  $\omega \in X^{-1}(\bigcup_I B_\alpha)$ , then  $X(\omega) \in \bigcup_I B_\alpha$ , so  $X(\omega) \in B_{\alpha_0}$  for some  $\alpha_0$ . But this says that  $\omega \in X^{-1}(B_{\alpha_0})$ , and hence  $\omega \in \bigcup_I X^{-1}(B_\alpha)$ . Conversely, if  $\omega \in \bigcup_I X^{-1}(B_\alpha)$ , there is  $\alpha_0$  such that  $X(\omega) \in B_{\alpha_0} \subseteq \bigcup_I B_\alpha$ , so that  $\omega \in X^{-1}(\bigcup_I B_\alpha)$ . ■

Functions whose range is finite play a distinguished role. A function  $X : \Omega \rightarrow \mathbb{R}$  is *simple* if there is a finite subset  $H$  of  $\mathbb{R}$  such that  $X(\omega) \in H$  for all  $\omega \in \Omega$ .

## 2.1 Fundamentals

Random variables are functions on the sample space for which inverse images of Borel sets are events.

### 2.1.1 Random variables

**Definition 2.3.** A *random variable* is a function  $X: \Omega \rightarrow \mathbb{R}$  such that  $X^{-1}(B) \in \mathcal{F}$  for every  $B \in \mathcal{B}(\mathbb{R})$ .  $\square$

For a random variable  $X$  and  $B \in \mathcal{B}(\mathbb{R})$ , we write  $\mathbf{1}(X \in B)$  for the indicator function of the event  $\{X \in B\}$ . More generally, if  $X_1, \dots, X_n$  are random variables, we put

$$\{X_1 \in B_1, \dots, X_n \in B_n\} = \bigcap_{i=1}^n \{X_i \in B_i\},$$

and we use  $\mathbf{1}(X_1 \in B_1, \dots, X_n \in B_n)$  to denote the indicator function of this event.

The technical requirement that sets  $\{X \in B\}$  be events is needed in order that probabilities

$$P\{X \in B\} \stackrel{\text{def}}{=} P(X^{-1}(B))$$

be defined. However, the key idea is that a random variable is a *function on the sample space*.

Two fundamental classes are indicator random variables, which take only the values 0 and 1, and simple random variables, which take finitely many values.

**Example 2.4 (Indicator random variable).** The indicator function of an event  $A$  is a random variable: for each  $B$ ,

$$\{\mathbf{1}_A \leq t\} = \begin{cases} \emptyset & \text{if } 0 \notin B, 1 \notin B \\ A^c & \text{if } 0 \in B, 1 \notin B \\ A & \text{if } 0 \notin B, 1 \in B \\ \Omega & \text{if } 0 \in B, 1 \in B, \end{cases}$$

and in each case  $\{\mathbf{1}_A \leq t\} \in \mathcal{F}$ .

Conversely, if  $A$  is subset of  $\Omega$  such that  $\mathbf{1}_A$  is a random variable, then since  $A = \{\mathbf{1}_A = 1\}$ ,  $A$  is an event. Thus, indicator functions link events and random variables.  $\square$

In particular, constant functions are random variables, no matter what the  $\sigma$ -algebra  $\mathcal{F}$ .

**Example 2.5 (Simple random variable).** A simple random variable  $X$  takes only finitely many values, and so has the form  $X = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$ , where the  $a_i$  are real numbers and the  $A_i$  are events. In such representations, we always suppose that the  $A_i$  constitute a partition of  $\Omega$ . We do not, however, require that the  $a_i$  be distinct or nonzero. That  $X$  is a random variable holds true, since for  $B \in \mathcal{B}(\mathbb{R})$ ,

$$\{X \in B\} = \bigcup \{A_i : a_i \in B\}.$$

The role of simple random variables as “elementary functions” in probability will emerge in this and later chapters.  $\square$

Random vectors, stochastic processes and complex-valued random variables are three generalizations of random variables.

### 2.1.2 Random vectors

The components of a random vector are random variables.

**Definition 2.6.** A random  $d$ -vector is a function  $X = (X_1, \dots, X_d)$  from  $\Omega$  to  $\mathbb{R}^d$  such that each component  $X_i$  is a random variable.  $\square$

Random vectors will sometimes be treated as “random elements” of  $\mathbb{R}^d$ , and other times as finite sequences of random variables.

### 2.1.3 Stochastic processes

A stochastic processes is any indexed family of random variables.

**Definition 2.7.** Given any set  $T$ , a stochastic process with index set  $T$  is a collection  $\{X_t : t \in T\}$  of random variables indexed by  $T$ .  $\square$

Typically (but not always), the index of a stochastic process represents time. Thus, a sequence  $(X_n)_{n \geq 0}$  is a *discrete time* stochastic process, and a family  $\{X_t : t \geq 0\}$  a *continuous time* stochastic process. Sometimes the *sample paths*  $t \mapsto X_t(\omega)$  are of central interest, to the point that a stochastic process is defined as a random element of some space of functions.

### 2.1.4 Complex-valued random variables

Random variables taking values in the set  $\mathbb{C}$  of complex numbers are needed for characteristic functions (Chapter 6). Their real and imaginary parts are (ordinary) random variables.

**Definition 2.8.** A function  $Z : \Omega \rightarrow \mathbb{C}$  is a *complex-valued random variable* if  $Z = X + iY$ , where  $X$  and  $Y$  are random variables.  $\square$

### 2.1.5 The $\sigma$ -algebra generated by a random variable

The family of events that are inverse images of Borel sets under a random variable is a  $\sigma$ -algebra on  $\Omega$ .

**Proposition 2.9.** Given a random variable  $X$ , the family

$$\sigma(X) = \{X^{-1}(B) : B \in \mathcal{B}(\mathbb{R})\}$$

is a  $\sigma$ -algebra on  $\Omega$ , known as the  *$\sigma$ -algebra generated by  $X$* .

**Proof:** Clearly  $\Omega = X^{-1}(\mathbb{R})$  belongs to  $\sigma(X)$ . If each of the sets  $X^{-1}(B_i)$  belongs to  $\sigma(X)$ , then since  $\mathcal{B}(\mathbb{R})$  is a  $\sigma$ -algebra,  $\bigcup_{i=1}^{\infty} B_i \in \mathcal{B}(\mathbb{R})$ , and since by Proposition 2.2,

$$\bigcup_{i=1}^{\infty} X^{-1}(B_i) = X^{-1}\left(\bigcup_{i=1}^{\infty} B_i\right),$$

we conclude that  $\bigcup_{i=1}^{\infty} X^{-1}(B_i) \in \sigma(X)$ .

That  $\sigma(X)$  is closed under countable intersection and complementation is proved in the same manner. ■

In fact,  $\sigma(X)$  is a  $\sigma$ -algebra for every function  $X: \Omega \rightarrow \mathbb{R}$ ;  $X$  is a random variable if and only if  $\sigma(X) \subseteq \mathcal{F}$ .

We continue the two key examples.

**Example 2.10 (Indicator random variable).** If  $A \in \mathcal{F}$ , then  $\sigma(\mathbf{1}_A) = \{A, A^c, \emptyset, \Omega\}$ . □

**Example 2.11 (Simple random variable).** For a simple random variable  $X = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$ , with  $a_1, \dots, a_n$  distinct,

$$\sigma(X) = \sigma(\{A_1, \dots, A_n\}) = \{\sum_{i \in I} A_i : I \subseteq \{1, \dots, n\}\}$$

regardless of the values of the  $a_i$ . □

Example 2.11 shows how  $\sigma(X)$  depends more on the “number” of values assumed by  $X$  than on precisely what those values are. Indeed, two random variables that are functions of each other generate the same  $\sigma$ -algebra.

We can extend to random variables  $X_1, \dots, X_d$ . For each  $d$ -dimensional Borel set  $B$ , let  $\{(X_1, \dots, X_d) \in B\}$  be the inverse image of  $B$  under the random vector  $(X_1, \dots, X_d)$ , which maps  $\Omega$  to  $\mathbb{R}^d$ . Then, the  $\sigma$ -algebra generated by  $X_1, \dots, X_d$  is

$$\sigma(X_1, \dots, X_d) = \left\{ \{(X_1, \dots, X_d) \in B\} : B \in \mathcal{B}(\mathbb{R}^d) \right\}. \quad (2.1)$$

### 2.1.6 Simplified criteria

To verify that a function  $X$  is a random variable, it is unnecessary to check that  $\{X \in B\} \in \mathcal{F}$  for all Borel sets  $B$ . The  $\sigma$ -algebra  $\sigma(X)$  is contained in  $\mathcal{F}$  once  $X^{-1}(B) \in \mathcal{F}$  for enough “elementary” Borel sets. In particular, it suffices that  $\{X \leq t\} = X^{-1}((-\infty, t])$  be an event for each  $t$ .

**Proposition 2.12.** A function  $X: \Omega \rightarrow \mathbb{R}$  is a random variable if  $\{X \leq t\} \in \mathcal{F}$  for every  $t \in \mathbb{R}$ .

**Proof:** Let  $\mathcal{G}$  be the family of Borel sets for which  $X^{-1}(B)$  belongs to  $\mathcal{F}$ . By an argument similar to that used to prove Proposition 2.9,  $\mathcal{G}$  is a  $\sigma$ -algebra on  $\mathbb{R}$ , and contains all intervals. Hence,  $\mathcal{G} = \mathcal{B}(\mathbb{R})$  by Theorem 1.17. ■

Similarly,  $X$  is a random variable if  $\{X > t\} \in \mathcal{F}$  for every  $t \in \mathbb{R}$ , or if  $\{X < t\} \in \mathcal{F}$  for every  $t \in \mathbb{R}$ , or, yet again, if  $\{X \geq t\} \in \mathcal{F}$  for every  $t \in \mathbb{R}$ .

## 2.2 Combining Random Variables

To work with random variables, we need assurance that algebraic, limiting and transformation operations, applied to them, yield other random variables. All of these operations are defined  $\omega$ -wise, that is, pointwise on random variables as functions on  $\Omega$ .

### 2.2.1 Algebraic operations

The set of random variables is closed under addition and scalar multiplication — and is, hence, a vector space, under maximum and minimum, and under multiplication and division.

**Proposition 2.13.** Let  $X$  and  $Y$  be random variables. Then,

- a)  $aX + bY$  is a random variable for all  $a, b \in \mathbb{R}$ .
- b)  $\max\{X, Y\}$  and  $\min\{X, Y\}$  are random variables.
- c)  $XY$  is a random variable.
- d) Provided that  $Y(\omega) \neq 0$  for each  $\omega$ ,  $X/Y$  is a random variable.

**Proof:** We apply Proposition 2.12 throughout.

- a) We show that  $X + Y$  and  $aX$  are random variables. For each  $t$ ,

$$\{X + Y < t\} = \bigcup_{r \in \mathbb{Q}} (\{X < r\} \cap \{Y < t - r\}),$$

where  $\mathbb{Q}$  is the countable set of rationals, and, hence,  $X + Y$  is a random variable.

If  $a > 0$ , then for each  $t$ ,  $\{aX \leq t\} = \{X \leq t/a\}$ , which is an event, while if  $a < 0$ ,  $\{aX \leq t\} = \{X \geq t/a\}$ , which is again an event. In either case, then,  $aX$  is a random variable.

b) For each  $t$ ,

$$\{\max\{X, Y\} \leq t\} = \{X \leq t\} \cap \{Y \leq t\},$$

from which it follows that  $\max\{X, Y\}$  is a random variable. Similarly,

$$\{\min\{X, Y\} \leq t\} = \{X \leq t\} \cup \{Y \leq t\},$$

which shows that  $\min\{X, Y\}$  is a random variable.

c) Although it also follows from Corollary 2.19, we show directly that  $X^2$  is a random variable: for  $t \geq 0$ ,

$$\{X^2 \leq t\} = \{-\sqrt{t} \leq X \leq \sqrt{t}\} = \{X \leq \sqrt{t}\} \setminus \{X < -\sqrt{t}\},$$

which belongs to  $\mathcal{F}$ . But then

$$XY = \frac{1}{2}[(X+Y)^2 - X^2 - Y^2]$$

is a random variable by a).

d) First of all,  $1/Y$  is a random variable since

$$\{1/Y \leq t\} = \begin{cases} \{Y \geq 1/t\} \cup \{Y \leq 0\} & \text{if } t \geq 0 \\ \{Y \leq 1/t\} & \text{if } t < 0. \end{cases}$$

Finally,  $X/Y = X(1/Y)$  is a random variable by c). ■

The properties in part b) of Proposition 2.13 mirror the closure properties of  $\mathcal{F}$ . Especially, by (1.1) and (1.2), maxima and minima of indicator functions correspond to union and intersection of events.

Before noting another consequence of Proposition 2.13, we introduce some important terminology.

**Definition 2.14.** Given a function  $X: \Omega \rightarrow \mathbb{R}$ , the *positive part* of  $X$  is the function

$$X^+ = \max\{X, 0\},$$

and the *negative part* of  $X$  is the function

$$X^- = -\min\{X, 0\}. \quad \square$$

These are positive functions with

$$X = X^+ - X^-$$

and

$$|X| = X^+ + X^-,$$

and comprise a canonical representation of  $X$  as the difference of positive functions.

**Corollary 2.15 (to Proposition 2.13).** Whenever  $X$  is a random variable, so are  $X^+$ ,  $X^-$  and  $|X|$ .  $\square$

## 2.2.2 Limiting operations

We next consider operations associated with sequences of random variables.

**Theorem 2.16.** Let  $X_1, X_2, \dots$  be random variables. Then,  $\sup_n X_n$ ,  $\inf_n X_n$ ,  $\limsup_n X_n$  and  $\liminf_n X_n$  are random variables. Consequently, if

$$X(\omega) = \lim_{n \rightarrow \infty} X_n(\omega)$$

exists for every  $\omega$ , then  $X$  is a random variable.

**Proof:** For each  $t$ ,

$$\{\sup_n X_n \leq t\} = \bigcap_n \{X_n \leq t\},$$

so that  $\sup_n X_n$  is a random variable. In the same way,

$$\{\inf_n X_n \geq t\} = \bigcap_n \{X_n \geq t\}$$

for each  $t$ , so that  $\inf_n X_n$  is a random variable. But then, the functions

$$\limsup_n X_n = \inf_k \sup_{m \geq k} X_m$$

and

$$\liminf_n X_n = \sup_k \inf_{m \geq k} X_m$$

are likewise random variables.

When  $X = \lim_n X_n$  exists,  $X = \limsup_n X_n$  as well.  $\blacksquare$

Series are but another form of limits.

**Corollary 2.17.** If  $X_1, X_2, \dots$  are random variables, then provided that  $X(\omega) = \sum_{n=1}^{\infty} X_n(\omega)$  converges for each  $\omega$ ,  $X$  is a random variable.  $\square$

**Technical Aside.** We tacitly assume in Theorem 2.16 that the functions there are finite-valued, which, of course, need not always be true. In a more general setting, one defines random variables to be functions taking values in  $\bar{\mathbb{R}}$ , satisfying  $\{X = \infty\} \in \mathcal{F}$ ,  $\{X = -\infty\} \in \mathcal{F}$  and  $\{X \in B\} \in \mathcal{F}$  for each  $B \in \mathcal{B}(\mathbb{R})$ .  $\square$

### 2.2.3 Transformations

Another important way of constructing random variables is as functions of other random variables.

**Theorem 2.18.** Let  $X_1, \dots, X_d$  be random variables and let  $g: \mathbb{R}^d \rightarrow \mathbb{R}$  be a Borel measurable function (Definition 2.53). Then,  $Y = g(X_1, \dots, X_d)$  is a random variable.

**Proof:** We show first that  $\{(X_1, \dots, X_d) \in B\} \in \mathcal{F}$  for every  $B \in \mathcal{B}(\mathbb{R}^d)$ . If  $B = \prod_{i=1}^d B_i$  is a rectangle ( $B_i \in \mathcal{B}(\mathbb{R})$  for each  $i$ ), then

$$\{(X_1, \dots, X_d) \in B\} = \{X_1 \in B_1, \dots, X_d \in B_d\} = \bigcap_{i=1}^d \{X_i \in B_i\}$$

is an event because the  $X_i$  are random variables. The proof of Proposition 2.9 shows that the family of sets  $B \in \mathcal{B}(\mathbb{R}^d)$  with  $\{(X_1, \dots, X_d) \in B\} \in \mathcal{F}$  is a  $\sigma$ -algebra, which equals  $\mathcal{B}(\mathbb{R}^d)$  by Theorem 1.17.

For  $B \in \mathcal{B}(\mathbb{R})$ ,

$$\{Y \in B\} = \{(X_1, \dots, X_d) \in g^{-1}(B)\}.$$

The result then follows, since  $g^{-1}(B) \in \mathcal{B}(\mathbb{R}^d)$ . ■

**Corollary 2.19.** Let  $X$  be a random variable and let  $g: \mathbb{R} \rightarrow \mathbb{R}$  be Borel measurable. Then,  $g(X)$  is a random variable. □

### 2.2.4 Approximation of positive random variables

The following approximation property is crucial to the definition of expectation (§4.1).

**Theorem 2.20.** Given a positive random variable  $X$ , there are simple random variables  $0 \leq X_1 \leq X_2 \leq \dots$  with  $X_n(\omega) \uparrow X(\omega)$  for every  $\omega$ .

**Proof:** For each  $n$ , let

$$X_n = \sum_{k=1}^{n2^n} \frac{k-1}{2^n} \mathbf{1}\left(\frac{k-1}{2^n} < X \leq \frac{k}{2^n}\right) + n\mathbf{1}(X > n).$$

Then, the claimed properties hold. On the event  $\{(k-1)/2^n < X \leq k/2^n\}$ ,  $X_n$  is equal to the left endpoint  $(k-1)/2^n$ , which ensures both monotonicity of the sequence and that  $X_n \leq X$  for each  $n$ . ■

### 2.2.5 Monotone class theorems

At times it is necessary to prove that some set of functions contains all random variables or all positive random variables. This is true provided that the set contains indicator functions and is closed under appropriate operations, as the following analogues of Theorem 1.17 demonstrate.

**Theorem 2.21 (Monotone class theorem).** Let  $\mathcal{S}$  be a  $\pi$ -system generating  $\mathcal{F}$  and let  $\mathbf{H}$  be a set of functions on  $\Omega$  such that

- i) The constant function 1 belongs to  $\mathbf{H}$ ;
- ii)  $\mathbf{1}_A \in \mathbf{H}$  for all  $A \in \mathcal{S}$ ;
- iii)  $\mathbf{H}$  is a vector space;
- iv) If  $X_n \in \mathbf{H}$  for each  $n$  and  $\sup_n X_n(\omega) < \infty$  for each  $\omega$ , then  $\sup_n X_n$  belongs to  $\mathbf{H}$ .

Then,  $\mathbf{H}$  contains all random variables.

**Proof:** To begin, let  $\mathcal{G} = \{A \in \mathcal{F}: \mathbf{1}_A \in \mathbf{H}\}$ . Then,  $\mathcal{S} \subseteq \mathcal{G}$  by ii),  $\Omega \in \mathcal{G}$  by i), and iv) implies that  $\mathcal{G}$  is closed under countable increasing unions. Therefore,  $\mathcal{G}$  is a  $d$ -system, and by Theorem 1.17,  $\mathcal{G} = d(\mathcal{S}) = \sigma(\mathcal{S}) = \mathcal{F}$ .

As a vector space,  $\mathbf{H}$  thus contains all simple random variables, and since by iv) it is closed under increasing limits, Theorem 2.20 implies that it contains all positive random variables. But if  $X$  is a random variable, the random variables  $X^+$  and  $X^-$  belong to  $\mathbf{H}$ . Finally, since  $\mathbf{H}$  is a vector space,  $X \in \mathbf{H}$  because  $X = X^+ - X^-$ . ■

**Theorem 2.22 (Monotone class theorem, bis).** Let  $\mathcal{S}$  be a  $\pi$ -system generating  $\mathcal{F}$  and let  $\mathbf{H}$  be a set of positive functions on  $\Omega$  such that

- i) The constant function 1 belongs to  $\mathbf{H}$ ;
- ii)  $\mathbf{1}_A \in \mathbf{H}$  for all  $A \in \mathcal{S}$ ;
- iii) If  $X, Y \in \mathbf{H}$  and  $a, b \in \mathbb{R}_+$ , then  $aX + bY \in \mathbf{H}$  ( $\mathbf{H}$  is a cone);
- iv) If  $X_n \in \mathbf{H}$  for each  $n$  and  $\sup_n X_n(\omega) < \infty$  for each  $\omega$ , then  $\sup_n X_n$  belongs to  $\mathbf{H}$ .

Then,  $\mathbf{H}$  contains all positive random variables. □

## 2.3 Distributions and Distribution Functions

The main importance of probabilities on  $\mathbb{R}$  or  $\mathbb{R}^d$  is that they are distributions of random variables.

### 2.3.1 Random variables

Associated with every random variable is a probability on  $\mathbb{R}$ .

**Proposition 2.23.** For a random variable  $X$ , the set function  $P_X(B) = P(X^{-1}(B))$  is a probability on  $\mathbb{R}$ .

**Proof:** Clearly  $P_X(B) \geq 0$  for all  $B$ , and  $P_X(\mathbb{R}) = P\{X \in \mathbb{R}\} = P(\Omega) = 1$ . For disjoint Borel sets  $B_1, B_2, \dots$ ,  $X^{-1}(B_1), X^{-1}(B_2), \dots$  are disjoint events by Proposition 2.2, and hence

$$\begin{aligned} P_X\left(\sum_{i=1}^{\infty} B_i\right) &= P\left(X^{-1}\left(\sum_{i=1}^{\infty} B_i\right)\right) = P\left(\sum_{i=1}^{\infty} X^{-1}(B_i)\right) \\ &= \sum_{i=1}^{\infty} P(X^{-1}(B_i)) \\ &= \sum_{i=1}^{\infty} P_X(B_i). \quad \blacksquare \end{aligned}$$

We then associate to a random variable a distribution, distribution function and survivor function.

**Definition 2.24.** Let  $X$  be a random variable.

- a) The *distribution* of  $X$  is the probability  $P_X(B) = P(X^{-1}(B))$ .
- b) The *distribution function* of  $X$  is  $F_X(t) = P_X((-\infty, t]) = P\{X \leq t\}$ .
- c) The *survivor function* of  $X$  is  $S_X(t) = 1 - F_X(t) = P\{X > t\}$ .  $\square$

We say that  $X$  is *discrete* or *absolutely continuous* if  $P_X$  is. A density function for  $P_X$ , if it exists, is called the *density of  $X$*  and denoted by  $f_X$ . Although only integrals of the density function literally are probabilities, the following heuristic interpretation is important to the formulation and understanding of numerous results:

$$P\{X \in (x, x + dx)\} \cong f_X(x) dx, \quad x \in \mathbb{R}. \quad (2.2)$$

Random variables with the same distribution function arise sufficiently often to merit special terminology.

**Definition 2.25.** Random variables  $X$  and  $Y$  are *identically distributed* if  $F_X = F_Y$  (equivalently,  $P_X = P_Y$ ). We denote this by  $X \stackrel{d}{=} Y$ .  $\square$

Equality in distribution of  $X$  and  $Y$  has no bearing on their equality as functions on  $\Omega$ . One can have  $X \stackrel{d}{=} Y$  even though  $P\{X = Y\} = 0$ . The relevant notion of equality for random variables as functions, indeed, is not that  $X(\omega) = Y(\omega)$  for every  $\omega$ , but only that  $P\{X = Y\} = 1$ . When this holds, in view of Definition 1.28, we say that  $X$  and  $Y$  are equal *almost surely*, which we denote by  $X \stackrel{\text{a.s.}}{=} Y$ .

### 2.3.2 Random vectors

The concepts just introduced extend to random vectors, as we now discuss.

**Definition 2.26.** Let  $X = (X_1, \dots, X_d)$  be a random  $d$ -vector.

- a) The *distribution* of  $X$  is the probability  $P_X(B) = P\{X \in B\}$  on  $\mathbb{R}^d$ .
- b) The *distribution function* of  $X$ , also known as the *joint distribution function* of  $X_1, \dots, X_d$ , is the function  $F_X: \mathbb{R}^d \rightarrow [0, 1]$  given by

$$F_X(t_1, \dots, t_d) = P\{X_1 \leq t_1, \dots, X_d \leq t_d\}. \quad \square$$

The distribution  $P_X$  is determined uniquely by  $F_X$ . The computational utility of multivariate distribution functions is (generally) much less than that of their one-dimensional counterparts. In principle, the distribution function of each component can be recovered from the joint distribution function, as the next result demonstrates, but the extent to which this is really feasible depends heavily on the form of  $F_X$ .

**Proposition 2.27.** Let  $X$  be a random  $d$ -vector. Then, for each  $i$  and  $t$ ,

$$F_{X_i}(t) = \lim_{t_j \rightarrow \infty, j \neq i} F_X(t_1, \dots, t_{i-1}, t, t_{i+1}, \dots, t_d).$$

**Proof:** As  $t_j \uparrow \infty$  for all  $j \neq i$ ,

$$\{X_1 \leq t_1, \dots, X_{i-1} \leq t_{i-1}, X_i \leq t, X_{i+1} \leq t_{i+1}, \dots, X_d \leq t_d\} \uparrow \{X_i \leq t\},$$

so the result follows from monotone continuity of  $P$ . ■

By analogy with the one-dimensional case, a random vector  $X$  is *discrete* if there is a countable subset  $C$  of  $\mathbb{R}^d$  such that  $P\{X \in C\} = 1$ , and *absolutely continuous* if there is a function  $f_X: \mathbb{R}^d \rightarrow \mathbb{R}_+$ , termed the *density* of  $X$  and *joint density* of  $X_1, \dots, X_d$ , such that

$$P\{X_1 \leq t_1, \dots, X_d \leq t_d\} = \int_{-\infty}^{t_1} \cdots \int_{-\infty}^{t_d} f_X(y_1, \dots, y_d) dy_1 \cdots dy_d. \quad (2.3)$$

Discreteness and absolute continuity of a random vector are inherited by its components. In the former case, the converse holds true as well.

**Proposition 2.28.** A random vector  $X = (X_1, \dots, X_d)$  is discrete if and only if  $X_1, \dots, X_d$  are discrete random variables.

**Proof:** If  $C \subset \mathbb{R}^d$  is countable and  $P\{X \in C\} = 1$ , then for each  $i$ ,  $P\{X_i \in C_i\} = 1$ , where  $C_i = \{x_i: x \in C\}$ . Conversely, if for each  $j$ ,

$C_j \subset \mathbb{R}$  is a countable set with  $P\{X_j \in C_j\} = 1$ , then with probability 1,  $X$  belongs to the countable set  $C = \prod_{j=1}^n C_j$ . ■

**Proposition 2.29.** If  $X = (X_1, \dots, X_d)$  is absolutely continuous, then for each  $i$ ,  $X_i$  is absolutely continuous, and

$$\begin{aligned} f_{X_i}(x) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(y_1, \dots, y_{i-1}, x, y_{i+1}, \dots, y_d) \\ &\quad \times dy_1 \cdots dy_{i-1} dy_{i+1} \cdots dy_d. \end{aligned} \quad (2.4)$$

**Proof:** From Proposition 2.27 and (2.3),

$$\begin{aligned} P\{X_i \leq t\} &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \int_{-\infty}^t \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(y_1, \dots, y_d) dy_1 \cdots dy_d \\ &= \int_{-\infty}^t \left[ \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f_X(y_1, \dots, y_{i-1}, x, y_{i+1}, \dots, y_d) \right. \\ &\quad \left. \times dy_1 \cdots dy_{i-1} dy_{i+1} \cdots dy_d \right] dx. \quad ■ \end{aligned}$$

Intuitively, (2.4) depicts “integrating out” the variables of the joint density other than that for  $X_i$ , and the result is termed a *marginal density function*. The procedure works more generally: any random “sub-vector”  $(X_{j_1}, \dots, X_{j_\ell})$  is absolutely continuous, and its joint density is obtained by integrating out the other variables in  $f_X$ .

The converse to Proposition 2.29 is not true: if  $X_1$  is uniformly distributed on  $[0, 1]$  and  $X_2 \equiv X_1$ , then  $(X_1, X_2)$  is not absolutely continuous even though the components are. The converse is true, however, when the components are independent, as shown in Corollary 3.6.

## 2.4 Key Random Variables and Distributions

We now introduce the important random variables and distributions.

### 2.4.1 Discrete random variables

Integer-valued random variables are the discrete random variables of greatest interest.

**Example 2.30 (Bernoulli distribution).** A random variable  $X$  has a *Bernoulli distribution* with parameter  $p \in [0, 1]$  if  $P\{X = 1\} = p = 1 - P\{X = 0\}$ . A Bernoulli distributed random variable  $X$  is the indicator function of the event  $\{X = 1\}$ . □

In Chapter 3 we relate Bernoulli and binomial distributions.

**Example 2.31 (Binomial distribution).** A random variable  $X$  has a *binomial distribution* with parameters  $n \geq 1$  and  $p \in [0, 1]$  if

$$P\{X = k\} = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, \dots, n.$$

To verify that these constitute legitimate probabilities, one must show that  $P\{X = k\} \geq 0$  for each  $k$ , which is obvious, and that  $\sum_{k=0}^n P\{X = k\} = 1$ , which follows from the binomial theorem:

$$\sum_{k=0}^n P\{X = k\} = \sum_{k=0}^n \binom{n}{k} p^k (1-p)^{n-k} = [p + (1-p)]^n = 1.$$

The binomial distribution with parameters  $n$  and  $p$  is denoted by  $B(n, p)$ .

We show in §3.6 that a random variable having a binomial distribution with parameters  $n$  and  $p$  can be interpreted as the number of successes in  $n$  independent trials, with success probability  $p$ , and as the sum of  $n$  independent, Bernoulli distributed random variables.  $\square$

In an infinite sequence of independent trials, the number of trials required to achieve the first success has a geometric distribution, as does the number of trials between any two consecutive successes.

**Example 2.32 (Geometric distribution).** A random variable  $X$  has a *geometric distribution* with parameter  $p \in (0, 1)$  if

$$P\{X = k\} = p(1-p)^{k-1}, \quad k \geq 1.$$

That these are allowable probabilities follows from

$$\sum_{k=1}^{\infty} P\{X = k\} = p \sum_{k=0}^{\infty} (1-p)^k = \frac{p}{1-(1-p)} = 1. \quad \square$$

Sums of independent, geometrically distributed random variables have negative binomial distributions.

**Example 2.33 (Negative binomial distribution).** A random variable  $X$  has a *negative binomial distribution* with parameters  $n \geq 1$  and  $p \in (0, 1)$  provided that

$$P\{X = k\} = \binom{k-1}{n-1} p^k (1-p)^{k-n}, \quad k \geq n.$$

The verification that  $\sum_{k=n}^{\infty} P\{X = k\} = 1$  can be done directly, but only with some effort; an easier approach is described in Chapter 3.  $\square$

Distribution	Parameters	Probabilities $P\{X = k\}$
Bernoulli	$p \in [0, 1]$	$p$ for $k = 1$ ; $1 - p$ for $k = 0$
Binomial	$n \geq 1, p \in [0, 1]$	$\binom{n}{k} p^k (1-p)^{n-k}, 0 \leq k \leq n$
Geometric	$p \in (0, 1)$	$p(1-p)^{k-1}, k \geq 1$
Negative binomial	$n \geq 1, p \in (0, 1)$	$\binom{k-1}{n-1} p^k (1-p)^{k-n}, k \geq n$
Poisson	$\lambda \in (0, \infty)$	$e^{-\lambda} \lambda^k / k!, k \geq 0$

Table 2.1. Discrete Distributions

We conclude with the Poisson distributions, perhaps the most important discrete distributions of all.

**Example 2.34 (Poisson distribution).** A random variable  $X$  has a *Poisson distribution* with parameter  $\lambda \in (0, \infty)$ , denoted by  $P(\lambda)$ , if

$$P\{X = k\} = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k \geq 0.$$

Since

$$\sum_{k=0}^{\infty} P\{X = k\} = e^{-\lambda} \sum_{k=0}^{\infty} \frac{\lambda^k}{k!} = e^{-\lambda} e^{\lambda} = 1,$$

this is an acceptable definition. Figure 1.2 shows the distribution function of a Poisson distributed random variable with  $\lambda = 1$ .

Poisson distributions correspond to numbers arrivals in fixed time intervals in Poisson processes (§3.6). Also, for large  $n$  and small  $p$ , a binomial distribution with parameters  $n$  and  $p$  is approximately a Poisson distribution with parameter  $\lambda = np$ , as described in §5.5.  $\square$

Table 2.1 summarizes the key discrete distributions.

## 2.4.2 Absolutely continuous random variables

The standard normal distribution — because of its role in the central limit theorem — is the most important of all probability distributions.

**Example 2.35 (Normal distribution).** A random variable  $X$  has the *standard normal distribution* if

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \in \mathbb{R}. \quad (2.5)$$

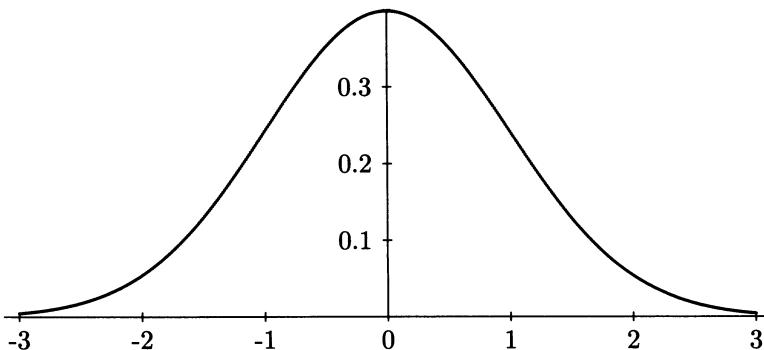


Figure 2.1. The Standard Normal Density Function

To verify that  $f_X$  is a density function, we observe that it is positive, and need to show that it integrates to one. With  $I = \int_{-\infty}^{\infty} f_X(x) dx$ , and by the change of variables  $(x, y) = (r \cos \theta, r \sin \theta)$ ,

$$\begin{aligned} I^2 &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} e^{-(x^2+y^2)/2} dx dy \\ &= \frac{1}{2\pi} \int_0^{2\pi} \left( \int_0^{\infty} r e^{-r^2/2} dr \right) d\theta \\ &= \frac{1}{2\pi} \int_0^{2\pi} d\theta, \\ &= 1. \end{aligned}$$

Figure 2.1 shows the standard normal density function pictorially.

More generally,  $Y$  has a *normal distribution* with parameters  $\mu \in \mathbb{R}$  and  $\sigma^2 > 0$  if

$$f_Y(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}, \quad x \in \mathbb{R}. \quad (2.6)$$

The computations required to show directly that (2.6) defines a density function are not difficult; nevertheless, is easier to observe that  $Y$  has density (2.6) if and only if  $X = (Y - \mu)/\sigma$  has the standard normal density (2.5), where  $\sigma$  is the positive square root of  $\sigma^2$ . This is verified in §5.

We denote by  $N(\mu, \sigma^2)$  the normal distribution with parameters  $\mu$  and  $\sigma^2$ ; the standard normal distribution is  $N(0, 1)$ .  $\square$

Interarrival times in a Poisson process have exponential distributions.

**Example 2.36 (Exponential distribution).** A positive random variable  $X$  has an *exponential distribution* with parameter  $\lambda \in (0, \infty)$ , denoted

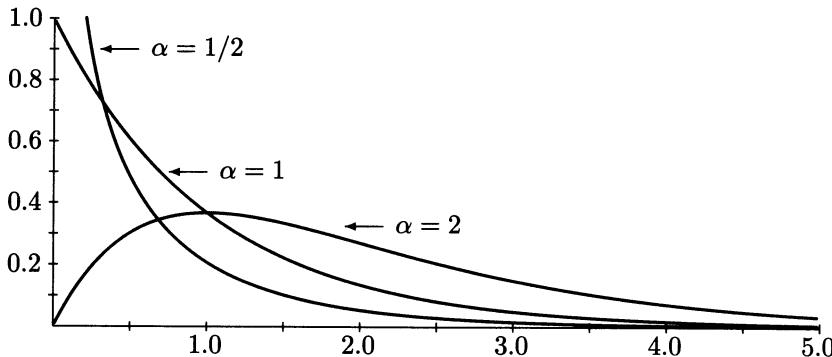


Figure 2.2. Some Gamma Density Functions ( $\lambda = 1$ )

by  $E(\lambda)$ , if

$$f_X(x) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & x < 0. \end{cases}$$

In situations where a density has qualitatively different forms for different  $x$ -values, one must pay careful attention to limits of integration (or, in the discrete case, summation). Thus,

$$\int_{-\infty}^{\infty} f_X(x) dx = \int_{-\infty}^0 0 dx + \int_0^{\infty} \lambda e^{-\lambda x} dx = \int_0^{\infty} \lambda e^{-\lambda x} dx = 1. \quad \square$$

Gamma distributions stand in the same relation to exponential as negative binomial to geometric: sums of independent, exponentially distributed random variables have gamma distributions.

**Example 2.37 (Gamma distribution).** A positive random variable  $X$  has a *gamma distribution* with parameters  $\alpha > 0, \lambda > 0$  if

$$f_X(x) = \frac{1}{\Gamma(\alpha)} \lambda^{\alpha} e^{-\lambda x} x^{\alpha-1}, \quad x \geq 0,$$

and  $f_X(x) = 0$  otherwise, where  $\Gamma(\alpha) = \int_0^{\infty} y^{\alpha-1} e^{-y} dy$  is Euler's gamma function, whose very definition (together with the change of variables  $y = \lambda x$ ) shows that  $\int_0^{\infty} f_X(x) dx = 1$ .

One terms  $\alpha$  the *shape* parameter and  $\lambda$  the *scale* parameter of the gamma distribution. Figure 2.2 shows the effect of varying  $\alpha$  with  $\lambda = 1$ .

Special cases of gamma distributions include *exponential distributions*, corresponding to  $\alpha = 1$ , *Erlang distributions*, which correspond to integer values of  $\alpha$ , and *Chi-squared ( $\chi^2$ ) distributions*, for which  $\lambda = 1/2$  and  $\alpha = k/2$  for some positive integer  $k$ . The sum of  $n$  independent random

Distribution	Parameters	Density Function $f_X(x)$
Standard normal	None	$\frac{1}{\sqrt{2\pi}} e^{-x^2/2}, x \in \mathbb{R}$
Normal	$\mu \in \mathbb{R}, \sigma^2 > 0$	$\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(x-\mu)^2/2\sigma^2}, x \in \mathbb{R}$
Exponential	$\lambda > 0$	$\lambda e^{-\lambda x}, x > 0$
Gamma	$\alpha > 0, \lambda > 0$	$\frac{1}{\Gamma(\alpha)} \lambda^\alpha e^{-\lambda x} x^{\alpha-1}, x > 0$
$\chi^2$	$k \in \mathbb{N}$	$\frac{1}{\Gamma(k/2)} \frac{1}{\sqrt{2^k}} e^{-x/2} x^{k/2-1}, x > 0$
Uniform on $[a, b]$	$a < b \in \mathbb{R}$	$1/(b-a), a \leq x \leq b$

Table 2.2. Absolutely Continuous Distributions

variables, each exponentially distributed with parameter  $\lambda$ , has an Erlang distribution with parameters  $n$  and  $\lambda$ . Chi-squared distributions are distributions of sums of squares of independent, normally distributed random variables. In this case,  $k$ , the number of normally distributed components, is referred to as the number of *degrees of freedom*.  $\square$

Table 2.2 summarizes the key absolutely continuous random variables. Uniform distributions, also shown there, were introduced in Example 1.37.

**Note on Notation.** Although not graceful grammatically, constructions of the form “Suppose that  $X \stackrel{d}{=} N(0, 1)$ ” to mean “Let  $X$  be normally distributed with parameters 0 and 1” are just too compact to eschew. Table 2.3 summarizes the distributions with notational abbreviations.  $\square$

### 2.4.3 Random vectors

Here we introduce multidimensional uniform distributions and bivariate normal distributions.

**Example 2.38 (*d*-dimensional uniform distribution).** A  $d$ -random vector  $X$  is *uniformly distributed* on  $[0, 1]^d$  if

$$f_X(x) = \begin{cases} 1 & x \in [0, 1]^d \\ 0 & \text{otherwise.} \end{cases}$$

It follows, using (2.4), that  $X_i \stackrel{d}{=} [0, 1]$  for each  $i$ .  $\square$

Distribution	Notation
Binomial with parameters $n, p$	$B(n, p)$
Exponential with parameter $\lambda$	$E(\lambda)$
Normal with parameters $\mu, \sigma^2$	$N(\mu, \sigma^2)$
Poisson with parameter $\lambda$	$P(\lambda)$
Uniform on $[a, b]$	$U[a, b]$

Table 2.3. Notation for Particular Distributions

**Example 2.39 (Bivariate standard normal distribution).** A random vector  $X = (X_1, X_2)$  has a *bivariate standard normal distribution* with parameter  $\rho \in (-1, 1)$  if

$$f_X(x) = \frac{1}{2\pi\sqrt{1-\rho^2}} \exp\left[-\frac{1}{2} \frac{x_1^2 - 2\rho x_1 x_2 + x_2^2}{1-\rho^2}\right], \quad x \in \mathbb{R}^2. \quad (2.7)$$

Both components have the standard normal density (2.5).  $\square$

## 2.5 Transformation Theory

Let  $X$  be a random variable with distribution function  $F_X$  and, possibly, density function  $f_X$ , let  $g$  be a Borel measurable function from  $\mathbb{R}$  into itself, and let  $Y = g(X)$ . We wish to know how the distribution function of  $Y$  relates to that of  $X$ , whether  $Y$  has a density, and, if so, how it relates to that of  $X$ .

### 2.5.1 Random variables

Other than the first, whose answer is furnished by Theorem 2.18, these questions admit useful answers only if  $g$  is invertible.

**Proposition 2.40.** Let  $Y = g(X)$ , where  $g$  is continuous and strictly increasing, and let  $h$  be the *pointwise inverse* of  $g$ . Then, for  $t \in \mathbb{R}$ ,

$$F_Y(t) = F_X(h(t)). \quad (2.8)$$

**Proof:** For each  $t$ ,

$$P\{Y \leq t\} = P\{g(X) \leq t\} = P\{X \leq h(t)\}. \quad \blacksquare$$

In general,  $g(X)$  need not be absolutely continuous even when  $X$  is, while if  $X$  is discrete, then so is  $g(X)$  regardless of  $g$ .

Here is the only broad result for the absolutely continuous case.

**Theorem 2.41.** Suppose that  $Y = g(X)$ , where  $X$  is absolutely continuous with density  $f_X$ , and that there is an open subset  $U$  of  $\mathbb{R}$  such that  $P\{X \in U\} = 1$ ,  $g$  is continuously differentiable on  $U$  and  $g'(x) \neq 0$  for all  $x \in U$ . Let  $h$  be the pointwise inverse of  $g$ , and let  $g(U) = \{g(x) : x \in U\}$ . Then,  $Y$  is absolutely continuous with

$$f_Y(y) = f_X(h(y)) |h'(y)|, \quad y \in g(U). \quad (2.9)$$

**Proof:** We may and do assume that  $g'(x) > 0$  for all  $x$ . By Proposition 2.40, for  $t \in \mathbb{R}$ , by the change of variables  $s = h(u)$

$$\mathsf{F}_Y(t) = \mathsf{F}_X(h(t)) = \int_{-\infty}^{h(t)} f_X(s) ds = \int_{-\infty}^t f_X(h(u)) h'(u) du. \quad \blacksquare$$

There is also a very nice heuristic derivation of (2.9), which is nearly a proof: recalling the interpretation (2.2) of a density function,

$$\begin{aligned} f_Y(y) dy &\cong P\{Y \in (y, y + dy)\} \\ &= P\{X \in g^{-1}((y, y + dy))\} \\ &= P\{X \in (h(y), h(y + dy))\} \\ &= P\{X \in (h(y), h(y) + h'(y) dy)\} \\ &\cong f_X(h(y)) h'(y) dy. \end{aligned} \quad (2.10)$$

To illustrate, suppose that  $Y = aX + b$  with  $a > 0$  and  $b \in \mathbb{R}$ . Then,  $h(y) = (y - b)/a$ , so for each  $t$ ,

$$\mathsf{F}_Y(t) = P\{aX + b \leq t\} = P\{X \leq (t - b)/a\} = \mathsf{F}_X((t - b)/a).$$

Consequently,  $Y$  has density

$$f_Y(t) = \frac{1}{a} f_X\left(\frac{t - b}{a}\right). \quad (2.11)$$

We illustrate with normal distributions.

**Example 2.42 (Normal distribution).** Suppose  $X \stackrel{\text{d}}{=} N(0, 1)$  and let  $Y = \sigma X + \mu$ , where  $\mu \in \mathbb{R}$  and  $\sigma > 0$ . Then, according to (2.11),

$$f_Y(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-(y-\mu)^2/2\sigma^2}, \quad y \in \mathbb{R}, \quad (2.12)$$

the normal density  $N(\mu, \sigma^2)$ . In particular, (2.12) is a density. Moreover,  $Y \stackrel{\text{d}}{=} N(\mu, \sigma^2)$  if and only if  $(Y - \mu)/\sigma \stackrel{\text{d}}{=} N(0, 1)$ .  $\square$

### 2.5.2 Random vectors

Recall that a random vector  $X = (X_1, \dots, X_d)$  is absolutely continuous if there exists a density function  $f_X$  on  $\mathbb{R}^d$  satisfying

$$F_X(t_1, \dots, t_d) = \int_{-\infty}^{t_1} \cdots \int_{-\infty}^{t_d} f_X(y_1, \dots, y_d) dy_1 \cdots dy_d.$$

Computation of the density of  $Y = g(X)$ , except for the special case that the components of  $X$  are independent (and then only for particular choices of  $g$ ), requires that  $g$  be invertible.

**Theorem 2.43.** Let  $X = (X_1, \dots, X_d)$  be absolutely continuous with density  $f_X$ , and let  $g$  be a Borel measurable function from  $\mathbb{R}^d$  to  $\mathbb{R}^d$  for which there exists an open set  $U$  such that  $P\{X \in U\} = 1$ ,  $g$  is one-to-one on  $U$ , and  $Jg(x) \neq 0$  for all  $x \in U$ , where  $Jg$  is the Jacobian of  $g$ , i.e., the determinant of the matrix of partial derivatives  $\partial g_i / \partial x_j$ . Let  $h$  be the pointwise inverse of  $g$ . Then,  $Y = g(X)$  is absolutely continuous with

$$f_Y(y) = f_X(h(y)) |Jh(y)|, \quad y \in g(U). \quad \square \quad (2.13)$$

A careful proof of Theorem 2.43 is surprisingly difficult. Here, instead, is an heuristic argument: for each  $y$ ,

$$\begin{aligned} f_Y(y) dy_1 \cdots dy_d &\cong P \left\{ Y \in \prod_{i=1}^d (y_i, y_i + dy_i) \right\} \\ &= P \left\{ X \in h \left( \prod_{i=1}^d (y_i, y_i + dy_i) \right) \right\} \\ &= f_X(h(y)) \times \text{volume of } h \left( \prod_{i=1}^d (y_i, y_i + dy_i) \right). \end{aligned}$$

Thus, we need to compute the volume of  $h \left( \prod_{i=1}^d (y_i, y_i + dy_i) \right)$ , but this is the volume spanned by the columns of the matrix

$$\begin{bmatrix} (\partial h_1 / \partial y_1) dy_1 & \cdots & (\partial h_1 / \partial y_d) dy_d \\ \vdots & & \vdots \\ (\partial h_d / \partial y_1) dy_1 & \cdots & (\partial h_d / \partial y_d) dy_d \end{bmatrix},$$

which is just  $|Jh(y)| dy_1 \cdots dy_d$ , and this gives (2.13).

For example, let  $g$  be an invertible affine mapping of  $\mathbb{R}^d$  into itself:  $g(x) = Ax + b$ , where  $A$  is a nonsingular  $d \times d$  matrix and  $b \in \mathbb{R}^d$ . Then, the inverse of  $g$  is given by  $h(y) = A^{-1}(y - b)$ , and hence  $|Jh(y)| = |\det A^{-1}| = |\det A|^{-1}$  for every  $y$ . Consequently, if  $X$  is a random  $n$ -vector with density  $f_X$ , then  $Y = AX + b$  is absolutely continuous, with density

$$f_Y(y) = |\det A|^{-1} f_X(A^{-1}(y - b)). \quad (2.14)$$

## 2.6 Random Variables with Prescribed Distributions

Here we present methods that construct explicitly individual random variables, random vectors or sequences of random variables with prescribed distribution functions. Specialized constructions for independent random variables are given in §3.3.

### 2.6.1 Individual random variables

**Proposition 2.44.** Let  $F$  be a distribution function on  $\mathbb{R}$ . Then, there exists a probability space  $(\Omega, \mathcal{F}, P)$  and a random variable  $X$  defined on it, such that  $\mathsf{F}_X = F$ .

**Proof:** Take  $(\Omega, \mathcal{F}) = (\mathbb{R}, \mathcal{B}(\mathbb{R}))$ , and let  $P$  be the unique probability such that  $\mathsf{F}_P = F$ , whose existence is guaranteed by Theorem 1.36. The identity function  $X(\omega) = \omega$  is a random variable, and for each  $t$ ,

$$P\{X \leq t\} = P((-\infty, t]) = \mathsf{F}_P(t) = F(t). \quad \blacksquare$$

An alternative construction, in Proposition 2.47, not only is more explicit, but also extends to independent random variables. It depends on a new concept.

**Definition 2.45.** The *inverse* of  $F$ , or *quantile function* associated with  $F$ , is the function defined by

$$F^{-1}(x) = \inf \{t: F(t) \geq x\}, \quad x \in (0, 1). \quad \square$$

Even though  $F$  need not be either continuous nor strictly increasing,  $F^{-1}$  always exists. As Figure 2.3 shows,  $F^{-1}$  jumps where  $F$  is flat and is flat where  $F$  jumps. Although not necessarily a pointwise inverse of  $F$ ,  $F^{-1}$  serves that role for many purposes. Here are the basic properties.

**Proposition 2.46.** Let  $F^{-1}$  be the inverse of  $F$ . Then,

- a) For each  $x$  and  $t$ ,  $F^{-1}(x) \leq t$  if and only if  $x \leq F(t)$ .
- b)  $F^{-1}$  is increasing and *left*-continuous.
- c) If  $F$  is continuous, then  $F(F^{-1}(x)) = x$  for all  $x \in (0, 1)$ .  $\square$

A random variable with distribution function  $F$  can be constructed by applying  $F^{-1}$  to a random variable uniformly distributed on  $[0, 1]$ , a process known as the *quantile transformation*.

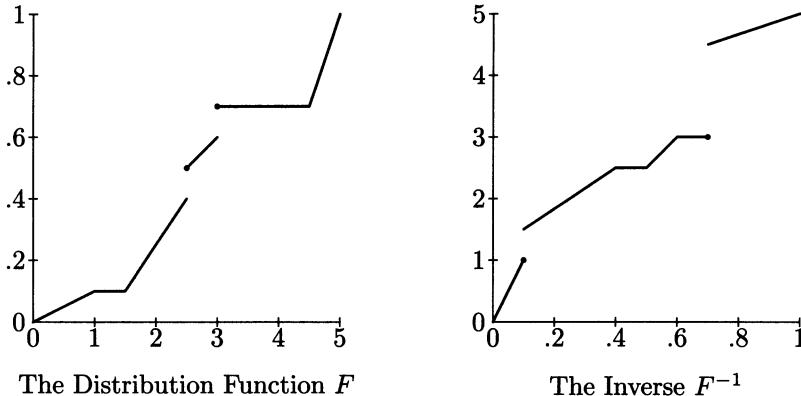


Figure 2.3. A Distribution Function and its Inverse

**Proposition 2.47 (Quantile transformation).** Let  $F$  be a distribution function on  $\mathbb{R}$  and suppose that  $U \stackrel{d}{=} U[0, 1]$ . Then,  $X = F^{-1}(U)$  has distribution function  $F$ .

**Proof:** Because it is monotone,  $F^{-1}$  is Borel measurable, so that  $F^{-1}(U)$  is a random variable. By part a) of Proposition 2.46, for each  $t \in \mathbb{R}$ ,

$$P\{X \leq t\} = P\{F^{-1}(U) \leq t\} = P\{U \leq F(t)\} = F(t). \blacksquare$$

Proposition 2.47 is the basis for many simulation procedures. “Random number” generators on computers produce numbers with properties (that mimic those) of independent random variables uniformly distributed on  $[0, 1]$ . Hence, random variables with distribution function  $F$  can be simulated by applying  $F^{-1}$  to the values produced by the random number generator. Feasibility of this technique, of course, depends on either having  $F^{-1}$  available in closed form or being able to approximate it numerically.

When  $F$  is continuous (not necessarily absolutely continuous), a converse to Proposition 2.47 holds as well.

**Proposition 2.48.** If  $F_X$  is continuous, then  $F_X(X) \stackrel{d}{=} U[0, 1]$ .

**Proof:** Let  $F = F_X$ . For  $x \in [0, 1]$ ,

$$P\{F(X) \geq x\} = P\{X \geq F^{-1}(x)\} = 1 - F(F^{-1}(x)) = 1 - x,$$

where the first equality is by part a) of Proposition 2.46, the second is by continuity of  $F$  (otherwise  $P\{F(X) \geq x\} = 1 - F(F^{-1}(x) - )$ ) and the third is by part c) of Proposition 2.46. ■

### 2.6.2 Random vectors

Construction of a random vector with an arbitrary distribution function is more complicated, and we omit the proof. For the case of independent components, see §3.3.

**Theorem 2.49.** Let  $F: \mathbb{R}^d \rightarrow [0, 1]$  be a  $d$ -dimensional distribution function. Then, there exists a probability space  $(\Omega, \mathcal{F}, P)$ , and a random vector  $X = (X_1, \dots, X_d)$  defined on it, such that  $F_X = F$ . □

### 2.6.3 Sequences of random variables

We wish to construct a sequence  $(X_k)$  of random variables with prescribed joint distribution functions  $F_n$ , in the sense that  $F_n$  is the distribution function of  $(X_1, \dots, X_n)$  for each  $n$ . The  $F_n$  must satisfy certain *consistency conditions*, since if such random variables exist, then for all choices of  $t_1, \dots, t_n$ ,

$$P\{X_1 \leq t_1, \dots, X_n \leq t_n\} = \lim_{t \rightarrow \infty} P\{X_1 \leq t_1, \dots, X_n \leq t_n, X_{n+1} \leq t\}.$$

In the next theorem, a centerpiece of probability, we see that this necessary condition is also sufficient.

**Theorem 2.50 (Kolmogorov extension theorem).** For each  $n$ , let  $F_n$  be a distribution function on  $\mathbb{R}^n$ , and suppose that

$$\lim_{t \rightarrow \infty} F_{n+1}(t_1, \dots, t_n, t) = F_n(t_1, \dots, t_n)$$

for each  $n$  and  $t_1, \dots, t_n$ . Then, there exists a probability space  $(\Omega, \mathcal{F}, P)$ , and a sequence  $(X_k)$  of random variables defined on it, such that  $F_n$  is the distribution function of  $(X_1, \dots, X_n)$  for each  $n$ . □

## 2.7 Complements

### 2.7.1 Measurability with respect to sub- $\sigma$ -algebras

Random variables are defined by the property that inverse images of Borel sets be events. At times, it is useful to require that these inverse images belong to some smaller  $\sigma$ -algebra. Let  $\mathcal{G}$  be a *sub- $\sigma$ -algebra* of  $\mathcal{F}$ :  $\mathcal{G}$  is a  $\sigma$ -algebra and  $A \in \mathcal{F}$  for all  $A \in \mathcal{G}$ .

**Definition 2.51.** A random variable  $Y$  is *measurable with respect to  $\mathcal{G}$*  if  $\{Y \in B\} \in \mathcal{G}$  for every  $B \in \mathcal{B}(\mathbb{R})$ .  $\square$

The most interesting case is that  $\mathcal{G}$  is the  $\sigma$ -algebra generated by random variables  $X_1, \dots, X_d$ , as defined by (2.1): random variables measurable with respect to  $\sigma(X_1, \dots, X_d)$  are functions of  $X_1, \dots, X_d$ .

**Theorem 2.52.** A random variable  $Y$  is measurable with respect to  $\sigma(X_1, \dots, X_d)$  if and only if there is a Borel measurable function  $g: \mathbb{R}^d \rightarrow \mathbb{R}$  such that  $Y = g(X_1, \dots, X_d)$ .  $\square$

## 2.7.2 Borel measurable functions

Borel measurable functions, like random variables, are defined via inverse images. In this case, inverse images of Borel sets must be Borel sets.

**Definition 2.53.** A function  $g: \mathbb{R} \rightarrow \mathbb{R}$  is *Borel measurable* if  $g^{-1}(B) \in \mathcal{B}(\mathbb{R})$  for every  $B \in \mathcal{B}(\mathbb{R})$ . A function  $g: \mathbb{R}^n \rightarrow \mathbb{R}$  is *Borel measurable* if  $g^{-1}(B) \in \mathcal{B}(\mathbb{R}^n)$  for every  $B \in \mathcal{B}(\mathbb{R})$ . A function  $g: \mathbb{R}^n \rightarrow \mathbb{R}^m$  is *Borel measurable* if  $g^{-1}(B) \in \mathcal{B}(\mathbb{R}^m)$  for every  $B \in \mathcal{B}(\mathbb{R}^m)$ .  $\square$

In order that  $g: \mathbb{R} \rightarrow \mathbb{R}$  be Borel measurable, it suffices that for every  $t$ ,  $g^{-1}((-\infty, t]) \in \mathcal{B}(\mathbb{R})$ . The same is true for real-valued functions  $g$  on  $\mathbb{R}^n$ . Also, a function  $g: \mathbb{R}^n \rightarrow \mathbb{R}^m$  is Borel measurable if and only if each of its components is Borel measurable as a function from  $\mathbb{R}^n$  to  $\mathbb{R}$ .

In the same way that all “reasonable” subsets of  $\mathbb{R}$  are Borel sets, all “reasonable” functions are Borel measurable. In particular, indicator functions, simple functions, monotone functions and continuous functions are Borel measurable. The class of Borel measurable functions has the same closure properties — under algebraic and limiting operations — as the family of random variables on a probability space  $(\Omega, \mathcal{F}, P)$ .

Versions of Theorem 2.21 and Theorem 2.22 hold as well.

**Theorem 2.54.** Let  $\mathcal{S}$  be a  $\pi$ -system generating  $\mathcal{B}(\mathbb{R})$  and let  $\mathbf{H}$  be a set of functions on  $\mathbb{R}$  such that

- i) The constant function 1 belongs to  $\mathbf{H}$ ;
- ii)  $\mathbf{1}_A \in \mathbf{H}$  for all  $A \in \mathcal{S}$ ;
- iii)  $\mathbf{H}$  is a vector space;
- iv) If  $g_n \in \mathbf{H}$  for each  $n$  and  $\sup_n g_n(t) < \infty$  for each  $t$ , then  $\sup_n g_n$  belongs to  $\mathbf{H}$ .

Then,  $\mathbf{H}$  contains all Borel measurable functions.  $\square$

**Theorem 2.55.** Let  $\mathcal{S}$  be a  $\pi$ -system generating  $\mathcal{B}(\mathbb{R})$  and let  $\mathbf{H}$  be a set of positive functions on  $\mathbb{R}$  such that

- i) The constant function 1 belongs to  $\mathbf{H}$ ;
- ii)  $\mathbf{1}_A \in \mathbf{H}$  for all  $A \in \mathcal{S}$ ;
- iii) If  $g, h \in \mathbf{H}$  and  $a, b \in \mathbb{R}_+$ , then  $ag + bh \in \mathbf{H}$  ( $\mathbf{H}$  is a cone);
- iv) If  $g_n \in \mathbf{H}$  for each  $n$  and  $\sup_n g_n(t) < \infty$  for each  $t$ , then  $\sup_n g_n$  belongs to  $\mathbf{H}$ .

Then,  $\mathbf{H}$  contains all positive Borel measurable functions.  $\square$

## 2.8 Exercises

**2.1.** Let  $\Omega = \{1, 2, 3, 4, 5, 6\}$ , and let  $\mathcal{F} = \sigma(\{1, 2, 3, 4\}, \{3, 4, 5, 6\})$ .

- a) List all sets in  $\mathcal{F}$ .
- b) Is the function

$$X(\omega) = \begin{cases} 2 & \omega = 1, 2, 3, 4 \\ 7 & \omega = 5, 6 \end{cases}$$

a random variable over  $(\Omega, \mathcal{F})$ ?

- c) Give an example of a function on  $\Omega$  that is *not* a random variable over  $(\Omega, \mathcal{F})$ .
- d) Show that there exists a probability  $P$  on  $(\Omega, \mathcal{F})$  such that  $P(A)$  is zero or one for all  $A \in \mathcal{F}$ , yet  $P$  is not a point mass.

**2.2.** Prove that if  $X$  and  $Y$  are random variables, then  $\{X \leq Y\}$ ,  $\{X < Y\}$  and  $\{X = Y\}$  are events.

**2.3.** Let  $X$  and  $Y$  be random variables and let  $A$  be an event. Prove that the function

$$Z(\omega) = \begin{cases} X(\omega) & \text{if } \omega \in A \\ Y(\omega) & \text{if } \omega \in A^c \end{cases}$$

is a random variable.

**2.4.** Let  $\mathcal{G} = \{A_1, \dots, A_n\}$  be a finite partition of  $\Omega$ , and let  $\mathcal{F} = \sigma(\mathcal{G})$ .

- a) Prove that a function  $X: \Omega \rightarrow \mathbb{R}$  is a random variable if and only if  $X$  is constant over each partition set  $A_i$ .
- b) Use part a) to show that provided  $\mathcal{F} \neq \mathcal{P}(\Omega)$ , there exist functions  $Y$  on  $\Omega$  such that  $|Y|$  is a random variable but  $Y$  is not.

**2.5.** Let  $X^+$  and  $X^-$  be the positive and negative parts of the function  $X: \Omega \rightarrow \mathbb{R}$ . Prove that  $X = X^+ - X^-$  and  $|X| = X^+ + X^-$ .

- 2.6.** Show that if  $X$  is discrete with values in the countable set  $C$ , then for every  $B \in \mathcal{B}(\mathbb{R})$ ,  $P\{X \in B\} = \sum_{a \in C \cap B} P\{X = a\}$ .
- 2.7.** Consider a random permutation of the integers  $\{1, \dots, n\}$ , with all  $n!$  permutations equally likely. For each  $i$ , let  $X_i$  be the integer in the  $i$ th position, and let  $A_i = \{X_i = i\}$ . (Physically, this means that there is a *match* in the  $i$ th position.)
- Use the inclusion-exclusion principle (Exercise 1.25) to show that the probability of at least one match is
- $$P\left(\bigcup_{i=1}^n A_i\right) = \sum_{k=1}^n (-1)^{k-1}/k!.$$
- Let  $p_n$  be the probability of no match. Show that  $\lim_n p_n = 1/e$ .
- 2.8.** Let  $X$  have distribution  $P(\lambda)$ . Show that the function  $i \mapsto P\{X = i\}$  is first increasing and then decreasing, with its maximum value at  $\lfloor \lambda \rfloor$ , the largest integer less than or equal to  $\lambda$ .
- 2.9.** Let  $X$  have density
- $$f_X(x) = \frac{1}{\pi(1+x^2)}, \quad x \in \mathbb{R},$$
- known as a *Cauchy density*. Show that
- $$\mathsf{F}_X(t) = 1/2 + (1/\pi) \arctan t.$$
- 2.10.** Show that the analogue of (2.8) for  $g$  continuous and strictly decreasing is
- $$P\{Y \leq t\} = 1 - \mathsf{F}_X(h(t) -).$$
- Let  $X$  have distribution function  $F$ . For each  $k$ , calculate the distribution function of  $|X|^k$ .
  - Let  $X$  be absolutely continuous. Compute the density of  $|X|$ .
- 2.12.** Let  $X$  have distribution function  $F$ . Calculate the distribution functions of  $X^+$  and  $X^-$ .
- 2.13.** Assume that  $X$  is positive and absolutely continuous with density  $f$  and distribution function  $F$  satisfying  $F(s) < 1$  for all  $s < \infty$ , and let  $H(s) = -\log(1 - F(s))$ .
- Prove that  $H$  is differentiable with derivative  $h = f/(1 - F)$ , which is termed the *hazard function* of  $T$  (or  $F$ ).

b) Prove that for each  $t$ ,

$$h(t) = \lim_{h \downarrow 0} \frac{1}{h} P\{X \leq t + h | X > t\}.$$

c) Prove that  $\int_0^\infty h(t) dt = \infty$ .

d) Prove that

$$P\{X > t + s | X > t\} = \exp \left[ - \int_t^{t+s} h(u) du \right]$$

for each  $t$  and  $s$ .

e) Prove that  $X \stackrel{d}{=} E(\lambda)$  if and only if  $h \equiv \lambda$ .

**2.14.** a) Prove that if  $X$  has a geometric distribution then

$$P\{X > n + k | X > n\} = P\{X > k\} \quad (2.15)$$

for each  $n$  and  $k$ . [If  $X$  is thought of as the discrete-valued lifetime of some system, then its distribution is *memoryless*: the probability of the system's surviving at least  $k$  additional time units does not depend on its age.]

b) Prove that if  $X$  is a positive and integer-valued and satisfies (2.15), then  $X$  has a geometric distribution.

**2.15.** a) Prove that if  $Y \stackrel{d}{=} E(\lambda)$ , then for all  $t$  and  $s$ ,

$$P\{Y > t + s | Y > t\} = P\{Y > s\}. \quad (2.16)$$

b) Prove that if  $Y$  is absolutely continuous, positive and satisfies (2.16), then  $Y$  has an exponential distribution. [Hint: Prove that the survivor function  $S_Y(t) = P\{Y > t\}$  fulfills the functional equation  $S_Y(t + s) = S_Y(t)S_Y(s)$  for  $s, t > 0$ .]

**2.16.** Let  $X$  have distribution  $N(0, 1)$ . Calculate the density of  $Y = e^X$ , which is said to have a *log normal distribution*.

**2.17.** Let  $Y = g(X)$ , where  $X$  is a random variable and  $g: \mathbb{R} \rightarrow \mathbb{R}$  is Borel measurable. Prove that  $\sigma(Y) \subseteq \sigma(X)$ . Conclude that if also  $X = h(Y)$  for some  $h$ , then  $\sigma(X) = \sigma(Y)$ .

**2.18.** Prove that if  $U \stackrel{d}{=} U[0, 1]$  and  $\lambda > 0$ , then

$$X = -(1/\lambda) \log U$$

satisfies  $X \stackrel{d}{=} E(\lambda)$ .

**2.19.** Calculate the density function of  $Y = 1/X - X$ , where  $X \stackrel{d}{=} U[0, 1]$ .

**2.20.** Suppose that  $Y = X^2$ .

a) Prove that for  $t \geq 0$ ,

$$F_Y(t) = F_X(\sqrt{t}) - F_X(-\sqrt{t}).$$

b) Prove that if  $X$  is absolutely continuous, then  $Y$  is absolutely continuous with

$$f_Y(y) = \frac{f_X(\sqrt{y}) + f_X(-\sqrt{y})}{2\sqrt{y}}, \quad y > 0.$$

c) Prove that if  $X \stackrel{d}{=} N(0, 1)$ , then  $Y$  has a  $\chi^2$  distribution with one degree of freedom.

**2.21.** Let  $(X_1, X_2)$  be a 2-dimensional random vector with distribution function  $F$ . Show that for  $a_1 \leq b_1$  and  $a_2 \leq b_2$ ,

$$\begin{aligned} P\{a_1 < X_1 \leq b_1, a_2 < X_2 \leq b_2\} \\ = F(b_1, b_2) - F(b_1, a_2) - F(a_1, b_2) + F(a_1, a_2). \end{aligned}$$

**2.22.** Let the random vector  $(X, Y)$  be uniformly distributed on the disk  $D_r = \{(x, y): x^2 + y^2 \leq r^2\}$  in  $\mathbb{R}^2$ . Calculate the joint distribution of  $R = \sqrt{X^2 + Y^2}$  and  $\Theta = \arctan Y/X$ .

# Chapter 3

## Independence

Independence, or the *absence of probabilistic interaction*, sets probability apart as a distinct mathematical theory.

### 3.1 Independent Random Variables

We define independence for random variables in terms of Borel sets, then develop specialized criteria, which involve only intervals.

#### 3.1.1 Fundamentals

**Definition 3.1.** Random variables  $X_1, \dots, X_n$  are *independent* if

$$P\{X_1 \in B_1, \dots, X_n \in B_n\} = \prod_{i=1}^n P\{X_i \in B_i\} \quad (3.1)$$

for all Borel sets  $B_1, \dots, B_n$ .

An infinite set of random variables is *independent* if every finite subset is independent.  $\square$

We also introduce a key abbreviation.

**Definition 3.2.** Random variables that are independent and have the same distribution function are said to be *i.i.d.*: *independent and identically distributed*.  $\square$

#### 3.1.2 Criteria for independence

Random variables are independent if and only if their joint distribution function is the product of their individual distribution functions. This

result affirms the general principle that definitions stated in terms of all Borel sets need only be checked for intervals  $(-\infty, t]$ .

**Theorem 3.3.** Random variables  $X_1, \dots, X_n$  are independent if and only if

$$\mathsf{F}_{X_1, \dots, X_n}(t_1, \dots, t_n) = \prod_{i=1}^n \mathsf{F}_{X_i}(t_i) \quad (3.2)$$

for all  $t_1, \dots, t_n \in \mathbb{R}$ .

**Proof:** If  $X_1, \dots, X_n$  are independent, then (3.2) holds, since by (3.1)

$$\begin{aligned} \mathsf{F}_{X_1, \dots, X_n}(t_1, \dots, t_n) &= P\{X_1 \in (-\infty, t_1], \dots, X_n \in (-\infty, t_n]\} \\ &= \prod_{i=1}^n P\{X_i \in (-\infty, t_i]\} \\ &= \prod_{i=1}^n \mathsf{F}_{X_i}(t_i). \end{aligned}$$

To prove sufficiency of (3.2), let  $\mathcal{B}_1$  be the family of sets  $B \in \mathcal{B}(\mathbb{R})$  such that

$$P\{X_1 \in B, X_2 \leq t_2, \dots, X_n \leq t_n\} = P\{X_1 \in B\} \prod_{i=2}^n P\{X_i \leq t_i\} \quad (3.3)$$

for all choices of  $t_2, \dots, t_n \in \mathbb{R}$ . Then, the  $\pi$ -system  $\mathcal{J} = \{(-\infty, t] : t \in \mathbb{R}\}$ , which generates  $\mathcal{B}(\mathbb{R})$ , is contained in  $\mathcal{B}_1$  by assumption, and we show momentarily that  $\mathcal{B}_1$  is a  $d$ -system. If this is so, then  $\mathcal{B}_1 = \sigma(\mathcal{J}) = \mathcal{B}(\mathbb{R})$  by Theorem 1.17, and therefore, (3.3) holds for all  $B \in \mathcal{B}(\mathbb{R})$ .

First of all,  $\mathbb{R} \in \mathcal{B}_1$  since the family  $\{X_2, \dots, X_n\}$  is independent. Next, if  $A \subseteq B \in \mathcal{B}_1$ , then for all  $t_2, \dots, t_n \in \mathbb{R}$ ,

$$\begin{aligned} P\{X_1 \in B \setminus A, X_2 \leq t_2, \dots, X_n \leq t_n\} \\ &= P\{X_1 \in B, X_2 \leq t_2, \dots, X_n \leq t_n\} \\ &\quad - P\{X_1 \in A, X_2 \leq t_2, \dots, X_n \leq t_n\} \\ &= P\{X_1 \in B\} \prod_{i=2}^n P\{X_i \leq t_i\} - P\{X_1 \in A\} \prod_{i=2}^n P\{X_i \leq t_i\} \\ &= P\{X_1 \in B \setminus A\} \prod_{i=2}^n P\{X_i \leq t_i\}, \end{aligned}$$

which confirms that  $B \setminus A \in \mathcal{B}_1$ . Finally, if  $B_k \in \mathcal{B}_1$  for each  $k$  and  $B_k \uparrow B$ ,

then  $B \in \mathcal{B}_1$  since for  $t_2, \dots, t_n \in \mathbb{R}$

$$\begin{aligned} & P\{X_1 \in B, X_2 \leq t_2, \dots, X_n \leq t_n\} \\ &= \lim_{k \rightarrow \infty} P\{X_1 \in B_k, X_2 \leq t_2, \dots, X_n \leq t_n\} \\ &= \lim_{k \rightarrow \infty} P\{X_1 \in B_k\} \prod_{i=2}^n P\{X_i \leq t_i\} \\ &= P\{X_1 \in B\} \prod_{i=2}^n P\{X_i \leq t_i\}. \end{aligned}$$

We continue this procedure iteratively until the same extension has been accomplished for  $X_2, \dots, X_n$ ; the next step is as follows. Let  $\mathcal{B}_2$  be the family of Borel sets  $B'$  for which

$$\begin{aligned} & P\{X_1 \in B, X_2 \in B', X_3 \leq t_3, \dots, X_n \leq t_n\} \\ &= P\{X_1 \in B\} P\{X_2 \in B'\} \prod_{i=3}^n P\{X_i \leq t_i\} \end{aligned}$$

for all  $B \in \mathcal{B}(\mathbb{R})$  and all  $t_3, \dots, t_n$ . Then, by (3.2) and the preceding argument,  $\mathcal{B}_2$  is a  $d$ -system containing  $\emptyset$ , and so  $\mathcal{B}_2 = \mathcal{B}(\mathbb{R})$ . ■

Specialized criteria for the discrete and absolutely continuous cases continue the motif introduced in Theorem 3.3.

**Theorem 3.4.** Discrete random variables  $X_1, \dots, X_n$ , taking values in the countable set  $C$ , are independent if and only if

$$P\{X_1 = a_1, \dots, X_n = a_n\} = \prod_{i=1}^n P\{X_i = a_i\} \quad (3.4)$$

for all  $a_1, \dots, a_n \in C$ .

**Proof:** Necessity of (3.4) is clear. To show sufficiency, we apply Theorem 3.3. For  $t_1, \dots, t_n \in \mathbb{R}$ ,

$$\begin{aligned} & P\{X_1 \leq t_1, \dots, X_n \leq t_n\} \\ &= \sum_{\{a_1 \in C: a_1 \leq t_1\}} \cdots \sum_{\{a_n \in C: a_n \leq t_n\}} P\{X_1 = a_1, \dots, X_n = a_n\} \\ &= \sum_{\{a_1 \in C: a_1 \leq t_1\}} \cdots \sum_{\{a_n \in C: a_n \leq t_n\}} \prod_{i=1}^n P\{X_i = a_i\} \\ &= \prod_{i=1}^n \sum_{\{a_i \in C: a_i \leq t_i\}} P\{X_i = t_i\} \\ &= \prod_{i=1}^n P\{X_i \leq t_i\}. \quad \blacksquare \end{aligned}$$

Absolutely continuous random variables are independent if and only if their joint density function is the product of their individual densities. This is one instance in which absolute continuity of a random vector follows from that of its components.

**Theorem 3.5.** Let  $X = (X_1, \dots, X_n)$  be an absolutely continuous random vector. Then,  $X_1, \dots, X_n$  are independent if and only if

$$f_X(y_1, \dots, y_n) = \prod_{i=1}^n f_{X_i}(y_i) \quad (3.5)$$

for all  $y_1, \dots, y_n \in \mathbb{R}$ .

**Proof:** If  $X_1, \dots, X_n$  are independent, then by (3.2) and (2.3),

$$\begin{aligned} \prod_{i=1}^n \int_{-\infty}^{t_i} f_{X_i}(y_i) dy_i &= \prod_{i=1}^n P\{X_i \leq t_i\} \\ &= P\{X_1 \leq t_1, \dots, X_n \leq t_n\} \\ &= \int_{-\infty}^{t_1} \cdots \int_{-\infty}^{t_n} f_X(y_1, \dots, y_n) dy_1 \cdots dy_n \end{aligned}$$

for all  $t_1, \dots, t_n$ . That is,

$$\int_{-\infty}^{t_1} \cdots \int_{-\infty}^{t_n} [f_X(y_1, \dots, y_n) - \prod_{i=1}^n f_{X_i}(y_i)] dy_1 \cdots dy_n = 0 \quad (3.6)$$

for all  $t_1, \dots, t_n$ , which implies (3.5) by differentiation with respect to  $t_1, \dots, t_n$ .

Conversely, if (3.5) holds, then

$$\begin{aligned} P\{X_1 \leq t_1, \dots, X_n \leq t_n\} &= \int_{-\infty}^{t_1} \cdots \int_{-\infty}^{t_n} [\prod_{i=1}^n f_{X_i}(y_i)] dy_1 \cdots dy_n \\ &= \prod_{i=1}^n \int_{-\infty}^{t_i} f_{X_i}(y_i) dy_i \\ &= \prod_{i=1}^n P\{X_i \leq t_i\}, \end{aligned}$$

which establishes independence by appeal to Theorem 3.3. ■

**Corollary 3.6.** If  $X_1, \dots, X_n$  are independent and absolutely continuous, then  $X = (X_1, \dots, X_n)$  is absolutely continuous.

**Proof:** In this case,

$$\begin{aligned} P\{X_1 \leq t_1, \dots, X_n \leq t_n\} &= \prod_{i=1}^n P\{X_i \leq t_i\} \\ &= \prod_{i=1}^n \int_{-\infty}^{t_i} f_{X_i}(y_i) dy_i \\ &= \int_{-\infty}^{t_1} \cdots \int_{-\infty}^{t_n} [\prod_{i=1}^n f_{X_i}(y_i)] dy_1 \cdots dy_n \end{aligned}$$

which shows that  $\prod_{i=1}^n f_{X_i}(y_i)$  fulfills (2.3), the defining criterion for the joint density  $f_X$ . ■

### 3.1.3 Examples

Independent random variables are inherent to certain probability structures.

**Example 3.7 (Binary expansions).** Let  $P$  be the uniform distribution on  $\Omega = [0, 1]$ . Each point of  $\Omega$  has a binary expansion

$$\omega = .X_1(\omega)X_2(\omega)\cdots,$$

where the  $X_i$  are functions from  $\Omega$  to  $\{0, 1\}$ . This expansion is unique provided that infinitely many of the  $X_i$  be equal to one. We now show that  $X_1, X_2, \dots$  are independent and Bernoulli distributed with parameter  $p = 1/2$ .

It is important not to forget to verify that the  $X_i$  are random variables, but this is easy:  $X_i$  is the indicator function of the union of  $2^{i-1}$  disjoint intervals, which is a Borel set. Moreover, each component of  $\{X_i = 1\}$  has length  $2^{-i}$ , so that

$$P\{X_i = 1\} = 2^{i-1} \times 2^{-i} = 1/2.$$

We now prove that  $X_1, \dots, X_n$  are independent for each  $n$ . Suppose that  $j_1, \dots, j_n$  are each zero or one. Then,  $\{X_1 = j_1, \dots, X_n = j_n\}$  consists of those  $\omega$  such that  $\omega = .j_1j_2\cdots j_nX_{n+1}(\omega)X_{n+2}(\omega)\cdots$ , and is, hence, an interval of length  $2^{-n}$ . Therefore,

$$P\{X_1 = j_1, \dots, X_n = j_n\} = 2^{-n} = \prod_{i=1}^n P\{X_i = j_i\},$$

which gives independence by Theorem 3.4.

The random variable  $U(\omega) = \omega$  is uniformly distributed on  $[0, 1]$ , and, hence, if  $X_1, X_2, \dots$  are independent and Bernoulli distributed with parameter  $p = 1/2$ , then  $\sum_{n=1}^{\infty} 2^{-n}X_n \stackrel{d}{=} U[0, 1]$ . □

**Example 3.8 (Multidimensional uniform distribution).** Suppose that  $P$  is the uniform distribution on  $[0, 1]^d$ . Then, the coordinate random variables  $U_i((\omega_1, \dots, \omega_n)) = \omega_i$  are independent, and each is uniformly distributed on  $[0, 1]$ . Indeed, for intervals  $I_1, \dots, I_n$ ,

$$\begin{aligned} P\{U_1 \in I_1, \dots, U_n \in I_n\} &= P\left(\prod_{j=1}^n I_j\right) = \prod_{j=1}^n |I_j| \\ &= \prod_{j=1}^n P\{U_j \in I_j\}. \quad \square \end{aligned}$$

In other cases, whether random variables are independent depends on the value of a parameter.

**Example 3.9 (Bivariate standard normal distribution).** Let  $(X, Y)$  have a bivariate normal distribution with density (2.7). Since  $X \stackrel{d}{=} Y \stackrel{d}{=} N(0, 1)$ , by Theorem 3.5 they are independent if and only if  $\rho = 0$ .  $\square$

## 3.2 Functions of Independent Random Variables

### 3.2.1 Transformation properties

Random variables that are functions of disjoint subsets of a family of independent random variables are independent. The name of the result reflects the property that the  $Y_\ell$  arise from disjoint blocks of the  $X_i$ . According to Definition 3.1, the same result holds true for (countably) infinite families and blocks.

**Theorem 3.10 (Disjoint blocks theorem).** Suppose that  $X_1, \dots, X_n$  are independent, let  $J_1, \dots, J_k$  be disjoint subsets of  $\{1, \dots, n\}$ , and for each  $\ell$ , let  $Y_\ell$  be a (Borel measurable) function  $g_\ell$  of  $X^{(\ell)} = \{X_i : i \in J_\ell\}$ . Then,  $Y_1, \dots, Y_k$  are independent.

**Proof:** For simplicity, we assume that  $k = 2$  and that  $J_1 = \{1, \dots, m\}$  and  $J_2 = \{m + 1, \dots, n\}$  for some  $m$ , so that  $Y_1 = g_1(X_1, \dots, X_m)$  and  $Y_2 = g_2(X_{m+1}, \dots, X_n)$ . Then, independence of  $Y_1$  and  $Y_2$  requires that

$$\begin{aligned} P\{X^{(1)} \in g_1^{-1}(B_1), X^{(2)} \in g_2^{-1}(B_2)\} \\ = P\{X^{(1)} \in g_1^{-1}(B_1)\}P\{X^{(2)} \in g_2^{-1}(B_2)\}, \end{aligned}$$

for all  $B_1$  and  $B_2$  in  $\mathcal{B}(\mathbb{R})$ , which is implied by the stronger condition

$$P\{X^{(1)} \in A_1, X^{(2)} \in A_2\} = P\{X^{(1)} \in A_1\}P\{X^{(2)} \in A_2\} \quad (3.7)$$

for all  $A_1 \in \mathcal{B}(\mathbb{R}_m)$  and  $A_2 \in \mathcal{B}(\mathbb{R}^{n-m})$ . If  $A_1 = \prod_{i=1}^m A_i^*$  and  $A_2 = \prod_{i=m+1}^N A_i^*$  are rectangles, then (3.7) reduces to

$$\begin{aligned} P\{X^{(1)} \in A_1, X^{(2)} \in A_2\} &= P\{X^{(1)} \in \prod_{i=1}^m A_i^*, X^{(2)} \in \prod_{i=m+1}^N A_i^*\} \\ &= P\{X_1 \in A_1^*, \dots, X_m \in A_m^*, X_{m+1} \in A_{m+1}^*, \dots, X_n \in A_n^*\} \\ &= \prod_{i=1}^m P\{X_i \in A_i^*\} \\ &= \prod_{i=1}^m P\{X_i \in A_i^*\} \times \prod_{i=m+1}^N P\{X_i \in A_i^*\} \\ &= P\{X^{(1)} \in A_1\} P\{X^{(2)} \in A_2\}. \end{aligned}$$

Validity for all Borel sets then follows, as in the proof of Theorem 3.3. ■

**Corollary 3.11.** Let  $X_1, \dots, X_n$  be independent, suppose that  $g_1, \dots, g_n$  are (Borel measurable) functions from  $\mathbb{R}$  to  $\mathbb{R}$ , and for each  $i$ , let  $Y_i = g_i(X_i)$ . Then,  $Y_1, \dots, Y_n$  are independent. □

### 3.2.2 Sums of independent random variables

Sums of independent random variables merit special consideration. We begin with the absolutely continuous case; the treatment of the subject concludes in §4.3.

**Theorem 3.12.** Let  $X$  and  $Y$  be independent and absolutely continuous. Then,  $X + Y$  is absolutely continuous and

$$f_{X+Y}(v) = \int_{-\infty}^{\infty} f_X(v-s) f_Y(s) ds, \quad v \in \mathbb{R}.$$

**Proof:** By Theorem 2.18, for each  $t$ ,

$$\begin{aligned} P\{X + Y \leq t\} &= P\{(X, Y) \in \{(u, s): u + s \leq t\}\} \\ &= \iint_{\{(u, s): u + s \leq t\}} f_{(X, Y)}(u, s) du ds \\ &= \iint_{\{(u, s): u + s \leq t\}} f_X(u) f_Y(s) du ds \\ &= \int_{-\infty}^{\infty} \left[ \int_{-\infty}^{t-s} f_X(u) du \right] f_Y(s) ds \\ &= \int_{-\infty}^{\infty} \left[ \int_{-\infty}^t f_X(v-s) dv \right] f_Y(s) ds \\ &= \int_{-\infty}^t \left[ \int_{-\infty}^{\infty} f_X(v-s) f_Y(s) ds \right] dv. \quad \blacksquare \end{aligned}$$

The density in Theorem 3.12 is accorded a special name.

**Definition 3.13.** The *convolution* of density functions  $f$  and  $g$  is the density  $f * g$  given by

$$f * g(t) = \int_{-\infty}^{\infty} f(t-s)g(s) ds. \quad (3.8)$$

(That  $f * g$  is a density follows from Theorem 3.12.)  $\square$

Convolution is *commutative*: for all densities  $f$  and  $g$ ,  $f * g = g * f$ ; and *associative*: for all densities  $f$ ,  $g$  and  $h$ ,  $f * (g * h) = (f * g) * h$ .

Once more, we illustrate for normal distributions.

**Example 3.14 (Normal distribution).** Suppose that  $X \stackrel{d}{=} N(\mu_X, \sigma_X^2)$  and  $Y \stackrel{d}{=} N(\mu_Y, \sigma_Y^2)$  are independent. Then, by Theorem 3.12,

$$\begin{aligned} f_{X+Y}(t) &= \int_{-\infty}^{\infty} f_X(t-s)f_Y(s) ds \\ &= \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi\sigma_X^2}} e^{-(t-s-\mu_X)^2/2\sigma_X^2} \frac{1}{\sqrt{2\pi\sigma_Y^2}} e^{-(s-\mu_Y)^2/2\sigma_Y^2} ds \\ &= \frac{1}{\sqrt{2\pi(\sigma_X^2 + \sigma_Y^2)}} e^{-(t - [\mu_X + \mu_Y])^2/2(\sigma_X^2 + \sigma_Y^2)}, \end{aligned}$$

so that  $X + Y \stackrel{d}{=} N(\mu_X + \mu_Y, \sigma_X^2 + \sigma_Y^2)$ .

More generally, if  $X_1, \dots, X_n$  are independent and normally distributed, then every linear combination  $\sum_{i=1}^n a_i X_i$  is normally distributed.  $\square$

Here is another important case.

**Proposition 3.15.** Let  $X$  and  $Y$  be positive and integer-valued. Then, for each  $n \in \mathbb{N}$ ,

$$P\{X + Y = n\} = \sum_{k=0}^n P\{X = k\}P\{Y = n - k\}.$$

**Proof:** By independence of  $X$  and  $Y$ ,

$$\begin{aligned} P\{X + Y = n\} &= P\left(\sum_{k=0}^n \{X = k, Y = n - k\}\right) \\ &= \sum_{k=0}^n P\{X = k, Y = n - k\} \\ &= \sum_{k=0}^n P\{X = k\}P\{Y = n - k\}. \quad \blacksquare \end{aligned}$$

In many ways, the Poisson distributions parallel the normal distributions, as the following example substantiates.

**Example 3.16 (Poisson distribution).** Let  $X \stackrel{d}{=} P(\lambda_X)$  and  $Y \stackrel{d}{=} P(\lambda_Y)$  be independent. Then,  $X + Y \stackrel{d}{=} P(\lambda_X + \lambda_Y)$ : by Proposition 3.15,

$$\begin{aligned} P\{X + Y = n\} &= \sum_{k=0}^n P\{X = k\}P\{Y = n - k\} \\ &= \sum_{k=0}^n e^{-\lambda_X} \frac{\lambda_X^k}{k!} e^{-\lambda_Y} \frac{\lambda_Y^{n-k}}{(n-k)!} \\ &= e^{-(\lambda_X + \lambda_Y)} \frac{(\lambda_X + \lambda_Y)^n}{n!}. \quad \square \end{aligned}$$

### 3.3 Constructing Independent Random Variables

This section complements §2.6 with explicit techniques for constructing independent random variables with prescribed distribution functions.

#### 3.3.1 Finite families

**Theorem 3.17.** Let  $F_1, \dots, F_n$  be distribution functions on  $\mathbb{R}$ . Then, there exists a probability space  $(\Omega, \mathcal{F}, P)$  and random variables  $X_1, \dots, X_n$  on it, such that  $X_1, \dots, X_n$  are independent and  $F_{X_i} = F_i$  for each  $i$ .

**Proof:** Let  $\Omega = [0, 1]^n$ , let  $\mathcal{F}$  be the  $\sigma$ -algebra of Borel subsets of  $\Omega$ , and let  $P$  be the uniform distribution on  $\Omega$ . Define the coordinate random variables  $U_i(\omega) = \omega_i$ , which, according to Example 3.8, are independent with distribution  $U[0, 1]$ . Then, the random variables  $X_i = F_i^{-1}(U_i)$  are independent by Corollary 3.11, while  $X_i$  has distribution function  $F_i$  by Proposition 2.47. ■

#### 3.3.2 Sequences

**Theorem 3.18.** Let  $F_1, F_2, \dots$  be distribution functions on  $\mathbb{R}$ . Then, there exists a probability space  $(\Omega, \mathcal{F}, P)$ , on which are defined independent random variables  $X_1, X_2, \dots$ , such that  $F_{X_i} = F_i$  for each  $i$ .

**Proof:** The main issue is to construct an i.i.d. sequence  $U_1, U_2, \dots$  with  $U_i \stackrel{d}{=} U[0, 1]$  for each  $i$ . We do this by constructing a double sequence  $(Z_{ij})$  of independent, Bernoulli distributed ( $p = 1/2$ ) random variables, then

putting  $U_i = \sum_{j=1}^{\infty} 2^{-j} Z_{ji}$ . The  $U_i$  are independent by the disjoint blocks theorem, and each is uniformly distributed by Example 3.7.

Let  $(\Omega, \mathcal{F}) = ([0, 1], \mathcal{B}([0, 1]))$ , with  $P$  the uniform distribution. By a variant of Example 3.7, we may write the binary expansion of  $\omega \in \Omega$  as

$$\omega = .Z_{11}(\omega)Z_{21}(\omega)Z_{12}(\omega)Z_{31}(\omega)Z_{22}(\omega)Z_{13}(\omega)\dots,$$

where the  $Z_{ij}, i, j \geq 1$ , are independent random variables, each having a Bernoulli distribution with parameter  $p = 1/2$ . This amounts simply to a re-labeling of the random variables  $X_i$  in Example 3.7:  $Z_{11} = X_1$ ,  $Z_{21} = X_2$ ,  $Z_{12} = X_3$ ,  $Z_{31} = X_4$  and so on. Visually, one is “snaking” through the two-dimensional lattice. The random variables

$$U_i(\omega) = \sum_{j=1}^{\infty} 2^{-j} Z_{ji}(\omega);$$

are independent by Theorem 3.10, and  $U_i \stackrel{d}{=} U[0, 1]$  by Example 3.7. Hence, by Proposition 2.47, we may take  $X_i = F_i^{-1}(U_i)$ . ■

## 3.4 Independent Events

Independence for events is independence of their indicator functions as random variables.

**Definition 3.19.** Events  $A_1, \dots, A_n$  are *independent* if their indicator functions are independent random variables.

An infinite collection of events is *independent* if every finite subcollection is independent. □

We have the following criterion for independence.

**Theorem 3.20.** Events  $A_1, \dots, A_n$  are independent if and only if

$$P(\bigcap_{i \in I} A_i) = \prod_{i \in I} P(A_i) \tag{3.9}$$

for every subset  $I$  of  $\{1, \dots, n\}$ .

**Proof:** Necessity. If  $A_1, \dots, A_n$  are independent, then for each  $I \subseteq \{1, \dots, n\}$ , the random variables  $\{\mathbf{1}_{A_i} : i \in I\}$  are independent, and hence,

$$P(\bigcap_{i \in I} A_i) = P\{\mathbf{1}_{A_i} = 1, i \in I\} = \prod_{i \in I} P\{\mathbf{1}_{A_i} = 1\} = \prod_{i \in I} P(A_i).$$

Sufficiency. By Theorem 3.4, to show that  $\mathbf{1}_{A_1}, \dots, \mathbf{1}_{A_n}$  are independent, it is enough to show that for each  $I \subseteq \{1, \dots, n\}$ ,

$$\begin{aligned} P(\{\mathbf{1}_{A_i} = 1, i \in I\} \cap \{\mathbf{1}_{A_j} = 0, j \notin I\}) \\ = \prod_{i \in I} P(A_i) \times \prod_{j \notin I} [1 - P(A_j)], \end{aligned} \quad (3.10)$$

which we show by induction on  $|I^c|$ , the cardinality of  $I^c$ . For  $|I^c| = 0$ ,  $I = \{1, \dots, n\}$ , and (3.10) is the same as (3.9). Assuming that (3.10) holds for all  $I$  such that  $|I^c| = k - 1$ , let  $I$  satisfy  $|I^c| = k$ , and let  $I' = I + \{j_0\}$  for some  $j_0 \in I^c$ . Then,

$$\begin{aligned} P(\{\mathbf{1}_{A_i} = 1, i \in I\} \cap \{\mathbf{1}_{A_j} = 0, j \notin I\}) \\ = P(\{\mathbf{1}_{A_i} = 1, i \in I\} \cap \{\mathbf{1}_{A_j} = 0, j \notin I'\}) \\ - P(\{\mathbf{1}_{A_i} = 1, i \in I'\} \cap \{\mathbf{1}_{A_j} = 0, j \notin I'\}) \\ = \prod_{i \in I} P(A_i) \times \prod_{j \notin I'} [1 - P(A_j)] - \prod_{i \in I'} P(A_i) \times \prod_{j \notin I'} [1 - P(A_j)] \end{aligned}$$

[by the induction hypothesis, since (3.10) holds for  $I'$ ]

$$= \prod_{i \in I} P(A_i) \times \prod_{j \notin I} [1 - P(A_j)],$$

so that (3.10) holds for  $I$ . ■

As a consequence, independence of events is equivalent to that of their complements.

**Corollary 3.21.** Events  $A_1, \dots, A_n$  are independent if and only if  $A_1^c, \dots, A_n^c$  are independent.

**Proof:** Independence of  $A_1, \dots, A_n$  means that  $\mathbf{1}_{A_1}, \dots, \mathbf{1}_{A_n}$  are independent random variables. This in turn implies, by Corollary 3.11, that  $\mathbf{1}_{A_1^c} = 1 - \mathbf{1}_{A_1}, \dots, \mathbf{1}_{A_n^c} = 1 - \mathbf{1}_{A_n}$  are independent, and, hence, that  $A_1^c, \dots, A_n^c$  are independent events.

The converse also holds true, since  $(A^c)^c = A$  for each event  $A$ . ■

For independent events, Theorem 1.27, the Borel-Cantelli lemma, has a converse.

**Theorem 3.22 (Borel-Cantelli lemma).** Let  $A_1, A_2, \dots$  be independent events such that  $\sum_{n=1}^{\infty} P(A_n) = \infty$ . Then,  $P\{A_n, \text{i.o.}\} = 1$ .

**Proof:** Since  $\{A_n, \text{i.o.}\} = \bigcap_{n=1}^{\infty} \bigcup_{m=n}^{\infty} A_m$ , we show that  $P\left(\bigcup_{m=n}^{\infty} A_m\right) = 1$  for each  $n$ . By Theorem 1.25,

$$\begin{aligned} P\left(\bigcup_{m=n}^{\infty} A_m\right) &= \lim_{k \rightarrow \infty} P\left(\bigcup_{m=n}^k A_m\right) \\ &= 1 - \lim_{k \rightarrow \infty} P\left(\bigcap_{m=n}^k A_m^c\right) \\ &= 1 - \lim_{k \rightarrow \infty} \prod_{m=n}^k [1 - P(A_m)] \end{aligned}$$

[by Corollary 3.21, since  $A_n^c, \dots, A_k^c$  are independent]

$$\geq 1 - \lim_{k \rightarrow \infty} \exp\left[-\sum_{m=n}^k P(A_m)\right]$$

since  $e^{-x} \geq 1 - x$ , and the limit is zero because  $\sum_{n=1}^{\infty} P(A_n) = \infty$ . ■

### 3.5 Occupancy Models

This section and the next illustrate and synthesize the fundamental concepts introduced so far: probability, random variables and independence.

Our setting is the classical context of “equally likely outcomes”: the sample space is a finite set  $\Omega$ , the family  $\mathcal{F}$  of events consists of all subsets of  $\Omega$ , and  $P$  is the uniform distribution on  $\Omega$ , given by

$$P(A) = |A|/|\Omega|,$$

where  $|A|$  (the cardinality of  $A$ ) is the number of outcomes in  $A$ . Computation of probabilities reduces, as for many calculations involving random walks, to counting numbers of outcomes in events of interest, the provenance of combinatorics. For concreteness, we focus on one specific, illustrative set of questions: occupancy problems. Instead of considering only determination of  $|A|$  for various subsets, we pursue a broader approach that incorporates construction of random variables representing physically meaningful functions of outcomes, and analysis of asymptotics.

Before proceeding any further, we present some fundamental counting rules.

**Counting Rules.** Let  $\Omega$  be a finite set with  $|\Omega| = n$ . Then,

1. The number of ways to choose  $k \leq n$  *distinct, ordered* elements from  $\Omega$  is  $(n)_k = n(n-1)\cdots(n-k+1)$ . In particular, the number of distinct orderings of  $\Omega$  is  $n! = n(n-1)\cdots 1$ .
2.  $|\Omega^k| = n^k$  for each  $k$ .

3. The number of size- $k$  subsets of  $\Omega$  is

$$\binom{n}{k} \stackrel{\text{def}}{=} \frac{n!}{k!(n-k)!}.$$

4. The number of subsets of  $\Omega$  is  $|\mathcal{P}(\Omega)| = 2^{|\Omega|}$ .  $\square$

### 3.5.1 Four occupancy models

Occupancy problems are formulated verbally as follows:  $k$  particles (balls) are distributed randomly among  $n$  distinguishable cells (boxes) in such a manner that “all configurations are equally likely.” The goal is to calculate various probabilities and other quantities of interest.

Because of ambiguities inherent in this description, to perform a mathematical analysis one needs to:

1. Decide what a “configuration” is, and specify an appropriate sample space  $\Omega$ .
2. Calculate  $|\Omega|$ .
3. Calculate  $|A|$ , and, hence,  $P(A) = |A|/|\Omega|$ , for various events  $A$ .
4. Define random variables of interest, perform calculations for them, and identify relationships — especially independence — among them.
5. Investigate asymptotics as  $n \rightarrow \infty$  and  $k \rightarrow \infty$ .

Reduction of a verbal description to a mathematical representation (ordinarily simplified) is a central part of mathematical modeling. As often happens, the one verbal description of occupancy leads to several distinct mathematical formulations. There are two key dichotomies: whether particles are distinguishable, and whether multiple occupancy of cells is permitted.

We now introduce the four models, three of which are named after pairs of distinguished physicists. (The fourth, strangely, seems to have no name.) In each, the  $\omega_i$  assume only nonnegative integer values.

**Model 3.23 (Maxwell-Boltzmann model).** Here, particles are distinguishable and multiple occupancy is allowed. The sample space is

$$\Omega = \{(\omega_1, \dots, \omega_k) : 1 \leq \omega_i \leq n \text{ for all } i\},$$

with  $\omega_i$  the cell in which particle  $i$  is located. Thus,  $|\Omega| = n^k$ .  $\square$

The next model, and the Fermi-Dirac model that follows, represent an “exclusion principle,” which prevents a cell’s being occupied by more than one particle.

**Model 3.24.** In the “model without a name,” particles are distinguishable, and multiple occupancy of cells is forbidden. Obviously, we must have  $k \leq n$ . The sample space is

$$\Omega = \{(\omega_1, \dots, \omega_k) : 1 \leq \omega_i \leq n \text{ for all } i, \omega_i \neq \omega_{i'} \text{ for } i \neq i'\},$$

where  $\omega_i$  is the cell in which particle  $i$  is located. Hence,  $|\Omega| = (n)_k$ .  $\square$

**Model 3.25 (Fermi-Dirac model).** For this model, particles are indistinguishable and multiple occupancy is forbidden. A configuration is simply the set of occupied cells, so we take

$$\Omega = \{(\omega_1, \dots, \omega_n) : \omega_j = 0 \text{ or } 1 \text{ for all } j \text{ and } \sum_{j=1}^n \omega_j = k\},$$

with cell  $j$  occupied if and only if  $\omega_j = 1$ , so that  $|\Omega| = \binom{n}{k}$ .  $\square$

**Model 3.26 (Bose-Einstein model).** This model depicts indistinguishable particles with multiple occupancy allowed. A configuration lists the number of particles in each cell, so that

$$\Omega = \{(\omega_1, \dots, \omega_n) : \omega_j \geq 0 \text{ for all } j \text{ and } \sum_{j=1}^n \omega_j = k\},$$

with  $\omega_j$  the number of balls in cell  $j$ .

Determination of  $|\Omega|$  is more difficult for this model than the others. The result, that

$$|\Omega| = \binom{n+k-1}{k}. \quad (3.11)$$

may be argued heuristically as follows. A configuration is an ordered arrangement of  $n+k-1$  symbols,  $k$  of which represent particles and  $n-1$  of which are boundaries between cells. Of the  $n+k-1$  symbols, the locations of the  $k$  particles may be chosen in  $\binom{n+k-1}{k}$  ways.  $\square$

### 3.5.2 Occupancy numbers

We now illustrate not only the explicit construction of random variables of physical interest, but also computation of their joint distributions, by examining occupancy numbers. Let  $X_j$  be the number of particles in cell  $j$ . Note carefully how the mathematical definitions of the  $X_j$  differ from model to model.

We begin by calculating the joint distribution of  $X_1, \dots, X_n$ .

**Proposition 3.27.** Suppose that  $k_j \geq 0$  for each  $j$  and  $\sum_{j=1}^n k_j = k$ .

- a) For the Maxwell-Boltzmann model,

$$P\{X_1 = k_1, \dots, X_n = k_n\} = \frac{k!}{k_1! \dots k_n!} n^{-k}. \quad (3.12)$$

- b) For Model 3.24 and the Fermi-Dirac model,

$$P\{X_1 = k_1, \dots, X_n = k_n\} = \binom{n}{k}^{-1} \quad (3.13)$$

for  $k_1, \dots, k_n$  each equal to zero or one.

- c) For the Bose-Einstein model,

$$P\{X_1 = k_1, \dots, X_n = k_n\} = \binom{n+k-1}{k}^{-1}. \quad (3.14)$$

**Proof:** For the Fermi-Dirac and Bose-Einstein models, the configuration is the set of occupancy numbers:  $\omega = (X_1(\omega), \dots, X_n(\omega))$ . Consequently,  $P\{X_1 = k_1, \dots, X_n = k_n\} = 1/|\Omega|$ , which gives c) and the second part of b).

Under the Maxwell-Boltzmann model,  $X_j(\omega) = \sum_{i=1}^k \mathbf{1}(\omega_i = j)$ , so

$$\begin{aligned} |\{X_1 = k_1, \dots, X_n = k_n\}| &= \left| \left\{ \omega : \sum_{i=1}^k \mathbf{1}(\omega_i = j) = k_j, j = 1, \dots, n \right\} \right| \\ &= \frac{k!}{k_1! \cdots k_n!}. \end{aligned}$$

Finally, for the first part of b), again  $X_j(\omega) = \sum_{i=1}^k \mathbf{1}(\omega_i = j)$ , but each  $k_j$  must be zero or one. This choice gives  $|\{X_1 = k_1, \dots, X_n = k_n\}| = k!$ , which shows that (3.13) holds. ■

Let us now consider the  $X_j$  individually. In principle, we could deduce  $P\{X_1 = \ell\}$  using Proposition 3.27, but direct computation, coupled with identification and application of independence properties, yields the results more easily.

**Proposition 3.28.** Let  $X_1$  be the number of particles in cell 1.

- a) For the Maxwell-Boltzmann model,  $X_1 \stackrel{d}{=} B(k, 1/n)$ :

$$P\{X_1 = \ell\} = \binom{k}{\ell} \left(\frac{1}{n}\right)^\ell \left(1 - \frac{1}{n}\right)^{k-\ell}, \quad \ell = 0, \dots, k. \quad (3.15)$$

- b) For Model 3.24 and the Fermi-Dirac model,  $X_1$  has a Bernoulli distribution with parameter  $k/n$ :

$$P\{X_1 = 1\} = 1 - P\{X_1 = 0\} = k/n. \quad (3.16)$$

- c) For the Bose-Einstein model, for  $\ell = 0, \dots, k$ ,

$$P\{X_1 = \ell\} = \binom{n+k-\ell-2}{k-\ell} / \binom{n+k-1}{k}. \quad (3.17)$$

**Proof:** a) Let  $Z_i(\omega) = \omega_i$  be the number of the cell containing particle  $i$ . Then, these random variables are i.i.d., with  $P\{Z_i = j\} = 1/n$ ,  $j = 1, \dots, n$ . Indeed,

$$P\{Z_i = j\} = P(\{\omega : \omega_i = j\}) = \frac{|\{\omega : \omega_i = j\}|}{n^k} = \frac{n^{k-1}}{n^k},$$

and independence follows from Theorem 3.4, since

$$P\{Z_1 = j_1, \dots, Z_k = j_k\} = \left(\frac{1}{n}\right)^k = \prod_{i=1}^k P\{Z_i = j_i\}.$$

In order that  $X_1 = \ell$ , exactly  $\ell$  of the  $Z_i$  must equal one. Consequently,

$$\begin{aligned} P\{X_1 = \ell\} &= \sum_{|B|=\ell} P\{Z_i = 1 \iff i \in B\} \\ &= \sum_{|B|=\ell} \left[ \prod_{i \in B} P\{Z_i = 1\} \times \prod_{i \notin B} P\{Z_i \neq 1\} \right] \\ &= \sum_{|B|=\ell} \left(\frac{1}{n}\right)^\ell \left(1 - \frac{1}{n}\right)^{k-\ell} \\ &= \binom{k}{\ell} \left(\frac{1}{n}\right)^\ell \left(1 - \frac{1}{n}\right)^{k-\ell}. \end{aligned}$$

b) For both cases, the number of outcomes for which  $X_1 = 1$  is the number of ways in which one particle may be chosen for cell 1 times the number of ways in which the remaining  $k - 1$  particles may be distributed among the other  $n - 1$  cells. Thus, for Model 3.24,

$$P\{X_1 = 1\} = k \frac{(n-1)_{k-1}}{(n)_k} = \frac{k}{n}.$$

A similar argument applies to the Fermi-Dirac model.

c) The reasoning used for Model 3.24 applies, but because particles are indistinguishable, there is only one way to choose  $\ell$  particles for cell one, and  $\binom{n-1+(k-\ell)-1}{k-\ell} = \binom{n+k-\ell-2}{k-\ell}$  ways to distribute the others. ■

### 3.5.3 Asymptotics

Even closed-form expressions such as (3.15) and (3.17) have limited usefulness. Their numerical values are difficult to calculate, even by computer, or to estimate. A standard response in these situations is to study *asymptotics*, with the goal of obtaining more tractable formulas, perhaps involving fewer parameters (for example,  $k/n$ , the number of particles per cell). We

conclude with two limit theorems, in which both  $k$  and  $n$  converge to infinity, but in such a manner that  $\lambda = \lim k/n$  exists, and is positive and finite: physically, the number of particles per cell stabilizes.

A key analytical tool is *Stirling's approximation* to  $k!$ :

$$\lim_{k \rightarrow \infty} \frac{k!}{\sqrt{2\pi} k^{k+1/2} e^{-k}} = 1,$$

which we abbreviate as

$$k! \cong \sqrt{2\pi} k^{k+1/2} e^{-k}. \quad (3.18)$$

Stirling's approximation is unavoidable in proving Theorem 3.30, and can be used in the proof of Theorem 3.29 as well, but it is easier not to.

For the Maxwell-Boltzmann model the asymptotic distribution of  $X_1$  is Poisson. Analytically, the following theorem states that a binomial probability  $\binom{k}{\ell} p^\ell (1-p)^{k-\ell}$  in which  $k$  is large and  $p$  is small (in this case,  $p = 1/n$ ) is approximately the Poisson probability  $e^{-\lambda} \lambda^\ell / \ell!$ , where  $\lambda = kp$ . See §5.5 and §6.4 for more general Poisson approximation theorems.

**Theorem 3.29.** For the Maxwell-Boltzmann model,

$$\lim_{\substack{n \rightarrow \infty, k \rightarrow \infty \\ k/n \rightarrow \lambda \in (0, \infty)}} P\{X_1 = \ell\} = e^{-\lambda} \frac{\lambda^\ell}{\ell!}, \quad \ell \geq 0.$$

**Proof:** By Proposition 3.28,

$$\begin{aligned} P\{X_1 = \ell\} &= \frac{k!}{\ell! (k-\ell)!} \left(\frac{1}{n}\right)^\ell \left(1 - \frac{1}{n}\right)^{k-\ell} \\ &\cong \frac{k(k-1)\dots(k-\ell+1)}{\ell!} \left(\frac{1}{n}\right)^\ell \left(1 - \frac{1}{n}\right)^{k-\ell} \\ &= \frac{k}{n} \dots \frac{k-\ell+1}{n} \frac{1}{\ell!} \left(1 - \frac{k/n}{k}\right)^{k-\ell} \\ &\rightarrow e^{-\lambda} \frac{\lambda^\ell}{\ell!}. \quad \blacksquare \end{aligned}$$

For the Bose-Einstein model the limiting distribution of  $X_1$  is a modified geometric distribution.

**Theorem 3.30.** For the Bose-Einstein model,

$$\lim_{\substack{n \rightarrow \infty, k \rightarrow \infty \\ k/n \rightarrow \lambda \in (0, \infty)}} P\{X_1 = \ell\} = \frac{1}{\lambda + 1} \left(\frac{\lambda}{\lambda + 1}\right)^\ell, \quad \ell \geq 0.$$

**Proof:** By Proposition 3.28,

$$\begin{aligned} P\{X_1 = \ell\} &= \frac{(n+k-\ell-2)! k! (n-1)!}{(k-\ell)! (n-2)! (n-k+1)!} \\ &\cong \frac{(n+k-\ell-2)^{n+k-\ell-3/2} k^{k+1/2} (n-1)^{n-1/2}}{(k-\ell)^{k-\ell+1/2} (n-2)^{n-3/2} (n+k-1)^{n+k-1/2}} \end{aligned}$$

[Stirling's approximation applies to each factorial; the exponentials and  $\sqrt{2\pi}$  terms cancel]

$$\begin{aligned} &= \frac{n-1}{n+k-1} \left( \frac{k-\ell}{n+k-\ell-2} \right)^\ell \left( \frac{n-1}{n-2} \right)^{n-3/2} \\ &\quad \times \left( \frac{n+k-\ell-2}{n+k-1} \right)^{n+k-3/2} \left( \frac{k}{k-\ell} \right)^{k+1/2} \\ &\rightarrow \frac{1}{\lambda+1} \left( \frac{\lambda}{\lambda+1} \right)^\ell \cdot e \cdot e^{-(\ell+1)} e^\ell. \quad \blacksquare \end{aligned}$$

## 3.6 Bernoulli and Poisson Processes

In this section we examine two important classes of stochastic processes, namely, Bernoulli and Poisson processes.

### 3.6.1 Bernoulli processes

Bernoulli processes count successes in repeated, independent trials, each of which has one of two possible outcomes, labeled for concreteness as “success” and “failure,” but without any implication that one is preferred to the other.

**Definition 3.31.** A *Bernoulli process* with parameter  $p$  is a sequence  $(X_i)$  of independent random variables, each having a Bernoulli distribution with parameter  $p$ :  $P\{X_i = 1\} = 1 - P\{X_i = 0\} = p$ .  $\square$

The interpretation is that  $i$  is a discrete time parameter, and that  $X_i = 1$  represents a success at time  $i$  and  $X_i = 0$  a failure. Let  $q = 1 - p$ .

In isolation a Bernoulli process is neither very deep nor very interesting. Two processes constructed from it have more complex, interesting structure. The *success counting process* counts cumulative numbers of successes:  $S_n = \sum_{i=1}^n X_i$  is the number of successes in the first  $n$  trials. The *success time process* gives times at which successes occur: for each  $k$ ,

$$T_k = \min \{n : S_n = k\}$$

is the “time” (trial number) at which the  $k$ th success occurs. Also, the random variables

$$U_k = T_k - T_{k-1}$$

are the times between successes.

We now derive properties of these processes, beginning with the distributions of the  $S_n$ . In the process, we relate Bernoulli and binomial distributions: sums of independent, Bernoulli distributed random variables have binomial distributions.

**Proposition 3.32.** For each  $n$ ,  $S_n \stackrel{d}{=} B(n, p)$ .

**Proof:** For  $k \leq n$ ,

$$\begin{aligned} P\{S_n = k\} &= P\left(\sum_{|J|=k} \{X_i = 1 \text{ for } i \in J, X_i = 0 \text{ for } i \notin J\}\right) \\ &= \sum_{|J|=k} P\{X_i = 1 \text{ for } i \in J, X_i = 0 \text{ for } i \notin J\} \\ &= \sum_{|J|=k} p^k q^{n-k} \\ &= \binom{n}{k} p^k q^{n-k}. \quad \blacksquare \end{aligned}$$

Next we calculate the distributions of the  $T_k$ .

**Proposition 3.33.** For each  $k$ ,  $T_k$  has a negative binomial distribution with parameters  $k$  and  $p$ :

$$P\{T_k = m\} = \binom{m-1}{k-1} p^k q^{m-k}, \quad m \geq k.$$

In particular,  $T_1 = U_1$  has a geometric distribution with parameter  $p$ .

**Proof:** For  $m \geq k$ ,

$$\{T_k = m\} = \{S_{m-1} = k-1, X_m = k\},$$

which is an event, so that  $T_k$  is a random variable. By Proposition 3.32,

$$\begin{aligned} P\{T_k = m\} &= P\{S_{m-1} = k-1, X_m = 1\} \\ &= P\{S_{m-1} = k-1\} P\{X_m = 1\} \end{aligned}$$

[by Theorem 3.10,  $S_{m-1} = \sum_{i=1}^{m-1} X_i$  and  $X_m$  are independent]

$$= \binom{m-1}{k-1} p^{k-1} q^{(m-1)-(k-1)} p. \quad \blacksquare$$

In particular, this resolves a point from Example 2.33:

$$\sum_{k=m}^{\infty} \binom{m-1}{k-1} p^k q^{m-k} = \sum_{k=m}^{\infty} P\{T_k = m\} = 1.$$

Albeit useful, Proposition 3.32 and Proposition 3.33 tell only part of the story. The following two results, which identify independence relationships within the sequences  $(S_n)$  and  $(T_k)$ , are more informative.

**Proposition 3.34.** The success counting process  $(S_n)$  has *independent and stationary increments*:

- a) For  $0 < n_1 < \dots < n_k$ , the random variables  $S_{n_1}, S_{n_2} - S_{n_1}, \dots, S_{n_k} - S_{n_{k-1}}$  are independent.
- b) For fixed  $j$ , the distribution of  $S_{k+j} - S_k$  is the same for all  $k$ .

**Proof:** a) By Theorem 3.10, increments in  $(S_n)$  over disjoint time intervals are functions of disjoint blocks of the  $X_i$ .

b) For each  $k$ ,

$$S_{k+j} - S_k = \sum_{i=k+1}^{k+j} X_i,$$

whose distribution does not depend on  $k$  because the  $X_i$  are identically distributed. ■

The success time sequence also has independent and stationary increments.

**Theorem 3.35.** The inter-success times  $U_1, U_2, \dots$  are i.i.d.

**Proof:** For integers  $m_1, m_2, \dots \geq 1$  and for each  $j$ , with  $\ell_j = \sum_{i=1}^j m_i$ ,

$$\begin{aligned} P\{U_1 = m_1, \dots, U_j = m_j\} \\ &= P\{S_{\ell_1-1} = 0, X_{\ell_1} = 1, S_{\ell_2-1} - S_{\ell_1} = 0, X_{\ell_2} = 1, \dots \\ &\quad S_{\ell_j-1} - S_{\ell_{j-1}} = 0, X_{\ell_j} = 1\} \\ &= P\{S_{\ell_1-1} = 0\}P\{X_{\ell_1} = 1\}P\{S_{\ell_2-1} - S_{\ell_1} = 0\} \\ &\quad \times P\{X_{m_1+m_2} = 1\} \cdots P\{S_{\ell_j-1} - S_{\ell_{j-1}} = 0\}P\{X_{\ell_j} = 1\} \\ &= P\{S_{m_1-1} = 0\}P\{X_1 = 1\}P\{S_{m_2-1} = 0\}P\{X_1 = 1\} \cdots \\ &\quad P\{S_{m_j-1} = 0\}P\{X_1 = 1\} \\ &= (q^{m_1-1}p)(q^{m_2-1}p) \cdots (q^{m_j-1}p) \\ &= P\{U_1 = m_1\} \cdots P\{U_j = m_j\}. \quad \blacksquare \end{aligned}$$

Theorem 3.35 exhibits the relationship between negative binomial and geometric distributions: a random variable whose distribution is negative binomial with parameters  $n$  and  $p$  is the sum of  $n$  independent random variables, each having a geometric distribution with parameter  $p$ .

### 3.6.2 Poisson processes

Poisson processes are continuous time analogues of Bernoulli processes. Within the milieu of stochastic processes, they are arrival counting processes. An *arrival counting process* is a stochastic process  $N = (N_t)_{t \geq 0}$  such that for each  $\omega$ ,  $N_0(\omega) = 0$  and the sample path  $t \mapsto N_t(\omega)$ , a function from  $\mathbb{R}_+$  to  $\mathbb{R}$ , is increasing and right-continuous, increases only by jumps of size 1, and has only finitely many jumps in any bounded time interval. The interpretation is that  $N_t$  is the number of “arrivals” in some system (for example, a queue), or the number of events of some kind that have occurred, during the interval  $(0, t]$ . More generally, for  $s < t$ ,  $N_t - N_s$  is the number of arrivals in  $(s, t]$ .

Associated with an arrival counting process are two important sequences of random variables, representing *times of arrivals* and *times between arrivals*. For each  $k$ ,

$$T_k = \inf \{t : N_t = k\}$$

is the time of the  $k$ th arrival. The random variables  $U_k = T_k - T_{k-1}$  (with  $T_0 = 0$ ) are then the times between successive arrivals.

We now define Poisson processes.

**Definition 3.36.** An arrival counting process  $N = (N_t)$  is a *Poisson process with rate  $\lambda > 0$*  if

- a)  $N$  has independent increments: for  $0 < t_1 < \dots < t_n$ , the random variables  $N_{t_1}, N_{t_2} - N_{t_1}, \dots, N_{t_n} - N_{t_{n-1}}$  are independent.
- b)  $N$  has stationary increments: for fixed  $s$ , the distribution of the increment  $N_{t+s} - N_t$  is the same for all  $t$ .
- c) For each  $t$ ,  $N_t$  has a Poisson distribution with parameter  $\lambda t$ .

In fact, part c) follows from a) and b) (see Billingsley, 1990), but rather than prove that here, we simply incorporate it into the definition.

Using this definition, it is straightforward to derive the joint distribution of the numbers of arrivals for any finite number of times.

**Proposition 3.37.** For  $0 < t_1 < \dots < t_n$  and  $k_1 \leq \dots \leq k_n$ ,

$$\begin{aligned} P\{N_{t_1} = k_1, \dots, N_{t_n} = k_n\} \\ = \prod_{\ell=1}^n \left[ e^{-\lambda(t_\ell - t_{\ell-1})} \frac{[\lambda(t_\ell - t_{\ell-1})]^{k_\ell - k_{\ell-1}}}{(k_\ell - k_{\ell-1})!} \right]. \end{aligned}$$

**Proof:** In the computation, observe how the independent increments property is used before that of stationary increments. With  $t_0 = 0$ ,

$$\begin{aligned} P\{N_{t_1} = k_1, \dots, N_{t_n} = k_n\} &= P\{N_{t_1} = k_1, N_{t_2} - N_{t_1} = k_2 - k_1, \dots, N_{t_n} - N_{t_{n-1}} = k_n - k_{n-1}\} \\ &= \prod_{\ell=1}^n P\{N_{t_\ell} - N_{t_{\ell-1}} = k_\ell - k_{\ell-1}\} \\ &= \prod_{\ell=1}^n P\{N_{t_\ell - t_{\ell-1}} = k_\ell - k_{\ell-1}\} \\ &= \prod_{\ell=1}^n \left[ e^{-\lambda(t_\ell - t_{\ell-1})} \frac{[\lambda(t_\ell - t_{\ell-1})]^{k_\ell - k_{\ell-1}}}{(k_\ell - k_{\ell-1})!} \right]. \quad \blacksquare \end{aligned}$$

Next, we consider distributional properties of the arrival and interarrival times. For each  $k$ ,  $T_k$ , the time of the  $k$ th arrival, has an Erlang distribution with parameters  $k$  and  $\lambda$  (Example 2.37). In particular,  $T_1 = U_1$  has an exponential distribution with parameter  $\lambda$ .

**Proposition 3.38.** For each  $k$ ,  $T_k$  is absolutely continuous with

$$f_{T_k}(t) = \frac{1}{(k-1)!} \lambda^k e^{-\lambda t} t^{k-1}, \quad t > 0.$$

**Proof:** The fundamental relationship is that for each  $k$  and  $t$ ,

$$\{T_k \leq t\} = \{N_t \geq k\}.$$

Not only does this imply that  $T_k$  is a random variable, but also

$$P\{T_k \leq t\} = P\{N_t \geq k\} = \sum_{\ell=k}^{\infty} e^{-\lambda t} \frac{(\lambda t)^\ell}{\ell!}.$$

From this expression, we conclude that

$$\begin{aligned} \frac{d}{dt} P\{T_k \leq t\} &= \frac{d}{dt} \sum_{\ell=k}^{\infty} e^{-\lambda t} \frac{(\lambda t)^\ell}{\ell!} \\ &= \sum_{\ell=k}^{\infty} e^{-\lambda t} \left[ \lambda \frac{(\lambda t)^{\ell-1}}{(\ell-1)!} - \lambda \frac{(\lambda t)^\ell}{\ell!} \right] \\ &= \frac{1}{(k-1)!} \lambda^k e^{-\lambda t} t^{k-1}. \quad \blacksquare \end{aligned}$$

The interarrival times in a Poisson process, like the times between successes in a Bernoulli process, are i.i.d.

**Theorem 3.39.** The interarrival times  $U_1, U_2, \dots$  are independent and each has distribution  $E(\lambda)$ .

**Proof:** It suffices to show that  $U_1, \dots, U_k$  are independent (and have the proper distribution) for each  $k$ . To this end, we first show that  $(T_1, \dots, T_k)$  has density

$$\tilde{f}_k(t_1, \dots, t_k) = \lambda^k e^{-\lambda t_k}, \quad t_1 < \dots < t_k. \quad (3.19)$$

Suppose that  $t_1 < \dots < t_k$  and that  $h_1, \dots, h_k$  are strictly positive, but small enough that the intervals  $(t_i, t_i + h_i]$ ,  $i = 1, \dots, k$ , are disjoint. Then,

$$\begin{aligned} P\{T_1 \in (t_1, t_1 + h_1), \dots, T_k \in (t_k, t_k + h_k)\} \\ &= P\{N_{t_1} = 0, N_{t_1+h_1} - N_{t_1} = 1, N_{t_2} - N_{t_1+h_1} = 0, \dots, \\ &\quad N_{t_k} - N_{t_{k-1}+h_{k-1}} = 0, N_{t_k+h_k} - N_{t_k} = 1\} \\ &= P\{N_{t_1} = 0\}P\{N_{t_1+h_1} - N_{t_1} = 1\}P\{N_{t_2} - N_{t_1+h_1} = 0\} \cdots \\ &\quad P\{N_{t_k} - N_{t_{k-1}+h_{k-1}} = 0\}P\{N_{t_k+h_k} - N_{t_k} = 1\} \\ &= P\{N_{t_1} = 0\}P\{N_{h_1} = 1\}P\{N_{t_2-t_1-h_1} = 0\} \cdots P\{N_{h_k} = 1\} \\ &= e^{-\lambda t_1} \times \lambda e^{-\lambda h_1} h_1 \times e^{-\lambda[t_2-(t_1+h_1)]} \times \cdots \\ &\quad \times e^{-\lambda[t_k-(t_{k-1}+h_{k-1})]} \times \lambda e^{-\lambda h_k} h_k \\ &= \lambda^k e^{-\lambda(t_k+h_k)} \prod_{i=1}^k h_i. \end{aligned}$$

Then, (3.19) follows by dividing both sides of this expression by  $\prod_{i=1}^k h_i$  and taking limits as  $\max_i h_i \rightarrow 0$ .

Now define  $g: \mathbb{R}_+^k \rightarrow \mathbb{R}_+^k$  by

$$g(t_1, \dots, t_k) = (t_1, t_2 - t_1, \dots, t_k - t_{k-1}),$$

so that  $(U_1, \dots, U_k) = g(T_1, \dots, T_k)$ . Then,  $g$  is invertible with inverse

$$h(u_1, \dots, u_k) = (u_1, u_1 + u_2, \dots, u_1 + \cdots + u_k),$$

and the Jacobian  $Jh$  is identically one, so that by (2.14), the density  $f_k$  of  $(U_1, \dots, U_k)$  is

$$\begin{aligned} f_k(u_1, \dots, u_k) &= \tilde{f}_k(h(u_1, \dots, u_k)) = \lambda^k e^{-\lambda(u_1+\cdots+u_k)} \\ &= \prod_{i=1}^k \lambda e^{-\lambda u_i}. \end{aligned}$$

Thus,  $U_i \stackrel{d}{=} E(\lambda)$ , and independence follows. ■

## 3.7 Complements

### 3.7.1 Independent $\sigma$ -algebras

We have adopted independence for random variables as our basic form and formulated independence for events in terms of it. A more general concept, independence for  $\sigma$ -algebras, subsumes both.

**Definition 3.40.** Sub- $\sigma$ -algebras  $\mathcal{G}_1, \dots, \mathcal{G}_n$  of  $\mathcal{F}$  are *independent* if

$$P(\bigcap_{i=1}^n A_i) = \prod_{i=1}^n P(A_i)$$

for all  $A_i \in \mathcal{G}_i$ ,  $i = 1, \dots, n$ .

An infinite collection of  $\sigma$ -algebras is *independent* if every finite subcollection is independent.  $\square$

In terms of this definition, random variables  $X_i$  are independent if the  $\sigma$ -algebras  $\sigma(X_i)$  are independent, and events  $A_i$  are independent if the  $\sigma$ -algebras  $\sigma(A_i) = \{\emptyset, A_i, A_i^c, \Omega\}$  are independent.

A simplified criterion for independence is valid for  $\sigma$ -algebras.

**Theorem 3.41.** For each  $i$  let  $S_i$  be a  $\pi$ -system generating  $\mathcal{G}_i$ . Then,  $\mathcal{G}_1, \dots, \mathcal{G}_n$  are independent if and only if  $P(\bigcap_{i=1}^n A_i) = \prod_{i=1}^n P(A_i)$  for all  $A_i \in S_i$ ,  $i = 1, \dots, n$ .  $\square$

### 3.7.2 Products of probability spaces

Let  $(\Omega_1, \mathcal{F}_1, P_1)$  and  $(\Omega_2, \mathcal{F}_2, P_2)$  be probability spaces. Here we construct their *product*, whose basic component is a probability  $P$  on the Cartesian product

$$\Omega_1 \times \Omega_2 = \{(\omega_1, \omega_2) : \omega_1 \in \Omega_1, \omega_2 \in \Omega_2\}$$

satisfying

$$P(A_1 \times A_2) = P_1(A_1)P_2(A_2)$$

for all rectangles  $A_1 \times A_2$ . As we see momentarily, the construction, which involves products, is intimately related to independence.

For the family of events we take the smallest  $\sigma$ -algebra containing the sets of interest, namely, the rectangles.

**Definition 3.42.** The *product* of  $\mathcal{F}_1$  and  $\mathcal{F}_2$  is the  $\sigma$ -algebra on  $\Omega_1 \times \Omega_2$  generated by the  $\pi$ -system  $\mathcal{S} = \{A_1 \times A_2 : A_1 \in \mathcal{F}_1, A_2 \in \mathcal{F}_2\}$  of rectangles. We denote the product  $\sigma$ -algebra by  $\mathcal{F}_1 \times \mathcal{F}_2$ .  $\square$

We now assert, without proof, existence of the product probability  $P$ .

**Theorem 3.43.** Given probability spaces  $(\Omega_1, \mathcal{F}_1, P_1)$  and  $(\Omega_2, \mathcal{F}_2, P_2)$ , there is a unique probability  $P$  on  $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2)$  such that

$$P(A_1 \times A_2) = P_1(A_1)P_2(A_2) \quad (3.20)$$

for all  $A_1 \in \mathcal{F}_1$  and  $A_2 \in \mathcal{F}_2$ .  $\square$

The probability  $P$  satisfying (3.20) is the *product* of  $P_1$  and  $P_2$ , and denoted by  $P_1 \times P_2$ . The probability space  $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2, P_1 \times P_2)$  is the *product* of  $(\Omega_1, \mathcal{F}_1, P_1)$  and  $(\Omega_2, \mathcal{F}_2, P_2)$ .

Products of probability spaces provide another vehicle for constructing independent random variables with prescribed distributions.

**Theorem 3.44.** Let  $F_1, \dots, F_n$  be distribution functions on  $\mathbb{R}$ . Then, there exist a probability space  $(\Omega, \mathcal{F}, P)$ , and random variables  $X_1, \dots, X_n$  defined on it, that are independent with  $\mathsf{F}_{X_i} = F_i$  for each  $i$ .

**Proof:** By the construction in Proposition 2.44, for each  $i$  there exists a probability space  $(\Omega_i, \mathcal{F}_i, P_i) = (\mathbb{R}, \mathcal{B}(\mathbb{R}), P_i)$ , on which the coordinate random variable  $\tilde{X}_i$  has distribution function  $F_i$ . We then take

$$(\Omega, \mathcal{F}, P) = \prod_{i=1}^n (\Omega_i, \mathcal{F}_i, P_i) = (\mathbb{R}^n, \mathcal{B}(\mathbb{R}^n), \prod_{i=1}^n P_i),$$

so that  $X_i(\omega_1, \dots, \omega_n) = \tilde{X}_i(\omega_i) = \omega_i$ . For  $t_1, \dots, t_n \in \mathbb{R}$ ,

$$\begin{aligned} P\{X_1 \leq t_1, \dots, X_n \leq t_n\} &= P((-\infty, t_1] \times \dots \times (-\infty, t_n]) \\ &= \prod_{i=1}^n P_i((-\infty, t_i]) \\ &= \prod_{i=1}^n P\{X_i \leq t_i\}. \quad \blacksquare \end{aligned}$$

## 3.8 Exercises

- 3.1. Let  $A_1, \dots, A_n$  be independent events. Prove that  $P(\bigcup_{i=1}^n A_i) = 1 - \prod_{i=1}^n [1 - P(A_i)]$ .
- 3.2. Show that an event whose probability is either zero or one is independent of every event, and that an event that is independent of itself has probability zero or one.
- 3.3. What is the minimum number of points a sample space must contain in order that there exist  $n$  independent events  $A_1, \dots, A_n$ , none of which has probability zero or one?

- 3.4.** Prove that  $A$  and  $B$  are independent if and only if  $P(B|A) = P(B)$ .
- 3.5.** Let  $P$  be the uniform distribution on  $[0, 1]$  and let  $X$  be the random variable given by  $X(\omega) = \omega$ .
- Prove or refute by counterexample: there exists no bounded random variable that is both independent of  $X$  and not constant (almost surely).
  - Let  $Y = X(1 - X)$ . Explicitly construct a random variable  $Z$  such that  $Z$  and  $Y$  are independent.
- 3.6.** Let  $X$  and  $Y$  be independent and geometrically distributed with parameter  $p$ . Prove that  $U = \min\{X, Y\}$  and  $V = X - Y$  are independent. (This property characterizes geometric distributions.)
- 3.7.** Let  $X \stackrel{d}{=} E(\lambda)$  and  $Y \stackrel{d}{=} E(\mu)$  be independent. Calculate the distribution of  $Z = \min\{X, Y\}$ .
- 3.8.** Let  $(X, Y)$  have the bivariate normal density function (2.7).
- Prove that  $X \stackrel{d}{=} Y \stackrel{d}{=} N(0, 1)$ .
  - Prove that  $X$  and  $Y$  are independent if and only if  $\rho = 0$ .
  - Calculate  $P\{X \geq 0, Y \geq 0\}$  as a function of  $\rho$ .
- 3.9.** Let  $X$  and  $Y$  be i.i.d. with continuous distribution function  $F$ . Prove that  $P\{X = Y\} = 0$  and that  $P\{X < Y\} = 1/2$ .
- 3.10.**
  - Suppose that  $P\{Y \leq t\}$  is zero or one for every  $t \in \mathbb{R}$ . Prove that there is a constant  $c$  such that  $P\{Y = c\} = 1$ , and conclude from this that  $P\{Y \in B\}$  is zero or one for every Borel set.
  - Let  $X$  be a random variable and let  $g: \mathbb{R} \rightarrow \mathbb{R}$  be a Borel measurable function such that  $X$  and  $Y = g(X)$  are independent. Prove that there is  $c \in \mathbb{R}$  such that  $P\{Y = c\} = 1$ .
- 3.11.** Let  $X$  and  $Y$  be independent and absolutely continuous with densities  $f_X$  and  $f_Y$ .
- Show that  $XY$  is absolutely continuous and calculate  $f_{XY}$ .
  - Assuming that  $P\{Y > 0\} = 1$ , show that  $X/Y$  is absolutely continuous and calculate  $f_{X/Y}$ .
- 3.12.** Let  $X_1, \dots, X_n$  be i.i.d. with density  $f$ .
- Calculate the distribution functions of  $M_n^* = \max\{X_1, \dots, X_n\}$  and  $M_n^{**} = \min\{X_1, \dots, X_n\}$ .
  - Show that  $M_n^*$  and  $M_n^{**}$  are absolutely continuous and calculate the density of each.

- 3.13.** Suppose that  $X \stackrel{d}{=} N(0, 1)$ . Show that  $Y = |X|$  and  $Z = \mathbf{1}(X > 0)$  are independent.
- 3.14.** Let  $X \stackrel{d}{=} B(n, p)$  and  $Y \stackrel{d}{=} B(m, p)$  be independent. Prove that  $X + Y \stackrel{d}{=} B(n + m, p)$ .
- 3.15.** Show that if  $X_1, \dots, X_n$  are independent and  $X_i$  has a negative binomial distribution with parameters  $k_i$  and  $p$ , then  $\sum_{i=1}^n X_i$  has a negative binomial distribution with parameters  $\sum_{i=1}^n k_i$  and  $p$ .
- 3.16.** Consider a group of  $n$  people.
- Calculate the probability that no two people have the same birthday. State carefully the assumptions you make regarding the probabilities and independence of various events.
  - Determine the minimum value of  $n$  for which the probability in a) is less than  $1/2$ .
- 3.17.** Let  $X_1, \dots, X_n$  be independent with distribution  $N(0, 1)$ . Prove that  $S = \sum_{i=1}^n X_i^2$  has a  $\chi^2$  distribution with  $n$  degrees of freedom:
- $$f_S(t) = \frac{1}{\Gamma(n/2)} \frac{1}{\sqrt{2^n}} e^{-t/2} t^{n/2-1}, \quad t > 0.$$
- 3.18.** Let  $R$  have the *Rayleigh density*
- $$f_R(r) = \frac{r}{\sigma^2} e^{-r^2/2\sigma^2}, \quad r > 0,$$
- where  $\sigma^2 > 0$ , and let  $\Theta$  be uniformly distributed on  $[-\pi, \pi]$ .
- Prove that  $X = R \cos \Theta$  and  $Y = R \sin \Theta$  are independent and each has distribution  $N(0, \sigma^2)$ .
  - Describe how to use this construction to simulate (pairs of) normally distributed random variables.
- 3.19.** Let  $X$  and  $Y$  be independent with distribution  $N(0, 1)$ . Show that  $(X + Y)/\sqrt{2} \stackrel{d}{=} N(0, 1)$  and that  $X/Y$  has a Cauchy distribution.
- 3.20.** Let  $X$  and  $Y$  be independent and Cauchy distributed. Show that  $(X + Y)/2$  also has a Cauchy distribution.
- 3.21.** Let  $(S_n)$  be a Bernoulli process with  $p = 1/2$ . Prove that the process  $X_n = 2S_n - n$  is a random walk.
- 3.22.** Let  $(X_n)$  be a Bernoulli process with parameter  $p \in (0, 1)$ , let  $q = 1 - p$ , and let  $(S_n)$  be the success counting process.

- a) For each  $n$ , let

$$W_n = \min \{k > n: X_k = 1\} - n$$

be the time from  $n$  until the next success after  $n$ , known as the *forward recurrence time* at  $n$ . Prove that  $W_n$  has a geometric distribution with parameter  $p$ .

- b) For each  $n$ , calculate the distribution of

$$V_n = \begin{cases} n - \max \{k < n: X_k = 1\} & \text{if } S_{n-1} > 0 \\ n & \text{if } S_{n-1} = 0 \end{cases},$$

the time elapsed at  $n$  since the last success before  $n$ , or *backward recurrence time* at  $n$ .

- c) Show that for each  $k$ ,  $\lim_{n \rightarrow \infty} P\{V_n = k\} = pq^{k-1}$ .  
d) Prove that  $V_n$  and  $W_n$  are independent for each  $n$ .

**3.23.** Let  $X_1, X_2, \dots$  be i.i.d. with continuous distribution function  $F$ .

- a) Show that for each permutation  $\pi$  of  $\{1, \dots, n\}$ ,

$$(X_1, \dots, X_n) \stackrel{\text{d}}{=} (X_{\pi 1}, \dots, X_{\pi n}).$$

- b) Show that  $P(A) = 0$ , where

$$A = \{\omega: X_n(\omega) = X_m(\omega) \text{ for some } n \neq m\}.$$

- c) Assume that the event  $A$  in part b) has been removed, so that the  $X_i$  are always distinct. For each  $n$ , let  $Y_n$  be the rank of  $X_n$  among  $X_1, \dots, X_n$ :  $Y_n = r$  if and only if exactly  $r - 1$  of the values  $X_1, \dots, X_{n-1}$  are less than  $X_n$ . Prove that

$$P\{Y_n = r\} = 1/n, \quad r = 1, \dots, n.$$

- d) Prove that the ranks  $Y_1, Y_2, \dots$  are independent.

**3.24.** Let  $X$  and  $Y$  be independent with distribution  $U[0, 1]$ . Calculate the density of  $X - Y$ .

**3.25.** Let  $X \stackrel{\text{d}}{=} P(\lambda)$  and  $Y \stackrel{\text{d}}{=} P(\mu)$  be independent. Show that the conditional distribution of  $X$  given  $X + Y$  is binomial: for  $k \leq n$ ,

$$P\{X = k | X + Y = n\} = \binom{n}{k} \left( \frac{\lambda}{\lambda + \mu} \right)^k \left( \frac{\mu}{\lambda + \mu} \right)^{n-k}.$$

**3.26.** Let  $X$  and  $Y$  be independent and geometrically distributed with parameter  $p$ , and let  $Z = \max\{X, Y\}$ .

- a) Calculate  $P\{Z = k\}$  for each  $k$ .
- b) For each  $j$  and  $k$ , calculate  $P\{X = j, Z = k\}$ ,  $P\{X = j|Z = k\}$  and  $P\{Z = k|X = j\}$ .

**3.27.** Suppose that  $f$  and  $g$  are density functions on  $\mathbb{R}$  for which there exists a constant  $c$  such that  $f(x)/g(x) \leq c$  for all  $x$ . Let  $Y_1, Y_2, \dots$  be i.i.d. with density  $g$ , and let  $U_1, U_2, \dots$  be independent with distribution  $U[0, 1]$  and independent of the  $Y_i$ .

- a) Let

$$N = \min \{i : U_i \leq f(Y_i)/cg(Y_i)\}.$$

Prove that  $X = Y_N$  has density  $f$ .

- b) Discuss how the property in a) could be applied to simulate random variables with density  $f$  if random variables with density  $g$  can be simulated easily.
- c) Use b) to simulate normally distributed random variables from random variables with distribution  $U(0, 1)$ . [Hint: Let  $g$  be the exponential density with parameter 1, and let

$$f(x) = \frac{2}{\sqrt{2\pi}} e^{-x^2/2}, \quad x \geq 0, \quad (3.21)$$

be the density of the absolute value of  $|X|$ , where  $X$  is the desired normally distributed random variable. Let  $\tilde{U}_1, \tilde{U}_2, \dots$  be uniformly distributed and independent of the  $U_i$ . Show that the random variables  $Y_i = -\log U_i$  are i.i.d. with density  $g$ . Use b) to simulate a random variable  $Z$  with the density  $f$  of (3.21). Finally, show that if  $W$  is independent of  $Z$  with  $P\{X = \pm 1\} = 1/2$ , then  $X = WZ$  is normally distributed.]

**3.28.** Devise a method for simulating random variables with a negative binomial distribution.

**3.29.** Prove that if  $X$  and  $Y$  are independent and either  $X$  or  $Y$  is absolutely continuous, then  $X + Y$  is absolutely continuous.

**3.30.** Let  $X_1, X_2, Y_1, Y_2$  be independent, and suppose that the  $X_i$  have distribution  $E(1)$  and the  $Y_i$  distribution  $E(1/\theta)$ , where  $\theta \in (0, \infty)$  is unknown. Let  $P_\theta$  denote probability under the stipulation that the value of the parameter is  $\theta$ . Let  $R_j$  be the rank of  $Y_j$  among the combined sample  $\{X_1, X_2, Y_1, Y_2\}$ , and let  $T = R_1 + R_2$ , which is known as *Wilcoxon's rank statistic*. Calculate

$$P_\theta\{T = k\}, \quad k = 3, \dots, 7,$$

first for  $\theta = 1$  and then for general  $\theta$ .

- 3.31.** Two points are chosen independently and at random in  $[0, 1]$ , and then a third is chosen at random between the first two. Let  $M$  denote the location of this middle point. Show that  $M$  is absolutely continuous and calculate its density.
- 3.32.** Let  $X$  and  $Y$  be independent and gamma distributed with parameters  $(\alpha, \lambda)$  and  $(\alpha + 1/2, \lambda)$ . Prove that  $2\sqrt{XY}$  has a gamma distribution with parameters  $(2\alpha, \lambda)$ .
- 3.33.** Two people independently toss a coin  $n$  times. Show that the probability that they obtain equal numbers of heads is the same as the probability that together they obtain  $n$  heads.
- 3.34.** Consider a machine that operates without failure for a random time  $X$ , and is then repaired, which requires a random time  $Y$ . Assume that  $X$  and  $Y$  are independent and exponentially distributed with parameters  $\lambda$  and  $\mu$ . Calculate the density of  $Z = X/(X + Y)$ , the proportion of time that the machine is in operation. [Hint: Calculate the joint density of  $X$  and  $X + Y$ .]
- 3.35.** Let  $X_1, \dots, X_n$  be independent with distribution  $E(1)$ . Let  $Y_k = \sum_{i=1}^k X_i$ , and define
- $$Z_i = \begin{cases} Y_i/Y_{i+1} & i = 1, \dots, n-1 \\ Y_n & i = n. \end{cases}$$

Prove that  $Z_1, \dots, Z_n$  are independent, and calculate their densities.

- 3.36.** Show that if  $X$  and  $Y$  are independent with distribution  $N(0, 1)$ , then  $X + Y$  and  $X - Y$  are independent.
- 3.37.** In the context of the occupation models of §5, let  $A$  be the event that no cell is empty. Show that

- a) For the Maxwell-Boltzmann model,

$$P(A) = \sum_{\ell=0}^n (-1)^\ell \binom{n}{\ell} \left(1 - \frac{\ell}{n}\right)^k.$$

- b) For the Bose-Einstein model,

$$P(A) = \binom{k-1}{n-1} / \binom{n+k-1}{k}.$$

# Chapter 4

## Expectation

Expectation is an averaging process for random variables. The expectation of a random variable is a weighted sum (more generally, integral) of its values, where the weight of each value is the probability of the set on which it is assumed. Although it will not be until §3 that we establish this, it is intuitively clear that if  $X$  is discrete with values in the countable set  $C$ , this weighted average is given by

$$E[X] = \sum_{t \in C} t P\{X = t\}. \quad (4.1)$$

Similarly, if  $X$  is absolutely continuous with density  $f_X$ , then by (2.2), the weighted average of its values is

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx. \quad (4.2)$$

In fact, neither (4.1) nor (4.2) is the definition of  $E[X]$ ; rather, these are key computational formulas. Also, convergence of sums and integrals is not guaranteed. When  $X \geq 0$ , the rules of arithmetic for the extended real numbers can be used, and it is permissible that  $E[X] = \infty$ , but finiteness is mandatory when  $X$  can assume both positive and negative values.

We may also approach expectation via a set of properties of averaging processes *qua* averaging processes:

1. **Constants preserved.** If  $X \equiv c$ , then  $E[X] = c$ .
2. **Monotonicity.** If  $X \leq Y$ , then  $E[X] \leq E[Y]$ .
3. **Linearity.** For  $a, b \in \mathbb{R}$ ,  $E[aX + bY] = aE[X] + bE[Y]$ .
4. **Continuity.** If  $X_n \rightarrow X$ , then  $E[X_n] \rightarrow E[X]$ .
5. **Relation to the probability.** For each event  $A$ ,  $E[\mathbf{1}_A] = P(A)$ .

With the exception of continuity, which also entails defining the sense in which  $X_n \rightarrow X$ , these will be realized fully. Continuity is not valid without restriction, but does hold true under two broadly applicable hypotheses, as shown in the monotone convergence theorem (Theorem 4.9) and the dominated convergence theorem (Theorem 4.16).

There is one point always to keep in mind, which motivates and explains many of the results and most of the ideas in the chapter: *expectation is to random variables as probability is to events*, so that properties of expectation extend those of probability.

## 4.1 Definition and Fundamental Properties

Let  $(\Omega, \mathcal{F}, P)$  be a probability space.

### 4.1.1 Simple random variables

Simple random variables are the “elementary functions,” whose expectations are defined explicitly. Recall that  $X$  is simple if it assumes only finitely many values, in which case  $X = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$ , where the  $a_i$  are real but not necessarily distinct, and the events  $A_i$  constitute partition of  $\Omega$ .

**Definition 4.1.** The *expectation* of  $X = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$  is

$$E[X] = \sum_{i=1}^n a_i P(A_i). \quad \square \quad (4.3)$$

Before proceeding, we observe that  $E[X]$  is well-defined, in the sense that all representations of  $X$  yield the same value for  $E[X]$ . That is, if

$$\sum_{i=1}^n a_i \mathbf{1}_{A_i} = \sum_{j=1}^m a'_j \mathbf{1}_{A'_j}$$

as functions on  $\Omega$ , where  $\{A_1, \dots, A_n\}$  and  $\{A'_1, \dots, A'_m\}$  are partitions of  $\Omega$ , then

$$\sum_{i=1}^n a_i P(A_i) = \sum_{j=1}^m a'_j P(A'_j).$$

From (4.3),

$$E[\mathbf{1}_A] = 1 \cdot P(A) + 0 \cdot P(A^c) = P(A),$$

so that the expectation of an indicator function is indeed the probability of the associated event. That  $E[c] = c$  is apparent.

Next, we verify linearity and monotonicity.

**Proposition 4.2.** If  $X, Y$  are simple and  $a, b \in \mathbb{R}$ , then  $aX + bY$  is simple and  $E[aX + bY] = aE[X] + bE[Y]$ .

**Proof:** If  $X = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$  and  $Y = \sum_{j=1}^m b_j \mathbf{1}_{B_j}$ , then

$$aX + bY = \sum_{i=1}^n \sum_{j=1}^m (a \cdot a_i + b \cdot b_j) \mathbf{1}_{A_i \cap B_j},$$

so that  $aX + bY$  is simple. By (4.3),

$$\begin{aligned} E[aX + bY] &= \sum_{i=1}^n \sum_{j=1}^m (a \cdot a_i + b \cdot b_j) P(A_i \cap B_j) \\ &= \sum_{i=1}^n (a \cdot a_i) \sum_{j=1}^m P(A_i \cap B_j) + \sum_{j=1}^m (b \cdot b_j) \sum_{i=1}^n P(A_i \cap B_j) \\ &= a \sum_{i=1}^n a_i P(A_i) + b \sum_{j=1}^m b_j P(B_j) \\ &= aE[X] + bE[Y]. \quad \blacksquare \end{aligned}$$

Monotonicity of expectation for simple random variables follows from linearity and the seemingly weaker property of *positivity*, i.e., that if  $X \geq 0$ , then  $E[X] \geq 0$ . Indeed, if  $X \leq Y$ , then

$$E[Y] - E[X] = E[Y - X] \geq 0.$$

This argument is valid provided only that  $E[Y] - E[X]$  is not of the form  $\infty - \infty$ , so that monotonicity holds whenever linearity and positivity do.

Even for simple random variables, continuity of expectation fails.

**Example 4.3.** Let  $P$  be the uniform distribution on  $[0, 1]$  and let

$$X_n = n \mathbf{1}_{(0, 1/n)}.$$

Then,  $X_n(\omega) \rightarrow 0$  for every  $\omega$ , but  $E[X_n] = 1$  for each  $n$ .  $\square$

We eventually establish two restricted forms of continuity: monotone continuity for increasing sequences of positive random variables and “dominated” continuity for integrable random variables.

### 4.1.2 Positive random variables

We now extend the definition of expectation to *all* positive random variables. The definition virtually forces monotone continuity: if  $X, X_1, X_2, \dots$  are positive and  $X_n \uparrow X$  (pointwise), then  $E[X_n] \uparrow E[X]$ . There is one complication, however: some expectations are infinite.

Recall from Theorem 2.20 that for every  $X \geq 0$  there are simple random variables  $0 \leq X_1 \leq X_2 \leq \dots$  such that  $X_n(\omega) \uparrow X(\omega)$  for each  $\omega$ . We simply define  $E[X]$  as the limit of the  $E[X_n]$ .

**Definition 4.4.** Let  $X$  be positive.

- a) The *expectation* of  $X$  is  $E[X] = \lim_{n \rightarrow \infty} E[X_n] \leq \infty$ , where the  $X_n$  are simple and positive with  $X_n \uparrow X$ .
- b) The *expectation of  $X$  over the event  $A$*  is  $E[X; A] \stackrel{\text{def}}{=} E[X\mathbf{1}_A]$ .  $\square$

Since  $0 \leq E[X_1] \leq E[X_2] \leq \dots$ , the limit defining  $E[X]$  exists in  $\overline{\mathbb{R}}_+$ , but may be infinite. We must show, though, that it does not depend on the approximating sequence  $(X_n)$ .

**Proposition 4.5.** If  $(X_n)$  and  $(\tilde{X}_k)$  are sequences of simple random variables increasing to  $X$ , then  $\lim_{n \rightarrow \infty} E[X_n] = \lim_{k \rightarrow \infty} E[\tilde{X}_k]$ .

**Proof:** We first show that if  $Y$  is simple and  $Y \leq X$ , then

$$E[Y] \leq \lim_{n \rightarrow \infty} E[X_n]. \quad (4.4)$$

Fix  $\varepsilon > 0$ , and for each  $n$ , define the event  $A_n = \{X_n > Y - \varepsilon\}$ . Since  $X_n \uparrow X \geq Y$ , the  $A_n$  increase to  $\Omega$ . For each  $n$ ,  $X_n \geq (Y - \varepsilon)\mathbf{1}_{A_n}$ , so that

$$\begin{aligned} E[X_n] &\geq E[(Y - \varepsilon)\mathbf{1}_{A_n}] = E[Y\mathbf{1}_{A_n}] - \varepsilon P(A_n) \\ &\geq E[Y] - E[Y\mathbf{1}_{A_n^c}] - \varepsilon \\ &\geq E[Y] - P(A_n^c) \max_{\omega \in \Omega} Y(\omega) - \varepsilon \end{aligned}$$

[ $Y$  is simple, so the “max” is truly a maximum, and is finite]

$$\rightarrow E[Y] - \varepsilon,$$

where the convergence holds because  $P(A_n^c) \rightarrow 0$  by continuity of  $P$ . But  $\varepsilon$  is arbitrary, and, hence, (4.4) holds.

For each  $k$ , (4.4) holds for  $Y = \tilde{X}_k$ , so that  $\lim_k E[\tilde{X}_k] \leq \lim_n E[X_n]$ , from which the conclusion follows by symmetry. ■

Proposition 4.5 also implies that if  $X$  is simple, then Definition 4.1 and Definition 4.4 yield the same value for  $E[X]$ , since the choice  $X_n = X$  for each  $n$  is permissible.

We now establish the fundamental properties, beginning with linearity.

**Proposition 4.6.** For  $X, Y \geq 0$  and  $a, b \in \mathbb{R}_+$ ,  $E[aX + bY] = aE[X] + bE[Y]$ .

**Proof:** Given simple random variables  $X_n$  and  $Y_n$  with  $X_n \uparrow X$  and  $Y_n \uparrow Y$ , the random variables  $aX_n + bY_n$  are likewise simple, and they increase to  $aX + bY$ . Therefore,

$$\begin{aligned} E[aX + bY] &= \lim_{n \rightarrow \infty} E[aX_n + bY_n] \\ &= \lim_{n \rightarrow \infty} (aE[X_n] + bE[Y_n]) \\ &= a \lim_{n \rightarrow \infty} E[X_n] + b \lim_{n \rightarrow \infty} E[Y_n] \\ &= aE[X] + bE[Y], \end{aligned}$$

where the second equality is by Proposition 4.2. ■

Of course,  $E[X] \geq 0$  for  $X \geq 0$  by construction, so positivity holds. Since linearity is valid as well, monotonicity follows.

**Corollary 4.7.** If  $0 \leq Y \leq X$ , then  $E[Y] \leq E[X]$ . □

The key to monotone continuity is the following result, which is famous and important in its own right.

**Theorem 4.8 (Fatou's lemma).** For  $X_n \geq 0$ ,

$$E[\liminf_{n \rightarrow \infty} X_n] \leq \liminf_{n \rightarrow \infty} E[X_n]. \quad (4.5)$$

**Proof:** For each  $m$ , let  $Z_m = \inf_{k \geq m} X_k$ , so that  $Z_m \uparrow \liminf_n X_n$  as  $m \rightarrow \infty$ . Given a simple random variable  $Y \leq \liminf_n X_n$  and  $\varepsilon > 0$ , then for each  $m$ ,

$$X_m \geq Z_m \geq (Y - \varepsilon) \mathbf{1}(Z_m \geq Y - \varepsilon).$$

By the argument used to prove Proposition 4.5,

$$\liminf_{m \rightarrow \infty} E[X_m] \geq \lim_{m \rightarrow \infty} E[(Y - \varepsilon); \{Z_m \geq Y - \varepsilon\}] \geq E[Y] - \varepsilon.$$

Since  $\varepsilon$  is arbitrary,

$$\liminf_m E[X_m] \geq E[Y].$$

In particular, this is true for simple random variables  $Y_k$  increasing to  $\liminf_n X_n$ , which exist by Theorem 2.20. Hence, by Definition 4.1,

$$E[\liminf_{n \rightarrow \infty} X_n] = \lim_{k \rightarrow \infty} E[Y_k] \leq \liminf_{m \rightarrow \infty} E[X_m]. \quad \blacksquare$$

The inequality in (4.5) can be strict: in Example 4.3,  $E[\liminf_n X_n] = 0$  but  $\liminf_n E[X_n] = 1$ .

The next theorem gives the first of two primary forms of continuity for expectation.

**Theorem 4.9 (Monotone convergence theorem).** Let  $X, X_1, X_2, \dots$  be positive with  $X_n(\omega) \uparrow X(\omega)$  for each  $\omega$ . Then,  $E[X_n] \uparrow E[X]$ .

**Proof:** We show that

$$E[X] \leq \liminf_n E[X_n] \leq \limsup_n E[X_n] \leq E[X],$$

which implies that  $\lim_n E[X_n] = E[X]$ . The middle inequality is, of course, always valid.

Since  $X_n \leq X$ ,  $E[X_n] \leq E[X]$  for each  $n$ , and, hence,  $\limsup_n E[X_n] \leq E[X]$ . Finally, Fatou's lemma implies that  $E[X] \leq \liminf_n E[X_n]$ . ■

Example 4.3 does not violate the monotone convergence theorem because in that instance it is not true that  $X_n \uparrow X$ .

Among implications of the monotone convergence theorem is linearity for convergent series of positive random variables.

**Theorem 4.10.** If  $Y_k \geq 0$  and  $\sum_{k=1}^{\infty} Y_k(\omega) < \infty$  for every  $\omega$ , then

$$E\left[\sum_{k=1}^{\infty} Y_k\right] = \sum_{k=1}^{\infty} E[Y_k]. \quad (4.6)$$

**Proof:** All that is needed is to apply the monotone convergence theorem to  $X_n = \sum_{k=1}^n Y_k$  and  $X = \sum_{k=1}^{\infty} Y_k$ . ■

It is illuminating to look at these three theorems when the random variables are indicator functions: one can see more explicitly how properties of expectation generalize those of probability. If  $X_n = \mathbf{1}_{A_n}$ , then since  $\liminf_n \mathbf{1}_{A_n} = \mathbf{1}_{\liminf_n A_n}$ , (4.5) becomes

$$\begin{aligned} P\left(\liminf_n A_n\right) &= E\left[\liminf_{n \rightarrow \infty} \mathbf{1}_{A_n}\right] \leq \liminf_{n \rightarrow \infty} E[\mathbf{1}_{A_n}] \\ &= \liminf_{n \rightarrow \infty} P(A_n), \end{aligned}$$

the expression, (1.15), used in Theorem 1.25 to prove continuity of  $P$ .

With  $X_n = \mathbf{1}_{A_n}$  and  $X = \mathbf{1}_A$ , where  $A_n \uparrow A$ , Theorem 4.9 yields

$$P(A_n) = E[\mathbf{1}_{A_n}] \uparrow E[\mathbf{1}_A] = P(A),$$

which is the monotone continuity of  $P$ .

Finally, suppose that in Theorem 4.10,  $Y_k = \mathbf{1}_{A_k}$ , where the  $A_k$  are disjoint, so that  $\sum_{k=1}^{\infty} Y_k = \mathbf{1}_{\sum_{k=1}^{\infty} A_k}$ . Then, (4.6) gives countable additivity of  $P$ :

$$P\left(\sum_{k=1}^{\infty} A_k\right) = E\left[\sum_{k=1}^{\infty} \mathbf{1}_{A_k}\right] = \sum_{k=1}^{\infty} E[\mathbf{1}_{A_k}] = \sum_{k=1}^{\infty} P(A_k).$$

Before moving on to integrable random variables, we derive a useful “converse” to the positivity of expectation.

**Proposition 4.11.** If  $X \geq 0$  and  $E[X] = 0$ , then  $X \stackrel{\text{a.s.}}{=} 0$ .

**Proof:** If  $P\{X > 0\} > 0$ , then by right-continuity of distribution functions there is  $\varepsilon > 0$  such that  $P\{X > \varepsilon\} > 0$ . But then  $X \geq \varepsilon \mathbf{1}(X > \varepsilon)$ , which implies that  $E[X] \geq \varepsilon P\{X > \varepsilon\} > 0$ . ■

### 4.1.3 Integrable random variables

Finally, we extend the definition of expectation to (some) random variables taking both positive and negative values. Recall from Definition 2.14 that the positive part of  $X$  is  $X^+ = \max\{X, 0\}$ , that the negative part of  $X$  is  $X^- = -\min\{X, 0\}$ , and that

$$X = X^+ - X^-, \quad (4.7)$$

$$|X| = X^+ + X^-. \quad (4.8)$$

Our extension preserves linearity by basing the definition on (4.7).

**Definition 4.12.** Let  $X$  be a random variable, not necessarily positive.

- a)  $X$  is *integrable* if  $E[|X|] < \infty$ .
- b) If  $X$  is integrable, the *expectation* of  $X$  is

$$E[X] = E[X^+] - E[X^-].$$

- c) For  $X$  integrable and  $A$  an event, the *expectation of  $X$  over  $A$*  is  $E[X; A] \stackrel{\text{def}}{=} E[X \mathbf{1}_A]$ . □

If  $X$  is integrable, then by (4.8),

$$E[X^+] + E[X^-] = E[|X|] < \infty,$$

so that both  $E[X^+]$  and  $E[X^-]$  are finite, and part b) of Definition 4.12 makes sense. (That is,  $E[X^+] - E[X^-]$  is not of the undefined form  $\infty - \infty$ .) Since  $|X \mathbf{1}_A| \leq |X|$ ,  $E[X; A]$  exists when  $E[X]$  does.

This definition coincides with earlier ones if  $X$  is simple, in which case integrability is automatic, or positive. For reasons that are clarified in §4, we denote by  $L^1$  the set of integrable random variables.

The important remaining results are linearity and a form of continuity known as the dominated convergence theorem.

**Theorem 4.13.** If  $X, Y$  belong to  $L^1$  and  $a, b \in \mathbb{R}$ , then  $aX + bY \in L^1$  and

$$E[aX + bY] = aE[X] + bE[Y].$$

**Proof:** That  $aX + bY$  belongs to  $L^1$  follows from the triangle inequality:  $|aX + bY| \leq |a| |X| + |b| |Y|$ , so that

$$E[|aX + bY|] \leq |a| E[|X|] + |b| E[|Y|].$$

We next prove that if  $Z_1$  and  $Z_2$  are positive and integrable, then

$$E[Z_1 - Z_2] = E[Z_1] - E[Z_2]. \quad (4.9)$$

Both  $Z_1 + (Z_1 - Z_2)^-$  and  $(Z_1 - Z_2)^+ + Z_2$  are equal and positive, so by linearity for positive random variables (Proposition 4.6),

$$E[Z_1] + E[(Z_1 - Z_2)^-] = E[(Z_1 - Z_2)^+] + E[Z_2].$$

By assumption and the first part of the theorem, all four terms in this expression are finite, and, hence, they may be re-arranged to give (4.9). But

$$X + Y = (X^+ + Y^+) - (X^- + Y^-), \quad (4.10)$$

and so

$$E[X + Y] = E[X] + E[Y]$$

follows from (4.9).

That  $E[aX] = aE[X]$  for  $a > 0$  is immediate, while for  $a < 0$ ,

$$\begin{aligned} E[aX] &\stackrel{\text{def}}{=} E[(aX)^+] - E[(aX)^-] = E[(-a)X^-] - E[(-a)X^+] \\ &= aE[X]. \blacksquare \end{aligned}$$

In particular,  $L^1$  is a vector space. Note that (4.10) is *not* the decomposition  $X + Y = (X + Y)^+ - (X + Y)^-$ , which is why (4.9) is needed.

We next note two important consequences of Theorem 4.13.

**Corollary 4.14.** For  $X \in L^1$ ,  $|E[X]| \leq E[|X|]$ .

**Proof:** By (4.8),

$$|E[X]| = |E[X^+] - E[X^-]| \leq E[X^+] + E[X^-] = E[|X|]. \blacksquare$$

**Corollary 4.15.** If  $X \leq Y \in L^1$ , then  $E[X] \leq E[Y]$ .  $\square$

We come now to the final form of continuity.

**Theorem 4.16 (Dominated convergence theorem).** Let  $X_1, X_2, \dots$  and  $X$  be integrable with  $X_n(\omega) \rightarrow X(\omega)$  for each  $\omega$ , and suppose that there is  $Y \in L^1$  such that  $|X_n| \leq Y$  for each  $n$ . Then,

$$\lim_{n \rightarrow \infty} E[X_n] = E[X].$$

**Proof:** The domination hypothesis implies that  $Y - X_n \geq 0$  for each  $n$ , and hence, by Theorem 4.8 and Theorem 4.13,

$$\begin{aligned} E[Y] - E[X] &= E[Y - X] = E[\liminf_{n \rightarrow \infty} (Y - X_n)] \\ &\leq \liminf_{n \rightarrow \infty} E[Y - X_n] \\ &= E[Y] + \liminf_{n \rightarrow \infty} (-E[X_n]) \\ &= E[Y] - \limsup_{n \rightarrow \infty} E[X_n]. \end{aligned}$$

Since  $Y \in L^1$ ,  $E[Y]$  may be subtracted, giving  $\limsup_n E[X_n] \leq E[X]$ .

In the same way, since also  $Y + X_n \geq 0$  for each  $n$ ,

$$\begin{aligned} E[Y] + E[X] &= E[Y + X] = E[\liminf_{n \rightarrow \infty} (Y + X_n)] \\ &\leq \liminf_{n \rightarrow \infty} E[Y + X_n] \\ &= E[Y] + \liminf_{n \rightarrow \infty} E[X_n], \end{aligned}$$

so that  $E[X] \leq \liminf_n E[X_n]$ .

Consequently,

$$E[X] \leq \liminf_{n \rightarrow \infty} E[X_n] \leq \limsup_{n \rightarrow \infty} E[X_n] \leq E[X],$$

which implies that  $\lim_{n \rightarrow \infty} E[X_n]$  exists and equals  $E[X]$ . ■

Example 4.3 does not violate the dominated convergence theorem. Any random variable  $Y$  dominating  $X_n = n\mathbf{1}_{(0,1/n)}$  for each  $n$  must satisfy  $Y \geq \sum_{n=1}^{\infty} n\mathbf{1}_{(1/(n+1),1/n)}$ , which implies that  $E[Y] = \infty$ .

Theorem 4.16 applies to indicator functions of events  $A_n \rightarrow A$ , which satisfy  $\mathbf{1}_{A_n} \rightarrow \mathbf{1}_A$ . With  $Y \equiv 1$  as the dominating function, we conclude that  $P(A_n) \rightarrow P(A)$ , the full-fledged continuity of  $P$ .

#### 4.1.4 Complex-valued random variables

Recall from Definition 2.8 that a complex-valued random variable has the form  $Z = X + iY$ , where  $X$  and  $Y$  are ordinary random variables. It is inescapable that one should define  $E[Z]$  as the complex number  $E[X] + iE[Y]$ . In order that this make sense,  $X$  and  $Y$  are must be integrable, but we impose the stronger condition that  $|Z| = \sqrt{X^2 + Y^2}$  be integrable.

**Definition 4.17.** A complex-valued random variable  $Z = X + iY$  is *integrable* if  $E[|Z|] = E[\sqrt{X^2 + Y^2}] < \infty$ , and in this case the *expectation* of  $Z$  is  $E[Z] = E[X] + iE[Y]$ . □

Among properties that remain valid are linearity, the dominated convergence theorem and key inequalities. In particular,  $|E[Z]| \leq E[|Z|]$ .

## 4.2 Integrals with respect to Distribution Functions

In this section, we define integrals with respect to distribution functions on  $\mathbb{R}$  and establish their properties. These integrals, known as *Lebesgue-Stieltjes integrals*, can be defined from scratch; however, they are really expectations with respect to probabilities on  $\mathbb{R}$ . Our principal application — to computation of expectations — appears in §3.

All functions appearing below are presumed to be Borel measurable.

### 4.2.1 Generalities

Recall from Theorem 1.36 that given a distribution function  $F$  on  $\mathbb{R}$ , there is a unique probability  $P_F$  on  $\mathbb{R}$  such that  $P_F((a, b]) = F(b) - F(a)$  for every interval  $(a, b]$ .

**Definition 4.18.** Let  $F$  be a distribution function on  $\mathbb{R}$ .

- a) For  $g$  a positive function, the *integral of  $g$  with respect to  $F$*  is

$$\int_{\mathbb{R}} g(x) dF(x) = E_F[g] \leq \infty,$$

where the expectation is that of  $g$  as a random variable on the probability space  $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_F)$ .

- b) A function  $g$  is *integrable with respect to  $F$*  if  $\int_{\mathbb{R}} |g(x)| dF(x) < \infty$ , and in this case, the *integral of  $g$  with respect to  $F$*  is

$$\int_{\mathbb{R}} g(x) dF(x) = \int_{\mathbb{R}} g^+(x) dF(x) - \int_{\mathbb{R}} g^-(x) dF(x).$$

- c) For  $g$  either positive or integrable and  $B \in \mathcal{B}(\mathbb{R})$ , the *integral of  $g$  over  $B$  with respect to  $F$*  is  $\int_B g(x) dF(x) = \int_{\mathbb{R}} \mathbf{1}_B(x)g(x) dF(x)$ .  $\square$

The properties of this integral are those of expectation, and do not require proof, but for completeness we list them. As is common, we abbreviate  $\int_{\mathbb{R}} g(x) dF(x)$  to  $\int g dF$ .

- a) **Constants preserved.** If  $g \equiv c$ , then  $\int g dF = c$ .
- b) **Relation to  $P_F$ .** For  $B \in \mathcal{B}(\mathbb{R})$ ,  $\int \mathbf{1}_B dF = P_F(B)$ .
- c) **Linearity.** If  $g, h$  are positive and  $a, b \in \mathbb{R}_+$ , or if  $g, h$  are integrable and  $a, b \in \mathbb{R}$ , then  $\int (ag + bh) dF = a \int g dF + b \int h dF$ .
- d) **Monotonicity.** If either  $0 \leq g \leq h$  or  $g$  and  $h$  are integrable and  $g \leq h$ , then  $\int g dF \leq \int h dF$ .

e) **Fatou's lemma.** If  $g_n \geq 0$  for each  $n$ , then

$$\int \liminf_{n \rightarrow \infty} g_n dF \leq \liminf_{n \rightarrow \infty} \int g_n dF.$$

f) **Monotone convergence theorem.** If  $0 \leq g_1 \leq g_2 \leq \dots$ , and if  $g_n(x) \uparrow g(x)$  for each  $x$ , then  $\int g_n dF \uparrow \int g dF$ .

g) **Dominated convergence theorem.** If  $g_n(x) \rightarrow g(x)$  for each  $x$ , and if there is an integrable function  $h$  such that  $|g_n| \leq h$  for each  $n$ , then  $\int g_n dF \rightarrow \int g dF$ .

In the two cases of primary interest — discrete and absolutely continuous distribution functions — these integrals reduce to sums and Riemann (more generally, Lebesgue) integrals.

### 4.2.2 Discrete distribution functions

Integrals with respect to a discrete distribution function are sums.

**Theorem 4.19.** If  $F(t) = \sum p_i \mathbf{1}(t_i \leq t)$ , then for each  $g \geq 0$ ,

$$\int g dF = \sum_i p_i g(t_i). \quad (4.11)$$

**Proof:** Suppose that  $g = \sum_{j=1}^m b_j \mathbf{1}_{B_j}$  is a simple random variable over  $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_F)$ . Then, since  $P_F = \sum_i p_i \varepsilon_{t_i}$ ,

$$\begin{aligned} E_F[g] &= \sum_{j=1}^m b_j P(B_j) = \sum_{j=1}^m b_j \sum_i p_i \mathbf{1}(t_i \in B_j) \\ &= \sum_i p_i \sum_{j=1}^m b_j \mathbf{1}_{B_j}(t_i) \\ &= \sum_i p_i g(t_i). \end{aligned}$$

If  $g \geq 0$  and  $g_n$  are simple with  $g_n \uparrow g$ , then (4.11) holds for each  $n$ , and therefore

$$E_F[g] = \lim_{n \rightarrow \infty} E_F[g_n] = \lim_{n \rightarrow \infty} \sum_i p_i g_n(t_i) = \sum_i p_i g(t_i).$$

Here, the first equality is by the monotone convergence theorem (Theorem 4.9) applied to the sequence  $(g_n)$  on  $(\mathbb{R}, \mathcal{B}(\mathbb{R}), P_F)$ , and the second is by the same theorem applied to the functions  $\tilde{g}_n(i) = g_n(t_i)$  and  $\tilde{g}(i) = g(t_i)$  on the probability space  $(\mathbb{N}, \mathcal{P}(\mathbb{N}), \tilde{P}_F)$ , where  $\tilde{P}_F(\{i\}) = p_i$  for each  $i$ . ■

**Corollary 4.20.** The function  $g$  is integrable with respect to  $F$  if and only if  $\sum_i p_i |g(t_i)| < \infty$ . In this case, (4.11) holds. □

### 4.2.3 Absolutely continuous distribution functions

Integrals with respect to an absolutely continuous distribution function are Riemann integrals.

**Theorem 4.21.** Suppose that  $F$  is absolutely continuous with piecewise continuous density  $f$ . Then, if  $g$  is positive and piecewise continuous,

$$\int g dF = \int_{-\infty}^{\infty} g(x)f(x) dx, \quad (4.12)$$

where the integral on the right-hand side is in the (improper) Riemann sense.

**Proof:** Suppose first that  $g = \sum_{j=1}^m b_j \mathbf{1}_{I_j}$  is a step function, where the  $I_j$  are intervals. Then,

$$\begin{aligned} \int g dF &= E_F[g] = \sum_{j=1}^m b_j P_F(I_j) = \sum_{j=1}^m b_j \int_{I_j} f(x) dx \\ &= \sum_{j=1}^m b_j \int_{-\infty}^{\infty} f(x) \mathbf{1}_{I_j}(x) dx \\ &= \int_{-\infty}^{\infty} \left[ \sum_{j=1}^m b_j \mathbf{1}_{I_j}(x) \right] f(x) dx \\ &= \int_{-\infty}^{\infty} g(x) f(x) dx. \end{aligned}$$

Validity of (4.12) for general  $g \geq 0$  follows by approximating with step functions  $g_n \uparrow g$ , which is possible since  $g$  is piecewise continuous, and then by appeal to the monotone convergence theorem. ■

**Corollary 4.22.** Let  $F$  and  $f$  be as in Theorem 4.21. Then, a piecewise continuous function  $g$  is integrable with respect to  $F$  if and only if  $\int_{-\infty}^{\infty} |g(x)|f(x) dx < \infty$ , and in this case (4.12) holds. □

### 4.2.4 Mixed distribution functions

Suppose that  $F = \alpha F_d + (1 - \alpha) F_a$  is a convex combination of a discrete distribution function  $F_d(t) = \sum_i p_i \mathbf{1}_{(t_i \leq t)}$  and an absolutely continuous distribution function  $F_a(t) = \int_{-\infty}^t f(x) dx$ . Then, by Theorem 4.19 and Theorem 4.21, the integral of  $g$  with respect to  $F$  is a corresponding convex combination of integrals with respect to  $F_d$  and  $F_a$ :

$$\begin{aligned} \int g dF &= \alpha \int g dF_d + (1 - \alpha) \int g dF_a \\ &= \alpha \sum_i p_i g(t_i) + (1 - \alpha) \int_{-\infty}^{\infty} g(x)f(x) dx. \end{aligned} \quad (4.13)$$

In order that this integral exist,  $g$  must be piecewise continuous and either positive or integrable with respect to both  $F_d$  and  $F_a$ .

## 4.3 Computation of Expectations

Other than the explicit definition of expectation for simple random variables, given by (4.3), we have as yet no methods for calculating expectations. This section is devoted to development of such methods, which involve integrals with respect to distribution functions.

### 4.3.1 Positive random variables

For  $X \geq 0$ ,  $E[X]$  is not only the integral of  $g(x) \equiv x$  with respect to  $F_X$ , but also the ordinary integral of  $1 - F_X(y) = P\{X > y\}$ .

**Theorem 4.23.** If  $X \geq 0$ , then

$$E[X] = \int_0^\infty x dF_X(x) = \int_0^\infty [1 - F_X(y)] dy. \quad (4.14)$$

**Proof:** To begin, we show that the first equality in (4.14) holds true when  $X = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$  is a simple. In this case,  $E[X] = \sum_{i=1}^n a_i P(A_i)$  by (4.3). On the other hand,  $F_X(t) = \sum_i P(A_i) \mathbf{1}(a_i \leq t)$ , so that  $\int_0^\infty x dF_X(x) = \sum_{i=1}^n a_i P(A_i)$  by (4.11).

We next prove that the second equality always holds:

$$\begin{aligned} \int_0^\infty x dF_X(x) &= \int_0^\infty \left[ \int_{(0,x)} dy \right] dF_X(x) \\ &= \int_0^\infty \left[ \int_{(y,\infty)} dF_X(x) \right] dy \\ &= \int_0^\infty [1 - F_X(y)] dy. \end{aligned}$$

The interchange of the order of integration is valid because the functions involved are positive; see the discussion of Fubini's theorem in §6.

Finally, we extend the first equality to all positive  $X$ . Let  $X_n$  be simple with  $X_n \uparrow X$ . Then, since  $1 - F_{X_n}(y) \uparrow 1 - F_X(y)$  for each  $y$ ,

$$\begin{aligned} E[X] &= \lim_{n \rightarrow \infty} E[X_n] = \lim_{n \rightarrow \infty} \int_0^\infty [1 - F_{X_n}(y)] dy \\ &= \int_0^\infty [1 - F_X(y)] dy \\ &= \int_{-\infty}^\infty x dF_X(x). \end{aligned}$$

Here the first and third equalities are by the monotone convergence theorem, and the fourth is by the second part of (4.14). ■

The case that  $X$  is integer-valued merits special mention.

**Corollary 4.24.** Let  $X$  be positive and integer-valued. Then,

$$E[X] = \sum_{n=1}^{\infty} nP\{X = n\} = \sum_{k=1}^{\infty} P\{X \geq k\}. \quad (4.15)$$

**Proof:** The first equality in (4.15) holds by Theorem 4.19, and the second also holds because for  $y \in [k-1, k)$ ,

$$1 - F_X(y) = P\{X > y\} = P\{X \geq k\},$$

so that

$$\int_0^{\infty} [1 - F_X(y)] dy = \sum_{k=1}^{\infty} \int_{k-1}^k [1 - F_X(y)] dy = \sum_{k=1}^{\infty} P\{X \geq k\}. \quad \blacksquare$$

### 4.3.2 Integrable random variables

We now derive an analogue of the first part of Theorem 4.23 for integrable random variables.

**Theorem 4.25.** For  $X \in L^1$ ,

$$E[X] = \int_{-\infty}^{\infty} x dF_X(x). \quad (4.16)$$

**Proof:** By Definition 4.12 and Theorem 4.23 (see also Exercise 2.12),

$$\begin{aligned} E[X] &= E[X^+] - E[X^-] = \int_0^{\infty} x dF_{X^+}(x) - \int_0^{\infty} x dF_{X^-}(x) \\ &= \int_{-\infty}^{\infty} x dF_X(x). \quad \blacksquare \end{aligned}$$

### 4.3.3 Functions of random variables

Let  $X$  be a random variable and let  $g: \mathbb{R} \rightarrow \mathbb{R}$  be a Borel measurable function. We wish to calculate  $E[g(X)]$ . Of course, Theorem 4.23 or Theorem 4.25 would apply directly if we were able to calculate the distribution function of  $g(X)$ , but we have seen in §2.5 that ordinarily this is not possible. Fortunately, neither is it necessary.

Discrete $X$	$E[g(X)] = \sum_i g(t_i)P\{X = t_i\}$
Absolutely Continuous $X$	$E[g(X)] = \int_{-\infty}^{\infty} g(x)f_X(x)dx$

Table 4.1. Computational Formulas for  $E[g(X)]$ 

**Theorem 4.26.** Let  $g$  be either positive or such that  $g(X) \in L^1$ . Then,

$$E[g(X)] = \int g(x) dF_X(x). \quad (4.17)$$

**Proof:** Suppose that  $g = \sum_{i=1}^n b_i \mathbf{1}_{B_i}$  is a simple function. Then, the random variable  $g(X) = \sum_{i=1}^n b_i \mathbf{1}(X \in B_i)$  is simple, and so

$$\begin{aligned} E[g(X)] &= \sum_{i=1}^n b_i P\{X \in B_i\} = \sum_{i=1}^n b_i \int \mathbf{1}_{B_i}(x) dF_X(x) \\ &= \int g(x) dF_X(x). \end{aligned}$$

If  $g \geq 0$  and the  $g_n$  are simple with  $g_n \uparrow g$ , then  $g_n(X) \uparrow g(X)$ , so

$$E[g(X)] = \lim_{n \rightarrow \infty} E[g_n(X)] = \lim_{n \rightarrow \infty} \int g_n(x) dF(x) = \int g(x) dF(x).$$

The first equality is by the monotone convergence theorem for expectations and the last by the monotone convergence theorem for integrals with respect to  $F$ .

Finally, if  $g(X)$  is integrable, then

$$\begin{aligned} E[g(X)] &= E[g^+(X)] - E[g^-(X)] \\ &= \int g^+(x) dF(x) - \int g^-(x) dF(x) \\ &= \int g(x) dF(x). \quad \blacksquare \end{aligned}$$

Table 4.1 summarizes the most important special cases of (4.17).

#### 4.3.4 Functions of random vectors

Except for the case of independent components, the only useful formulas are those for the discrete and absolutely continuous cases. For the sake of simplicity, we prove only the more difficult of the two results.

**Theorem 4.27.** Let  $X_1, \dots, X_n$  be discrete, with values in the countable set  $C$ , and let  $g: \mathbb{R}^n \rightarrow \mathbb{R}$  be Borel measurable. Then, if either  $g \geq 0$  or  $g(X_1, \dots, X_n) \in L^1$ ,

$$E[g(X_1, \dots, X_n)] = \sum_{t_1 \in C} \cdots \sum_{t_n \in C} g(t_1, \dots, t_n) P\{X_1 = t_1, \dots, X_n = t_n\}. \quad \blacksquare$$

**Theorem 4.28.** Let  $X = (X_1, \dots, X_n)$  be an absolutely continuous random vector with density  $f_X$ , and let  $g: \mathbb{R}^n \rightarrow \mathbb{R}$  be Borel measurable. Then, if either  $g \geq 0$  or  $g(X_1, \dots, X_n) \in L^1$ ,

$$E[g(X_1, \dots, X_n)] = \int_{\mathbb{R}^n} g(x_1, \dots, x_n) f_X(x_1, \dots, x_n) dx_1 \cdots dx_n. \quad (4.18)$$

**Proof:** We apply Theorem 2.55. Let  $\mathcal{J}$  be the  $\pi$ -system of rectangles  $B = \prod_{i=1}^n (-\infty, t_i]$ , where  $t_1, \dots, t_n \in \mathbb{R}$ , which generates  $\mathcal{B}(\mathbb{R}^n)$  by Definition 1.16, and let  $\mathbf{H}$  be the set of positive, Borel measurable functions  $g: \mathbb{R}^n \rightarrow \mathbb{R}_+$  satisfying (4.18). Then, evidently,  $1 \in \mathbf{H}$ . Next,  $\mathbf{H}$  is a cone, since whenever  $f, g \in \mathbf{H}$  and  $a, b > 0$ ,

$$\begin{aligned} E[(af + bg)(X_1, \dots, X_n)] &= aE[f(X_1, \dots, X_n)] + bE[g(X_1, \dots, X_n)] \\ &= a \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \dots, x_n) f_X(x_1, \dots, x_n) dx_1 \cdots dx_n \\ &\quad + b \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \dots, x_n) f_X(x_1, \dots, x_n) dx_1 \cdots dx_n \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} (af + bg)(x_1, \dots, x_n) f_X(x_1, \dots, x_n) dx_1 \cdots dx_n. \end{aligned}$$

Finally, if  $g_k \in \mathbf{H}$  for each  $k$  and  $g_k \uparrow g$  (and  $g$  is finite-valued), then

$$\begin{aligned} E[g(X_1, \dots, X_n)] &= \lim_{k \rightarrow \infty} E[g_k(X_1, \dots, X_n)] \\ &= \lim_{k \rightarrow \infty} \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g_k(x_1, \dots, x_n) f_X(x_1, \dots, x_n) dx_1 \cdots dx_n \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \dots, x_n) f_X(x_1, \dots, x_n) dx_1 \cdots dx_n, \end{aligned}$$

where the first equality is by the monotone convergence theorem for expectations, the second holds since  $g_n \in \mathbf{H}$  and the third is by the monotone convergence theorem for Lebesgue integrals. Hence, by Theorem 2.55,  $\mathbf{H}$  contains all positive, Borel measurable functions.

Finally, for  $g$  such that  $g(X_1, \dots, X_n) \in L^1$ ,

$$\begin{aligned}
E[g(X_1, \dots, X_n)] &= E[(g(X_1, \dots, X_n))^+] - E[(g(X_1, \dots, X_n))^-] \\
&= E[g^+(X_1, \dots, X_n)] - E[g^-(X_1, \dots, X_n)] \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g^+(x_1, \dots, x_n) f_X(x_1, \dots, x_n) dx_1 \cdots dx_n \\
&\quad - \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g^-(x_1, \dots, x_n) f_X(x_1, \dots, x_n) dx_1 \cdots dx_n \\
&= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} g(x_1, \dots, x_n) f_X(x_1, \dots, x_n) dx_1 \cdots dx_n. \quad \blacksquare
\end{aligned}$$

### 4.3.5 Functions of independent random variables

When the components of a random vector are independent, more specific computational formulas obtain. We state the result for two random variables, even though it generalizes in an obvious manner.

**Theorem 4.29.** Let  $X$  and  $Y$  be independent, and let  $h: \mathbb{R}^2 \rightarrow \mathbb{R}_+$  be Borel measurable. Then,

$$\begin{aligned}
E[h(X, Y)] &= \int \left[ \int h(x, y) dF_X(x) \right] dF_Y(y) \quad (4.19) \\
&= \int \left[ \int h(x, y) dF_Y(y) \right] dF_X(x).
\end{aligned}$$

**Proof:** We apply Theorem 2.55. Let  $\mathbf{H}$  be the set of functions  $h: \mathbb{R}^2 \rightarrow \mathbb{R}_+$  satisfying (4.19), and let  $\mathcal{S}$  be the  $\pi$ -system of rectangles  $B_1 \times B_2$ , where  $B_1$  and  $B_2$  belong to  $\mathcal{B}(\mathbb{R})$ . By the reasoning used to prove Theorem 4.28,  $\mathbf{H}$  satisfies conditions i), iii) and iv) of Theorem 2.55, and since  $\sigma(\mathcal{S}) = \mathcal{B}((\mathbb{R}^2)$ , it now suffices to show that (4.19) holds when  $h = \mathbf{1}_B$  for some  $B \in \mathcal{S}$ . This, however, is easy. On the one hand,

$$E[h(X, Y)] = P\{X \in B_1, Y \in B_2\} = P\{X \in B_1\}P\{Y \in B_2\},$$

while also

$$\int h(x, y) dF_X(x) = P\{X \in B_1\}\mathbf{1}_{B_2}(y),$$

for each  $y$ , so that

$$\int \left[ \int h(x, y) dF_X(x) \right] dF_Y(y) = P\{X \in B_1\}P\{Y \in B_2\}. \quad \blacksquare$$

In particular, the expectation of the product of functions of independent random variables is the product of their expectations. Ordinarily the product of integrable random variables need not be integrable, but Corollary 4.30 identifies one instance in which this is so.

**Corollary 4.30.** Let  $X, Y$  be independent, and let  $g_1, g_2$  be positive functions. Then,

$$E[g_1(X)g_2(Y)] = E[g_1(X)] E[g_2(Y)]. \quad (4.20)$$

If  $g_1(X)$  and  $g_2(Y)$  are integrable, then  $g_1(X)g_2(Y) \in L^1$  and (4.20) holds.

**Proof:** For  $g_1, g_2 \geq 0$ , it suffices to apply (4.19) with  $h(x, y) = g_1(x)g_2(y)$ :

$$\begin{aligned} E[g_1(X)g_2(Y)] &= \int \left[ \int g_1(x)g_2(y) d\mathbb{F}_X(x) \right] d\mathbb{F}_Y(y) \\ &= \int g_1(x) d\mathbb{F}_X(x) \int g_2(y) d\mathbb{F}_Y(y) \\ &= E[g_1(X)] E[g_2(Y)]. \end{aligned}$$

If  $g_1(X)$  and  $g_2(Y)$  are integrable, then by (4.20),

$$E[|g_1(X)g_2(Y)|] = E[|g_1(X)|]E[|g_2(Y)|] < \infty.$$

That (4.20) extends is then computational. ■

### 4.3.6 Sums of independent random variables

We now derive the distribution function of the sum of independent random variables.

**Theorem 4.31.** Let  $X$  and  $Y$  be independent. Then,

$$\mathbb{F}_{X+Y}(t) = \int \mathbb{F}_X(t-y) d\mathbb{F}_Y(y) = \int \mathbb{F}_Y(t-x) d\mathbb{F}_X(x), \quad t \in \mathbb{R}.$$

**Proof:** With  $t$  fixed, take  $h(x, y) = \mathbf{1}_{(-\infty, t]}(x+y)$  in (4.19), leading to

$$\begin{aligned} P\{X+Y \leq t\} &= E[\mathbf{1}_{(-\infty, t]}(X+Y)] \\ &= \int \left[ \int \mathbf{1}_{(-\infty, t]}(x+y) d\mathbb{F}_X(x) \right] d\mathbb{F}_Y(y) \\ &= \int \left[ \int \mathbf{1}_{(-\infty, t-y]}(x) d\mathbb{F}_X(x) \right] d\mathbb{F}_Y(y) \\ &= \int \mathbb{F}_X(t-y) d\mathbb{F}_Y(y). \quad ■ \end{aligned}$$

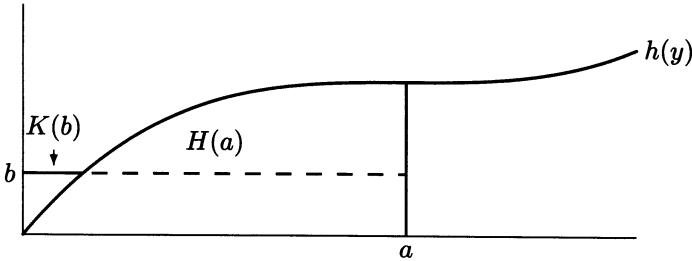


Figure 4.1. The Proof of Young's Inequality

The operation in Theorem 4.31, as in Theorem 3.12 for densities, is termed convolution.

**Definition 4.32.** The *convolution* of distribution functions  $F$  and  $G$  on  $\mathbb{R}$  is the distribution function

$$F * G(t) = \int F(t - y) dG(y) = \int G(t - x) dF(x). \quad \square$$

## 4.4 $L^p$ Spaces and Inequalities

The main results in this section are inequalities involving expectations.

### 4.4.1 $L^p$ spaces

The space  $L^p$  consists of all random variables whose  $p$ th absolute power is integrable.

**Definition 4.33.** For  $1 \leq p < \infty$ ,  $L^p$  denotes the set of random variables  $X$  such that  $E[|X|^p] < \infty$ .  $\square$

This notation is consistent with the “ $L^1$ ” introduced in §1.

### 4.4.2 Key inequalities

We proceed to a suite of important and extremely useful inequalities involving expectations. They are based on the following analytical result.

**Lemma 4.34 (Young's inequality).** Let  $h: \mathbb{R}_+ \rightarrow \mathbb{R}_+$  be continuous and strictly increasing with  $h(0) = 0$  and  $h(\infty) = \infty$ , let  $k$  be the pointwise inverse of  $h$ , and define  $H(x) = \int_0^x h(y) dy$  and  $K(x) = \int_0^x k(y) dy$ . Then, for all  $a, b \in \mathbb{R}_+$ ,

$$ab \leq H(a) + K(b).$$

**Proof:** There are situations where one picture is worth a thousand words. This is one — see Figure 4.1. ■

Here is the first major inequality.

**Theorem 4.35 (Hölder's inequality).** Suppose that  $p, q > 1$  satisfy

$$\frac{1}{p} + \frac{1}{q} = 1, \quad (4.21)$$

that  $X \in L^p$  and that  $Y \in L^q$ . Then,  $XY \in L^1$  and

$$E[|XY|] \leq E[|X|^p]^{1/p} E[|Y|^q]^{1/q}. \quad (4.22)$$

**Proof:** If  $E[|X|^p] = 0$  or  $E[|Y|^q] = 0$ , then  $XY \stackrel{\text{a.s.}}{=} 0$  by Proposition 4.11, and (4.22) holds trivially, so we assume that both are strictly positive.

By Young's inequality (Lemma 4.34), applied to the function  $h(x) = x^{p-1}$ , for  $a, b > 0$ ,

$$ab \leq \frac{a^p}{p} + \frac{b^q}{q}.$$

Substituting  $a = |X|/E[|X|^p]^{1/p}$  and  $b = |Y|/E[|Y|^q]^{1/q}$  gives

$$\frac{|XY|}{E[|X|^p]^{1/p} E[|Y|^q]^{1/q}} \leq \frac{|X|^p}{pE[|X|^p]} + \frac{|Y|^q}{qE[|Y|^q]}, \quad (4.23)$$

and (4.22) follows by taking expectations in (4.23) and applying (4.21). ■

Values  $p$  and  $q$  satisfying (4.21) are *conjugate exponents*. The case  $p = q = 2$  is especially important.

**Corollary 4.36 (Cauchy-Schwarz inequality).** If  $X$  and  $Y$  belong to  $L^2$ , then  $XY \in L^1$  and

$$E[|XY|] \leq \sqrt{E[X^2]E[Y^2]}. \quad \square$$

Also, the spaces  $L^p$  decrease as  $p$  increases, but more is true:  $E[|X|^p]^{1/p}$  is an increasing function of  $p$ .

**Corollary 4.37 (Lyapunov's inequality).** For  $1 \leq r \leq s$ ,  $L^s \subseteq L^r$ , and for  $X \in L^s$

$$E[|X|^r]^{1/r} \leq E[|X|^s]^{1/s}. \quad (4.24)$$

**Proof:** It suffices to apply (4.22) to  $X \in L^r$  with  $p = s/r$  and  $Y \equiv 1$ . ■

In particular, taking  $s = 2$  and  $r = 1$  gives

$$E[X^2] \leq E[X]^2,$$

which also can be deduced from the Cauchy-Schwarz inequality, as well as Jensen's inequality (Theorem 4.39 below).

The next result shows that each  $L^p$  space is a vector space, and more. In particular, with  $\|X\|_p = E[|X|^p]^{1/p}$ , the function

$$d(X, Y) = \|X - Y\|_p = E[|X - Y|^p]^{1/p}$$

is positive and satisfies the triangle inequality. This function is not, though, a metric in general, since  $d(X, Y) = 0$  does not imply that  $X = Y$  as functions on  $\Omega$ , but only that  $X \xrightarrow{\text{a.s.}} Y$ .

**Theorem 4.38 (Minkowski's inequality).** Suppose that  $p \geq 1$  and that  $X$  and  $Y$  belong to  $L^p$ . Then,  $X + Y \in L^p$  and

$$E[|X + Y|^p]^{1/p} \leq E[|X|^p]^{1/p} + E[|Y|^p]^{1/p}. \quad (4.25)$$

**Proof:** For  $p = 1$ , this is immediate by the triangle inequality.

For  $p > 1$ , let  $q = p/(p - 1)$ , so that (4.21) is satisfied. Then, by the triangle equality followed by Hölder's inequality (4.22),

$$\begin{aligned} E[|X + Y|^p] &\leq E[|X + Y|^{p-1}|X|] + E[|X + Y|^{p-1}|Y|] \\ &\leq E[|X|^p]^{1/p} E[|X + Y|^{q(p-1)}]^{1/q} \\ &\quad + E[|Y|^p]^{1/p} E[|X + Y|^{q(p-1)}]^{1/q} \\ &= E[|X + Y|^p]^{1/q} \left( E[|X|^p]^{1/p} + E[|Y|^p]^{1/p} \right), \end{aligned}$$

where at the last step we used the property that  $q(p - 1) = p$ . Then, (4.25) follows by dividing both sides of the last expression by  $E[|X + Y|^p]^{1/q}$  and using the property that  $1 - 1/q = 1/p$ . ■

The next inequality pertains to expectations of convex functions of random variables.

**Theorem 4.39 (Jensen's inequality).** Let  $g$  be convex and suppose that  $X$  and  $g(X)$  are integrable. Then,

$$g(E[X]) \leq E[g(X)]. \quad (4.26)$$

**Proof:** By convexity, there exists  $a \in \mathbb{R}$  (in nice cases,  $a = g'(E[X])$ ; otherwise  $a$  is a one-sided derivative) such that for all  $x \in \mathbb{R}$ ,

$$g(x) \geq g(E[X]) + a(x - E[X]).$$

Thus,

$$g(X) \geq g(E[X]) + a(X - E[X]),$$

and taking expectations gives

$$E[g(X)] \geq g(E[X]) + aE[X - E[X]] = g(E[X]). \blacksquare$$

The final result uses expectations to provide upper bounds on tail probabilities for random variables.

**Theorem 4.40 (Chebyshev's inequality).** Let  $X$  be positive, and let  $g$  be a positive, increasing function on  $\mathbb{R}_+$ . Then, for each  $a > 0$ ,

$$P\{X \geq a\} \leq \frac{E[g(X)]}{g(a)}. \quad (4.27)$$

**Proof:** Taking expectations in the inequality

$$g(X) \geq g(X)\mathbf{1}(X \geq a) \geq g(a)\mathbf{1}(X \geq a)$$

gives

$$E[g(X)] \geq E[g(X); \{X \geq a\}] = g(a)P\{X \geq a\}. \blacksquare$$

Several special cases of (4.27) occur frequently in probability:

$$X \in L^1 \Rightarrow P\{|X| \geq a\} \leq \frac{E[|X|]}{a} \quad (4.28)$$

$$X \in L^p \Rightarrow P\{|X| \geq a\} \leq \frac{E[|X|^p]}{a^p} \quad (4.29)$$

$$X \in L^2 \Rightarrow P\{|X - E[X]| \geq a\} \leq \frac{\text{Var}(X)}{a^2} \quad (4.30)$$

$$X \geq 0 \Rightarrow P\{X \geq a\} \leq \frac{E[e^{tX}]}{e^{ta}}. \quad (4.31)$$

Each holds for all  $a > 0$  and the last is valid for all positive  $t$ . The variance of  $X$ ,  $\text{Var}(X)$ , which appears in (4.30), will be defined momentarily.

**Note on Terminology.** There is no uniform nomenclature for these inequalities. We refer to (4.27) through (4.31) all as *Chebyshev's inequality*. Alternatives include *Markov's inequality* for (4.28) alone; *Markov's inequality* for both (4.28) and (4.29); and *Chebyshev's inequality* or the *Chebyshev-Bienaym  inequality* for (4.30) alone.  $\square$

Table 4.2 summarizes the major inequalities.

Name	Conditions	Statement
Hölder	$X \in L^p, Y \in L^q$	$E[ XY ] \leq \ X\ _p \ Y\ _q$
Cauchy-Schwarz	$X, Y \in L^2$	$E[ XY ] \leq \sqrt{E[X^2]} E[Y^2]$
Lyapunov	$X \in L^s; r \leq s$	$\ X\ _r \leq \ X\ _s$
Minkowski	$X, Y \in L^p$	$\ X + Y\ _p \leq \ X\ _p + \ Y\ _p$
Jensen	$g$ convex; $X, g(X) \in L^1$	$g(E[X]) \leq E[g(X)]$
Chebyshev	$X \geq 0, g \geq 0$ and ↑	$P\{X \geq a\} \leq E[g(X)]/g(a)$

Table 4.2. The Key Inequalities

## 4.5 Moments

We now define moments for random variables and vectors.

### 4.5.1 Moments of random variables

Moments of random variables are expectations of powers. The first moment is the expectation; it and second moment (more precisely, the second central moment, or variance) are crucial throughout probability and statistics.

**Definition 4.41.** Let  $X$  be a random variable.

- a) For  $k$  a positive integer and if  $X \in L^k$ , the  $k$ th *moment* of  $X$  is  $E[X^k]$ .
- b) If  $X \in L^1$ , the *mean* of  $X$  is  $E[X]$ , the expectation of  $X$ .
- c) For  $k$  a positive integer and if  $X \in L^k$ , the  $k$ th *central moment* of  $X$  is  $E[(X - E[X])^k]$ . □

That central moments exist under the conditions stated in Definition 4.41 follows from Corollary 4.37:  $L^k \subseteq L^1$  for each  $k \geq 1$ .

Moments can be calculated via distribution functions: by Theorem 4.26,

$$E[X^k] = \int x^k dF_X(x).$$

Thus, we can then speak sensibly of the moments of a distribution function, for example, the mean and variance.

### 4.5.2 Variance and standard deviation

The second central moment plays a particularly important role.

**Definition 4.42.** Suppose that  $X \in L^2$ .

- a) The *variance* of  $X$  is the second central moment:

$$\text{Var}(X) = E[(X - E[X])^2]. \quad (4.32)$$

- b) The *standard deviation* of  $X$  is  $\sqrt{\text{Var}(X)}$ .  $\square$

The greater the variability of a random variable about its mean, either in terms of size or in terms of likelihood, the higher its variance. The variance of  $X \in L^2$  can also be expressed as

$$\text{Var}(X) = E[X^2] - E[X]^2, \quad (4.33)$$

which is more convenient than (4.32) for computational purposes.

Variance is additive for independent random variables.

**Proposition 4.43.** If  $X, Y \in L^2$  are independent, then

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y).$$

**Proof:** By (4.33), and since

$$E[XY] = E[X]E[Y]$$

by independence,

$$\begin{aligned} \text{Var}(X + Y) &= E[(X + Y)^2] - E[X + Y]^2 \\ &= E[X^2] + 2E[XY] + E[Y^2] \\ &\quad - (E[X]^2 + 2E[X]E[Y] + E[Y]^2) \\ &= E[X^2] - E[X]^2 + E[Y^2] - E[Y]^2 \\ &= \text{Var}(X) + \text{Var}(Y). \quad \blacksquare \end{aligned}$$

For the parametrized families of distributions introduced in §2.4, the parameters relate the mean and variance as shown in Table 4.3, which parallels Tables 2.1 and 2.2. Note in particular the similarity between geometric and exponential distributions.

Distribution	Parameters	Mean	Variance
Bernoulli	$p \in [0, 1]$	$p$	$p(1 - p)$
Binomial	$n \in \mathbb{N}, p \in [0, 1]$	$np$	$np(1 - p)$
Geometric	$p \in (0, 1)$	$1/p$	$1/p^2$
Negative binomial	$n \geq 1, p \in (0, 1)$	$n/p$	$n/p^2$
Poisson	$\lambda \in (0, \infty)$	$\lambda$	$\lambda$
Uniform on $[a, b]$	$a < b \in \mathbb{R}$	$(a + b)/2$	$(b - a)/12$
Standard normal	None	0	1
Normal	$\mu \in \mathbb{R}, \sigma^2 > 0$	$\mu$	$\sigma^2$
Exponential	$\lambda > 0$	$1/\lambda$	$1/\lambda^2$
Gamma	$\alpha > 0, \lambda > 0$	$\alpha/\lambda$	$\alpha/\lambda^2$
$\chi^2$	$k \in \mathbb{N}$	$k$	$2k$

Table 4.3. Moments of Key Distributions

### 4.5.3 Covariance and correlation

“Product moments” of pairs of random variables are important as well.

**Definition 4.44.** Given  $X$  and  $Y \in L^2$ ,

- a) The *covariance* of  $X$  and  $Y$  is

$$\text{Cov}(X, Y) = E[(X - E[X])(Y - E[Y])].$$

- b) Provided that  $\text{Var}(X) > 0$  and  $\text{Var}(Y) > 0$ , the *correlation* between  $X$  and  $Y$  is

$$\text{Corr}(X, Y) = \frac{\text{Cov}(X, Y)}{\sqrt{\text{Var}(X) \text{Var}(Y)}}.$$

- c)  $X$  and  $Y$  are *uncorrelated* if  $\text{Corr}(X, Y) = 0$ .  $\square$

The more useful expression for covariance, analogous to (4.33), is

$$\text{Cov}(X, Y) = E[XY] - E[X]E[Y]. \quad (4.34)$$

Correlation measures linear association between random variables, but *only* linear association:  $X$  and  $Y$  can be functionally dependent but uncorrelated notwithstanding. By the Cauchy-Schwarz inequality (Corollary 4.36),  $|\text{Corr}(X, Y)| \leq 1$ , with equality if and only if there are  $a, b \in \mathbb{R}$  such that  $Y \stackrel{\text{a.s.}}{=} aX + b$ . Independent random variables in  $L^2$  are uncorrelated.

#### 4.5.4 Moments of random vectors

Moments of random vectors are defined component-wise. No new theory is involved, although the mean is a vector and covariances are given by a matrix.

**Definition 4.45.** Let  $X = (X_1, \dots, X_d)$  be a random  $d$ -vector.

- a) Provided that  $E[X_i] < \infty$  for each  $i$ , the *mean* of  $X$  is the  $d$ -vector

$$\mu_X = (E[X_1], \dots, E[X_d]).$$

- b) Provided that  $E[X_i^2] < \infty$  for each  $i$ , the *covariance matrix* of  $X$  is the  $d \times d$  matrix

$$C_X^2(i, j) = \text{Cov}(X_i, X_j). \quad \square$$

The covariance matrix  $C_X$  has the properties that  $C_X^2(i, i) = \text{Var}(X_i)$  for each  $i$ , that it is symmetric: for each  $i$  and  $j$ ,  $C_X^2(i, j) = C_X^2(j, i)$ , and that it is positive definite: for every  $d$ -vector  $x$ ,

$$\sum_{i=1}^d \sum_{j=1}^d x_i C_X^2(i, j) x_j \geq 0.$$

#### 4.5.5 Multivariate normal distributions

Here we present the full definition for a random vector to be normally distributed: it is obtained via an affine transformation of a random vector with i.i.d. components, each with the standard normal distribution.

Let  $\mu$  be an element of  $\mathbb{R}^n$  and let  $C^2$  be a symmetric, positive definite, nonsingular  $n \times n$  matrix. We use the notation  $A^T$  for the transpose of a matrix  $A$ .

**Definition 4.46.** A random vector  $X = (X_1, \dots, X_n)$  has a *multivariate normal distribution* with *mean vector*  $\mu$  and *covariance matrix*  $C^2$  if  $X = CY + \mu$ , where  $Y_1, \dots, Y_n$  are i.i.d. with distribution  $N(0, 1)$  and where  $C$  is the unique matrix satisfying  $C^T C = C^2$  (which exists because  $C^2$  is symmetric and positive definite).  $\square$

If this is the case, we say that  $X$  has distribution  $N(\mu, C^2)$ . Such random vectors may be characterized as follows.

**Proposition 4.47.** Let  $X$  have the multivariate normal distribution  $N(\mu, C^2)$ . Then,

- a)  $E[X_i] = \mu_i$  for each  $i$ .

b)  $\text{Cov}(X_i, X_j) = C^2(i, j)$  for each  $i$  and  $j$ .

c)  $X$  is absolutely continuous with density

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^n \det C^2}} \exp\left[-\frac{1}{2}(x - \mu)^T [C^2]^{-1}(x - \mu)\right]. \quad (4.35)$$

**Proof:** a) If  $X \stackrel{\text{d}}{=} N(\mu, C^2)$ , then  $X_i = \sum_{j=1}^n C(i, j)Y_j + \mu_i$  is normally distributed by Example 3.14, and  $E[X_i] = \sum_{j=1}^n C(i, j)E[Y_j] + \mu_i = \mu_i$ .

b) To calculate covariances, we may assume that  $\mu = 0$ , so that

$$\begin{aligned} \text{Cov}(X_i, X_k) &= E[X_i X_k] = E\left[\sum_{j=1}^n C(i, j)Y_j \cdot \sum_{\ell=1}^n C(k, \ell)Y_\ell\right] \\ &= \sum_{j=1}^n \sum_{\ell=1}^n C(i, j)C(k, \ell)E[Y_j Y_\ell] \\ &= \sum_{j=1}^n C(i, j)C(k, j) \\ &= C^2(i, k). \end{aligned}$$

c) Finally, since  $Y = (Y_1, \dots, Y_n)$  has density  $\prod_{i=1}^n (1/\sqrt{2\pi}) e^{-y_i^2/2}$ ,  $X$  has density (4.35) by Theorem 2.43. ■

## 4.6 Complements

### 4.6.1 Integration with respect to Lebesgue measure

Integration with respect to Lebesgue measure (§1.6) is conceptually the same as expectation. The integral is defined via the same procedure, and the properties are nearly identical.

**Definition 4.48.** Let  $f$  be a Borel measurable function on  $\mathbb{R}$ .

- a) If  $f = \sum_{i=1}^n b_i \mathbf{1}_{B_i}$  is positive and simple, the *Lebesgue integral* of  $f$  is  $\int f(x) dx = \sum_{i=1}^n b_i \lambda(B_i) \leq \infty$ .
- b) The *Lebesgue integral* of  $f \geq 0$  is  $\int f(x) dx = \lim_{n \rightarrow \infty} \int f_n(x) dx$ , where the  $f_n$  are positive, simple functions such that  $f_n \uparrow f$ .
- c)  $f$  is *integrable* if  $\int |f(x)| dx < \infty$ , and in this case the *Lebesgue integral* of  $f$  is  $\int f(x) dx = \int f^+(x) dx - \int f^-(x) dx$ . We denote by  $L^1$  the set of integrable functions.
- d) If  $f$  either positive or integrable and  $B \in \mathcal{B}(\mathbb{R})$ , the *Lebesgue integral* of  $f$  over  $B$  is  $\int_B f(x) dx = \int f(x) \mathbf{1}_B(x) dx$ . □

The restriction to positive simple functions in part a) is necessary because some of the  $B_i$  may have infinite Lebesgue measure, in which case allowing the  $b_i$  to be both positive and negative could lead to indeterminate expressions of the form  $\infty - \infty$ . Even simple functions — nonzero constants in particular — can have infinite Lebesgue integrals.

Here are the major properties.

- a) **Linearity.** If  $f, g$  are positive and  $a, b \in \mathbb{R}_+$  or if  $f$  and  $g$  are integrable and  $a, b \in \mathbb{R}$ , then  $\int (af + bg)(x) dx = a \int f(x) dx + b \int g(x) dx$ .
- b) **Monotonicity.** If either  $0 \leq f \leq g$  or  $f$  and  $g$  are integrable and  $f \leq g$ , then  $\int f(x) dx \leq \int g(x) dx$ .
- c) **Fatou's lemma.** If  $f_n \geq 0$  for each  $n$ , then

$$\int \liminf_{n \rightarrow \infty} f_n(x) dx \leq \liminf_{n \rightarrow \infty} \int f_n(x) dx.$$

- d) **Monotone convergence theorem.** If  $f, f_1, f_2, \dots$  are positive with  $f_n(x) \uparrow f(x)$  for each  $x$ , then  $\int f_n(x) dx \uparrow \int f(x) dx$ .
- e) **Monotonicity.** If  $f, g \in L^1$  and  $f \leq g$ , then  $\int f(x) dx \leq \int g(x) dx$ .
- f) **Dominated convergence theorem.** If  $f, f_1, f_2, \dots$  are integrable with  $f_n(x) \rightarrow f(x)$  for each  $x$ , and if there is  $g \in L^1$  with  $|f_n(x)| \leq g(x)$  for all  $n$  and  $x$ , then  $\int f_n(x) dx \rightarrow \int f(x) dx$ .  $\square$

Theorem 4.21 remains valid if  $f$  and  $g$  are merely Borel measurable, provided that the integrals be Lebesgue integrals.

Lebesgue and Riemann integrals are related: whenever the Riemann integral exists, so does the Lebesgue integral and the two are equal.

### 4.6.2 Expectation for product probabilities

Let  $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2, P_1 \times P_2)$  be the product of the probability spaces  $(\Omega_1, \mathcal{F}_1, P_1)$  and  $(\Omega_2, \mathcal{F}_2, P_2)$ , as defined in §3.7. We are concerned here with computation of expectations with respect to  $P_1 \times P_2$ . Under either positivity or integrability, these expectations, which are double integrals, can be computed as iterated integrals instead, in either order. In particular, the two iterated integrals are equal, so that interchange of the order of integration is permissible.

**Theorem 4.49 (Fubini's theorem).** Let  $X$  be a positive random variable on  $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2, P_1 \times P_2)$ . Then,

- a) The function  $\omega_1 \mapsto Y_1(\omega_1) = E_{P_2}[X(\omega_1, \cdot)]$  is a random variable over  $(\Omega_1, \mathcal{F}_1, P_1)$ .
- b) The function  $\omega_2 \mapsto Y_2(\omega_2) = E_{P_1}[X(\cdot, \omega_2)]$  is a random variable over  $(\Omega_2, \mathcal{F}_2, P_2)$ .

$$\text{c) } E_{P_1}[Y_1] = E_{P_1 \times P_2}[X] = E_{P_2}[Y_2].$$

**Proof:** Let  $\mathbf{H}$  be the set of positive random variables on  $(\Omega_1 \times \Omega_2, \mathcal{F}_1 \times \mathcal{F}_2, P_1 \times P_2)$  satisfying a), b) and c). We show that  $X \in \mathbf{H}$  for all  $X \geq 0$  by appeal to Theorem 2.22.

To this end, let  $\mathcal{S}$  be the  $\pi$ -system of rectangles  $A = A_1 \times A_2$ , where  $A_1 \in \mathcal{F}_1$  and  $A_2 \in \mathcal{F}_2$ , which generates  $\mathcal{F}_1 \times \mathcal{F}_2$  by Definition 3.42. For  $X = \mathbf{1}_A$ , with  $A = A_1 \times A_2 \in \mathcal{S}$ ,

$$Y_1(\omega_1) = \mathbf{1}_{A_1}(\omega_1)P_2(A_2),$$

which is a simple random variable on  $\Omega_1, \mathcal{F}_1, P_1$ , and, moreover,

$$E_{P_1}[Y_1] = P_1(A_1)P_2(A_2).$$

In the same way,  $Y_2(\omega_2) = \mathbf{1}_{A_2}(\omega_2)P_1(A_1)$  and

$$E_{P_2}[Y_2] = P_1(A_1)P_2(A_2).$$

Finally,

$$E_{P_1 \times P_2}[X] = P_1 \times P_2(A_1 \times A_2) = P_1(A_1)P_2(A_2),$$

and, hence, a), b) and c) are fulfilled.

That  $\mathbf{H}$  satisfies the hypotheses of Theorem 2.22 holds true by Theorem 2.16, linearity of expectation and the monotone convergence theorem, the latter two applied to  $P_1$ ,  $P_2$  and  $P_1 \times P_2$ . ■

The version for integrable random variables is more complicated to state, but its content is essentially the same. This is a case when only “almost sure” conclusions can be drawn.

**Theorem 4.50 (Fubini's theorem, bis).** Suppose that  $X$  belongs to  $L^1(P_1 \times P_2)$ . Then,

- a) For  $P_1$ -almost every  $\omega_1$ , the random variable  $X(\omega_1, \cdot)$  belongs to  $L^1(P_2)$ , and the function

$$Y_1(\omega_1) = \begin{cases} E_{P_2}[X(\omega_1, \cdot)] & \text{if } X(\omega_1, \cdot) \in L^1(P_2) \\ 0 & \text{otherwise} \end{cases}$$

is a random variable belonging to  $L^1(P_1)$ .

- b) For  $P_2$ -almost every  $\omega_2$ , the random variable  $X(\cdot, \omega_2)$  belongs to  $L^1(P_1)$ , and the function

$$Y_2(\omega_2) = \begin{cases} E_{P_1}[X(\cdot, \omega_2)] & \text{if } X(\cdot, \omega_2) \in L^1(P_1) \\ 0 & \text{otherwise} \end{cases}$$

is a random variable belonging to  $L^1(P_2)$ .

- c)  $E_{P_1}[Y_1] = E_{P_1 \times P_2}[X] = E_{P_2}[Y_2]$ . □

## 4.7 Exercises

**4.1.** Verify the entries in Table 4.3.

**4.2.** A point  $m$  is a *median* of  $X$  if  $P\{X \geq m\} \geq 1/2$  and  $P\{X \leq m\} \geq 1/2$ . Prove that if  $m_0$  is a median of  $X \in L^1$ , then

$$E[|X - m_0|] \leq E[|X - a|]$$

for all  $a \in \mathbb{R}$ .

**4.3.** Give a direct proof that for a positive, integer-valued random variable  $X$ ,

$$\sum_{n=1}^{\infty} nP\{X = n\} = \sum_{k=1}^{\infty} P\{X \geq k\}.$$

**4.4.** Suppose that  $X, Y \in L^2$  and equality holds in the Cauchy-Schwarz inequality:  $E[|XY|] = \sqrt{E[X^2]E[Y^2]}$ . Prove that there is a constant  $c$  such that  $X \stackrel{\text{a.s.}}{=} cY$ . Conclude from this that if  $X, Y \in L^2$  and  $|\text{Corr}(X, Y)| = 1$ , then there are  $a, b \in \mathbb{R}$  such that  $Y \stackrel{\text{a.s.}}{=} aX + b$ .

**4.5.** Let  $F$  be a distribution function on  $\mathbb{R}_+$  with  $F(0) < 1$  and finite mean  $m$ . Prove that the function

$$H_F(t) = \frac{1}{m} \int_0^t [1 - F(y)] dy$$

is a distribution function, and calculate its mean and variance.

**4.6.** Calculate  $E[1/(X + 1)]$  when  $X \stackrel{d}{=} P(\lambda)$  and when  $X \stackrel{d}{=} B(n, p)$ .

**4.7.** Modify Example 4.3 to produce positive random variables  $X_n$  such that  $E[\liminf_n X_n] = 0$  but  $\liminf_n E[X_n] = \infty$ .

**4.8.** Let  $X, Y$  and  $Z$  be independent with distribution  $U[0, 1]$ . Calculate the probability that roots of the quadratic equation  $Xt^2 + Yt + Z = 0$  are real.

**4.9.** Prove that for  $X \geq 0$ ,

$$\sum_{k=1}^{\infty} P\{X > k\} \leq E[X] \leq \sum_{k=0}^{\infty} P\{X > k\}. \quad (4.36)$$

**4.10.** Let  $X_1, X_2, \dots$  be independent with continuous distribution function  $F$ , and for each  $n$ , let  $A_n = \{X_n > \max_{k < n} X_k\}$  be the event that a *record* occurs at time  $n$ , in the sense that  $X_n$  exceeds each of  $X_1, \dots, X_{n-1}$ .

- a) Prove that  $A_1, A_2, \dots$  are independent, and that  $P\{A_n\} = 1/n$  for each  $n$ .
- b) Prove that  $P\{A_n, \text{i.o.}\} = 1$ .
- c) Let  $N_n$  be the time of the next record after  $n$ . Show that for  $k \geq 1$ ,

$$P\{N_n = n + k\} = \frac{n}{(n+k-1)(n+k)}$$

and use this to show that  $E[N_n] = \infty$ .

- 4.11.** Let  $Z_1, \dots, Z_k$  be i.i.d. with  $P\{Z_i = j\} = 1/J$ ,  $j = 1, \dots, J$ , where  $J$  is fixed, and let  $X$  be the number of distinct values among  $Z_1, \dots, Z_k$ . Prove that  $X$  is a random variable and calculate  $E[X]$ . [Hint: This can be done without computing the distribution of  $X$ .]

- 4.12.** Let  $\{A_1, \dots, A_n\}$  be a finite partition of  $\Omega$ . Suppose that we know which of  $A_1, \dots, A_n$  has occurred, and wish to predict whether some other event  $B$  has occurred. Since we know the values of  $\mathbf{1}_{A_1}, \dots, \mathbf{1}_{A_n}$ , it makes sense to use a predictor that is a function of them. Suppose that we confine ourselves to *linear* predictors of the form  $Y = \sum_{i=1}^n a_i \mathbf{1}_{A_i}$ , whose accuracy we assess via the virtually universal criterion of *mean squared error*:

$$\text{MSE}(Y) = E[(\mathbf{1}_B - Y)^2].$$

Determine the values of  $a_1, \dots, a_n$  that minimize  $\text{MSE}(Y)$ .

- 4.13.** Prove that Theorem 4.10 implies Fatou's lemma (Theorem 4.8).  
**4.14.** Give an example of random variables  $X$  and  $Y$  that are uncorrelated but for which there exists a function  $g$  such that  $Y = g(X)$ .  
**4.15.** Prove that Proposition 4.43 remains valid if  $X$  and  $Y$  are only uncorrelated, rather than independent.

- 4.16.** Let  $X$  have finite variance. Show that

$$\arg \min_{a \in \mathbb{R}} E[(X - a)^2] = E[X]$$

and that the minimum value of  $E[(X - a)^2]$  is  $\text{Var}(X)$ .

- 4.17.** Prove that there *do not exist* random variables  $X$ ,  $Y$  and  $Z$  such that  $\text{Corr}(X, Y) = \text{Corr}(Y, Z) = \text{Corr}(Z, X) = -1$ .  
**4.18.** Prove that if  $V, W \in L^2$  and  $(V, W) \stackrel{\text{d}}{=} (-V, W)$ , then  $V$  and  $W$  are uncorrelated.  
**4.19.** Prove that if  $\text{Var}(X) = 0$ , then  $X \stackrel{\text{a.s.}}{=} E[X]$ .

**4.20.** Let  $A$  be an event with  $P(A) > 0$ . Show that if  $X$  is positive or integrable, then  $E[X|A]$ , the expectation of  $X$  with respect to the probability  $P_A(B) = P(B|A)$ , is given by  $E[X|A] \stackrel{\text{def}}{=} E[X; A]/P(A)$ .

**4.21.** Verify (4.33) and (4.34).

**4.22.** Prove that for  $X \in L^2$  and  $a, b \in \mathbb{R}$ ,  $\text{Var}(aX + b) = a^2\text{Var}(X)$ .

**4.23.** Prove that for  $X, Y \in L^2$ ,

$$\text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y).$$

**4.24.** Prove that a  $E[|X|^p] < \infty$  if and only if  $\sum_{k=1}^{\infty} k^{p-1}P\{|X| \geq k\} < \infty$ .

**4.25.** Suppose  $X \stackrel{d}{=} P(\lambda)$ .

- a) Use Chebyshev's inequality to show that  $P\{X \geq 2\lambda\} \leq 1/\lambda$ .
- b) Show that  $P\{X \geq 2\lambda\} \leq (e/4)^{\lambda}$ , and compare this bound to that in part a).

**4.26.** Let  $X_1, \dots, X_n$  be independent with distribution  $U[0, 1]$ . Let  $M^* = \max\{X_1, \dots, X_n\}$  and  $M^{**} = \min\{X_1, \dots, X_n\}$ . Calculate  $E[M^*]$ ,  $\text{Var}(M^*)$  and  $\text{Corr}(M^*, M^{**})$ .

**4.27.** Let  $X_1, \dots, X_n$  be i.i.d. with  $E[|X_1|^3] < \infty$ . Show that

$$\begin{aligned} E\left[\left(\sum_{i=1}^n X_i\right)^3\right] &= nE[X_1^3] + 3n(n-1)E[X_1^2]E[X_1] \\ &\quad + n(n-1)(n-2)E[X_1]^3. \end{aligned}$$

**4.28.** Let  $X_1, \dots, X_n$  be i.i.d. with  $E[X_1] = E[X_1^3] = 0$ . Prove that the *sample mean*

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

and *sample variance*

$$S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

are uncorrelated.

**4.29.** In the context of Exercise 4.28, prove that if the  $X_i \stackrel{d}{=} N(0, 1)$  for each  $i$ , then  $\bar{X}$  and  $S^2$  are independent.

**4.30.** Prove *Cantelli's inequality*: if  $\text{Var}(X) < \infty$ , then for  $a > 0$ ,

$$P\{|X - E[X]| > a\} \leq \frac{2\text{Var}(X)}{a^2 + \text{Var}(X)}.$$

When is this a better bound than Chebyshev's inequality?

**4.31.** Let  $X_1, X_2, \dots$  be i.i.d. and assume only strictly positive integer values, and suppose that  $a = E[X_1]$  and  $b = E[1/X_1]$  are finite. Let  $S_m = \sum_{i=1}^m X_i$ .

- a) Show that for  $m \leq n$ ,  $E[S_m/S_n] = m/n$ .
- b) Show that  $E[1/S_n] < \infty$  for each  $n$ , and that for  $m > n$ ,

$$E[S_m/S_n] = 1 + (m-n)aE[1/S_n].$$

- c) Show that  $E[S_m/S_n] \geq m/n$  for all  $m$  and  $n$ .

**4.32.** Suppose that  $E[X^2] < \infty$ , and for each  $c$ , let  $X^c = \min\{X, c\}$ . Prove that  $\text{Var}(X^c) \leq \text{Var}(X)$ .

**4.33.** Let  $X_1, X_2, \dots$  be independent and Bernoulli distributed with parameter  $p$ , and suppose that  $N$  is geometrically distributed with parameter  $r$  and independent of the  $X_i$ . Let  $S = \sum_{i=1}^N X_i$ .

- a) Calculate  $P\{S = k\}$  for each  $k$ .
- b) Compute  $\text{Cov}(S, N)$ .

**4.34.** The lifetimes of  $n$  computer systems are assumed to be independent and exponentially distributed with mean  $\theta$ .

- a) Calculate the density function of  $L$ , the lifetime of the system that survives the longest.
- b) Show that  $E[L] = \theta \sum_{i=1}^n (1/i)$  and  $\text{Var}(L) = \theta^2 \sum_{i=1}^n (1/i^2)$ .
- c) Show that the times between failures are independent.

**4.35.** Without evaluating the expectation, show that if  $X \stackrel{d}{=} U[0, 2]$ , then  $E[X \log X] \geq 0$ .

**4.36.** Show that given  $a > 0$ , there exists a random variable  $X$  taking values in  $\{-1, 0, 1\}$ , such that

$$P\{|X - E[X]| \geq a\} = \frac{\text{Var}(X)}{a^2}.$$

What does this imply about possible improvements to Chebyshev's inequality?

**4.37.** Show that for  $X_i$  and  $Y_j$  belonging to  $L^2$  and  $a_i, b_j \in \mathbb{R}$ ,

$$\text{Cov}\left(\sum_{i=1}^n a_i X_i, \sum_{j=1}^m b_j Y_j\right) = \sum_{i=1}^n \sum_{j=1}^m a_i b_j \text{Cov}(X_i, Y_j).$$

**4.38.** For each of the four occupancy models in §3.5, calculate the mean and variance of the number of empty cells.

- 4.39.** Consider  $n$  independent repetitions of an experiment with  $r$  possible outcomes having probabilities  $p_1, \dots, p_r$ . For  $j = 1, \dots, r$ , let  $Z_j$  be the number of experiments resulting in outcome  $j$ . Calculate the mean vector and covariance matrix of the random vector  $(Z_1, \dots, Z_r)$ .

- 4.40.** Suppose that  $X, Y \in L^2$ .

- a) For each  $\theta \in [0, 2\pi]$ , define

$$X_\theta = X \cos \theta - Y \sin \theta$$

$$Y_\theta = X \sin \theta + Y \cos \theta.$$

Show that there is at least one value of  $\theta$  for which  $X_\theta$  and  $Y_\theta$  are uncorrelated.

- b) Show that if  $|\text{Corr}(X, Y)| < 1$  there is a linear transformation  $L$  such that  $(X', Y') = L(X, Y)$  satisfies  $\text{Var}(X') = \text{Var}(Y') = 1$  and  $\text{Corr}(X', Y') = 0$ .

- 4.41.** Suppose that  $E[X] = E[Y] = 0$  and  $\text{Var}(X) = \text{Var}(Y) = 1$ , and let  $\rho = \text{Corr}(X, Y)$ . Assume that  $0 < \rho < 1$ . Determine all values of  $a$  and  $b$  for which  $X - aY$  and  $Y - bX$  are uncorrelated, and plot them.

- 4.42.** Suppose that  $X$  and  $Y$  are either both positive or both integrable, and that  $E[X; A] = E[Y; A]$  for all  $A \in \mathcal{F}$ . Prove that  $X \stackrel{\text{a.s.}}{=} Y$ .

# Chapter 5

# Convergence of Random Variables

In this chapter, we develop the tools needed to describe and apply convergence of sequences of random variables. Almost sure convergence, because of its relationship to pointwise convergence, and convergence in distribution, because of its being the easiest to establish, are the most important and useful. Both generalize meaningfully and naturally to random vectors. Convergence in probability is significant for weak laws of large numbers, and in statistics, for certain forms of consistency. Quadratic mean convergence and  $L^1$  convergence are used to establish convergence of moments, as well as in martingale theory (Chapter 9).

## 5.1 Modes of Convergence

Let  $X, X_1, X_2, \dots$  be random variables on  $(\Omega, \mathcal{F}, P)$ . Four of the modes of convergence pertain to the  $X_n$  and  $X$  as functions on  $\Omega$ , while the fifth has to do with convergence of distribution functions.

### 5.1.1 Convergence of random variables as functions

Almost sure convergence, also known as convergence with probability one, is the probabilistic version of pointwise convergence.

**Definition 5.1.** The sequence  $(X_n)$  converges to  $X$  *almost surely*, denoted by  $X_n \xrightarrow{\text{a.s.}} X$ , if

$$P(\{\omega: X_n(\omega) \rightarrow X(\omega)\}) = 1. \quad \square$$

Convergence in probability means that the probability that  $|X_n - X|$  exceeds any prescribed, strictly positive value converges to zero.

**Definition 5.2.** The sequence  $(X_n)$  converges to  $X$  in probability, denoted by  $X_n \xrightarrow{P} X$ , if

$$\lim_{n \rightarrow \infty} P\{|X_n - X| > \varepsilon\} = 0$$

for every  $\varepsilon > 0$ .  $\square$

Quadratic mean convergence, which is also referred to as mean square convergence and  $L^2$  convergence, requires that  $E[(X_n - X)^2] \rightarrow 0$ . Recall that  $L^2$  is the vector space of random variables  $X$  for which  $E[X^2] < \infty$ .

**Definition 5.3.** Suppose that  $X, X_1, X_2, \dots$  belong to  $L^2$ . Then,  $(X_n)$  converges to  $X$  in quadratic mean if

$$\lim_{n \rightarrow \infty} E[(X_n - X)^2] = 0.$$

This is denoted by  $X_n \xrightarrow{\text{q.m.}} X$  or  $X_n \xrightarrow{L^2} X$ .  $\square$

Mean, or  $L^1$ , convergence, is analogous, but with first moments. The space  $L^1$  consists of those random variables  $X$  such that  $E[|X|] < \infty$ .

**Definition 5.4.** For  $X, X_1, X_2, \dots \in L^1$ ,  $(X_n)$  converges to  $X$  in  $L^1$  if

$$\lim_{n \rightarrow \infty} E[|X_n - X|] = 0.$$

We denote this by  $X_n \xrightarrow{L^1} X$ .  $\square$

### 5.1.2 Convergence of distribution functions

The final form of convergence is not convergence of random variables *per se*, but, rather, their distribution functions.

**Definition 5.5.** The sequence  $(X_n)$  converges to  $X$  in distribution if

$$\lim_{n \rightarrow \infty} F_{X_n}(t) = F_X(t)$$

for all  $t$  at which  $F_X$  is continuous. This is denoted by  $X_n \xrightarrow{d} X$ .  $\square$

This definition of convergence in distribution is cumbersome because of the proviso regarding continuity points of the limit distribution function  $F_X$ . In Theorem 5.8 and Chapter 6 we develop more tractable criteria.

Mode	Defining Condition
Almost sure	$P(\{\omega: X_n(\omega) \rightarrow X(\omega)\}) = 1$
In probability	$P\{ X_n - X  > \varepsilon\} \rightarrow 0$ for all $\varepsilon > 0$
Quadratic mean	$E[(X_n - X)^2] \rightarrow 0$
$L^1$	$E[ X_n - X ] \rightarrow 0$
In distribution	$F_{X_n}(t) \rightarrow F_X(t)$ at continuity points $t$ of $F_X$

Table 5.1. Definitions of Convergence for Random Variables

### 5.1.3 Alternative criteria

In several results below, we use the following criterion for “non-convergence” of a real sequence: for  $x, x_1, x_2, \dots \in \mathbb{R}$ ,  $x_n \not\rightarrow x$  if and only if there is  $\varepsilon > 0$  such that  $|x_n - x| > \varepsilon$  for infinitely many values of  $n$ .

We begin with an equivalent form of almost sure convergence that illuminates its relationship to convergence in probability:  $X_n \xrightarrow{\text{a.s.}} X$  if and only if  $\sup_{k \geq n} |X_k - X| \xrightarrow{\text{P}} 0$ .

**Proposition 5.6.** We have  $X_n \xrightarrow{\text{a.s.}} X$  if and only if for every  $\varepsilon > 0$ ,

$$\lim_{n \rightarrow \infty} P\{\sup_{k \geq n} |X_k - X| > \varepsilon\} = 0. \quad (5.1)$$

**Proof:** For  $\varepsilon > 0$ , let  $A^\varepsilon = \{|X_n - X| > \varepsilon, \text{i.o.}\}$ ; then

$$\begin{aligned} P(A^\varepsilon) &= \lim_{n \rightarrow \infty} P(\bigcup_{k=n}^{\infty} \{|X_k - X| > \varepsilon\}) \\ &= \lim_{n \rightarrow \infty} P\{\sup_{k \geq n} |X_k - X| > \varepsilon\}. \end{aligned} \quad (5.2)$$

Necessity: If  $X_n \xrightarrow{\text{a.s.}} X$ , then  $P(A^\varepsilon) = 0$  for each  $\varepsilon$ , so that (5.2) implies that (5.1) holds.

Sufficiency: If (5.1) holds, then by (5.2),  $P(A^\varepsilon) = 0$  for each  $\varepsilon$ , and, hence, by Boole’s inequality (Proposition 1.24),

$$P(\{\omega: X_n(\omega) \not\rightarrow X(\omega)\}) = P\left(\bigcup_{m=1}^{\infty} A^{1/m}\right) \leq \sum_{m=1}^{\infty} P(A^{1/m}) = 0. \quad \blacksquare$$

The next result presents a form of convergence, termed *complete convergence*, which is stronger than almost sure convergence, and sometimes more convenient to establish.

**Proposition 5.7.** If  $\sum_{n=1}^{\infty} P\{|X_n - X| > \varepsilon\} < \infty$  for every  $\varepsilon > 0$ , then  $X_n \xrightarrow{\text{a.s.}} X$ .

**Proof:** We retain the notation in the proof of Proposition 5.6. By the Borel-Cantelli lemma (Theorem 1.27),  $P(A^\varepsilon) = 0$  for every  $\varepsilon$ , in light of which almost sure convergence follows. ■

Thus,  $X_n \xrightarrow{P} X$  if and only if the probabilities  $P\{|X_n - X| > \varepsilon\}$  converge to 0, while  $X_n \xrightarrow{\text{a.s.}} X$  if (but not only if) the convergence is fast enough that their sum is finite.

The following criterion for convergence in distribution is superior to Definition 5.5 in that one need not deal with continuity points of the limit distribution function. Denote by  $\mathbf{C}$  the set of bounded, continuous functions  $f: \mathbb{R} \rightarrow \mathbb{R}$ .

**Theorem 5.8.** We have  $X_n \xrightarrow{d} X$  if and only if

$$E[f(X_n)] \rightarrow E[f(X)]$$

for every  $f \in \mathbf{C}$ .

**Proof:** We employ two forms of approximation: of bounded, continuous functions by step functions (linear combinations of indicator functions of intervals), and of indicator functions of intervals by functions in  $\mathbf{C}$ .

Sufficiency: Suppose that  $X_n \xrightarrow{d} X$ . Given  $f \in \mathbf{C}$ , let  $M = \sup_x |f(x)|$ , which is finite. Then, given  $\varepsilon > 0$ , choose  $K > 0$  such that  $\pm K$  are continuity points of  $F_X$  and such that  $P\{|X| > K\} < \varepsilon/M$ . This is possible since  $F_X$  has at most countably discontinuities and since  $\lim_{x \rightarrow \infty} P\{|X| > x\} = 0$ . By convergence in distribution,  $P\{|X_n| > K\} < 2\varepsilon/M$  for all large values of  $n$ . With  $K$  and  $\varepsilon$  remaining fixed, there is a step function  $g = \sum_{i=1}^k a_i \mathbf{1}_{(x_{i-1}, x_i]}$ , with  $-K = x_0 < \dots < x_k = K$ , such that each  $x_i$  is a continuity point of  $F_X$  and  $\sup_{x \in [-K, K]} |f(x) - g(x)| < \varepsilon$ . Then, for  $n$  sufficiently large,

$$\begin{aligned} & |E[f(X_n)] - E[f(X)]| \\ & \leq |E[f(X_n); \{|X_n| \leq K\}] - E[f(X); \{|X| \leq K\}]| \\ & \quad + E[|f(X_n)|; \{|X_n| > K\}] + E[|f(X)|; \{|X| > K\}] \\ & \leq \left| E[f(X_n); \{|X_n| \leq K\}] - E[f(X); \{|X| \leq K\}] \right| + 3\varepsilon \\ & \leq 3\varepsilon + |E[f(X_n); \{|X_n| \leq K\}] - E[g(X_n)]| \\ & \quad + |E[f(X); \{|X| \leq K\}] - E[g(X)]| \\ & \quad + |E[g(X_n)] - E[g(X)]| \\ & \leq 5\varepsilon + |E[g(X_n)] - E[g(X)]|. \end{aligned}$$

But, because  $X_n \xrightarrow{d} X$  and the  $x_i$  are continuity points of  $\mathbb{F}_X$ ,

$$\begin{aligned} E[g(X_n)] &= \sum_{i=1}^k a_i [\mathbb{F}_{X_n}(x_i) - \mathbb{F}_{X_n}(x_{i-1})] \\ &\rightarrow \sum_{i=1}^k a_i [\mathbb{F}_X(x_i) - \mathbb{F}_X(x_{i-1})] \\ &= E[g(X)], \end{aligned}$$

and, hence,  $E[f(X_n)] \rightarrow E[f(X)]$ .

Necessity: Suppose that  $E[f(X_n)] \rightarrow E[f(X)]$  for every  $f \in \mathbf{C}$ . Given a continuity point  $t$  of  $\mathbb{F}_X$  and  $\varepsilon > 0$ , there exists  $f \in \mathbf{C}$  such that  $\mathbf{1}_{(-\infty, t]} \leq f \leq \mathbf{1}_{(-\infty, t+\varepsilon]}$  (for example, one may take  $f(x) = 1$  for  $x \leq t$ ,  $f(x) = 0$  for  $x \geq t + \varepsilon$  and  $f$  to be linear between  $t$  and  $t + \varepsilon$ ). Then, since

$$\mathbb{F}_{X_n}(t) = E[\mathbf{1}_{(-\infty, t]}(X_n)] \leq E[f(X_n)]$$

and

$$E[f(X)] \leq E[\mathbf{1}_{(-\infty, t+\varepsilon)}(X)] = \mathbb{F}_X(t + \varepsilon),$$

letting  $n \rightarrow \infty$  gives

$$\begin{aligned} \limsup_{n \rightarrow \infty} \mathbb{F}_{X_n}(t) &\leq \lim_{n \rightarrow \infty} E[f(X_n)] = E[f(X)] \leq E[\mathbf{1}_{(-\infty, t+\varepsilon)}(X)] \\ &= \mathbb{F}_X(t + \varepsilon), \end{aligned}$$

where the first equality is by Theorem 5.8. Then, letting  $\varepsilon \downarrow 0$  and invoking right-continuity of  $\mathbb{F}_X$  yields

$$\limsup_{n \rightarrow \infty} \mathbb{F}_{X_n}(t) \leq \mathbb{F}_X(t).$$

Next, given  $\varepsilon$ , let  $f \in \mathbf{C}$  be such that  $\mathbf{1}_{(-\infty, t-\varepsilon]} \leq f \leq \mathbf{1}_{(-\infty, t]}$ . Then,

$$\mathbb{F}_X(t - \varepsilon) \leq E[f(X)] = \lim_{n \rightarrow \infty} E[f(X_n)] \leq \liminf_{n \rightarrow \infty} \mathbb{F}_{X_n}(t),$$

and consequently, because  $t$  is a continuity point of  $\mathbb{F}_X$ ,

$$\mathbb{F}_X(t) = \mathbb{F}_X(t-) \leq \liminf_{n \rightarrow \infty} \mathbb{F}_{X_n}(t).$$

Thus,

$$\mathbb{F}_X(t) \leq \liminf_{n \rightarrow \infty} \mathbb{F}_{X_n}(t) \leq \limsup_{n \rightarrow \infty} \mathbb{F}_{X_n}(t) \leq \mathbb{F}_X(t),$$

which completes the proof that  $\mathbb{F}_{X_n}(t) \rightarrow \mathbb{F}_X(t)$ . ■

In the proof of Theorem 5.8, the continuous functions used to approximate indicator functions can be taken to be arbitrarily smooth.

**Corollary 5.9.** Let  $k$  be a fixed integer. If  $E[f(X_n)] \rightarrow E[f(X)]$  for every function  $f$  belonging to the space  $\mathbf{C}^{(k)}$  of bounded,  $k$ -times uniformly continuously differentiable functions from  $\mathbb{R}$  to  $\mathbb{R}$ , then  $X_n \xrightarrow{d} X$ . □

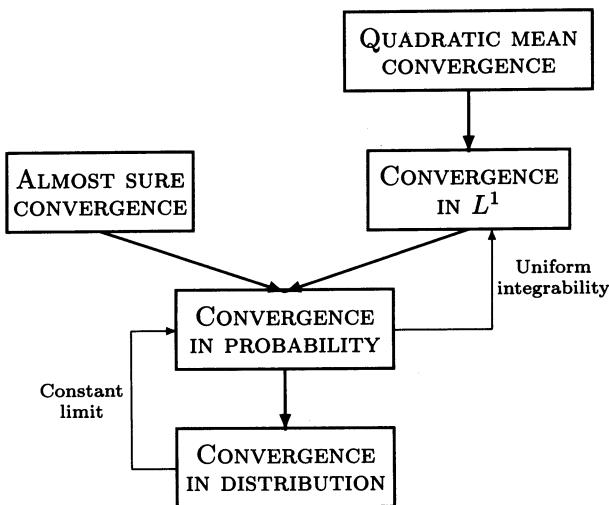


Figure 5.1. Implications Among Forms of Convergence

## 5.2 Relationships Among the Modes

Given the plethora of definitions in §1, it is natural to wonder about implications among them.

### 5.2.1 Implications always valid

Figure 5.1 depicts the implications that are always valid, and two of limited validity as well. None of the other implications holds in general.

**Proposition 5.10.** If  $X_n \xrightarrow{\text{a.s.}} X$ , then  $X_n \xrightarrow{P} X$ .

**Proof:** Suppose that  $\varepsilon > 0$ , and for each  $n$ , let  $A_n = \{|X_n - X| > \varepsilon\}$ . Then, convergence in probability follows:

$$\lim_{m \rightarrow \infty} P(A_m) \leq \lim_{m \rightarrow \infty} P(\bigcup_{k=m}^{\infty} A_k) = P\{A_n, \text{i.o.}\} = 0. \quad \blacksquare$$

**Proposition 5.11.** If  $X_n \xrightarrow{\text{q.m.}} X$ , then  $X_n \xrightarrow{L^1} X$ .

**Proof:** By the Cauchy-Schwarz inequality (Corollary 4.36),

$$E[|X_n - X|] \leq \sqrt{E[(X_n - X)^2]},$$

so that  $E[|X_n - X|] \rightarrow 0$  if  $E[(X_n - X)^2] \rightarrow 0$ .  $\blacksquare$

**Proposition 5.12.** If  $X_n \xrightarrow{L^1} X$ , then  $X_n \xrightarrow{P} X$ .

**Proof:** By Chebyshev's inequality (4.28), for each  $\varepsilon > 0$ ,

$$P\{|X_n - X| > \varepsilon\} \leq \frac{1}{\varepsilon} E[|X_n - X|].$$

Therefore,  $E[|X_n - X|] \rightarrow 0$  implies that  $P\{|X_n - X| > \varepsilon\} \rightarrow 0$ . ■

**Proposition 5.13.** If  $X_n \xrightarrow{P} X$ , then  $X_n \xrightarrow{d} X$ .

**Proof:** Let  $t$  be a continuity point of  $F_X$ . Then, for  $\varepsilon > 0$  and each  $n$ ,

$$\begin{aligned} F_{X_n}(t) &= P\{X_n \leq t\} \\ &= P\{X_n \leq t, |X_n - X| \leq \varepsilon\} + P\{X_n \leq t, |X_n - X| > \varepsilon\} \\ &\leq P\{X \leq t + \varepsilon\} + P\{|X_n - X| > \varepsilon\} \\ &= F_X(t + \varepsilon) + P\{|X_n - X| > \varepsilon\}. \end{aligned}$$

Letting  $n \rightarrow \infty$  gives  $\limsup_n F_{X_n}(t) \leq F_X(t + \varepsilon)$  by convergence in probability, and then, letting  $\varepsilon \downarrow 0$  yields  $\limsup_n F_{X_n}(t) \leq F_X(t)$ , where we have also used right-continuity of  $F_X$ .

In the same way, for each  $\varepsilon$  and  $n$ ,

$$\begin{aligned} F_X(t - \varepsilon) &= P\{X \leq t - \varepsilon\} \\ &= P\{X \leq t - \varepsilon, |X_n - X| \leq \varepsilon\} + P\{X \leq t - \varepsilon, |X_n - X| > \varepsilon\} \\ &\leq F_{X_n}(t) + P\{|X_n - X| > \varepsilon\}. \end{aligned}$$

Letting  $n \rightarrow \infty$  and then  $\varepsilon \downarrow 0$ , and invoking both convergence in probability and the assumed continuity of  $F_X$  at  $t$ , gives  $F_X(t) \leq \liminf_n F_{X_n}(t)$ , which shows that  $F_{X_n}(t) \rightarrow F_X(t)$ . ■

### 5.2.2 Counterexamples

Counterexamples to all other implications among the modes of convergence can be constructed on the probability space  $([0, 1], \mathcal{B}([0, 1]), P)$ , where  $P$  is the uniform distribution.

- a) The sequence  $X_n = n \mathbf{1}_{(0,1/n)}$  converges to zero almost surely, and, hence, in probability and in distribution, but not in  $L^1$ , since  $E[X_n] = n P\{X_n = n\} = 1$  for each  $n$ , and so also not in quadratic mean.
- b) The sequence  $X_n = \sqrt{n} \mathbf{1}_{(0,1/n)}$  converges to zero almost surely and in  $L^1$ , since  $E[X_n] = 1/\sqrt{n}$ , but not in quadratic mean, because  $E[X_n^2] = 1$  for every  $n$ .
- c) The sequence  $X_1 = 2 \cdot \mathbf{1}_{[0,1/2)}, X_2 = 2 \cdot \mathbf{1}_{[1/2,1)}, X_3 = 3 \cdot \mathbf{1}_{[0,1/3)}, X_4 = 3 \cdot \mathbf{1}_{[1/3,2/3)}, X_5 = 3 \cdot \mathbf{1}_{[2/3,1)}, \dots$  converges to zero in probability, but neither almost surely ( $\liminf_n X_n \stackrel{\text{a.s.}}{\equiv} 0$ , while  $\limsup_n X_n \stackrel{\text{a.s.}}{\equiv} \infty$ ) nor in  $L^1$ , since  $E[X_n] = 1$  for all  $n$ .

- d) The sequence  $X_n = \mathbf{1}_{[0,1/2+1/n]}$  converges to  $X = \mathbf{1}_{[1/2,1]}$  only in distribution. Indeed,

$$\mathsf{F}_{X_n}(t) = \begin{cases} 0 & t < 0 \\ 1/2 - 1/n & 0 \leq t < 1 \\ 1 & t \geq 1, \end{cases}$$

does converge pointwise to

$$\mathsf{F}_X(t) = \begin{cases} 0 & t < 0 \\ 1/2 & 0 \leq t < 1 \\ 1 & t \geq 1. \end{cases}$$

On the other hand,  $P\{|X_n - X| > 1/2\} = 1 - 1/n \rightarrow 1$ , so that there is no convergence in probability.

- e) If  $X_n \stackrel{\text{d}}{=} U[1/2 - 1/n, 1/2 + 1/n]$  and  $X \stackrel{\text{a.s.}}{=} 1/2$ , then  $X_n \xrightarrow{\text{d}} X$ , even though  $\mathsf{F}_{X_n}(1/2) = 1/2$  for each  $n$ , and these values do not converge to  $\mathsf{F}_X(1/2) = 1$ . There is no contradiction because  $\mathsf{F}_X$  is not continuous at  $1/2$ .

### 5.2.3 Implications of restricted validity

“Implications of restricted validity” hold true in the presence of additional assumptions satisfied in many cases of interest. We present two: convergence in distribution to a constant random variable implies convergence in probability, and convergence in probability and uniform integrability imply convergence in  $L^1$ .

**Proposition 5.14.** If  $X_n \xrightarrow{\text{d}} c$ , then  $X_n \xrightarrow{\text{P}} c$ .

**Proof:** For  $\varepsilon > 0$  and each  $n$ ,

$$\begin{aligned} P\{|X_n - c| > \varepsilon\} &= P\{X_n < c - \varepsilon\} + P\{X_n > c + \varepsilon\} \\ &\leq \mathsf{F}_{X_n}(c - \varepsilon) + [1 - \mathsf{F}_{X_n}(c + \varepsilon)] \\ &\rightarrow \mathsf{F}_c(c - \varepsilon) + [1 - \mathsf{F}_c(c + \varepsilon)] \\ &= 0, \end{aligned}$$

since  $c - \varepsilon$  and  $c + \varepsilon$  are both continuity points of  $\mathsf{F}_c$ . ■

Uniform integrability is used to deduce  $L^1$  convergence from convergence in probability.

**Definition 5.15.** A sequence  $(X_n)$  is *uniformly integrable* if  $X_n \in L^1$  for each  $n$  and if

$$\lim_{a \rightarrow \infty} \sup_n E[|X_n|; \{|X_n| > a\}] = 0. \quad \square \quad (5.3)$$

The “uniformity” in the definition arises from the property that if  $Y \in L^1$ , then

$$\lim_{a \rightarrow \infty} E[|Y|; \{|Y| > a\}] = 0$$

by the dominated convergence theorem (Theorem 4.16), since the random variables  $|Y| \mathbf{1}(|Y| > a)$  are dominated by  $Y \in L^1$  and converge to zero. Indeed, any sequence  $(X_n)$  dominated by an element of  $L^1$  is uniformly integrable.

The next result characterizes uniform integrability.

**Proposition 5.16.** The sequence  $(X_n)$  is uniformly integrable if and only if

i)  $\sup_n E[|X_n|] < \infty$

ii)  $(X_n)$  is *uniformly absolutely continuous*: for each  $\varepsilon > 0$  there is  $\delta > 0$  such that

$$\sup_n E[|X_n|; A] < \varepsilon$$

whenever  $P(A) < \delta$ .

**Proof:** Necessity: If  $(X_n)$  is uniformly integrable, then for each  $a > 0$ ,

$$\sup_n E[|X_n|] \leq \sup_n (E[|X_n|; \{|X_n| > a\}] + a),$$

which is finite. To prove uniform absolute continuity, given  $\varepsilon > 0$ , first choose  $a$  such that

$$\sup_n E[|X_n|; \{|X_n| > a\}] < \varepsilon/2,$$

and then, choose  $\delta$  such that  $a\delta < \varepsilon/2$ . Then, if  $P(A) < \delta$ ,

$$E[|X_n|; A] \leq E[|X_n|; \{|X_n| > a\}] + aP(A) \leq \varepsilon.$$

Sufficiency: Let  $\varepsilon > 0$  be fixed, and let  $\delta$  be chosen as in the statement of the proposition. By Chebyshev's inequality,

$$\sup_n P\{|X_n| > a\} \leq \sup_n E[|X_n|]/a,$$

which is at most  $\delta$  if  $a$  is large enough, but then, uniform absolute continuity gives  $\sup_n E[|X_n|; \{|X_n| > a\}] < \varepsilon$ , which proves uniform integrability. ■

Convergence in probability, in the presence of uniform integrability, entails  $L^1$  convergence. The hypotheses of the following theorem are weaker in two respects than those of the dominated convergence theorem: only convergence in probability is assumed, and the domination condition is relaxed

to uniform integrability. Since  $X_n \xrightarrow{L^1} X$  implies  $E[X_n] \rightarrow E[X]$ , Theorem 5.17 provides another set of conditions under which the expectation of the limit is the limit of the expectations.

**Theorem 5.17.** For  $X, X_1, X_2, \dots \in L^1$ , the following are equivalent:

- a)  $X_n \xrightarrow{P} X$  and  $(X_n)$  is uniformly integrable.
- b)  $X_n \xrightarrow{L^1} X$ .

**Proof:** a)  $\Rightarrow$  b): For each  $n$  and  $\varepsilon$ ,

$$\begin{aligned} E[|X_n - X|] &\leq E[|X_n - X|; \{|X_n - X| \leq \varepsilon\}] \\ &\quad + E[|X_n - X|; \{|X_n - X| > \varepsilon\}] \\ &\leq \varepsilon + E[|X_n|; \{|X_n - X| > \varepsilon\}] + E[|X|; \{|X_n - X| > \varepsilon\}]. \end{aligned}$$

As  $n \rightarrow \infty$ , the second term converges to zero by Proposition 5.16, which is applicable because  $P\{|X_n - X| > \varepsilon\} \rightarrow 0$  by convergence in probability. The third term converges to zero because  $X \in L^1$ , since this means that  $\{X\}$  is uniformly integrable.

b)  $\Rightarrow$  a): That  $X_n \xrightarrow{P} X$  follows from Proposition 5.12. To verify uniform integrability, we apply Proposition 5.16. Since  $E[|X_n - X|] \rightarrow 0$ ,

$$\sup_n E[|X_n|] \leq \sup_n E[|X_n - X|] + E[|X|] < \infty.$$

Given  $\varepsilon > 0$  choose  $N$  such that  $E[|X_n - X|] < \varepsilon$  for  $n \geq N$ . Fixing  $N$  as well, choose  $\delta > 0$  such that whenever  $P(A) < \delta$ ,  $\sup_{n \leq N} E[|X_n|; A] < \delta$  and  $E[|X|; A] < \delta$ . Then, whenever  $P(A) < \delta$  and  $n$  is large enough,

$$\sup_n E[|X_n|; A] \leq \sup_n E[|X_n - X|; A] + E[|X|; A] \leq 2\varepsilon. \blacksquare$$

#### 5.2.4 Implications involving subsequences

One final implication is limited in the strength of the conclusion: if  $X_n \xrightarrow{P} X$ , there is a subsequence  $(X_{n'})$  such that  $X_{n'} \xrightarrow{\text{a.s.}} X$ . The converse holds true as well, and is useful for establishing that convergence in probability is preserved under transformations.

**Proposition 5.18.** The sequence  $(X_n)$  converges in probability to  $X$  if and only if each subsequence  $(X_{n'})$  contains a further subsequence  $(X_{n''})$  such that  $X_{n''} \xrightarrow{\text{a.s.}} X$ .

**Proof:** Suppose that  $X_n \xrightarrow{P} X$ ; to prove the proposition it is enough to show that there is a subsequence  $(X_{n'})$  such that  $X_{n'} \xrightarrow{\text{a.s.}} X$ . For each  $n$ , choose  $n'$  so that  $P\{|X_{n'} - X| > 2^{-n}\} \leq 2^{-n}$ . Consequently, for  $\varepsilon > 0$ ,

$$\begin{aligned} \sum_{n'} P\{|X_{n'} - X| > \varepsilon\} &\leq \sum_{n': 2^{-n} > \varepsilon} P\{|X_{n'} - X| > \varepsilon\} \\ &\quad + \sum_{n': 2^{-n} \leq \varepsilon} P\{|X_{n'} - X| > 2^{-n}\}, \end{aligned}$$

which is finite, so that  $X_{n'} \xrightarrow{\text{a.s.}} X$  by Proposition 5.7.

Conversely, to show that  $X_n \xrightarrow{P} X$ , it suffices to show that every subsequence  $(X_{n'})$  contains a further subsequence  $(X_{n''})$  such that  $X_{n''} \xrightarrow{P} X$ , but, in fact,  $(X_{n'})$  contains a subsequence  $(X_{n''})$  with  $X_{n''} \xrightarrow{\text{a.s.}} X$ . ■

## 5.3 Convergence under Transformations

We next consider preservation of convergence under algebraic operations, then under continuous mappings of the random variables.

### 5.3.1 Algebraic operations

Almost sure convergence, convergence in probability, quadratic mean convergence and  $L^1$  convergence are preserved under addition.

**Theorem 5.19.** Let  $X, X_n, Y$  and  $Y_n$  be random variables.

- a) If  $X_n \xrightarrow{\text{a.s.}} X$  and  $Y_n \xrightarrow{\text{a.s.}} Y$ , then  $X_n + Y_n \xrightarrow{\text{a.s.}} X + Y$ .
- b) If  $X_n \xrightarrow{P} X$  and  $Y_n \xrightarrow{P} Y$ , then  $X_n + Y_n \xrightarrow{P} X + Y$ .
- c) If  $X_n \xrightarrow{\text{q.m.}} X$  and  $Y_n \xrightarrow{\text{q.m.}} Y$ , then  $X_n + Y_n \xrightarrow{\text{q.m.}} X + Y$ .
- d) If  $X_n \xrightarrow{L^1} X$  and  $Y_n \xrightarrow{L^1} Y$ , then  $X_n + Y_n \xrightarrow{L^1} X + Y$ .

**Proof:** a) Since if  $X_n(\omega) \rightarrow X(\omega)$  and  $Y_n(\omega) \rightarrow Y(\omega)$ , then necessarily  $(X_n + Y_n)(\omega) \rightarrow (X + Y)(\omega)$ .

- b) For each  $\varepsilon$ ,

$$\begin{aligned} \{|X_n - X| \leq \varepsilon/2\} \cap \{|Y_n - Y| \leq \varepsilon/2\} \\ \subseteq \{(X_n + Y_n) - (X + Y) \leq \varepsilon\}, \end{aligned}$$

which implies that

$$\begin{aligned} P\{|(X_n + Y_n) - (X + Y)| > \varepsilon\} \\ \leq P\{|X_n - X| > \varepsilon/2\} + P\{|Y_n - Y| > \varepsilon/2\}, \end{aligned}$$

and the two terms on the right-hand side converge to zero by assumption.

c) By Minkowski's inequality (Theorem 4.38),

$$E[|(X_n + Y_n) - (X + Y)|] \leq E[|X_n - X|] + E[|Y_n - Y|].$$

d) Also by Minkowski's inequality,

$$\begin{aligned} & \sqrt{E[((X_n + Y_n) - (X + Y))^2]} \\ & \leq \sqrt{E[(X_n - X)^2]} + \sqrt{E[(Y_n - Y)^2]}. \quad \blacksquare \end{aligned}$$

It does not follow from  $X_n \xrightarrow{d} X$  and  $Y_n \xrightarrow{d} Y$  that  $X_n + Y_n \xrightarrow{d} X + Y$ , but this does hold true, however, if one of the limits is a constant.

**Theorem 5.20 (Slutsky's theorem).** If  $X_n \xrightarrow{d} X$  and  $Y_n \xrightarrow{d} c \in \mathbb{R}$ , then  $X_n + Y_n \xrightarrow{d} X + c$ .

**Proof:** By Corollary 5.9, it suffices to show that

$$E[f(X_n + Y_n)] \rightarrow E[f(X + c)]$$

for every bounded, uniformly continuous function  $f$ . By uniform continuity, given  $\varepsilon > 0$ , there is  $\delta > 0$  such that  $|f(x) - f(y)| < \varepsilon$  whenever  $|x - y| < \delta$ . Therefore, with  $M = \sup_x |f(x)|$ , which is finite,

$$\begin{aligned} & |E[f(X_n + Y_n)] - E[f(X + c)]| \\ & \leq E[|f(X_n + Y_n) - f(X_n + c)|; \{|Y_n - c| > \delta\}] \\ & \quad + E[|f(X_n + Y_n) - f(X_n + c)|; \{|Y_n - c| \leq \delta\}] \\ & \quad + |E[f(X_n + c)] - E[f(X + c)]| \\ & \leq 2MP\{|Y_n - c| > \delta\} + \varepsilon + |E[f(X_n + c)] - E[f(X + c)]|. \end{aligned}$$

Since  $Y_n \xrightarrow{P} c$ , the first term converges to zero as  $n \rightarrow \infty$ . To complete the proof it remains to show that  $X_n + c \xrightarrow{d} X + c$ . For  $f$  bounded and continuous, so is  $h(x) = f(x + c)$ . Thus,

$$E[f(X_n + c)] = E[h(X_n)] \rightarrow E[h(X)] = E[f(X + c)],$$

and, hence,  $X_n + c \xrightarrow{d} X + c$  by Theorem 5.8. ■

Convergence almost surely and in probability are preserved under multiplication; so is convergence in distribution, provided that one limit factor is constant. Quadratic mean convergence of products does not hold in general, since  $XY$  need not belong to  $L^2$  when  $X$  and  $Y$  do. However, the

product of random variables in  $L^2$  belongs to  $L^1$ , and  $L^2$  convergence of factors implies  $L^1$  convergence of products.

**Theorem 5.21.** Let  $X, X_n, Y$  and  $Y_n$  be random variables.

- a) If  $X_n \xrightarrow{\text{a.s.}} X$  and  $Y_n \xrightarrow{\text{a.s.}} Y$ , then  $X_n Y_n \xrightarrow{\text{a.s.}} XY$ .
- b) If  $X_n \xrightarrow{P} X$  and  $Y_n \xrightarrow{P} Y$ , then  $X_n Y_n \xrightarrow{P} XY$ .
- c) If  $X_n \xrightarrow{\text{q.m.}} X$  and  $Y_n \xrightarrow{\text{q.m.}} Y$ , then  $X_n Y_n \xrightarrow{L^1} XY$ .

**Proof:** a) This argument is the same “ $\omega$ -wise” reasoning used in Theorem 5.19: if  $X_n(\omega) \rightarrow X(\omega)$  and  $Y_n(\omega) \rightarrow Y(\omega)$ , then  $(X_n Y_n)(\omega) \rightarrow (XY)(\omega)$ .

b) We apply Proposition 5.18. Given the subsequence  $(n')$  of  $\mathbb{N}$ , since  $X_n \xrightarrow{P} X$  and  $Y_n \xrightarrow{P} Y$ , there is a further subsequence  $(n'')$  such that  $X_{n''} \xrightarrow{\text{a.s.}} X$  and  $Y_{n''} \xrightarrow{\text{a.s.}} Y$ , which by a) implies that  $X_{n''} Y_{n''} \xrightarrow{\text{a.s.}} XY$ . But then, another application of Proposition 5.18 gives that  $X_n Y_n \xrightarrow{P} XY$ .

c) By Minkowski’s inequality and the Cauchy-Schwarz inequality,

$$\begin{aligned} E[|X_n Y_n - XY|] &\leq E[|X_n Y_n - X_n Y|] + E[|X_n Y - XY|] \\ &\leq \sqrt{E[X_n^2]E[(Y_n - Y)^2]} + \sqrt{E[Y^2]E[(X_n - X)^2]}. \end{aligned}$$

Since  $X_n \xrightarrow{\text{q.m.}} X$  and  $Y \in L^2$ , the second term converges to zero. For the first term, we need, in addition to  $Y_n \xrightarrow{\text{q.m.}} Y$ , the property that  $X_n \xrightarrow{\text{q.m.}} X$  implies that  $E[X_n^2] \rightarrow E[X^2]$ . ■

Here is the analogue of Theorem 5.20 for multiplication.

**Theorem 5.22 (Slutsky’s theorem, bis).** If  $X_n \xrightarrow{d} X$  and  $Y_n \xrightarrow{d} c \in \mathbb{R}$ , then  $X_n Y_n \xrightarrow{d} cX$ .

**Proof:** As in Theorem 5.20, it suffices to show that

$$E[f(X_n Y_n)] \rightarrow E[f(cX)]$$

for every bounded, uniformly continuous function  $f$ . First, choose  $K$  such that  $\pm K$  are continuity points of  $F_X$  and  $P\{|X| > K\} < \varepsilon$ , which means that  $P\{|X_n| > K\} < 2\varepsilon$  for all sufficiently large values of  $n$ . Then, given  $\varepsilon > 0$ , there is  $\delta > 0$  such that  $|f(x) - f(y)| < \varepsilon$  whenever  $|x - y| < \delta$ .

With  $M = \sup_x |f(x)|$ , which is finite,

$$\begin{aligned}
& |E[f(X_n Y_n)] - E[f(cX)]| \\
& \leq E[|f(X_n Y_n) - f(cX_n)|; \{|Y_n - c| > \delta/K\}] \\
& \quad + E[|f(X_n Y_n) - f(cX_n)|; \{|Y_n - c| \leq \delta/K, |X_n| > K\}] \\
& \quad + E[|f(X_n Y_n) - f(cX_n)|; \{|Y_n - c| \leq \delta/K, |X_n| \leq K\}] \\
& \quad + |E[f(cX_n)] - E[f(cX)]| \\
& \leq 2MP\{|Y_n - c| > \delta/K\} + 2MP\{|X_n| > K\} + \varepsilon \\
& \quad + |E[f(cX_n)] - E[f(cX)]| \\
& \leq 2MP\{|Y_n - c| > \delta\} + 3\varepsilon + |E[f(cX_n)] - E[f(cX)]|,
\end{aligned}$$

provided that  $n$  is large. Hence, it remains only to show  $cX_n \xrightarrow{d} cX$ , but  $h(x) = f(cx)$  is bounded and continuous, so that

$$E[f(cX_n)] = E[h(X_n)] \rightarrow E[h(X)] = E[f(cX)],$$

and  $cX_n \xrightarrow{d} cX$  by Theorem 5.8. ■

### 5.3.2 Continuous mappings

Convergence almost surely, in probability and in distribution are preserved under continuous mappings.

**Theorem 5.23.** Let  $g: \mathbb{R} \rightarrow \mathbb{R}$  be continuous.

- a) If  $X_n \xrightarrow{\text{a.s.}} X$ , then  $g(X_n) \xrightarrow{\text{a.s.}} g(X)$ .
- b) If  $X_n \xrightarrow{P} X$ , then  $g(X_n) \xrightarrow{P} g(X)$ .
- c) If  $X_n \xrightarrow{d} X$ , then  $g(X_n) \xrightarrow{d} g(X)$ .

**Proof:** a) This is immediate by “ $\omega$ -wise” reasoning. If  $X_n(\omega) \rightarrow X(\omega)$ , then  $g(X_n(\omega)) \rightarrow g(X(\omega))$ .

b) Again, we apply Proposition 5.18. Given  $(n')$ , there is a subsequence  $(n'')$  such that  $X_{n''} \xrightarrow{\text{a.s.}} X$ , which implies that  $g(X_{n''}) \xrightarrow{\text{a.s.}} g(X)$  by continuity of  $g$ . Hence,  $g(X_n) \xrightarrow{P} g(X)$  by still another application of Proposition 5.18.

c) If  $f$  is bounded and continuous, so is  $f \circ g$ . Therefore,

$$E[f(g(X_n))] = E[(f \circ g)(X_n)] \rightarrow E[(f \circ g)(X)] = E[f(g(X))],$$

giving convergence in distribution by Theorem 5.8. ■

## 5.4 Convergence of Random Vectors

Convergence for random vectors parallels that for random variables. We reserve subscripts for the sequence index, while arguments in parentheses denote components of vectors. Thus,  $x_n(i)$  is the  $i$ th component of the vector  $x_n$ . For  $x \in \mathbb{R}^k$ , let  $\|x\|_2 = \sqrt{\sum_{i=1}^k x(i)^2}$  denote the Euclidean norm of  $x$ , and let  $\|x\|_1 = \sum_{i=1}^k |x_i|$  be the  $\ell^1$ -norm of  $x$ .

### 5.4.1 Convergence of random vectors as functions

The four function-based modes of convergence extend with no substantive alterations.

**Definition 5.24.** Given random  $k$ -vectors  $X, X_1, X_2, \dots$ ,

- a)  $(X_n)$  converges to  $X$  *almost surely* if  $P\{|X_n - X| \rightarrow 0\} = 1$ .
- b)  $(X_n)$  converges to  $X$  *in probability* if  $\lim_{n \rightarrow \infty} P\{|X_n - X| > \varepsilon\} = 0$  for every  $\varepsilon > 0$ .
- c)  $(X_n)$  converges to  $X$  in *quadratic mean* if  $\lim_{n \rightarrow \infty} E[\|X_n - X\|_2] = 0$ .
- d)  $(X_n)$  converges to  $X$  in  $L^1$  (or in *mean*) if  $\lim_{n \rightarrow \infty} E[\|X_n - X\|_1] = 0$ .

We employ the same notation as previously:  $X_n \xrightarrow{\text{a.s.}} X, \dots$   $\square$

Any need to develop more theory is obviated by the following result.

**Proposition 5.25.** For convergence almost surely, in probability, in quadratic mean, or in  $L^1$ ,  $X_n \rightarrow X$  if and only if, for each component  $i$ ,  $X_n(i) \rightarrow X(i)$  in the same sense.  $\square$

### 5.4.2 Convergence in distribution

Because of the relative intractability of multi-dimensional distribution functions, we adopt the criterion from Theorem 5.8 as the definition of convergence in distribution for random vectors.

**Definition 5.26.** Given random  $k$ -vectors  $X, X_1, X_2, \dots$ ,  $(X_n)$  converges to  $X$  *in distribution* if  $\lim_{n \rightarrow \infty} E[f(X_n)] = E[f(X)]$  for every bounded, continuous function  $f: \mathbb{R}^k \rightarrow \mathbb{R}$ .  $\square$

Only an incomplete analogue of Proposition 5.25 holds for convergence in distribution: convergence of components is implied by, but need not imply, convergence of random vectors.

**Proposition 5.27.** If  $X_n \xrightarrow{d} X$ , then  $X_n(i) \xrightarrow{d} X(i)$  for each  $i$ .  $\square$

The “converse” is that random vectors converge in distribution if and only if every linear combination of their components converges in distribution.

**Theorem 5.28 (Cramér-Wold device).** Suppose that  $X, X_1, X_2, \dots$  are random  $k$ -vectors. Then,  $X_n \xrightarrow{d} X$  if and only if for all choices of  $a = (a(1), \dots, a(k)) \in \mathbb{R}^k$ ,

$$\sum_{i=1}^k a(i)X_n(i) \xrightarrow{d} \sum_{i=1}^k a(i)X(i). \quad (5.4)$$

**Proof:** We prove here that convergence in distribution implies (5.4); the converse is proved in §6.5 using characteristic functions. For  $f$  a bounded, continuous function on  $\mathbb{R}$ ,  $\tilde{f}(x) = f\left(\sum_{i=1}^k a(i)x(i)\right)$  is a bounded, continuous function on  $\mathbb{R}^k$ , and, hence,

$$\begin{aligned} E\left[f\left(\sum_{i=1}^k a(i)X_n(i)\right)\right] &= E[\tilde{f}(X_n)] \rightarrow E[\tilde{f}(X)] \\ &= E\left[f\left(\sum_{i=1}^k a(i)X(i)\right)\right] \end{aligned}$$

by Definition 5.26. Theorem 5.8 applies to give (5.4). ■

### 5.4.3 Continuous mappings

As with random variables, convergence almost surely, in probability and in distribution are preserved by continuous mappings.

**Theorem 5.29.** Let  $g$  be a continuous mapping of  $\mathbb{R}^k$  into  $\mathbb{R}^m$ .

- a) If  $X_n \xrightarrow{\text{a.s.}} X$ , then  $g(X_n) \xrightarrow{\text{a.s.}} g(X)$ .
- b) If  $X_n \xrightarrow{P} X$ , then  $g(X_n) \xrightarrow{P} g(X)$ .
- c) If  $X_n \xrightarrow{d} X$ , then  $g(X_n) \xrightarrow{d} g(X)$ . □

## 5.5 Limit Theorems for Bernoulli Summands

This section illustrates the convergence theory via limit theorems for sums of independent, Bernoulli distributed random variables. Generalizations of these, the heart of classical probability, are presented in Chapter 7. Here we exploit special structure and do not need the more powerful tools developed in later chapters.

Let  $(X_n)$  be a Bernoulli process with parameter  $p \in (0, 1)$ :  $X_1, X_2, \dots$  are i.i.d. random variables with  $P\{X_i = 1\} = 1 - P\{X_i = 0\} = p$ , and let

$q = 1 - p$ . (See §3.6 for details.) In this section we study the asymptotic behavior of the success counting process  $S_n = \sum_{i=1}^n X_i$ . We recall from Proposition 3.32 that  $S_n \xrightarrow{d} B(n, p)$ :

$$P\{S_n = k\} = \binom{n}{k} p^k (1-p)^{n-k}, \quad 0 \leq k \leq n.$$

Thus, results in this section, especially the local central limit theorem, describe asymptotic behavior of binomial probabilities.

### 5.5.1 Laws of large numbers

Laws of large numbers establish convergence of empirical averages  $S_n/n$  to  $p = E[X_1]$ , and are of two main varieties: *weak laws of large numbers* state that  $S_n/n$  converges in probability to  $p$ , while *strong laws of large numbers* state that  $S_n/n$  converges almost surely to  $p$ . Of course, in this case the latter implies the former, but in other contexts, weak laws of large numbers can be proved under less stringent hypotheses than strong laws.

We begin with an intermediate result, showing convergence in quadratic mean, which *does not* follow from almost sure convergence. Recall that  $E[X_i] = p$  and  $\text{Var}(X_i) = pq$ , so that  $E[S_n] = np$  by linearity and  $\text{Var}(S_n) = npq$  by Proposition 4.43.

**Theorem 5.30 (Weak law of large numbers).** As  $n \rightarrow \infty$ , we have  $S_n/n \xrightarrow{\text{q.m.}} p$ , and, hence,  $S_n/n \xrightarrow{P} p$ .

**Proof:** For each  $n$ ,

$$E[(S_n/n - p)^2] = \frac{E[(S_n - np)^2]}{n^2} = \frac{\text{Var}(S_n)}{n^2} = \frac{pq}{n}. \blacksquare$$

Next is the strong law of large numbers.

**Theorem 5.31 (Strong law of large numbers).** We have  $S_n/n \xrightarrow{\text{a.s.}} p$ .

**Proof:** For  $\varepsilon > 0$ ,

$$\begin{aligned} P\{|S_n/n - p| > \varepsilon\} &= P\{|S_n - np| > n\varepsilon\} \\ &\leq \frac{E[(S_n - np)^4]}{n^4 \varepsilon^4} \end{aligned}$$

[by Chebyshev's inequality (4.27) with  $g(x) = x^4$ ]

$$\begin{aligned} &\leq \frac{3n(n-1)\text{Var}(X_1)^2}{n^4 \varepsilon^4} + \frac{nE[(X_1 - p)^4]}{n^4 \varepsilon^4} \\ &= O(1/n^2), \end{aligned}$$

The penultimate step is done by expanding the fourth power of  $S_n - np = \sum_{i=1}^n (X_i - p)$ . The resultant expectation contains  $n$  terms of the form  $E[(X_i - p)^4]$  and  $6\binom{n}{2}$  terms of the form  $E[(X_i - p)^2(X_j - p)^2]$ , which, by independence of the  $X_i$ , are equal to  $E[(X_i - p)^2]E[(X_j - p)^2] = \text{Var}(X_1)^2$ . There are also  $4n(n - 1)$  terms  $E[(X_i - p)^3(X_j - p)]$ ,  $6n(n - 1)(n - 2)$  terms  $E[(X_i - p)^2(X_j - p)(X_k - p)]$  and  $n(n - 1)(n - 2)(n - 3)$  terms  $E[(X_i - p)(X_j - p)(X_k - p)(X_\ell - p)]$ , all of which are zero by independence.

Hence,  $\sum_{n=1}^{\infty} P\{|S_n/n - p| > \varepsilon\} < \infty$ , and consequently,  $S_n/n \xrightarrow{\text{a.s.}} p$  by Proposition 5.7. ■

One can invoke Theorem 5.31 to support the frequency interpretation of probability. If the  $X_i$  are the indicator functions of some event  $A$  under independent replications of a random experiment, then  $(1/n) \sum_{i=1}^n X_i$  is the empirical frequency of  $A$  in the first  $n$  replications, and the strong law of large numbers implies that with probability one these empirical frequencies converge to  $P(A)$ , the probability of  $A$ .

It is informative to note that in the proofs of Theorem 5.30 and Theorem 5.31, only the moments of the  $X_i$  were used, and not their being Bernoulli-distributed. These theorems, therefore, remain valid more generally.

**Theorem 5.32 (Weak law of large numbers).** Let  $X_1, X_2, \dots$  be i.i.d. with  $E[X_1^2] < \infty$ . Then,  $(1/n) \sum_{i=1}^n X_i \xrightarrow{\text{q.m.}} E[X_1]$ . □

In fact, in Theorem 5.32, the  $X_i$  need only be uncorrelated, since in this case we still have  $\text{Var}(S_n) = n\text{Var}(X_1)$ .

**Theorem 5.33 (Strong law of large numbers).** Let  $X_1, X_2, \dots$  be i.i.d. with  $E[X_1^4] < \infty$ . Then,  $(1/n) \sum_{i=1}^n X_i \xrightarrow{\text{a.s.}} E[X_1]$ . □

A more general law of large numbers, requiring only that  $E[|X_1|] < \infty$ , is presented in §7.2.

### 5.5.2 Central limit theorems

Central limit theorems are perhaps the most useful in probability. They state, in this case, that for large  $n$ ,  $S_n$  has approximately a normal distribution with mean  $np = E[S_n]$  and variance  $npq = \text{Var}(S_n)$ , or, more precisely, that  $[S_n - np]/\sqrt{npq}$  has approximately a standard normal distribution.

The local central limit theorem provides an approximation to individual binomial probabilities in terms of the standard normal density. Let

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

be the standard normal density function.

**Theorem 5.34 (DeMoivre-Laplace local limit theorem).** Suppose that  $0 \leq k_n \leq n$  for each  $n$ , and that

$$x_n \stackrel{\text{def}}{=} \frac{k_n - np}{\sqrt{npq}} = o(n^{1/6}), \quad n \rightarrow \infty. \quad (5.5)$$

Then,

$$\lim_{n \rightarrow \infty} \frac{\sqrt{npq} P\{S_n = k_n\}}{\phi(x_n)} = 1. \quad (5.6)$$

**Proof:** For notational simplicity, let  $k = k_n$ ,  $j = n - k$ ,  $x = x_n$  and

$$b(n, p; k) = P\{S_n = k\} = \binom{n}{k} p^k (1-p)^{n-k}.$$

By Stirling's approximation (3.18), since  $n$ ,  $k$  and  $j$  all converge to infinity,

$$b(n, p; k) \cong \sqrt{\frac{n}{2\pi kj}} \left(\frac{k}{np}\right)^{-k} \left(\frac{j}{nq}\right)^{-j},$$

and, consequently,

$$\begin{aligned} \frac{\sqrt{npq} b(n, p; k)}{\phi(x)} &\cong \sqrt{\frac{n^2 pq}{kj}} \left(\frac{k}{np}\right)^{-k} \left(\frac{j}{nq}\right)^{-j} e^{x^2/2} \\ &\cong \left(\frac{k}{np}\right)^{-k} \left(\frac{j}{nq}\right)^{-j} e^{x^2/2}, \end{aligned}$$

since, by (5.5),  $k/n \rightarrow p$  and  $j/n \rightarrow q$ .

Hence, it suffices to show that

$$b_n \stackrel{\text{def}}{=} \left(\frac{k}{np}\right)^{-k} \left(\frac{j}{nq}\right)^{-j} e^{x^2/2} \rightarrow 1,$$

which we do by showing that  $\log b_n \rightarrow 0$ . Indeed,

$$\begin{aligned} \log b_n &= -k \log \left(\frac{k}{np}\right) - j \log \left(\frac{j}{nq}\right) + \frac{x^2}{2} \\ &= -[np + x\sqrt{npq}] \log [1 + x\sqrt{q/p}] \\ &\quad - [nq - x\sqrt{npq}] \log [1 - x\sqrt{p/q}] \\ &\quad + \frac{x^2}{2} \end{aligned}$$

[since  $k = np + x\sqrt{npq}$  and  $j = nq - x\sqrt{npq}$ ]

$$\begin{aligned} &\cong -[np + x\sqrt{npq}] \left[ x\sqrt{\frac{q}{np}} - \frac{1}{2}x^2 \frac{q}{np} \right] \\ &\quad - [nq - x\sqrt{npq}] \left[ -x\sqrt{\frac{p}{nq}} - \frac{1}{2}x^2 \frac{p}{nq} \right] \\ &\quad + \frac{x^2}{2} \end{aligned}$$

[since for  $y \rightarrow 0$ ,  $\log(1+y) \cong y - y^2/2$ ]

$$\begin{aligned} &= -x\sqrt{npq} + \frac{1}{2}x^2p - x^2p + \frac{1}{2}x^3q\sqrt{\frac{q}{np}} \\ &\quad + x\sqrt{npq} + \frac{1}{2}x^2 - x^2q - \frac{1}{2}x^3q\sqrt{\frac{p}{nq}} \\ &\quad + \frac{x^2}{2} \\ &= \frac{1}{2} \frac{x^3}{\sqrt{n}} \left[ q\sqrt{\frac{q}{p}} - p\sqrt{\frac{p}{q}} \right], \end{aligned}$$

which converges to zero by (5.5). ■

**Technical Aside.** The proof of Theorem 5.34 shows that the convergence in (5.6) is uniform in values of  $k$  satisfying  $|k - np| = o(n^{2/3})$ . □

Thus, for large values of  $n$  and values of  $k$  not differing too drastically from  $np = E[S_n]$ , the binomial distribution of  $S_n$  can be approximated by a normal distribution. Theorem 5.35 makes this more explicit.

**Theorem 5.35 (DeMoivre-Laplace global limit theorem).** As  $n \rightarrow \infty$ ,

$$\frac{S_n - np}{\sqrt{npq}} \xrightarrow{d} Z,$$

where  $Z \stackrel{d}{=} N(0, 1)$ .

**Proof:** We show that for  $a < b$ ,

$$\lim_{n \rightarrow \infty} P \left\{ a < \frac{S_n - np}{\sqrt{npq}} < b \right\} = \int_a^b \phi(x) dx.$$

For fixed  $n$  and each integer  $k$ , let  $x_k$  satisfy  $k = np + x_k\sqrt{npq}$ , and put  $\Delta x_k \stackrel{\text{def}}{=} x_k - x_{k-1} = 1/\sqrt{npq}$ . Then, for fixed  $T$ , by Theorem 5.34 and the discussion following it,

$$P \{ S_n = np + x_k\sqrt{npq} \} = \frac{1}{\sqrt{2\pi}} e^{-x_k^2/2} \Delta x_k [1 + \varepsilon_n(x_k)],$$

where  $\varepsilon_n(x_k) \rightarrow 0$  uniformly in  $k$  such that  $|x_k| \leq T$ . Therefore,

$$\begin{aligned} P\left\{a < \frac{S_n - np}{\sqrt{npq}} < b\right\} &= \sum_{a < x_k \leq b} P\{S_n = np + x_k \sqrt{npq}\} \\ &= \sum_{a < x_k \leq b} \frac{1}{\sqrt{2\pi}} e^{-x_k^2/2} \Delta x_k \\ &\quad + \sum_{a < x_k \leq b} \frac{1}{\sqrt{2\pi}} e^{-x_k^2/2} \Delta x_k \varepsilon_n(x_k) \\ &\rightarrow \int_a^b \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx. \quad \blacksquare \end{aligned}$$

In situations such as Theorem 5.35, it is convenient to write

$$\frac{S_n - np}{\sqrt{npq}} \xrightarrow{d} N(0, 1). \quad (5.7)$$

We do this hereafter, with similar usage for other distributions with special notations (Table 2.3).

The transformation  $S_n \mapsto S_n^* = (S_n - E[S_n])/\sqrt{\text{Var}(S_n)}$  appearing in Theorem 5.35 is known as *standardization*, since  $E[S_n^*] = 0$  and  $\text{Var}(S_n^*) = 1$ , the first two moments of the standard normal distribution.

### 5.5.3 The Poisson limit theorem

In the central limit theorems, although  $n$  converged to infinity,  $p$  remained fixed. These theorems provide useful approximations to binomial probabilities for values of  $p$  close to neither zero nor one. These approximations are not good, though, for small values of  $p$ , so here we consider the effect of simultaneously allowing  $n \rightarrow \infty$  and  $p \rightarrow 0$ . Theorem 3.29 contains a special case of the following result.

**Theorem 5.36 (Poisson limit theorem).** Let  $Y_n \stackrel{d}{=} B(n, p_n)$  for each  $n$ , and suppose that  $np_n \rightarrow \lambda$ . Then,  $Y_n \xrightarrow{d} P(\lambda)$ .

**Proof:** It suffices to show that  $P\{Y_n = k\} \rightarrow e^{-\lambda} \lambda^k / k!$  for each  $k \geq 0$ :

$$\begin{aligned} P\{Y_n = k\} &= \binom{n}{k} p_n^k (1 - p_n)^{n-k} \\ &= \frac{n(n-1) \cdots (n-k+1)}{k!} p_n^k (1 - p_n)^{n-k} \\ &\cong \frac{1}{k!} (np_n)^k \left(1 - \frac{np_n}{n}\right)^{n-k} \\ &\rightarrow e^{-\lambda} \frac{\lambda^k}{k!}. \quad \blacksquare \end{aligned}$$

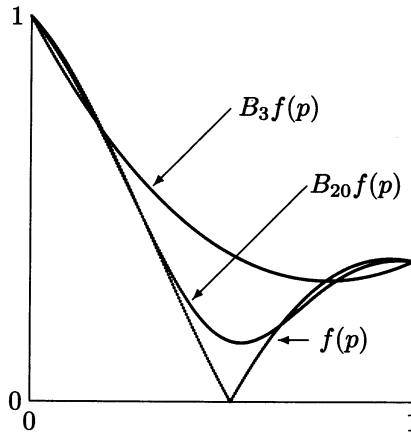


Figure 5.2. Bernstein Polynomials for  $f(p) = e^{-p}|\cos 3p|$

#### 5.5.4 Approximation of continuous functions

One beauty of probability is that it often leads to insightful derivations of results in other areas of mathematics. Although we do not pursue this connection in detail, which is particularly profound in number theory, partial differential equations and potential theory, we illustrate by showing that every continuous function on  $[0, 1]$  can be approximated uniformly by polynomials depending on its values at only finitely many points.

We first define the approximating polynomials.

**Definition 5.37.** For a continuous function  $f$  on  $[0, 1]$  and  $n \geq 1$ , the  $n$ th Bernstein polynomial of  $f$  is the function

$$B_n f(p) = \sum_{k=0}^n f(k/n) \binom{n}{k} p^k (1-p)^{n-k}, \quad p \in [0, 1]. \quad \square$$

For each  $n$ ,  $B_n f(\cdot)$  is a polynomial of degree at most  $n$ , whose computation requires only the values  $f(0), f(1/n), \dots, f(1)$ . Figure 5.2 illustrates for  $f(p) = e^{-p} |\cos 3p|$ , with  $n = 3$  and  $n = 20$ .

Note that  $B_n f(p) = E_p[f(S_n/n)]$ , where  $P_p$  is the probability under which  $(X_i)$  is a Bernoulli process with parameter  $p$ . This observation is crucial to proving the following approximation theorem.

**Theorem 5.38 (Weierstrass approximation theorem).** For  $f$  a continuous function on  $[0, 1]$ ,

$$\lim_{n \rightarrow \infty} \sup_{0 \leq p \leq 1} |B_n f(p) - f(p)| = 0.$$

**Proof:** Since  $f$  is bounded and uniformly continuous, there is  $M$  such that  $|f(x)| \leq M$  for all  $x \in [0, 1]$ , and given  $\varepsilon > 0$  there is  $\delta > 0$  such that  $|f(x) - f(y)| < \varepsilon/2$  whenever  $|x - y| < \delta$ . Then, for each  $p$ ,

$$\begin{aligned} |B_n f(p) - f(p)| &\leq E_p[|f(S_n/n) - f(p)|] \\ &= E_p[|f(S_n/n) - f(p)| ; \{|S_n/n - p| \leq \delta\}] \\ &\quad + E_p[|f(S_n/n) - f(p)| ; \{|S_n/n - p| > \delta\}] \\ &\leq \varepsilon/2 + 2M P_p \{|S_n/n - p| > \delta\} \\ &\leq \varepsilon/2 + 2M \frac{p(1-p)}{n\delta^2} \end{aligned}$$

[by Chebyshev's inequality]

$$\leq \varepsilon/2 + \frac{M}{2n\delta^2}$$

uniformly in  $p$ , since  $p(1-p) \leq 1/4$  for all  $p$ . Thus,

$$\sup_p |B_n f(p) - f(p)| \leq \varepsilon$$

for  $n$  sufficiently large. ■

## 5.6 Complements

### 5.6.1 $L^p$ Convergence of random variables

Convergence in quadratic mean and in  $L^1$  are special cases of  $L^p$  convergence. Recall from Definition 4.33 that  $L^p$  is the vector space of random variables  $X$  for which the  $L^p$  norm of  $X$ ,  $\|X\|_p \stackrel{\text{def}}{=} E[|X|^p]^{1/p}$  is finite, and that  $d_p(X, Y) = \|X - Y\|_p$  is a pseudo-metric on  $L^p$ . Convergence in  $L^p$  is simply convergence with respect to this metric. Let  $p \in [1, \infty)$  be fixed.

**Definition 5.39.** Suppose that  $X, X_1, X_2, \dots$  belong to  $L^p$ . Then,  $(X_n)$  converges to  $X$  in  $L^p$ , denoted by  $X_n \xrightarrow{L^p} X$ , if

$$\lim_{n \rightarrow \infty} E[|X_n - X|^p] = 0. \quad \square$$

The following properties are then valid.

**Theorem 5.40.** Suppose that  $X, X_1, X_2, \dots$  belong to  $L^p$ . Then,

- a) If  $X_n \xrightarrow{L^p} X$ , then  $X_n \xrightarrow{P} X$ .

- b) If  $1 \leq q \leq p$  and  $X_n \xrightarrow{L^p} X$ , then  $X_n \xrightarrow{L^q} X$ .

**Proof:** It suffices to prove b), for if  $X_n \xrightarrow{L^p} X$ , then in particular,  $X_n \xrightarrow{L^1} X$ , which yields convergence in probability by Proposition 5.12. But b) is a consequence of Lyapunov's inequality (4.24):  $\|X - Y\|_q \leq \|X - Y\|_p$  whenever  $q \leq p$ . ■

## 5.7 Exercises

- 5.1. Prove that for all five forms of convergence the limit is unique: if  $X_n \rightarrow X$  and  $X_n \rightarrow Y$ , then  $X \stackrel{\text{a.s.}}{\rightarrow} Y$  for the first four and  $X \stackrel{d}{\rightarrow} Y$  for convergence in distribution.
- 5.2. Prove that the four function-based forms are compatible with the vector space structure of the family of random variables in the sense that  $X_n \rightarrow X$  if and only if  $X_n - X \rightarrow 0$ .
- 5.3. Let  $X_1, X_2, \dots$  be independent. Prove that the probability of the event  $\{\omega: (X_n(\omega)) \text{ converges}\}$  is either zero or one, and that if the probability is one, then the limit  $X$  is a constant (almost surely).
- 5.4. a) Prove that the function

$$\rho(A, B) = P(A \Delta B), \quad A, B \in \mathcal{F},$$

is a pseudo-metric:  $\rho(A, B) \geq 0$ , with equality if and only if  $P(A \Delta B) = 0$ ,  $\rho(A, B) = \rho(B, A)$ , and

$$\rho(A, C) \leq \rho(A, B) + \rho(B, C)$$

for all  $A, B$  and  $C$ .

- b) Prove that  $\rho(A_n, A) \rightarrow 0$  if and only if  $\mathbf{1}_{A_n} \xrightarrow{\text{q.m.}} \mathbf{1}_A$ .
- 5.5. Prove that  $\mathbf{1}_{A_n} \xrightarrow{d} \mathbf{1}_A$  if and only if  $P(A_n) \rightarrow P(A)$ .
- 5.6. a) Prove that if  $X_n \xrightarrow{L^1} X$ , then  $E[X_n] \rightarrow E[X]$ .  
 b) Prove that if  $X_n \xrightarrow{\text{q.m.}} X$ , then  $E[X_n^2] \rightarrow E[X^2]$ .
- 5.7. Let  $X, X_1, X_2, \dots$  be positive and integer-valued. Prove that  $X_n \xrightarrow{d} X$  if and only if

$$P\{X_n = k\} \rightarrow P\{X = k\}$$

for all  $k \in \mathbb{N}$ .

**5.8.** For each  $n$ , suppose that  $X_n$  is uniformly distributed on  $\{0, \dots, n\}$ :

$$P\{X_n = k/n\} = \frac{1}{n+1}, \quad k = 0, \dots, n.$$

Prove that  $X_n \xrightarrow{d} U[0, 1]$ , so that continuous uniform distributions are limits of discrete.

**5.9.** Let  $Z_1, Z_2, \dots$  be i.i.d. and positive, with density  $f$  satisfying

$$\lambda \stackrel{\text{def}}{=} \lim_{x \downarrow 0} f(x) > 0.$$

Prove that

$$n \min\{Z_1, \dots, Z_n\} \xrightarrow{d} E(\lambda).$$

**5.10.** Let  $X_1, X_2, \dots$  be independent with

$$\begin{aligned} P\{X_k = k^2\} &= 1/k^2 \\ P\{X_k = -1\} &= 1 - 1/k^2. \end{aligned}$$

Prove that  $\sum_{k=1}^n X_k \xrightarrow{\text{a.s.}} -\infty$ .

**5.11.** Let  $X_1, X_2, \dots$  be pairwise uncorrelated with mean 0 and partial sums  $S_n = \sum_{k=1}^n X_k$ . Prove that if there is a constant  $c$  such that  $\text{Var}(X_k) \leq c$  for every  $k$ , then  $S_n/n^\alpha \xrightarrow{\text{q.m.}} 0$  for all  $\alpha > 1/2$ .

**5.12.** Prove that for  $a \in \mathbb{R}$ ,  $\limsup_n X_n \xrightarrow{\text{a.s.}} a$  if and only if for every  $\varepsilon > 0$ ,

$$\begin{aligned} P\{X_n > a + \varepsilon, \text{ i.o.}\} &= 0 \\ P\{X_n > a - \varepsilon, \text{ i.o.}\} &= 1. \end{aligned}$$

**5.13.** Let  $X_1, X_2, \dots$  be i.i.d. with distribution  $N(0, 1)$ .

a) Show that for  $x > 0$ ,

$$\frac{x}{1+x^2} e^{-x^2/2} \leq \int_x^\infty e^{-y^2/2} dy \leq \frac{1}{x} e^{-x^2/2}. \quad (5.8)$$

b) Use the right-hand inequality in (5.8) and the first Borel-Cantelli lemma (Theorem 1.27) to prove that for every  $\varepsilon > 0$ ,

$$P\{X_n/\sqrt{2 \log n} > 1 + \varepsilon, \text{ i.o.}\} = 0,$$

c) Use the left-hand inequality in (5.8) and the second Borel-Cantelli lemma (Theorem 3.22) to show that for every  $\varepsilon > 0$ ,

$$P\{X_n/\sqrt{2 \log n} > 1 - \varepsilon, \text{ i.o.}\} = 1.$$

- d) Explain why b) and c) mean that

$$P\{\limsup_{n \rightarrow \infty} X_n / \sqrt{2 \log n} = 1\} = 1.$$

**5.14.** Let  $N = (N_t)$  be a Poisson process with rate  $\lambda$ .

- a) For each  $t$ , let

$$W_t = T_{N_t+1} - t$$

be the *forward recurrence time* at  $t$ , that is, the time between  $t$  and the first arrival after  $t$ , which is arrival number  $N_t + 1$  and occurs at time  $T_{N_t+1}$ . Prove that  $W_t \xrightarrow{d} E(\lambda)$ .

- b) For each  $t$ , let

$$V_t = \begin{cases} t - T_{N_t} & \text{if } N_t > 0 \\ t & \text{if } N_t = 0 \end{cases}$$

be the *backward recurrence time* at  $t$ , the time that has elapsed at  $t$  since the last arrival prior to  $t$  (or  $t$  if there have been no arrivals). Prove that  $V_t$  has the mixed distribution in Example 1.42 and calculate  $E[V_t]$ .

- c) Prove that as  $t \rightarrow \infty$ ,  $V_t \xrightarrow{d} E(\lambda)$ .  
d) Prove that  $W_t$  and  $V_t$  are independent for each  $t$ .  
e) Let  $I_t = V_t + W_t$  be the length of the interarrival interval containing  $t$ . Prove that

$$\lim_{t \rightarrow \infty} E[I_t] = 2/\lambda = 2E[U_1].$$

(This is sometimes called the waiting time paradox, but the behavior is not really paradoxical: a specific point in time is more likely to lie in a long than a short interarrival interval.)

**5.15.** Prove that if  $X_1 \leq X_2 \leq \dots$  and  $X_n \xrightarrow{P} X$ , then  $X_n \xrightarrow{\text{a.s.}} X$ .

**5.16.** Let  $X, X_1, X_2, \dots$  be positive. Prove that if  $X_n \xrightarrow{P} X$  and  $E[X_n] \rightarrow E[X]$ , then  $X_n \xrightarrow{L^1} X$ .

**5.17.** Prove the *local limit theorem* for Poisson probabilities. Let

$$p(k; \lambda) = e^{-\lambda} \frac{\lambda^k}{k!},$$

and suppose that  $\lambda \rightarrow \infty$  and  $k \rightarrow \infty$  in such a manner that  $x = (k - \lambda)/\sqrt{\lambda}$  remains bounded. Then, with  $\phi$  the standard normal density,

$$\lim \frac{\sqrt{\lambda} p(k; \lambda)}{\phi(x)} = 1.$$

**5.18.** Prove that  $X_n \xrightarrow{L^1} X$  if and only if

$$\sup_{A \in \mathcal{F}} |E[X_n; A] - E[X; A]| \rightarrow 0.$$

**5.19.** Let  $X_1, X_2, \dots$  be independent with distribution  $P(1)$ . Prove that

$$\limsup_{n \rightarrow \infty} \frac{X_n(\log \log n)}{\log n} \stackrel{\text{a.s.}}{\rightarrow} 1.$$

**5.20.** Let  $(X_n)$  be a Bernoulli process with parameter  $p$ , let  $(S_n)$  be the success counting process, and for each  $n$ , let

$$S_n^* = \frac{S_n - np}{\sqrt{np(1-p)}}.$$

a) Prove that for  $\alpha > 1$ ,

$$P \left\{ |S_n^*| > \sqrt{2\alpha \log n}, \text{ i.o.} \right\} = 0.$$

b) Use a) to show that for  $\beta > 1/2$ ,

$$\frac{S_n - np}{n^\beta} \xrightarrow{\text{a.s.}} 0.$$

**5.21.** Let  $(X_n)$  be a sequence for which there exists  $Y \in L^1$  such that  $|X_n| \leq Y$  for all  $n$ . Prove that  $(X_n)$  is uniformly integrable.

**5.22.** Let  $(X_n)$  be a sequence for which there exists an increasing function  $f: \mathbb{R}_+ \rightarrow \mathbb{R}_+$ , with  $f(x)/x \rightarrow \infty$  as  $x \rightarrow \infty$ , and

$$\sup_n E[f(|X_n|)] < \infty.$$

Prove that  $(X_n)$  is uniformly integrable.

**5.23.** Let  $X_1, X_2, \dots$  be independent with distribution  $U[0, 1]$ . Prove that as  $n \rightarrow \infty$ ,

$$n(1 - \max \{X_1, \dots, X_n\}) \xrightarrow{\text{d}} E(1).$$

**5.24.** Suppose that for each  $n$ , let  $X_n \stackrel{\text{d}}{=} B(n, p)$ , and with  $k > 0$ , let  $A_n = \{X_n \geq k\}$ .

a) Calculate  $\lim_{n \rightarrow \infty} P(A_n)$  under the assumptions that  $np_n \rightarrow \lambda \in (0, \infty)$ , that  $np_n \rightarrow 0$  and that  $np_n \rightarrow \infty$ .

b) Establish a condition on  $p_n$  under which  $\sum_{n=1}^{\infty} P(A_n) < \infty$ .

- 5.25.** Use the construction from the proof of Theorem 5.8 to show that  $X_1, \dots, X_n$  are independent if and only if

$$E\left[\prod_{i=1}^n g_i(X_i)\right] = \prod_{i=1}^n E[g_i(X_i)]$$

for all bounded continuous functions  $g_1, \dots, g_n: \mathbb{R} \rightarrow \mathbb{R}$ .

- 5.26.** Construct random vectors  $X, X_1, X_2, \dots$  such that  $X_n(i) \xrightarrow{d} X(i)$  for each  $i$ , but  $(X_n)$  does not converge in distribution to  $X$ .
- 5.27.** Prove that if  $X_n \xrightarrow{L^1} X$ , there is a subsequence  $(X_{n'})$  with  $X_{n'} \xrightarrow{\text{a.s.}} X$ .

# Chapter 6

# Characteristic Functions

The characteristic function of a random variable is a complex-valued function calculated from its distribution function, but is more tractable in many ways, primarily because of its superior smoothness properties. The characteristic function uniquely determines the distribution function, so that recognizing the characteristic function of a random variable identifies its distribution function. The density function, if it exists, can be recovered algorithmically from the characteristic function. Characteristic functions convert convolution to the simpler operation of pointwise multiplication. Moments of a random variable are derivatives at zero of its characteristic function, while existence of even-order derivatives of the characteristic function implies existence of the corresponding moments. Finally, random variables converge in distribution if and only if their characteristic functions converge pointwise.

## 6.1 Definition and Basic Properties

We first review some notation and properties for complex numbers. Given  $z = x + iy \in \mathbb{C}$ , the *real part* of  $z$  is  $\Re z = x$ ; the *imaginary part* of  $z$  is  $\Im z = y$ ; the *complex conjugate* of  $z$  is  $\bar{z} = x - iy$ , and  $z$  is real if and only if  $\bar{z} = z$ . The *modulus* of  $z$  is  $|z| = \sqrt{x^2 + y^2}$ . Note also that  $|z|^2 = z\bar{z}$ . We will employ *Euler's formula*:  $e^{it} = \cos t + i \sin t$ , and alternative forms.

### 6.1.1 Fundamentals

Let  $(\Omega, \mathcal{F}, P)$  be a probability space.

**Definition 6.1.** The *characteristic function* of a random variable  $X$  is the function  $\varphi_X : \mathbb{R} \rightarrow \mathbb{C}$  defined by

$$\varphi_X(t) = E[e^{itX}] = \int_{-\infty}^{\infty} e^{itx} d\mathbb{F}_X(x). \quad \square$$

Distribution	Parameters	Characteristic Function
Constant	$c$	$\varphi(t) = e^{itc}$
Bernoulli	$p$	$\varphi(t) = 1 - p + pe^{it}$
Binomial	$n, p$	$\varphi(t) = (1 - p + pe^{it})^n$
Geometric	$p$	$\varphi(t) = pe^{it} / (1 - [1 - p]e^{it})$
Negative binomial	$m, p$	$\varphi(t) = [pe^{it} / (1 - [1 - p]e^{it})]^m$
Poisson	$\lambda$	$\varphi(t) = e^{\lambda(e^{it}-1)}$
Standard normal		$\varphi(t) = e^{-t^2/2}$
Normal	$\mu, \sigma^2$	$\varphi(t) = e^{\mu it - \sigma^2 t^2 / 2}$
Exponential	$\lambda$	$\varphi(t) = \lambda / (\lambda - it)$
Gamma	$\alpha, \lambda$	$\varphi(t) = [\lambda / (\lambda - it)]^\alpha$
Uniform on $[-a, a]$	$a$	$\varphi(t) = (\sin at) / at$

Table 6.1. Characteristic Functions of Key Distributions

Similarly, one can also define the characteristic function of a distribution function  $F$ :

$$\varphi_F(t) = \int_{-\infty}^{\infty} e^{itx} dF(x).$$

The characteristic function always exists, since for all  $t$ ,  $|\varphi_X(t)| \leq E[|e^{itX}|] = 1$ . Table 6.1 shows the key examples.

### 6.1.2 Elementary properties

**Proposition 6.2.** The characteristic function  $\varphi_X$  is uniformly continuous and  $\varphi_X(-t) = \overline{\varphi_X(t)}$  for each  $t$ .

**Proof:** For each  $h$ ,

$$\begin{aligned} |\varphi_X(t+h) - \varphi_X(t)| &= |E[e^{itX}(e^{ihX} - 1)]| \leq E[|e^{itX}| |e^{ihX} - 1|] \\ &= E[|e^{ihX} - 1|] \end{aligned}$$

uniformly in  $t$ . As  $h \rightarrow 0$ ,  $E[|e^{ihX} - 1|] \rightarrow 0$  by the dominated convergence theorem (Theorem 4.16).

The second statement is computational: for each  $t$ ,

$$\varphi_X(-t) = E[e^{-itX}] = E[\overline{e^{itX}}] = \overline{E[e^{itX}]} = \overline{\varphi_X(t)}. \blacksquare$$

Thus,  $\varphi_{-X}$  is the complex conjugate of  $\varphi_X$ :

$$\varphi_{-X}(t) = \varphi_X(-t) = \overline{\varphi_X(t)}, \quad t \in \mathbb{R}. \quad (6.1)$$

Many of our examples in this chapter involve normal distributions.

**Example 6.3 (Normal distribution).** If  $X \stackrel{d}{=} N(0, 1)$ , then

$$\begin{aligned} \varphi_X(t) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{itx} e^{-x^2/2} dx \\ &= e^{-t^2/2} \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-(x-it)^2/2} dx \\ &= e^{-t^2/2}, \end{aligned}$$

where the second equality is by completion of the square. The “correct” way to prove that the integral is (real and equal to) one is via Cauchy’s theorem and limits; we use instead the heuristic argument that it is the integral of the density of a normal distribution with (imaginary) mean  $it$  and variance one, and, hence, equal to one.

More generally, if  $Y \stackrel{d}{=} N(\mu, \sigma^2)$ , then  $Y = \sigma X + \mu$ , where  $X \stackrel{d}{=} N(0, 1)$ , and (by Exercise 6.1),

$$\varphi_Y(t) = e^{it\mu} \varphi_X(\sigma t) = e^{it\mu - \sigma^2 t^2/2}. \quad \square$$

The following result, in conjunction with the uniqueness theorems in §2, is one of the most powerful properties of characteristic functions: the difficult-to-calculate operation of convolution for distribution or density functions becomes pointwise multiplication of characteristic functions.

**Theorem 6.4.** If  $X$  and  $Y$  are independent, then  $\varphi_{X+Y} = \varphi_X \varphi_Y$ .

**Proof:** For each  $t$ ,

$$\varphi_{X+Y}(t) = E[e^{it(X+Y)}] = E[e^{itX} e^{itY}] = E[e^{itX}] E[e^{itY}],$$

where the last equality is by Corollary 4.30. ■

If  $X \stackrel{d}{=} N(0, \sigma_X^2)$  and  $Y \stackrel{d}{=} N(0, \sigma_Y^2)$  are independent, then Theorem 6.4 gives

$$\varphi_{X+Y}(t) = e^{-(\sigma_X^2 + \sigma_Y^2)t^2/2},$$

the characteristic function of the normal distribution  $N(0, \sigma_X^2 + \sigma_Y^2)$ . Does this imply that  $X + Y$  is normally distributed? Not yet, but the results in the next section justify this conclusion.

## 6.2 Inversion and Uniqueness Theorems

The key result in this section is Theorem 6.5, the inversion theorem for characteristic functions. However, Theorem 6.6, to the effect that random variables with the same characteristic function are identically distributed, is more useful, since most “inversions” are effected by recognition.

### 6.2.1 The inversion theorem

**Theorem 6.5.** Whenever  $a < b \in \mathbb{R}$  are continuity points of  $\mathsf{F}_X$ ,

$$P\{a < X < b\} = \lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \varphi_X(t) dt. \quad (6.2)$$

**Proof:** The proof requires the trigonometric identity

$$\int_0^\infty \frac{\sin \alpha x}{x} dx = (\operatorname{sgn} \alpha) \frac{\pi}{2}, \quad (6.3)$$

where  $\operatorname{sgn} \alpha$ , the *signum* of  $\alpha$ , is -1, 0 or 1 according as  $\alpha < 0$ ,  $\alpha = 0$  or  $\alpha > 0$ . For a proof, see Chung (1974).

We now verify (6.2). For fixed  $a, b$  and  $T$ ,

$$\begin{aligned} & \frac{1}{2\pi} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \varphi_X(t) dt \\ &= \int_{-\infty}^\infty \left[ \frac{1}{2\pi} \int_{-T}^T \frac{e^{it(x-a)} - e^{it(x-b)}}{it} dt \right] d\mathsf{F}_X(x) \\ &= \int_{-\infty}^\infty \left[ \frac{1}{\pi} \int_0^T \frac{\sin t(x-a)}{t} dt - \frac{1}{\pi} \int_0^T \frac{\sin t(x-b)}{t} dt \right] d\mathsf{F}_X(x) \\ &\rightarrow \int_{-\infty}^\infty \left[ \frac{1}{\pi} \int_0^\infty \frac{\sin t(x-a)}{t} dt - \frac{1}{\pi} \int_0^\infty \frac{\sin t(x-b)}{t} dt \right] d\mathsf{F}_X(x) \end{aligned}$$

[as  $T \rightarrow \infty$ , by the dominated convergence theorem]

$$\begin{aligned} &= \int_{(-\infty, a)} \left[ \frac{1}{\pi} \frac{-\pi}{2} - \frac{1}{\pi} \frac{-\pi}{2} \right] d\mathsf{F}_X(x) + \int_{\{a\}} \left[ 0 - \frac{1}{\pi} \frac{-\pi}{2} \right] d\mathsf{F}_X(x) \\ &\quad + \int_{(a, b)} \left[ \frac{1}{\pi} \frac{\pi}{2} - \frac{1}{\pi} \frac{-\pi}{2} \right] d\mathsf{F}_X(x) \\ &\quad + \int_{\{b\}} \left[ \frac{1}{\pi} \frac{\pi}{2} - 0 \right] d\mathsf{F}_X(x) + \int_{(b, \infty)} \left[ \frac{1}{\pi} \frac{\pi}{2} - \frac{1}{\pi} \frac{\pi}{2} \right] d\mathsf{F}_X(x) \end{aligned}$$

[by (6.3)]

$$= \int_{(a,b)} \left[ \frac{1}{\pi} \frac{\pi}{2} - \frac{1}{\pi} \frac{-\pi}{2} \right] dF_X(x),$$

where the final equality holds because  $a$  and  $b$  are continuity points of  $F_X$ . This last expression, however, is just (6.2). ■

### 6.2.2 The uniqueness theorem

The most important consequence of Theorem 6.5 is that the distribution of a random variable is determined uniquely by its characteristic function.

**Theorem 6.6.** If  $\varphi_X(t) = \varphi_Y(t)$  for all  $t$ , then  $X \stackrel{d}{=} Y$ .

**Proof:** Theorem 6.5 implies that if  $a$  and  $b$  are continuity points of  $F_X$  and  $F_Y$ , then

$$F_X(b) - F_X(a) = F_Y(b) - F_Y(a).$$

Since a distribution function has at most countably many discontinuities, it follows (by letting  $a \rightarrow -\infty$ ) that

$$F_X(b) = F_Y(b)$$

for all common points  $b$  of continuity, and, hence, that  $F_X = F_Y$ . ■

Hence, random variables with the same characteristic function are identically distributed. It is indeed true that if

$$\varphi_Z(t) = e^{-(\sigma_X^2 + \sigma_Y^2)t^2/2},$$

then  $Z \stackrel{d}{=} N(0, \sigma_X^2 + \sigma_Y^2)$ .

### 6.2.3 Specialized inversion theorems

We first consider absolutely continuous random variables, and derive the only really usable inversion algorithm for characteristic functions.

**Theorem 6.7 (Fourier inversion theorem).** If

$$\int_{-\infty}^{\infty} |\varphi_X(t)| dt < \infty, \quad (6.4)$$

then  $X$  is absolutely continuous with density

$$f_X(x) = \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi_X(t) dt. \quad (6.5)$$

**Proof:** Given (6.4), we may invoke the dominated convergence theorem to take limits in (6.2), obtaining

$$\begin{aligned}\mathsf{F}_X(b) - \mathsf{F}_X(a) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \frac{e^{-ita} - e^{-itb}}{it} \varphi_X(t) dt \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \left[ \int_a^b e^{-itx} dx \right] \varphi_X(t) dt \\ &= \int_a^b \left[ \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} \varphi_X(t) dt \right] dx. \quad \blacksquare\end{aligned}$$

Again, we illustrate for normal distributions.

**Example 6.8 (Normal distribution).** If  $X \stackrel{d}{=} N(0, 1)$ , with  $\varphi_X(t) = e^{-t^2/2}$ , then (6.5) applies, and hence

$$\begin{aligned}f_X(x) &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-itx} e^{-t^2/2} dt \\ &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-(x+it)^2/2} dt \\ &= \frac{1}{\sqrt{2\pi}} e^{-x^2/2}. \quad \square\end{aligned}$$

There is no simple criterion for discreteness of in terms of characteristic functions, but one can recover individual probabilities  $P\{X = x\}$ .

**Proposition 6.9.** For each  $x \in \mathbb{R}$ ,

$$P\{X = x\} = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T e^{-itx} \varphi_X(t) dt.$$

**Proof:** Computations similar to those in the proof of Theorem 6.5 give

$$\begin{aligned}\frac{1}{2T} \int_{-T}^T e^{-itx} \varphi_X(t) dt &= \int_{\{x\}} d\mathsf{F}_X(y) + \int_{\{x\}^c} \frac{\sin T(y-x)}{T(y-x)} d\mathsf{F}_X(y) \\ &\rightarrow \int_{\{x\}} d\mathsf{F}_X(y) \\ &= P\{X = x\}. \quad \blacksquare\end{aligned}$$

For integer-valued random variables we can improve Proposition 6.9.

**Corollary 6.10.** If  $X$  is integer-valued, then

$$P\{X = n\} = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-int} \varphi_X(t) dt, \quad n \in \mathbb{Z}. \quad (6.6)$$

**Proof:** In this case,  $\varphi_X$  is periodic with period  $2\pi$ , so that (6.6) follows from Proposition 6.9. ■

## 6.3 Moments and Taylor Expansions

Here we consider two related, but conceptually different: problems: computation of moments that are known (by other means) to exist, and establishing that moments exist.

### 6.3.1 Calculation of moments known to exist

Assuming that  $E[X^k]$  exists, it can be calculated from the  $k$ th derivative of  $\varphi_X$  at zero.

**Theorem 6.11.** If  $E[|X|^k] < \infty$ , then the derivative  $\varphi_X^{(k)}$  exists and

$$E[X^k] = i^{-k} \varphi_X^{(k)}(0). \quad (6.7)$$

**Proof:** If  $X \in L^k$ , then by the dominated convergence theorem it is permissible to differentiate  $\varphi_X(t) = E[e^{itX}]$  inside the expectation, with the result that

$$\varphi_X^{(k)}(t) = E[(iX)^k e^{itX}]$$

for all  $t$ , from which (6.7) follows.

For concreteness, here is more detail for  $k = 1$ . We have

$$\varphi'_X(t) = \lim_{h \rightarrow 0} \frac{\varphi_X(t+h) - \varphi_X(t)}{h} = \lim_{h \rightarrow 0} \frac{1}{h} E[e^{i(t+h)X} - e^{itX}].$$

The random variables  $Z_h = (1/h)[e^{i(t+h)X} - e^{itX}]$  converge to  $Z = iX e^{itX}$  as  $h \rightarrow 0$  and, moreover, since  $|e^{ity} - e^{itz}| \leq |y - z||t|$ ,

$$|Z_h| = \frac{|e^{i(t+h)X} - e^{itX}|}{|h|} \leq \frac{|h||X|}{|h|} = |X|,$$

so that the  $Z_h$  are dominated by  $|X| \in L^1$ . Hence,

$$\lim_{h \rightarrow 0} E[Z_h] = E[\lim_{h \rightarrow 0} Z_h] = E[(iX)e^{itX}]$$

by the dominated convergence theorem. ■

In particular, for  $X \in L^2$ ,  $E[X] = -i\varphi'_X(0)$  and  $E[X^2] = -\varphi''_X(0)$ .

### 6.3.2 Establishing existence of moments

For even  $k$ , existence of  $\varphi_X^{(k)}(0)$  implies that  $E[X^k] < \infty$ .

**Theorem 6.12.** Let  $k$  be an even integer, and suppose that  $\varphi_X^{(k)}(0)$  exists. Then,  $E[|X|^k] < \infty$ .

**Proof:** We do the proof first for  $k = 2$ , and then deduce the general case by induction.

We need the following result from analysis: given that  $\varphi_X''(0)$  exists,

$$\varphi_X''(0) = \lim_{h \downarrow 0} \frac{\varphi_X(h) + \varphi_X(-h) - 2\varphi_X(0)}{h^2}.$$

Since for  $y \downarrow 0$ ,  $1 - \cos y \cong y^2/2 + O(y^4)$ ,

$$\begin{aligned}\varphi_X''(0) &= \lim_{h \downarrow 0} 2 \int_{-\infty}^{\infty} \frac{\cos hx - 1}{h^2} dF_X(x) \\ &= 2 \int_{-\infty}^{\infty} \lim_{h \downarrow 0} \frac{\cos hx - 1}{h^2} dF_X(x) \\ &= - \int_{-\infty}^{\infty} x^2 dF_X(x) \\ &= -E[X^2].\end{aligned}$$

Suppose now that  $E[|X|^{2k-2}] < \infty$  and that  $\varphi_X^{(2k)}(0)$  exists. Then,  $\varphi_X^{(2k-2)}(t)$  exists for all  $t$  in some neighborhood  $U$  of 0, on which, furthermore,  $\varphi_X^{(2k-2)}$  is continuous. If we put

$$\tilde{G}(x) = \int_{-\infty}^x y^{2k-2} dF_X(y),$$

then  $G(x) = \tilde{G}(x)/\tilde{G}(\infty)$  is a distribution function and

$$\varphi_G(t) = \frac{(-1)^{k-1} \varphi_X^{(2k-2)}(t)}{\tilde{G}(\infty)}.$$

Hence,  $\varphi_G''$  exists in a neighborhood of the origin, and by the case  $k = 2$ ,

$$\varphi_G''(0) = -\frac{1}{\tilde{G}(\infty)} \int x^2 dG(x) = -\frac{1}{\tilde{G}(\infty)} \int x^{2k} dF_X(x),$$

which yields

$$(-1)^k \varphi_X^{(2k)}(0) = \int x^{2k} dF_X(x).$$

This reasoning is, of course, invalid if  $\tilde{G} \equiv 0$ , but this can happen only if  $X \stackrel{\text{a.s.}}{\equiv} 0$ , in which case  $\varphi_X \equiv 1$  and the theorem evidently holds. ■

Yet again, we use normal distributions to illustrate.

**Example 6.13 (Normal distribution).** Suppose  $X \stackrel{d}{=} N(0, 1)$ . Then,

$$\begin{aligned}\varphi'_X(t) &= -te^{-t^2/2} \\ \varphi''_X(t) &= -e^{-t^2/2} + t^2 e^{-t^2/2}.\end{aligned}$$

Thus,  $E[X^2]$  exists, and by Theorem 6.11,  $E[X] = 0$  and  $E[X^2] = 1$ .  $\square$

### 6.3.3 Taylor expansions of characteristic functions

The Taylor expansion for characteristic functions follows. It is crucial to the proofs of several limit theorems.

**Theorem 6.14.** If  $E[|X|^k] < \infty$  for some integer  $k \geq 1$ , then

$$\varphi_X(t) = \sum_{j=0}^k \frac{(it)^j}{j!} E[X^j] + o(|t|^k), \quad t \rightarrow 0. \quad \square \quad (6.8)$$

## 6.4 Continuity Theorems and Applications

As in §3, the problems treated here, while related, are distinct conceptually. They are the following. Given random variables  $X, X_1, X_2, \dots$ , does convergence of  $\varphi_{X_n}$  to  $\varphi_X$  imply that  $X_n \xrightarrow{d} X$ ? And, given  $X_1, X_2, \dots$ , does existence of  $\varphi = \lim_n \varphi_{X_n}$  imply existence of  $X$  such that  $X_n \xrightarrow{d} X$ ? In the first, case the putative limit  $X$  is known, while in the second, it is not.

### 6.4.1 Convergence in distribution

This is the easier and more useful of the two continuity theorems.

**Theorem 6.15.** We have  $X_n \xrightarrow{d} X$  if and only if

$$\varphi_{X_n}(t) \rightarrow \varphi_X(t)$$

for each  $t \in \mathbb{R}$ .

**Proof:** Necessity: If  $X_n \xrightarrow{d} X$ , then since for each  $t$ ,  $\cos tx$  and  $\sin tx$  are bounded, continuous functions of  $x$ , Theorem 5.8 gives

$$\begin{aligned}\varphi_{X_n}(t) &= E[\cos tX_n] + iE[\sin tX_n] \rightarrow E[\cos tX] + iE[\sin tX] \\ &= \varphi_X(t).\end{aligned}$$

Sufficiency: It suffices to show that  $\mathbb{F}_{X_n}(b) - \mathbb{F}_{X_n}(a) \rightarrow \mathbb{F}_X(b) - \mathbb{F}_X(a)$  for all choices of  $a < b$  that are continuity points of  $\mathbb{F}_X$  for every  $n$  and of  $\mathbb{F}_X$  as well. Given this, (6.2) implies that

$$\begin{aligned}\mathbb{F}_X(b) - \mathbb{F}_X(a) &= \lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \varphi_X(t) dt \\ &= \lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \left[ \lim_{n \rightarrow \infty} \varphi_{X_n}(t) \right] dt \\ &= \lim_{n \rightarrow \infty} \lim_{T \rightarrow \infty} \frac{1}{2\pi} \int_{-T}^T \frac{e^{-ita} - e^{-itb}}{it} \varphi_{X_n}(t) dt \\ &= \lim_{n \rightarrow \infty} [\mathbb{F}_{X_n}(b) - \mathbb{F}_{X_n}(a)]\end{aligned}$$

by the dominated convergence theorem and Theorem 6.5. ■

### 6.4.2 The Lévy continuity theorem

Unlike Theorem 6.15, the next theorem does not entail advance knowledge of the limit  $X$ . The proof requires a preliminary result, whose analytical derivation (see Chow/Teicher, 1988) we omit.

**Lemma 6.16.** There is  $K \in \mathbb{R}$  such that for each  $X$ ,

$$P\{|X| \geq 1/a\} \leq \frac{K}{a} \int_0^a [1 - \Re \varphi_X(t)] dt$$

for all  $a > 0$ , where  $\Re \varphi_X$  denotes the real part of  $\varphi_X$ . □

The Lévy continuity theorem establishes that the pointwise limit of characteristic functions is a characteristic function, provided that it is continuous at zero.

**Theorem 6.17 (Lévy continuity theorem).** If  $\varphi(t) = \lim_{n \rightarrow \infty} \varphi_{X_n}(t)$  exists for every  $t \in \mathbb{R}$ , and  $\varphi$  is continuous at zero, then there is  $X$  such that  $\varphi_X = \varphi$  and  $X_n \xrightarrow{d} X$ .

**Proof:** We show that there exists a distribution function  $F$  such that every subsequence  $(X_{n'})$  admits a further subsequence  $(X_{n''})$  satisfying  $X_{n''} \xrightarrow{d} F$ .

Given  $(X_{n'})$ , by Helly's theorem (Theorem 6.32), there exist a subsequence  $(X_{n''})$  and a function  $F$  satisfying the conditions stated there, such that  $\mathbb{F}_{X_{n''}}(t) \rightarrow F(t)$  at all continuity points  $t$  of  $F$ . However, we do not know yet that  $F$  is a distribution function, but only that  $F(\infty) \leq 1$ . We will use Lemma 6.16 to show that  $F$  actually is a distribution function. Once this is done,  $F$ , which satisfies  $\varphi_F = \varphi$ , does not depend on  $(X_{n'})$  by the uniqueness theorem, and the proof will be complete.

Fix  $\varepsilon > 0$ . Then, since  $\varphi$  is continuous and  $\varphi(0) = \lim_n \varphi_{X_n}(0) = 1$ , there is  $a$  such that  $\pm 1/a$  are continuity points of  $F$  and such that  $0 < t < a$  implies that  $|1 - \Re\varphi(t)| < \varepsilon/2K$ , where  $K$  is the constant appearing in Lemma 6.16. Consequently,

$$\frac{K}{a} \int_0^a [1 - \Re\varphi(t)] dt < \frac{\varepsilon}{2},$$

so that, since  $\varphi_{X_{n''}} \rightarrow \varphi$ ,

$$\frac{K}{a} \int_0^a [1 - \Re\varphi_{X_{n''}}(t)] dt < \varepsilon,$$

for all sufficiently large values of  $n''$ . Hence, by Lemma 6.16,

$$F(1/a) - F(-1/a) = 1 - \lim_{n'' \rightarrow \infty} P\{|X_{n''}| > 1/a\} \geq 1 - \varepsilon.$$

Thus,  $F(\infty) - F(-\infty) = 1$ , and  $F$  is indeed a distribution function. ■

**Example 6.18 (Uniform distribution).** The function

$$\varphi(t) = \frac{\sin t}{t} = \lim_{n \rightarrow \infty} \prod_{i=1}^n \cos(t/2^i)$$

is a characteristic function, since for each  $n$ ,  $\varphi_n(t) = \prod_{i=1}^n \cos(t/2^i)$  is the characteristic function of  $\sum_{i=1}^n 2^{-i} X_i$ , where  $X_1, X_2, \dots$  are i.i.d. with  $P\{X_i = 1\} = P\{X_i = -1\} = 1/2$ , and since  $\varphi$  is continuous at zero. From Table 6.1,  $\varphi = \varphi_{U[-1,1]}$ , so we have shown that  $\sum_{i=1}^{\infty} 2^{-i} X_i \stackrel{d}{=} U[-1, 1]$ . □

### 6.4.3 Application to classical limit theorems

Let  $X_1, X_2, \dots$  be i.i.d., and for each  $n$ , let  $S_n = \sum_{i=1}^n X_i$ . To illustrate the power of Theorem 6.15, we use it to prove two classical limit theorems for the partial sums  $S_n$ . In addition, we give another proof of the Poisson approximation for binomial probabilities.

The weak law of large numbers generalizes Theorem 5.30, in which the summands are Bernoulli distributed.

**Theorem 6.19 (Weak law of large numbers).** If  $E[|X_1|]$  is finite, then  $S_n/n \xrightarrow{P} E[X_1]$ .

**Proof:** Let  $\mu = E[X_1]$ . By Theorem 6.15, Theorem 6.6 and the property that convergence in distribution to a constant implies convergence in probability (Proposition 5.14), it suffices to show that  $\varphi_{S_n/n}(t) \rightarrow e^{it\mu}$  for each  $t \in \mathbb{R}$ . But

$$\varphi_{S_n/n}(t) = \varphi_{S_n}\left(\frac{t}{n}\right) = \varphi_X\left(\frac{t}{n}\right)^n = \left[1 + \frac{it\mu}{n} + o\left(\frac{1}{n}\right)\right]^n \rightarrow e^{it\mu},$$

where the second equality is by Theorem 6.4, since the  $X_i$  are i.i.d., and the third is by Theorem 6.14 with  $k = 1$ . ■

The central limit theorem extends the DeMoivre-Laplace global limit theorem (Theorem 5.35), in which the  $S_n$  have binomial distributions.

**Theorem 6.20 (Central limit theorem).** If  $0 < \sigma^2 = \text{Var}(X_1) < \infty$  and  $E[X_1] = 0$ , then  $S_n/\sigma\sqrt{n} \xrightarrow{\text{d}} N(0, 1)$ .

**Proof:** By Theorem 6.15, Theorem 6.6 and the property that  $\varphi_{N(0,1)}(t) = e^{-t^2/2}$  (Table 6.1), we need only show that  $\varphi_{S_n/\sigma\sqrt{n}}(t) \rightarrow e^{-t^2/2}$  for each  $t$ . Again we use the Taylor expansion (6.8), this time with  $k = 2$ :

$$\begin{aligned}\varphi_{S_n/\sigma\sqrt{n}}(t) &= \varphi_X\left(\frac{t}{\sigma\sqrt{n}}\right)^n \\ &= \left[1 + i\frac{t}{\sigma\sqrt{n}}E[X_1] - \frac{t^2}{2n\sigma^2}E[X_1^2] + o\left(\frac{1}{n}\right)\right]^n \\ &= \left[1 - \frac{t^2}{2n} + o\left(\frac{1}{n}\right)\right]^n,\end{aligned}$$

since  $E[X_1] = 0$  and  $\sigma^2 = E[X_1^2]$ , and this converges to  $e^{-t^2/2}$ . ■

The final application is another proof of the Poisson limit theorem for binomial probabilities, given already in §5.5 as Theorem 5.36.

**Theorem 6.21 (Poisson limit theorem, bis).** Suppose that for each  $n$ ,  $Y_n \stackrel{\text{d}}{=} B(n, p)$  and that  $np_n \rightarrow \lambda \in (0, \infty)$ . Then,  $Y_n \xrightarrow{\text{d}} P(\lambda)$ .

**Proof:** Once again, by appeal to Theorem 6.15, Theorem 6.6 and Table 6.1, it is enough to verify that

$$[1 - p_n + p_n e^{it}]^n = \varphi_{Y_n}(t) \rightarrow \varphi_Y(t) = e^{\lambda(e^{it} - 1)}$$

for each  $t$ , but this is nearly immediate:

$$[1 - p_n + p_n e^{it}]^n \cong \left[1 - \frac{\lambda}{n} + \frac{\lambda}{n} e^{it}\right]^n \rightarrow e^{\lambda(e^{it} - 1)}. \quad ■$$

## 6.5 Other Transforms

Here, we consider characteristic functions for random vectors and other transforms for random variables, namely, Laplace transforms, moment generating functions and generating functions.

### 6.5.1 Characteristic functions of random vectors

We begin by recalling that the inner product of  $x, y \in \mathbb{R}^k$  is given by  $\langle x, y \rangle = \sum_{j=1}^k x(j)y(j)$ . The characteristic function of a random vector is defined in the following manner.

**Definition 6.22.** The *characteristic function* of a random  $k$ -vector  $X$ , or *joint characteristic function* of  $X_1, \dots, X_k$ , is the function  $\varphi_X$  defined for  $t = (t(1), \dots, t(k)) \in \mathbb{R}^k$  by

$$\varphi_X(t) = E[e^{i\langle t, X \rangle}] = E\left[\exp\left(i \sum_{j=1}^k t(j)X(j)\right)\right]. \quad \square$$

The properties, with one exception, are *exactly* those of one-dimensional characteristic functions.

**Theorem 6.23.** Let  $\varphi_X$  denote the characteristic function of the random  $k$ -vector  $X$ . Then,

- a)  $\varphi_X$  is uniformly continuous.
- b)  $\varphi_X(-t) = \overline{\varphi_X(t)} = \varphi_{-X}(t)$ , where  $-t = (-t(1), \dots, -t(k))$ .
- c) If  $X$  and  $Y$  are independent, then  $\varphi_{X+Y} = \varphi_X \varphi_Y$ .
- d)  $X \stackrel{d}{=} Y$  if and only if  $\varphi_X(t) = \varphi_Y(t)$  for all  $t \in \mathbb{R}^k$ .
- e) The components  $X(1), \dots, X(k)$  are independent if and only if

$$\varphi_X(t) = \prod_{j=1}^k \varphi_{X(j)}(t(j)), \quad t \in \mathbb{R}^k. \quad (6.9)$$

- f)  $X_n \xrightarrow{d} X$  if and only if  $\varphi_{X_n}(t) \rightarrow \varphi_X(t)$  for all  $t \in \mathbb{R}^k$ .
- g) Let  $X_1, X_2, \dots$  be random vectors for which  $\varphi(t) = \lim_{n \rightarrow \infty} \varphi_{X_n}(t)$  exists for all  $t \in \mathbb{R}^k$ . If  $\varphi$  is continuous at  $0 \in \mathbb{R}^k$ , then there is a random vector  $X$ , with  $\varphi_X = \varphi$ , such that  $X_n \xrightarrow{d} X$ .

**Proof:** We prove (sketchily) only the sufficiency of (6.9) as a criterion for independence. Suppose, for simplicity, that  $k = 2$ . The same pattern of argument used to prove Theorem 6.5 shows that if  $a_i < b_i$  are continuity points of  $F_{X_i}$ , then

$$\begin{aligned} P\{a_1 < X_1 < b_1, a_2 < X_2 < b_2\} \\ = \lim_{T_1 \rightarrow \infty} \lim_{T_2 \rightarrow \infty} \frac{1}{2\pi} \int_{-T_1}^{T_1} \frac{e^{-it_1 a_1} - e^{-it_1 b_1}}{it_1} \frac{1}{2\pi} \int_{-T_2}^{T_2} \frac{e^{-it_2 a_2} - e^{-it_2 b_2}}{it_2} \\ \times \varphi_{X_1, X_2}(t_1, t_2) dt_1 dt_2. \end{aligned}$$

Thus, if (6.9) holds, then by two applications of (6.2), we conclude that

$$\begin{aligned} P\{a_1 < X_1 < b_1, a_2 < X_2 < b_2\} \\ = P\{a_1 < X_1 < b_1\}P\{a_2 < X_2 < b_2\} \end{aligned}$$

for continuity points  $a_1, b_1, a_2, b_2$ , giving independence of  $X_1$  and  $X_2$ . ■

Multi-dimensional characteristic functions allow completion of the proof of Theorem 5.28.

**Theorem 6.24 (Cramér-Wold device).** Let  $X, X_1, X_2, \dots$  be random  $k$ -vectors. Then,  $X_n \xrightarrow{d} X$  if and only if

$$\sum_{j=1}^k t(j)X_n(j) \xrightarrow{d} \sum_{j=1}^k t(j)X(j) \quad (6.10)$$

for all  $t = (t(1), \dots, t(k)) \in \mathbb{R}^k$ .

**Proof:** With  $t$  fixed, let  $Y_n = \sum_{j=1}^k t(j)X_n(j)$ , with  $Y$  defined analogously. Then, (6.10) and Theorem 6.15 imply that  $\varphi_{Y_n}(s) \rightarrow \varphi_Y(s)$  for all  $s \in \mathbb{R}$ . The particular choice  $s = 1$  gives  $\varphi_{X_n}(t) \rightarrow \varphi_X(t)$ , and since  $t$  is arbitrary, this gives  $X_n \xrightarrow{d} X$ . ■

### 6.5.2 Laplace transforms

Laplace transforms exist for all positive random variables.

**Definition 6.25.** The *Laplace transform* of a random variable  $X \geq 0$  is the function  $\ell_X : \mathbb{R}_+ \rightarrow \mathbb{R}_+$  given by  $\ell_X(t) = E[e^{-tX}]$ . □

The properties mirror those of characteristic functions.

**Theorem 6.26.** Let  $\ell_X$  be the Laplace transform of  $X \geq 0$ . Then,

- a)  $\ell_X$  is uniformly continuous and  $0 < \ell_X(t) \leq \ell_X(0) = 1$  for all  $t$ .
- b) If  $X$  and  $Y$  are positive and independent, then  $\ell_{X+Y} = \ell_X \ell_Y$ .
- c) If  $X$  and  $Y$  are positive and  $\ell_X(t) = \ell_Y(t)$  for all  $t$  belonging to an open interval in  $\mathbb{R}_+$ , then  $X \stackrel{d}{=} Y$ .
- d) If  $X \geq 0$  and  $E[X^k] < \infty$ , then the derivative  $\ell_X^{(k)}$  exists and  $E[X^k] = (-1)^k \ell_X^{(k)}(0)$ .
- e) Given positive random variables  $X, X_1, X_2, \dots, X_n \xrightarrow{d} X$  if and only if  $\ell_{X_n}(t) \rightarrow \ell_X(t)$  for all  $t \in \mathbb{R}_+$ . □

### 6.5.3 Moment generating functions

Moment generating functions are of the same ilk as characteristic functions, but without the “clean” conditions for existence.

**Definition 6.27.** The *moment generating function* of  $X$  is the function  $\psi_X(t) = E[e^{tX}]$ , provided that the expectation exists for all  $t$  in some neighborhood of the origin.  $\square$

Existence of a moment generating function implies that of moments of *all orders*, which are again derivatives at zero.

**Proposition 6.28.** If  $\psi_X$  exists, then for each  $k$ ,  $E[|X|^k] < \infty$  and  $E[X^k] = \psi_X^{(k)}(0)$ .  $\square$

In particular, the distribution of any random variable whose moment generating function exists is uniquely determined by its moments. Other relevant properties are the following.

**Proposition 6.29.** With  $\psi_X$  the moment generating function of  $X$ ,

- a) If  $X$  and  $Y$  are independent, then  $\psi_{X+Y} = \psi_X\psi_Y$ .
- b) If  $\psi_X(t) = \psi_Y(t) < \infty$  for all  $t$  belonging to an open interval in  $\mathbb{R}$ , then  $X \stackrel{d}{=} Y$ .
- c) Let  $X, X_1, X_2, \dots$  be random variables whose moment generating functions all exist in some neighborhood  $U$  of  $0 \in \mathbb{R}$ . Then,  $X_n \xrightarrow{d} X$  if and only if  $\psi_{X_n}(t) \rightarrow \psi_X(t)$  for all  $t \in U$ .  $\square$

### 6.5.4 Generating functions

Generating functions are useful mainly for positive, integer-valued random variables.

**Definition 6.30.** The *generating function* of a positive, integer-valued random variable  $X$  is the function on  $[-1, 1]$  defined by  $\zeta_X(u) = E[u^X]$ .  $\square$

The properties are those of the other transforms, except that moments are computed from derivatives of the generating function at  $u = 1$ . Derivatives at  $u = 0$  yield the probabilities  $P\{X = k\}$ ; see Exercise 6.26.

**Proposition 6.31.** With  $\zeta_X$  the generating function of  $X$ ,

- a) If  $X$  and  $Y$  are independent, then  $\zeta_{X+Y} = \zeta_X\zeta_Y$ .
- b) If  $\zeta_X(u) = \zeta_Y(u)$  for all  $u$  belonging to an open interval containing the origin, then  $X \stackrel{d}{=} Y$ .

c) If  $E[X^k] < \infty$ , then the derivative  $\zeta_X^{(k)}$  exists and

$$\zeta_X^{(k)}(1) = E[X(X - 1) \cdots (X - k + 1)].$$

d) If  $X, X_1, X_2, \dots$  are positive and integer-valued, then  $X_n \xrightarrow{d} X$  if and only if  $\zeta_{X_n}(u) \rightarrow \zeta_X(u)$  for all  $u \in [-1, 1]$ .  $\square$

In fact, more is true than part d) states literally:  $E[X] = \lim_{u \uparrow 1} \zeta'_X(u)$  regardless of whether the limit is finite, with similar results holding for higher moments.

## 6.6 Complements

### 6.6.1 Helly's theorem

Every sequence of distribution functions admits a subsequence that converges at continuity points of the limit. The limit  $G$ , however, may be only a *subdistribution function*, with  $G(\infty) < 1$ . It may even be that  $G \equiv 0$ , which occurs, for example, if for each  $n$ ,  $F_n$  is the uniform distribution on  $[0, n]$ .

**Theorem 6.32 (Helly's selection theorem).** Let  $(F_n)$  be a sequence of distribution functions; then there exists a subsequence  $(F_{n'})$  and a function  $G: \mathbb{R} \rightarrow [0, 1]$  such that  $G$  is right-continuous and increasing,  $G(-\infty) = 0$ ,  $G(\infty) \leq 1$ , and  $F_{n'}(t) \rightarrow G(t)$  at all continuity points  $t$  of  $G$ .

**Proof:** Since  $0 \leq F_n \leq 1$  for each  $n$ , we may use compactness of  $[0, 1]$  and Cantor's diagonal argument to construct a subsequence  $(F_{n'})$  for which  $\tilde{G}(r) = \lim_{n' \rightarrow \infty} F_{n'}(r)$  exists for every rational  $r$ . That is, expressing  $\mathbb{Q}$  as a sequence  $(r_i)$ , we construct a subsequence  $(F_{n,1})$  converging at  $r_1$ , a further subsequence  $(F_{n,2})$  converging at  $r_2$  (and also at  $r_1$ ), and so on, and then take the “diagonal” subsequence  $F_{n,n}$ .

By monotonicity of the  $F_n$ ,  $\tilde{G}$  is an increasing function on  $\mathbb{Q}$ , and, hence, the function  $G(t) = \lim_{r \downarrow t, r \in \mathbb{Q}} \tilde{G}(r)$  is increasing and right-continuous.

That  $F_{n'}(t) \rightarrow G(t)$  if  $t$  is a continuity point of  $G$  is shown in the following manner. For  $\varepsilon > 0$ , there exist  $r, r'$  and  $r''$  in  $\mathbb{Q}$  such that  $r < r' < t < r''$  and  $G(r'') - G(r) < \varepsilon$ . Then,

$$G(r) \leq \tilde{G}(r') \leq G(t) \leq \tilde{G}(r'') \leq G(r'') \leq G(r) + \varepsilon,$$

while also  $F_{n'}(r') \rightarrow G(r')$ ,  $F_{n'}(r'') \rightarrow G(r'')$ , and  $F_{n'}(r') \leq F_{n'}(t) \leq F_{n'}(r'')$ . The proof follows from these statements.  $\blacksquare$

## 6.7 Exercises

- 6.1.** Prove that  $\varphi_{aX+b}(t) = e^{itb}\varphi_X(at)$  for all  $t \in \mathbb{R}$ .
- 6.2.** Let  $X_1, \dots, X_n$  be i.i.d. with distribution  $N(0, 1)$ .
- Prove that  $\bar{X}_n = (1/n) \sum_{i=1}^n X_i$  has distribution  $N(0, 1)$ .
  - Prove that  $S^2 = \sum_{i=1}^n X_i^2$  has a  $\chi^2$  distribution with  $n$  degrees of freedom.
- 6.3.** Let  $X$  be independent of  $Y$  and  $Z$ . Does it follow from  $X+Y \stackrel{d}{=} X+Z$  that  $Y \stackrel{d}{=} Z$ ? (*Hint:* In terms of characteristic functions, this question asks whether  $\varphi_X\varphi_Y = \varphi_X\varphi_Z$  implies that  $\varphi_Y = \varphi_Z$ .)
- 6.4.** Let  $X$  and  $Y$  be i.i.d. with mean 0 and variance 1. Prove that if  $X+Y$  and  $X-Y$  are independent, then  $X \stackrel{d}{=} Y \stackrel{d}{=} N(0, 1)$ .
- 6.5.** Prove that a convex combination of characteristic functions is itself a characteristic function. That is, suppose that  $\varphi_1, \varphi_2, \dots$  are characteristic functions, and that  $p_1, p_2, \dots$  are positive real numbers with  $\sum_{k=1}^{\infty} p_k = 1$ ; then  $\psi(t) = \sum_{k=1}^{\infty} p_k \varphi_k(t)$  is a characteristic function.
- 6.6.** Let  $h$  be a function on  $\mathbb{R}$  for which  $h''(0)$  exists. Prove that
- $$h''(0) = \lim_{t \downarrow 0} \frac{h(t) + h(-t) - 2h(0)}{t^2}.$$
- 6.7.**
- Let  $X_1, X_2, \dots$  be i.i.d., let  $N \stackrel{d}{=} P(\lambda)$  be independent of the  $X_i$ , and let  $Z = \sum_{i=1}^N X_i$ . Prove that  $\varphi_Z = e^{\lambda(\varphi_X - 1)}$ .
  - Use a) to express the expectation and variance of  $Z$  in terms of those of the  $X_i$  (and  $\lambda$ ).
- 6.8.** Suppose that  $Y_\lambda \stackrel{d}{=} P(\lambda)$ ,  $\lambda > 0$ . Use characteristic functions to prove that  $[Y_\lambda - \lambda]/\sqrt{\lambda} \stackrel{d}{\rightarrow} N(0, 1)$  as  $\lambda \rightarrow \infty$ .
- 6.9.** Let  $(N_t)$  be a Poisson process with rate  $\lambda$ , as defined in §3.6. Prove the central limit for  $N$ : as  $t \rightarrow \infty$ ,  $\sqrt{t}[N_t/t - \lambda] \stackrel{d}{\rightarrow} N(0, \lambda)$ .
- 6.10.** Prove that if  $(S_n)$  is a random walk, then for each  $n$ ,  $\varphi_{S_n}(t) = (\cos t)^n$ .
- 6.11.** A random variable  $X$  is *symmetrically distributed* (or just *symmetric*) if  $X \stackrel{d}{=} -X$ .
- Show that  $X$  is symmetric if and only if

$$\mathsf{F}_X(t) = \mathsf{F}_{-X}(t) = 1 - \mathsf{F}_X((-t) - ), \quad t \in \mathbb{R}.$$

- b) Show that if  $X$  is absolutely continuous, then  $X$  is symmetric if and only if

$$f_X(x) = f_X(-x), \quad x \in \mathbb{R}.$$

- c) Show that  $X$  is symmetric if and only if  $\varphi_X(t)$  is real for all  $t$ .

**6.12.** Prove that if  $X$  and  $Y$  are i.i.d., then  $\varphi_{X-Y}(t) = |\varphi_X(t)|^2$ , so that  $X - Y$  is symmetric.

**6.13.** Prove that for each  $X$ ,

$$\sum_{x \in \mathbb{R}} P\{X = x\}^2 = \lim_{T \rightarrow \infty} \frac{1}{2T} \int_{-T}^T |\varphi_X(t)|^2 dt.$$

**6.14.** A random variable  $X$  has a *lattice distribution* if there exist  $a, d \in \mathbb{R}$  such that

$$P\{X \in \{a + nd : n \in \mathbb{Z}\}\} = 1.$$

Show that  $X$  has a lattice distribution if and only if  $|\varphi_X(t)| = 1$  for some  $t \neq 0$ .

**6.15.** Let  $U_1, U_2, \dots$  be i.i.d. with distribution  $E(\lambda)$ .

- a) Calculate  $\varphi_{U_1}$  and use it to show that  $E[U_1] = 1/\lambda$ .  
 b) Use characteristic functions to show that  $T_k = \sum_{i=1}^k U_i$  has density function

$$f_{T_k}(t) = \frac{1}{(k-1)!} \lambda^k e^{-\lambda t} t^{k-1}.$$

- c) Thinking of the  $U_i$  as interarrival times in an arrival process (see §3.6) and the  $T_k$  as arrival times, let  $N_t = \sum_{k=1}^{\infty} \mathbf{1}(T_k \leq t)$  be the arrival counting process. Prove that for each  $k$  and  $t$ ,

$$P\{N_t = k\} = F_{T_k}(t) - F_{T_{k-1}}(t).$$

- d) Use c) to show that  $\varphi_{N_t}(u) = e^{\lambda t(e^{iu}-1)}$  for each  $t$  and  $u$ , and conclude that  $N_t$  has a Poisson distribution with mean  $\lambda t$ .

**6.16.** A random variable  $X$  has a *Cauchy distribution* if its density is

$$f_X(x) = \frac{1}{\pi(1+x^2)}, \quad x \in \mathbb{R}.$$

Calculate  $\varphi_X$ , and use the result to show that  $E[|X|] = \infty$ .

**6.17.** Let  $X_1, X_2, \dots$  be i.i.d. and integer-valued, with partial sums  $S_n = \sum_{i=1}^n X_i$ . Prove that for each  $n$  and  $k$ ,

$$P\{S_n = k\} = \frac{1}{2\pi} \int_{-\pi}^{\pi} e^{-ikt} \varphi_X(t)^n dt.$$

**6.18.** Let  $X_1, X_2, \dots$  be i.i.d. and positive, with continuous distribution function  $F$ . Recall from Exercise 4.10 that a *record* occurs at time  $k$  if  $X_k > \max\{X_1, \dots, X_{k-1}\}$ . Suppose that  $0 < \lambda < \mu$ , and for each  $n$ , let  $N_n(\lambda, \mu)$  be the number of records at times  $\lfloor n\lambda \rfloor, \dots, \lfloor n\mu \rfloor$ .

- a) Show that

$$E[e^{-tN_n(\lambda, \mu)}] = \exp \left[ \sum_{j=\lfloor n\lambda \rfloor}^{\lfloor n\mu \rfloor} \log \left( 1 - \frac{1 - e^{-t}}{j} \right) \right].$$

- b) Prove that as  $n \rightarrow \infty$ ,  $N_n(\lambda, \mu) \xrightarrow{d} P(\log(\mu/\lambda))$ .

**6.19.** For each  $n$ , let  $X_{n1}, \dots, X_{nn}$  be i.i.d. with distribution  $U[-n, n]$ , representing the positions of randomly located bodies of mass  $m > 0$ . Assuming an inverse square law of gravitational attraction, the force exerted on a unit mass at the origin is then

$$Y_n = Gm \sum_{i=1}^n \frac{\operatorname{sgn} X_{ni}}{X_{ni}^2},$$

where  $G$  is the gravitational constant.

- a) Show that  $Y_n \xrightarrow{d} Y$ , where

$$\varphi_Y(t) = \exp \left[ - \int_0^\infty \left[ [1 - \cos \left( \frac{Gmt}{x^2} \right)] \right] dx \right].$$

- b) Show that there is a constant  $c$  such that  $\varphi_Y(t) = e^{-c\sqrt{|t|}}$ .  
 c) Show that if instead there is an inverse  $p$ th power law of attraction, with  $p > 1/2$ , then  $Y_n \xrightarrow{d} Y$ , where  $\varphi_Y(t) = e^{-c|t|^{1/p}}$  for some  $c > 0$ .

**6.20.** For each  $n$ , suppose that  $X_n \stackrel{d}{=} U[-n, n]$ .

- a) Show that  $\varphi_{X_n}(t) = (\sin nt)/nt$ .  
 b) Show that

$$\lim_{n \rightarrow \infty} \varphi_{X_n}(t) = \begin{cases} 1 & t = 0 \\ 0 & t \neq 0. \end{cases}$$

- c) Prove that there is no subsequence  $(X_{n'})$  that converges in distribution.  
d) Discuss b) and c) in the context of Theorem 6.17.

**6.21.** Suppose that

$$\mathsf{F}_X(t) = \frac{e^t}{1 + e^t}, \quad t \in \mathbb{R}.$$

Calculate the characteristic function, expectation and variance of  $X$ .

**6.22.** Let  $X$  and  $Y$  be independent. Show that

$$\varphi_{XY}(t) = \int_{-\infty}^{\infty} \varphi_Y(tx) d\mathsf{F}_X(x) = \int_{-\infty}^{\infty} \varphi_X(ty) d\mathsf{F}_Y(y).$$

**6.23.** Calculate the density of a random variable  $X$  with  $\varphi_X(t) = [1 - |t|]^+$ .

**6.24.** Suppose that

$$\varphi_X(t) = \frac{3 \sin t}{t^3} - \frac{3 \cos t}{t^2}, \quad t \neq 0.$$

- a) Show that  $X$  is symmetric (Exercise 6.11).  
b) Show that for  $n \geq 1$ ,  $E[X^{2n}] = 3/(2n+1)(2n+3)$ .  
c) Prove that  $P\{|X| > 1\} = 0$ .  
d) Show that  $X$  is absolutely continuous.

**6.25.** Show that if  $E[|X|^k|Y|^\ell] < \infty$ , then

$$E[X^k Y^\ell] = i^{k+\ell} \frac{\partial^k}{\partial t_1^k} \frac{\partial^\ell}{\partial t_2^\ell} \varphi_{X,Y}(t_1, t_2) \Big|_{t_1=t_2=0},$$

where  $\varphi_{X,Y}$  is the joint characteristic function of  $X$  and  $Y$ .

**6.26.** Let  $X$  be a positive and integer-valued. Show that for each  $k$ ,

$$P\{X = k\} = k! \zeta_X^{(k)}(0).$$

**6.27.** Let  $X$  have moment generating function  $\psi_X$ . Prove that the *cumulant generating function*

$$\xi_X(t) = \log \psi_X(t)$$

satisfies  $E[X] = \xi'_X(0)$  and  $\text{Var}(X) = \xi''_X(0)$ .

# Chapter 7

# Classical Limit Theorems

This chapter treats the celebrated limit theorems for sums of independent random variables that culminate the “classical era” of probability, together with some of their many applications.

Several of the results are established via “ $\omega$ -wise reasoning.” That is, certain properties are shown to be satisfied for all  $\omega$  in some event of probability one, and then real analysis is used to show that each such  $\omega$  possesses other properties of interest.

## 7.1 Series of Independent Random Variables

In this section, we study almost sure convergence of series of independent random variables; “convergence” of the series means absolute convergence. In order to prove almost sure convergence, we employ the following Cauchy-like criterion, which is closely related to Proposition 5.6.

**Lemma 7.1.** The sequence  $(Y_k)$  converges almost surely if and only if

$$\lim_{n \rightarrow \infty} P \left\{ \sup_{k \geq 1} |Y_{n+k} - Y_n| > \varepsilon \right\} = 0$$

for every  $\varepsilon > 0$ .  $\square$

### 7.1.1 Kolmogorov’s inequality

Kolmogorov’s inequality is a *maximal inequality*: it provides an upper bound on tail probabilities for the maximum of partial sums of independent random variables.

**Theorem 7.2 (Kolmogorov’s inequality).** Let  $X_1, \dots, X_n$  be independent with  $E[X_i] = 0$  and  $0 < \sigma_i^2 = \text{Var}(X_i) < \infty$  for each  $i$ . Let

$S_k = \sum_{i=1}^k X_i$ . Then, for each  $\varepsilon > 0$ ,

$$P\{\max_{1 \leq k \leq n} |S_k| > \varepsilon\} \leq \frac{1}{\varepsilon^2} \sum_{i=1}^n \sigma_i^2. \quad (7.1)$$

Conversely, if there is  $c$  such that  $P\{|X_k| \leq c\} = 1$  for each  $k$ , then for each  $\varepsilon$ ,

$$P\{\max_{1 \leq k \leq n} |S_k| > \varepsilon\} \geq 1 - (c + \varepsilon)^2 / \sum_{i=1}^n \sigma_i^2. \quad (7.2)$$

**Proof:** The pattern of argument, a *first step decomposition*, is exceedingly useful in probability. In words, if  $\max_{k \leq n} |S_k| > \varepsilon$ , then there is a smallest value of  $k$  for which  $|S_k| > \varepsilon$ . Thus, the event

$$A = \{\max_{k \leq n} |S_k| > \varepsilon\} = \bigcup_{k=1}^n \{|S_k| \geq \varepsilon\}$$

is the disjoint union of the events

$$A_k = \{|S_1| \leq \varepsilon, \dots, |S_{k-1}| \leq \varepsilon, |S_k| > \varepsilon\}, \quad k = 1, \dots, n.$$

Consequently, since  $E[X_i] = 0$ ,

$$\sum_{i=1}^n \sigma_i^2 = \text{Var}(S_n) = E[S_n^2] \geq E[S_n^2; A] = \sum_{k=1}^n E[S_n^2; A_k].$$

For each  $k$ , with  $S_n = S_k + (S_n - S_k)$ ,

$$\begin{aligned} E[S_n^2; A_k] &= E[S_k^2; A_k] + 2E[(S_n - S_k)S_k; A_k] + E[(S_n - S_k)^2; A_k] \\ &\geq E[S_k^2; A_k] + 2E[(S_n - S_k)S_k; A_k] \\ &= E[S_k^2; A_k] \end{aligned}$$

[by the disjoint blocks theorem,  $S_n - S_k$  and  $S_k \mathbf{1}_{A_k}$  are independent, so that  $E[(S_n - S_k)S_k; A_k] = E[S_n - S_k]E[S_k; A_k] = 0$ ]

$$\geq \varepsilon^2 P(A_k),$$

since  $|S_k| \geq \varepsilon$  on  $A_k$ . Therefore,

$$\sum_{i=1}^n \sigma_i^2 \geq \varepsilon^2 \sum_{k=1}^n P(A_k) = \varepsilon^2 P(A).$$

To prove (7.2), we observe that

$$E[S_n^2; A] = E[S_n^2] - E[S_n^2; A^c] \geq E[S_n^2] - \varepsilon^2 + \varepsilon^2 P(A).$$

But,

$$\begin{aligned} E[S_n^2; A] &= \sum_{k=1}^n E[(S_k + [S_n - S_k])^2; A_k] \\ &= \sum_{k=1}^n E[S_k^2; A_k] + \sum_{k=1}^n E[(S_n - S_k)^2; A_k] \\ &\leq (\varepsilon + c)^2 \sum_{k=1}^n P(A_k) + \sum_{k=1}^n P(A_k) \sum_{j=k+1}^n \sigma_j^2 \end{aligned}$$

[on the event  $A_k$ ,  $|S_{k-1}| \leq \varepsilon$  and  $|S_k| \leq |S_{k-1}| + |X_k| \leq \varepsilon + c$ ]

$$\leq P(A) \{(\varepsilon + c)^2 + E[S_n^2]\}.$$

Hence,

$$P(A) \geq \frac{E[S_n^2] - \varepsilon}{(\varepsilon + c)^2 + E[S_n^2] - \varepsilon^2} \geq 1 - \frac{(\varepsilon + c)^2}{E[S_n^2]}. \quad \blacksquare$$

### 7.1.2 The three series theorem

The main result concerning almost sure convergence of series is the three series theorem, which provides necessary and sufficient conditions for almost sure convergence. However, the key result is Theorem 7.3, which is also used to prove the strong law of large numbers.

**Theorem 7.3.** Let  $X_1, X_2, \dots$  be independent with  $E[X_k] = 0$  for each  $k$ . If

$$\sum_{k=1}^{\infty} E[X_k^2] = \sum_{k=1}^{\infty} \text{Var}(X_k) < \infty, \quad (7.3)$$

then  $\sum_{k=1}^{\infty} X_k$  converges almost surely.

**Proof:** We prove that the partial sums  $S_n = \sum_{k=1}^n X_k$  satisfy the hypotheses of Lemma 7.1. For  $\varepsilon > 0$ , by Kolmogorov's inequality

$$\begin{aligned} P\{\sup_{k \geq 1} |S_{n+k} - S_n| > \varepsilon\} &= \lim_{m \rightarrow \infty} P\{\max_{k \leq m} |S_{n+k} - S_n| > \varepsilon\} \\ &\leq \lim_{m \rightarrow \infty} \frac{1}{\varepsilon^2} \sum_{k=n+1}^{n+m} E[X_k^2] \\ &= \frac{1}{\varepsilon^2} \sum_{k=n+1}^{\infty} E[X_k^2], \end{aligned}$$

which converges to zero as  $n \rightarrow \infty$  by (7.3). ■

Here is an example that has already been considered.

**Example 7.4.** Suppose that  $P\{X_k = \pm 2^{-k}\} = 1/2$  for each  $k$ . Since  $E[X_k] = 0$  and  $E[X_k^2] = 2^{-2k}$ , Theorem 7.3 implies that  $S = \sum_{k=1}^{\infty} X_k$  converges almost surely. By Example 3.7,  $S \xrightarrow{d} U[-1, 1]$ .  $\square$

To prove the three series theorem, we employ a form of truncation for random variables. Given  $X$  and  $c > 0$ , let

$$X^c = X \mathbf{1}(|X| \leq c) = \begin{cases} X & \text{if } |X| \leq c \\ 0 & \text{otherwise} \end{cases}$$

be the *truncation* of  $X$  at  $c$ , obtained by setting to zero values of  $|X|$  that exceed  $c$ .

We now establish that almost sure convergence of  $\sum_{k=1}^{\infty} X_k$  is equivalent to that of three series involving the truncations  $X_k^c$ .

**Theorem 7.5 (Three series theorem).** Let  $X_1, X_2, \dots$  be independent. If, for some  $c > 0$ ,

$$\sum_{k=1}^{\infty} P\{|X_k| > c\} < \infty \quad (7.4)$$

$$\sum_{k=1}^{\infty} |E[X_k^c]| < \infty \quad (7.5)$$

$$\sum_{k=1}^{\infty} \text{Var}(X_k^c) < \infty, \quad (7.6)$$

then  $\sum_{k=1}^{\infty} X_k$  converges almost surely.

Conversely, if  $\sum_{k=1}^{\infty} X_k$  converges almost surely, then (7.4), (7.5) and (7.6) hold for every  $c > 0$ .

**Proof:** Sufficiency: By (7.6) and Theorem 7.3,  $\sum_{k=1}^{\infty} (X_k^c - E[X_k^c])$  converges almost surely, which, together with (7.5), implies that  $\sum_{k=1}^{\infty} X_k^c$  converges almost surely. Finally, by (7.4) and the Borel-Cantelli lemma (Theorem 1.27),  $P\{X_k \neq X_k^c, \text{i.o.}\} = 0$ .

We now reason  $\omega$ -wise. For outcomes  $\omega$  in an event of probability one,

1.  $X_k(\omega) = X_k^c(\omega)$  for all sufficiently large values of  $k$ ,
2.  $\sum_{k=1}^{\infty} X_k^c(\omega)$  converges,

and for every such  $\omega$ ,  $\sum_{k=1}^{\infty} X_k(\omega)$  converges as well.

Necessity: Assume that  $\sum_{k=1}^{\infty} X_k$  converges almost surely. Then,  $X_k \xrightarrow{\text{a.s.}} 0$ , and, hence,  $P\{|X_k| > c, \text{i.o.}\} = 0$  for every  $c$ , so that (7.4) holds by

the second Borel-Cantelli lemma (Theorem 3.22). Also, again by  $\omega$ -wise reasoning, almost sure convergence of  $\sum_{k=1}^{\infty} X_k$  implies that of  $\sum_{k=1}^{\infty} X_k^c$  for each  $c$ .

To complete the proof, we use a symmetrization argument. With  $c$  now fixed, let  $\tilde{X}_1, \tilde{X}_2, \dots$  be i.i.d. with the same distributions as the  $X_k^c$ , and assume that the sequences  $(X_k^c)$  and  $(\tilde{X}_k)$  are independent. Then, almost sure convergence of  $\sum_{k=1}^{\infty} X_k^c$  implies that of  $\sum_{k=1}^{\infty} (X_k^c - \tilde{X}_k)$ .

But  $E[X_k^c - \tilde{X}_k] = 0$  and  $P\{|X_k^c - \tilde{X}_k| \leq 2c\} = 1$ , so that by (7.2), with  $S_n^* = \sum_{k=1}^n (X_k^c - \tilde{X}_k)$ , and with  $\varepsilon > 0$ ,

$$P\left\{\sup_{k \geq 1} |S_{n+k}^* - S_n^*| > \varepsilon\right\} \geq 1 - \frac{(c + \varepsilon)^2}{\sum_{k=1}^{\infty} \text{Var}(X_k^c - \tilde{X}_k)}.$$

If  $\sum_{k=1}^{\infty} \text{Var}(X_k^c - \tilde{X}_k)$  were infinite, we would have

$$P\left\{\sup_{k \geq 1} |S_{n+k}^* - S_n^*| \geq \varepsilon\right\} = 1,$$

contradicting almost sure convergence of  $\sum_{k=1}^{\infty} (X_k^c - \tilde{X}_k)$ , because the latter implies via Lemma 7.1 that once  $n$  is large enough

$$P\left\{\sup_{k \geq 1} |S_{n+k} - S_n| > \varepsilon\right\} < 1/2.$$

Consequently,

$$\sum_{k=1}^{\infty} \text{Var}(X_k^c) = \frac{1}{2} \sum_{k=1}^{\infty} \text{Var}(X_k^c - \tilde{X}_k) < \infty,$$

and (7.6) holds. This implies by Theorem 7.3 that  $\sum_{k=1}^{\infty} (X_k^c - E[X_k^c])$  converges almost surely, so that  $\sum_{k=1}^{\infty} E[X_k^c]$  does, which is (7.5). ■

## 7.2 The Strong Law of Large Numbers

Let  $X_1, X_2, \dots$  be i.i.d. with partial sums  $S_n = \sum_{i=1}^n X_i$ . In this section, we show that  $S_n/n \xrightarrow{\text{a.s.}} E[X_1]$  if (and only if, but we do not prove this)  $E[|X_1|] < \infty$ . The proof depends on an analytical link between convergence of series and convergence of normalized partial sums.

**Lemma 7.6 (Kronecker's lemma).** Let  $x_1, x_2, \dots$  be real numbers for  $\sum_{k=1}^{\infty} x_k/k$  converges, not necessarily absolutely. Then,

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n x_k = 0.$$

**Proof:** Let  $s_n = \sum_{k=1}^n (x_k/k)$  and  $s = \lim_n s_n$ . The argument employs “summation-by-parts.” For each  $n$ ,

$$\begin{aligned}\frac{1}{n} \sum_{k=1}^n x_k &= \frac{1}{n} \sum_{k=1}^n k(s_k - s_{k-1}) = \frac{1}{n} \sum_{k=1}^n \left( \sum_{j=0}^{k-1} 1 \right) (s_k - s_{k-1}) \\ &= \frac{1}{n} \sum_{j=0}^{n-1} \sum_{k=j+1}^n (s_k - s_{k-1}) \\ &= \frac{1}{n} \sum_{j=0}^{n-1} (s_n - s_j) \\ &= s_n - \frac{1}{n} \sum_{j=1}^{n-1} s_j,\end{aligned}$$

which converges to  $s - s = 0$ , since if a real sequence  $(y_n)$  converges, then its Cesàro averages  $(1/n) \sum_{i=1}^n y_i$  converge to the same limit. ■

Here is the first of the three great theorems.

**Theorem 7.7 (Strong law of large numbers).** Let  $X_1, X_2, \dots$  be i.i.d. If  $E[|X_1|] < \infty$ , then  $S_n/n \xrightarrow{\text{a.s.}} E[X_1]$ .

**Proof:** Without loss of generality, we assume that  $E[X_1] = 0$ , for otherwise we replace the  $X_i$  by  $X_i - E[X_1]$ .

The basis of the proof is a truncation that allows us to use Theorem 7.3, followed by  $\omega$ -wise application of Kronecker’s lemma (Lemma 7.6). To begin with, with  $Y_k = X_k \mathbf{1}(|X_k| \leq k)$ , we show that

$$\sum_{k=1}^{\infty} P\{X_k \neq Y_k\} < \infty \tag{7.7}$$

$$\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n E[Y_k] = 0 \tag{7.8}$$

and

$$\sum_{k=1}^{\infty} \frac{\text{Var}(Y_k)}{k^2} < \infty. \tag{7.9}$$

The proof of (7.7) is straightforward: since the  $X_k$  are identically distributed, then by (4.36),

$$\begin{aligned}\sum_{k=1}^{\infty} P\{X_k \neq Y_k\} &= \sum_{k=1}^{\infty} P\{|X_k| \geq k\} = \sum_{k=1}^{\infty} P\{|X_1| \geq k\} \\ &\leq E[|X_1|] + 1.\end{aligned}$$

For (7.8), it suffices to show that  $E[Y_k] \rightarrow 0$ . Again, using the property that the  $X_k$  are identically distributed,

$$E[Y_k] = E[X_k; \{|X_k| \leq k\}] = E[X_1; \{|X_1| \leq k\}] \rightarrow E[X_1] = 0$$

by the dominated convergence theorem.

This leaves the proof of (7.9). For each  $k$ ,

$$\begin{aligned} \text{Var}(Y_k) &\leq E[Y_k^2] = E[X_k^2; \{|X_k| \leq k\}] \\ &= E[X_1^2; \{|X_1| \leq k\}] \\ &= \sum_{j=1}^k E[X_1^2; \{j-1 < |X_1| \leq j\}], \end{aligned}$$

so that

$$\begin{aligned} \sum_{k=1}^{\infty} \frac{\text{Var}(Y_k)}{k^2} &\leq \sum_{k=1}^{\infty} \frac{1}{k^2} \sum_{j=1}^k E[X_1^2; \{j-1 < |X_1| \leq j\}] \\ &\leq \sum_{j=1}^{\infty} E[X_1^2; \{j-1 < |X_1| \leq j\}] \sum_{k=j}^{\infty} \frac{1}{k^2} \\ &\leq 2 \sum_{j=1}^{\infty} \frac{1}{j} E[X_1^2; \{j-1 < |X_1| \leq j\}] \\ &\leq 2E[|X_1|]. \end{aligned}$$

The proof now proceeds in the following manner. By (7.9) and Theorem 7.3, the series  $\sum_{k=1}^{\infty} (Y_k - E[Y_k])/k$  converges for almost surely. Hence, by Kronecker's lemma, applied  $\omega$ -wise, on an event of probability one,

$$\frac{1}{n} \sum_{k=1}^n (Y_k - E[Y_k]) \xrightarrow{\text{a.s.}} 0, \quad (7.10)$$

But then, (7.10) and (7.8) imply that on an event of probability one,

$$\frac{1}{n} \sum_{k=1}^n Y_k(\omega) \rightarrow 0. \quad (7.11)$$

Finally, (7.7) and the Borel-Cantelli lemma imply that for  $\omega$  in an event of probability one,

$$X_k(\omega) = Y_k(\omega) \text{ for all sufficiently large values of } k. \quad (7.12)$$

Any  $\omega$  satisfying both (7.11) and (7.12) satisfies  $S_n(\omega)/n \rightarrow 0$ . ■

There is a converse as well, which we state without proof.

**Theorem 7.8.** Let  $X_1, X_2, \dots$  be i.i.d. If  $E[|X_1|] = \infty$ , then for every  $a \in \mathbb{R}$ ,

$$\limsup_{n \rightarrow \infty} |S_n/n - a| \stackrel{\text{a.s.}}{=} \infty. \quad \square$$

### 7.3 The Central Limit Theorem

Let  $X_1, X_2, \dots$  be independent, but not necessarily identically distributed, with  $E[X_i^2] < \infty$  and  $E[X_i] = 0$  for each  $i$ , and let

$$\begin{aligned} S_n &= \sum_{i=1}^n X_i, \\ s_i^2 &= \text{Var}(X_i), \\ \sigma_n^2 &= \sum_{i=1}^n s_i^2 = \text{Var}(S_n). \end{aligned}$$

We are interested in identifying conditions implying that

$$\frac{S_n}{\sigma_n} = \frac{S_n - E[S_n]}{\sqrt{\text{Var}(S_n)}} \xrightarrow{\text{d}} N(0, 1), \quad (7.13)$$

in which case we say that the  $X_i$  satisfy the central limit theorem. Before developing these, however, we consider why the limit in (7.13) is a normal distribution.

Normality is forced by the nature of the limiting process. To understand this, let the  $X_i$  be i.i.d. with variance  $\sigma^2$ . Suppose that  $S_n/\sigma\sqrt{n} \xrightarrow{\text{d}} Z$ , in which case,  $S_{2n}/\sigma\sqrt{2n} \xrightarrow{\text{d}} Z$  as well. But also

$$\begin{aligned} \frac{S_{2n}}{\sigma\sqrt{2n}} &= \frac{X_1 + X_3 + \cdots + X_{2n-1}}{\sigma\sqrt{2n}} + \frac{X_2 + \cdots + X_{2n}}{\sigma\sqrt{2n}} \\ &\xrightarrow{\text{d}} \frac{Z_1 + Z_2}{\sqrt{2}}, \end{aligned}$$

where  $Z_1$  and  $Z_2$  are independent and  $Z_1 \stackrel{\text{d}}{=} Z_2 \stackrel{\text{d}}{=} Z$ . That  $Z$  must be normally distributed is a consequence of the following result.

**Proposition 7.9.** Let  $Z_1$  and  $Z_2$  be i.i.d. with  $E[Z_i] = 0$  and  $0 < \sigma^2 = \text{Var}(Z_i) < \infty$ . If

$$\frac{Z_1 + Z_2}{\sqrt{2}} \stackrel{\text{d}}{=} Z_1,$$

then  $Z_1 \stackrel{\text{d}}{=} N(0, \sigma^2)$ .

**Proof:** Let  $\varphi$  be the characteristic function of  $Z_1$  and  $Z_2$ . By hypothesis,  $\varphi(t) = \varphi(t/\sqrt{2})^2$  and, hence, by induction,

$$\varphi(t) = \varphi(t/2^{n/2})^{2^n} = \left[1 - \frac{t^2\sigma^2}{2 \cdot 2^n} + o\left(\frac{1}{2^n}\right)\right]^{2^n} \rightarrow e^{-\sigma^2 t^2/2},$$

where the expansion is by Theorem 6.14. Thus,  $Z_1 \xrightarrow{d} N(0, \sigma^2)$ . ■

Consequently, the limit in (7.13) must be normal. Once this is so, two things must happen in order that  $S_n/\sigma_n \xrightarrow{d} N(0, 1)$ . First, the variances  $\sigma_n$  must converge to  $\infty$ . If not,  $\sum_{k=1}^{\infty} \text{Var}(X_k) < \infty$ , and Theorem 7.3 implies that  $\sum_{k=1}^{\infty} X_k$  converges almost surely, in which case  $S_n$  converges to a finite limit, which can have any distribution. In this case there is “not enough randomness” to engender a normally distributed limit. Second, the asymptotic behavior of  $S_n$  must reflect *only* the property that  $S_n$  is the sum of independent random variables, and not their specific distributions. This is credible if  $X_1, \dots, X_n$  are all small compared to  $S_n$ , vanishingly so in the limit as  $n \rightarrow \infty$ . We term this property *uniform asymptotic negligibility*, and it is the key aspect of two conditions — the Lyapunov and Lindeberg conditions — under which the central limit theorem holds.

### 7.3.1 The Lyapunov condition

The Lyapunov condition is sufficient but not necessary for the  $X_i$  to satisfy the central limit theorem, while the Lindeberg condition (Definition 7.12) is, in the presence of one minor additional assumption, both necessary and sufficient.

**Definition 7.10 (Lyapunov condition).** The sequence  $(X_i)$  satisfies the *Lyapunov condition* if

$$E[|X_i|^3] < \infty$$

for each  $i$  and

$$\lim_{n \rightarrow \infty} \frac{1}{\sigma_n^3} \sum_{i=1}^n E[|X_i|^3] = 0. \quad \square \quad (7.14)$$

The Lyapunov condition entails that  $X_1, \dots, X_n$  be uniformly asymptotically negligible as  $n \rightarrow \infty$ , relative to their sum  $S_n$ . The most succinct expression of this is (7.17):  $\max_{k \leq n} s_k^2/\sigma_n^2 \rightarrow 0$ . This allows the asymptotic behavior of the  $S_n$  to reflect only their being the sum of independent random variables, so that the central limit is valid.

**Theorem 7.11 (Central limit theorem).** If  $(X_i)$  satisfies the Lyapunov condition, then  $S_n/\sigma_n \xrightarrow{d} N(0, 1)$  as  $n \rightarrow \infty$ .

**Proof:** We employ characteristic functions. By Theorem 6.15, it suffices to show that  $\varphi_{S_n/\sigma_n}(t) \rightarrow e^{-t^2/2}$  for each  $t \in \mathbb{R}$ . For each  $n$  and  $t$ ,

$$\begin{aligned} & \left| \varphi_{S_n/\sigma_n}(t) - e^{-t^2/2} \right| \\ &= \left| \prod_{k=1}^n \varphi_{X_k}(t/\sigma_n) - \prod_{k=1}^n e^{-t^2 s_k^2 / 2\sigma_n^2} \right| \\ &\leq \sum_{k=1}^n \left| \varphi_{X_k}(t/\sigma_n) - e^{-t^2 s_k^2 / 2\sigma_n^2} \right| \\ &\cong \sum_{k=1}^n \left| 1 - \frac{t^2 s_k^2}{2\sigma_n^2} + \frac{i^3 t^3 E[X_k^3]}{6\sigma_n^3} - \left[ 1 - \frac{t^2 s_k^2}{2\sigma_n^2} + \frac{t^4 s_k^4}{8\sigma_n^4} \right] \right| \end{aligned}$$

[by the Taylor expansion for characteristic functions in Theorem 6.14, and the expansion  $e^{-u^2/2} \cong 1 - u^2/2 + (1/2)(u^2/2)^2$ ]

$$\begin{aligned} &\leq \frac{|t|^3}{6\sigma_n^3} \sum_{k=1}^n E[|X_k|^3] + \frac{t^4}{8\sigma_n^4} \sum_{k=1}^n s_k^4 \\ &\leq \frac{|t|^3}{6\sigma_n^3} \sum_{k=1}^n E[|X_k|^3] + \frac{t^4}{8} \max_{k \leq n} \frac{s_k^2}{\sigma_n^2} \frac{1}{\sigma_n^2} \sum_{k=1}^n s_k^2 \\ &\leq \frac{|t|^3}{6\sigma_n^3} \sum_{k=1}^n E[|X_k|^3] + \frac{t^4}{8} \max_{k \leq n} \frac{s_k^2}{\sigma_n^2}. \end{aligned}$$

The Lyapunov condition is the statement that the first term in this last expression converges to zero as  $n \rightarrow \infty$ . Regarding the second term, for each  $n$ , let  $j_n = \arg \max_{j=1, \dots, n} s_j^2 / \sigma_n^2$ . Then,

$$(s_{j_n}/\sigma_n)^{3/2} = E[(X_{j_n}/\sigma_n)^2]^{3/2} \leq E[|X_{j_n}/\sigma_n|^3] \leq \frac{1}{\sigma_n^3} \sum_{k=1}^n E[|X_k|^3],$$

where the inequality is by Lyapunov's inequality (4.24), and this converges to zero by (7.14). ■

### 7.3.2 The Lindeberg condition

The Lindeberg condition is sufficient in order that the  $X_i$  satisfy the central limit theorem, and nearly necessary as well.

**Definition 7.12 (Lindeberg condition).** The sequence  $(X_i)$  satisfies the *Lindeberg condition* if for every  $\varepsilon > 0$

$$\lim_{n \rightarrow \infty} \frac{1}{\sigma_n^2} \sum_{i=1}^n E[X_i^2; \{|X_i| > \varepsilon \sigma_n\}] = 0. \quad \square \quad (7.15)$$

A sequence satisfying the Lindeberg condition is uniformly asymptotically negligible.

**Proposition 7.13.** If  $(X_i)$  satisfies the Lindeberg condition, then

$$\max_{k \leq n} \frac{X_k}{\sigma_n} \xrightarrow{\text{P}} 0 \quad (7.16)$$

$$\max_{k \leq n} \frac{s_k^2}{\sigma_n^2} \rightarrow 0. \quad (7.17)$$

**Proof:** By Boole's inequality (Proposition 1.24), for  $\varepsilon > 0$ ,

$$\begin{aligned} P\{\max_{k \leq n} |X_k|/\sigma_n > \varepsilon\} &= P(\bigcup_{k=1}^n \{|X_k| > \varepsilon\sigma_n\}) \\ &\leq \sum_{k=1}^n P\{|X_k| > \varepsilon\sigma_n\} \\ &= \sum_{k=1}^n E[\mathbf{1}(|X_k| > \varepsilon\sigma_n)] \\ &\leq \sum_{k=1}^n E\left[\frac{X_k^2}{\varepsilon^2\sigma_n^2}; \{|X_k| > \varepsilon\sigma_n\}\right], \end{aligned}$$

which converges to zero by (7.15).

To prove (7.17), define  $j_n = \arg \max_{k \leq n} s_k^2/\sigma_n^2$ . Then, for  $\varepsilon > 0$ ,

$$\begin{aligned} s_{j_n}^2/\sigma_n^2 &= E[X_{j_n}^2/\sigma_n^2] \\ &= E[X_{j_n}^2/\sigma_n^2; \{|X_{j_n}| \leq \varepsilon\sigma_n\}] + E[X_{j_n}^2/\sigma_n^2; \{|X_{j_n}| > \varepsilon\sigma_n\}] \\ &\leq \varepsilon^2 + \frac{1}{\sigma_n^2} \sum_{k=1}^n E[X_k^2; \{|X_k| > \varepsilon\sigma_n\}], \end{aligned}$$

which converges to zero by the Lindeberg condition. ■

The Lindeberg condition is implied by the Lyapunov condition.

**Proposition 7.14.** If  $(X_i)$  satisfies the Lyapunov condition, then  $(X_i)$  satisfies the Lindeberg condition. □

Here are two fairly broad examples.

**Example 7.15 (Uniformly bounded sequence).** If there is a constant  $c$  such that  $P\{|X_k| \leq c\} = 1$  for each  $k$ , and if  $\sigma_n^2 \rightarrow \infty$ , then the Lindeberg condition is satisfied. Indeed, given  $\varepsilon$ ,  $\sum_{k=1}^n E[X_k^2; \{|X_k| > \varepsilon\sigma_n\}] = 0$  once  $n$  is large enough that  $\varepsilon\sigma_n > c$ . □

**Example 7.16 (i.i.d. sequence).** If  $X_1, X_2, \dots$  are i.i.d. with variance  $s^2 \in (0, \infty)$ , then the Lindeberg condition is satisfied. In this case,  $\sigma_n^2 = ns^2$  for each  $n$ , so that for each  $\varepsilon$ ,

$$\frac{1}{\sigma_n^2} \sum_{k=1}^n E[X_k^2; \{|X_k| > \varepsilon \sigma_n\}] = \frac{1}{s^2} E[|X_1|^2; \{|X_1| > \varepsilon s\sqrt{n}\}],$$

which converges to zero by the dominated convergence theorem.  $\square$

Next is the second great limit theorem.

**Theorem 7.17 (Lindeberg-Feller central limit theorem).** If  $(X_i)$  satisfies the Lindeberg condition, then  $S_n/\sigma_n \xrightarrow{d} N(0, 1)$ .

**Proof:** By Theorem 5.8, it suffices to show that for every function  $f: \mathbb{R} \rightarrow \mathbb{R}$  with three bounded, uniformly continuous derivatives,

$$E[f(S_n/\sigma_n)] \rightarrow E[f(Z)], \quad (7.18)$$

where  $Z \stackrel{d}{=} N(0, 1)$ .

Assuming this much smoothness provides a Taylor expansion to which (ultimately) the Lindeberg condition is applicable. Specifically, with

$$g(h) \stackrel{\text{def}}{=} \sup_{h \in \mathbb{R}} \left| f(x + h) - \left[ f(x) + hf'(x) + \frac{h^2}{2} f''(x) \right] \right|, \quad (7.19)$$

there is a constant  $K$ , depending only on  $f$ , such that for all  $h$ ,

$$g(h) \leq K \min \{h^2, |h|^3\}. \quad (7.20)$$

In (7.20),  $h^2$  gives a better bound for large values of  $|h|$ , while  $|h|^3$  is a better bound for small values of  $|h|$ .

The argument is based on successively replacing the  $X_i$  by independent, normally distributed random variables with the same variances. To this end, let  $Z_1, Z_2, \dots$  be independent with  $Z_k \stackrel{d}{=} N(0, s_k^2)$  for each  $k$ , such that the sequences  $(X_i)$  and  $(Z_k)$  are independent. For  $k \leq n$ , define

$$U_n^{(k)} = X_1 + \cdots + X_{k-1} + Z_{k+1} + \cdots + Z_n;$$

note that there is no summand (yet) with index  $k$ . For each  $n$  and  $k$ ,  $U_n^{(k)}$ ,  $X_k$  and  $Z_k$  are independent by the disjoint blocks theorem. Moreover,

$$U_n^{(n)} + X_n = S_n \quad (7.21)$$

$$U_n^{(1)} + Z_1 = \sum_{k=1}^n Z_k \stackrel{d}{=} N(0, \sigma_n^2) \quad (7.22)$$

$$U_n^{(k)} + X_k = U_n^{(k+1)} + Z_{k+1}.$$

The remainder of the proof is a series of inequalities. First of all, by the triangle inequality, (7.21) and (7.22),

$$\begin{aligned} & |E[f(S_n/\sigma_n)] - E[f(Z)]| \\ &= \left| E\left[f\left(\frac{U_n^{(n)} + X_n}{\sigma_n}\right)\right] - E\left[f\left(\frac{U_n^{(1)} + Z_1}{\sigma_n}\right)\right] \right| \\ &\leq \sum_{k=1}^n \left| E\left[f\left(\frac{U_n^{(k)} + X_k}{\sigma_n}\right)\right] - E\left[f\left(\frac{U_n^{(k)} + Z_k}{\sigma_n}\right)\right] \right|. \end{aligned}$$

Next, for each  $n$  and  $k$ ,

$$\begin{aligned} & \left| E\left[f\left(\frac{U_n^{(k)} + X_k}{\sigma_n}\right)\right] - E\left[f\left(\frac{U_n^{(k)} + Z_k}{\sigma_n}\right)\right] \right| \\ &\leq E\left[\left| f\left(\frac{U_n^{(k)} + X_k}{\sigma_n}\right) - f\left(\frac{U_n^{(k)} + Z_k}{\sigma_n}\right) \right|\right] \\ &\leq E\left[\left| f\left(\frac{U_n^{(k)} + X_k}{\sigma_n}\right) - f\left(\frac{U_n^{(k)}}{\sigma_n}\right) \right|\right] \\ &\quad + E\left[\left| f\left(\frac{U_n^{(k)}}{\sigma_n}\right) - f\left(\frac{U_n^{(k)} + Z_k}{\sigma_n}\right) \right|\right] \\ &\leq E[g(X_k/\sigma_n)] + E[g(Z_k/\sigma_n)], \end{aligned}$$

where  $g$  is defined in (7.19), which we have applied with  $x = U_n^{(k)}/\sigma_n$  and  $h = X_k/\sigma_n$ , then with  $x = U_n^{(k)}/\sigma_n$  and  $h = Z_k/\sigma_n$ . Hence,

$$|E[f(S_n/\sigma_n)] - E[f(Z)]| \leq \sum_{k=1}^n \{E[g(X_k/\sigma_n)] + E[g(Z_k/\sigma_n)]\}. \quad (7.23)$$

We now apply (7.20) to estimate the terms on the right-hand side of (7.23). Given  $\varepsilon$ , for each  $n$  and  $k$ ,

$$\begin{aligned} & E[g(X_k/\sigma_n)] \\ &= E[g(X_k/\sigma_n); \{|X_k| \leq \varepsilon\sigma_n\}] + E[g(X_k/\sigma_n); \{|X_k| > \varepsilon\sigma_n\}] \\ &\leq KE[|X_k|^3/\sigma_n^3; \{|X_k| \leq \varepsilon\sigma_n\}] + \frac{K}{\sigma_n^2} E[X_k^2; \{|X_k| > \varepsilon\sigma_n\}] \end{aligned}$$

[by (7.20), with the  $|h|^3$  bound applied to the first term and the  $h^2$  bound to the second]

$$\leq \frac{K\varepsilon s_k^2}{\sigma_n^2} + \frac{K}{\sigma_n^2} E[X_k^2; \{|X_k| > \varepsilon\sigma_n\}],$$

so that

$$\sum_{k=1}^n E[g(X_k/\sigma_n)] \leq K\varepsilon + \frac{K}{\sigma_n^2} \sum_{k=1}^n E[X_k^2; \{|X_k| > \varepsilon\sigma_n\}].$$

By the Lindeberg condition, the first term in (7.23) can be made arbitrarily small by choosing  $n$  large enough.

Finally, to dispose of the second term in (7.23), by virtue of the argument in the preceding paragraph, it suffices to show that  $(Z_k)$  satisfies the Lindeberg condition. In fact, this sequence fulfills the Lyapunov condition, which implies by Proposition 7.14 that the Lindeberg condition holds as well. For each  $k$ ,  $Z_k \stackrel{d}{=} s_k Z$ , so that

$$\begin{aligned} \frac{1}{\sigma_n^3} \sum_{k=1}^n E[|Z_k|^3] &= \frac{1}{\sigma_n^3} \sum_{k=1}^n E[|s_k Z|^3] = E[|Z|^3] \frac{1}{\sigma_n^3} \sum_{k=1}^n |s_k|^3 \\ &\leq E[|Z|^3] \max_{k \leq n} \frac{s_k}{\sigma_n} \frac{1}{\sigma_n^2} \sum_{k=1}^n s_k^2 \\ &= E[|Z|^3] \max_{k \leq n} \frac{s_k}{\sigma_n}, \end{aligned}$$

which converges to zero by (7.17), while the latter holds because  $(X_i)$  satisfies the Lindeberg condition. ■

Without further restrictions, the Lindeberg condition is not necessary for validity of the central limit theorem. If, for each  $k$ ,  $X_k \stackrel{d}{=} N(0, 2^k)$ , then  $\sigma_n^2 \cong 2^{n+1}$  and  $\sigma_n^2/s_n^2 \rightarrow 1/2$ , so that (7.17) fails, and hence, so does the Lindeberg condition. However,  $S_n/\sigma_n \stackrel{d}{=} N(0, 1)$  for each  $n$ .

Once we stipulate (7.17), though, the Lindeberg condition is necessary.

**Theorem 7.18.** If  $X_1, X_2, \dots$  are independent with  $\max_{k \leq n} s_k^2/\sigma_n^2 \rightarrow 0$ , and if  $S_n/\sigma_n \stackrel{d}{\rightarrow} N(0, 1)$ , then  $(X_i)$  satisfies the Lindeberg condition. □

## 7.4 The Law of the Iterated Logarithm

The law of the iterated logarithm, the third great limit theorem, gives the precise rate of growth of the partial sums  $S_n$ : that rate is  $\sqrt{2n \log \log n}$ , from which the theorem takes its name.

The maximal inequality associated with the law of the iterated logarithm asserts that the maximum of the first  $n$  partial sums of independent, symmetric random variables (see Exercise 6.11) cannot be big unless the  $n$ th partial sum is at least moderately large as well.

**Proposition 7.19.** Let  $Y_1, \dots, Y_n$  be independent and symmetric with partial sums  $T_k = \sum_{i=1}^k Y_i$ . Then, for each  $a > 0$ ,

$$P\{\max_{k \leq n} T_k > a\} \leq 2P\{T_n > a\}.$$

**Proof:** As in the proof of Kolmogorov's inequality, we use a first step decomposition. Let  $A = \{\max_{k \leq n} T_k > a\}$  and

$$A_k = \{T_1 \leq a, \dots, T_{k-1} \leq a, T_k > a\},$$

so that  $A = \sum_{k=1}^n A_k$ . Then,  $\{T_n > a\} \subseteq A$ , and, consequently,

$$P\{T_n > a\} = \sum_{k=1}^n P(\{T_n > a\} \cap A_k) \geq \sum_{k=1}^n P(A_k \cap \{\sum_{i=k+1}^n Y_i \geq 0\})$$

[if  $S_k > a$ , which occurs on  $A_k$ , and  $\sum_{i=k+1}^n Y_i \geq 0$ , then  $T_n > a$ ]

$$= \sum_{k=1}^n P(A_k)P\{\sum_{i=k+1}^n Y_i \geq 0\}$$

[by the disjoint blocks theorem,  $A_k$  and  $\{\sum_{i=k+1}^n Y_i \geq 0\}$  are independent]

$$\begin{aligned} &\geq \frac{1}{2} \sum_{k=1}^n P(A_k) \\ &= \frac{1}{2} P(A), \end{aligned}$$

where the last inequality is by symmetry of the  $Y_i$ . ■

### 7.4.1 Normally distributed summands

Although the law of the iterated logarithm is valid more generally, we prove it solely for normally distributed summands. The analytical key is a pair of inequalities for the tail of the standard normal density function.

**Lemma 7.20.** For each  $t > 0$ ,

$$\frac{t}{1+t^2} e^{-t^2/2} \leq \int_t^\infty e^{-x^2/2} dx \leq \frac{1}{t} e^{-t^2/2}. \quad (7.24)$$

**Proof:** The genesis of both parts is the inequality

$$e^{-t^2/2} = t \int_t^\infty e^{-x^2/2} dx + \int_t^\infty \left[ \int_x^\infty e^{-y^2/2} dy \right] dx, \quad (7.25)$$

which is proved by computation:

$$\begin{aligned} \int_t^\infty \left[ \int_x^\infty e^{-y^2/2} dy \right] &= \int_t^\infty \left[ \int_t^y dx \right] e^{-y^2/2} dy \\ &= \int_t^\infty (y - t) e^{-y^2/2} dy \\ &= e^{-t^2/2} - t \int_t^\infty e^{-y^2/2} dy. \end{aligned}$$

Given (7.25), the right-hand inequality in (7.24) is immediate:

$$e^{-t^2/2} \geq t \int_t^\infty e^{-x^2/2} dx$$

since the second term on the right-hand side of (7.25) is positive.

Finally, substituting  $\int_x^\infty e^{-y^2/2} dy \leq (1/x)e^{-x^2/2}$  in (7.25) gives

$$\begin{aligned} e^{-t^2/2} &\leq t \int_t^\infty e^{-x^2/2} dx + \int_t^\infty \frac{1}{x} e^{-x^2/2} dx \\ &\leq t \int_t^\infty e^{-x^2/2} dx + \int_t^\infty \frac{1}{t} e^{-x^2/2} dx \\ &= \left( t + \frac{1}{t} \right) \int_t^\infty e^{-x^2/2} dx, \end{aligned}$$

which is the left-hand inequality in (7.24). ■

**Theorem 7.21 (Law of the iterated logarithm).** If  $X_1, X_2, \dots$  are i.i.d. with distribution  $N(0, 1)$ , then

$$\limsup_{n \rightarrow \infty} \frac{S_n}{\sqrt{2n \log \log n}} \stackrel{\text{a.s.}}{=} 1. \quad (7.26)$$

**Proof:** We first prove the *upper bound* that for every  $\varepsilon > 0$ ,

$$P \left\{ S_n > (1 + \varepsilon) \sqrt{2n \log \log n}, \text{ i.o.} \right\} \stackrel{\text{a.s.}}{=} 0, \quad (7.27)$$

and then the *lower bound* that for every  $\varepsilon > 0$ ,

$$P \left\{ S_n > (1 - \varepsilon) \sqrt{2n \log \log n}, \text{ i.o.} \right\} \stackrel{\text{a.s.}}{=} 1. \quad (7.28)$$

Both are proved by recourse to a Borel-Cantelli lemma; the latter is harder because of the need to devise independent events, the sum of whose probabilities diverges.

Upper bound: Let  $\varepsilon$  be fixed, and let  $n_k = \lfloor (1 + \varepsilon)^k \rfloor$ , where  $\lfloor x \rfloor$  is the integer part of  $x$  (the largest integer not exceeding  $x$ ). This sequence grows geometrically, even if just barely so. Also, the events

$$A_k = \left\{ S_n > (1 + \varepsilon) \sqrt{2n \log \log n} \text{ for some } n \in (n_k, n_{k+1}] \right\},$$

have the property that

$$\left\{ S_n > (1 + \varepsilon) \sqrt{2n \log \log n}, \text{ i.o.} \right\} = \{ A_k, \text{ i.o.} \}.$$

By the Borel-Cantelli lemma (Theorem 1.27), (7.27) holds provided that  $\sum_{k=1}^{\infty} P(A_k) < \infty$ , which we prove by showing that  $P(A_k) = O(k^{-(1+\varepsilon)})$ .

For each  $k$ ,

$$\begin{aligned} P(A_k) &\leq P \left\{ S_n > \sqrt{2n_k \log \log n_k} \text{ for some } n \in (n_k, n_{k+1}] \right\} \\ &\leq P \left\{ S_n > \sqrt{2n_k \log \log n_k} \text{ for some } n \leq n_{k+1} \right\} \\ &\leq 2P \left\{ S_{n_{k+1}} > \sqrt{2n_k \log \log n_k} \right\} \end{aligned}$$

[by Proposition 7.19]

$$\begin{aligned} &= 2P \left\{ S_{n_{k+1}} / \sqrt{n_{k+1}} > (1 + \varepsilon) \sqrt{2n_k \log \log n_k / n_{k+1}} \right\} \\ &\leq C \exp \left[ -\frac{1}{2}(1 + \varepsilon)^2 \frac{2n_k \log \log n_k}{n_{k+1}} \right] \end{aligned}$$

[by (7.24), where  $C$  is a constant, since  $S_{n_{k+1}} / \sqrt{n_{k+1}} \xrightarrow{d} N(0, 1)$ ]

$$\begin{aligned} &= O \left( e^{-(1+\varepsilon) \log \log n_k} \right) \\ &= O \left( (\log n_k)^{-(1+\varepsilon)} \right) \end{aligned}$$

[since  $n_k/n_{k+1} \rightarrow 1$ ]

$$= O \left( k^{-(1+\varepsilon)} \right).$$

Lower bound: We first note that (7.28) holds for given  $\varepsilon$  if, for any subsequence  $n_k$ ,

$$P \left\{ S_{n_k} > (1 - \varepsilon) \sqrt{2n_k \log \log n_k}, \text{ i.o.} \right\} \stackrel{\text{a.s.}}{=} 1.$$

Given  $\varepsilon$ , let an integer  $N > 1$  and  $c < 1$  be chosen so that

$$c\sqrt{1 - 1/N} - 2/\sqrt{N} > 1 - \varepsilon,$$

which is true if  $c$  is near enough to 1 and  $N$  is large enough. Then, let  $n_k = N^k$  (a very sparse geometric sequence),  $\Delta n_k = n_k - n_{k-1}$ , and

$$A_k = \left\{ S_{n_k} - S_{n_{k-1}} > c\sqrt{\Delta n_k \log \log \Delta n_k} \right\}.$$

The  $A_k$  are independent events by the disjoint blocks theorem (Theorem 3.10) and by calculations similar to those for the upper bound, appealing instead to the left-hand inequality in (7.24), one obtains that for some constant  $C$ ,  $P(A_k) \geq Ck^{-c^2}$ . Consequently,  $\sum_{k=1}^{\infty} P(A_k) = \infty$ , so that  $P\{A_k, \text{i.o.}\} = 1$  by Theorem 3.22, the second Borel-Cantelli lemma.

But since the  $X_k$  are symmetric, the upper bound applies to the sequence  $(-X_k)$  with  $\varepsilon = 1$ . In view of the preceding paragraph, there is an event  $A$ , with  $P(A) = 1$ , on which

1.  $S_{n_k} - S_{n_{k-1}} > c\sqrt{2\Delta n_k \log \log \Delta n_k}$  for infinitely many values of  $k$ .
2.  $S_{n_j} > -2\sqrt{n_j \log \log n_j}$  for all sufficiently large values of  $j$ .

Hence, for  $\omega \in A$ , for infinitely many values of  $k$ ,

$$\begin{aligned} S_{n_k}(\omega) &> S_{n_{k-1}}(\omega) + c\sqrt{2\Delta n_k \log \log \Delta n_k} \\ &\geq -2\sqrt{2n_k \log \log n_k} + c\sqrt{2\Delta n_k \log \log \Delta n_k} \\ &> (1 - \varepsilon)\sqrt{2n_k \log \log n_k}. \quad \blacksquare \end{aligned}$$

### 7.4.2 More general versions

Intuitively at least, it is plausible that random variables satisfying the central limit theorem also satisfy the law of the iterated logarithm, since the estimates given in Lemma 7.20, which are exact for normally distributed summands, will be “valid asymptotically” by the central limit theorem. Under appropriate assumptions, this hold true, but is a deep and difficult result, which we state without proof.

**Theorem 7.22 (Law of the iterated logarithm).** Suppose that  $X_1, X_2, \dots$  are i.i.d. with  $E[X_1] = 0$  and  $0 < \sigma^2 = \text{Var}(X_1) < \infty$ . Then,

$$\limsup_{n \rightarrow \infty} \frac{S_n}{\sqrt{2n \log \log n}} \stackrel{\text{a.s.}}{=} \sigma. \quad \square$$

## 7.5 Applications of the Limit Theorems

Applications of the strong law of large numbers, central limit theorem and law of the iterated logarithm are ubiquitous in probability and statistics. We present but a few.

### 7.5.1 Monte Carlo integration

Let  $h$  be a continuous (or even just Borel measurable) function on  $[0, 1]$ . It is necessary in a variety of settings to approximate the Riemann (or Lebesgue) integral  $\int_0^1 h(x) dx$ . There are, to be sure, many quadrature methods available, but the “Monte Carlo” technique is one of the simplest. Moreover, even though it may not be among the best methods for functions on  $[0, 1]$ , it extends to multi-dimensional integrals.

The basis of the Monte Carlo method is the property that if  $U \stackrel{d}{=} U[0, 1]$ , then, by Theorem 4.26,

$$E[h(U)] = \int_0^1 h(x) dx.$$

In addition, as discussed in §2.6, random number generators on computers produce values whose properties mimic those of i.i.d. uniform random variables, so the technique is natural for computer implementation.

**Theorem 7.23.** Let  $h$  be a function on  $[0, 1]$  with  $\int_0^1 |h(x)| dx < \infty$ , and let  $U_1, U_2, \dots$  be independent with distribution  $U[0, 1]$ . Then,  $(1/n) \sum_{i=1}^n h(U_i) \xrightarrow{\text{a.s.}} \int_0^1 h(x) dx$ .

**Proof:** We apply Theorem 7.7 to the random variables  $X_i = h(U_i)$ , which are also i.i.d. It applies because  $E[|X_1|] = \int_0^1 |h(x)| dx < \infty$ . ■

### 7.5.2 Maximum likelihood estimation

Let  $X_1, X_2, \dots$  be i.i.d. random variables, the “data.” Suppose that the  $X_i$  have density function  $f(\cdot, \theta)$ , where  $\theta \in \mathbb{R}$  is an unknown parameter, which we wish to determine. For concreteness, one might think of  $\theta$  as the mean of a normal distribution with known variance 1.

Maximum likelihood estimation is a central concept in statistics. Suppose that we have observed the data  $X_1, \dots, X_n$  and wish to estimate  $\theta$ . On the basis of (2.2) and (3.5), the function  $\Lambda_n(\theta) = \prod_{i=1}^n f(X_i, \theta)$ , known as the *likelihood function*, can be interpreted as the probability of the observations when the value of the parameter is  $\theta$ . The *maximum likelihood estimator* is the value of  $\theta$  that maximizes the likelihood function. This value,

$$\hat{\theta}_n = \arg \max_{\theta \in \mathbb{R}} \Lambda_n(\theta),$$

is a random variable because it is a function of  $X_1, \dots, X_n$ .

The maximum likelihood estimator depends as well on the *sample size*  $n$ , and it is natural to consider the large sample (asymptotic) behavior of these estimators as  $n \rightarrow \infty$ . Under widely satisfied assumptions, maximum likelihood estimators are *consistent*, in the sense that  $\hat{\theta}_n \xrightarrow{P} \theta$  as

$n \rightarrow \infty$ , and *asymptotically normal*: there is a constant  $\sigma^2(\theta)$  such that  $\sqrt{n}[\hat{\theta}_n - \theta] \xrightarrow{d} N(0, \sigma^2(\theta))$  as  $n \rightarrow \infty$ . We sketch how the limit theorems of probability are used to establish these properties.

Rather than the likelihood functions, one deals instead with *log-likelihood functions*  $L_n(\theta) = \sum_{i=1}^n \log f(X_i, \theta)$ , which are sums of i.i.d. random variables, as well as their first two derivatives with respect to  $\theta$ :

$$L'_n(\theta) = \sum_{i=1}^n \frac{f'(X_i, \theta)}{f(X_i, \theta)}, \quad (7.29)$$

known as the *score function*, and  $L''_n(\theta)$ . Under regularity conditions, the maximum likelihood estimator  $\hat{\theta}_n$  satisfies the *likelihood equation*

$$L'_n(\hat{\theta}_n) = 0. \quad (7.30)$$

Consistency of maximum likelihood estimators under general conditions is established in the following manner. In specific cases where these conditions fail, special structure may suffice; see Exercise 7.25.

**Theorem 7.24.** Suppose that

- i) The mapping  $\theta \mapsto f(x, \theta)$  is continuous for (almost) every  $x \in \mathbb{R}$ .
- ii) For each  $\theta$  and  $\gamma > 0$ ,

$$k_\theta(\gamma) \stackrel{\text{def}}{=} \inf_{|\theta' - \theta| > \gamma} \int_{-\infty}^{\infty} \left[ \sqrt{f(x, \theta)} - \sqrt{f(x, \theta')} \right]^2 dx > 0. \quad (7.31)$$

- iii) For each  $\theta$ ,  $\omega_\theta(\delta) \rightarrow 0$  as  $\delta \rightarrow 0$ , where

$$\omega_\theta(\delta) \stackrel{\text{def}}{=} \left[ \int_{-\infty}^{\infty} \sup_{|h| \leq \delta} \left( \sqrt{f(x, \theta)} - \sqrt{f(x, \theta + h)} \right)^2 dx \right]^{1/2}. \quad (7.32)$$

- iv) For each  $\theta$ ,  $\lim_{c \rightarrow \infty} \int_{-\infty}^{\infty} \sup_{|u| \geq c} \left[ \sqrt{f(x, \theta)} \sqrt{f(x, \theta + u)} \right] dx < 1$ .

Then,  $\hat{\theta}_n \xrightarrow{P} \theta$ .

**Proof:** With

$$Z_n(u) = \prod_{i=1}^n \frac{f(X_i, \theta + u)}{f(X_i, \theta)},$$

we will show that  $\sup_{|u| > \gamma} Z_n(u) \xrightarrow{P} 0$  for each  $\gamma > 0$ . Once this is established, consistency ensues. Indeed, if  $|\hat{\theta} - \theta| > \gamma$ , then  $\sup_{|u| > \gamma} Z_n(u)$  exceeds  $Z_n(v)$  for every  $v$  with  $|v| \leq \gamma$ . Therefore,

$$\begin{aligned} P\{|\hat{\theta} - \theta| > \gamma\} &\leq P\left\{\sup_{|u| > \gamma} Z_n(u) > Z_n(0)\right\} \\ &= P\left\{\sup_{|u| > \gamma} Z_n(u) > 1\right\}, \end{aligned}$$

yielding  $\hat{\theta} \xrightarrow{P} \theta$ .

We now establish that  $\sup_{|u|>\gamma} Z_n(u) \xrightarrow{P} 0$ . For simplicity, suppose that  $\theta$  lies in a known compact subset  $\Theta$  of  $\mathbb{R}$ . The extensions to  $\Theta = \mathbb{R}$  and  $\Theta = \mathbb{R}^d$ , the former straightforward (although it is where condition iv) is used), and the latter different only notationally, are discussed in Ibragimov/Has'minskii (1981).

With  $\gamma$  fixed, let  $J = [u_0 - \delta, u_0 + \delta]$  be an interval disjoint from  $[-\gamma/2, \gamma/2]$ . We will derive an upper bound for  $E[\sup_{u \in J} Z_n(u)]$ . For each  $j$  and for  $u \in J$ ,

$$\begin{aligned} \sqrt{f(X_j, \theta + u)} &\leq \sqrt{f(X_j, \theta + u_0)} \\ &+ \sup_{|h| \leq \delta} \left| \sqrt{f(X_j, \theta + u_0 + h)} - \sqrt{f(X_j, \theta + u_0)} \right| \end{aligned}$$

and consequently, where  $X$  also has density  $f(\cdot, \theta)$  (Note the presence of the  $n$ th root on the left-hand side!),

$$\begin{aligned} E \left[ \sup_{u \in J} \sqrt{Z_n(u)} \right]^{1/n} &\leq E \left[ \frac{\sqrt{f(X, \theta + u_0)}}{\sqrt{f(X, \theta)}} \right] \\ &+ E \left[ \frac{\sup_{|h| \leq \delta} \left| \sqrt{f(X, \theta + u_0 + h)} - \sqrt{f(X, \theta + u_0)} \right|}{\sqrt{f(X, \theta)}} \right]. \end{aligned}$$

We consider these terms individually.

Regarding the first,

$$\begin{aligned} E \left[ \frac{\sqrt{f(X, \theta + u_0)}}{\sqrt{f(X, \theta)}} \right] &= \int \frac{\sqrt{f(x, \theta + u_0)}}{\sqrt{f(x, \theta)}} f(x, \theta) dx \\ &= \int \sqrt{f(x, \theta + u_0)} \sqrt{f(x, \theta)} dx \\ &= \frac{1}{2} \left[ \int f(x, \theta + u_0) dx + \int f(x, \theta) dx \right. \\ &\quad \left. - \int (\sqrt{f(x, \theta + u_0)} - \sqrt{f(x, \theta)})^2 dx \right] \\ &= \frac{1}{2} \left[ 2 - \int (\sqrt{f(x, \theta + u_0)} - \sqrt{f(x, \theta)})^2 dx \right] \\ &\leq 1 - \frac{1}{2} k_\theta(\gamma/2). \end{aligned}$$

By the Cauchy-Schwarz inequality,

$$\begin{aligned} E & \left[ \frac{\sup_{|h| \leq \delta} |\sqrt{f(X, \theta + u_0 + h)} - \sqrt{f(X, \theta + u_0)}|}{\sqrt{f(X, \theta)}} \right] \\ & \leq E \left[ \left( \frac{\sup_{|h| \leq \delta} |\sqrt{f(X, \theta + u_0 + h)} - \sqrt{f(X, \theta + u_0)}|}{\sqrt{f(X, \theta)}} \right)^2 \right]^{1/2} \\ & = \int \frac{\sup_{|h| \leq \delta} |\sqrt{f(x, \theta + u_0 + h)} - \sqrt{f(x, \theta + u_0)}|^2}{f(x, \theta)} f(x, \theta) dx \\ & \leq \omega_{\theta+u_0}(\delta) \end{aligned}$$

Thus, by (7.31) and (7.32) (using also the inequality  $1+a \leq e^a$ ,  $a \in \mathbb{R}$ ),

$$E \left[ \sup_{u \in J} \sqrt{Z_n(u)} \right] \leq \exp \left[ -n \left( \frac{1}{2} k_\theta(\gamma/2) - \omega_{\theta+u_0}(\delta) \right) \right].$$

For each  $v \in \Theta \setminus [-\gamma, \gamma]$ , then, there is an interval  $J(v)$ , centered at  $v$ , for which

$$\sup_{u \in J(v)} Z_n(u) \xrightarrow{P} 0. \quad (7.33)$$

By compactness of  $\Theta$ , the class of intervals  $J(v)$ ,  $v \in \Theta \setminus [-\gamma, \gamma]$ , contains a finite subset  $J(v_1), \dots, J(v_k)$  such that  $\Theta \setminus [-\gamma, \gamma] \subseteq \bigcup_{\ell=1}^k J(v_\ell)$ . Since

$$\sup_{|u| > \gamma} Z_n(u) \leq \sum_{\ell=1}^k \sup_{u \in J(v_\ell)} Z_n(u),$$

(7.33) and Theorem 5.19 yield  $\sup_{|u| > \gamma} Z_n(u) \xrightarrow{P} 0$ . ■

Asymptotic normality of the estimation error  $\hat{\theta}_n - \theta$  is demonstrated in the following central limit theorem.

**Theorem 7.25.** Suppose that the assumptions of Theorem 7.24 are fulfilled, that the maximum likelihood estimators satisfy the likelihood equation (7.30) and that the *Fisher information*

$$I(\theta) \stackrel{\text{def}}{=} E \left[ \left( \frac{d}{d\theta} \log f(X_1, \theta) \right)^2 \right]$$

is finite. Then,  $\sqrt{n}[\hat{\theta}_n - \theta] \xrightarrow{d} N(0, 1/I(\theta))$ .

**Proof:** We use a Taylor expansion of the score function, which is valid under the indicated assumptions:

$$L'_n(\theta) = L'_n(\hat{\theta}) + [\hat{\theta} - \theta] L''_n(\theta^*) = [\hat{\theta} - \theta] L''_n(\theta^*),$$

where  $\theta^*$  lies between  $\hat{\theta}$  and  $\theta$ . That is,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{f'(X_i, \theta)}{f(X_i, \theta)} = \sqrt{n}[\hat{\theta} - \theta] \frac{1}{n} \sum_{i=1}^n \frac{d^2}{d\theta^2} \log f(X_i, \theta^*).$$

By the weak law of large numbers and consistency (Theorem 7.24),

$$\frac{1}{n} \sum_{i=1}^n \frac{d^2}{d\theta^2} \log f(X_i, \theta^*) \xrightarrow{P} E_\theta \left[ \frac{d^2}{d\theta^2} \log f(X_1, \theta) \right] = -I(\theta).$$

The random variables  $f'(X_i, \theta)/f(X_i, \theta)$  are i.i.d. with mean zero and finite  $P_\theta$ -variance  $I(\theta)$ , so by Theorem 7.17,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{f'(X_i, \theta)}{f(X_i, \theta)} \xrightarrow{d} N(0, I(\theta)).$$

Thus, by Theorem 5.22,

$$\begin{aligned} \sqrt{n}[\hat{\theta}_n - \theta] &= \frac{(1/\sqrt{n}) \sum_{i=1}^n f'(X_i, \theta)/f(X_i, \theta^*)}{(1/n) \sum_{i=1}^n (d^2/d\theta^2) \log f(X_i, \theta^*)} \\ &\xrightarrow{d} -\frac{N(0, I(\theta))}{I(\theta)} \\ &\stackrel{d}{=} N(0, 1/I(\theta)). \quad \blacksquare \end{aligned}$$

### 7.5.3 Empirical distribution functions

Let  $X_1, X_2, \dots$  be i.i.d. with *entirely unknown* distribution function  $F$ , which we seek to estimate based on observation of the  $X_i$ .

In the absence of even partial knowledge of  $F$ , only very general tools are available. Since for each  $t$ ,  $F(t) = P\{X_1 \leq t\} = E[1(X_1 \leq t)]$ , a plausible estimator of this value, given the data  $X_1, \dots, X_n$ , is the relative frequency with which the  $X_i$  are less than or equal to  $t$ .

**Definition 7.26.** The *empirical distribution function* for  $X_1, \dots, X_n$  is

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(X_i \leq t), \quad t \in \mathbb{R}. \quad \square$$

For each  $n$ ,  $\hat{F}_n$  is a random distribution function that jumps by  $1/n$  at each of the data values  $X_1, \dots, X_n$ .

Theorem 5.31 (or Theorem 7.7) provides one justification for empirical distribution functions: for *fixed*  $t$ ,

$$\hat{F}_n(t) = \frac{1}{n} \sum_{k=1}^n \mathbf{1}(X_k \leq t) \xrightarrow{\text{a.s.}} E[\mathbf{1}(X_1 \leq t)] = F(t). \quad (7.34)$$

In fact, the convergence takes place uniformly in  $t$ , as we now establish. Once again, the reasoning occurs “ $\omega$ -wise,” and here is the relevant analytical preliminary. See Chung (1974) for the proof.

**Lemma 7.27.** Let  $G, G_1, G_2, \dots$  be distribution functions on  $\mathbb{R}$  with

- i)  $G_n(r) \rightarrow G(r)$  for every rational  $r$ .
- ii)  $\Delta G_n(t) \rightarrow \Delta G(t)$  ( $\stackrel{\text{def}}{=} G(t) - G(t-)$ ) for every  $t$  such that  $\Delta G(t) > 0$ .

Then,  $G_n \rightarrow G$  uniformly.  $\square$

The next result applies this lemma to establish uniform convergence of empirical distribution functions to the “true” distribution function of the data.

**Theorem 7.28 (Glivenko-Cantelli theorem).** Let  $X_1, X_2, \dots$  be i.i.d. with distribution function  $F$ . Then,

$$\sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)| \xrightarrow{\text{a.s.}} 0. \quad (7.35)$$

**Proof:** Again reasoning  $\omega$ -wise, we show that with probability one, the hypotheses of Lemma 7.27 are fulfilled by  $G_n = \hat{F}_n$  and  $G = F$ , which implies (7.35).

Fix  $r \in \mathbb{Q}$ . Then, by (7.34), there is an event  $A_r$  with  $P(A_r) = 1$  such that for  $\omega \in A_r$ ,  $\hat{F}_n(r, \omega) \rightarrow F(r)$ .

Similarly, if  $\Delta F(t) > 0$ , and there are only countably many values of  $t$  for which this occurs, then since  $\Delta \hat{F}_n(t) = (1/n) \sum_{i=1}^n \mathbf{1}(X_i = t)$ , Theorem 7.7 implies that  $\Delta \hat{F}_n(t) \xrightarrow{\text{a.s.}} E[\mathbf{1}(X_1 = t)] = \Delta F(t)$ , so that there is an event  $B_t$  with  $P(B_t) = 1$  such that for  $\omega \in B_t$ ,  $\Delta \hat{F}_n(t, \omega) \rightarrow \Delta F(t)$ .

But then the event  $C = (\bigcap_{r \in \mathbb{Q}} A_r) \cap (\bigcap_{t: \Delta F(t) > 0} B_t)$  is almost sure, and for  $\omega \in C$ ,  $G_n = \hat{F}_n(\omega)$  and  $G = F$  satisfy the hypotheses of Lemma 7.27. ■

The central limit theorem corresponding to (7.34) follows from any of the central limit theorems we have presented. For fixed  $t$ ,

$$\sqrt{n} [\hat{F}_n(t) - F(t)] \xrightarrow{d} N(0, F(t)[1 - F(t)]). \quad (7.36)$$

Here, without proof, is the central limit theorem that accompanies the Glivenko-Cantelli theorem.

**Theorem 7.29 (Kolmogorov-Smirnov theorem).** Let  $X_1, X_2, \dots$  be i.i.d. with continuous distribution function  $F$ . Then,

$$\sqrt{n} \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)| \xrightarrow{d} Z,$$

where

$$P\{Z \leq t\} = 1 - 2 \sum_{j=1}^{\infty} (-1)^{j+1} e^{-2j^2 t^2}, \quad t > 0. \quad \square$$

There is even a law of the iterated logarithm for empirical distribution functions, whose proof we also omit.

**Theorem 7.30.** Under the context and assumptions of Theorem 7.29,

$$\limsup_{n \rightarrow \infty} \frac{\sqrt{n} \sup_{t \in \mathbb{R}} |\hat{F}_n(t) - F(t)|}{\sqrt{2 \log \log n}} \stackrel{\text{a.s.}}{=} \sup_{x \in \mathbb{R}} \sqrt{F(x)[1 - F(x)]}. \quad \square$$

#### 7.5.4 Random sums of independent random variables

Let  $X_1, X_2, \dots$  be i.i.d. with  $E[X_i] = 0$ , variance  $\sigma^2 \in (0, \infty)$ ; and partial sums  $S_n = \sum_{i=1}^n X_i$ ; let  $N_1 \leq N_2 \leq \dots$  be positive and integer-valued with  $N_n \xrightarrow{\text{a.s.}} \infty$ . Then, the process  $S_{N_n} = \sum_{i=1}^{N_n} X_i$  is a *random sum of random variables* and we are interested in whether it satisfies the strong law of large numbers and central limit theorem. No assumption is made regarding independence (or lack thereof) between  $(X_i)$  and  $(N_n)$ .

If  $(N_n)$  satisfies a strong law of large numbers, then  $(S_{N_n})$  does as well.

**Proposition 7.31.** If there are constants  $a_n \uparrow \infty$  and  $c > 0$  such that  $N_n/a_n \xrightarrow{\text{a.s.}} c$ , then  $S_{N_n}/N_n \xrightarrow{\text{a.s.}} 0$  and  $S_{N_n}/a_n \xrightarrow{\text{a.s.}} 0$ .

**Proof:** Once more, we reason  $\omega$ -wise. For  $\omega$  belonging to a set of probability one,  $N_n(\omega)/a_n \rightarrow c$  (in particular,  $N_n(\omega) \rightarrow \infty$ ) and  $S_k(\omega)/k \rightarrow 0$ . For such an  $\omega$ , it is trivial that  $S_{N_n(\omega)}(\omega)/N_n(\omega) \rightarrow 0$ . Moreover,

$$\frac{S_{N_n(\omega)}(\omega)}{a_n} = \frac{N_n(\omega)}{a_n} \frac{S_{N_n(\omega)}}{N_n(\omega)} \rightarrow c \times 0 = 0. \quad \blacksquare$$

If  $(N_n)$  satisfies a weak law of large numbers, then  $(S_{N_n})$  satisfies the central limit theorem.

**Theorem 7.32.** Assume that  $\sigma^2 = 1$ . If there are constants  $a_n \uparrow \infty$  and  $c > 0$  such that  $N_n/a_n \xrightarrow{P} c$ , then  $S_{N_n}/\sqrt{N_n} \xrightarrow{d} N(0, 1)$  and  $S_{N_n}/\sqrt{a_n} \xrightarrow{d} N(0, c^2)$ .

**Proof:** Let  $k_n = \lfloor ca_n \rfloor$  be the integer part of  $ca_n$ , so that  $N_n/k_n \xrightarrow{P} c$ . For each  $n$ ,

$$\frac{S_{N_n}}{\sqrt{N_n}} = \sqrt{\frac{k_n}{N_n}} \left[ \frac{S_{k_n}}{\sqrt{k_n}} + \frac{S_{N_n} - S_{k_n}}{\sqrt{k_n}} \right],$$

and since  $S_{k_n}/\sqrt{k_n} \xrightarrow{d} N(0, 1)$ , it suffices, by Slutsky's theorem (Theorem 5.20), to show that  $(S_{N_n} - S_{k_n})/\sqrt{k_n} \xrightarrow{P} 0$ . Given  $\varepsilon, \delta > 0$ ,

$$\begin{aligned} P\left\{|S_{N_n} - S_{k_n}| > \varepsilon\sqrt{k_n}\right\} \\ = P\left\{|S_{N_n} - S_{k_n}| > \varepsilon\sqrt{k_n}, |N_n - k_n| \leq \delta k_n\right\} \\ + P\left\{|S_{N_n} - S_{k_n}| > \varepsilon\sqrt{k_n}, |N_n - k_n| > \delta k_n\right\} \\ \leq P\left\{|S_{N_n} - S_{k_n}| > \varepsilon\sqrt{k_n}, |N_n - k_n| \leq \delta k_n\right\} \\ + P\{|N_n - k_n| > \delta k_n\}, \end{aligned}$$

and the second term converges to zero since  $N_n/k_n \xrightarrow{P} 1$ .

Regarding the first term,

$$\begin{aligned} P\left\{|S_{N_n} - S_{k_n}| > \varepsilon\sqrt{k_n}, |N_n - k_n| \leq \delta k_n\right\} \\ \leq P\left\{\max_{k_n \leq j \leq (1+\delta)k_n} |S_j - S_{k_n}| > \varepsilon\sqrt{k_n}\right\} \\ + P\left\{\max_{(1-\delta)k_n \leq j \leq k_n} |S_j - S_{k_n}| > \varepsilon\sqrt{k_n}\right\} \\ \leq \frac{(1+\delta)k_n - k_n}{k_n \varepsilon^2} + \frac{k_n - (1-\delta)k_n}{k_n \varepsilon^2} \end{aligned}$$

[by Kolmogorov's inequality (Theorem 7.2)]

$$\leq \frac{2\delta}{\varepsilon^2}. \blacksquare$$

### 7.5.5 Renewal processes

A renewal process is a probabilistic model of a system that renews itself at random times. For example, in reliability theory (one of the principal applications), one may study a device that is installed and operates until it fails, then is replaced (instantaneously, for simplicity) by another device with identical characteristics, and so on. Renewal processes are arrival counting processes in which the interarrival times are i.i.d., but not necessarily exponentially distributed, as they are in a Poisson process.

**Definition 7.93.** A *renewal process* is an arrival counting process  $N = (N_t)_{t \geq 0}$  in which the interarrival times  $U_1, U_2, \dots$  are i.i.d.  $\square$

We recall interpretations and notation from §3.6:  $U_k$  is the time between the  $(k-1)$ st and  $k$ th arrivals;  $T_k = \sum_{i=1}^k U_i$  is the time of the  $k$ th arrival; and for each  $t$ ,  $N_t = \sum_{k=1}^{\infty} \mathbf{1}(T_k \leq t)$  is the number of arrivals in  $(0, t]$ .

The following result uses Theorem 7.7 to establish a strong law of large numbers for the arrival counting process  $(N_t)$ .

**Theorem 7.34.** If  $0 < m = E[U_1] < \infty$ , then as  $t \rightarrow \infty$ ,  $N_t/t \xrightarrow{\text{a.s.}} 1/m$ .

**Proof:** On an event of probability one, we have both  $T_n(\omega)/n \rightarrow m$  as  $n \rightarrow \infty$ , by Theorem 7.7, and  $N_t(\omega) \rightarrow \infty$  as  $t \rightarrow \infty$ . For each  $t$ ,  $T_{N_t}$  is the time of the last arrival before  $t$  and  $T_{N_t+1}$  is the time of the first arrival after  $t$ . Since these satisfy  $T_{N_t} \leq t \leq T_{N_t+1}$ ,

$$\frac{T_{N_t}}{N_t} \leq \frac{t}{N_t} \leq \frac{T_{N_t+1}}{N_t+1} \frac{N_t+1}{N_t}.$$

For  $\omega$  as above, the outer terms in this expression both converge to  $m$ . ■

One further tool is needed to derive the associated central limit theorem. The backward recurrence time at  $t$  is the time that has elapsed since the last arrival before  $t$ , provided there has been an arrival before  $t$ , and is set equal to  $t$  otherwise.

**Definition 7.35.** The *backward recurrence time* at  $t$  is

$$V_t = \begin{cases} t - T_{N_t} & \text{if } N_t > 0 \\ t & \text{if } N_t = 0. \end{cases} \quad \square$$

Here is the central limit theorem for renewal processes.

**Theorem 7.36.** Let  $(N_t)$  be a renewal process for which the interarrival variance  $\sigma^2 \stackrel{\text{def}}{=} \text{Var}(U_i)$  satisfies  $0 < \sigma^2 < \infty$ . Then, as  $t \rightarrow \infty$ ,

$$\sqrt{t}[N_t/t - 1/m] \xrightarrow{\text{d}} N(0, \sigma^2/m^3).$$

**Proof:** By Theorem 7.32, applied to  $X_i = U_i - m$  and  $(N_t)$ ,

$$\frac{1}{\sqrt{N_t}} \sum_{i=1}^{N_t} [m - U_i] \xrightarrow{\text{d}} N(0, \sigma^2).$$

But,

$$\begin{aligned} \frac{1}{\sqrt{N_t}} \sum_{i=1}^{N_t} [m - U_i] &= \frac{mN_t - T_{N_t}}{\sqrt{N_t}} \\ &= m\sqrt{\frac{t}{N_t}} \left( \sqrt{t} \left[ \frac{N_t}{t} - \frac{1}{m} \right] \right) + \sqrt{\frac{t}{N_t}} - \frac{V_t}{\sqrt{t}}. \end{aligned}$$

By Theorem 7.34,  $\sqrt{t}/N_t \xrightarrow{\text{a.s.}} \sqrt{m}$ , so, again by Slutsky's theorem, it suffices that  $V_t/\sqrt{t} \xrightarrow{\text{P}} 0$ . This statement, which can be shown in a variety of ways, depends ultimately on existence of  $V_\infty \in L^1$  such that  $V_t \xrightarrow{\text{d}} V_\infty$  as  $t \rightarrow \infty$ . (For details, see Karr, 1991.) Then, by Chebyshev's inequality,

$$P\{V_t/t \geq \varepsilon\} \leq \frac{E[V_t]}{\varepsilon\sqrt{t}} \cong \frac{E[V_\infty]}{\varepsilon\sqrt{t}} \rightarrow 0. \quad \blacksquare$$

## 7.6 Complements

### 7.6.1 The Berry-Esséen theorem

Let  $X_1, X_2, \dots$  be i.i.d. with  $E[X_i] = 0$ ,  $\text{Var}(X_i) = 1$  and partial sums  $S_n = \sum_{i=1}^n X_i$ . With

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x e^{-y^2/2} dy$$

denoting the standard normal distribution function, the central limit theorem (Theorem 6.20 suffices in this case) implies that

$$P\{S_n/\sqrt{n} \leq x\} - \Phi(x) \rightarrow 0$$

for each  $x$ . Here, we concerned with the rate of this convergence, uniformly in  $x$ .

Without stronger moment assumptions, no (uniform) estimate of the rate of convergence is possible, but for each additional moment (beyond the second) that exists, and with the addition of a further term to the expansion of  $P\{S_n/\sqrt{n} \leq x\}$  (the first term is  $\Phi(x)$ ) the rate of convergence improves by one factor of  $1/\sqrt{n}$ . The Berry-Esséen theorem verifies the first of these steps under the assumption that  $E[|X_1|^3] < \infty$ .

We begin with a key analytical preliminary: Esséen's smoothing lemma, which bounds the uniform distance between two distribution functions in terms of an integrated distance between their characteristic functions. For the proof, see Chow/Teicher (1988).

**Lemma 7.37 (Smoothing lemma).** Let  $F$  and  $G$  be distribution functions on  $\mathbb{R}$ , the latter absolutely continuous with density  $g$  satisfying  $|g| \leq M$  for some  $M$ . Then, for each  $T > 0$ ,

$$\sup_{x \in \mathbb{R}} |F(x) - G(x)| \leq \frac{2}{\pi} \int_0^T \left| \frac{\varphi_F(t) - \varphi_G(t)}{t} \right| dt + \frac{24M}{\pi T}. \quad \square \quad (7.37)$$

We come now to the main result.

**Theorem 7.38 (Berry–Esséen theorem).** There exists an absolute constant  $C$  (not depending on the distribution of the  $X_i$ ) such that

$$\sup_{x \in \mathbb{R}} |P\{S_n/\sqrt{n} \leq x\} - \Phi(x)| \leq C \frac{E[|X_1|^3]}{\sqrt{n}}$$

for all  $n$ .

**Proof:** Let  $\gamma = E[|X_1|^3]$ . Since  $\text{Var}((X_i)) = 1$ ,  $\gamma \geq 1$  by Lyapunov's inequality (Corollary Corollary 4.37), and we assume that  $\gamma < \infty$ .

It suffices to show that for  $|t| \leq \sqrt{n}/5\gamma$ ,

$$|\varphi_{S_n/\sqrt{n}}(t) - e^{-t^2/2}| \leq \frac{7}{6} \frac{\gamma}{\sqrt{n}} |t|^3 e^{-t^2/4}, \quad (7.38)$$

because once we have (7.38), Lemma 7.37, applied with  $T = \sqrt{n}/5\gamma$ , gives

$$\begin{aligned} \sup_x \left| P\left\{ \frac{S_n}{\sqrt{n}} \leq x \right\} - \Phi(x) \right| &\leq \frac{2}{\pi} \int_0^{\sqrt{n}/(5\gamma)} \frac{7}{6} \frac{\gamma}{\sqrt{n}} t^2 e^{-t^2/4} dt \\ &\quad + \frac{24}{\pi} \frac{1}{\sqrt{2\pi}} \frac{5\gamma}{\sqrt{n}} \\ &\leq \frac{C\gamma}{\sqrt{n}} \end{aligned}$$

for appropriately chosen  $C$ . By the Taylor expansion for characteristic functions in Theorem 6.14, there is  $\theta$  with  $|\theta| \leq 1$  such that

$$\varphi_{X_1}(t/\sqrt{n}) = 1 - \frac{t^2}{2n} + \frac{\theta\gamma t^3}{6n^{3/2}}.$$

Hence, for  $|t| \leq \sqrt{n}/5\gamma$ ,  $|1 - \varphi_{X_1}(1 - 1/\sqrt{n})| \leq 25$ , which implies that the logarithms in the expression

$$\varphi_{S_n/\sqrt{n}}(t) = \exp[n \log \varphi_{X_1}(t/\sqrt{n})]$$

are well-defined. But, we also have

$$\log \varphi_{X_1}(t/\sqrt{n}) = -\frac{t^2}{2n} + \frac{\theta t^3}{6n^{3/2}} (\log \varphi_{X_1})'''(\theta_1 t/\sqrt{n})$$

for some  $\theta_1$ . Moreover,  $|(\log \varphi_{X_1})'''| \leq 7\gamma$ , and, hence, for  $|t| \leq \sqrt{n}/5\gamma$ ,

$$\begin{aligned} |\varphi_{X_1}(t/\sqrt{n})^n - e^{-t^2/2}| &\leq |e^{n \log \varphi_{X_1}(t/\sqrt{n})} - e^{-t^2/2}| \\ &\leq \frac{7}{6} \frac{\gamma |t|^3}{\sqrt{n}} \exp\left[-\frac{t^2}{2} + \frac{7}{6} \frac{|t|^3 \gamma}{\sqrt{n}}\right] \end{aligned}$$

[via the inequality  $|e^z - 1| \leq |z|e^{|z|}$ ]

$$\leq \frac{7}{6} \frac{\gamma |t|^3}{\sqrt{n}} e^{-t^2/4}. \blacksquare$$

## 7.7 Exercises

**7.1.** Prove Lemma 7.1.

- 7.2.** Let  $X_1, X_2, \dots$  be independent with  $E[X_i] = 0$  and  $\sigma_i^2 = \text{Var}(X_i) < \infty$  for each  $i$ , and let  $S_n = \sum_{i=1}^n X_i$ . Prove that if  $\sum_{k=1}^{\infty} \sigma_k^2/k^2 < \infty$ , then  $S_n/n \xrightarrow{\text{a.s.}} 0$ . [This is a strong law of large numbers for independent but not necessarily identically distributed summands.]
- 7.3.** Let  $X_1, X_2, \dots$  be i.i.d. with distribution  $U[0, 1]$ . Prove that, almost surely,  $\sum_{n=1}^{\infty} \prod_{i=1}^n X_i < \infty$ .
- 7.4.** Let  $X_1, X_2, \dots$  be i.i.d., with values in a finite set  $B$ , and let  $f_0(y) = P\{X_k = y\}$ ,  $y \in B$ . Let  $f_1 \neq f_0$  be a second probability on  $B$ , and for each  $n$ , let  $Z_n = \prod_{k=1}^n f_1(X_k)/f_0(X_k)$ . Prove that  $Z_n \xrightarrow{\text{a.s.}} 0$ . [Hint: Apply the strong law of large numbers to the process  $Y_n = \log Z_n$ .]
- 7.5.** Let  $X_1, X_2, \dots$  be independent. Prove that if  $\sum_{k=1}^{\infty} X_k$  converges in probability, then the convergence also occurs almost surely.
- 7.6.** Let  $X_1, X_2, \dots$  be i.i.d. with  $P\{X_k = 0\} = P\{X_k = 2\} = 1/2$ .
- Prove that  $Z = \sum_{k=1}^{\infty} X_k/3^k$  converges almost surely.
  - Prove that  $Z$  has the Cantor distribution of Example 1.51.
  - Use parts a) and b) of this exercise to calculate the mean and variance of the Cantor distribution. [Since the Cantor distribution is neither discrete nor absolutely continuous, the computational techniques in §4.3 do not apply.]
- 7.7.** Let  $X_1, X_2, \dots$  be i.i.d., with values in  $\{1, \dots, \ell\}$  for some  $\ell$ , and suppose that  $p_k \stackrel{\text{def}}{=} P\{X_i = k\} > 0$  for each  $k$ . Define  $N_n(k) = \sum_{i=1}^n \mathbf{1}(X_i = k)$ , the number of  $X_1, \dots, X_n$  whose value is  $k$ , and let  $\Pi_n = \prod_{k=1}^{\ell} p_k^{N_n(k)}$ . Prove that the sequence  $(1/n) \log \Pi_n$  converges almost surely and calculate the limit.
- 7.8.** Show that in Theorem 7.7,  $(S_n/n)$  is uniformly integrable, so that  $S_n/n \xrightarrow{\text{L}^1} E[X_1]$ .
- 7.9.** Prove that if the  $X_k$  are i.i.d. with  $E[|X_1|^3] < \infty$ , then the Lyapunov condition is satisfied.
- 7.10.** Prove Proposition 7.14.
- 7.11.** Let  $X_1, X_2, \dots$  be independent with  $E[X_i] = 0$  for each  $i$ . Prove that if there is  $\delta > 0$  for which  $E[|X_i|^{2+\delta}] < \infty$  for each  $i$  and

$$\lim_{n \rightarrow \infty} \frac{1}{\sigma_n^{2+\delta}} \sum_{i=1}^n E[|X_i|^{2+\delta}] = 0,$$

then  $(X_n)$  satisfies the Lindeberg condition. (This is a more general Lyapunov condition, and implies a version of Theorem 7.11, although the proof is more delicate than for  $\delta = 1$ .)

- 7.12.** Let  $X_1, X_2, \dots$  be independent with  $X_k \stackrel{d}{=} U[-a_k, a_k]$  for each  $k$ , where  $0 < a_k \leq 1$ . Prove that  $(X_k)$  satisfies the Lindeberg condition if and only if  $\sum_{k=1}^{\infty} a_k^2 = \infty$ .
- 7.13.** Let  $X_1, X_2, \dots$  be independent. In which of the following situations does  $(X_k)$  satisfy the Lindeberg condition?
- $P\{X_k = k\} = P\{X_k = -k\} = 1/2$ .
  - $P\{X_k = 2^{k/2}\} = P\{X_k = -2^{k/2}\} = 1/2$ .
  - $P\{X_k = k\} = P\{X_k = -k\} = 1/2k$  and  $P\{X_k = 0\} = 1 - 1/k$ .
- 7.14.** Let  $X_1, X_2, \dots$  be independent with  $E[X_k] = 0$  and  $s_k^2 = \text{Var}(X_k)$  for each  $k$ , and let  $S_n = \sum_{k=1}^n X_k$  and  $\sigma_n^2 = \sum_{k=1}^n s_k^2$  for each  $n$ . Prove that if there are constants  $c_k$  such that  $P\{|X_k| \leq c_k\} = 1$  for each  $k$  and if  $\sigma_n^2 \rightarrow \infty$  and  $\max_{k \leq n} c_k/\sigma_n \rightarrow 0$ , then  $(X_k)$  satisfies the Lindeberg condition.
- 7.15.** Let  $Y_1, Y_2, \dots$  be i.i.d. with mean 0 and variance 1, and let  $Z_k$  be independent, independent of  $(Y_i)$ , such that

$$\begin{aligned} P\{Z_j = \pm j\} &= 1/2j^2 \\ P\{Z_j = 0\} &= 1 - 1/j^2. \end{aligned}$$

Let  $X_i = Y_i + Z_i$  and  $S_n = \sum_{i=1}^n X_i$ . Prove that  $S_n/\sqrt{n} \xrightarrow{d} N(0, 1)$ , but that  $(X_n)$  does not satisfy the Lindeberg condition.

- 7.16.** Use the central limit theorem to show that

$$\lim_{n \rightarrow \infty} \sum_{k=0}^n e^{-n} \frac{n^k}{k!} = \frac{1}{2}$$

and

$$\lim_{n \rightarrow \infty} \frac{1}{(n-1)!} \int_0^n t^{n-1} e^{-t} dt = \frac{1}{2}.$$

- 7.17.** Let  $X_1, X_2, \dots$  be i.i.d. with distribution function

$$F(x) = \begin{cases} (1/2)e^{-x^2} & x < 0 \\ 1 - (1/2)e^{-x^2} & x \geq 0. \end{cases}$$

Let  $S_n = \sum_{i=1}^n X_i$  be the partial sums.

- a) Show that  $\varphi_{X_1}(t) = 2 \int_0^\infty x(\cos tx)e^{-x^2} dx$ .

b) Prove that  $\limsup_{n \rightarrow \infty} X_n / \sqrt{\log n} \xrightarrow{\text{a.s.}} 1$ .

c) Prove that  $S_n / \sqrt{n} \xrightarrow{d} N(0, 1)$  as  $n \rightarrow \infty$ .

- 7.18.** Let  $X_1, X_2, \dots$  be i.i.d. with  $E[X_1] = \mu$ ,  $\text{Var}(X_1) = \sigma^2$  and finite fourth central moment  $\mu_4 \stackrel{\text{def}}{=} E[(X_1 - \mu)^4]$ . Define the “sample variance”

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X}_n)^2,$$

where  $\bar{X}_n = (1/n) \sum_{i=1}^n X_i$  is the sample mean. Prove that

$$\frac{\sqrt{n}[S_n^2 - \sigma^2]}{\sqrt{\mu_4 - \sigma^4}} \xrightarrow{d} N(0, 1).$$

- 7.19.** Let  $X_1, X_2, \dots$  be i.i.d. with mean 0 and variance 1. Prove that

$$\frac{\sum_{i=1}^n X_i}{\sqrt{\sum_{j=1}^n X_j^2}} \xrightarrow{d} N(0, 1).$$

- 7.20.** Let  $X_1, X_2, \dots$  be i.i.d. with density  $f(x) = (1/|x|^3)\mathbf{1}(|x| \geq 1)$ , and let  $S_n = \sum_{i=1}^n X_i$ . Prove that  $S_n / \sqrt{n \log n} \xrightarrow{d} N(0, 1)$ .

- 7.21.** Let  $Y_1, Y_2, \dots$  be random variables such that for some constant  $c$ ,  $\sqrt{n}[Y_n - c] \xrightarrow{d} N(0, 1)$ , and suppose that  $h: \mathbb{R} \rightarrow \mathbb{R}$  is twice continuously differentiable in some neighborhood of  $c$ . Prove that

$$\sqrt{n}[h(Y_n) - h(c)] \xrightarrow{d} N(0, h'(c)^2).$$

This technique, the *Δ-method*, is widely used to derive asymptotic properties of statistics based on i.i.d. data.

- 7.22.** This exercise describes some alternative methods of Monte Carlo numerical integration. Let  $h$  be a function on  $[0, 1]$  with  $0 \leq h(x) \leq 1$  for all  $x$ , whose integral  $I(h) \stackrel{\text{def}}{=} \int_0^1 h(x) dx$  we wish to compute (i.e., approximate) numerically. The procedure described in §5, which we term the “standard” method, uses estimators

$$\hat{I}_s(h) = \frac{1}{n} \sum_{i=1}^n h(U_i),$$

where the  $U_i$  are independent with distribution  $U[0, 1]$ . For simplicity, the sample size  $n$  is suppressed from the notation.

The “hit or miss” method employs estimators

$$\hat{I}_{hm}(h) = \frac{1}{n} \sum_{i=1}^n \mathbf{1}(U_i \leq h(V_i)),$$

where  $(V_i)$  is a second i.i.d. sequence of uniform random variables independent of  $(U_i)$ . The rational is that  $\hat{I}_{\text{hm}}(h)$  is the fraction of pairs  $(U_i, V_i)$  under the graph of  $h$ , whose area is  $I(h)$ .

The method of “antithetic variables” is based on estimators

$$\hat{I}_{\text{av}}(h) = \frac{1}{2n} \sum_{i=1}^n [h(U_i) + h(1 - U_i)].$$

- a) Show that the hit or miss and antithetic variables methods are consistent: as  $n \rightarrow \infty$ ,  $\hat{I}_{\text{hm}}(h) \xrightarrow{\text{a.s.}} I(h)$  and  $\hat{I}_{\text{av}}(h) \xrightarrow{\text{a.s.}} I(h)$ .
- b) Show that there exist constants  $\sigma_s^2(h)$ ,  $\sigma_{\text{hm}}^2(h)$  and  $\sigma_{\text{av}}^2(h)$ , such that

$$\begin{aligned}\sqrt{n} [\hat{I}_s(h) - I(h)] &\xrightarrow{\text{d}} N(0, \sigma_s^2(h)) \\ \sqrt{n} [\hat{I}_{\text{hm}}(h) - I(h)] &\xrightarrow{\text{d}} N(0, \sigma_{\text{hm}}^2(h)) \\ \sqrt{n} [\hat{I}_{\text{av}}(h) - I(h)] &\xrightarrow{\text{d}} N(0, \sigma_{\text{av}}^2(h))\end{aligned}$$

- c) For each function  $h$ , determine the method for which  $\sigma^2(h)$  is smallest. This method, all other things being equal, is then the preferred method for estimation of  $I(h)$ .
- 7.23.** Let  $X_1, \dots, X_n$  be independent with distribution  $N(\theta, 1)$ . Show that the maximum likelihood estimator of  $\theta$  is  $\hat{\theta} = (1/n) \sum_{i=1}^n X_i$ , the sample mean.
- 7.24.** Let  $X_1, \dots, X_n$  be independent and Bernoulli distributed with parameter  $p$ . Show that the maximum likelihood estimator of  $p$  is  $\hat{p} = (1/n) \sum_{i=1}^n X_i$ .
- 7.25.** Let  $X_1, X_2, \dots$  be independent with distribution  $U[\theta - 1/2, \theta + 1/2]$ , and define  $M_n^* = \max \{X_1, \dots, X_n\}$  and  $M_n^{**} = \min \{X_1, \dots, X_n\}$ .
- a) Show that for each  $n$ , every point in  $[M_n^* - 1/2, M_n^{**} + 1/2]$  is a maximum likelihood estimator of  $\theta$ .
  - b) Show that  $M_n^{**} + 1/2$  and  $M_n^* - 1/2$  are both consistent estimators of  $\theta$ , in the sense of almost sure convergence.
- 7.26.** Verify (7.36).
- 7.27.** Let  $X_1, X_2, \dots$  be i.i.d. with continuous distribution function  $F$ . Prove that the distribution of the *Kolmogorov-Smirnov statistic*

$$D_n = \sup_{x \in \mathbb{R}} |\hat{F}_n(x) - F(x)|$$

does not depend on  $F$ . [Hint: consider the quantile transformation described in Proposition 2.48.]

- 7.28.** Suppose that the  $X_i$  are exponentially distributed with parameter 1. Prove that

$$\max\{X_1, \dots, X_n\} - \log n \xrightarrow{d} Z,$$

where  $Z$  satisfies  $P\{Z \leq x\} = e^{-e^{-x}}$ ,  $x \in \mathbb{R}$ .

- 7.29.** Let  $X_1, X_2, \dots$  be i.i.d. with distribution function  $F$  satisfying  $F(t) < 1$  for all  $t < \infty$ , and for each  $n$ , let  $M_n^* = \max\{X_1, \dots, X_n\}$ . Suppose that there are constants  $c, \alpha > 0$  such that  $1 - F(x) \cong cx^{-\alpha}$  as  $x \rightarrow \infty$ . Show that  $M_n^*/\sqrt[n]{cn} \xrightarrow{d} Z$ , where  $Z$  is a random variable with  $P\{Z \leq x\} = e^{-x^{-\alpha}}$ ,  $x > 0$ .

- 7.30.** Suppose that  $(X_i)$  and  $(M_n^*)$  are as in Exercise 7.29, and that there is  $b \in \mathbb{R}$  such that  $F(b) = 1$  but  $F(x) < 1$  for all  $x < b$ .

- a) Prove that  $M_n^* \xrightarrow{P} b$ .
- b) Suppose that there are  $c, \alpha > 0$  such that  $1 - F(x) \cong c(b - x)^\alpha$  as  $x \uparrow b$ . Prove that there is a random variable  $Z$  such that  $\sqrt[n]{n}(b - M_n^*) \xrightarrow{d} Z$ , and identify the limit.

- 7.31.** Let  $X_1, \dots, X_n$  be i.i.d. with mean 0 and variance 1, and let  $S_k = \sum_{i=1}^k X_i$ . Use a first step decomposition to prove that for each  $a > 0$ ,

$$P\{\max_{k \leq n} S_k \geq a\} \leq 2P\{S_n \geq a - \sqrt{2n}\}.$$

In the following exercises, let  $(X_n)$  be a random walk, let  $Z_n$  be the number of visits to the origin in steps  $1, \dots, n$ , let  $T_j^0$  be the time of the  $j$ th return to the origin, let  $T^k$  be the first passage time to level  $k$ , and let  $M_n^* = \max\{X_1, \dots, X_n\}$ . See the Prelude for details.

- 7.32.** Prove that for each  $n$  and  $j$ ,  $P\{Z_{n2} = j\} = P\{X_{2n-j} = j\}$ , and use this to show that for  $x \geq 0$ ,

$$\lim_{n \rightarrow \infty} P\{Z_n/\sqrt{n} \leq x\} = 2 \int_0^x \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy.$$

- 7.33.** Prove that for  $x \geq 0$ ,

$$\lim_{j \rightarrow \infty} P\{T_j^0/j^2 \leq x\} = \int_0^x \frac{1}{\sqrt{2\pi}} \frac{1}{y^{3/2}} e^{-1/2y} dy,$$

and conclude from this, using (0.17), that for  $x \geq 0$ ,

$$\lim_{k \rightarrow \infty} P\{T^k/k^2 \leq x\} = \int_0^x \frac{1}{\sqrt{2\pi}} \frac{1}{y^{3/2}} e^{-1/2y} dy.$$

- 7.34.** Use (0.21) and (0.19) to show that for  $x \geq 0$ ,

$$\lim_{n \rightarrow \infty} P\{M_n^*/\sqrt{n} \leq x\} = 2 \int_0^x \frac{1}{\sqrt{2\pi}} e^{-y^2/2} dy.$$

## Chapter 8

# Prediction and Conditional Expectation

Let  $(\Omega, \mathcal{F}, P)$  be a probability space, and let  $L^2$  be the vector space of random variables  $Z$  with  $E[Z^2] < \infty$ . Let  $X \in L^2$  be an unobservable random variable, whose value we wish to predict from observation of other random variables  $Y_1, \dots, Y_n$ . (For example,  $X$  may be the value of a stochastic process at some time in the future, or a spatial process over a region where it cannot be observed). In order to use knowledge of  $Y_1, \dots, Y_n$  to predict  $X$ , the predictor must be function of  $Y_1, \dots, Y_n$ , and, in particular, is a random variable.

To solve such a prediction problem, we must first specify a set  $V$  of allowable predictors. Often, these are all random variables in  $L^2$  that are functions of  $Y_1, \dots, Y_n$ , but more restricted situations that facilitate computation, for example, allowing only linear functions of the observable random variables, are also of interest. Second, we must identify the “best” predictor according to some optimality criterion. Our criterion is *mean squared error* (MSE). The mean squared error of a predictor  $Z$  is

$$\text{MSE}(Z) \stackrel{\text{def}}{=} E[(Z - X)^2].$$

Assuming it exists, the optimal predictor,

$$\hat{X} \stackrel{\text{def}}{=} \arg \min_{Z \in V} \text{MSE}(Z),$$

the *minimum mean squared error* (MMSE) predictor of  $X$ , then satisfies

$$E[(\hat{X} - X)^2] \leq E[(Z - X)^2]$$

for all  $Z \in V$ .

Issues for prediction problems are existence, characterization and computation of MMSE predictors, and properties of MMSE predictors as functions of  $X$  and as functions of  $Y_1, \dots, Y_n$ . The latter are initially explored in Theorem 8.23, and treated in detail in Chapter 9.

## 8.1 Prediction in $L^2$

The space  $L^2$ , albeit infinite-dimensional, has effectively the same geometry as finite-dimensional Euclidean spaces. Many familiar concepts and results carry over, not only in spirit, but also in substance. Both  $L^2$  and Euclidean spaces have defined on them an inner product, a norm (whose square measures length) derived from the inner product, and a metric derived from the norm. Like Euclidean spaces,  $L^2$  is a complete metric space: every Cauchy sequence converges. Completeness ensures existence of solutions to MMSE prediction problems. Inner products lead as well to orthogonality and orthogonal decompositions. The latter represent a random variable as the sum of its projection onto a given subspace, and a second random variable orthogonal to the entire subspace. Moreover, the orthogonal projection is MMSE predictor within the subspace.

### 8.1.1 The inner product and norm

The basic properties of the inner product and norm on  $L^2$  are analogous to those of the inner product  $\langle x, y \rangle = \sum_{i=1}^d x_i y_i$  and norm  $\|x\| = \sqrt{\langle x, x \rangle}$  on  $\mathbb{R}^d$ .

**Definition 8.1.** The *inner product* of  $X, Y \in L^2$  is

$$\langle X, Y \rangle = E[XY]. \quad \square$$

Existence of  $\langle X, Y \rangle$  is guaranteed by the Cauchy-Schwarz inequality:

$$|\langle X, Y \rangle| \leq \sqrt{E[X^2]E[Y^2]}.$$

Linearity of expectation implies that the inner product on  $L^2$  is symmetric and linear in each argument: for  $X, Y, Z \in L^2$  and  $a \in \mathbb{R}$ ,

$$\begin{aligned}\langle X, Y \rangle &= \langle Y, X \rangle \\ \langle X + Y, Z \rangle &= \langle X, Z \rangle + \langle Y, Z \rangle \\ \langle aX, Y \rangle &= a\langle X, Y \rangle.\end{aligned}$$

Now, we pursue a concept introduced in Chapter 5.

**Definition 8.2.** The *norm* of  $X \in L^2$  is

$$\|X\| \stackrel{\text{def}}{=} \sqrt{\langle X, X \rangle} = \sqrt{E[X^2]}. \quad \square$$

The properties of the norm are those of Euclidean length.

**Proposition 8.3.** Let  $\|X\|$  be the norm of  $X \in L^2$ . Then,

a)  $\|X\| \geq 0$ , and  $\|X\| = 0$  if and only if  $X \stackrel{\text{a.s.}}{\equiv} 0$ .

b) For  $X, Y \in L^2$ , there hold the *triangle inequality*, *parallelogram law* and *polarization identity*:

$$\|X + Y\| \leq \|X\| + \|Y\| \quad (8.1)$$

$$\|X + Y\|^2 + \|X - Y\|^2 = 2(\|X\|^2 + \|Y\|^2) \quad (8.2)$$

$$\langle X, Y \rangle = \frac{1}{4} (\|X + Y\|^2 - \|X - Y\|^2). \quad (8.3)$$

**Proof:** a) The first property follows from positivity of expectation, while the second is a consequence of Proposition 4.11.

b) The triangle inequality is just Minkowski's inequality.

For the parallelogram law (whose name arises from its geometric interpretation in terms of the lengths of the sides and diagonals of a parallelogram), we observe that

$$\begin{aligned} \|X + Y\|^2 + \|X - Y\|^2 &= E[(X + Y)^2] + E[(X - Y)^2] \\ &= E[X^2] + 2E[XY] + E[Y^2] \\ &\quad + E[X^2] - 2E[XY] + E[Y^2] \\ &= 2(E[X^2] + E[Y^2]) \\ &= 2(\|X\|^2 + \|Y\|^2). \end{aligned}$$

The polarization identity is proved similarly. ■

Then, the pair  $(L^2, \langle \cdot, \cdot \rangle)$  is termed an *inner product space*, and  $(L^2, \|\cdot\|)$  is a *normed linear space*.

### 8.1.2 $L^2$ as metric space

The “metric” on  $L^2$  is given by

$$d(X, Y) \stackrel{\text{def}}{=} \|X - Y\| = \sqrt{\langle X - Y, X - Y \rangle} = \sqrt{E[(X - Y)^2]}.$$

That  $d$  satisfies the triangle inequality is simply a restatement of the triangle inequality (8.1) for norms. However,  $d(X, Y) = 0$  does not imply that  $X \equiv Y$  as functions on  $\Omega$ , but only that  $X \stackrel{\text{a.s.}}{\equiv} Y$ . To circumvent this difficulty, we “identify” random variables that are equal almost surely (i.e., treat them as if they were the same function on  $\Omega$ ), in which case  $(L^2, d)$  becomes a metric space.

Convergence with respect to  $d$  is quadratic mean convergence (Definition 5.3):  $d(X_n, X) \rightarrow 0$  if and only if  $X_n \xrightarrow{\text{q.m.}} X$ .

We first note elementary properties of this metric.

**Proposition 8.4.** With respect to  $d$ , the function  $X \mapsto \|X\|$  from  $L^2$  into  $\mathbb{R}_+$  is uniformly continuous, and for fixed  $Y \in L^2$ , the function  $X \mapsto \langle X, Y \rangle$  is uniformly continuous.

**Proof:** Two applications of the triangle inequality yield

$$\begin{aligned}\|X\| &= \|(X - Y) + Y\| \leq \|X - Y\| + \|Y\| \\ \|Y\| &= \|(Y - X) + X\| \leq \|Y - X\| + \|X\|,\end{aligned}$$

and these expressions combine to produce

$$|\|X\| - \|Y\|| \leq \|X - Y\|.$$

By linearity of the inner product and the Cauchy-Schwarz inequality,

$$|\langle X, Y \rangle - \langle Z, Y \rangle| = |\langle X - Z, Y \rangle| \leq \|X - Z\| \cdot \|Y\|. \blacksquare$$

The most important property of  $L^2$  *qua* metric space is that it is complete: every Cauchy sequence converges to an element of  $L^2$ .

**Theorem 8.5.** The metric space  $(L^2, d)$  is complete.

**Proof:** Let  $(X_n)$  be a Cauchy sequence in  $L^2$ :  $\lim_{n,m \rightarrow \infty} \|X_n - X_m\| = 0$ . Then, there is a subsequence  $(X_{n'})$  such that  $\|X_{(n+1)'} - X_{n'}\| < 2^{-n}$  for each  $n$ . With  $X_0 = 0$ , we show that the series  $Y = \sum_{n=1}^{\infty} |X_{(n+1)'} - X_{n'}|$  converges almost surely. Indeed,

$$\begin{aligned}E[Y^2] &= E\left[\lim_{m \rightarrow \infty} \left(\sum_{n=1}^m |X_{(n+1)'} - X_{n'}|\right)^2\right] \\ &= \lim_{m \rightarrow \infty} \left\| \sum_{n=1}^m |X_{(n+1)'} - X_{n'}| \right\|^2\end{aligned}$$

[by the monotone convergence theorem]

$$\leq \left( \lim_{m \rightarrow \infty} \sum_{n=1}^m \|X_{(n+1)'} - X_{n'}\| \right)^2,$$

by the triangle inequality, and this last expression is finite by the construction of  $(n')$ . Hence,  $P\{Y < \infty\} = 1$ , and for every  $\omega$  such that  $Y(\omega) < \infty$ ,  $Z(\omega) = \lim_n X_{n'}(\omega)$  exists. Since  $|Z| \leq Y$  and  $E[Y^2] < \infty$ ,  $Z$  belongs to  $L^2$  as well.

To complete the proof, we show that  $X_{n'} \xrightarrow{\text{q.m.}} Z$ . Given  $\varepsilon > 0$ , for  $m$  and  $n$  sufficiently large, we have  $\|X_m - X_{n'}\| < \varepsilon$ . Hence, by Fatou's lemma (Theorem 4.8), for large values of  $m$ ,

$$\begin{aligned} E[(X_m - Z)^2] &= E[\liminf_n (X_m - X_{n'})^2] \\ &\leq \liminf_{n \rightarrow \infty} E[(X_m - X_{n'})^2] \\ &= \liminf_{n \rightarrow \infty} \|X_m - X_{n'}\|^2 \\ &< \varepsilon^2, \end{aligned}$$

so that  $\limsup_m E[(X_m - Z)^2] < \varepsilon^2$ , and, hence,  $X_m \xrightarrow{\text{q.m.}} Z$ . ■

Thus,  $L^2$  is a *Hilbert space*, a complete inner product space.

### 8.1.3 Orthogonality and orthonormality

Orthogonality, which is analogous to perpendicularity in Euclidean space, generalizes “uncorrelatedness.” In particular, mean zero random variables are orthogonal if and only if they are uncorrelated.

**Definition 8.6.** Random variables  $X, Y \in L^2$  are *orthogonal* if  $\langle X, Y \rangle = 0$ . We denote this by  $X \perp Y$ . □

We next define orthogonal and orthonormal subsets of  $L^2$ , as well as orthogonal complements. An orthogonal set is the analogue of a set of mutually perpendicular vectors in Euclidean space, while, in addition, each member of an orthonormal set has length one. The orthogonal complement of a subset  $K$  of  $L^2$  consists of those random variables in  $L^2$  orthogonal to every element of  $K$ .

**Definition 8.7.** Let  $K$  be a subset of  $L^2$ .

- a)  $K$  is *orthogonal* if  $X \perp Y$  for all  $X, Y \in K$  such that  $X \neq Y$ .
- b)  $K$  is *orthonormal* if  $K$  is orthogonal and  $\|X\| = 1$  for each  $X \in K$ .
- c) The *orthogonal complement* of  $K \subseteq L^2$  is

$$K^\perp = \{X \in L^2 : X \perp Y \text{ for all } Y \in K\}. \quad \square$$

Regardless of  $K$ , its orthogonal complement is a closed subspace.

**Proposition 8.8.** For  $K \subseteq L^2$ ,  $K^\perp$  is a closed subspace of  $L^2$ .

**Proof:** That  $K^\perp$  is a subspace is a consequence of linearity of the inner product. For  $X, Z \in K^\perp$  and  $a, b \in \mathbb{R}$ ,

$$\langle aX + bZ, Y \rangle = a\langle X, Y \rangle + b\langle Z, Y \rangle = 0$$

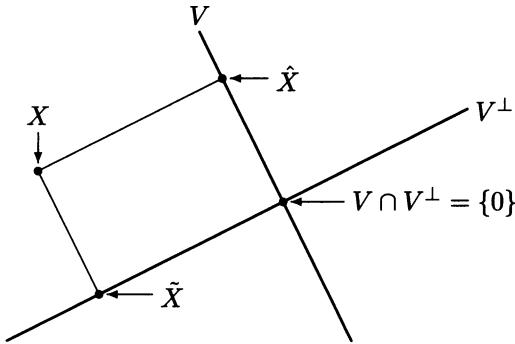


Figure 8.1. Orthogonal Decompositions and Projections

for all  $Y \in K$ , and, hence,  $aX + bZ \in K^\perp$ .

For fixed  $Y \in L^2$ ,  $\{Y\}^\perp = \{X : X \perp Y\}$  is the inverse image under the continuous mapping  $X \mapsto \langle X, Y \rangle$  (Proposition 8.4) of the closed set  $\{0\}$ , and, hence, is closed. Then,  $K^\perp = \bigcap_{Y \in K} \{Y\}^\perp$ , the intersection of closed sets, is closed as well. ■

Only  $X = 0$  is orthogonal to itself, so for  $V$  a subspace of  $L^2$ ,

$$V \cap V^\perp = \{0\}. \quad (8.4)$$

Figure 8.1 depicts (among other things) a subspace  $V$  of  $\mathbb{R}^2$  and its orthogonal complement.

**Technical Aside.** In finite-dimensional vector spaces, subspaces are automatically closed, as are finite-dimensional subspaces of infinite-dimensional vector spaces. In general, an infinite-dimensional subspace of a vector space need not be closed. For this reason, closure must be assumed in situations such as Proposition 8.8 and Theorem 8.9. □

#### 8.1.4 The orthogonal decomposition theorem

Given a closed subspace  $V$  of  $L^2$ , every  $X \in L^2$  can be decomposed uniquely as the sum of random variables  $\hat{X} \in V$  and  $\tilde{X} \in V^\perp$ ;  $\hat{X}$  is the perpendicular projection of  $X$  onto  $V$  and, hence, the MMSE predictor of  $X$  within  $V$ .

**Theorem 8.9 (Orthogonal decomposition theorem).** Suppose that  $V$  is a closed subspace of  $L^2$ . Then,

- a) Every  $X \in L^2$  has a unique decomposition

$$X = \hat{X} + \tilde{X}, \quad (8.5)$$

with  $\hat{X} \in V$  and  $\tilde{X} \in V^\perp$ . Moreover,  $\hat{X}$ , which is called the *orthogonal projection* of  $X$  onto  $V$ , is the element of  $V$  closest to  $X$ :

$$\hat{X} = \arg \min_{Z \in V} \|Z - X\|. \quad (8.6)$$

b) The mappings  $X \mapsto \hat{X}$  and  $X \mapsto \tilde{X}$  of  $L^2$  onto  $V$  and  $V^\perp$  are linear.

**Proof:** a) We first show that for each  $X$ , there is a unique element  $\hat{X}$  of  $V$  satisfying (8.6). To this end, let  $d = \inf \{\|X - Z\| : Z \in V\}$ . Then, there is a sequence  $(Z_n)$  in  $V$  such that  $\|Z_n - X\|^2 \leq d^2 + 1/n$  for each  $n$ . Moreover, this sequence is Cauchy, since for each  $n$  and  $m$ ,

$$\|Z_n - Z_m\|^2 = 2(\|Z_m - X\|^2 + \|Z_n - X\|^2) - 4 \left\| \frac{Z_m + Z_n}{2} - X \right\|^2$$

[by the parallelogram law, since  $\|Z_m + Z_n - 2X\|^2 = 4\|(Z_m + Z_n)/2 - X\|^2$ ]

$$\leq 2(d^2 + 1/m + d^2 + 1/n) - 4d^2$$

[ $V$  is a subspace of  $L^2$  and  $Z_n, Z_m \in V$ , so that  $(Z_m + Z_n)/2 \in V$ ]

$$= 2(1/m + 1/n),$$

which converges to zero as  $m$  and  $n$  converge to infinity.

Hence, by completeness of  $L^2$  (Theorem 8.5), there is  $\hat{X} \in L^2$  such that  $Z_n \xrightarrow{\text{q.m.}} \hat{X}$ . Continuity of the norm implies that  $\|\hat{X} - X\| = \lim_n \|Z_n - X\| = d$ , so that  $\hat{X}$  satisfies (8.6). Finally,  $\hat{X} \in V$  because  $V$  is closed.

Next, we show that  $\tilde{X} = X - \hat{X}$  belongs to  $V^\perp$ . For  $Z \in V$  and  $a \in \mathbb{R}$ , (8.6) implies that  $\|X - (\hat{X} + aZ)\|^2 \geq \|\hat{X} - X\|^2$ . Thus,

$$E[(\hat{X} - X)^2] + 2aE[(\hat{X} - X)Z] + a^2E[Z^2] \geq E[(\hat{X} - X)^2],$$

so that

$$2aE[(\hat{X} - X)Z] \geq -a^2E[Z^2].$$

However, unless  $E[(\hat{X} - X)Z] = 0$ , this inequality fails for  $a$  either positive and sufficiently close to zero or negative and sufficiently close to zero. Hence, the decomposition (8.5) holds.

Uniqueness of the decomposition (8.5) is based on (8.4). If  $X = \hat{X}_1 + \tilde{X}_1 = \hat{X}_2 + \tilde{X}_2$  with  $\hat{X}_1, \hat{X}_2 \in V$  and  $\tilde{X}_1, \tilde{X}_2 \in V^\perp$ , then  $\hat{X}_1 - \hat{X}_2$  and  $\tilde{X}_2 - \tilde{X}_1$  are equal, the former belongs to  $V$  and the latter belongs to  $V^\perp$ , and so both are zero by (8.4).

b) Linearity of the operators  $X \mapsto \hat{X}$  and  $X \mapsto \tilde{X}$  is shown similarly (and simultaneously). Since

$$(aX + bY)^\wedge + (aX + bY)^\sim = aX + bY = a\hat{X} + a\tilde{X} + b\hat{Y} + b\tilde{Y},$$

then

$$(aX + bY)^\sim - a\hat{X} - b\tilde{Y} = -(aX + bY)^\sim + a\tilde{X} + b\tilde{Y}.$$

The left-hand side of this expression belongs to  $V$ , while the right-hand side belongs to  $V^\perp$ , and so both are zero. ■

**Corollary 8.10.** If  $\hat{X}$  is the orthogonal projection of  $X$  onto a closed subspace  $V$  of  $L^2$ , then

$$\|\hat{X}\| \leq \|X\|.$$

**Proof:** Since  $\hat{X} \perp \tilde{X}$ ,

$$\|X\|^2 = \|\hat{X}\|^2 + \|\tilde{X}\|^2 \geq \|\hat{X}\|^2. \blacksquare$$

The orthogonal decomposition theorem is illustrated in Figure 8.1.

### 8.1.5 Computation of MMSE predictors

When  $V$  consists of all functions of observable random variables  $Y_1, \dots, Y_n$ , MMSE predictors are conditional expectations, whose computation is discussed in §6.

Direct computation of (other) MMSE predictors is based on the property that the “prediction error”  $\tilde{X} = \hat{X} - X$  belongs to  $V^\perp$ , which implies that  $\hat{X}$  is the solution to the system of equations

$$\langle \hat{X} - X, Z \rangle = 0, \quad Z \in V. \tag{8.7}$$

In reality, these can be solved only when  $V$  is finite-dimensional, of which the most important case is linear prediction.

### 8.1.6 Linear prediction

Suppose that  $V$  is the subspace of all linear functions of observable random variables  $Y_1, \dots, Y_n \in L^2$ : each  $Z \in V$  has the form  $Z = \sum_{i=1}^n a_i Y_i$ , for  $a_1, \dots, a_n \in \mathbb{R}$ . This formulation includes affine functions as well, since we can always take, for example,  $Y_n \equiv 1$ .

**Proposition 8.11.** The MMSE linear predictor of  $X \in L^2$  within  $V$  is  $\hat{X} = \sum_{i=1}^n a_i^* Y_i$ , where  $a_1^*, \dots, a_n^*$  satisfy the *normal equations*

$$\sum_{j=1}^n \langle Y_j, Y_i \rangle a_j^* = \langle Y_i, X \rangle, \quad i = 1, \dots, n. \tag{8.8}$$

**Proof:** It is evident that  $\hat{X}$  satisfies (8.7) if and only if  $(\hat{X} - X) \perp Y_i$  for each  $i$ . With  $\hat{X} = \sum_{j=1}^k a_j^* Y_j$ , then for each  $i$ ,

$$\begin{aligned} 0 &= E[(\hat{X} - X)Y_i] = E\left[\left(\sum_{j=1}^k a_j^* Y_j - X\right) Y_i\right] \\ &= \sum_{j=1}^k a_j^* E[Y_j Y_i] - E[XY_i] \\ &= \sum_{j=1}^k \langle Y_i, Y_j \rangle a_j^* - \langle Y_i, X \rangle. \quad \blacksquare \end{aligned}$$

Since  $Y_1, \dots, Y_n$  have not been stipulated to be linearly independent, the normal equations (8.8) may have multiple solutions, although all define the same element of  $V$ . If  $\{Y_1, \dots, Y_n\}$  is orthonormal, then (8.8) can be solved in closed form:

$$\hat{X} = \sum_{i=1}^n \langle X, Y_i \rangle Y_i. \quad (8.9)$$

## 8.2 Conditional Expectation Given a Finite Set of Random Variables

Conditional expectation is not only important within probability, but also central to the Markov processes, and to statistics as well. Conditional expectations are MMSE predictors among functions of the observable random variables.

### 8.2.1 Basics

Let  $Y_1, \dots, Y_n$  be random variables. The conditional expectation of a random variable  $X$  given  $Y_1, \dots, Y_n$  has two defining properties. First, it is a function of  $Y_1, \dots, Y_n$ . Second, it is indistinguishable from  $X$  in an integrated sense: its expectation equals that of  $X$  over any event determined by  $Y_1, \dots, Y_n$ . Its most important property, though, is that it is the MMSE predictor of  $X$  among functions of  $Y_1, \dots, Y_n$ .

We state the definition generally, but develop the theory first for the case that  $X \in L^2$ , which allows application of results from §1.

**Definition 8.12.** The *conditional expectation of  $X$  given  $Y_1, \dots, Y_n$*  is a random variable  $E[X|Y_1, \dots, Y_n]$  satisfying

- a)  $E[X|Y_1, \dots, Y_n] = h(Y_1, \dots, Y_n)$  for some function  $h: \mathbb{R}^n \rightarrow \mathbb{R}$ .
- b) For every event  $A \in \sigma(Y_1, \dots, Y_n)$ ,

$$E[E[X|Y_1, \dots, Y_n]; A] = E[X; A]. \quad \square \quad (8.10)$$

We also recall from Theorem 2.18 that each event  $A \in \sigma(Y_1, \dots, Y_n)$  has the form  $A = \{(Y_1, \dots, Y_n) \in B\}$  for some  $B \in \mathcal{B}(\mathbb{R}^n)$ .

Before proceeding, we note one useful consequence of (8.10) and examine some examples.

**Proposition 8.13.** We have  $E[E[X|Y_1, \dots, Y_n]] = E[X]$ .

**Proof:** It suffices to take  $A = \Omega$  in (8.10). ■

Often Proposition 8.13 is the most effective means of computing  $E[X]$ , which is the expectation of *any* conditional expectation of  $X$ .

## 8.2.2 Examples

The interpretations of the examples are vivid in terms of prediction: as established in Theorem 8.19,  $E[X|Y_1, \dots, Y_n]$  is the MMSE predictor of  $X$  among all random variables in  $L^2$  that are functions of  $Y_1, \dots, Y_n$ .

Prediction of a constant is, of course, trivial.

**Example 8.14 (Constants).** Conditional expectation preserves constants: if  $X \equiv c$ , then  $E[X|Y_1, \dots, Y_n] \equiv c$  satisfies Definition 8.12. □

The conditional expectation of a random variable  $X$  independent of  $Y_1, \dots, Y_n$  is also constant, but for a different reason, and equal to  $E[X]$ . In prediction terms, knowledge of  $Y_1, \dots, Y_n$  is not useful in predicting the value of  $X$ , which illuminates the nature of independence.

**Example 8.15 (Independence).** If  $X$  and the random vector  $Y = (Y_1, \dots, Y_n)$  are independent, then  $E[X|Y_1, \dots, Y_n] \equiv E[X]$ , since for  $A \in \sigma(Y_1, \dots, Y_n)$ ,

$$E[X; A] = E[X]P(A) = E[E[X]; A]. \quad \square$$

At the other extreme, if  $X$  is a function of  $Y_1, \dots, Y_n$ , then once they are known, so is  $X$ , which is thus its own conditional expectation.

**Example 8.16 (Functional dependence).** If  $X = h(Y_1, \dots, Y_n)$  for some  $h$ , then  $X$  fulfills Definition 8.12, and so  $E[X|Y_1, \dots, Y_n] = X$ . □

The final example exploits closure properties of the family of Poisson distributions under convolution.

**Example 8.17 (Poisson distribution).** Let  $X \stackrel{d}{=} P(\lambda)$  and  $Y \stackrel{d}{=} P(\mu)$  be independent, and let  $Z = X + Y$ . Then,

$$E[X|Z] = \frac{\lambda}{\lambda + \mu} Z.$$

Since  $\sigma(Z)$  is generated by the countable partition  $\{\{Z = k\} : k \in \mathbb{N}\}$ , it suffices to show that for each  $k$ ,

$$E[X; \{Z = k\}] = E\left[\frac{\lambda}{\lambda + \mu} Z; \{Z = k\}\right] = \frac{k\lambda}{\lambda + \mu} e^{-(\lambda+\mu)} \frac{(\lambda + \mu)^k}{k!},$$

where the second equality reflects the property that  $Z \stackrel{d}{=} P(\lambda + \mu)$  (Example 3.16). But this is computational:

$$\begin{aligned} E[X; \{Z = k\}] &= E[X; \{X + Y = k\}] \\ &= \sum_{i=0}^{\infty} \sum_{j=0}^{\infty} i \mathbf{1}(i + j = k) \left(e^{-\lambda} \frac{\lambda^i}{i!}\right) \left(e^{-\mu} \frac{\mu^j}{j!}\right) \\ &= e^{-(\lambda+\mu)} \frac{(\lambda + \mu)^k}{k!} \sum_{i=0}^k i \binom{k}{i} \left(\frac{\lambda}{\lambda + \mu}\right)^i \left(\frac{\mu}{\lambda + \mu}\right)^{k-i} \\ &= \frac{k\lambda}{\lambda + \mu} e^{-(\lambda+\mu)} \frac{(\lambda + \mu)^k}{k!}. \end{aligned}$$

Note that even though  $X$  is integer-valued,  $E[X|Z]$  is not.  $\square$

### 8.2.3 Conditional probability

Conditional probability stands in the same relation to conditional expectation as probability to expectation.

**Definition 8.18.** The *conditional probability* of  $B \in \mathcal{F}$  given the  $Y_1, \dots, Y_n$  is  $P\{B|Y_1, \dots, Y_n\} = E[\mathbf{1}_B|Y_1, \dots, Y_n]$ .  $\square$

For every  $A \in \sigma(Y_1, \dots, Y_n)$ , then,

$$P(A \cap B) = E[P\{B|Y_1, \dots, Y_n\}; A]. \quad (8.11)$$

## 8.3 Conditional Expectation for $X \in L^2$

We now examine conditional expectations  $E[X|Y_1, \dots, Y_n]$  for  $X \in L^2$ .

### 8.3.1 Conditional expectation as MMSE prediction

For  $X \in L^2$ ,  $E[X|Y_1, \dots, Y_n]$  is the orthogonal projection of  $X$  onto

$$V(Y_1, \dots, Y_n) \stackrel{\text{def}}{=} \{Z \in L^2 : Z = h(Y_1, \dots, Y_n) \text{ for some } h: \mathbb{R}^n \rightarrow \mathbb{R}\},$$

the closed subspace of all random variables in  $L^2$  that are functions of  $Y_1, \dots, Y_n$ . In particular,  $E[X|Y_1, \dots, Y_n]$  solves the problem of MMSE prediction of  $X$  given  $V(Y_1, \dots, Y_n)$ .

**Theorem 8.19.** If  $X \in L^2$ , then  $E[X|Y_1, \dots, Y_n]$  is the orthogonal projection of  $X$  onto  $V(Y_1, \dots, Y_n)$ .

**Proof:** We show that Definition 8.12 is satisfied by  $\hat{X}$ , the orthogonal projection of  $X$  onto  $V(Y_1, \dots, Y_n)$ . Since  $\hat{X} \in V(Y_1, \dots, Y_n)$  by construction, only (8.10) needs verification, and it holds because  $X - \hat{X} \in V(Y_1, \dots, Y_n)^\perp$ . Indeed, for  $A \in \sigma(Y_1, \dots, Y_n)$ ,  $\mathbf{1}_A \in V(Y_1, \dots, Y_n)$ , and, hence,  $(X - \hat{X}) \perp \mathbf{1}_A$ :

$$E[X; A] - E[\hat{X}; A] = E[(X - \hat{X})\mathbf{1}_A] = 0.$$

To complete the argument, we show that  $E[X|Y_1, \dots, Y_n]$  is unique. If  $X^*$  also satisfies (8.10), then  $E[(\hat{X} - X^*)\mathbf{1}_A] = 0$  for all  $A \in V(Y_1, \dots, Y_n)$ . By linearity of expectation,  $E[(\hat{X} - X^*)Z] = 0$  for all simple random variables  $Z$  measurable with respect to  $\sigma(Y_1, \dots, Y_n)$ . Choosing simple random variables converging to  $\hat{X} - X^*$  gives  $E[(\hat{X} - X^*)^2] = 0$ , which shows that  $X^* = \hat{X}$ . ■

Contained in Theorem 8.19 is a useful extension of (8.10): as an member of  $V(Y_1, \dots, Y_n)^\perp$ ,  $X - E[X|Y_1, \dots, Y_n]$  is orthogonal to *every* element of  $V(Y_1, \dots, Y_n)$ , not merely indicator functions.

**Corollary 8.20.** For  $X \in L^2$  and  $Z \in V(Y_1, \dots, Y_n)$ ,

$$E[E[X|Y_1, \dots, Y_n]Z] = E[XZ]. \quad \square \quad (8.12)$$

### 8.3.2 Properties of conditional expectation

Conditional expectation is norm-reducing, and, hence, can be interpreted as a contraction operation.

**Proposition 8.21.** For  $X \in L^2$ ,  $\|E[X|Y_1, \dots, Y_n]\| \leq \|X\|$ .

**Proof:** Theorem 8.9 and Theorem 8.19 together imply that

$$\begin{aligned} \|X\|^2 &= \|E[X|Y_1, \dots, Y_n]\|^2 + \|X - E[X|Y_1, \dots, Y_n]\|^2 \\ &\geq \|E[X|Y_1, \dots, Y_n]\|^2. \quad \blacksquare \end{aligned}$$

Conditional expectation is positive, linear and continuous in  $X$ .

**Proposition 8.22.** With  $Y_1, \dots, Y_n$  fixed,

- a) If  $X \in L^2$  is positive, then  $E[X|Y_1, \dots, Y_n] \geq 0$ .

b) For  $X_1, X_2 \in L^2$  and  $a, b \in \mathbb{R}$ ,

$$E[aX_1 + bX_2 | Y_1, \dots, Y_n] = aE[X_1 | Y_1, \dots, Y_n] + bE[X_2 | Y_1, \dots, Y_n].$$

c) The function  $X \mapsto E[X | Y_1, \dots, Y_n]$  is uniformly continuous on  $L^2$ .

**Proof:** a) This property is unexpectedly subtle. The random variable

$$W = E[X | Y_1, \dots, Y_n] \mathbf{1}(E[X | Y_1, \dots, Y_n] < 0)$$

assumes only negative values, but by (8.10), since  $\{E[X | Y_1, \dots, Y_n] < 0\} \in \sigma(Y_1, \dots, Y_n)$ , it also satisfies

$$\begin{aligned} E[W] &= E[E[X | Y_1, \dots, Y_n]; \{E[X | Y_1, \dots, Y_n] < 0\}] \\ &= E[X; \{E[X | Y_1, \dots, Y_n] < 0\}]. \end{aligned}$$

Because  $X \geq 0$ , this last quantity cannot be negative, so that  $W$  must be equal to zero almost surely, and, hence,  $P\{E[X | Y_1, \dots, Y_n] < 0\} = 0$ .

b) Linearity of conditional expectation reprises that of orthogonal projections. With “ $\hat{\cdot}$ ” denoting orthogonal projection onto  $V(Y_1, \dots, Y_n)$ ,

$$\begin{aligned} E[aX_1 + bX_2 | Y_1, \dots, Y_n] &= (aX_1 + bX_2)^{\hat{\cdot}} \\ &= a\hat{X}_1 + b\hat{X}_2 \\ &= aE[X_1 | Y_1, \dots, Y_n] + bE[X_2 | Y_1, \dots, Y_n]. \end{aligned}$$

c) For each  $X$  and  $X'$ ,

$$\begin{aligned} \|E[X | Y_1, \dots, Y_n] - E[X' | Y_1, \dots, Y_n]\| &= \|E[X - X' | Y_1, \dots, Y_n]\| \\ &\leq \|X - X'\| \end{aligned}$$

by linearity of conditional expectation and Proposition 8.21. ■

We conclude with two properties that are peculiar to conditional expectation, and also extremely important. The first is a reduction property for iterated conditional expectations: conditioning first on  $Y_1, \dots, Y_n$ , and then on  $Y_1, \dots, Y_m$ , with  $m < n$ , is the same as having conditioned on  $Y_1, \dots, Y_m$  to begin with.

**Theorem 8.23.** For  $X \in L^2$  and  $m < n$ ,

$$E[E[X | Y_1, \dots, Y_n] | Y_1, \dots, Y_m] = E[X | Y_1, \dots, Y_m]. \quad (8.13)$$

**Proof:** With  $Z = E[X | Y_1, \dots, Y_n]$ , we show that  $Z' = E[X | Y_1, \dots, Y_m]$  satisfies the defining conditions of  $E[Z | Y_1, \dots, Y_m]$ . Evidently,  $Z'$  is a function of  $Y_1, \dots, Y_m$ , while for  $A \in \sigma(Y_1, \dots, Y_m)$ ,

$$E[Z; A] = E[X; A] = E[Z'; A],$$

where the first equality is by the definition of  $Z$  and the inclusion

$$\sigma(Y_1, \dots, Y_m) \subseteq \sigma(Y_1, \dots, Y_n),$$

which holds since  $m < n$ , while the second is by the definition of  $Z'$ . ■

Sometimes (8.13) is an effective means of computing  $E[X|Y_1, \dots, Y_m]$ , in the same way that Proposition 8.13 is useful for computing  $E[X]$ .

Factors that are functions of  $Y_1, \dots, Y_n$  play the same role as constants with regard to ordinary expectations: they can be moved “outside the conditional expectation.” In prediction terms, this is obvious: to predict  $ZX$  when  $Z$  is a function of  $Y_1, \dots, Y_n$ , there is no need to predict  $Z$ , whose value is known, so  $ZX$  should be predicted by  $Z$  times the predictor of  $X$ .

**Theorem 8.24.** If  $Z \in V(Y_1, \dots, Y_n)$  is *bounded*, i.e.,  $Z = h(Y_1, \dots, Y_n)$  for some bounded function  $h: \mathbb{R}^n \rightarrow \mathbb{R}$ , then for all  $X \in L^2$ ,

$$E[ZX|Y_1, \dots, Y_n] = ZE[X|Y_1, \dots, Y_n]. \quad (8.14)$$

**Proof:** Again, let “ $\hat{\cdot}$ ” denote orthogonal projection onto  $V(Y_1, \dots, Y_n)$ . In view of Theorem 8.19, we need to show that  $(ZX)\hat{\cdot} = Z\hat{X}$ , and for this it suffices, by the uniqueness property in Theorem 8.9, to show that  $Z\hat{X} \in V(Y_1, \dots, Y_n)$  and that  $(ZX - Z\hat{X}) \in V(Y_1, \dots, Y_n)^\perp$ . The first is immediate:  $Z$  is a function of  $Y_1, \dots, Y_n$ , and hence so is  $Z\hat{X}$ , while boundedness of  $Z$  gives  $Z\hat{X} \in L^2$ :

$$E[(Z\hat{X})^2] \leq \max_{\omega} |Z(\omega)|^2 E[X^2] < \infty.$$

For  $Y \in V(Y_1, \dots, Y_n)$ , since  $ZY \in V(Y_1, \dots, Y_n)$ ,

$$E[(ZX - Z\hat{X})Y] = E[(X - \hat{X})ZY] = 0$$

by (8.12), so that  $(ZX - Z\hat{X}) \in V(Y_1, \dots, Y_n)^\perp$ . ■

## 8.4 Conditional Expectation for Positive and Integrable Random Variables

We now extend the definition of conditional expectation to the same random variables for which expectation is defined: those that are either positive or integrable. The extension parallels the process used in §4.1 to define expectation. Properties are all analogous either to those of ordinary expectations (for example, the monotone convergence theorem) or to those of conditional expectations for  $X \in L^2$  (for example, the reduction property).

Definition 8.12, of course, was stated in generality, so the primary issue is existence. Let  $Y_1, \dots, Y_n$  be fixed.

**Theorem 8.25.** For every  $X \geq 0$ ,  $E[X|Y_1, \dots, Y_n]$  exists and is unique.

**Proof:** Let  $X_k$  be simple random variables increasing to  $X$ , and put

$$E[X|Y_1, \dots, Y_n] = \lim_{k \rightarrow \infty} E[X_k|Y_1, \dots, Y_n].$$

Each of  $E[X_k|Y_1, \dots, Y_n]$  exists by Theorem 8.19, since  $X_k \in L^2$ , while the limit exists almost surely by Proposition 8.22, since  $(E[X_k|Y_1, \dots, Y_n])$  is an increasing sequence.

Obviously,  $E[X|Y_1, \dots, Y_n]$  is a function of  $Y_1, \dots, Y_n$ , so it remains to verify that it satisfies (8.10). For  $A \in \sigma(Y_1, \dots, Y_n)$ ,

$$\begin{aligned} E[E[X|Y_1, \dots, Y_n]; A] &= E[\lim_{k \rightarrow \infty} E[X_k|Y_1, \dots, Y_n]; A] \\ &= \lim_{k \rightarrow \infty} E[E[X_k|Y_1, \dots, Y_n]; A] \\ &= \lim_{k \rightarrow \infty} E[X_k; A] \\ &= E[X; A] \end{aligned}$$

by two applications of the monotone convergence theorem. ■

Existence of  $E[X|Y_1, \dots, Y_n]$  for  $X \in L^1$  is via the positive and negative parts of  $X$ , in parallel to the definition  $E[X] = E[X^+] - E[X^-]$ .

**Theorem 8.26.** For every  $X \in L^1$ ,  $E[X|Y_1, \dots, Y_n]$  exists, is unique and belongs to  $L^1$ . Moreover,

$$|E[X|Y_1, \dots, Y_n]| \leq E[|X| | Y_1, \dots, Y_n]. \quad (8.15)$$

**Proof:** The definition, as intimated above, is

$$E[X|Y_1, \dots, Y_n] = E[X^+|Y_1, \dots, Y_n] - E[X^-|Y_1, \dots, Y_n],$$

where  $X^+$  and  $X^-$  are the positive and negative parts of  $X$ . This is evidently a function of  $Y_1, \dots, Y_n$ , while for  $A \in \sigma(Y_1, \dots, Y_n)$ ,

$$\begin{aligned} E[E[X|Y_1, \dots, Y_n]; A] &= E[(E[X^+|Y_1, \dots, Y_n] - E[X^-|Y_1, \dots, Y_n]); A] \\ &= E[X^+; A] - E[X^-; A] \\ &= E[X; A]. \end{aligned}$$

That (8.15) holds follows from the identity  $|X| = X^+ + X^-$ . ■

Properties of conditional expectation are established in the same manner as Theorem 8.25 and Theorem 8.26.

**Proposition 8.27.** If either  $X_1, X_2, a_1$  and  $a_2$  are positive or  $X_1, X_2$  are integrable and  $a_1, a_2 \in \mathbb{R}$ , then

$$\begin{aligned} E[a_1X_1 + a_2X_2|Y_1, \dots, Y_n] \\ = a_1E[X_1|Y_1, \dots, Y_n] + a_2E[X_2|Y_1, \dots, Y_n]. \quad \square \end{aligned}$$

**Theorem 8.28 (Fatou's lemma).** If  $X_k \geq 0$  for each  $k$ , then

$$E\left[\liminf_{k \rightarrow \infty} X_k | Y_1, \dots, Y_n\right] \leq \liminf_{k \rightarrow \infty} E[X_k | Y_1, \dots, Y_n]. \quad \square$$

**Theorem 8.29 (Monotone convergence theorem).** If  $0 \leq X_k \uparrow X$ , then

$$E[X_k | Y_1, \dots, Y_n] \uparrow E[X | Y_1, \dots, Y_n]. \quad \square$$

**Theorem 8.30.** If  $X_k \geq 0$  for each  $k$  and  $\sum_{k=1}^{\infty} X_k(\omega) < \infty$  for (almost) every  $\omega$ , then

$$E\left[\sum_{k=1}^{\infty} X_k | Y_1, \dots, Y_n\right] = \sum_{k=1}^{\infty} E[X_k | Y_1, \dots, Y_n]. \quad \square$$

**Theorem 8.31 (Dominated convergence theorem).** If  $X_k \xrightarrow{\text{a.s.}} X$  and if there is  $Z \in L^1$  such that  $|X_k| \leq Z$  almost surely for each  $k$ , then

$$E[X_k | Y_1, \dots, Y_n] \xrightarrow{\text{a.s.}} E[X | Y_1, \dots, Y_n]. \quad \square$$

The proof of Jensen's inequality differs sufficiently from that for ordinary expectations (Theorem 4.39) to merit inclusion.

**Theorem 8.32 (Jensen's inequality).** If  $g: \mathbb{R} \rightarrow \mathbb{R}$  is convex and  $X$  and  $g(X)$  are integrable, then

$$g(E[X | Y_1, \dots, Y_n]) \leq E[g(X) | Y_1, \dots, Y_n].$$

**Proof:** By convexity of  $g$ , there is a countable family of affine functions  $h_i(x) = a_i x + b_i$  such that  $g(x) = \sup_i(a_i x + b_i)$  for all  $x \in \mathbb{R}$ . But then, by linearity of conditional expectation,

$$\begin{aligned} a_j E[X | Y_1, \dots, Y_n] + b_j &= E[a_j X + b_j | Y_1, \dots, Y_n] \\ &\leq E[\sup_i(a_i X + b_i) | Y_1, \dots, Y_n] \\ &= E[g(X) | Y_1, \dots, Y_n], \end{aligned}$$

for each  $j$ , and consequently,

$$\begin{aligned} g(E[X|Y_1, \dots, Y_n]) &= \sup_j (a_j E[X|Y_1, \dots, Y_n] + b_j) \\ &\leq E[g(X)|Y_1, \dots, Y_n]. \quad \blacksquare \end{aligned}$$

Finally, we extend the “peculiar properties.”

**Theorem 8.33.** If  $X$  either is positive or belongs to  $L^1$ , then for  $m \leq n$ ,

$$E[E[X|Y_1, \dots, Y_n]|Y_1, \dots, Y_m] = E[X|Y_1, \dots, Y_m]. \quad \square$$

**Theorem 8.34.** If  $Z = h(Y_1, \dots, Y_n)$  for some function  $h: \mathbb{R}^n \rightarrow \mathbb{R}$ , and if either  $X$  and  $Z$  are positive, or  $X \in L^1$  and  $Z$  is bounded, then

$$E[ZX|Y_1, \dots, Y_n] = ZE[X|Y_1, \dots, Y_n]. \quad \square$$

## 8.5 Conditional Distributions

Before considering computation of conditional expectations, we introduce conditional distributions and conditional distribution functions, whose role *vis-à-vis* conditional expectations parallels that of distribution functions in regard to expectations.

### 8.5.1 Generalities

Here is the basic definition.

**Definition 8.35.** The *conditional distribution of  $X$  given  $Y_1, \dots, Y_n$*  is a function  $P_{X|Y_1, \dots, Y_n}: \mathcal{B}(\mathbb{R}) \times \mathbb{R}^n \rightarrow [0, 1]$ , whose values are written as  $P_{X|Y_1, \dots, Y_n}(B|y_1, \dots, y_n)$ , such that

- a) With  $y_1, \dots, y_n$  fixed,  $B \mapsto P_{X|Y_1, \dots, Y_n}(B|y_1, \dots, y_n)$  is a probability on  $\mathbb{R}$ .
- b) With  $B$  fixed,  $(y_1, \dots, y_n) \mapsto P_{X|Y_1, \dots, Y_n}(B|y_1, \dots, y_n)$  is a Borel measurable function and

$$P_{X|Y_1, \dots, Y_n}(B|Y_1, \dots, Y_n) = P\{X \in B|Y_1, \dots, Y_n\}. \quad \square$$

The notation can be confusing, and it is especially important to be sure of what is really random. The  $X$  and  $Y_1, \dots, Y_n$  in the subscript of  $P_{X|Y_1, \dots, Y_n}(B|y_1, \dots, y_n)$  serve *only to specify what random variables are*

*involved.* They are not functions on  $\Omega$ . On the other hand, on the right-hand side of the equation

$$P\{X \in B|Y_1, \dots, Y_n\} = P_{X|Y_1, \dots, Y_n}(B|Y_1, \dots, Y_n), \quad (8.16)$$

the  $Y_1, \dots, Y_n$  in the second argument of the function  $(B, y_1, \dots, y_n) \mapsto P_{X|Y_1, \dots, Y_n}(B|y_1, \dots, y_n)$  are functions on  $\Omega$ , so that this expression is the function of  $\omega$  obtained by substituting  $Y_1(\omega), \dots, Y_n(\omega)$  for  $y_1, \dots, y_n$ .

Although we do not give a proof, the conditional distribution  $P_{X|Y_1, \dots, Y_n}$  exists and is unique.

For fixed  $y_1, \dots, y_n$  the probability  $B \mapsto P_{X|Y_1, \dots, Y_n}(B|y_1, \dots, y_n)$  on  $\mathcal{B}(\mathbb{R})$  has associated with it a distribution function.

**Definition 8.36.** The *conditional distribution function of  $X$  given  $Y_1, \dots, Y_n$*  is the function

$$\mathsf{F}_{X|Y_1, \dots, Y_n}(t|y_1, \dots, y_n) = P_{X|Y_1, \dots, Y_n}((-\infty, t]|y_1, \dots, y_n). \quad \square \quad (8.17)$$

Hence,  $t \mapsto \mathsf{F}_{X|Y_1, \dots, Y_n}(t|y_1, \dots, y_n)$  is a distribution function for each  $y_1, \dots, y_n$ , and for each  $t$ ,

$$\mathsf{F}_{X|Y_1, \dots, Y_n}(t|Y_1, \dots, Y_n) = P\{X \leq t|Y_1, \dots, Y_n\}.$$

There are two cases where computation of conditional distributions and distribution functions is feasible, namely, that  $Y_1, \dots, Y_n$  are discrete, and that the random vector  $(X, Y_1, \dots, Y_n)$  is absolutely continuous.

### 8.5.2 Discrete random variables

**Theorem 8.37.** Suppose that  $Y_1, \dots, Y_n$  are discrete with values in the countable set  $C$ . Then,

$$P_{X|Y_1, \dots, Y_n}(B|y_1, \dots, y_n) = P\{X \in B|Y_1 = y_1, \dots, Y_n = y_n\}, \quad (8.18)$$

provided that  $P\{Y_1 = y_1, \dots, Y_n = y_n\} > 0$ , in which case the conditional probability is given by Definition 1.43, and which is zero otherwise.

**Proof:** According to Exercise 1.23, if  $P\{Y_1 = y_1, \dots, Y_n = y_n\} > 0$ , then the mapping  $A \mapsto P\{A|Y_1 = y_1, \dots, Y_n = y_n\}$  is a probability on  $(\Omega, \mathcal{F})$ , and by Proposition 2.23,  $B \mapsto P\{X \in B|Y_1 = y_1, \dots, Y_n = y_n\}$  is, therefore, a probability on  $\mathcal{B}(\mathbb{R})$ . For  $B' \in \mathcal{B}(\mathbb{R}^n)$ ,

$$\begin{aligned} & P\{X \in B, (Y_1, \dots, Y_n) \in B'\} \\ &= \sum_{(y_1, \dots, y_n) \in B'} P\{X \in B, (Y_1, \dots, Y_n) = (y_1, \dots, y_n)\} \end{aligned}$$

[the sum is restricted to  $(y_1, \dots, y_n)$  for which  $P\{Y_1 = y_1, \dots, Y_n = y_n\}$  is strictly positive]

$$\begin{aligned} &= \sum_{(y_1, \dots, y_n) \in B'} P\{X \in B | Y_1 = y_1, \dots, Y_n = y_n\} P\{Y_1 = y_1, \dots, Y_n = y_n\} \\ &= \sum_{(y_1, \dots, y_n) \in B'} P_{X|Y_1, \dots, Y_n}(B | y_1, \dots, y_n) P\{Y_1 = y_1, \dots, Y_n = y_n\} \\ &= E[P_{X|Y_1, \dots, Y_n}(B | Y_1, \dots, Y_n); \{(Y_1, \dots, Y_n) \in B'\}]. \end{aligned}$$

Thus,  $P_{X|Y_1, \dots, Y_n}(B | Y_1, \dots, Y_n)$  satisfies the definition of the conditional probability  $P\{X \in B | Y_1, \dots, Y_n\}$ . ■

### 8.5.3 Absolutely continuous random variables

We begin by defining conditional density functions.

**Definition 8.38.** Assume that the random  $(n+1)$ -vector  $(X, Y_1, \dots, Y_n)$  is absolutely continuous with density  $f$ . The *conditional density of  $X$  given  $Y_1, \dots, Y_n$*  is the function

$$f_{X|Y_1, \dots, Y_n}(x | y_1, \dots, y_n) = \frac{f(x, y_1, \dots, y_n)}{\int_{-\infty}^{\infty} f(z, y_1, \dots, y_n) dz}. \quad \square \quad (8.19)$$

As joint density of  $X$  and  $Y_1, \dots, Y_n$ ,  $f$  has the interpretation that

$$P\{X = x, Y_1 = y_1, \dots, Y_n = y_n\} \cong f(x, y_1, \dots, y_n) dx dy_1 \cdots dy_n.$$

Similarly,  $f_Y(y_1, \dots, y_n) = \int_{-\infty}^{\infty} f(z, y_1, \dots, y_n) dz$ , the marginal density of  $Y_1, \dots, Y_n$ , has the interpretation that

$$P\{Y_1 = y_1, \dots, Y_n = y_n\} \cong f_Y(y_1, \dots, y_n) dy_1 \cdots dy_n.$$

Consequently, by analogy to the definition of conditional probability given an event, the conditional density  $f_{X|Y_1, \dots, Y_n}$  has the interpretation

$$\begin{aligned} &f_{X|Y_1, \dots, Y_n}(x | y_1, \dots, y_n) \\ &\cong \frac{P\{X = x, Y_1 = y_1, \dots, Y_n = y_n\} dx dy_1 \cdots dy_n}{P\{Y_1 = y_1, \dots, Y_n = y_n\} dy_1 \cdots dy_n} \\ &\cong P\{X = x | Y_1 = y_1, \dots, Y_n = y_n\} dx. \end{aligned}$$

The conditional distribution and conditional distribution function of  $X$  given  $Y_1, \dots, Y_n$  are obtained by integrating the conditional density.

**Theorem 8.39.** Assume that  $(X, Y_1, \dots, Y_n)$  is absolutely continuous. Then,

$$P_{X|Y_1, \dots, Y_n}(B | y_1, \dots, y_n) = \int_B f_{X|Y_1, \dots, Y_n}(x | y_1, \dots, y_n) dx.$$

Consequently,

$$\mathsf{F}_{X|Y_1, \dots, Y_n}(t|y_1, \dots, y_n) = \int_{-\infty}^t f_{X|Y_1, \dots, Y_n}(x|y_1, \dots, y_n) dx.$$

**Proof:** For simplicity, let  $Y = (Y_1, \dots, Y_n)$  and  $y = (y_1, \dots, y_n)$ , and let  $B \in \mathcal{B}(\mathbb{R})$  be fixed. Then, evidently,  $\int_B f_{X|Y}(x|y) dx$  is a measurable function of  $y$ , while for  $B' \in \mathcal{B}(\mathbb{R}^n)$ , Theorem 4.28 gives

$$\begin{aligned} P\{X \in B, Y \in B'\} &= \int_{B \times B'} f(x, y) dx dy \\ &= \int_{B \times B'} [f_{X|Y}(x|y) f_Y(y)] dx dy \\ &= \int_{B'} \left[ \int_B f_{X|Y}(x|y) dx \right] f_Y(y) dy \end{aligned}$$

[since  $f_Y(y) = \int_{-\infty}^{\infty} f(z, y) dz$  is the (marginal) density of  $Y$  by (2.4)]

$$= E[\int_B f_{X|Y}(x|Y) dx; \{Y \in B'\}]$$

by another application of Theorem 4.28, so that  $\int_B f_{X|Y}(x|Y) dx$  satisfies the definition of  $P\{X \in B|Y\}$ . ■

## 8.6 Computational Techniques

Our goal now is to compute conditional expectations  $E[g(X)|Y_1, \dots, Y_n]$ . According to Definition 8.12,  $E[g(X)|Y_1, \dots, Y_n]$  is some function  $h$  of  $Y_1, \dots, Y_n$ , so the point is to determine *which function*.

### 8.6.1 General results

In the same way that expectations are integrals with respect to distribution functions (see §4.3), conditional expectations are integrals with respect to conditional distribution functions.

**Theorem 8.40.** If either  $g \geq 0$  or  $g(X) \in L^1$ , then

$$E[g(X)|Y_1, \dots, Y_n] = \int_{-\infty}^{\infty} g(t) d\mathsf{F}_{X|Y_1, \dots, Y_n}(t|Y_1, \dots, Y_n). \quad (8.20)$$

**Proof:** We use the usual approximation argument. Again, to simplify the notation, let  $Y = (Y_1, \dots, Y_n)$  and  $y = (y_1, \dots, y_n)$ .

To begin with, if  $g$  is the indicator function of  $B \in \mathcal{B}(\mathbb{R})$ , then (8.20) reduces to (8.16).

If  $g = \sum_{i=1}^n a_i \mathbf{1}_{B_i}$  is a simple function, then by part a) of Definition 8.35, for each  $y$ ,

$$\int_{-\infty}^{\infty} g(t) dF_{X|Y}(t|y) = \sum_{i=1}^n a_i \int_{B_i} dF_{X|Y}(t|y) = \sum_{i=1}^n a_i P_{X|Y}(B_i|y).$$

Consequently, for  $B \in \mathcal{B}(\mathbb{R}^n)$ ,

$$\begin{aligned} E[g(X); \{Y \in B\}] &= E\left[\sum_{i=1}^n a_i \mathbf{1}(X \in B_i); \{Y \in B\}\right] \\ &= \sum_{i=1}^n a_i E[\mathbf{1}(X \in B_i); \{Y \in B\}] \\ &= \sum_{i=1}^n a_i E[P\{X \in B_i|Y\}; \{Y \in B\}] \\ &= \sum_{i=1}^n a_i E[P_{X|Y}(B_i|Y); \{Y \in B\}] \\ &= E\left[\sum_{i=1}^n a_i P_{X|Y}(B_i|Y); \{Y \in B\}\right] \\ &= E\left[\int_{-\infty}^{\infty} g(t) dF_{X|Y}(t|y); \{Y \in B\}\right], \end{aligned}$$

so that  $\int_{-\infty}^{\infty} g(t) dF_{X|Y}(t|y)$  satisfies the definition of  $E[g(X)|Y]$ .

If  $g, g_n \geq 0$  and the  $g_n$  are simple with  $g_n \uparrow g$ , then for each  $B$ ,

$$\begin{aligned} E[g(X); \{Y \in B\}] &= \lim_{n \rightarrow \infty} E[g_n(X); \{Y \in B\}] \\ &= \lim_{n \rightarrow \infty} E\left[\int_{-\infty}^{\infty} g_n(t) dF_{X|Y}(t|y); \{Y \in B\}\right] \\ &= E\left[\int_{-\infty}^{\infty} g(t) dF_{X|Y}(t|y); \{Y \in B\}\right]. \end{aligned}$$

The first equality is by the monotone convergence theorem for expectations, and the last is by the monotone convergence theorem for integrals with respect to distribution functions. Thus, once more,  $\int_{-\infty}^{\infty} g(t) dF_{X|Y}(t|y)$  satisfies the definition of  $E[g(X)|Y]$ .

Finally, if  $g(X) \in L^1$ , then

$$\begin{aligned} E[g(X); \{Y \in B\}] &= E[g^+(X); \{Y \in B\}] - E[g^-(X); \{Y \in B\}] \\ &= E\left[\int_{-\infty}^{\infty} g^+(t) dF_{X|Y}(t|y); \{Y \in B\}\right] \\ &\quad - E\left[\int_{-\infty}^{\infty} g^-(t) dF_{X|Y}(t|y); \{Y \in B\}\right] \\ &= E\left[\int_{-\infty}^{\infty} g(t) dF_{X|Y}(t|y); \{Y \in B\}\right], \end{aligned}$$

and still (8.20) holds. ■

In particular, if either  $X \geq 0$  or  $X \in L^1$ , then

$$E[X|Y_1, \dots, Y_n] = \int_{-\infty}^{\infty} t dF_{X|Y_1, \dots, Y_n}(t|Y_1, \dots, Y_n), \quad (8.21)$$

which is analogous to the computational relationships (4.14) and (4.16) for ordinary expectations.

### 8.6.2 Special cases

Two cases of interest, as in §5, are those where  $Y_1, \dots, Y_n$  are discrete and where  $(X, Y_1, \dots, Y_n)$  is absolutely continuous. The results follow immediately from Theorem 8.40; no proofs are necessary.

**Proposition 8.41.** If  $Y_1, \dots, Y_n$  are discrete, then for  $X$  and functions  $g$  such that either positive  $g \geq 0$  or  $g(X) \in L^1$ ,  $E[g(X)|Y_1, \dots, Y_n] = h(Y_1, \dots, Y_n)$ , where

$$h(y_1, \dots, y_n) = \int_0^{\infty} g(t) dF_{X|Y_1, \dots, Y_n}(t|Y_1 = y_1, \dots, Y_n = y_n), \quad (8.22)$$

provided that  $P\{Y_1 = y_1, \dots, Y_n = y_n\} > 0$ , and is zero otherwise.  $\square$

**Proposition 8.42.** If  $(X, Y_1, \dots, Y_n)$  is absolutely continuous, then  $E[g(X)|Y_1, \dots, Y_n] = h(Y_1, \dots, Y_n)$  for the function  $h$  given by

$$h(y_1, \dots, y_n) = \int_{-\infty}^{\infty} g(x) f_{X|Y_1, \dots, Y_n}(x|y_1, \dots, y_n) dx. \quad \square \quad (8.23)$$

## 8.7 Complements

### 8.7.1 Mixed conditional distributions

If the distributions  $B \mapsto P_{X|Y_1, \dots, Y_n}(B|y_1, \dots, y_n)$  are mixed, in the sense of having both discrete and absolutely continuous components, then in principle no new difficulties emerge. It is crucial, however, that the “mixing proportions” can depend on  $y_1, \dots, y_n$ . Thus, the conditional distribution  $P_{X|Y_1, \dots, Y_n}$  is termed mixed if there exist a function  $\alpha: \mathbb{R}^n \rightarrow [0, 1]$  and for each  $(y_1, \dots, y_n) \in \mathbb{R}^n$ , a discrete probability  $\sum_i p_i(y_1, \dots, y_n) \varepsilon_{t_i(y_1, \dots, y_n)}$ , where the masses  $p_i(y_1, \dots, y_n)$  and the locations  $t_i(y_1, \dots, y_n)$  depend (in a Borel measurable manner) on  $(y_1, \dots, y_n)$ , and a function  $f: \mathbb{R}^{n+1} \rightarrow \mathbb{R}_+$  with  $\int_{-\infty}^{\infty} f(x|y_1, \dots, y_n) dx = 1$ , such that

$$\begin{aligned} P_{X|Y_1, \dots, Y_n}(B|y_1, \dots, y_n) \\ = \alpha(y_1, \dots, y_n) \sum_{i: t_i(y_1, \dots, y_n) \in B} p_i(y_1, \dots, y_n) \\ + [1 - \alpha(y_1, \dots, y_n)] \int_B f(x|y_1, \dots, y_n) dx \end{aligned}$$

for all  $B$ . Then,

$$\begin{aligned} F_{X|Y_1, \dots, Y_n}(t|y_1, \dots, y_n) \\ = \alpha(y_1, \dots, y_n) \sum_{i: t_i(y_1, \dots, y_n) \leq t} p_i(y_1, \dots, y_n) \\ + [1 - \alpha(y_1, \dots, y_n)] \int_{-\infty}^t f(x|y_1, \dots, y_n) dx. \end{aligned}$$

### 8.7.2 Conditional expectation given a $\sigma$ -algebra

One can define conditional expectation given  $\mathcal{H}$  a sub- $\sigma$ -algebra  $\mathcal{H}$  of  $\mathcal{F}$ . Recall from Definition 2.51 that a random variable  $Y$  is measurable with respect to  $\mathcal{H}$  if  $\{Y \in B\} \in \mathcal{H}$  for every  $B \in \mathcal{B}(\mathbb{R})$ .

**Definition 8.43.** The *conditional expectation of  $X$  given  $\mathcal{H}$*  is a random variable  $E[X|\mathcal{H}]$  such that

- a)  $E[X|\mathcal{H}]$  is measurable with respect to  $\mathcal{H}$ .
- b) For all  $A \in \mathcal{H}$ ,  $E[E[X|\mathcal{H}]; A] = E[X; A]$ .  $\square$

This reduces to the conditional expectation studied in §2 when  $\mathcal{H} = \sigma(Y_1, \dots, Y_n)$ . The properties are essentially the same.

**Theorem 8.44.** For  $X \in L^2$ ,  $E[X|\mathcal{H}]$  is the orthogonal projection of  $X$  onto the closed subspace

$$V(\mathcal{H}) \stackrel{\text{def}}{=} \{Z \in L^2 : Z \text{ is measurable with respect to } \mathcal{H}\}.$$

Thus,  $\|E[X|\mathcal{H}] - X\| \leq \|Z - X\|$  for all  $Z \in V(\mathcal{H})$ .  $\square$

**Corollary 8.45.** For  $X \in L^2$ ,  $E[X|\mathcal{H}] \in L^2$  and  $\|E[X|\mathcal{H}]\| \leq \|X\|$ .  $\square$

**Theorem 8.46.** For  $X$  either positive or belonging to  $L^1$ ,  $E[X|\mathcal{H}]$  exists and is unique.  $\square$

**Corollary 8.47.** For  $X \in L^1$ ,  $|E[X|\mathcal{H}]| \leq E[|X||\mathcal{H}]$ .  $\square$

## 8.8 Exercises

**8.1.** Prove the polarization identity (8.3).

**8.2.** Show that  $X, Y \in L^2$  are orthogonal if and only if  $\|X + Y\|^2 = \|X\|^2 + \|Y\|^2$ . (This is the Pythagorean law.)

**8.3.** Show that an orthonormal subset of  $L^2$  is linearly independent.

**8.4.** Use the normal equations (8.8) to show that for  $X \in L^2$ ,

$$\arg \min_{a \in \mathbb{R}} E[(X - a)^2] = E[X].$$

**8.5.** Consider a prediction problem in which the set of allowable predictors is  $V = \{aY + b: a, b \in \mathbb{R}\}$ , where  $Y \in L^2$  and  $\sigma^2 = \text{Var}(Y) > 0$ .

- a) Show that  $Z_1 = (Y - E[Y])/\sigma$  and  $Z_2 \equiv 1$  constitute an orthonormal basis of  $V$ .
- b) Show that for  $X \in L^2$ , the MMSE predictor of  $X$  within  $V$  is

$$\hat{X} = \frac{\text{Cov}(X, Y)}{\sigma^2} (Y - E[Y]) + E[X].$$

**8.6.** Prove that if  $X$  and  $Y$  are independent, then for each function  $h$ ,  $E[h(X, Y)|Y] = \int h(x, Y) dF_X(x)$ .

**8.7.** Let  $X, Y \in L^2$  be such that  $E[X|Y] = Y$  and  $E[Y|X] = X$ . Prove that  $X \stackrel{\text{a.s.}}{=} Y$ .

**8.8.** Use Example 8.17 to show that if  $(N_t)$  is a Poisson process with rate  $\lambda$ , then for  $s < t$ ,  $E[N_s|N_t] = (s/t)N_t$ .

**8.9.** Prove that for bounded random variables  $X$  and  $Z$ ,

$$E[E[X|Y_1, \dots, Y_n]Z] = E[XE[Z|Y_1, \dots, Y_n]].$$

**8.10.** Suppose that  $X \stackrel{\text{d}}{=} E(1)$ . For each  $t$ , calculate  $E[X|\min\{X, t\}]$  and  $E[X|\max\{X, t\}]$ .

**8.11.** For  $X \in L^2$ , define the *conditional variance of  $X$  given  $Y_1, \dots, Y_n$*  as

$$\text{Var}(X|Y_1, \dots, Y_n) = E[(X - E[X|Y_1, \dots, Y_n])^2|Y_1, \dots, Y_n].$$

Prove that

$$\text{Var}(X) = E[\text{Var}(X|Y_1, \dots, Y_n)] + \text{Var}(E[X|Y_1, \dots, Y_n]).$$

**8.12.** Let  $(X, Y)$  have a bivariate normal distribution with correlation  $\rho$ . Calculate  $E[X|Y]$  and  $E[Y|X]$ .

**8.13.** Let  $X_1, X_2, \dots$  be i.i.d. with  $E[|X_1|] < \infty$ , and let  $S_n = \sum_{i=1}^n X_i$ . Prove that for each  $n$ ,  $E[X_1|S_n, S_{n+1}, \dots] = S_n/n$ , and hence that  $E[X_1|S_n, S_{n+1}, \dots] \xrightarrow{\text{a.s.}} E[X_1]$ .

**8.14.** Prove that a random variable  $X$  and random vector  $Y = (Y_1, \dots, Y_n)$  are independent if and only if  $E[g(X)|Y_1, \dots, Y_n] = E[g(X)]$  for all bounded functions  $g: \mathbb{R} \rightarrow \mathbb{R}$ .

**8.15.** Let  $Y_1, \dots, Y_{n+1}$  be random variables. Prove that for every  $X \in L^2$ ,

$$E[(E[X|Y_1, \dots, Y_{n+1}] - X)^2] \leq E[(E[X|Y_1, \dots, Y_n] - X)^2],$$

and interpret this in terms of prediction.

**8.16.** Let  $P$  be the uniform distribution on  $[0, 1]$ , and let  $Y$  be the random variable  $Y(\omega) = \omega(1 - \omega)$ . Calculate  $E[X|Y]$  for each positive  $X$ .

**8.17.** Let  $X_1, X_2, \dots$  be i.i.d. with  $E[X_k] = \mu$  and  $\sigma^2 = \text{Var}(X_k) < \infty$ . Define the partial sums  $S_n = \sum_{i=1}^n X_i$ , with  $S_0 = 0$ . Let  $N$  be positive, integer-valued and independent of  $(X_n)$ , and consider the random sum  $S_N = \sum_{i=1}^N X_i$ .

a) Show that

$$E[S_N|N] = \mu N,$$

and, hence, that  $E[S_N] = \mu E[N]$ .

b) Show that

$$E[S_N^2|N] = \sigma^2 N + \mu^2 N^2,$$

and conclude from this that

$$E[S_N^2] = \sigma^2 E[N] + \mu^2 E[N^2]$$

and

$$\text{Var}(S_N) = \sigma^2 E[N] + \mu^2 \text{Var}(N).$$

c) Show that for each  $t$ ,  $E[e^{itS_N}|N] = \varphi_{X_1}(t)^N$ , so that

$$\varphi_{S_N}(t) = E[\varphi_{X_1}(t)^N] = \zeta_N(\varphi_{X_1}(t)).$$

**8.18.** Suppose that  $(X, Y_1, \dots, Y_n)$  has a multivariate normal distribution (Definition 4.46). Show that the orthogonal projection  $\hat{X}$  of  $X$  onto the subspace

$$V'(Y_1, \dots, Y_n) = \left\{ \sum_{i=1}^n a_i Y_i : a_1, \dots, a_n \in \mathbb{R} \right\}$$

satisfies the definition of  $E[X|Y_1, \dots, Y_n]$ . In this case, prediction and linear prediction are synonymous.

**8.19.** Suppose that  $\{A_1, A_2, \dots\}$  is a countable partition of  $\Omega$  with  $P(A_i) > 0$  for each  $i$ , so that the  $\sigma$ -algebra  $\mathcal{H}$  generated by this partition consists of all unions of some of the  $A_i$ :  $\mathcal{H} = \{\sum_{i \in I} A_i : I \subseteq \mathbb{N}\}$ . Show that for each  $X \geq 0$ ,  $E[X|\mathcal{H}] = \sum_{i=1}^{\infty} E[X|A_i] \mathbf{1}_{A_i}$ , where  $E[X|A_i]$  is the expectation of  $X$  with respect to the probability  $P(\cdot|A_i)$  (Exercise 1.23).

**8.20.** Suppose that  $X \stackrel{d}{=} P(\lambda)$  and that conditional on  $X$ ,  $Y \stackrel{d}{=} N(0, X)$ .

- a) Calculate the characteristic function of  $Y$ .
- b) Prove that  $Y$  is not absolutely continuous.
- c) Show that as  $\lambda \rightarrow \infty$ ,  $Y/\lambda \xrightarrow{d} N(0, 1)$ .

**8.21.** Let  $Y \stackrel{d}{=} E(1)$  and suppose that conditional on the value  $y$  of  $Y$ ,  $X \stackrel{d}{=} P(y)$ . Calculate the distribution of  $X$ .

**8.22.** Show that for  $m < n$ ,

$$E[E[X|Y_1, \dots, Y_m] | Y_1, \dots, Y_n] = E[X|Y_1, \dots, Y_m].$$

# Chapter 9

## Martingales

Martingales originated as mathematical models of fair gambling games, and because of the variety of situations in which they arise and the powerful theory that has been developed for them, have become central objects in probability and statistics. Along with Markov processes and stationary processes, they are one of three key departures from independence.

### 9.1 Fundamentals

Let  $X = (X_n)_{n \geq 0}$  and  $Y = (Y_n)_{n \geq 0}$  be sequences of random variables, or discrete time stochastic processes, over  $(\Omega, \mathcal{F}, P)$ .

We first define what it means for  $X$  to be adapted to  $Y$ . Thinking of  $Y$  as a fundamental random phenomenon, for each  $n$ ,  $Y_0, \dots, Y_n$  comprise the history, past and present, at time  $n$ . In order that  $X$  be adapted to  $Y$ ,  $X_n$  must be an observable part of this history.

**Definition 9.1.** The sequence  $X$  is *adapted to  $Y$*  if for each  $n$ ,

$$X_n = h_n(Y_0, \dots, Y_n)$$

for some function  $h_n: \mathbb{R}^{n+1} \rightarrow \mathbb{R}$ .  $\square$

By Theorem 2.52,  $X$  is adapted if for each  $n$ ,  $X_n$  is measurable with respect to  $\sigma(Y_0, \dots, Y_n)$ .

#### 9.1.1 Definitions

**Definition 9.2.** Suppose that  $X$  is adapted to  $Y$  and that  $E[|X_n|] < \infty$  for each  $n$ .

a)  $X$  is a *martingale with respect to  $Y$*  if for each  $n$ ,

$$E[X_{n+1}|Y_0, \dots, Y_n] = X_n. \quad (9.1)$$

b)  $X$  is a *submartingale with respect to  $Y$*  if for each  $n$ ,

$$E[X_{n+1}|Y_0, \dots, Y_n] \geq X_n. \quad (9.2)$$

c)  $X$  is a *supermartingale with respect to  $Y$*  if for each  $n$ ,

$$E[X_{n+1}|Y_0, \dots, Y_n] \leq X_n. \quad \square \quad (9.3)$$

In order to minimize the proliferation of “with respect to” qualifiers, we omit them when there is no danger of confusion.

Obviously, submartingales and supermartingales are virtually the same, save for a change of sign ( $X$  is a submartingale if and only if  $-X$  is a supermartingale), and it is not necessary to state results for both. Here we confine attention to submartingales.

Martingales are mathematical models of fair games. In this context,  $X_n$  is a gambler’s fortune after  $n$  plays of the game, and (9.1) shows that the MMSE predictor of the next fortune  $X_{n+1}$  is just the current fortune  $X_n$ . Submartingales, then, model super-fair games, which favor the gambler, and supermartingales are models of sub-fair games, which are disadvantageous to the gambler.

**Etymological Note.** The word “martingale” is the most interesting in probability. Among other things, it means a gambling system in which losing bets are doubled (which led to its being introduced in probability), a part of a horse’s harness, a part of the rigging of a sailing ship, and a belt on the back of a man’s coat.  $\square$

All three defining properties — (9.1), (9.2) and (9.3) — extend further than one step into the future. Thus, if  $X$  is a martingale and  $m \geq 1$ ,

$$E[X_{n+m}|Y_0, \dots, Y_n] = X_n; \quad (9.4)$$

if  $X$  is a submartingale, then

$$E[X_{n+m}|Y_0, \dots, Y_n] \geq X_n \quad (9.5)$$

for all  $m \geq 1$ ; and if  $X$  is a supermartingale and  $m \geq 1$ , then

$$E[X_{n+m}|Y_0, \dots, Y_n] \leq X_n. \quad (9.6)$$

The expectation consequences of these properties are important as well. For a martingale,

$$E[X_n] = E[X_0] \quad (9.7)$$

for all  $n$ , for a submartingale,

$$E[X_{n+m}] \geq E[X_n]$$

for all  $n$  and  $m$ , and for a supermartingale,

$$E[X_{n+m}] \leq E[X_n]$$

for all  $n$  and  $m$ .

### 9.1.2 Examples

The first example comprises all uniformly integrable martingales.

**Example 9.3 (Successive conditional expectations).** Suppose that  $E[|X|] < \infty$ , and let  $Y$  be any sequence. Then, the sequence  $X_n = E[X|Y_0, \dots, Y_n]$  is a martingale. Evidently  $X$  is adapted to  $Y$ : by definition,  $E[X|Y_0, \dots, Y_n]$  is a function of  $Y_0, \dots, Y_n$ . Moreover,

$$E[|X_n|] = E[|E[X|Y_0, \dots, Y_n]|] < \infty.$$

Finally, for each  $n$ , Theorem 8.33 implies that

$$\begin{aligned} E[X_{n+1}|Y_0, \dots, Y_n] &= E[E[X|Y_0, \dots, Y_{n+1}]|Y_0, \dots, Y_n] \\ &= E[X|Y_0, \dots, Y_n] \\ &= X_n. \quad \square \end{aligned}$$

The second example shows that martingales generalize partial sums of independent, mean zero random variables.

**Example 9.4 (Sums of independent random variables).** Let  $Y_1, Y_2, \dots$  be independent with  $E[Y_k] = 0$  for each  $k$ , and let  $X_n = \sum_{k=1}^n Y_k$ , with  $X_0 = Y_0 = 0$ . Then,  $X$  is a martingale with respect to  $Y$ . Indeed,  $X_n$  is a function of  $Y_0, \dots, Y_n$  by construction, and for each  $n$ ,  $E[|X_n|] \leq \sum_{k=1}^n E[|Y_k|] < \infty$ . Finally, for each  $n$ ,

$$\begin{aligned} E[X_{n+1}|Y_0, \dots, Y_n] &= E[X_n + Y_{n+1}|Y_0, \dots, Y_n] \\ &= X_n + E[Y_{n+1}|Y_0, \dots, Y_n] \\ &= X_n + E[Y_{n+1}], \end{aligned}$$

by Example 8.15, since  $Y_{n+1}$  is independent of  $Y_0, \dots, Y_n$ . But  $E[Y_{n+1}] = 0$ , and, hence, (9.1) holds.  $\square$

**Example 9.5 (Products of independent random variables).** Suppose that  $Y_1, Y_2, \dots$  are independent with  $E[Y_k] = 1$  for each  $k$ , and let  $X_n = \prod_{i=1}^n Y_i$ , with  $X_0 = Y_0 = 1$ . Then,  $X$  is a martingale with respect to  $Y$ . First,  $X_n$  is a function of  $Y_0, \dots, Y_n$  by construction. By independence,  $E[|X_n|] = E[\prod_{i=1}^n |Y_i|] = \prod_{i=1}^n E[|Y_i|] < \infty$ . Finally, for each  $n$ ,

$$\begin{aligned} E[X_{n+1}|Y_0, \dots, Y_n] &= E[X_n Y_{n+1}|Y_0, \dots, Y_n] \\ &= X_n E[Y_{n+1}|Y_0, \dots, Y_n] \\ &= X_n E[Y_{n+1}] \\ &= X_n \end{aligned}$$

since  $Y_{n+1}$  is independent of  $Y_0, \dots, Y_n$  and  $E[Y_{n+1}] = 1$ .  $\square$

**Example 9.6 (Likelihood ratios).** Let  $f$  and  $g$  be density functions on  $\mathbb{R}$  such that  $f \neq g$ , and with (for simplicity)  $f(x) > 0$  and  $g(x) > 0$  for all  $x$ . Let  $Y_1, Y_2, \dots$  be i.i.d. random variables, whose density is either  $f$  or  $g$ , but which of the two is the true density is not known. A central problem in statistical inference is to decide on the basis of the data  $Y_k$  whether the density is  $f$  or  $g$ .

A key tool is the *likelihood ratio* sequence

$$X_n = \prod_{k=1}^n \frac{g(Y_k)}{f(Y_k)},$$

where  $X_0 \equiv 1$ . The rationale is that if  $g$  were the true density, then  $\prod_{k=1}^n g(Y_k)$  would be the probability, in the sense of (2.2), of having observed the values  $Y_1, \dots, Y_n$ , which would be  $\prod_{k=1}^n f(Y_k)$  if  $f$  were the true density. Intuitively, random variables are likely to lie in sets where their density function is large, so if  $g$  were the density, then the  $X_n$  would be large, while if  $f$  were, then the  $X_n$  would be small. A more precise statement is that if the  $Y_k$  have density  $g$ , then  $X$  is a submartingale with respect to  $Y$ , while if the  $Y_k$  have density  $f$ , then  $X$  is a martingale.

The first of these properties is a special case of Example 9.5. With  $P_f$  a probability under which the  $Y_k$  have density  $f$ , for each  $n$ ,

$$\begin{aligned} E_f[X_{n+1}|Y_0, \dots, Y_n] &= E_f[X_n g(Y_{n+1})/f(Y_{n+1})|Y_0, \dots, Y_n] \\ &= X_n E_f[g(Y_{n+1})/f(Y_{n+1})] \\ &= X_n \int_{-\infty}^{\infty} [g(x)/f(x)] f(x) dx, \end{aligned}$$

which is simply  $X_n$ . Hence,  $X$  is a martingale over  $(\Omega, \mathcal{F}, P_f)$ .

On the other hand, let the  $Y_k$  have density  $g$  under  $P_g$ . Then,

$$\begin{aligned} E_g[X_{n+1}|Y_0, \dots, Y_n] &= E_g[X_n g(Y_{n+1})/f(Y_{n+1})|Y_0, \dots, Y_n] \\ &= X_n E_g[g(Y_{n+1})/f(Y_{n+1})] \\ &= X_n \int_{-\infty}^{\infty} [g(x)/f(x)] g(x) dx \\ &= X_n E_f\left[\left(g(Y_1)/f(Y_1)\right)^2\right] \\ &\geq X_n E_f[g(Y_1)/f(Y_1)]^2 \\ &= X_n, \end{aligned}$$

with the penultimate step by the Cauchy-Schwarz inequality or Jensen's inequality. Thus,  $X$  is a submartingale over  $(\Omega, \mathcal{F}, P_g)$ .

We see in §5 that under  $P_f$ ,  $X_n \xrightarrow{\text{a.s.}} 0$ , while under  $P_g$ ,  $X_n \xrightarrow{\text{a.s.}} \infty$ .  $\square$

**Example 9.7 (Pólya urn scheme).** Consider an urn that initially contains one red ball and one black ball. At each time  $n = 1, 2, \dots$ , a

ball is drawn, its color noted, and it and another ball of the same color are placed back into the urn. Thus, after  $n$  draws the urn contains  $n+2$  balls. Let  $Y_n$  be the number of black balls in the urn after  $n$  draws and  $X_n = Y_n/(n+2)$  the *fraction* of black balls. Then,  $X$  is a martingale with respect to  $Y$ .

Only (9.1) needs to be checked, since  $X_n$  is a function of  $Y_n$  and since  $0 \leq X_n \leq 1$ . Noting that the possible values of  $Y_n$  are  $1, \dots, n+1$ , we have

$$\begin{aligned} P\{Y_{n+1} = k+1 | Y_1 = i_1, \dots, Y_{n-1} = i_{n-1}, Y_n = k\} &= \frac{k}{n+2} \\ P\{Y_{n+1} = k | Y_1 = i_1, \dots, Y_{n-1} = i_{n-1}, Y_n = k\} &= \frac{n+2-k}{n+2}, \end{aligned}$$

so that

$$E[X_{n+1} | Y_0, \dots, Y_n] = \frac{Y_n + 1}{n+3} X_n + \frac{Y_n}{n+3} (1 - X_n) = X_n. \quad \square$$

**Example 9.8 (Branching processes).** Let  $\{Z_{ni} : n \geq 0, i \geq 1\}$  be i.i.d., positive and integer-valued with  $0 < m = E[Z_{ni}] < \infty$ . Let  $Y_0 = 1$ , and for each  $n$ , put  $Y_{n+1} = \sum_{i=1}^{Y_n} Z_{ni}$ , with  $Y_{n+1} = 0$  if  $Y_n = 0$ . The stochastic process  $Y$  is a *branching process*, and can be used to model phenomena such as populations, disease epidemics and nuclear reactions. The interpretation is that  $Y_n$  is the size of the  $n$ th generation in a population whose members reproduce independently, with  $Z_{ni}$  the number of progeny of the  $i$ th member of that generation. Extinction of the population prior to time  $n$  corresponds to the event  $\{Y_n = 0\}$ .

The process  $X_n = Y_n/m^n$  is a martingale with respect to  $Y$ . First of all,  $X_n$  is trivially a function of  $Y_n$ . For each  $n$ ,  $\{Y_0, \dots, Y_n\}$  and  $\{Z_{ni} : i \geq 1\}$  are independent, and, hence,

$$E[Y_{n+1} | Y_0, \dots, Y_n] = E\left[\sum_{i=1}^{Y_n} Z_{ni} \mid Y_n\right] = mY_n. \quad (9.8)$$

Consequently,  $E[Y_{n+1}] = E[Y_n]$ , so that  $E[Y_{n+1}] = m^{n+1}$  by induction. Finally, by (9.8),  $E[X_{n+1} | Y_0, \dots, Y_n] = mY_n/m^{n+1} = X_n$ .  $\square$

### 9.1.3 Compositions and transformations

The classes of martingales, supermartingales and submartingales are closed under linear combinations, although in the latter two cases the scalar multipliers must be positive.

**Proposition 9.9.** Let  $Y$  be fixed.

- a) If  $X$  and  $Z$  are martingales and if  $a, b \in \mathbb{R}$ , then  $aX + bZ = (aX_n + bZ_n)$  is a martingale.

- b) If  $X$  and  $Z$  are submartingales and if  $a, b \geq 0$ , then  $aX + bZ$  is a submartingale.
- c) If  $X$  and  $Z$  are supermartingales and if  $a, b \geq 0$ , then  $aX + bZ$  is a supermartingale.

**Proof:** We prove part a) for illustration. If  $X_n$  and  $Z_n$  are functions of  $Y_0, \dots, Y_n$ , then so is  $aX_n + bZ_n$ , and

$$E[|aX_n + bZ_n|] \leq |a| E[|X_n|] + |b| E[|Z_n|] < \infty.$$

Finally, by linearity of conditional expectation

$$\begin{aligned} E[aX_{n+1} + bZ_{n+1}|Y_0, \dots, Y_n] \\ = aE[X_{n+1}|Y_0, \dots, Y_n] + bE[Z_{n+1}|Y_0, \dots, Y_n] \\ = aX_n + bZ_n. \end{aligned}$$

The other parts are proved similarly. ■

A convex function of a martingale is a submartingale, as is a convex, increasing function of a submartingale.

**Proposition 9.10.** If  $X$  is a martingale and  $g$  is a convex function such that  $E[|g(X_n)|] < \infty$  for each  $n$ , then  $g(X) = (g(X_n))$  is a submartingale. If  $X$  is a submartingale and  $g$  is increasing and convex with  $E[|g(X_n)|] < \infty$  for each  $n$ , then  $g(X)$  is a submartingale.

**Proof:** These are consequences of Jensen's inequality for conditional expectations (Theorem 8.32). For the martingale case,

$$E[g(X_{n+1})|Y_0, \dots, Y_n] \geq g(E[X_{n+1}|Y_0, \dots, Y_n]) = g(X_n)$$

When  $X$  is a submartingale, then by monotonicity of  $g$ ,

$$E[g(X_{n+1})|Y_0, \dots, Y_n] \geq g(E[X_{n+1}|Y_0, \dots, Y_n]) \geq g(X_n). \quad ■$$

In particular, if  $X$  is a martingale and if  $E[X_n^2] < \infty$  for each  $n$ , then  $X^2$  is a submartingale.

## 9.2 Stopping Times

A *random time* is a random variable taking values in the set  $\{0, 1, \dots, \infty\}$ . Stopping times are random times whose occurrence can be determined without prescient knowledge of the future. The term arose in gambling, since the rule whereby a gambler ceases to play must be a stopping time.

**Definition 9.11.** A *stopping time* of  $Y$  is a random variable  $T$  taking values in  $\bar{\mathbb{N}} = \{0, 1, \dots, \infty\}$  such that for each  $k$ , the event  $\{T \leq k\}$  belongs to  $\sigma(Y_0, \dots, Y_k)$ .  $\square$

That is,  $T$  is a stopping time of  $Y$  if and only if for each  $k$  there is a function  $h_k: \mathbb{R}^{k+1} \rightarrow \{0, 1\}$  such that  $\mathbf{1}(T \leq k) = h_k(Y_0, \dots, Y_k)$ .

Physically, whether a stopping time  $T$  has occurred by a fixed time  $k$  is determined by the observable history  $Y_0, \dots, Y_k$  at time  $k$ , and does not require knowledge of the future.

Evidently, constants are stopping times: for  $T \equiv n$ ,  $\{T \leq k\}$  is  $\emptyset$  if  $k < n$  and  $\Omega$  if  $k \geq n$ , which belongs to  $\sigma(Y_0, \dots, Y_k)$  in either case.

Stopping times are one class of random variables that must be permitted to assume the value infinity. As the examples discussed momentarily confirm, a stopping time is the time at which an event connected with  $Y$  occurs. Should this event not occur, the stopping time is taken to be infinite.

**Example 9.12 (First entry time).** The *first entry time* of  $B \in \mathcal{B}(\mathbb{R})$ ,

$$T_B^Y = \begin{cases} \min \{n: Y_n \in B\} & \text{if } Y_n \in B \text{ for some } n \\ \infty & \text{if } Y_n \notin B \text{ for all } n, \end{cases}$$

is a stopping time of  $Y$ . Indeed, we simply take

$$h_k(y_0, \dots, y_k) = \max_{j=0, \dots, k} \mathbf{1}_B(y_j).$$

More generally, if  $X$  is adapted to  $Y$ , then for  $B \in \mathcal{B}(\mathbb{R})$ ,  $T_B^X$ , the first entry time of  $B$  by  $X$ , is a stopping time not only of  $X$  but also of  $Y$ .  $\square$

The last time at which some event occurs is *not* a stopping time.

**Example 9.13 (Last exit time).** The *last exit time* of  $B \in \mathcal{B}(\mathbb{R})$ ,

$$L_B^Y = \begin{cases} \sup \{n: Y_n \in B\} & \text{if } Y_n \in B \text{ for some } n \\ 0 & \text{if } Y_n \notin B \text{ for all } n, \end{cases}$$

is *not* a stopping time of  $Y$  (in general), since it is impossible to know that  $Y$  will not return to  $B$  after some time  $k$  without knowing the entire future  $Y_{k+1}, Y_{k+2}, \dots$ .  $\square$

The minimum, maximum and sum of stopping times are stopping times.

**Proposition 9.14.** If  $T_1$  and  $T_2$  are stopping times of  $Y$ , then so are  $T_1 \wedge T_2 = \min \{T_1, T_2\}$ ,  $T_1 \vee T_2 = \max \{T_1, T_2\}$  and  $T_1 + T_2$ .  $\square$

Given a sequence  $X$  adapted to  $Y$  and a finite stopping time  $T$  of  $Y$ , the random variable

$$X_T \stackrel{\text{def}}{=} \sum_{k=0}^{\infty} X_k \mathbf{1}(T = k)$$

is the state of  $X$  at the random time  $T$  in precisely the same way that  $X_n$  is the state at the deterministic time  $n$ .

### 9.3 Optional Sampling Theorems

Optional sampling theorems extend the martingale property to stopping times. We begin with a useful characterization of martingales that involves stopping times but not conditional expectations.

**Proposition 9.15 (Komatsu's lemma).** Let  $X$  be an adapted, integrable sequence such that for every bounded stopping time  $T$  of  $Y$ ,

$$E[X_T] = E[X_0]. \quad (9.9)$$

Then,  $X$  is a martingale with respect to  $Y$ .

**Proof:** Let  $n$  and  $A \in \sigma(Y_0, \dots, Y_n)$  be fixed, and let

$$T(\omega) = \begin{cases} n & \text{if } \omega \in A \\ n+1 & \text{otherwise,} \end{cases}$$

which defines a stopping time of  $Y$ , since

$$\{T \leq k\} = \begin{cases} \emptyset & \text{if } k < n \\ A & \text{if } k = n \\ \Omega & \text{if } k \geq n+1. \end{cases}$$

In the first and third cases,  $\{T \leq k\}$  clearly belongs to  $\sigma(Y_0, \dots, Y_k)$ , while in the second, this holds true since  $A \in \sigma(Y_0, \dots, Y_n)$ .

We then apply (9.9) successively to  $T$  and to the bounded stopping time  $T' \equiv n+1$ :

$$\begin{aligned} E[X_n; A] + E[X_{n+1}; A^c] &= E[X_T] \\ &= E[X_0] \\ &= E[X_{n+1}] \\ &= E[X_{n+1}; A] + E[X_{n+1}; A^c]. \end{aligned}$$

Therefore,  $E[X_{n+1}; A] = E[X_n; A]$ , so that  $X$  is a martingale. ■

### 9.3.1 Optional sampling theorems for martingales

We now consider converses to Proposition 9.15, identifying conditions on a martingale  $X$  and stopping time  $T$  implying that  $E[X_T] = E[X_0]$ . These results, known as *optional sampling theorems*, demonstrate that a fair game cannot be made unfair by sampling it at stopping times. We present four theorems, which increasingly shift the burden from the stopping time to the martingale.

For bounded stopping times, no restriction need be imposed on the martingale.

**Theorem 9.16.** Let  $X$  be a martingale with respect to  $Y$  and let  $T$  be a *bounded* stopping time of  $Y$ . Then,  $E[X_T] = E[X_0]$ .

**Proof:** This is merely a computation: if  $P\{T \leq n\} = 1$ , then

$$\begin{aligned} E[X_T] &= E\left[\sum_{k=0}^n X_k \mathbf{1}(T = k)\right] \\ &= \sum_{k=0}^n E\left[E[X_n | Y_0, \dots, Y_k]; \{T = k\}\right] \end{aligned}$$

[by (9.4), since  $X_k = E[X_n | Y_0, \dots, Y_k]$ ]

$$\begin{aligned} &= \sum_{k=0}^n E[X_n; \{T = k\}] \\ &= E[X_n], \end{aligned}$$

which equals  $E[X_0]$  by (9.7). ■

The next result is less important in its own right than as a progenitor of other, more useful versions.

**Theorem 9.17.** Let  $X$  be a martingale and let  $T$  be a *finite* stopping time such that

$$E[|X_T|] < \infty \tag{9.10}$$

and

$$\lim_{k \rightarrow \infty} E[X_k; \{T > k\}] = 0. \tag{9.11}$$

Then,  $E[X_T] = E[X_0]$ .

**Proof:** For each  $k$ ,  $T \wedge k = \min\{T, k\}$  is a bounded stopping time, so by Theorem 9.16,  $E[X_0] = E[X_{T \wedge k}]$  for each  $k$ . Hence,

$$\begin{aligned} E[X_0] &= \lim_{k \rightarrow \infty} E[X_{T \wedge k}] \\ &= \lim_{k \rightarrow \infty} E[X_T; \{T \leq k\}] + \lim_{k \rightarrow \infty} E[X_k; \{T > k\}]. \end{aligned}$$

Here, the second term converges to zero by (9.11), while  $E[X_T; \{T \leq k\}] \rightarrow E[X_T]$  by the dominated convergence theorem, which applies by (9.10). ■

The next theorem is a consequence of Theorem 9.17, but its conditions are satisfied sufficiently often to merit its being set apart.

**Theorem 9.18.** Let  $X$  be a martingale for which there is a constant  $c < \infty$  such that

$$E[|X_{n+1} - X_n| | Y_0, \dots, Y_n] \leq c \quad (9.12)$$

for each  $n$  (almost surely). Then,  $E[X_T] = E[X_0]$  for every stopping time  $T$  such that  $E[T] < \infty$ .

**Proof:** We verify that (9.10) and (9.11) are fulfilled, so that this theorem follows from Theorem 9.17.

For (9.10),

$$\begin{aligned} E[|X_T|] &= E\left[\left|\sum_{k=0}^{\infty} (X_{k+1} - X_k) \mathbf{1}(T > k)\right|\right] \\ &\leq \sum_{k=0}^{\infty} E[|X_{k+1} - X_k|; \{T > k\}] \end{aligned}$$

[by the triangle inequality and Fatou's lemma]

$$\begin{aligned} &= \sum_{k=0}^{\infty} E\left[E[|X_{k+1} - X_k| | Y_0, \dots, Y_k]; \{T > k\}\right] \\ &\leq c \sum_{k=0}^{\infty} P\{T > k\} \end{aligned}$$

by (9.12), and this equals  $cE[T]$  by (4.15).

Similarly,

$$\begin{aligned} |E[X_k; \{T > k\}]| &\leq E[|X_k|; \{T > k\}] \leq cE[k; \{T > k\}] \\ &\leq cE[T; \{T > k\}], \end{aligned}$$

and this last expression converges to zero by the dominated convergence theorem, which applies since  $E[T] < \infty$ . ■

Before presenting the final optional sampling theorem, we recall from Definition 5.15 that  $(X_n)$  is *uniformly integrable* if

$$\lim_{a \rightarrow \infty} \sup_n E[|X_n|; \{|X_n| > a\}] = 0.$$

For uniformly integrable martingales, the optional sampling theorem holds for all finite stopping times.

**Theorem 9.19.** Let  $X$  be a martingale such that  $(X_n)$  is uniformly integrable. Then,  $E[X_T] = E[X_0]$  for every finite stopping time  $T$ .

**Proof:** Once more, we verify that (9.10) and (9.11) hold. For each  $k$ ,

$$\begin{aligned} E[|X_T|] &= E[|X_T|; \{T \leq k\}] + E[|X_T|; \{T > k\}] \\ &\leq \sum_{n=0}^k E[|X_n|] + \sup_n E[|X_n|; \{T > k\}]. \end{aligned}$$

The second term converges to zero by uniform integrability, and, hence,  $E[|X_T|]$  is finite. For (9.11),

$$|E[X_k; \{T > k\}]| \leq \sup_n E[|X_n|; \{T > k\}],$$

which converges to zero by uniform integrability. ■

The optional sampling theorems are valid for submartingales as well.

### 9.3.2 Applications of optional sampling theorems

We consider two applications of the optional sampling theorems: the *gambler's ruin problem* and *Wald's identity*.

We begin with the gambler's ruin problem. Let  $Y_1, Y_2, \dots$  be i.i.d. with  $P\{Y_i = \pm 1\} = 1/2$  for each  $i$ , and let  $X_n = \sum_{i=1}^n Y_i$ , with  $X_0 = 0$ . We interpret the  $Y_i$  as outcomes of a series of fair gambles, so that  $X_n$  is the net change in the fortune of the gambler after  $n$  independent gambles. By Example 9.4,  $X$  is a martingale with respect to  $Y$ . Also, of course,  $X$  is a random walk.

Denoting by  $a$  the initial assets of the gambler and by  $b$  those of the opponent, suppose that the game continues until the stopping time

$$T = \min \{n : X_n = b \text{ or } X_n = -a\}$$

at which one player or the other is ruined. We calculate  $P\{X_T = -a\}$ , the probability that the gambler is ruined.

**Proposition 9.20.** We have  $P\{T < \infty\} = 1$ , and

$$P\{X_T = -a\} = 1 - P\{X_T = b\} = \frac{b}{a+b}. \quad (9.13)$$

**Proof:** Finiteness of  $T$  is established most quickly by appeal to the law of the iterated logarithm (Theorem 7.22), which implies that  $\limsup_n X_n \stackrel{\text{a.s.}}{=} \infty$ . Increments in  $X$  are all  $\pm 1$ , and so  $X$  cannot exceed  $b$  without passing through  $b$ , and, hence,  $P\{T < \infty\} \geq P\{T_{\{b\}}^X < \infty\} = 1$ .

We next verify that (9.10) and (9.11) are fulfilled. The former holds because  $X_T \in \{a, b\}$ , so that  $E[|X_T|] \leq \max \{a, b\} < \infty$ , while for the latter, since on  $\{T > k\}$ ,  $X_k \in \{-a + 1, \dots, 0, \dots, b - 1\}$ , it follows that

$$|E[X_k; \{T > k\}]| \leq \max \{a, b\} P\{T > k\},$$

which converges to zero because  $T$  is finite almost surely. Hence Theorem 9.17 applies, with the result that

$$0 = E[X_0] = E[X_T] = -aP\{X_T = -a\} + b(1 - P\{X_T = -a\}),$$

which is easily solved to yield (9.13). ■

We now consider Wald's identity for sums of independent random variables. Let  $Y_1, Y_2, \dots$  be i.i.d. with  $E[|Y_k|] < \infty$ , and let  $S_n = \sum_{i=1}^n Y_i$  be the partial sum process, with  $S_0 = 0$ . Exercise 8.17 shows that if  $N$  is a positive, integer-valued and independent of the  $Y_k$ , then  $E[S_N] = E[N]E[Y_1]$ . The same conclusion obtains if  $N$  is replaced by a finite stopping time.

**Proposition 9.21 (Wald's identity).** If  $T$  is a stopping time of  $Y$  with  $E[T] < \infty$ , then

$$E[S_T] = E[T]E[Y_1].$$

**Proof:** The sequence  $X_n = S_n - nE[Y_1]$  is a mean zero martingale, and satisfies the hypotheses of Theorem 9.18 since

$$E[|X_{n+1} - X_n| | Y_0, \dots, Y_n] = E[|Y_{n+1} - E[Y_1]| | Y_0, \dots, Y_n] \leq 2E[|Y_1|],$$

which is finite. Consequently,

$$0 = E[X_0] = E[X_T] = E[S_T - TE[Y_1]]. \quad \blacksquare$$

Wald's identity is valid for moment generating functions as well.

**Proposition 9.22 (Wald's identity, bis).** Assume that the moment generating function  $\psi(t) = E[e^{tY_1}]$  exists for all  $t \in \mathbb{R}$ . Then, for every stopping time  $T$  of  $Y$  with  $E[T] < \infty$ , and each  $t \in \mathbb{R}$ ,

$$E[\psi(t)^{-T} \prod_{i=1}^T e^{tY_i}] = 1.$$

**Proof:** With  $t$  fixed, let

$$X_n = \psi(t)^{-n} \prod_{i=1}^n e^{tY_i}.$$

Then,  $X$  is a martingale by Example 9.5, and the hypotheses of Theorem 9.18 are satisfied. ■

## 9.4 Martingale Convergence Theorems

In a variety of situations, martingales converge almost surely. Theorems to this effect are known generically as martingale convergence theorems. In particular, a martingale  $X$  converges almost surely to an integrable random variable  $X_\infty$  under each of the following assumptions:

1. **Positivity:**  $X_n \geq 0$  for each  $n$ .
2.  **$L^1$  boundedness:**  $\sup_n E[|X_n|] < \infty$ .
3. **Uniform integrability:**  $(X_n)$  is uniformly integrable.

In the latter case,  $X_n \xrightarrow{L^1} X_\infty$  as well, and the martingale comprises the successive conditional expectations of the limit, as in Example 9.3.

### 9.4.1 Upcrossings and almost sure convergence

Here, we introduce upcrossings, the key device for proving almost sure convergence of submartingales.

Let  $x = (x_n)$  be a sequence of real numbers, and suppose that  $a < b$ . With  $x_0 = 0$ , let

$$t_1 = \inf \{n: x_n \leq a\}$$

be the first time at which  $x$  lies below  $a$ , and let

$$t_2 = \inf \{n > t_1: x_n \geq b\}$$

be the first index *after*  $t_1$  at which  $x$  is above  $b$ . Continuing, we define integers  $t_j$  recursively:

$$\begin{aligned} t_{2k-1} &= \inf \{n > t_{2k-2}: x_n \leq a\} \\ t_{2k} &= \inf \{n > t_{2k-1}: x_n \geq b\}. \end{aligned}$$

In case  $\{n > t_{j-1}: \dots\} = \emptyset$ ,  $t_j$  and all of its successors are set equal to infinity. Physically, the indices  $t_{2k}$  mark *completed upcrossings* of the interval  $(a, b)$  from below  $a$  to above  $b$ . Figure 9.1 illustrates.

We further define

$$U_m^x(a, b) = \sup \{k: t_{2k} \leq m\},$$

the number of *upcrossings* of  $(a, b)$  by the finite sequence  $x_0, \dots, x_m$ , and

$$U^x(a, b) = \lim_{m \rightarrow \infty} U_m^x(a, b) \leq \infty.$$

The limit exists in  $\overline{\mathbb{R}}_+$  since  $U_m^x(a, b)$  is increasing in  $m$ , and is the number of *upcrossings* of  $(a, b)$  by the sequence  $(x_n)$ .

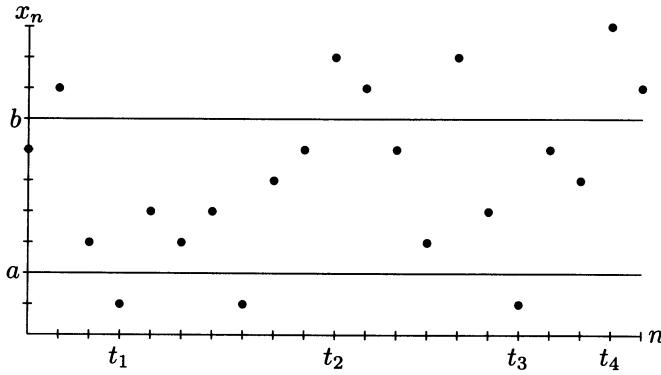


Figure 9.1. Upcrossings for a Sequence  $x_0, \dots, x_{20}$

The key point is that non-convergence of  $(x_n)$  entails that the number of upcrossings of some interval must be infinite. Conversely, a sequence for which the number of upcrossings of every interval is finite converges, although the limit may be infinite.

**Lemma 9.23.** Let  $(x_n)$  be a real sequence such that  $U^x(a, b) < \infty$  for all  $a < b$ . Then,  $(x_n)$  converges in  $\overline{\mathbb{R}}$ .

**Proof:** If  $(x_n)$  does not converge, there are real numbers  $a < b$  such that  $\liminf_n x_n < a < b < \limsup_n x_n$ , which implies that  $U^x(a, b) = \infty$ . ■

We can also define upcrossings for a sequence  $X$  of random variables. In this case the  $t_j$  become stopping times  $T_j$  of  $X$ , and of any sequence  $Y$  to which  $X$  is adapted, and the  $U_m^X(a, b)$  and  $U^X(a, b)$  are random variables, although the latter may assume the value  $\infty$ . Lemma 9.23 has the following analogue, yet another criterion for almost sure convergence.

**Lemma 9.24 (Upcrossing lemma).** If  $E[U^X(a, b)] < \infty$  for all  $a < b$ , then  $(X_n)$  converges almost surely. □

#### 9.4.2 Almost sure convergence of submartingales

The main convergence theorem is for submartingales, and shows that in order to achieve convergence it suffices to curtail the tendency of a submartingale  $X$  to grow, for which it is enough to assume that  $\sup_n E[X_n^+] < \infty$ .

The key technical tools are upcrossings and a switching principle for submartingales: the process constructed by following one submartingale  $Z^{(1)}$  until a stopping time  $T$ , then switching to a second submartingale  $Z^{(2)}$ , is again a submartingale, provided that the jump when the switch occurs be upward.

We first prove the switching principle.

**Lemma 9.25 (Switching principle).** Let  $Z^{(1)}$  and  $Z^{(2)}$  be submartingales with respect to  $Y$  and let  $T$  be a stopping time of  $Y$ , not necessarily finite, such that  $Z_T^{(1)} \leq Z_T^{(2)}$  on the event  $\{T < \infty\}$ . Then, the process

$$Z_n = Z_n^{(1)} \mathbf{1}(T > n) + Z_n^{(2)} \mathbf{1}(T \leq n) = \begin{cases} Z_n^{(1)} & n < T \\ Z_n^{(2)} & n \geq T \end{cases}$$

is a submartingale.

**Proof:** That  $Z$  is adapted is evident from its definition, and  $E[|Z_n|] \leq E[|Z_n^{(1)}|] + E[|Z_n^{(2)}|] < \infty$  for each  $n$ , so only the submartingale property (9.2) requires detailed verification. For each  $n$ ,

$$\begin{aligned} E[Z_{n+1} | Y_0, \dots, Y_n] &= E\left[Z_{n+1}^{(1)} \mathbf{1}(T > n+1) + Z_{n+1}^{(2)} \mathbf{1}(T \leq n+1) \mid Y_0, \dots, Y_n\right] \\ &= E\left[Z_{n+1}^{(1)} \mathbf{1}(T > n+1) + Z_{n+1}^{(2)} \mathbf{1}(T = n+1) + Z_{n+1}^{(2)} \mathbf{1}(T \leq n) \mid Y_0, \dots, Y_n\right] \\ &\geq E\left[Z_{n+1}^{(1)} \mathbf{1}(T > n+1) + Z_{n+1}^{(1)} \mathbf{1}(T = n+1) + Z_{n+1}^{(2)} \mathbf{1}(T \leq n) \mid Y_0, \dots, Y_n\right] \end{aligned}$$

[by the assumption that  $Z_T^{(1)} \leq Z_T^{(2)}$  provided that  $T < \infty$ ]

$$\begin{aligned} &= E\left[Z_{n+1}^{(1)} \mathbf{1}(T > n) + Z_{n+1}^{(2)} \mathbf{1}(T \leq n) \mid Y_0, \dots, Y_n\right] \\ &= \mathbf{1}(T > n)E[Z_{n+1}^{(1)} \mid Y_0, \dots, Y_n] + \mathbf{1}(T \leq n)E[Z_{n+1}^{(2)} \mid Y_0, \dots, Y_n] \end{aligned}$$

[since the events  $\{T > n\}$  and  $\{T \leq n\}$  belong to  $\sigma(Y_0, \dots, Y_n)$ ]

$$\begin{aligned} &\geq \mathbf{1}(T > n)Z_n^{(1)} + \mathbf{1}(T \leq n)Z_n^{(2)} \\ &= Z_n, \end{aligned}$$

where the last inequality holds because the processes  $Z^{(1)}$  and  $Z^{(2)}$  are submartingales. ■

We now prove the convergence theorem for submartingales, by showing that  $E[U^X(a, b)] < \infty$  for all  $a$  and  $b$ .

**Theorem 9.26 (Submartingale convergence theorem).** Let  $X$  be a submartingale with respect to  $Y$ . If

$$\sup_n E[X_n^+] < \infty, \tag{9.14}$$

then there is  $X_\infty \in L^1$  such that  $X_n \xrightarrow{\text{a.s.}} X_\infty$ .

**Proof:** The key point in the proof is the *upcrossing inequality*: for  $a < b$  and for each  $n$ ,

$$E[U_n^X(a, b)] \leq \frac{E[(X_n - a)^+] - E[(X_0 - a)^+]}{b - a}. \quad (9.15)$$

Given this, (9.14) and the monotone convergence theorem imply that

$$E[U^X(a, b)] \leq \frac{\sup_n E[(X_n - a)^+] - E[(X_0 - a)^+]}{b - a} < \infty,$$

so that existence of  $X_\infty$  is a consequence of Lemma 9.24.

To establish (9.15), we apply Lemma 9.25 repeatedly. With  $a$  and  $b$  fixed, consider the sequence  $(Z_n)$  defined by

$$Z_n = \begin{cases} (X_n - a) - (b - a)U_n^X(a, b) & \text{if } T_{2k} \leq n < T_{2k+1} \text{ for some } k \\ -(b - a)U_n^X(a, b) & \text{otherwise.} \end{cases}$$

We claim that  $Z$  is a submartingale with respect to  $Y$ . Indeed, if we define, for each  $k$ , processes

$$\begin{aligned} Z_n^{(k)} &= (X_n - a) - k(b - a), & n \in \mathbb{N} \\ \tilde{Z}_n^{(k)} &= -k(b - a), & n \in \mathbb{N}, \end{aligned}$$

then each of these is a submartingale, and  $Z$  is constructed by switching, in the manner of Lemma 9.25, from  $Z^{(0)}$  to  $\tilde{Z}^{(0)}$  at  $T_1$ , from  $\tilde{Z}^{(0)}$  to  $Z^{(1)}$  at  $T_2$ , from  $Z^{(1)}$  to  $\tilde{Z}^{(1)}$  at  $T_3$ , and so on. All switches that occur are upward, and, hence,  $Z$  is a submartingale by Lemma 9.25.

Moreover,  $Z_0 = (X_0 - a)^+$ , while for each  $m$ ,

$$Z_m \leq (X_m - a)^+ - (b - a)U_m^X(a, b),$$

and consequently,

$$\begin{aligned} E[(X_0 - a)^+] &= E[Z_0] \leq E[Z_m] \\ &\leq E[(X_m - a)^+] - (b - a)E[U_m^X(a, b)]. \end{aligned}$$

Hence, (9.15) holds, and so  $X_\infty$  exists by Lemma 9.24.

Finally,  $X_\infty \in L^1$  by Fatou's lemma. ■

### 9.4.3 Almost sure convergence of martingales

Theorem 9.26 leads to two important criteria for almost sure convergence of martingales.

**Theorem 9.27 (Martingale convergence theorem).** Let  $X$  be a martingale with respect to  $Y$  such that either

i)  $X_n \geq 0$  for each  $n$ , or

ii)  $X$  is  $L^1$ -bounded:

$$\sup_n E[|X_n|] < \infty. \quad (9.16)$$

Then, there is  $X_\infty \in L^1$  such that  $X_n \xrightarrow{\text{a.s.}} X_\infty$ .

**Proof:** In the first instance,  $-X$  is a submartingale satisfying the assumptions of Theorem 9.26, while if (9.16) holds, then  $X$  itself fulfills the hypotheses of the same theorem, since a martingale is a submartingale. ■

#### 9.4.4 Uniformly integrable martingales

The final convergence theorem shows that a uniformly integrable martingale not only converges almost surely and in  $L^1$ , but also has the “successive conditional expectations” form from Example 9.3.

**Theorem 9.28 (Martingale convergence theorem, bis).** Let  $X$  be a martingale with respect to  $Y$ . Then, the following are equivalent:

- a)  $(X_n)$  is uniformly integrable.
- b) There is  $X_\infty$  such that  $X_n \xrightarrow{\text{a.s.}} X_\infty$  and  $X_n \xrightarrow{L^1} X_\infty$ .
- c) There is  $X_\infty \in L^1$  such that for each  $n \in \mathbb{N}$ ,

$$X_n = E[X_\infty | Y_0, \dots, Y_n]. \quad (9.17)$$

**Proof:** a)  $\Rightarrow$  b): By Proposition 5.16, uniform integrability entails (9.16), and, hence, there is  $X_\infty \in L^1$  such that  $X_n \xrightarrow{\text{a.s.}} X_\infty$ . But then,  $X_n \xrightarrow{L^1} X_\infty$  by Theorem 5.17.

b)  $\Rightarrow$  c): Fix  $n$  and suppose that  $A \in \sigma(Y_0, \dots, Y_n)$ . Then, for each  $m$ ,

$$E[X_n; A] = E[E[X_{n+m} | Y_0, \dots, Y_n]; A] = E[X_{n+m}; A],$$

and taking limits as  $m \rightarrow \infty$  gives  $E[X_n; A] = E[X_\infty; A]$ . The interchange of limit and expectation is by the dominated convergence theorem, since  $X_{n+m} \xrightarrow{L^1} X_\infty$  by b). Hence, (9.17) holds.

c)  $\Rightarrow$  a): We need to show that the sequence  $(E[X_\infty | Y_0, \dots, Y_n])_{n \geq 0}$  is uniformly integrable, which we do by verifying Definition 5.15 and applying Proposition 5.16. For  $a > 0$ ,

$$\begin{aligned} & E\left[\left|E[X_\infty | Y_0, \dots, Y_n]\right| ; \{|E[X_\infty | Y_0, \dots, Y_n]| > a\}\right] \\ & \leq E\left[E[|X_\infty| | Y_0, \dots, Y_n] ; \{|E[X_\infty | Y_0, \dots, Y_n]| > a\}\right] \\ & \leq E[|X_\infty| ; \{|E[X_\infty | Y_0, \dots, Y_n]| > a\}] \end{aligned}$$

since  $\{|E[X_\infty|Y_0, \dots, Y_n]| > a\} \in \sigma(Y_0, \dots, Y_n)$ . Because  $X_\infty \in L^1$ , in view of Proposition 5.16, it suffices to show that

$$\sup_n P \{ |E[X_\infty|Y_0, \dots, Y_n]| > a \} \rightarrow 0$$

as  $a \rightarrow \infty$ . By Chebyshev's inequality,

$$\begin{aligned} P \{ |E[X_\infty|Y_0, \dots, Y_n]| > a \} &\leq P \{ E[|X_\infty| | Y_0, \dots, Y_n] > a \} \\ &\leq \frac{1}{a} E[E[|X_\infty| | Y_0, \dots, Y_n]] \\ &= \frac{1}{a} E[|X_\infty|], \end{aligned}$$

uniformly in  $n$ , and  $E[|X_\infty|]/a$  does converge to zero. ■

## 9.5 Applications of Convergence Theorems

Probability theory and real analysis are replete with important applications of the martingale convergence theorem.

### 9.5.1 The Radon-Nikodym theorem

Let  $(\Omega, \mathcal{F}, P)$  be a probability space, and let  $Z$  be a positive random variable with  $E[Z] = 1$ . The set function  $P'(A) = E_P[Z; A]$  is a probability on  $(\Omega, \mathcal{F})$ , and, moreover, if  $P(A) = 0$ , then  $P'(A) = 0$  as well, a relationship to which we assign a special name.

**Definition 9.29.** A probability  $P'$  on  $(\Omega, \mathcal{F})$  is *absolutely continuous with respect to  $P$*  if  $P'(A) = 0$  for all events  $A$  such that  $P(A) = 0$ . We denote this by  $P' \ll P$ . □

The following characterization clarifies the “continuity” aspect of absolute continuity. See also Proposition 5.16, which connects the property with uniform integrability.

**Proposition 9.30.** We have  $P' \ll P$  if and only if for each  $\varepsilon > 0$  there is  $\delta > 0$  such that  $P'(A) < \varepsilon$  for all events  $A$  with  $P(A) < \delta$ . □

The Radon-Nikodym theorem, one of the most celebrated results in measure theory, shows that *every* probability  $P' \ll P$  has the form  $P'(A) = E_P[Z; A]$  for some positive random variable  $Z$ . Here, we give a martingale proof for a case that, albeit special, often obtains in practice.

**Definition 9.31.** The  $\sigma$ -algebra  $\mathcal{F}$  is *countably generated* if there is a sequence  $(A_n)$  of events such that  $\mathcal{F} = \sigma(A_1, A_2, \dots)$ . □

Any  $\sigma$ -algebra generated by a countable partition in the manner of Example 1.14 is countably generated, as is the Borel  $\sigma$ -algebra  $\mathcal{B}(\mathbb{R})$ , which is generated by the countable family of intervals with rational endpoints, albeit not by a countable partition of  $\mathbb{R}$ .

**Theorem 9.32 (Radon-Nikodym theorem, bis).** Assume that  $\mathcal{F}$  is countably generated, and let  $P'$  be a probability on  $(\Omega, \mathcal{F})$  with  $P' \ll P$ . Then, there is a positive random variable  $Z$  such that

$$P'(A) = E_P[Z; A], \quad A \in \mathcal{F}. \quad (9.18)$$

**Proof:** With  $(A_k)$  generating  $\mathcal{F}$ , for each  $n$ , let  $\{A_{n1}, \dots, A_{nk_n}\}$  be the finite partition of  $\Omega$  induced by  $A_1, \dots, A_n$ , and define

$$X_n = \sum_{i=1}^{k_n} \frac{P'(A_{ni})}{P(A_{ni})} \mathbf{1}_{A_{ni}},$$

where  $0/0 = 0$  by convention. Absolute continuity of  $P'$  with respect to  $P$  ensures that the denominator cannot be zero unless the numerator is zero as well. Thus,  $X_n$  is a bounded random variable for each  $n$ .

Moreover,  $X$  is a martingale (with respect to itself) over  $(\Omega, \mathcal{F}, P)$ . Indeed, for each  $n$ ,

$$E_P[X_n] = \sum_{i=1}^{k_n} \frac{P'(A_{ni})}{P(A_{ni})} P(A_{ni}) = 1.$$

Since the partitions  $\{A_{ni}\}$  are successive refinements of one another, to show that  $E_P[X_{n+1}|X_0, \dots, X_n] = X_n$ , it suffices to show that

$$E_P[X_{n+1}; A_{ni}] = E_P[X_n; A_{ni}] = P(A_{ni}), \quad i = 1, \dots, k_n,$$

but this also is computational:

$$\begin{aligned} E_P[X_{n+1}; A_{ni}] &= \sum_{j=1}^{k_{n+1}} \frac{P'(A_{n+1,j})}{P(A_{n+1,j})} P(A_{n+1,j} \cap A_{ni}) \\ &= \sum_{j: A_{n+1,j} \subseteq A_{ni}} P'(A_{n+1,j}) \\ &= P'(A_{ni}). \end{aligned}$$

Since  $X$  is positive, Theorem 9.27 implies that there is  $X_\infty$  such that  $X_n \xrightarrow{\text{a.s.}} X_\infty$ .

We now show that  $(X_n)$  is  $P$ -uniformly integrable. To this end, we first note that Chebyshev's inequality implies that

$$P\{X_n > a\} \leq E_P[X_n]/a = 1/a$$

for each  $n$  and  $a$ . Next, we appeal to Proposition 9.30:  $P' \ll P$  implies that given  $\varepsilon > 0$  there is  $\delta > 0$  such that  $P'(A) < \varepsilon$  for all  $A$  such that  $P(A) < \delta$ . In particular, for  $a$  sufficiently large that  $1/a < \delta$ ,

$$\sup_n E_P[X_n; \{X_n > a\}] \leq \sup_n P' \{X_n > a\} < \varepsilon,$$

as required. By Theorem 9.28,

$$X_n = E_P[X_\infty | X_0, \dots, X_n]$$

for each  $n$ . In particular, for each  $n$  and  $i$ ,

$$E_P[X_n; A_{ni}] = E_P[E_P[X_\infty | X_0, \dots, X_n]; A_{ni}] = E_P[X_\infty; A_{ni}],$$

so that  $Z = X_\infty$  satisfies (9.18) by Theorem 1.17. ■

Note how “physical” this proof is. The  $X_n$  are difference quotients of  $P'$  with respect to  $P$  over increasingly fine partitions, so that we can interpret the random variable  $Z$  in (9.18), known as the *Radon-Nikodym derivative of  $P'$  with respect to  $P$* , as the limit of difference quotients:

$$Z(\omega) = \lim_{A \downarrow \{\omega\}} \frac{P'(A)}{P(A)}.$$

### 9.5.2 Zero-one laws

Let  $Y_1, Y_2, \dots$  be independent random variables. It is no coincidence that many events having to do with asymptotic behavior of the sequence  $(Y_n)$  have probability one or zero. In fact, all events not depending on any finite number of the  $Y_i$  have this property.

The tail  $\sigma$ -algebra consists of those events that are expressible in terms of  $Y_{n+1}, Y_{n+2}, \dots$  for every  $n$ , and, hence, describes various aspects of the asymptotic behavior of  $(Y_n)$ .

**Definition 9.33.** The *tail  $\sigma$ -algebra* of  $Y$  is

$$\mathcal{T} = \bigcap_{n=1}^{\infty} \sigma(Y_{n+1}, Y_{n+2}, \dots).$$

Events belonging to  $\mathcal{T}$  are *tail events* of  $Y$ . □

If  $B_n \in \mathcal{B}(\mathbb{R})$  for each  $n$ , then  $\{Y_n \in B_n, \text{ i.o.}\}$  is a tail event. The event  $\{\sum_n Y_n \text{ converges}\}$  belongs to  $\mathcal{T}$ . The random variables  $\liminf_n Y_n$ ,  $\limsup_n Y_n$  and, provided it exists,  $\lim_n Y_n$  are measurable with respect to  $\mathcal{T}$ . If it exists,  $\lim_{n \rightarrow \infty} (1/n) \sum_{i=1}^n Y_i$  is measurable with respect to  $\mathcal{T}$ .

**Theorem 9.34 (Kolmogorov zero-one law).** Let  $Y_1, Y_2, \dots$  be independent with tail  $\sigma$ -algebra  $\mathcal{T}$ . Then,  $P(A)$  is zero or one for every  $A \in \mathcal{T}$ .

**Proof:** Suppose that  $A$  is tail event. Then, for each  $n$ ,

$$P(A)^2 = E[\mathbf{1}_A]E[\mathbf{1}_A|Y_0, \dots, Y_n] = E[\mathbf{1}_A E[\mathbf{1}_A|Y_0, \dots, Y_n]]$$

because the random variables  $\mathbf{1}_A$  (which, since  $A$  is a tail event, depends on  $Y_{n+1}, Y_{n+2}, \dots$ ) and  $E[\mathbf{1}_A|Y_0, \dots, Y_n]$  (which is a function of  $Y_0, \dots, Y_n$ ) are independent by the disjoint blocks theorem. By Theorem 9.28,

$$E[\mathbf{1}_A|Y_0, \dots, Y_n] \xrightarrow{\text{L}^1} E[\mathbf{1}_A|Y_1, Y_2, \dots] = \mathbf{1}_A,$$

and therefore,

$$E[\mathbf{1}_A E[\mathbf{1}_A|Y_0, \dots, Y_n]] \rightarrow E[\mathbf{1}_A \mathbf{1}_A] = P(A).$$

That is, for  $A \in \mathcal{T}$ ,  $P(A) = P(A)^2$ , so  $P(A)$  must be either zero or one. ■

A broader concept is that of symmetric events, which are invariant under any re-labeling of finitely many of the  $Y_k$ .

**Definition 9.35.** An event  $A \in \sigma(Y_1, Y_2, \dots)$  with  $\mathbf{1}_A = h(Y_1, Y_2, \dots)$  for some function  $h: \mathbb{R}^N \rightarrow \{0, 1\}$  is *symmetric* if for every  $n$  and every permutation  $\sigma$  of  $\{1, \dots, n\}$ ,  $\mathbf{1}_A = h(Y_{\sigma(1)}, Y_{\sigma(2)}, \dots, Y_{\sigma(n)}, Y_{n+1}, \dots)$ . The *symmetric*  $\sigma$ -algebra associated with  $(Y_n)$  is the  $\sigma$ -algebra  $\mathcal{S}$  consisting of all symmetric events. □

That  $\mathcal{S}$  actually is a  $\sigma$ -algebra is straightforward to verify, as is the property that it contains the tail  $\sigma$ -algebra  $\mathcal{T}$ : if  $A$  is a tail event and  $n$  is fixed, then, since  $A \in \sigma(Y_{n+1}, Y_{n+2}, \dots)$ ,  $\mathbf{1}_A$  does not depend on any permutation of the indices of  $Y_1, \dots, Y_n$ . The inclusion is strict, moreover. For example, assuming that the series converges almost surely,  $\sum_n Y_n$  is measurable with respect to  $\mathcal{S}$ , but not  $\mathcal{T}$ .

When  $Y_1, Y_2, \dots$  are i.i.d. (note the addition of the “identically distributed” hypothesis), the symmetric  $\sigma$ -algebra contains only events whose probability is zero or one, as the following companion of Theorem 9.34 demonstrates.

**Theorem 9.36 (Hewitt-Savage zero-one law).** Let  $Y_1, Y_2, \dots$  be i.i.d. with symmetric  $\sigma$ -algebra  $\mathcal{S}$ . Then,  $P(A)$  is zero or one for every  $A \in \mathcal{S}$ . □

### 9.5.3 Likelihood ratios

As in Example 9.6, let  $f$  and  $g$  be distinct density functions on  $\mathbb{R}$  such that  $f(x) > 0$  and  $g(x) > 0$  for all  $x$ , and let  $Y_1, Y_2, \dots$  be i.i.d. Let  $P_f$  be a probability under which the  $Y_k$  have density  $f$  and let  $P_g$  be one under which they have density  $g$ . It was shown there that the likelihood ratio process  $X_n = \prod_{k=1}^n [g(Y_k)/f(Y_k)]$ , where  $X_0 = 1$ , is a martingale with respect to  $Y$  if the  $Y_k$  have density  $f$  and a submartingale with respect to  $Y$  if the  $Y_k$  have density  $g$ . Here, we show that in the former case,  $X_n \xrightarrow{\text{a.s.}} 0$ , while in the latter,  $X_n \xrightarrow{\text{a.s.}} \infty$ .

**Theorem 9.37.** Under  $P_f$ ,  $X_n \xrightarrow{\text{a.s.}} 0$ , while under  $P_g$ ,  $X_n \xrightarrow{\text{a.s.}} \infty$ .

**Proof:** Under  $P_f$ ,  $X$  is a positive martingale, and by Theorem 9.27 there is a  $X_\infty \geq 0$  such that  $X_n \xrightarrow{\text{a.s.}} X_\infty$ . For each  $k$ ,

$$\alpha \stackrel{\text{def}}{=} E_f [\sqrt{g(Y_k)/f(Y_k)}] < E_f [g(Y_k)/f(Y_k)]^{1/2} = 1,$$

where the inequality is strict because  $f \neq g$ . But then, by Fatou's lemma,

$$E_f [\sqrt{X_\infty}] \leq \liminf_{n \rightarrow \infty} E_f [\sqrt{X_n}] = \liminf_{n \rightarrow \infty} \alpha^n = 0,$$

which implies that

$$P_f \{X_\infty = 0\} = 1.$$

With respect to  $P_g$ , the sequence  $(1/X_n)$  is a positive martingale converging to zero almost surely by a), and, hence,

$$P_g \{X_n \rightarrow \infty\} = 1. \blacksquare$$

## 9.6 Complements

### 9.6.1 Conditioning on $Y_0, \dots, Y_T$

Let  $T$  be a stopping time of  $Y$ . For each fixed  $n$ , we can think of  $Y_0, \dots, Y_n$  as the history (past and present) of  $Y$  at time  $n$ , so that for  $X \in L^2$ , conditional expectations  $E[X|Y_0, \dots, Y_n]$  are MMSE predictors of  $X$  given this history. We can also condition on the history of  $Y$  at  $T$ .

A primary motivation is to extend the martingale property, as expressed in (9.4), to stopping times  $S \leq T$ , leading to relationships

$$E[X_T|Y_0, \dots, Y_S] = X_S,$$

which generalize the definition of a martingale.

There are two issues: definition of events prior to  $T$  and of conditional expectations given  $Y_0, \dots, Y_T$ .

**Definition 9.38.** An event  $A$  is *prior to  $T$*  if for each  $k$ ,

$$A \cap \{T \leq k\} \in \sigma(Y_0, \dots, Y_k). \quad \square$$

That is,  $A$  is prior to  $T$  if for each  $k$ , the simultaneous occurrence of  $A$  and the event  $\{T \leq k\}$  is part of the observable history at  $k$ .

The family

$$\sigma(Y_0, \dots, Y_T) \stackrel{\text{def}}{=} \{A: A \text{ is prior to } T\}$$

is a  $\sigma$ -algebra, with respect to which  $T$  is measurable, with respect to which  $X_T$  is measurable for any process  $X$  adapted to  $Y$ , and such that if  $T$  and  $S$  are stopping times with  $S \leq T$ , then  $\sigma(Y_0, \dots, Y_S) \subseteq \sigma(Y_0, \dots, Y_T)$ . Finally, if  $T \equiv n$ , then  $\sigma(Y_0, \dots, Y_T) = \sigma(Y_0, \dots, Y_n)$ .

We define conditional expectations  $E[X|Y_0, \dots, Y_T]$  in a “physical” and computationally useful way, making use of the property that on the event  $\{T = n\}$ ,  $\sigma(Y_0, \dots, Y_T)$ , the past at  $T$ , is  $\sigma(Y_0, \dots, Y_n)$ , the past at  $n$ .

**Definition 9.39.** For  $X$  either positive or integrable, the *conditional expectation of  $X$  given  $Y_0, \dots, Y_T$*  is the random variable

$$E[X|Y_0, \dots, Y_T] = \sum_{n=0}^{\infty} E[X|Y_0, \dots, Y_n] \mathbf{1}(T = n). \quad \square$$

This conditional expectation has the right properties. It is measurable with respect to  $\sigma(Y_0, \dots, Y_T)$  by construction, while for  $A \in \sigma(Y_0, \dots, Y_T)$ ,

$$\begin{aligned} E[X; A] &= \sum_{k=0}^{\infty} E[X; A \cap \{T = k\}] \\ &= \sum_{k=0}^{\infty} E[E[X|Y_0, \dots, Y_k]; A \cap \{T = k\}] \end{aligned}$$

[since by definition,  $A \cap \{T = k\} \in \sigma(Y_0, \dots, Y_k)\]$

$$\begin{aligned} &= E[\sum_{k=0}^{\infty} E[X|Y_0, \dots, Y_k] \mathbf{1}(T = k); A] \\ &= E[E[X|Y_0, \dots, Y_T]; A]. \end{aligned}$$

The other properties of conditional expectations as functions of  $X$  carry over, and even the two “peculiar properties” hold.

**Theorem 9.40.** Let  $S$  and  $T$  be stopping times of  $Y$  such that  $S \leq T$ , and let  $X$  be either positive or integrable. Then,

$$E[E[X|Y_0, \dots, Y_T]|Y_0, \dots, Y_S] = E[X|Y_0, \dots, Y_S]. \quad \square$$

**Theorem 9.41.** Let  $T$  be a stopping time of  $Y$  and suppose that either  $X$  is positive and  $Z$  is positive and measurable with respect to  $\sigma(Y_0, \dots, Y_T)$ , or that  $X$  is integrable and  $Z$  is bounded and measurable with respect to  $\sigma(Y_0, \dots, Y_T)$ . Then,

$$E[XZ|Y_0, \dots, Y_T] = ZE[X|Y_0, \dots, Y_T]. \quad \square$$

The optional sampling theorems in §3 extend in the following manner. For simplicity, we consider only Theorem 9.17, the key theorem of the four.

**Theorem 9.42.** Let  $X$  be a martingale with respect to  $Y$  and let  $S$  and  $T$  be stopping times of  $Y$  such that  $S \leq T \leq n$  for some constant  $n$ . Then,

$$E[X_T|Y_0, \dots, Y_S] = X_S.$$

**Proof:** Suppose that  $A \in \sigma(Y_0, \dots, Y_S)$ ; then

$$\begin{aligned} E[X_T; A] &= \sum_{k=0}^n E[X_T; A \cap \{S = k\}] \\ &= \sum_{k=0}^n \sum_{j=k}^n E[X_j; A \cap \{S = k\} \cap \{T = j\}] \end{aligned}$$

[since  $T \geq S$ ]

$$= \sum_{k=0}^n \sum_{j=k}^n E[X_n; A \cap \{S = k\} \cap \{T = j\}]$$

[ $A \cap \{S = k\} \in \sigma(Y_0, \dots, Y_k)$  because  $S$  is a stopping time and by the definition of  $\sigma(Y_0, \dots, Y_S)$ , while  $\{T = j\} \in \sigma(Y_0, \dots, Y_j)$ , and because  $X$  is a martingale]

$$\begin{aligned} &= \sum_{k=0}^n E[X_n; A \cap \{S = k\}] \\ &= \sum_{k=0}^n E[X_k; A \cap \{S = k\}] \end{aligned}$$

[since  $A \cap \{S = k\} \in \sigma(Y_0, \dots, Y_k)$  and  $X_k = E[X_n|Y_0, \dots, Y_k]$ ]

$$= E[X_S; A]. \quad \blacksquare$$

### 9.6.2 Martingales with respect to filtrations

Conditional expectations with respect to  $\sigma$ -algebras lead to broader classes of martingales, supermartingales and submartingales.

First of all, the base sequence  $(Y_n)$  is replaced by an increasing family of  $\sigma$ -algebras.

**Definition 9.43.** A *filtration* on  $(\Omega, \mathcal{F})$  is a family  $\mathcal{H} = (\mathcal{H}_n)_{n \geq 0}$  of sub- $\sigma$ -algebras of  $\mathcal{F}$  such that  $\mathcal{H}_n \subseteq \mathcal{H}_{n+1}$  for each  $n$ .  $\square$

One interprets  $\mathcal{H}_n$  as the observed history (past and present) at time  $n$ . In particular, a sequence  $Z$  is *adapted to  $\mathcal{H}$*  if for each  $n$ ,  $Z_n$  is measurable with respect to  $\mathcal{H}_n$ .

The definitions from §1 generalize in the following manner.

**Definition 9.44.** Let  $X$  be adapted to  $\mathcal{H}$  and integrable.

- a)  $X$  is a *martingale with respect to  $\mathcal{H}$*  for each  $n$ ,  $E[X_{n+1} | \mathcal{H}_n] = X_n$ .
- b)  $X$  is a *submartingale with respect to  $\mathcal{H}$*  if for each  $n$ ,  $E[X_{n+1} | \mathcal{H}_n] \geq X_n$ .
- c)  $X$  is a *supermartingale with respect to  $\mathcal{H}$*  if for each  $n$ ,  $E[X_{n+1} | \mathcal{H}_n] \leq X_n$ .  $\square$

One can also define a stopping time of a filtration.

**Definition 9.45.** A *stopping time* of  $\mathcal{H}$  is a random variable  $T$  taking values in  $\bar{\mathbb{N}} = \{0, 1, \dots, \infty\}$  such that for each  $k$ ,  $\{T \leq k\} \in \mathcal{H}_k$ .  $\square$

The optional sampling theorems, as well as the submartingale and martingale convergence theorems in §3 and 4, remain valid with essentially only notational emendations.

### 9.6.3 Reversed martingales

In reversed martingales, the conditioning occurs with respect to *decreasing* families of random variables.

**Definition 9.46.** The sequence  $X$  is a *reversed martingale with respect to  $Y$*  if for each  $n$ ,  $X_n$  is measurable with respect to  $\sigma(Y_n, Y_{n+1}, \dots)$ ,  $E[|X_n|] < \infty$ , and

$$E[X_n | Y_{n+1}, Y_{n+2}, \dots] = X_{n+1}. \quad \square$$

Reversed submartingales and reversed supermartingales are defined in the obvious manner.

The convergence theory for reversed martingales is cleaner than for martingales: reversed martingales converge almost surely *and* in  $L^1$ .

**Theorem 9.47 (Reversed martingale convergence theorem).**

If  $X$  is a reversed martingale with respect to  $Y$ , then  $(X_n)$  is uniformly integrable and there is  $X_\infty \in L^1$ , measurable with respect to the tail  $\sigma$ -algebra  $\mathcal{T} = \bigcap_n \sigma(Y_n, Y_{n+1}, \dots)$ , such that  $X_n \xrightarrow{\text{a.s.}} X_\infty$  and  $X_n \xrightarrow{L^1} X_\infty$ .

**Proof:** For each  $n$ ,  $X_n = E[X_1|Y_{n+1}, Y_{n+2}, \dots]$ , and the argument used to prove uniform integrability of  $(E[X_\infty|Y_0, \dots, Y_n])$  in Theorem 9.28 works here.

Reasoning as in Theorem 9.26, for  $a < b$  and each  $n$ ,

$$E[U_n^X(a, b)] \leq \frac{E[(X_n - a)^+]}{b - a},$$

which implies existence of  $X_\infty$  such that  $X_n \xrightarrow{\text{a.s.}} X_\infty$ . Finally,  $X_n \xrightarrow{L^1} X_\infty$  by uniform integrability. ■

As an application, we give a martingale proof of the strong law of large numbers, which shows both almost sure and  $L^1$  convergence.

**Theorem 9.48 (Strong law of large numbers, bis).** Let  $Y_1, Y_2, \dots$  be i.i.d. with  $E[|Y_1|] < \infty$  and partial sums  $S_n = \sum_{i=1}^n Y_i$ . Then,  $S_n/n \rightarrow E[Y_1]$  almost surely and in  $L^1$ .

**Proof:** The key point is that because the  $Y_i$  are i.i.d., for each  $n$ ,

$$S_n/n = E[Y_1|S_n, Y_{n+1}, Y_{n+2}, \dots] = E[Y_1|S_n, S_{n+1}, \dots],$$

so that  $(S_n/n)$  is a reversed martingale. By Theorem 9.47, there is  $X_\infty \in L^1$  and such that  $X_n \xrightarrow{\text{a.s.}} X_\infty$  and  $X_n \xrightarrow{L^1} X_\infty$ .

But  $X_\infty$  is measurable with respect to the tail  $\sigma$ -algebra  $\mathcal{T}$  of  $(Y_k)$ , and by the Kolmogorov zero-one law (Theorem 9.34),  $X_\infty$  must be constant:  $X_\infty \stackrel{\text{a.s.}}{=} E[X_\infty]$ . Finally,  $L^1$  convergence gives  $E[X_\infty] = \lim_n E[S_n/n] = E[Y_1]$ . ■

## 9.7 Exercises

- 9.1. Let  $X$  be a submartingale with respect to  $Y$ . Prove that if  $E[X_n] = E[X_0]$  for all  $n$ , then  $X$  is a martingale.
- 9.2. Prove parts b) and c) of Proposition 9.9.
- 9.3. Verify (9.4), (9.5) and (9.6).

- 9.4.** Show that  $X$  is a martingale with respect to  $Y$  if and only if it is both a submartingale and a supermartingale.
- 9.5.** Describe the limiting behavior of the branching process martingale  $X_n = Y_n/m^n$  in Example 9.8.
- 9.6.** Prove that a random time  $T$  is a stopping time of  $Y$  if and only if  $\{T = k\} \in \sigma(Y_0, \dots, Y_k)$  for each  $k$ .
- 9.7.** Prove Proposition 9.14.
- 9.8.** Prove that if  $X$  is adapted to  $Y$ , then for  $B \in \mathcal{B}(\mathbb{R})$ , the first entry time  $T_B^X$  of  $B$  by  $X$  is a stopping time of  $Y$ .
- 9.9.** Let  $Y_1, Y_2, \dots$  be independent, and suppose that  $E[Y_k] = 0$  and  $0 < s_k^2 = \text{Var}(Y_k) < \infty$  for each  $k$ , and let  $S_n = \sum_{k=1}^n Y_k$  and  $\sigma_n^2 = \sum_{k=1}^n s_k^2$  for each  $k$ . Prove that  $X_n = S_n^2 - \sigma_n^2$  is a martingale.
- 9.10.** Consider a gambler engaged in a fair game modeled by the i.i.d. sequence  $(Y_k)$ , where  $P\{Y_k = 1\} = P\{Y_k = -1\} = 1/2$  for each  $k$ , with  $Y_k$  corresponding to the gambler's winning and  $Y_k = -1$  to losing. Assume that the gambler bets in the following manner: initially one unit is bet; after a loss the stake is doubled, and after a win, one unit is bet. Let  $X_n$  be the gambler's net gain after  $n$  plays of the game. Show that  $(X_n)$  is a martingale.
- 9.11.** Let  $X$  be a martingale with  $E[X_n] = 0$  and  $E[X_n^2] < \infty$  for each  $n$ .
- Prove that for  $k \neq j$ ,  $(X_k - X_{k-1}) \perp (X_j - X_{j-1})$ . Thus, a martingale with  $E[X_n^2] < \infty$  for each  $n$  has *orthogonal increments*.
  - Prove that for each  $n$ ,  $E[X_n^2] = \sum_{k=1}^n E[(X_k - X_{k-1})^2]$ .
- 9.12.** Let  $(Y_n)$  be a sequence for which there exist constants  $\alpha$  and  $\beta$  with
- $$E[Y_{n+1}|Y_0, \dots, Y_n] = \alpha Y_n + \beta Y_{n-1}$$
- for each  $n$ . Show that there exists a real number  $a$  such that  $X_n = aY_n + Y_{n-1}$  is a martingale with respect to  $Y$ .
- 9.13.** Let  $Y_1, Y_2, \dots$  be i.i.d. with  $E[|Y_1|] < \infty$  and partial sums  $S_n = \sum_{i=1}^n Y_i$ , and let  $T$  be a stopping time of  $Y$  with  $E[T] < \infty$ . Prove Wald's identity  $E[S_T] = E[Y_1]E[T]$  by verifying and using the identity  $S_T = \sum_{k=1}^{\infty} Y_k \mathbf{1}(T \geq k)$ .
- 9.14.** Show that the martingale  $X$  associated with the Pólya urn scheme of Example 9.7 is uniformly integrable, and, hence, that the limit  $X_\infty$  exists and  $X_\infty \stackrel{d}{=} U[0, 1]$ .

- 9.15.** Prove that if  $X$  is martingale such that  $\sup_n E[X_n^2] < \infty$  (which implies existence of the almost sure limit  $X_\infty$  by Theorem 9.27), then  $X_n \xrightarrow{\text{q.m.}} X_\infty$ .
- 9.16.** Use Exercise 9.9 to show that in the gambler's ruin problem,  $E[T] = ab$ .
- 9.17.** Let  $X$  be a martingale with respect to  $Y$  and let  $Z = (Z_k)$  be a sequence of random variables such that for each  $n$ ,  $Z_n$  is measurable with respect to  $\sigma(Y_0, \dots, Y_{n-1})$ . That is, the value of  $Z$  at  $n$  is determined by the *strict past*  $Y_0, \dots, Y_{n-1}$ ; such a sequence is termed *predictable* with respect to  $Y$ . Show that the process
- $$W_n = \sum_{k=1}^n Z_k [X_k - X_{k-1}]$$
- (with  $W_0 = 0$ ) is a martingale with respect to  $Y$ .
- 9.18.** Let  $X$  be a submartingale with respect to  $Y$ . Prove that for each  $n$  and each  $a > 0$ ,
- $$P\{\max_{k \leq n} X_k \geq a\} \leq \frac{1}{a} E[X_n; \{\max_{k \leq n} X_k > a\}] \leq \frac{E[X_n^+]}{a}.$$
- 9.19.** Let  $(X_n)$  be a martingale such that for some  $p > 1$ ,  $E[|X_n|^p] < \infty$  for all  $n$ . Use Exercise 9.18 to show that for each  $n$ ,
- $$E[\max_{k \leq n} |X_k|] \leq \frac{p}{p-1} E[|X_n|^p]^{1/p}.$$
- 9.20.** Suppose that for each  $n$ ,  $X_n$  represents the assets of an insurance company at the start of year  $n$ , and satisfies  $X_{n+1} = X_n + b - Y_n$ , where  $b$  is a positive constant (the premiums collected each year) and  $Y_n$ , the claims in year  $n$ , has distribution  $N(\mu, \sigma^2)$ , where  $\mu < b$ . Assume that  $Y_1, Y_2, \dots$  are independent, and that  $X_0 \equiv 1$ . The company is ruined if ever  $X_n < 0$ . Show that  $P\{X_n < 0 \text{ for some } n\} \leq e^{-2(b-\mu)/\sigma^2}$ .
- 9.21.** Let  $(X_n)$  be a martingale such that for some  $K$ ,  $E[(X_n - X_{n-1})^2] \leq K$  for all  $n$ . Prove that  $X_n/n \xrightarrow{\text{P}} 0$ .
- 9.22.** Let  $(\Omega, \mathcal{F}, P)$  be a probability space, and let  $Z$  be a positive random variable with  $E[Z] = 1$  and define the probability  $P'(A) = E[Z; A]$ .
- Prove that  $E'[X] = E[XZ]$  for each  $X \geq 0$ .
  - Prove that for each  $X \geq 0$ ,

$$E'[X|Y_1, \dots, Y_n] = \frac{E[XZ|Y_1, \dots, Y_n]}{E[Z|Y_1, \dots, Y_n]}.$$

# Appendix A

## Notation

This index contains only the notation used (relatively speaking) throughout the book. Symbols with highly localized usage are omitted, as are “standard” symbols such as  $\pi$ ,  $e$  and  $i$ .

### GREEK LETTERS

Symbol	Usage	Page
$\zeta_x$	Generating function	177
$\sigma(\cdot)$	$\sigma$ -algebra generated by $(\cdot)$	22
$\sigma^2$	Variance	123
$\varphi_X, \varphi_F$	Characteristic function	163
$\psi_X$	Moment generating function	177
$\Omega$	Sample space	16
$\omega$	Outcome	16

### SCRIPT LETTERS

Symbol	Usage	Page
$\mathcal{B}(S)$	Borel $\sigma$ -algebra on $S$	22
$\mathcal{F}$	$\sigma$ -algebra of events	21
$\mathcal{P}(\Omega)$	Family of all subsets of $\Omega$	21

## ROMAN LETTERS

Symbol	Usage	Page
$B(n, p)$	Binomial distribution	54
$C^2$	Covariance matrix	126
$\text{Corr}(X, Y)$	Correlation	124
$\text{Cov}(X, Y)$	Covariance	124
$E[X]$	Expectation	102, 103, 107
$E[X; A]$	Expectation over an event	103, 107
$E[X Y]$	Conditional expectation	225
$E[X \mathcal{H}]$	Conditional expectation	239
$E(\lambda)$	Exponential distribution	57
$F, F_P, F_X$	Distribution function	29, 52
$F^{-1}$	Inverse of a distribution function	63
$F_{X Y}(x y)$	Conditional distribution function	234
$f, f_P, f_X$	Density function	33, 52
$L^1$	Integrable random variables	107
$L^p$	Random variables with $E[ X ^p] < \infty$	119
$\ell_X$	Laplace transform	176
$\text{MSE}(\cdot)$	Mean squared error of $(\cdot)$	217
$N(0, 1)$	Standard normal distribution	57
$N(\mu, \sigma^2)$	Normal distribution	57
$P$	Probability	23
$P(B A)$	Conditional probability	35
$P\{B Y\}$	Conditional probability	227
$P\{B \mathcal{H}\}$	Conditional probability	239
$P_{X Y}(B y)$	Conditional distribution	233
$P(\lambda)$	Poisson distribution	55
$S_X$	Survivor function	52
$U[a, b]$	Uniform distribution	31
$\text{Var}(X)$	Variance	123

## SYMBOLS

Symbol	Usage	Page
$\mathbf{1}_A, \mathbf{1}(\cdot)$	Indicator function	18
$\stackrel{\text{a.s.}}{=}$	Equality almost surely	52
$\stackrel{d}{=}$	Equality in distribution	52
$\stackrel{\text{a.s.}}{\rightarrow}$	Almost sure convergence	135
$\stackrel{P}{\rightarrow}$	Convergence in probability	135
$\stackrel{\text{q.m.}}{\rightarrow}$	Quadratic mean convergence	136
$\stackrel{L^1}{\rightarrow}$	$L^1$ convergence	136
$\stackrel{d}{\rightarrow}$	Convergence in distribution	136
$F * G$	Convolution	118
$f * g$	Convolution	78
$X^+$	Positive part	48
$X^-$	Negative part	48
$\ X\ _p$	$L^p$ norm	157
$\langle X, Y \rangle$	Inner product	218
$X \perp Y$	Orthogonality	221
$X^{-1}(B)$	Inverse image	43
$\lfloor x \rfloor$	Integer part of $x$	
$\stackrel{\text{def}}{=}$	Equal by definition	

### SETS OF NUMBERS

Symbol	Usage
$\mathbb{C}$	Complex numbers
$\mathbb{N}$	Positive integers 0,1, ...
$\mathbb{Q}$	Rational numbers
$\mathbb{R}$	Real line $(-\infty, \infty)$
$\mathbb{R}_+$	Positive real half-line $[0, \infty)$
$\overline{\mathbb{R}}$	Extended real line $[-\infty, \infty]$
$\overline{\mathbb{R}}_+$	Extended positive real half-line $[0, \infty]$
$\mathbb{Z}$	Integers

## Appendix B

# Named Objects

This is a guide to named objects, such as conditions, inequalities and results, that appear in the book.

Object	Page
Bayes' theorem	36
Berry–Esséen theorem	210
Boole's inequality	26
Borel-Cantelli lemma	27, 81
Cauchy-Schwarz inequality	120
Central limit theorem	152, 154, 174, 191, 194
Chebyshev's inequality	122
Continuity theorem	171, 172
Cramér-Wold device	150, 176
DeMoivre-Laplace central limit theorem	152, 154
Disjoint blocks theorem	76
Dominated convergence theorem	108, 232
Fatou's lemma	105, 231
Fubini's theorem	128
Glivenko-Cantelli theorem	206
Helly's selection theorem	178
Hölder's inequality	120
Jensen's inequality	121, 232

Object	Page
Kolmogorov extension theorem	65
Kolmogorov's inequality	183
Kolmogorov-Smirnov theorem	206
Law of the iterated logarithm	198, 200
Law of total probability	35
Lindeberg condition	192
Lindeberg-Feller central limit theorem	194, 196
Lyapunov condition	191
Lyapunov's inequality	120
Markov's inequality	122
Martingale convergence theorem	258, 259
Minkowski's inequality	121
Monotone class theorem	23, 51
Monotone convergence theorem	105, 232
Optional sampling theorem	251, 252
Poisson limit theorem	155, 174
Radon-Nikodym theorem	261
Slutsky's theorem	146, 147
Strong law of large numbers	151, 188
Submartingale convergence theorem	257
Three series theorem	186
Wald's identity	254
Weak law of large numbers	151, 173
Weierstrass approximation theorem	156
Zero-one law	262, 263

# Bibliography

- Apostol, T. (1974). *Mathematical Analysis*, 2nd ed. Addison–Wesley, Reading, MA.
- Bickel, P. J., and Doksum, K. A. (1977). *Mathematical Statistics*. Holden–Day, San Francisco.
- Billingsley, P. (1990). *Probability and Measure*, 2nd ed. Wiley, New York.
- Breiman, L. (1968). *Probability*. Addison–Wesley, Reading, MA.
- Chow, Y. S., and Teicher, H. (1988). *Probability Theory: Independence, Interchangeability, Martingales*, 2nd ed. Springer–Verlag, New York.
- Chung, K. L. (1974). *A Course in Probability Theory*, 2nd ed. Academic Press, New York.
- Durrett, R. (1991). *Probability: Theory and Examples*. Brooks/Cole, Pacific Grove, CA.
- Feller, W. (1967). *An Introduction to Probability Theory and its Applications*, Volume I, 3rd edition. Wiley, New York.
- Feller, W. (1966). *An Introduction to Probability Theory and its Applications*, Volume II, 3rd edition. Wiley, New York.
- Ibragimov, I. A., and Has'minskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer–Verlag, New York.
- Karr, A. F. (1991). *Point Processes and their Statistical Inference*, 2nd edition. Dekker, New York.
- Lehmann, E. L. (1986). *Testing Statistical Hypotheses*, 2nd ed. Wiley, New York.
- Lehmann, E. L. (1983). *The Theory of Point Estimation*. Wiley, New York.

- Rudin, W. (1975). *Principles of Mathematical Analysis*, 3rd ed. McGraw–Hill, New York.
- Rudin, W. (1974). *Real and Complex Analysis*, 2nd ed. McGraw–Hill, New York.
- Shiryayev, A. N. (1984). *Probability*. Springer–Verlag, New York.
- Woodroffe, M. (1975). *Probability with Applications*. McGraw–Hill, New York.

# Index

- $L^1$  convergence, 136, 140, 141, 144, 145, 149, 158, 160, 162  
 $L^2$ , Completeness, 220  
 $L^2$ , Convergence, 220  
 $L^2$ , Inner product, 218–220  
 $L^2$ , Metric, 219, 220  
 $L^2$ , Norm, 218–220  
 $L^p$  convergence, 157  
 $L^p$  space, 119, 120  
 $\Delta$ -method, 214  
 $\chi^2$  distribution, 58, 59, 97, 125, 179  
 $\pi$ -system, 21  
 $d$ -system, 21  
 $\sigma$ -algebra, 21  
 $\sigma$ -algebra, Generated by random variable(s), 46, 66  
 $\sigma$ -algebra, Generated by sets, 22  
 $\sigma$ -algebras, Independent, 94  
Absolute continuity, 260, 261  
Adapted sequence, 243  
Almost sure convergence, 135, 137, 138, 140, 141, 144, 145, 147–150, 159, 160, 162, 183  
Almost surely, 28  
Arc sine distribution, 14  
Arc sine law, 13  
Arrival counting process, 91  
Backward recurrence time, 98, 209  
Bayes' theorem, 36  
Bernoulli distribution, 54, 56, 125, 164, 215  
Bernoulli process, 88–90, 97, 150, 151, 153, 154, 156, 161  
Bernstein polynomial, 156  
Berry-Esséen theorem, 211  
Binomial distribution, 55, 56, 60, 85, 87, 89, 97, 98, 125, 130, 155, 164, 174  
Birthday problem, 97  
Bonferroni's inequality, 41  
Boole's inequality, 26  
Borel  $\sigma$ -algebra, 22, 23  
Borel measurable function, 66  
Borel set, 22, 23  
Borel-Cantelli lemma, 27, 81  
Bose-Einstein model, 84, 85, 87, 100  
Branching process, 247, 269  
Cantelli's inequality, 132  
Cantor distribution, 39, 212  
Cauchy distribution, 68, 97, 180  
Cauchy-Schwarz inequality, 120, 123, 130  
Central limit theorem, 12, 153, 154, 174, 191, 194, 196, 211  
Characteristic function, 163–172, 175  
Chebyshev's inequality, 122, 123, 133  
Complete convergence, 137  
Complex-valued random variable, 45, 109  
Conditional density function, 235, 238  
Conditional distribution, 233–235  
Conditional distribution function, 234, 236

- Conditional expectation, 225, 226, 228–231, 233, 236, 238, 239, 265, 266
- Conditional probability, 6, 35, 36, 96, 132, 227, 238, 241
- Conditional variance, 240
- Continuity theorem, 171, 172
- Convergence in distribution, 136, 138–142, 146–150, 158, 171, 172, 175–177
- Convergence in probability, 136, 137, 140–142, 144, 145, 147–150, 157, 160
- Convolution, 78, 119
- Correlation, 125
- Countable additivity, 2, 24
- Counting rules, 82
- Coupon collector’s problem, 42
- Covariance, 125
- Covariance matrix, 126
- Cramér-Wold device, 150, 176
- Cumulant generating function, 182
- DeMoivre-Laplace limit theorem, 153, 154
- Density function, 13, 33, 34, 52–54
- Density function, Marginal, 54
- Disjoint blocks theorem, 76
- Disjontification, 26, 40
- Distribution function, 29–32, 38, 52, 53
- Distribution function, Integral with respect to, 110–112
- Distribution function, Inverse, 63, 64
- Distribution functions, Convergence, 136, 178
- Dominated convergence theorem, 108, 110, 232
- Empirical distribution function, 205–207, 215
- Erlang distribution, 58, 92
- Event, 16, 18, 23
- Event, Almost sure, 28
- Event, Null, 28
- Events, Independent, 80, 81, 94–96
- Expectation, 5, 101–109, 113–117, 128, 129
- Expectation, Over an event, 107
- Exponential distribution, 33, 57, 59, 60, 69, 92, 96, 99, 100, 125, 164, 240
- Extended real number system, 36
- Fatou’s lemma, 105, 110, 232
- Fermi-Dirac model, 84, 85
- Filtration, 267
- First entry time, 249
- First passage time, 9, 10, 216
- Fubini’s theorem, 128, 129
- Gambler’s ruin problem, 253, 270
- Gamma distribution, 58, 59, 100, 125, 164
- Generating function, 177
- Geometric distribution, 55, 56, 69, 87, 89, 96, 98, 125, 164
- Givenko-Cantelli theorem, 206
- Hölder’s inequality, 120, 123
- Hazard function, 68
- Helly’s theorem, 178
- Hewitt-Savage zero-one law, 263
- Inclusion-exclusion principle, 41
- Independence, Events, 80, 94
- Independence, Random variables, 71, 94
- Independence,  $\sigma$ -algebras, 94
- Indicator function, 19, 40
- Inverse image, 43
- Jensen’s inequality, 121, 123, 232
- Joint distribution function, 53
- Kolmogorov extension theorem, 65
- Kolmogorov zero-one law, 263
- Kolmogorov’s inequality, 183

- Kolmogorov-Smirnov limit theorem, 206  
Kolmogorov-Smirnov statistic, 215  
Komatsu's lemma, 250  
Kronecker's lemma, 187  
  
Laplace transform, 176  
Last exit time, 249  
Law of the iterated logarithm, 198, 200  
Law of total probability, 35  
Lebesgue measure, 38  
Lebesgue measure, Integral with respect to, 127, 128  
Lebesgue null set, 38  
Likelihood ratio, 212, 246, 264  
Lindeberg condition, 192–194, 196, 212, 213  
Lindeberg-Feller central limit theorem, 194, 196  
Log normal distribution, 69  
Lyapunov condition, 191, 193, 212  
Lyapunov's inequality, 120, 123  
  
Marginal density function, 54  
Markov's inequality, 122  
Martingale, 243, 244, 247, 248, 250, 267  
Martingale convergence theorem, 255, 257–259, 268  
Martingale, Reversed, 267  
Martingale, Uniformly integrable, 245, 253, 259  
Maximum likelihood estimator, 201, 202, 204, 215  
Maxwell-Boltzmann model, 83–85, 87, 100  
Mean, 123, 126  
Mean squared error (MSE), 217  
Measure, 37  
Median, 130  
Minkowski's inequality, 121, 123  
MMSE predictor, 217, 222, 224, 228, 239–241  
Moment, 123  
  
Moment generating function, 177  
Monotone class argument, 23  
Monotone class theorem, 23, 51, 66  
Monotone convergence theorem, 106, 110, 232  
Monte Carlo integration, 201, 214  
  
Negative binomial distribution, 55, 56, 89, 97, 99, 125, 164  
Normal distribution, 13, 56, 57, 59–61, 69, 78, 97, 99, 100, 125, 132, 159, 164, 165, 168, 171, 179, 190, 197, 215  
Normal distribution, Bivariate, 60, 76, 96, 240  
Normal distribution, Multivariate, 126, 241  
Normal equations, 224, 240  
  
Occupancy model, 83, 84, 100  
Optional sampling theorem, 251–253, 266  
Orthogonal decomposition theorem, 222, 224  
Outcome, 16  
  
Pólya urn scheme, 42, 246, 269  
Point mass, 24, 31  
Poisson distribution, 32, 56, 60, 68, 79, 87, 91, 98, 125, 130, 132, 155, 160, 161, 164, 174, 179  
Poisson limit theorem, 155, 174  
Poisson process, 91–93, 160, 179, 180, 240  
Probability, 2, 24–28, 106, 109  
Probability on  $\mathbb{R}$ , 28, 29, 31, 34, 39  
Probability on  $\mathbb{R}$ , Absolutely continuous, 33, 34  
Probability on  $\mathbb{R}$ , Discrete, 32  
Probability on  $\mathbb{R}$ , Singular, 38  
Probability space, 24

- Probability, Conditional, 35  
 Product  $\sigma$ -algebra, 94  
 Product probability, 95, 128, 129  
 Quadratic mean convergence, 136, 140, 141, 145, 147, 149, 158, 220  
 Quantile function, 63  
 Quantile transformation, 63  
 Radon-Nikodym theorem, 261  
 Random experiment, 16  
 Random sum of random variables, 133, 179, 207, 241  
 Random variable, 1, 44, 47, 50, 52  
 Random variable,  $\sigma$ -algebra generated by, 46  
 Random variable, Indicator, 44  
 Random variable, Integrable, 107  
 Random variable, Lattice, 180  
 Random variable, Simple, 45, 50  
 Random variable, Symmetric, 179  
 Random variables, Convergence, 49, 135, 136, 157  
 Random variables, i.i.d., 71  
 Random variables, Identically distributed, 52  
 Random variables, Independent, 3, 5, 71–74, 76, 77, 79, 94, 95, 117, 118, 165, 175  
 Random variables, Orthogonal, 221  
 Random variables, Series, 49, 106, 185, 186, 212  
 Random variables, Uncorrelated, 125, 131  
 Random vector, 45, 53  
 Random vectors, Convergence, 149  
 Random walk, 1, 13, 97, 179, 216  
 Ranks, 98, 130  
 Rayleigh distribution, 97  
 Reflection principle, 7, 8, 10  
 Renewal process, 208, 209  
 Reversed martingale, 267, 268  
 Sample mean, 132, 215  
 Sample space, 16  
 Sample standard deviation, 132  
 Sample variance, 214  
 Set operations, 18–20, 40  
 Sets, Convergence, 20, 40  
 Simple function, 43  
 Slutsky's theorem, 146, 147  
 Standard deviation, 124  
 Standardization, 155  
 Stirling's approximation, 9, 87  
 Stochastic process, 45  
 Stopping time, 249, 267, 269  
 Stopping time, Event prior to, 265  
 Strong law of large numbers, 12, 151, 152, 161, 188, 190, 212, 240, 268  
 Submartingale, 243, 244, 248, 267  
 Supermartingale, 243, 244, 267  
 Survivor function, 31, 52  
 Symmetric  $\sigma$ -algebra, 263  
 Symmetric event, 263  
 Tail  $\sigma$ -algebra, 262  
 Tail event, 262  
 Tail probability, 31  
 Three series theorem, 186  
 Uniform absolute continuity, 143  
 Uniform distribution, 24, 31, 59, 60, 76, 82, 125, 164, 173, 215  
 Uniform integrability, 142–144, 161, 212  
 Upcrossing, 255, 256  
 Variance, 124, 131  
 Wald's identity, 241, 254, 269  
 Weak law of large numbers, 151, 152, 173  
 Weierstrass approximation theorem, 156  
 Wilcoxon rank statistic, 99  
 Young's inequality, 119  
 Zero-one law, 263

<i>Keyfitz</i>	Applied Mathematical Demography, Second Edition
<i>Kiefer</i>	Introduction to Statistical Inference
<i>Kokoska and Nevison</i>	Statistical Tables and Formulae
<i>Lindman</i>	Analysis of Variance in Experimental Design
<i>Madansky</i>	Prescriptions for Working Statisticians
<i>McPherson</i>	Statistics in Scientific Investigation: Its Basis, Application, and Interpretation
<i>Nguyen and Rogers</i>	Fundamentals of Mathematical Statistics: Volume I: Probability for Statistics
<i>Nguyen and Rogers</i>	Fundamentals of Mathematical Statistics: Volume II: Statistical Inference
<i>Noether</i>	Introduction to Statistics: The Nonparametric Way
<i>Peters</i>	Counting for Something: Statistical Principles and Persona
<i>Pfeiffer</i>	Probability for Applications
<i>Pitman</i>	Probability
<i>Santner and Duffy</i>	The Statistical Analysis of Discrete Data
<i>Saville and Wood</i>	Statistical Methods: The Geometric Approach
<i>Sen and Srivastava</i>	Regression Analysis: Theory, Methods, and Applications
<i>Whittle</i>	Probability via Expectation, Third Edition
<i>Zacks</i>	Introduction to Reliability Analysis: Probability Models and Statistical Methods