JAMES DAVIDSON

STOCHASTIC LIMIT THEORY

AN INTRODUCTION FOR ECONOMETRICIANS • SECOND EDITION

OXFORD

Stochastic Limit Theory

Stochastic Limit Theory

An Introduction for Econometricians

Second Edition

JAMES DAVIDSON

OXFORD
UNIVERSITY PRESS

OXFORD
UNIVERSITY PRESS

Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

First Edition published in 1994
Second Edition published in 2021

Impression: 1

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2021939064

ISBN 978–0–19–284450–7

DOI: 10.1093/oso/9780192844507.001.0001

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

For
Lynette,
Julia, and Nicola

. . . what in me is dark
Illumine, what is low raise and support,
That, to the height of this great argument,
I may assert Eternal Providence,
And justify the ways of God to men.
Paradise Lost, Book I, 16–20

In the formal construction of a course in the theory of probability, limit theorems appear as a kind of superstructure over elementary chapters, in which all problems have finite, purely arithmetical character. In reality, however, the epistemological value of the theory of probability is revealed only by limit theorems. Moreover, without limit theorems it is impossible to understand the real content of the primary concept of all our sciences—the concept of probability.

B. V. Gnedenko and A. N. Kolmogorov
Limit Theorems for Sums of Independent Random Variables

Contents

I. MATHEMATICS

II. PROBABILITY

III. THEORY OF STOCHASTIC PROCESSES

IV. THE LAW OF LARGE NUMBERS

V. THE CENTRAL LIMIT THEOREM

VI. THE FUNCTIONAL CENTRAL LIMIT THEOREM

From Preface to the First Edition

Recent years have seen a marked increase in the mathematical sophistication of econometric research. While the theory of linear parametric models which forms the backbone of the subject makes an extensive and clever use of matrix algebra, the statistical prerequisites of this theory are comparatively simple. But now that these models are pretty thoroughly understood, research is concentrated increasingly on the less tractable questions, such as nonlinear and nonparametric estimation and nonstationary data generation processes. The standard econometrics texts are no longer an adequate guide to this new technical literature, and a sound understanding of the probabilistic foundations of the subject is becoming less and less of a luxury.

The asymptotic theory traditionally taught to students of econometrics is founded on a small body of classical limit theorems, such as Khinchine's weak law of large numbers and the Lindeberg–Lévy central limit theorem, relevant to the stationary and independent data case. To deal with linear stochastic difference equations, appeal can be made to the results of Mann and Wald ([131]), but even these are rooted in the assumption of independent and identically distributed disturbances. This foundation has become increasingly inadequate to sustain the expanding edifice of econometric inference techniques, and recent years have seen a systematic attempt to construct a less restrictive limit theory. Hall and Heyde's *Martingale Limit Theory and its Application* ([88]) is an important landmark, as are a series of papers by econometricians including among others Halbert White, Ronald Gallant, Donald Andrews, and Herman Bierens. This work introduced to the econometrics profession pioneering research into limit theory under dependence, done in the preceding decades by probabilists such as J. L. Doob, I. A. Ibragimov, Patrick Billingsley, Robert Serfling, Murray Rosenblatt, and Donald McLeish.

These latter authors devised various concepts of limited dependence for general nonstationary time series. The concept of a martingale has a long history in probability, but it was primarily Doob's *Stochastic Processes* ([60]) that brought it to prominence as a tool of limit theory. Martingale processes behave like the wealth of a gambler who undertakes a succession of fair bets; the differences of a martingale (the net winning at each step) are unpredictable from lagged information. Powerful limit theorems are available for martingale difference sequences involving no further restrictions on the dependence of the process. Ibragimov and Rosenblatt respectively defined strong mixing and uniform mixing as characterizations of 'limited memory', or independence at long range. McLeish defined the notion of a mixingale, the asymptotic counterpart of a martingale difference, becoming unpredictable m steps ahead as m becomes large. This is a weaker property than mixing because it involves only low-order moments of the distribution, but mixingales possess most of those attributes of mixing processes needed to make limit theorems work. Very important from the econometrician's point of view is the property dubbed by Gallant and White ([76]) near-epoch dependence from a phrase in one of McLeish's papers, although the idea itself goes back to Billingsley ([21]) and Ibragimov ([103]). The mixing property may not be preserved by transformations of sequences involving an infinite number of lags, but near-epoch dependence is a condition under which the outputs of a dynamic econometric model can be shown, given some further conditions, to be mixingales when the inputs are mixing. Applications of these results are increasingly in evidence in the econometric literature; Gallant and White's monograph provides an excellent survey of the possibilities.

Limit theorems impose restrictions on the amount of dependence between sequence coordinates, and on their marginal distributions. Typically, the probability of outliers must be controlled by requiring the existence of higher-order moments, but there are almost always trade-offs between dependence and moment restrictions, allowing one to buy more of one at the price of less of the other. The fun of proving limit theorems has been to see how far out the envelope of sufficient conditions can be stretched, in one direction or another. To complicate matters, one can get results both by putting limits on the rate of approach to independence (the rate of mixing) and by limiting the type of dependence (the martingale approach), as well as by combining both types of constraint (the mixingale approach). The results now available are remarkably powerful, judged by the yardstick of the classical theory. Proofs of necessity are elusive and the limits to the envelope are not yet known with certainty, but they probably lie not too far beyond the currently charted points.

Perhaps the major development in time-series econometrics in the 1980s has been the theory of cointegration, and dealing with the distributions of estimators when time series are generated by unit root processes also requires a new type of limit theory. The essential extra ingredient of this theory is the functional central limit theorem (FCLT). The proof of these weak convergence results calls for a limit theory for the space of functions, which throws up some interesting problems which have no counterpart in ordinary probability. These ideas were pioneered by Russian probabilists in the 1950s, notably A. V. Skorokhod and Yu. V. Prokhorov. It turns out that FCLTs hold under properties generally similar to those for the ordinary CLT (though with a crucial difference), and they can be analysed with the same kind of tools, imposing limitations on dependence and outliers

The probabilistic literature which deals with issues of this kind has been seen as accessible to practising econometricians only with difficulty. Few concessions are made to the nonspecialist, and the concerns of probabilists, statisticians, and econometricians are frequently different. Textbooks on stochastic processes (Cox and Miller [35] is a distinguished example) often give prominence to topics that econometricians would regard as fairly specialized (e.g. Markov chains, processes in continuous time), while the treatment of important issues like nonstationarity gets tucked away under the heading of advanced or optional material if not omitted altogether. Probability texts are written for students of mathematics and assume a familiarity with the folklore of the subject that econometricians may lack. The intellectual investment required is one that students and practitioners are often, quite reasonably, disinclined to make.

It is with issues of this sort in mind that the present book has been written. The first objective has been to provide a coherent and unified account of modern asymptotic theory, which can function as both a course text, and as a work of reference. The second has been to provide a grounding in the requisite mathematics and probability theory, making the treatment sufficiently self-contained that even readers with limited mathematical training might make use of it. This is not to say that the material is elementary. Even when the mathematics is mastered, the reasoning can be intricate and demand a degree of patience to absorb. Proofs for nearly all the results are provided, but readers should never hesitate to pass over these when they impede progress. The book is also intended to be useful as a reference for students and researchers who only wish to know basic things, like the meaning of technical terms, and the variety of limit results available. But, that said, it will not have succeeded in its aim unless the reader is sometimes stimulated to gain a deeper understanding of the material—if for no better reason, because this is a theory abounding in mathematical elegance, and technical ingenuity which is often dazzling.

London
June 1994

Preface to the Second Edition

In preparing a new edition of this book I have kept two considerations in mind. The first is to be true to the intended aim of the original; to provide students of econometrics with a window into asymptotic theory without assuming too much prior knowledge, while at the same time never compromising rigour for those readers willing to embrace it. The second is to make a better book by expanding the range of topics, increasing the depth of coverage of existing topics and updating superseded results. All the proofs have been reviewed for possible improvements, clarifications, or corrections, and a few have been replaced, although little material has been removed altogether. There has, however, been a good deal of reorganization of topics. Some material has moved chapters in the course of getting expanded. Many of the updates have benefitted from my joint work with Robert de Jong, whose contribution is most warmly acknowledged. These include improved results for the strong law of large numbers and the central limit theorem under dependence, the functional central limit theorem, and convergence of stochastic integrals.

Among topics newly treated or treated more fully here are the properties of Brownian motion, Skorokhod embedding, the law of the iterated logarithm, infinite divisibility, the α-stable class of limit distributions with infinite variance, and weak convergence to limits in the space of càdlàg functions. The infinite variance case was neglected in the first edition simply because it did not feature prominently in the literature the book was aiming to survey. However, these ideas are truthfully less important for use in applications than for the insight they provide into the necessary and sufficient conditions for the central limit theorem. A major omission from the works referenced in the first edition is Gnedenko and Kolmogorov's *Limit Distributions for Sums of Independent Random Variables* ([82]), as translated into English and annotated by Kai Lai Chung. This is a wonderful book—comprehensive, clear and highly readable. While the main focus of the present work is and remains on dependent data, one must not overlook the fact that all the problems had to be solved first for the independence case and this was the major achievement. Paul Lévy in Paris and Alexander Khinchine and Andrey Kolmogorov in Moscow were the pioneers of this field in the interwar period, and it is the Soviet school that has put its indelible stamp on the subject, with Yuri Prokhorov and especially Anatoli Skorokhod notable among the following generation of contributors. It is a curious and poignant fact that probability theory has proved to be one of the rare positive contributions of Stalinism to the human condition. But we should not overlook that another rising star of the interwar period, Jósef Marcinkiewicz ([133]), had the misfortune to be Polish and was murdered in the Katyn Forest by the same regime.

It is of course Patrick Billingsley's *Convergence of Probability Measures* ([21]) that has been the pre-eminent reference in this field and a major inspiration for the present work. Billingsley's second edition of 1999 ([22]) is a thorough rewrite that adds a range of new topics but also regrettably omits some others—almost a new book, in fact. Since the first edition is so widely cited in the literature it is actually unfortunate that it is no longer in print, since the second is not a perfect substitute. Where necessary, both editions are cited here.

Among the other important books that have appeared or been reissued or simply come to light in the years since the first edition, mention should be made of Gikhman and Skorokhod [78], [79]; van de Vaart [181]; van de Vaart and Wellner [182]; Jacod and Shiryaev [107]; Samorodnitsky and Taqqu [163]; and Csörgő and Révész [37]. These are serious monographs attempting the same kind of thing as the present book, but generally at a more sophisticated and advanced level and without the same concessions to nonspecialists. Alternative versions of many of the results following can be found therein and of course much more besides, although this book retains its special focus on the 'near-epoch dependence on mixing' characterization of serial dependence. Of the innumerable useful textbooks of probability and statistics now available the following recent acquisitions, listed roughly in order of accessibility and usefulness to this reader at least, are deserving of mention: Gut [87], Williams [190], Durrett [64], Ash [10], Gnedenko [80], Kallenberg [109]. It is also most necessary to give credit to the key texts that informed the first edition. The immortal classics include Kolmogorov [114], Cramér [36], Doob [60], Feller [74], and Loève [121]; beyond these the essential sources include Billingsley [23], Hall and Heyde [88], Parthasarathy [141], Breiman [26], Ibragimov and Linnik [105], Dudley [63], Gallant and White [76], Pollard [147], Lukacs [122], Révész [158], Shiryaev [169]; and for the mathematics, especially Royden [162], Dieudonné [57], Halmos [91], and Apostol [9].

Much the biggest change in working environment between now and the time of the first edition of 25 years ago is of course the Internet. Nearly all the materials needed, certainly all of the classic texts and monographs as well as (given a university account) the journal literature, can now be ferreted out online either for free or at a reasonable charge. One can never be too sure of the copyright position, but for a work long out of print such as the Gnedenko/Kolmogorov volume cited above, such access is indispensable. Dare one say it, even Wikipedia articles can be surprisingly useful for refreshing the memory on bits and bobs of obscure mathematics, even if one could not make the claim for much of the other content on that site. Recalling now the hours spent in the gloomy basement of Senate House (University of London) library back then, this is a wonderfully liberating way to work.

I must take this opportunity to acknowledge and thank the people who commented encouragingly and helpfully on the first edition, over the years. The names of Stéphane Gregoir, Kairat Mynbaev, Jean-Pierre Dion, Harry Kelejian, Giuseppe Cavaliere, and Bent Nielsen come especially to mind, among many others. The compliment that I must admit to valuing above all is the citation in Patrick Billingsley's second edition of *Convergence of Probability Measures.*

Regarding the new edition: I must first thank Jenny Firth, who expertly re-keyed the original text in LaTeX and without whose efforts this project would have been much harder to bring to fruition. Oxford University Press accepted the proposal in 2018, and thanks are due to Katie Bishop and Adam Swallow in their successive roles as commissioning editors. The first edition was delivered to OUP back in 1994 as a parcel of A4 camera-ready pages, but times and technology have changed; thanks are also due to Henry Clarke and Saraswathi Ethiraju and his team, who have laboured with me to produce this internet-friendly text.

Exeter
June 2021

Mathematical Symbols and Abbreviations

In the text, the symbol □ is used to terminate examples and definitions and also theorems and lemmas unless the proof follows directly. The symbol ■ terminates proofs. References to numbered expressions are enclosed in parentheses. References to numbered theorems, examples etc. are given in bold face. References to chapter sections are preceded by §.

In statements of theorems, roman numbers (i), (ii), (iii), . . . are used to indicate the parts of a multi-part result. Lower case letters (a), (b), (c), . . . are used to itemize the assumptions or conditions specified in a theorem and also the components of a definition.

The page numbers below refer to fuller definitions or examples of use, as appropriate.

$\lvert\cdot\rvert$	absolute value	22
$\lVert\cdot\rVert_p$	L_p-norm	177
$\lVert\cdot\rVert$	Euclidean norm;	27
	also fineness (of a partition)	603
$\Rightarrow$	weak convergence (of measures);	495, 639
	also implication	22
$\uparrow, \downarrow$	monotone convergence	28
$\to$	convergence	28
$\overset{\text{a.s.}}{\to}, \to_{\text{a.s.}}$	almost sure convergence	239
$\overset{\text{d}}{\to}, \to_{\text{d}}$	convergence in distribution	495
$\overset{L_p}{\to}, \to_{L_p}$	convergence in L_p norm	411
$\overset{\text{pr}}{\to}, \to_{\text{pr}}$	convergence in probability	405
$\mapsto$	mapping, function	6
$\circ$	composition of mappings	8
$\sim$	asymptotic equality (of sequences)	39
$\overset{\text{d}}{\sim}, \sim_{\text{d}}$	equivalence in distribution (of r.v.s)	164
$\simeq$	equivalence in order of magnitude (of sequences)	39
$\dot{+}$	addition modulo 1	66
$-, \backslash$	set difference	4
$\leq, \geq$	partial ordering, inequality	6
$<, >$	strict ordering, strict inequality	6

$\ll$	order of magnitude inequality (of sequences);	39
	also absolutely continuous (of measures)	96
$\perp$	mutually singular (of measures)	96
$1_A(\cdot)$, $1(A)$	indicator function	75
2^X	power set of X	15
a.e.	almost everywhere	56
AR	autoregressive process	300
ARMA	autoregressive moving average process	263
a.s., a.s.$[\mu]$	almost surely, (with resp. to p.m. μ)	149
A^c	complement of A	4
$\bar{A}$, $(A)^-$	closure of A	23,107
A^o	interior of A	23,107
α_m	strong mixing coefficient	290
$\aleph_0$	aleph-nought (cardinality of $\mathbb{N}$)	9
$\forall$	‘for every’	14
$\mathcal{B}$	Borel field	19
CLT	central limit theorem	520
ch.f.	characteristic function	214
c.d.f.	cumulative distribution function	155
$C_{[0,1]}$	continuous functions on the unit interval	602
$\mathbb{C}$	complex plane	215
$\subseteq$, $\supseteq$	set containment	3
$\subset$, $\supset$	strict containment	3
$\chi^2(\nu)$	chi-squared distribution with ν degrees of freedom	165
$d(x,y)$	distance between x and y	104
$D_{[0,1]}$	càdlàg functions on the unit interval	668
$\mathbb{D}$	dyadic rationals	31
$\triangle$	symmetric difference of sets	4
Δ	difference operator $(1-L)$	263
∂A	boundary of A	23,107
$\in$	set membership	3
ess sup	essential supremum	154
$\mathrm{E}(\cdot)$	expectation	171
$\mathrm{E}(\cdot \mid x)$	conditional expectation (on variable x)	191
$\mathrm{E}(\cdot \mid \mathcal{G})$	conditional expectation (on σ-field $\mathcal{G}$)	195
$\exists$	‘there exists’	17
f^+, f^-	positive, negative parts of f	81
FCLT	functional central limit theorem	661

$F(\cdot)$	cumulative distribution function	155
$\phi_X(\cdot)$	characteristic function of X	214
ϕ_m	uniform mixing coefficient	290
$\Gamma(\cdot)$	gamma function	162
iff	‘if and only if’	6
inf	infimum	14
i.i.d.	independently and identically distributed	256
i.o.	infinitely often	401
in pr.	in probability	405
LIE	law of iterated expectations	197
LIL	law of the iterated logarithm	586
LLN	law of large numbers	414
lim	limit (sets); *also* limit (numbers)	14, 28
limsup, $\overline{\lim}$	superior limit (sets); *also* superior limit (numbers)	15, 30
liminf, $\underline{\lim}$	inferior limit (sets); *also* inferior limit (numbers)	15, 30
L	lag operator	263
$L(n)$	slowly varying function	45
L_p-NED	near-epoch dependent in L_p-norm	368
MA	moving average process	261
$m(\cdot)$	Lebesgue measure	52
m.d.	martingale difference	316
m.g.f.	moment-generating function	214
m.s.	mean square	411
$\mathbb{M}$	space of measures	638
$\mathrm{N}(\mu,\sigma^2)$	Gaussian distribution with mean μ and variance σ^2	164
$\mathbb{N}$	natural numbers	9
$\mathbb{N}_0$	$\mathbb{N}\cup\{0\}$	9
$\cap, \bigcap$	intersection	4
$m \wedge n$	minimum of m and n	358
$O(\cdot)$	‘big Oh’, order of magnitude relation	39
$o(\cdot)$	‘little Oh’, strict order of magnitude relation	39
$O_p(\cdot)$	stochastic order relation	249
$o_p(\cdot)$	strict stochastic order relation	249
$\emptyset$	null set	9
p.d.f.	probability density function	161
p.m.	probability measure	147
$P(\cdot)$	probability	147
$P(\cdot\vert A)$	conditional probability (on event A)	150
$P(\cdot\vert\mathcal{G})$	conditional probability (on σ-field $\mathcal{G}$)	150

$\Pi, \prod$	multiple product of numbers;	219
	also multiple Cartesian product of sets;	5
	also partition of an interval	82
π	product measure	90
$\pi_t(\cdot)$	coordinate projection	600
$\mathbb{Q}$	rational numbers	10
r.v.	random variable	154
$R_{[0,1]}$	real valued functions on [0,1]	599
xRy	binary relation	5
$\mathbb{R}$	real line	11
$\mathbb{R}^+$	non-negative reals	12
$\bar{\mathbb{R}}$	extended real line, $\mathbb{R} \cup \{-\infty, +\infty\}$	14
$\mathbb{R}^k$	k-dimensional Euclidean space	13
sgn(x)	1 if $x = 0$, $x/\|x\|$ otherwise (sign of x)	215
s.e.	stochastic equicontinuity	482
s.s.e.	strong stochastic equicontinuity	482
SLLN	strong law of large numbers	413
sup	supremum	14
$S(x, \varepsilon)$	ε-neighbourhood, sphere	23
S_n	sum of random variables	415
s_n^2	variance of S_n	520
$S_\alpha(\delta^{1/\alpha}, \beta, \nu)$	α-stable family	517
$S_\alpha(\beta)$	$S_\alpha(1, \beta, 0)$	517
$\sigma(\mathcal{C})$	σ-field generated by collection $\mathcal{C}$	18
$\sigma(X)$	σ-field generated by r.v. X	193
$\int f \mathrm{d}x$	Lebesgue integral	81
$\int f \mathrm{d}\mu$, $\int f \mathrm{d}F$	Lebesgue–Stieltjes integral	81
$\int f \mathrm{d}P$	expected value (integral with resp. to p.m.)	171
$\Sigma, \sum$	sum	37
$t(\nu)$	Student's t distribution with ν d.f.	165
$T\omega$	shift transformation	271
$\cup, \bigcup$	union	4
U$[a, b]$	uniform distribution on interval$[a, b]$	164
$\vee, \bigvee$	union of σ-fields	20
$m \vee n$	maximum of m and n	358
WLLN	weak law of large numbers	413
w.p.1	with probability 1	149
$w(f, \delta)$, $w_x(\delta)$	modulus of continuity	123
$w'_x(\delta)$	modulus of continuity for D	670
$w''_x(\delta)$	alternate modulus of continuity for D	671

$\bar{X}_n$	sample mean of sequence $\{X_t\}_1^n$	413
$\times$	binary Cartesian product	5
$\otimes$	σ-field of product sets	68
	also Kronecker product	746
$\mathbb{Z}$	integers	10
$\bar{z} = a - \mathrm{i}b$	complex conjugate of $z = a + \mathrm{i}b$	215
$\{\cdot\}$	set designation;	3
	also sequence, array	28
$\{\cdot\}_1^\infty, \{\cdot\}_{-\infty}^\infty$	infinite sequences	28
$\{\{\cdot\}\}$	array	48
$[x]$	largest integer $\leq x$	10
$[a, b]$	closed interval bounded by a, b	12
(a, b)	open interval bounded by a, b	12
$(\Omega, \mathcal{F})$	measurable space	51
$(\Omega, \mathcal{F}, \mu)$	measure space	51
$(\Omega, \mathcal{F}^\mu, \bar{\mu})$	complete measure space	56
$(\Omega, \mathcal{F}, P)$	probability space	147
$(\mathbb{S}, d)$	metric space	104
$(\mathbb{X}, \tau)$	topological space	126

Common Usages

$A, B, C, D \dots$	sets
$X, Y, Z \dots$	random variables
$\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z} \dots$	random vectors
$f, g, h \dots$	functions
$\varepsilon, \delta, \eta$	positive constants
B, M	bounding constants
$\mathcal{A}, \mathcal{C}, \mathcal{D}, \mathcal{V} \dots$	collections of subsets
$\mathcal{F}, \mathcal{G}, \mathcal{H} \dots$	σ-fields
$\mathbb{S}, \mathbb{T}, \mathbb{X} \dots$	spaces
$\mu, \nu \dots$	measures
d, ρ	metrics
τ	topology

PART I

MATHEMATICS

1
Sets and Numbers

1.1 Basic Set Theory

By tradition the approach to the theory of sets is either 'axiomatic' or 'naïve'. In the former case rules are developed to avoid such paradoxical constructions as Bertrand Russell's notorious "set of all sets that are not members of themselves", although with little explicit reference to what a 'set' actually is. In the preface to his excellent little book *Naïve Set Theory* ([90]), Paul Halmos describes his treatment as "axiomatic set theory from a naïve point of view" which probably strikes the right balance. This chapter is definitely on the naïve side since no axioms are proposed and the focus is practical, being largely concerned with defining terminology and introducing notation. A *set* is a specified collection of objects and in this book the objects in question are most often numbers of some sort, although they may also be functions or other sets and sometimes wholly arbitrary, to be determined by the context in which the theory is applied.

In any analysis there is a set that defines the universe of discourse, containing all the objects under consideration. In what follows sets denoted A, B, etc. are subsets of a set X with generic element x. Often the term *space* is used to denote the universal set, particularly when, as in nearly every case of interest, the number of elements is infinite. Set membership is denoted by the symbol '$\in$', $x \in A$ meaning 'x belongs to the set A'. To show sets A and B have the same elements one writes $A = B$. The usual way to define a set is by a descriptive statement enclosed in braces, so that for example $A = \{x : x \in B\}$ defines membership of A in terms of membership of B and is an alternative way of writing $A = B$.

Another way to denote set membership is by *labels*. If a set has n elements the usual option is to write $A = \{x_i, i = 1, \dots, n\}$, but any set of labels will do. The statement $A = \{x_\alpha, \alpha \in C\}$ says that A is the set of elements bearing labels α contained in another set C, called the *index set* for A. The labels (indices) need not be numbers and can be any convenient objects at all. Sets whose elements are sets (the word 'collection' tends to be preferred in this context) are denoted by upper-case script characters. $A \in \mathcal{C}$ denotes that the set A is in the collection $\mathcal{C}$, or using indices one could write $\mathcal{C} = \{A_\alpha : \alpha \in C\}$.

B is called a *subset* of A, written $B \subseteq A$, if all the elements of B are also elements of A. If B is a proper subset of A, ruling out $B = A$, the relation is written $B \subset A$.

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
DOI: 10.1093/oso/9780192844507.003.0001

The *union* of A and B is the set whose elements belong to either or both sets, written $A \cup B$. The union of a collection $\mathcal{C}$, the set of elements belonging to one or more $A \in \mathcal{C}$, is denoted $\bigcup_{A \in \mathcal{C}} A$, or alternatively $\bigcup_{\alpha \in C} A_\alpha$ for the union of the collection $\{A_\alpha : \alpha \in C\}$. The *intersection* of A and B is the set of elements belonging to both, written $A \cap B$. The intersection of a collection $\mathcal{C}$ is the set of elements common to all the sets in $\mathcal{C}$, written $\bigcap_{A \in \mathcal{C}} A$ or $\bigcap_{\alpha \in \mathcal{C}} A_\alpha$. In particular, the union and intersection of $\{A_1, A_2, \dots, A_n\}$ are written $\bigcup_{i=1}^n A_i$ and $\bigcap_{i=1}^n A_i$. When the index set is implicit or unspecified, write just $\bigcup_\alpha A_\alpha$, $\bigcap_i A_i$ or similar.

The difference of sets A and B, written $A - B$ or by some authors $A \backslash B$, is the set of elements belonging to A but not to B, that is to say $A - B = A \cap B^c$. $X - A$ is the *complement* of A in X, also denoted A^c when X is understood. The *symmetric difference* of two sets is $A \triangle B = (A - B) \cup (B - A)$. Observe the implication, that $A^c \triangle B^c = A \triangle B$. The *null set* (or *empty set*) is $\emptyset = X^c$, the set with no elements. Sets with no elements in common (having empty intersection) are called *disjoint*. A *partition* of a set is a collection of disjoint subsets whose union is the set, such that each of its elements belongs to one and only one member of the collection.

Here are the basic rules of set algebra. Unions and intersections obey commutative, associative, and distributive laws:

$$A \cup B = B \cup A \tag{1.1}$$
$$A \cap B = B \cap A \tag{1.2}$$
$$(A \cup B) \cup C = A \cup (B \cup C) \tag{1.3}$$
$$(A \cap B) \cap C = A \cap (B \cap C) \tag{1.4}$$
$$A \cap (B \cup C) = (A \cap B) \cup (A \cap C) \tag{1.5}$$
$$A \cup (B \cap C) = (A \cup B) \cap (A \cup C). \tag{1.6}$$

There are also rules relating to complements known as de Morgan's laws:

$$(A \cup B)^c = A^c \cap B^c \tag{1.7}$$
$$(A \cap B)^c = A^c \cup B^c. \tag{1.8}$$

The distributive and de Morgan laws extend to general collections, as follows.

1.1 Theorem Let $\mathcal{C}$ be a collection of sets and B a set. Then

(i) $\left(\bigcup_{A \in \mathcal{C}} A\right) \cap B = \bigcup_{A \in \mathcal{C}} (A \cap B)$

(ii) $\left(\bigcap_{A \in \mathcal{C}} A\right) \cup B = \bigcap_{A \in \mathcal{C}} (A \cup B)$

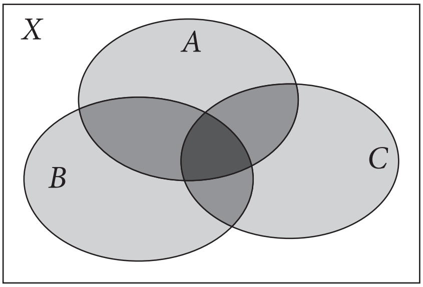

Figure 1.1

(iii) $\left(\bigcup_{A\in\mathcal{C}} A\right)^c = \bigcap_{A\in\mathcal{C}} A^c$

(iv) $\left(\bigcap_{A\in\mathcal{C}} A\right)^c = \bigcup_{A\in\mathcal{C}} A^c$. □

Venn diagrams, illustrated in Figure 1.1, are a useful device for clarifying relationships between subsets.

The *Cartesian product* of two sets A and B, written $A \times B$, is the set of all possible ordered pairs of elements, the first taken from A and the second from B; write $A \times B = \{(x,y) : x \in A, y \in B\}$. For a collection of n sets the Cartesian product is the set of all the n-tuples (ordered sets of n elements, with the ith element drawn from A_i) and is written

$$\prod_{i=1}^{n} A_i = \{(x_1, x_2, \dots, x_n) : x_i \in A_i, i = 1, \dots, n\}. \tag{1.9}$$

If one of the factor sets A_i is empty, $\Pi_{i=1}^{n} A_i$ is also empty.

Product sets are important in a variety of different contexts in mathematics. Some of these are readily appreciated; for example, sets whose elements are n-vectors of real numbers are products of copies of the real line (see §1.4). But product sets are also central to the mathematical formalization of the notion of relationship between set elements.

Thus: a *relation* R on a set A is any subset of $A \times A$. If $(x,y) \in R$ this is written xRy. R is said to be

reflexive iff xRx,
symmetric iff xRy implies yRx,
antisymmetric iff xRy and yRx implies $x = y$,
transitive iff xRy and yRz implies xRz,

where in each case the indicated condition holds for every x, y, and $z \in A$, as the case may be. (Note: 'iff' means 'if and only if').

An *equivalence relation* is a relation that is reflexive, symmetric and transitive. Given an equivalence relation R on A, the *equivalence class* of an element $x \in A$ is the set $E_x = \{y \in A : xRy\}$. If E_x and E_y are the equivalence classes of elements x and y, then either $E_x \cap E_y = \emptyset$, or $E_x = E_y$. In other words, the equivalence classes of the elements of A form a partition of A. The equality relation $x = y$ is the obvious example of an equivalence relation, but by no means the only one.

A partial ordering is any relation that is reflexive, antisymmetric, and transitive. Partial orderings are usually denoted by the symbols $\leq$ or $\geq$, with the understanding that $x \geq y$ is the same as $y \leq x$. To every partial ordering there corresponds a *strict ordering*, defined by the omission of the elements (x, x) for all $x \in A$. Strict orderings, usually denoted by $<$ or $>$, are not reflexive or antisymmetric, but they are transitive. A set A is said to be *linearly ordered* by a partial ordering $\leq$ if one of the relations $x < y$, $x > y$, or $x = y$ holds for every pair $(x, y) \in A \times A$. If there exist elements $a \in A$ and $b \in A$ such that $a \leq x$ for all $x \in A$, or $x \leq b$ for all $x \in A$, a and b are called respectively the *smallest* and *largest* elements of A. A linearly ordered set A is called *well-ordered* if every subset of A contains a smallest element. It is of course in sets whose elements are numbers that the ordering concept is most familiar.

1.2 Mappings

Consider two sets X and Y, which can be thought of as representing the universal sets for a pair of related problems. The following bundle of definitions contains the basic ideas about relationships between the elements of such sets. A *mapping* (or *transformation* or *function*)

$T : X \mapsto Y$

is a rule that associates each element of X with a unique element of Y; in other words, for each $x \in X$ there exists a specified element $y \in Y$, denoted $T(x)$. X is called the *domain* of the mapping and Y the *codomain*. The set

$$G_T = \big\{(x, y) : x \in X,\ y = T(x)\big\} \subseteq X \times Y \tag{1.10}$$

is called the *graph* of T. For $A \subseteq X$ the set

$$T(A) = \{T(x) : x \in A\} \subseteq Y \tag{1.11}$$

is called the *image* of A under T. For $B \subseteq Y$ the set

$$T^{-1}(B) = \{x : T(x) \in B\} \subseteq X \tag{1.12}$$

is called the *inverse image* of B under T. The set $T(X)$ is called the *range* of T and if $T(X) = Y$ the mapping is said to be from X *onto* Y and otherwise *into* Y. If each y is the image of one and only one $x \in X$, so that $T(x_1) = T(x_2)$ if and *only* if $x_1 = x_2$, the mapping is said to be *one-to-one,* or 1–1.

The notions of mapping and graph are really interchangeable and it is permissible to say that the graph *is* the mapping, but it is convenient to keep a distinction in mind between the rule and the subset of $X \times Y$ which it generates. The term *function* is usually reserved for cases when the codomain is the set of real numbers (see §1.4). The term *correspondence* is used for a rule connecting elements of X to elements of Y where the latter are not necessarily unique. T^{-1} is a correspondence, but not a mapping unless T is one-to-one. However, the term *one-to-one correspondence* is often used specifically, in certain contexts that will arise below, to refer to a mapping that is both 1–1 and onto. If partial orderings are defined on both X and Y, a mapping is called *order-preserving* if $T(x_1) \leq T(x_2)$ iff $x_1 \leq x_2$. On the other hand, if X is partially ordered by $\leq$, a 1–1 mapping *induces* a partial ordering on the codomain, defined by '$T(x_1) \leq T(x_2)$ iff $x_1 \leq x_2$'. If the mapping is also onto, a linear ordering on X induces a linear ordering on Y.

The next theorem is a miscellany of useful facts about mappings.

1.2 Theorem

(i) For a collection $\{A_\alpha \subseteq X\}$, $T\left(\bigcup_\alpha A_\alpha\right) = \bigcup_\alpha T(A_\alpha)$;
(ii) for a collection $\{B_\alpha \subseteq Y\}$, $T^{-1}\left(\bigcup_\alpha B_\alpha\right) = \bigcup_\alpha T^{-1}(B_\alpha)$;
(iii) for $B \subseteq Y$, $T^{-1}(B^c) = T^{-1}(B)^c$;
(iv) for $A \subseteq X$, $A \subseteq T^{-1}\big(T(A)\big)$;
(v) for $B \subseteq Y$, $T\big(T^{-1}(B)\big) \subseteq B$. □

Here, $T^{-1}(B)^c$ means $X - T^{-1}(B)$. Using de Morgan's laws, properties (ii) and (iii) are easily extended to the inverse images of intersections and differences; for example, the inverse images of disjoint sets are also disjoint. However, $Y - T(A) = T(A)^c \neq T(A^c)$, in general. Parts (iv) and (v) are illustrated in Figure 1.2, where X and Y both correspond to the real line, A and B are intervals of the line, and T is a function of a real variable.

When T is a 1–1 correspondence (1–1 and onto) so is T^{-1}. These properties then hold symmetrically and the inclusion relations of parts (iv) and (v) also become equalities for all $A \subseteq X$ and $B \subseteq Y$.

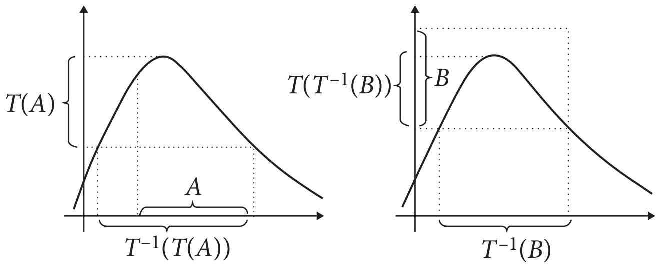

Figure 1.2

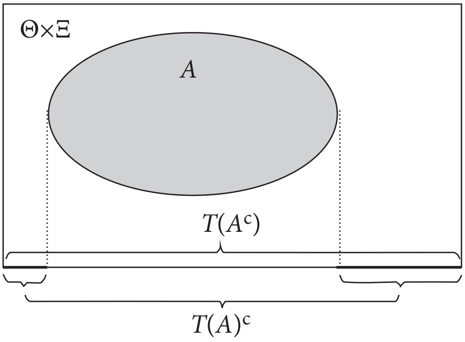

Figure 1.3

1.3 Example If $X = \Theta \times \Xi$ is a product space, having as elements the ordered pairs $x = (\theta, \xi)$, the mapping

$$T : \Theta \times \Xi \mapsto \Xi$$

defined by $T(\theta, \xi) = \xi$ is called the *projection mapping* onto Ξ. The projection of a set $A \subseteq \Theta \times \Xi$ onto Ξ (respectively, Θ) is the set consisting of the second (resp., first) members of each pair in A. On the other hand, for a set $B \in \Xi$, $T^{-1}(B) = \Theta \times B$. It is a useful exercise to verify **1.2** for this case and also to check that $T(A)^c \neq T(A^c)$ in general. In Figure 1.3, Θ and Ξ are line segments and $\Theta \times \Xi$ is a rectangle in the plane. Here, $T(A)^c$ is the union of the indicated line segments, whereas $T(A^c) = \Xi$. □

If Z is a third set and

$$U : Y \mapsto Z$$

is a further mapping, the composite mapping

$$U \circ T : X \mapsto Z$$

takes each $x \in X$ to the element $U\big(T(x)\big) \in Z$. $U \circ T$ operates as a simple transformation from X to Z and **1.2** applies to this case. For $C \subseteq Z$,

$$(U \circ T)^{-1}(C) = T^{-1}\big(U^{-1}(C)\big).$$

1.3 Countable Sets

The number of elements contained in a set is called the *cardinality* or *cardinal number* of the set. The notion of 'number' in this context is not a primitive one, but can be reduced to fundamentals by what is called the 'pigeonhole' principle. A set A is said to be *equipotent* with a set B if there exists a 1–1 correspondence connecting A and B. Think in terms of taking an element from each set and placing the pair in a pigeonhole. Equipotency means that such a procedure can never exhaust one set before the other.

Now, think of the number 0 as being just a name for the null set, $\varnothing$. Let the number 1 be the name for the set that has a single element, the number 0. Let 2 denote the set whose elements are the numbers 0 and 1. And proceeding recursively, let n be the name for the set $\{0, \ldots, n-1\}$. Then, the statement that a set A has n elements, or has cardinal number n, can be interpreted to mean that A is equipotent with the set n. The set of *natural numbers*, denoted $\mathbb{N}$, is the collection $\{n : n = 1, 2, 3, \ldots\}$. This collection is well ordered by the relation usually denoted $\leq$, where $n \leq m$ actually means the same as $n \subseteq m$ under this definition of a number.

Set theory is trivial when the number of elements in the set is finite, but formalization becomes indispensable for dealing with sets having an infinite number of elements. The set of natural numbers $\mathbb{N}$ is a case in point. If n is a member so is $n + 1$ and this is true for every n. Nonetheless a cardinal number is formally assigned to $\mathbb{N}$ and is represented by the symbol $\aleph_0$ ('aleph-nought').

When the elements of an infinite set can be put into a one-to-one correspondence with the natural numbers, the set is said to have cardinal number $\aleph_0$, but more commonly to be *countable* or, equivalently, *denumerable*. Countability of a set requires that a scheme can be devised for labelling each element with a unique element of $\mathbb{N}$. This imposes a well-ordering on the elements, such that there is a 'first' element labelled 1 and so on, although this ordering may have significance or be arbitrary, depending on the circumstances. It is the pigeonhole principle that matters here; that each element has its own unique label.

With infinite sets, everyday notions of size and quantity tend to break down. Augmenting the natural numbers by the number 0 defines the set $\mathbb{N}_0 = \{0, 1, 2, 3, \ldots\}$. The commonplace observation that $\mathbb{N}_0$ has 'one more' element than $\mathbb{N}$ is contradicted by the fact that $\mathbb{N}$ and $\mathbb{N}_0$ are equipotent (label $n - 1 \in \mathbb{N}_0$ by $n \in \mathbb{N}$). Still more surprisingly, the set of even numbers, $\mathbb{E} = \{2n, n \in \mathbb{N}\}$, also has an obvious labelling scheme demonstrating equipotency with $\mathbb{N}$. The naïve idea that there are 'twice as many' elements in $\mathbb{N}$ as in $\mathbb{E}$ is without logical foundation. *Every* infinite subset A of $\mathbb{N}$ has a natural well-ordering and is equipotent with $\mathbb{N}$ itself, the label of an element $x \in A$ being the cardinal number of the set $\{y \in A : y \le x\}$.

Turning to sets apparently 'larger' than $\mathbb{N}$, consider the *integers*, $\mathbb{Z} = \{\dots, -1, 0, 1, 2, \dots\}$, the set containing the signed whole numbers and zero. These are linearly ordered although not well ordered. They can, however, be paired with the natural numbers using the 'zigzag' scheme:

$(1,0), (2,1), (3,-1), (4,2), \dots, \big(n, [n/2](-1)^n\big), \dots,$

where $[x]$ denotes the largest whole number below x. Thus, $\mathbb{N}$ and $\mathbb{Z}$ are equipotent.

Then there are the *rational numbers*,

$$\mathbb{Q} = \{x : x = a/b, a \in \mathbb{Z}, b \in \mathbb{Z}, b \neq 0\}. \tag{1.13}$$

1.4 Theorem $\mathbb{Q}$ is a countable set.

Proof Construct a 1–1 correspondence between $\mathbb{Z} \times \mathbb{Z}$ and $\mathbb{N}$. A 1–1 correspondence between $\mathbb{Z} \times \mathbb{Z}$ and $\mathbb{Z} \times \mathbb{N}$ is obtained by the method just used to show $\mathbb{Z}$ countable and one between $\mathbb{Z} \times \mathbb{N}$ and $\mathbb{N} \times \mathbb{N}$ is got by the same method. Then note that the number $2^a3^b \in \mathbb{N}$ is *uniquely* associated with each pair $(a, b) \in \mathbb{N} \times \mathbb{N}$. The rule for recovering a and b from 2^a3^b is 'get a as the number of divisions by 2 required to get an odd number and the number so obtained is 3^b'. The collection $\{2^a3^b : a \in \mathbb{N}, b \in \mathbb{N}\} \subset \mathbb{N}$ is equipotent with $\mathbb{N}$ itself, as shown in the preceding paragraph. The composition of all these mappings is the desired correspondence. ∎

Generalizing this type of argument leads to the following fundamental result.

1.5 Theorem The union of a countable collection of countable sets is a countable set. □

The concept of a *sequence* is fundamental to all the topics in this book. A sequence can be thought of as a mapping whose domain is a well-ordered countable set, the index set. Since there is always an order-preserving 1–1 mapping from $\mathbb{N}$ to the index set, there is usually no loss of generality in considering the composite mapping and thinking of $\mathbb{N}$ itself as the domain. Another way to characterize a sequence is as the graph of the mapping, that is, a countable collection of pairs having the ordering conferred on it by the elements of the domain. The ranges of the sequences considered below typically contain either sets or real numbers; the associated theory for these cases is to be found respectively in §1.5 and §2.1.

The term sequence may also be applied to mappings having $\mathbb{Z}$ or another linearly ordered set as index set. This usage broadens the notion, since while such sets can be re-indexed by $\mathbb{N}$ (see above) this cannot be done while preserving the original ordering.

1.4 The Real Continuum

The real-number continuum $\mathbb{R}$ is such a complex object that no single statement of definition can do it justice. One can emphasize the ordinal and arithmetic properties of the reals, or their geometrical interpretation as the distances of points on a line from the origin (the point zero). But from a set-theoretic point of view, the essentials are captured by defining $\mathbb{R}$ as the set of countably infinite sequences of decimal digits, having a decimal point inserted at exactly one position in the sequence and possibly preceded by a minus sign.

Thus, the real number x can be written in the form

$$x = m(x)10^{p(x)}\sum_{i=1}^{\infty} d_i(x)10^{-i} \tag{1.14}$$

where the sequence $\{d_1(x), d_2(x), \ldots\}$ consists of decimal digits (elements of the set $\{0, 1, 2, \ldots, 9\}$), $p(x) \in \mathbb{N}_0$ denotes the position of the decimal point in the string (the decimal exponent), and $m(x) = +1$ if $x \geq 0$ and -1 otherwise (the sign). When $d_i(x) = 0$ for all but a finite number of terms, the decimal expansion of x is said to terminate and the final 0s are conventionally omitted from the representation.

The representation of x by (1.14) is not always unique and there exists a 1–1 correspondence between elements of $\mathbb{R}$ and sequences $\{m, p, d_1, d_2, d_{3,\ldots}\}$ only after certain of the latter are excluded. To eliminate arbitrary leading zeros stipulate that $d_1 \neq 0$ unless $p = 0$. And since for example $0.49999\ldots$ (the sequence of 9s not terminating) is the same number as 0.5, take the terminating representation of a number and exclude sequences having $d_i = 9$ in all but a finite number of places. $\mathbb{R}$ is of course linearly ordered and in terms of (1.14) the ordering corresponds to the lexicographic ordering of the sequences $\{m, mp, md_1, md_2, md_3 \ldots\}$.

The choice of base 10 in the definition is of course merely conventional. The 'base-D' representations

$$x = m(x)D^{p(x)}\sum_{i=1}^{\infty} d_i(x)D^{-i} \tag{1.15}$$

where the d_i are elements of a set of D ordered digits for any choice of natural number $D \geq 2$ are all equally valid. Of these alternatives the most important is the binary (base 2) representation where $D = 2$, the d_i are the binary digits 0 and 1, and $p(x)$ is the binary exponent. Octal ($D = 8$) and hexadecimal ($D = 16$) numbers are used in computing. If $D > 10$ the usual set of decimal digits has to be augmented, the hexadecimal digits being $0, \dots, 9, A, \dots, F$.

The integers have the representation in (1.14) with the strings terminating after $p(x)$ digits. The rationals are also elements of $\mathbb{R}$, being those which either terminate after a finite number of places, or else cycle repeatedly through a finite sequence of digits beyond a certain finite point. The real numbers that are not rational are called *irrational*. The irrational numbers are overwhelmingly more numerous than the rationals, representing a higher order of infinity. The following is the famous 'diagonal' argument of Georg Cantor.

1.6 Theorem The set $\mathbb{R}$ is uncountable.

Proof Assume a 1–1 correspondence between $\mathbb{R}$ and $\mathbb{N}$ exists. Now construct a real number in the following way. Let the first digit be different from that of the real number labelled 1, the second digit be different from that of the real number labelled 2, and in general the n^{th} digit be different from that of the real number labelled n, for every n. This number is different from every member of the labelled collection and hence it has no label. Since this construction can be performed for any labelling scheme, the assumption is contradicted. ∎

The cardinal number of $\mathbb{R}$ is denoted c and Theorem **1.6** says that $\aleph_0 < c$.

The linear ordering on $\mathbb{R}$ is of interest chiefly since it provides the basis for constructing the fundamental subsets of $\mathbb{R}$, the *intervals*. The set $A = \{x : \alpha < x < b\}$ is called an *open* interval since it does not contain the end points, whereas the interval $B = \{x : a \le x \le b\}$ is said to be *closed*. Common notations are $[a, b]$, (a, b), $(a, b]$, and $[a, b)$ to denote closed, open, and half-open intervals. A set containing just a single point a is called a *singleton*, written $\{a\}$. Unbounded intervals such as $C = \{x : a < x\}$, defined by a single boundary point, are written $(a, +\infty)$, $(-\infty, b)$ and $[a, +\infty)$, $(-\infty, b]$ for the open and closed cases respectively, where the infinities $+\infty$ and $-\infty$ are the fictional 'points' (not elements of $\mathbb{R}$) with the respective properties $x < +\infty$ and $x > -\infty$, for all $x \in \mathbb{R}$. An important example is the positive half-line $[0, +\infty)$, denoted subsequently by $\mathbb{R}^+$.

1.7 Theorem Every open interval is uncountable.

Proof Let the interval in question be (a, b). If $a < b$, there exists $n \ge 0$ such that the $(n+1)^{\text{th}}$ term of the sequence $(m, mp, md_1, md_2, \dots)$ in the expansion of (1.14) defining b exceeds that in the corresponding sequence for a, whereas the first n digits of each sequence are the same. The elements of (a, b) are those reals whose expansions generate the same initial sequence, with the $(n+1)^{\text{th}}$ terms neither exceeding that of b nor being exceeded by that of a. If a and b are distinct, n is finite. The result follows on applying the diagonal argument in **1.6** to these expansions, beginning at position $n+2$. ∎

Other useful results concerning $\mathbb{R}$ and its intervals include the following.

1.8 Theorem The points of any open interval are equipotent with $\mathbb{R}$.

Proof This might be proved by elaborating the argument of **1.7**, but it is simpler just to exhibit a 1–1 mapping from $\mathbb{R}$ onto (a,b). For example, the function

$$y = \frac{a+b}{2} + \frac{(b-a)x}{2(1+|x|)} \tag{1.16}$$

for $x \in \mathbb{R}$ fulfils the requirement. ∎

The real plane $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$ is the space whose elements are all the possible pairs of real numbers.

1.9 Theorem $\mathbb{R}^2$ is equipotent with $\mathbb{R}$.

Proof In view of the last theorem, it suffices to show that the unit interval $[0,1]$ is equipotent with the unit square $[0,1]^2$. Given points $x \in [0,1]$ and $y \in [0,1]$ whose decimal expansions from (1.14) have in each case $m = 1$ and $p = 0$, define the decimal expansion of the point $z \in [0,1]$ according to the rule

$$d_i(z) = \begin{cases} d_{(i+1)/2}(x), & i \text{ odd} \\ d_{i/2}(y), & i \text{ even} \end{cases}, \ i = 1,2,3,\dots. \tag{1.17}$$

In words, construct z by taking a digit from x and y alternately. Such a z exists for every pair x and y and given z, x and y can be uniquely recovered by setting

$$d_i(x) = d_{2i-1}(z),\ d_i(y) = d_{2i}(z), i = 1,2,3,\dots. \tag{1.18}$$

This defines a 1–1 mapping from $[0,1]^2$ onto $[0,1]$, as required. ∎

This argument can be extended from $\mathbb{R}^2$ to $\mathbb{R}^k$, the space of k-tuples of reals, for any $k \in \mathbb{N}$.

1.10 Theorem Every open interval contains a rational number.

Proof This is equivalent to the proposition that if $x < y$, there exists rational r with $x < r < y$. First suppose $x \geq 0$. Choose q as the smallest integer exceeding $1/(y - x)$, such that $qy > qx + 1$ and choose p as the smallest integer exceeding qy. Then $x < (p - 1)/q < y$. For the case $x < 0$ choose an integer $n > -x$ and then $x < r - n < y$, where r is the rational satisfying $n + x < r < n + y$, found as above. ∎

1.11 Corollary Every collection of disjoint open intervals is countable.

Proof Since each open interval contains a rational appearing in no other interval disjoint with it, a set of disjoint open intervals can be placed in 1–1 correspondence with a subset of the rationals. ∎

The *supremum* of a set $A \subset \mathbb{R}$, when it exists, is the smallest number y such that $x \leq y$ for every $x \in A$, written $\sup A$. The *infimum* of A, when it exists, is the largest number y such that $x \geq y$ for every $x \in A$, written $\inf A$.[1] These may or may not be elements of A. In particular, $\inf[a, b] = \inf(a, b) = a$ and $\sup[a, b] = \sup(a, b) = b$. Open intervals do not possess largest or smallest elements. However, every subset of $\mathbb{R}$ which is bounded above (resp. below) has a supremum (resp. infimum). While unbounded sets in $\mathbb{R}$ lack suprema and/or infima, it is customary to define the set $\bar{\mathbb{R}} = \mathbb{R} \cup \{-\infty, +\infty\}$, called the *extended* real line. In $\bar{\mathbb{R}}$, every set has a supremum, either a finite real number or $+\infty$, and similarly, every set has an infimum. The notation $\bar{\mathbb{R}}^+$ is used in the sequel to denote $\mathbb{R}^+ \cup \{+\infty\}$.

1.5 Sequences of Sets

Set sequences $\{A_1, A_2, A_3, \dots\}$ are written, variously, as $\{A_n : n \in \mathbb{N}\}$, $\{A_n\}_1^\infty$, or just $\{A_n\}$ when the context is clear.

A *monotone* sequence is one that is either *non-decreasing*, with each member of the sequence being contained in its successor ($A_n \subseteq A_{n+1}, \forall n$), or *non-increasing*, with each member containing its successor ($A_{n+1} \subseteq A_n, \forall n$). (Note, $\forall$ is the ubiquitous shorthand for "for every".) When the inclusion is strict, speak of increasing (resp. decreasing) sequences with $\subset$ (resp. $\supset$) replacing $\subseteq$ (resp. $\supseteq$). For a non-decreasing sequence, define the set $A = \bigcup_{n=1}^\infty A_n$ and for a non-increasing sequence the set $A = \bigcap_{n=1}^\infty A_n = \{\bigcup_{n=1}^\infty A_n^c\}^c$. These sets are called the *limits* of the respective sequences, indicated by $A_n \uparrow A$ or $A_n \downarrow A$ and also in general by $A_n \to A$ and $\lim_{n\to\infty} A_n = A$.

[1] Some older literature uses the terms *least upper bound* (L.U.B.) for supremum and *greatest lower bound* (G.L.B.) for infimum.

1.12 Example The sequence $\{[0, 1/n], n \in \mathbb{N}\}$ is decreasing and has as limit the singleton $\{0\}$. In fact, $\lim_{n\to\infty}[0, 1/n) = \{0\}$ also, whereas $\lim_{n\to\infty}(0, 1/n] = \emptyset$. The decreasing sequence of open intervals, $\{(a - 1/n, b + 1/n), n \in \mathbb{N}\}$, has as its limit the closed interval $[a, b]$. On the other hand, the sequence of closed intervals $\{[a + 1/n, b - 1/n], n \in \mathbb{N}\}$ (the set being empty if the lower bound exceeds the upper) is increasing and its limit is (a, b). □

Consider an arbitrary sequence $\{A_n\}$. The sequence $B_n = \bigcup_{m=n}^{\infty} A_m$ is non-increasing, so that $B = \lim_{n\to\infty} B_n$ exists. This set is called the *superior limit* of the sequence $\{A_n\}$, written $\operatorname{limsup}_n A_n$, and also as $\overline{\lim}_n A_n$. Similarly, the limit of the non-decreasing sequence $C_n = \bigcap_{m=n}^{\infty} A_m$ is called the *inferior limit* of the sequence, written $\operatorname{liminf}_n A_n$, or $\underline{\lim}_n A_n$. Formally: for a sequence $\{A_n, n \in \mathbb{N}\}$,

$$\operatorname*{limsup}_n A_n = \bigcap_{n=1}^{\infty}\left(\bigcup_{m=n}^{\infty} A_m\right) \tag{1.19}$$

$$\operatorname*{liminf}_n A_n = \bigcup_{n=1}^{\infty}\left(\bigcap_{m=n}^{\infty} A_m\right). \tag{1.20}$$

De Morgan's laws imply that $\operatorname{liminf}_n A_n = (\operatorname{limsup}_n A_n^c)^c$. The limsup is the set of elements contained in *infinitely many* of the A_n, while the liminf is the set belonging to *all but a finite number* of the A_n, that is, to every member of the sequence from some point onwards.

These concepts provide a criterion for convergence of set sequences in general. $\operatorname{liminf}_n A_n \subseteq \operatorname{limsup}_n A_n$ and if these two sets differ there are elements that belong to infinitely many of the A_n, but also do *not* belong to infinitely many of them. Such a sequence is not convergent. On the other hand, if $\operatorname{liminf}_n A_n = \operatorname{limsup}_n A_n = A$, the elements of A belong to infinitely many of the A_n and do not belong to at most a finite number of them. Then the sequence $\{A_n\}$ is said to converge to A and A is called the limit of the sequence.

1.6 Classes of Subsets

The set of all the subsets of X is called the *power set* of X, denoted 2^X. The power set of a set with n elements has 2^n elements, which accounts for its name and representation. In the case of a countable set, the power set in thought of formally as having $2^{\aleph_0}$ elements. One of the fundamental facts of set theory is that the number of subsets of a given set strictly exceeds the number of its elements. For finite sets this is obvious, but when extended to countable sets it amounts to the claim that $2^{\aleph_0} > \aleph_0$.

1.13 Theorem $2^{\aleph_0} = c$.

Proof The proposition is proved by showing that $2^{\mathbb{N}}$ is equipotent with $\mathbb{R}$, or equivalently (in view of **1.8**) with the unit interval $[0,1]$. For a set $A \in 2^{\mathbb{N}}$, construct the sequence of binary digits $\{d_1, d_2, d_3, \ldots\}$ according to the rule, '$d_n = 1$ if $n \in A$, $d_n = 0$ otherwise'. Using formula (1.15) with $D = 2$, $m = 1$, and $q = 0$, let this sequence define an element x_A of $[0,1]$ (the case where $d_n = 1$ for all n defines 1). On the other hand, for any element $x \in [0,1]$, construct the set $A_x \in 2^{\mathbb{N}}$ according to the rule, 'include n in A_x if and only if the nth digit in the binary expansion of x is a 1'. These constructions define a 1–1 correspondence between $2^{\mathbb{N}}$ and $[0,1]$. ∎

When studying the subsets of a given set, particularly their measure-theoretic properties, the power set is often too big for anything very interesting or useful to be said about it. The idea behind the following definitions is to specify subsets of 2^X that are large enough to be interesting, but whose characteristics may be more tractable. This is done by choosing a base collection of sets with known properties and then specifying certain operations for creating new sets from existing ones. These operations permit an interesting diversity of class members to be generated, but important properties of the sets may be deduced from those of the base collection, as the following examples show.

1.14 Definition A *ring* $\mathcal{R}$ is a nonempty class of subsets of X satisfying

(a) $\varnothing \in \mathcal{R}$

(b) if A and $B \in \mathcal{R}$ then $A \cup B \in \mathcal{R}$, $A \cap B \in \mathcal{R}$, and $A - B \in \mathcal{R}$. □

To define a ring, specify an arbitrary basic collection $\mathcal{C}$ which must include $\varnothing$ and then declare that any sets that can be generated by the specified operations also belong to the class. A ring is said to be *closed* under the operations of union, intersection, and difference.

Rings lack a crucial piece of structure, for there is no requirement for the set X itself to be a member. If X is included, a ring becomes a *field,* or synonymously an *algebra*. Since $X - A = A^c$, this amounts to including all complements and, in view of the de Morgan laws, specifying the inclusion of intersections and differences becomes redundant.

1.15 Definition A field $\mathcal{F}$ is a class of subsets of X satisfying

(a) $X \in \mathcal{F}$

(b) if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$

(c) if A and $B \in \mathcal{F}$ then $A \cup B \in \mathcal{F}$. □

A field is said to be *closed* under complementation and finite union and hence under intersections and differences too; none of these operations can take one outside the class.

These classes can be very complex and also very trivial. The simplest case of a ring is $\{\varnothing\}$. The smallest possible field is $\{X, \varnothing\}$. Scarcely less trivial is the field $\{X, A, A^c, \varnothing\}$, where A is any subset of X. What makes any class of sets interesting, or not, is the collection $\mathcal{C}$ of sets it is declared to contain, the 'seed' for the class. The smallest field containing $\mathcal{C}$ is called 'the field generated by $\mathcal{C}$'.

Rings and fields are natural classes in the sense of being defined in terms of the simple set operations, but their structure is rather restrictive for some of the applications in probability. More inclusive definitions, carefully tailored to include some important cases, are as follows.

1.16 Definition A *semi-ring* $\mathcal{S}$ is a nonempty class of subsets of X satisfying

(a) $\varnothing \in \mathcal{S}$

(b) if $A, B \in \mathcal{S}$ then $A \cap B \in \mathcal{S}$

(c) if $A, B \in \mathcal{S}$ and $A \subseteq B$, $\exists\ n < \infty$ such that $B - A = \bigcup_{j=1}^{n} C_j$, where $C_j \in \mathcal{S}$ and $C_j \cap C_{j'} = \varnothing$ for each $j \neq j'$. □

(Note that $\exists$ is the ubiquitous shorthand for "there exists".) More succinctly, condition (c) says that the difference of two $\mathcal{S}$-sets has a finite partition into $\mathcal{S}$-sets.

1.17 Definition A *semi-algebra* $\mathcal{S}$ is a class of subsets of X satisfying

(a) $X \in \mathcal{S}$

(b) if $A, B \in \mathcal{S}$ then $A \cap B \in \mathcal{S}$

(c) if $A \in \mathcal{S}$, $\exists\ n < \infty$ such that $A^c = \bigcup_{j=1}^{n} C_j$, where $C_j \in \mathcal{S}$ and $C_j \cap C_{j'} = \varnothing$ for each $j \neq j'$. □

A semi-ring containing X is a semi-algebra.

1.18 Example Let $X = \mathbb{R}$ and consider the class of all the half-open intervals $I = (a, b]$ for $-\infty < a < b < +\infty$, together with the empty set. If $I_1 = (a_1, b_1]$ and $I_2 = (a_2, b_2]$, then $I_1 \cap I_2$ is one of I_1, I_2, $(a_1, b_2]$, $(a_2, b_1]$, and $\varnothing$. Also if $I_1 \subseteq I_2$ so that $a_2 \leq a_1$ and $b_1 \leq b_2$, then $I_2 - I_1$ is one of $\varnothing$, $(a_2, a_1]$, $(b_1, b_2]$, $(a_2, a_1] \cup (b_1, b_2]$, and I_2. The conditions defining a semi-ring are therefore satisfied, although not those defining a ring. Now let $\mathbb{R}$ be a member of the class and follow **1.17**. The half-open intervals, plus the unbounded intervals of the form $(-\infty, b]$ and $(a, +\infty)$, plus $\varnothing$ and $\mathbb{R}$, constitute a semi-algebra. □

1.7 Sigma Fields

A field contains the complements and finite unions and the qualifier *finite* deserves explanation. It is clear that $A_1, \dots, A_n \in \mathcal{F}$ implies that $\bigcup_{j=1}^{n} A_j \in \mathcal{F}$ by a simple n-fold iteration of pairwise union. But, given the constructive nature of the definition, it is not legitimate without a further stipulation to assume that such an operation can be taken to the limit. Making this additional stipulation is what defines a σ-field.

1.19 Definition A σ-field (σ-algebra) $\mathcal{F}$ is a class of subsets of X satisfying

(a) $X \in \mathcal{F}$

(b) if $A \in \mathcal{F}$ then $A^c \in \mathcal{F}$

(c) if $\{A_n, n \in \mathbb{N}\}$ is a sequence of $\mathcal{F}$-sets, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$. □

A σ-field is closed under the operations of complementation and *countable* union and hence, by the de Morgan laws, of countable intersection also. A σ-ring can be defined similarly, although this is not a concept needed in the sequel. Given a collection of sets $\mathcal{C}$, the intersection of all the σ-fields containing $\mathcal{C}$ is called the σ-field *generated* by $\mathcal{C}$, customarily denoted $\sigma(\mathcal{C})$.

The following theorem establishes a basic fact about σ-fields.

1.20 Theorem If $\mathcal{C}$ is a finite collection $\sigma(\mathcal{C})$ is finite, otherwise $\sigma(\mathcal{C})$ is always uncountable.

Proof Define the relation R between elements of X by 'xRy iff x and y are elements of the same sets of $\mathcal{C}$'. R is an equivalence relation and hence defines an equivalence class $\mathcal{E}$ of disjoint subsets. Each set of $\mathcal{E}$ is the intersection of all the $\mathcal{C}$-sets containing its elements and the complements of the remainder. (For example, see Figure 1.1. For this collection of regions of $\mathbb{R}^2$, $\mathcal{E}$ is the partition defined by the complete network of set boundaries.) If $\mathcal{C}$ contains n sets, $\mathcal{E}$ contains at most 2^n sets and $\sigma(\mathcal{C})$, in this case the collection of all unions of $\mathcal{E}$-sets, contains at most 2^{2^n} sets. This proves the first part of the theorem.

Let $\mathcal{C}$ be infinite. If it is uncountable then so is $\sigma(\mathcal{C})$ and there is nothing more to show, so assume $\mathcal{C}$ is countable. In this case every set in $\mathcal{E}$ is a countable intersection of $\mathcal{C}$-sets or the complements of $\mathcal{C}$-sets, hence $\mathcal{E} \subseteq \sigma(\mathcal{C})$ and hence also $\mathcal{U}(\mathcal{E}) \subseteq \sigma(\mathcal{C})$, where $\mathcal{U}(\mathcal{E})$ is the collection of all the countable unions of $\mathcal{E}$-sets. If $\mathcal{U}(\mathcal{E})$ is uncountable the same will be true of $\sigma(\mathcal{C})$. Assume that $\mathcal{E}$ is countable since otherwise there is nothing more to show. So let the sets of $\mathcal{E}$ be indexed by $\mathbb{N}$. Then every union of $\mathcal{E}$-sets corresponds uniquely with a subset of $\mathbb{N}$ and every subset of $\mathbb{N}$ corresponds uniquely to a union of $\mathcal{E}$-sets. In other words, the elements of $\mathcal{U}(\mathcal{E})$

are equipotent with those of $2^{\mathbb{N}}$, which are uncountable by **1.13**. This completes the proof. ∎

1.21 Example Let $X = \mathbb{R}$ and let $\mathcal{C} = \{(-\infty, r],\ r \in \mathbb{Q}\}$, the collection of *closed half-lines with rational endpoints.* $\sigma(\mathcal{C})$ is called the *Borel field* of $\mathbb{R}$, generally denoted $\mathcal{B}$. A number of different base collections generate $\mathcal{B}$. Since countable unions of open intervals can be closed intervals and vice versa (compare **1.12**), the set of open half-lines, $\{(-\infty, r),\ r \in \mathbb{Q}\}$, will also serve. Or, letting $\{r_n\}$ be a decreasing sequence of rational numbers with $r_n \downarrow x$,

$$(-\infty, x] = \bigcap_{n=1}^{\infty} (-\infty, r_n]. \tag{1.21}$$

Such a sequence exists for any $x \in \mathbb{R}$ (see **2.5**) and hence the same σ-field is generated by the (uncountable) collection of half-lines with real endpoints, $\{(-\infty, x],\ x \in \mathbb{R}\}$. It easily follows that various other collections generate $\mathcal{B}$, including the open intervals of $\mathbb{R}$, the closed intervals, and the half-open intervals. □

1.22 Example Let $X = \bar{\mathbb{R}}$, the extended real line. The Borel field of $\bar{\mathbb{R}}$ is easily given. It is

$$\bar{\mathcal{B}} = \{B, B \cup \{+\infty\}, B \cup \{-\infty\}, B \cup \{+\infty\} \cup \{-\infty\} : B \in \mathcal{B}\}$$

where $\mathcal{B}$ is the Borel field of $\mathbb{R}$. $\bar{\mathcal{B}}$ is the σ-field generated by the collection $\mathcal{C}$ of **1.21** augmented by the sets $\{-\infty\}$ and $\bar{\mathbb{R}}$. □

1.23 Example Given an interval I of the line, the class $\mathcal{B}_I = \{B \cap I : B \in \mathcal{B}\}$ is called the restriction of $\mathcal{B}$ to I, or the Borel field on I. In fact, $\mathcal{B}_I$ is the σ-field generated from the collection $\mathcal{C} = \{(-\infty, r] \cap I : r \in \mathbb{Q}\}$. □

1.24 Example Consider the Cartesian product $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$, defining what is often called the *Euclidean plane*. $\mathcal{B}^2$, the Borel field of $\mathbb{R}^2$, is the σ-field generated from the collection of 'half-planes' $\{(-\infty, x] \times (-\infty, y] : x \in \mathbb{Q}, y \in \mathbb{Q}\}$. Other collections generating $\mathcal{B}^2$ include the rectangles (Cartesian products of the intervals), open, closed, or half-open as the case may be and with rational or real end points. Any region of the plane that can be constructed as a countable union of rectangles is a Borel set, including circles, ellipses and geometrical figures of arbitrary form. □

The last example has a natural generalization to $\mathcal{B}^k$, the Borel field of the space of real k-tuples $\mathbb{R}^k$.

Notice how $\sigma(\mathcal{C})$ has been defined 'from the outside'. It might be thought that $\sigma(\mathcal{C})$ could be defined 'from the inside', in terms of a specified sequence of the operations of complementation and countable union applied to the elements of $\mathcal{C}$. But, despite the constructive nature of the definitions, **1.20** suggests how this may be impossible. Suppose $\mathcal{A}_1$ is the set that contains $\mathcal{C}$ together with the complement of every set in $\mathcal{C}$ and all the finite and countable unions of the sets of $\mathcal{C}$. Of course, $\mathcal{A}_1$ is not $\sigma(\mathcal{C})$ because it does not contain the complements of the unions. So let $\mathcal{A}_2$ be the set containing $\mathcal{A}_1$ together with all the complements and finite and countable unions of the sets in $\mathcal{A}_1$. Defining $\mathcal{A}_3, \mathcal{A}_4, \ldots$ in the same manner, it might be thought that the monotone sequence $\{\mathcal{A}_n\}$ would approach $\sigma(\mathcal{C})$ as $n \to \infty$; but in fact this is not so. In the case of the class $\mathcal{B}_{[0,1]}$, for example, it can be shown that $\mathcal{A}_\infty$ is strictly smaller than $\sigma(\mathcal{C})$ (see Billingsley [23], pp. 32–34). On the other hand, $\sigma(\mathcal{C})$ may be smaller than 2^X. This fact is demonstrated, again for $\mathcal{B}_{[0,1]}$, in §3.4.

The union of two σ-fields (the set of elements contained in either or both of them) is not generally a σ-field, for the unions of the sets from one field with those from the other are not guaranteed to belong to it. The concept of union for σ-fields is therefore extended by adding in these sets. Given σ-fields $\mathcal{F}$ and $\mathcal{G}$, the smallest σ-field containing all the elements of $\mathcal{F}$ and all the elements of $\mathcal{G}$ is denoted $\mathcal{F} \vee \mathcal{G}$, called the union of $\mathcal{F}$ and $\mathcal{G}$. On the other hand, $\mathcal{F} \cap \mathcal{G} = \{A : A \in \mathcal{F} \text{ and } A \in \mathcal{G}\}$ is a σ-field, although for uniformity the notation $\mathcal{F} \wedge \mathcal{G}$ may be used for such intersections. Formally, $\mathcal{F} \wedge \mathcal{G}$ denotes the largest of the σ-fields whose elements belong to both $\mathcal{F}$ and $\mathcal{G}$. Both of these operations generalize to the countable case, so that for a sequence of σ-fields $\mathcal{F}_n$, $n = 1, 2, 3, \ldots$ the notations $\bigvee_{n=1}^\infty \mathcal{F}_n$ and $\bigcap_{n=1}^\infty \mathcal{F}_n$ are defined.

Without going prematurely into too many details, it can be said that a large part of the intellectual labour in probability and measure theory is devoted to proving that particular classes of sets are σ-fields. Problems of this kind arise throughout this book. It is usually not too hard to show that $A^c \in \mathcal{F}$ whenever $A \in \mathcal{F}$, but the requirement to show that a class contains the countable unions can be tough to fulfil. The following material can be helpful in this connection.

A *monotone class* $\mathcal{M}$ is a class of sets such that, if $\{A_n\}$ is a monotone sequence with limit A and $A_n \in \mathcal{M}$ for all n, then $A \in \mathcal{M}$. If $\{A_n\}$ is non-decreasing, then $A = \bigcup_{n=1}^\infty A_n$. If it is non-increasing, then $A = \bigcap_{n=1}^\infty A_n$. The next theorem shows that to determine whether a set is a σ-field it is sufficient to consider whether the limits of monotone sequences belong to it, which should often be easier to establish than the general case.

1.25 Theorem $\mathcal{F}$ is a σ-field iff it is both a field and a monotone class.

Proof The 'only if' part of the theorem is immediate. For the 'if' part, define $A_n = \bigcup_{m=1}^{n} E_m$, for any sequence $\{E_m \in \mathcal{F}, m \in \mathbb{N}\}$. Since $\mathcal{F}$ is a field, $A_n \in \mathcal{F}$ for any finite n. But $\{A_n, n \in \mathbb{N}\}$ is a monotone sequence with limit $\bigcup_{n=1}^{\infty} A_n \in \mathcal{F}$, by assumption. $\bigcup_{n=1}^{\infty} A_n = \bigcup_{m=1}^{\infty} E_m$, so the theorem follows. ∎

A useful trick that plays a special role in probability theory is Dynkin's π-λ theorem. To develop this result, define two new classes of subsets of X.

1.26 Definition A class $\mathcal{L}$ is a λ-system (a Dynkin system) if

(a) $X \in \mathcal{L}$

(b) if A and $B \in \mathcal{L}$ and $B \subseteq A$, then $A - B \in \mathcal{L}$

(c) if $\{A_n \in \mathcal{L}\}$ is a non-decreasing sequence and $A_n \uparrow A$, then $A \in \mathcal{L}$. □

Conditions (a) and (b) imply that a λ-system is closed under complementation (put $A = X$). Moreover, since (b) implies that $B_n = A_{n+1} - A_n \in \mathcal{L}$ for each n, (c) implies that a countable union of *disjoint* $\mathcal{L}$-sets is in $\mathcal{L}$. In fact, these implications hold in both directions.

1.27 Theorem A class $\mathcal{L}$ is a λ-system iff

(a) $X \in \mathcal{L}$

(b) if $B \in \mathcal{L}$ then $B^c \in \mathcal{L}$

(c) if $\{A_n \in \mathcal{L}\}$ is a disjoint sequence with $A_n \cap A_m = \emptyset$ for $n \neq m$, then $\bigcup_{n=1}^{\infty} A_n \in \mathcal{L}$. □

In particular, note that a σ-field is a λ-system, since the last condition of closure under countable union holds for general sequences in that case, not just disjoint sequences.

1.28 Definition A class $\mathcal{P}$ is a π-system if A and $B \in \mathcal{P}$ implies $A \cap B \in \mathcal{P}$. □

A class that is both a π-system and a λ-system is in fact a σ-field. This follows by **1.25**, because a λ-system is a monotone class by **1.26**(c) and by de Morgan's laws is closed under unions if closed under both intersections and complementation.

Dynkin's π-λ theorem, which is as follows, provides a neat way of showing a class of sets is a σ-field given some more easily established properties.

1.29 Theorem If $\mathcal{P}$ is a π-system, $\mathcal{L}$ is a λ-system, and $\mathcal{P} \subseteq \mathcal{L}$, then $\sigma(\mathcal{P}) \subseteq \mathcal{L}$.

Proof Let $\lambda(\mathcal{P})$ denote the smallest λ-system containing $\mathcal{P}$ (the intersection of all the λ-systems containing $\mathcal{P}$), so that in particular, $\lambda(\mathcal{P}) \subseteq \mathcal{L}$. The object is to show $\lambda(\mathcal{P})$ is a π-system. By the remarks above, it will then follow that $\lambda(\mathcal{P})$ is a σ-field and hence that $\sigma(\mathcal{P}) \subseteq \lambda(\mathcal{P}) \subseteq \mathcal{L}$, as required.

For a set $A \in \lambda(\mathcal{P})$, let $\mathcal{G}_A$ denote the class of sets B such that $A \cap B \in \lambda(\mathcal{P})$. $\mathcal{G}_A$ is a λ-system. To show this, first note $X \in \mathcal{G}_A$ so that condition **1.26**(a) is satisfied. Let $B_1, B_2 \in \mathcal{G}_A$ and $B_1 \subset B_2$; then $A \cap B_1 \in \lambda(\mathcal{P})$ and $A \cap B_2 \in \lambda(\mathcal{P})$ and $(A \cap B_1) \subset (A \cap B_2)$, which implies that

$$(A \cap B_2) - (A \cap B_1) = A \cap (B_2 - B_1) \in \lambda(\mathcal{P}). \tag{1.22}$$

But this means that $B_2 - B_1 \in \mathcal{G}_A$ and condition **1.26**(b) is satisfied. Lastly, suppose $A \cap B_i \in \lambda(\mathcal{P})$ for each $i = 1, 2, \ldots$ and $B_i \uparrow B$. Then $A \cap B \in \lambda(\mathcal{P})$ by **1.26**(c), which means that **1.26**(c) holds for $\mathcal{G}_A$ and $\mathcal{G}_A$ is a λ-system as asserted.

Suppose $A \in \mathcal{P}$. Then $B \in \mathcal{P}$ implies $A \cap B \in \mathcal{P}$ ($\mathcal{P}$ is a π-system) and since $\mathcal{P} \subseteq \lambda(\mathcal{P})$, this further implies $B \in \mathcal{G}_A$. Hence $\mathcal{P} \subseteq \mathcal{G}_A$. Since $\mathcal{G}_A$ is a λ-system and $\lambda(\mathcal{P})$ is the smallest λ-system containing $\mathcal{P}$, $\lambda(\mathcal{P}) \subseteq \mathcal{G}_A$ in this case. So, when $A \in \mathcal{P}$, $B \in \lambda(\mathcal{P})$ implies $B \in \mathcal{G}_A$ and hence $A \cap B \in \lambda(\mathcal{P})$.

Summarize the last conclusion as:

$$\{A \in \mathcal{P}, B \in \lambda(\mathcal{P})\} \Rightarrow \{A \cap B \in \lambda(\mathcal{P})\}. \tag{1.23}$$

Now define $\mathcal{G}_B$ by analogy with $\mathcal{G}_A$ so that

$$\{B \in \lambda(\mathcal{P}), A \cap B \in \lambda(\mathcal{P})\} \Rightarrow \{A \in \mathcal{G}_B\} \tag{1.24}$$

and (1.23) and (1.24) together yield $\mathcal{P} \subseteq \mathcal{G}_B$. Since $\mathcal{G}_B$ is also a λ-system by the same argument as held for $\mathcal{G}_A$ and contains $\mathcal{P}$, $\lambda(\mathcal{P}) \subseteq \mathcal{G}_B$ by definition of $\lambda(\mathcal{P})$.

Thus, suppose $B \in \lambda(\mathcal{P})$ and $C \in \lambda(\mathcal{P})$. Then $C \in \mathcal{G}_B$, which means that $B \cap C \in \lambda(\mathcal{P})$. So $\lambda(\mathcal{P})$ is a π-system as required. ∎

1.8 The Topology of the Real Line

This section revisits the subject matter of §1.4 to treat rigorously the idea of 'nearness' as it applies to points of the line. The distance between a pair of points, $x, y \in \mathbb{R}$, is defined as the non-negative real number $|x - y|$, what is formally called the *Euclidean* distance. Chapters 5 and 6 examine the generalization of this theory to non-Euclidean spaces and find not only that most aspects of the theory have a natural generalization but also that the concept of distance itself can be dispensed with in their development. The fact that this is a special case of a very powerful general theory may be helpful in making sense of certain ideas, the definition of compactness for example, that can otherwise appear a little puzzling at first sight.

An *ε-neighbourhood* of a point $x \in \mathbb{R}$ is a set $S(x, \varepsilon) = \{y : |x - y| < \varepsilon\}$, for some $\varepsilon > 0$. An *open set* is a set $A \subseteq \mathbb{R}$ such that for each $x \in A$ there exists for some $\varepsilon > 0$ an ε-neighbourhood which is a subset of A. The open intervals defined in §1.4 are open sets since if $a < x < b$, $\varepsilon = \min\{|b - x|, |a - x|\} > 0$ satisfies the definition. $\mathbb{R}$ and $\varnothing$ are also open sets on the definition.

The concept of an open set is subtle and often gives beginners some difficulty. Naïve intuition strongly favours the notion that in any bounded set of points there ought to be one that is 'next to' a point outside the set. But open sets are sets that do not have this property and there is no shortage of them in $\mathbb{R}$. A complete understanding of the issues calls for the additional concepts of *Cauchy sequence* and *limit,* to appear in §2.1 below. Doubters are invited to suspend their disbelief for the moment and just take the definition at face value.

The collection of all the open sets of $\mathbb{R}$ is known as the *topology* of $\mathbb{R}$. More precisely this is the *usual topology* of $\mathbb{R}$ since other ways of defining open sets of $\mathbb{R}$ can be devised, although these will not concern us. (See Chapter 6 for more information on these matters.) More generally, subsets of $\mathbb{R}$ can be discussed from a topological standpoint although the term *subspace* rather than subset is usually preferred in this context. If $A \subseteq \mathbb{S} \subseteq \mathbb{R}$, A is said to be *open in* $\mathbb{S}$ if for each $x \in A$ there exists $S(x, \varepsilon)$, $\varepsilon > 0$, such that $S(x, \varepsilon) \cap \mathbb{S}$ is a subset of A. Thus, the interval $[0, \frac{1}{2})$ is not open in $\mathbb{R}$, but it is open in $[0, 1]$. These sets define the *relative topology* on $\mathbb{S}$, that is, the topology on $\mathbb{S}$ relative to $\mathbb{R}$. The following result is an immediate consequence of the definition.

1.30 Theorem If A is open in $\mathbb{R}$, $A \cap \mathbb{S}$ is open in the relative topology on $\mathbb{S}$. □

A *closure point* of a set A is a point $x \in \mathbb{R}$ such that for every $\varepsilon > 0$ the set $A \cap S(x, \varepsilon)$ is not empty. The closure points of A are not necessarily elements of A, open sets being a case in point. The set of closure points of A is called the closure of A and will be denoted $\bar{A}$, or sometimes $(A)^-$ if the set is defined by an expression. On the other hand, an *accumulation point* of A is a point $x \in \mathbb{R}$ which is a closure point of the set $A - \{x\}$. An accumulation point has other points of A arbitrarily close to it, and if x is a closure point of A and $x \notin A$ it must also be an accumulation point. A closure point that is not an accumulation point (the former definition being satisfied because each ε-neighbourhood of x contains x itself) is an *isolated* point of A.

A *boundary point* of a set A is a point $x \in \bar{A}$ such that the set $A^c \cap S(x, \varepsilon)$ is not empty for any $\varepsilon > 0$. The set of boundary points of A is denoted ∂A and $\bar{A} = A \cup \partial A$. The *interior* of A is the set $A^o = A - \partial A$. A *closed* set is one containing all its closure points, that is, a set A such that $\bar{A} = A$. For an open interval $A = (a, b) \subset \mathbb{R}$,

$\bar{A} = [a, b]$. Every point of (a, b) is a closure point and a and b are also closure points, not belonging to (a, b). They are the boundary points of both (a, b) and $[a, b]$.

1.31 Theorem The complement of an open set in $\mathbb{R}$ is a closed set. □

This gives an alternative definition of a closed set. According to the definitions, $\emptyset$ (the empty set) and $\mathbb{R}$ are both open *and* closed. The half-line $(-\infty, x]$ is the complement of the open set $(x, +\infty)$ and is hence closed. This result extends to relative topologies:

1.32 Theorem If A is open in $\mathbb{S} \subset \mathbb{R}$, then $\mathbb{S} - A$ is closed in $\mathbb{S}$. □

In particular, a corollary to **1.30** is that if B is closed in $\mathbb{R}$ then $\mathbb{S} \cap B$ is closed in $\mathbb{S}$. But, for example, the interval $[\frac{1}{2}, 1)$ is not closed in $\mathbb{R}$ although it is closed in the set $(0, 1)$ since its complement $(0, \frac{1}{2})$ is open in $(0, 1)$.

Some additional properties of open sets are given in the following theorems.

1.33 Theorem

(i) The union of a collection of open sets is open.
(ii) If A and B are open, then $A \cap B$ is open. □

This result is proved in a more general context below, as **5.4**. Arbitrary intersections of open sets *need not* be open. See **1.12** for a counterexample.

1.34 Theorem Every open set $A \subseteq \mathbb{R}$ is the union of a countable collection of disjoint open intervals.

Proof Consider a collection $\{S(x, \varepsilon_x), x \in A\}$ where for each x, $\varepsilon_x > 0$ is chosen small enough that $S(x, \varepsilon_x) \subseteq A$. Then $\bigcup_{x \in A} S(x, \varepsilon_x) \subseteq A$, but, since necessarily $A \subseteq \bigcup_{x \in A} S(x, \varepsilon_x)$, it follows that $\bigcup_{x \in A} S(x, \varepsilon_x) = A$. This shows that A is a union of open intervals.

Now define a relation R for elements of A, such that xRy if there exists an open interval $I \subseteq A$ with $x \in I$ and $y \in I$. Every $x \in A$ is contained in some interval by the preceding argument, so that xRx for all $x \in A$. The symmetry of R is obvious. Lastly, if $x, y \in I \subseteq A$ and $y, z \in I' \subseteq A$, $I \cap I'$ is nonempty and hence $I \cup I'$ is also an open interval, so R is transitive. Hence R is an equivalence relation and the intervals I are an equivalence class partitioning A. Thus, A is union of disjoint open intervals. The theorem now follows from **1.11**. ∎

Recall from **1.21** that $\mathcal{B}$, the Borel field of $\mathbb{R}$, is the σ-field of sets generated by both the open and the closed half-lines. Since every interval is the intersection of a half-line (open or closed) with the complement of another half-line, **1.31** and **1.34** yield directly the following important fact.

1.35 Theorem $\mathcal{B}$ contains the open sets and the closed sets of $\mathbb{R}$. □

A collection $\mathcal{C}$ is called a *covering* for a set $A \subseteq \mathbb{R}$ if $A \subseteq \bigcup_{B \in \mathcal{C}} B$. If each B is an open set, it is called an *open covering*. It's important to appreciate the subtle implications of this rather simple idea. There are various ways to cover a set of which perhaps the most natural is a partition. Consider Figure 1.1 where the eight regions defined by the various set boundaries jointly cover X. However, this is obviously not an open covering since it involves complements and the complements of open sets are closed. The boundary points have to go somewhere! An open covering contains no boundary points so the collection of covering sets cannot be disjoint.

The following is *Lindelöf's covering theorem.*

1.36 Theorem If $\mathcal{C}$ is any collection of open subsets of $\mathbb{R}$, there is a countable subcollection $\{B_i \in \mathcal{C}, i \in \mathbb{N}\}$ such that

$$\bigcup_{B \in \mathcal{C}} B = \bigcup_{i=1}^{\infty} B_i. \tag{1.25}$$

Proof Consider the collection $\mathcal{S} = \{S_k = S(r_k, s_k), r_k \in \mathbb{Q}, s_k \in \mathbb{Q}^+\}$; that is, the collection of all neighbourhoods of rational points of $\mathbb{R}$ having rational radii. The set $\mathbb{Q} \times \mathbb{Q}^+$ is countable by **1.5** and hence $\mathcal{S}$ is countable; in other words, indexing by $k \in \mathbb{N}$ exhausts the set. For any open set $B \subseteq \mathbb{R}$ and point $x \in B$ there is a set $S_k \in \mathcal{S}$ such that $x \in S_k \subseteq B$. Since x has an ε-neighbourhood inside B by definition, the desired S_k is found by setting s_k to any rational from the open interval $(0, \frac{1}{2}\varepsilon)$ for $\varepsilon > 0$ sufficiently small and then choosing $r_k \in S(x, \frac{1}{4}\varepsilon)$, as is possible by **1.10**.

For each $x \in \bigcup_{B \in \mathcal{C}} B$ choose a member of $\mathcal{S}$, say $S_{k(x)}$, satisfying $x \in S_{k(x)} \subseteq B$ for some $B \in \mathcal{C}$. Letting $k(x)$ be the smallest index that satisfies the requirement gives an unambiguous choice. The distinct members of this collection form a set that covers $\bigcup_{B \in \mathcal{C}} B$, but is a subset of $\mathcal{S}$ and hence countable. Labelling the indices of this set as $k_1, k_2, \ldots$, choose B_i as any member of $\mathcal{C}$ containing S_{k_i}. Clearly, $\bigcup_{i=1}^{\infty} B_i$ is a countable covering for $\bigcup_{i=1}^{\infty} S_{k_i}$ and hence also for $\bigcup_{B \in \mathcal{C}} B$. ∎

It follows that if $\mathcal{C}$ is a covering for a set in $\mathbb{R}$ it contains a countable *subcovering*. This is sometimes called the Lindelöf property.

The concept of a covering leads on to the crucial notion of *compactness.* A set A is said to be *compact* if every open covering of A contains a *finite* subcovering. The words that matter in this definition are 'every' and 'open'. Any open covering that has $\mathbb{R}$ as a member obviously contains a finite subcovering. But for a set to be compact, there must be no way to construct an irreducible, infinite, open covering. Moreover, every interval has an irreducible infinite cover, consisting of the singleton sets of its individual points; but these sets are not open.

1.37 Example Consider the half-open interval $(0, 1]$. An open covering is the countable collection $\{(1/n, 1], n \in \mathbb{N}\}$. It is easy to see that there is no finite subcollection covering $(0, 1]$ in this case, so $(0, 1]$ is not compact. □

A set A is *bounded* if $A \subseteq S(x, \varepsilon)$ for some $x \in A$ and $\varepsilon > 0$. The idea here is that ε is a possibly large but finite number. In other words, a bounded set must be containable within a finite interval. The following fundamental result, known as the *Heine–Borel theorem*, provides an alternative definition of compactness in $\mathbb{R}$. The proof is found as a case of **5.12** below.

1.38 Theorem A set in $\mathbb{R}$ is compact iff it is closed and bounded. □

A subset B of A is said to be *dense* in A if $B \subseteq A \subseteq \bar{B}$. Readers may think they know what is implied here after studying the following theorem, but denseness is a slightly tricky notion. See also **2.5** and the remarks following before coming to any premature conclusions.

1.39 Theorem Let A be an interval of $\mathbb{R}$ and $C \subseteq A$ be a countable set. Then $A - C$ is dense in A.

Proof By **1.7**, each neighbourhood of a point in A contains an uncountable number of points. Hence for each $x \in A$ (whether or not $x \in C$), the set $(A - C) \cap S(x, \varepsilon)$ is not empty for every $\varepsilon > 0$, so that x is a closure point of $A - C$. Thus,

$$A - C \subseteq (A - C) \cup C = A \subseteq \overline{A - C}. \quad \blacksquare$$

The k-fold Cartesian product of $\mathbb{R}$ with copies of itself generates what is called *Euclidean k-space*, $\mathbb{R}^k$. The points of $\mathbb{R}^k$ have the interpretation of k-vectors, or ordered k-tuples of real numbers, $\boldsymbol{x} = (x_1, x_2, \ldots, x_k)'$. All the concepts defined above for sets in $\mathbb{R}$ generalize directly to $\mathbb{R}^k$. The only modification required is to replace the scalars x and y by vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ and define an ε-neighbourhood in a new way. Let $||\boldsymbol{x}-\boldsymbol{y}||$ denote the Euclidean distance between $\boldsymbol{x}$ and $\boldsymbol{y}$ where $||\boldsymbol{a}|| = \left(\sum_{i=1}^{k} a_i^2\right)^{1/2}$ is the length (Euclidean norm) of the vector $\boldsymbol{a} = (a_1, \ldots, a_k)'$. Then define $S(\boldsymbol{x}, \varepsilon) = \{\boldsymbol{y} : ||\boldsymbol{x}-\boldsymbol{y}|| < \varepsilon\}$, for some $\varepsilon > 0$. An open set A of $\mathbb{R}^2$ is one in which every point $\boldsymbol{x} \in A$ can be contained in an open disk with positive radius centred on $\boldsymbol{x}$. In $\mathbb{R}^3$ the open disk becomes an open sphere; and so on.

2

Limits, Sequences, and Sums

2.1 Sequences and Limits

A *real sequence* is a mapping from $\mathbb{N}$ into $\mathbb{R}$. The elements of the domain are called the *indices* and those of the range variously the *terms, members,* or *coordinates* of the sequence. A sequence may be denoted $\{x_n, n \in \mathbb{N}\}$ or more briefly by $\{x_n\}_1^\infty$, or just by $\{x_n\}$ when the context is clear.

$\{x_n\}_1^\infty$ is said to *converge* to a limit x if for every $\varepsilon > 0$ there is an integer N_ε for which

$$|x_n - x| < \varepsilon \text{ for all } n > N_\varepsilon. \tag{2.1}$$

Write $x_n \to x$, or $x = \lim_{n\to\infty} x_n$. When a sequence is tending to $+\infty$ or $-\infty$ it is often said to *diverge*, but it may also be said to converge in $\bar{\mathbb{R}}$, to distinguish those cases when it does not approach *any* fixed value but is always wandering.

A sequence is *monotone* (non-decreasing, increasing, non-increasing, or decreasing) if the respective inequalities $x_n \le x_{n+1}$, $x_n < x_{n+1}$, $x_n \ge x_{n+1}$, or $x_n > x_{n+1}$ hold for every n. To indicate that a monotone sequence is converging, one may write for emphasis either $x_n \uparrow x$ or $x_n \downarrow x$, as appropriate, although $x_n \to x$ will also do in both cases. The following result does not require elaboration.

2.1 Theorem Every monotone sequence in a compact set converges. □

A sequence that does not converge may nonetheless visit the same point an infinite number of times, so exhibiting a kind of convergent behaviour. If $\{x_n, n \in \mathbb{N}\}$ is a real sequence, a *subsequence* is $\{x_{n_k}, k \in \mathbb{N}\}$ where $\{n_k, k \in \mathbb{N}\}$ is any increasing sequence of positive integers. If there exists a subsequence $\{x_{n_k}, k \in \mathbb{N}\}$ and a constant c such that $x_{n_k} \to c$ as $k \to \infty$, c is called a *cluster point* of the sequence. For example, the sequence $\{(-1)^n, n = 1, 2, 3, \ldots\}$ does not converge, but the subsequence obtained by taking only even values of n converges trivially. c is usually a finite constant, but allowing the notion of convergence in $\bar{\mathbb{R}}$, $+\infty$ and $-\infty$ may be cluster points. If a subsequence is convergent, then so is any subsequence *of* the subsequence, defined as $\{x_{m_k}, k \in \mathbb{N}\}$ where $\{m_k\}$ is an increasing sequence whose members are also members of $\{n_k\}$.

The concept of a subsequence is often useful in arguments concerning convergence. A typical line of reasoning employs a two-pronged attack; first identify a

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0002

convergent subsequence (a monotone sequence, perhaps); then use other characteristics of the sequence to show that the cluster point is actually a limit. Especially useful in this connection is the knowledge that the members of the sequence are points in a compact set. Such sequences cannot diverge to infinity, since the set is bounded; and because the set is closed, any limit points or cluster points that exist must be in the set.

The result that every bounded sequence in $\mathbb{R}$ has a cluster point is the commonly cited version of the *Bolzano–Weierstrass theorem*. Of course, it is in compact sets of $\mathbb{R}$ that sequences are necessarily bounded. The term *sequential compactness* is often used to describe sets having the Bolzano–Weierstrass property and in $\mathbb{R}$ this is equivalent to compactness, as the following theorem establishes.

2.2 Theorem Every sequence in a compact subset of $\mathbb{R}$ has at least one cluster point.

Proof A monotone sequence converges in a compact set by **2.1**, and every sequence $\{x_n, n \in \mathbb{N}\}$ has a monotone subsequence. Thus, define a subsequence $\{x_{n_k}\}$ by setting $n_1 = 1$ and for $k = 1,2,3,\ldots$ letting $x_{n_{k+1}} = \sup_{n \geq n_k} x_n$ if there exists a finite n_{k+1} satisfying this condition; otherwise let the subsequence terminate at n_k. This subsequence is non-increasing. If it terminates, the subsequence $\{x_n, n \geq n_k\}$ must contain a non-decreasing subsequence. A monotone subsequence therefore exists in every case. ∎

Extending the same line of reasoning, the following is one of the most useful attributes of compact sets.

2.3 Theorem A sequence in a compact subset of $\mathbb{R}$ either has two or more cluster points, or it converges.

Proof Suppose that c is the unique cluster point of the sequence $\{x_n\}$, but that $x_n \nrightarrow c$. Then there is an infinite set of integers $\{n_k, k \in \mathbb{N}\}$ such that $|x_{n_k} - c| \geq \varepsilon$ for some $\varepsilon > 0$. Define a sequence $\{y_k\}$ by setting $y_k = x_{n_k}$. Since $\{y_k\}$ is also a sequence on a compact set, it has a cluster point c' which by construction is different from c. But c' is also a cluster point of $\{x_n\}$, of which $\{y_k\}$ is a subsequence, which is a contradiction. Hence, $x_n \to c$. ∎

2.4 Example Consider the sequence of powers, $\{1, x, x^2, x^3, \ldots, x^n, \ldots\}$ or more formally $\{x^n, n \in \mathbb{N}_0\}$, where x is a real number. In the case $|x| < 1$, this sequence converges to zero, $\{|x^n|\}$ being monotone on the compact interval [0,1]. The condition specified in (2.1) is satisfied for $N_\varepsilon \geq \log(\varepsilon)/\log|x|$ in this case. If $x = 1$ it converges to 1, trivially. If $x > 1$ it diverges in $\mathbb{R}$, but converges in $\bar{\mathbb{R}}$ to $+\infty$.

If $x = -1$ it neither converges nor diverges, but oscillates between cluster points $+1$ and -1. Finally, if $x < -1$ the sequence diverges in $\mathbb{R}$, but does not converge in $\bar{\mathbb{R}}$. Ultimately, it oscillates between the cluster points $+\infty$ and $-\infty$. □

The following concepts characterize the asymptotic behaviour of a real sequence whether or not it has a limit. The *superior limit* of a sequence $\{x_n\}$ is

$$\limsup_n x_n = \inf_n \sup_{m \geq n} x_m. \tag{2.2}$$

(Alternative notation: $\overline{\lim}_n x_n$). The limsup is the *eventual* upper bound of a sequence. Think of $\{\sup_{m \geq n} x_m, n = 1, 2, \dots\}$ as the sequence of the largest values the sequence takes beyond the point n. This may be $+\infty$ for every n, but in all cases it must be a non-increasing sequence having a limit, either $+\infty$ or a finite real number; this limit is the limsup of the sequence. A link with the corresponding concept for set sequences is that if $x_n = \sup A_n$ for some sequence of sets $\{A_n \subseteq \mathbb{R}\}$, then limsup $x_n = \sup A$, where $A = \limsup_n A_n$. The *inferior limit* is defined likewise, as the eventual lower bound:

$$\liminf_n x_n = -\limsup_n (-x_n) = \sup_n \inf_{m \geq n} x_m \tag{2.3}$$

also written $\underline{\lim}_n x_n$. Necessarily, $\liminf_n x_n \leq \limsup_n x_n$. When the limsup and liminf of a sequence are equal the sequence is convergent and the limit is equal to their common value. If both equal $+\infty$, or $-\infty$, the sequence converges in $\bar{\mathbb{R}}$.

The usual application of these concepts is in arguments to establish the value of a limit. It may not be permissible to *assume* the existence of the limit, but the limsup and liminf always exist. The trick is to derive these and show them to be equal. For this purpose, it is sufficient in view of the above inequality to show $\liminf_n x_n \geq \limsup_n x_n$. This type of argument often appears in the sequel.

To determine whether a sequence converges it is not necessary to know what the limit is; the relationship between sequence coordinates 'in the tail' (as n becomes large) is sufficient for this purpose. The *Cauchy criterion* for convergence of a real sequence states that $\{x_n\}$ converges iff for every $\varepsilon > 0$ $\exists$ N_ε such that $|x_n - x_m| < \varepsilon$ whenever $n > N_\varepsilon$ and $m > N_\varepsilon$. A sequence satisfying this criterion is called a *Cauchy sequence.* Any sequence satisfying (2.1) is a Cauchy sequence and, conversely, a real Cauchy sequence must possess a limit in $\mathbb{R}$. The two definitions are therefore equivalent (in $\mathbb{R}$, at least) but the Cauchy condition may be easier to verify in practice.

The limit of a Cauchy sequence whose members all belong to a set A is by definition a closure point of A, though it need not itself belong to A. Conversely, for every accumulation point x of a set A there must exist a Cauchy sequence in the set whose limit is x. Construct such a sequence by taking one point from each of the sequence of sets,

$$\{A \cap S(x, 1/n), n = 1, 2, 3, \ldots\}$$

none of which are empty by definition. The term *limit point* is sometimes used synonymously with accumulation point.

The following is a fundamental property of the reals.

2.5 Theorem Every real number is the limit of a Cauchy sequence of rationals.

Proof For finite n let x_n be a number whose decimal expansion consists only of zeros beyond the n^{th} place in the sequence. If the decimal point appears at position m, with $m > n$, then x_n is an integer. If $m \leq n$, removing the decimal point produces a finite integer a and $x_n = a/10^{n-m}$, so x_n is rational. Given any real x a sequence of rationals $\{x_n\}$ is obtained by replacing with a zero every digit in the decimal expansion of x beyond the n^{th}, for $n = 1, 2, \ldots$ Since $|x_{n+1} - x_n| < 10^{-n}$, $\{x_n\}$ is a Cauchy sequence and $x_n \to x$ as $n \to \infty$. ∎

The sequence exhibited is increasing, but a decreasing sequence can also be constructed, as $\{-y_n\}$ where $\{y_n\}$ is an increasing sequence tending to $-x$. If x is itself rational this construction works by putting $x_n = x$ for every n, which trivially defines a Cauchy sequence, but certain arguments such as in **2.6** below depend on having $x_n \neq x$ for every n. To satisfy this requirement choose the 'non-terminating' representation of the number; for example, instead of 1 take 0.9999999...and consider the sequence {0.9, 0.99, 0.999,...}. This does not work for the point 0, but then one can choose {0.1, 0.01, 0.001,...}.

One interesting corollary of **2.5** is that, since every ε-neighbourhood of a real number must contain a rational, $\mathbb{Q}$ is dense in $\mathbb{R}$. It was also shown in **1.39** that $\mathbb{R} - \mathbb{Q}$ is dense in $\mathbb{R}$ since $\mathbb{Q}$ is countable. Take care not to jump to the conclusion that because a set is dense its complement must be 'sparse'.

Another version of this proof, at least for points of the interval $[0, 1]$, is got by using the binary expansion of a real number. The *dyadic rationals* are the set

$$\mathbb{D} = \{i/2^n, i = 1, \ldots 2^n - 1, n \in \mathbb{N}\}. \tag{2.4}$$

The dyadic rationals corresponding to a finite n define a covering of $[0, 1]$ by intervals of width $1/2^n$, which are bisected each time n is incremented. For any $x \in [0, 1]$, a point of the set $\{i/2^n, i = 1, \ldots, 2^n - 1\}$ is contained in $S(x, \varepsilon\}$ for $\varepsilon < 2/2^n$ and the set defined by (2.4) includes these points for every finite n. The dyadic rationals are therefore dense in $[0, 1]$, although they are still rational numbers and of a smaller order of infinity than the points of $[0, 1]$ itself. Since it defines the limit of a sequence of finite partitions of an interval, $\mathbb{D}$ is a convenient analytic tool that will often appear in the sequel.

Another useful application of these ideas is to set limits in $\mathbb{R}$.

2.6 Theorem Every open interval is the limit of a sequence of closed subintervals with rational endpoints.

Proof If (a, b) is the interval, with $a < b$, choose Cauchy sequences of rationals $a_n \downarrow a$ and $b_n \uparrow b$, with $a_1 < b_1$ (always possible by **1.10**). By definition, for every $x \in (a, b)$ there exists $N \geq 1$ such that $x \in [a_n, b_n]$ for all $n \geq N$ and hence $(a, b) \subseteq \liminf_n [a_n, b_n]$ by the definition in §1.5. On the other hand, since $a_n > a$ and $b > b_n$, $(a, b)^c \subseteq [a_n, b_n]^c$ for all $n \geq 1$, so that $(a, b)^c \subseteq \liminf_n [a_n, b_n]^c$. This is equivalent by de Morgan's laws to $\limsup_n [a_n, b_n] \subseteq (a, b)$. Hence $\lim_n [a_n, b_n]$ exists and is equal to (a, b). ∎

This shows that the limits of sequences of open sets need not be open, nor the limits of sequences of closed sets closed (take complements above). The only hard and fast rules are the following corollaries of **1.33**(i): the limit of a non-decreasing sequence of open sets is open and (by complements) the limit of a non-increasing sequence of closed sets is closed.

2.2 Functions and Continuity

A *function* of a real variable is a mapping $f : \mathbb{S} \mapsto \mathbb{T}$, where $\mathbb{S} \subseteq \mathbb{R}$ and $\mathbb{T} \subseteq \mathbb{R}$. Specifying a subset of $\mathbb{R}$ as the codomain implies without loss of generality that $f(\mathbb{S}) = \mathbb{T}$, such that the mapping is *onto* $\mathbb{T}$.

Consider the image in $\mathbb{T}$, under f, of a Cauchy sequence $\{x_n\}$ in $\mathbb{S}$ converging to x. If the image of every such sequence converging to $x \in \mathbb{S}$ is a Cauchy sequence in $\mathbb{T}$ converging to $f(x)$, the function is said to be *continuous* at x. Continuity is formally defined, without invoking sequences explicitly, using the $\varepsilon - \delta$ approach. f is continuous at the point $x \in \mathbb{S}$ if for any $\varepsilon > 0$ $\exists$ $\delta > 0$ such that $|y - x| < \delta$ implies $|f(y) - f(x)| < \varepsilon$ whenever $y \in \mathbb{S}$. The choice of δ here may depend on x. If f is continuous at every point of $\mathbb{S}$, it is simply said to be continuous on $\mathbb{S}$.

Perhaps the chief reason why continuity matters is the following result.

2.7 Theorem If $f : \mathbb{S} \mapsto \mathbb{T}$ is continuous at all points of $\mathbb{S}$, $f^{-1}(A)$ is open in $\mathbb{S}$ whenever A is open in $\mathbb{T}$ and $f^{-1}(A)$ is closed in $\mathbb{S}$ whenever A is closed in $\mathbb{T}$. □

This important result has several generalizations, of which one, the extension to vector functions, is given in the next section. A proof will be given in a still more general context below; see **5.19**.

Continuity does *not* ensure that $f(A)$ is open when A is open; compare for example the sets A and $T(A)$ in Figure 1.2 (page 8). A mapping taking open sets to open sets is called an *open mapping*, although, since $f(A^c) \neq f(A)^c$ in general it cannot be assumed that an open mapping is also a closed mapping, taking closed sets to closed sets. However, a *homeomorphism* is a function which is 1–1 onto, continuous, and has a continuous inverse. If f is a homeomorphism so is f^{-1} and hence by **2.7** it is both an open mapping and a closed mapping. It therefore preserves the structure of neighbourhoods so that, if two points are close in the domain, their images are always close in the range. Such a transformation amounts to a relabelling of axes.

If $f(x+h)$ has a limit as $h \downarrow 0$, this is denoted $f(x+)$. Likewise, $f(x-)$ denotes the limit of $f(x-h)$. It is not necessary to have $x \in \mathbb{S}$ for these limits to exist, but if $f(x)$ exists, there is a weaker notion of continuity at x. f is said to be *right-continuous* at the point $x \in \mathbb{S}$ if for any $\varepsilon > 0$ $\exists$ $\delta > 0$ such that, whenever $0 \leq h < \delta$ and $x + h \in \mathbb{S}$,

$$|f(x+h) - f(x)| < \varepsilon. \tag{2.5}$$

It is said to be *left-continuous* at x if for any $\varepsilon > 0$ $\exists$ $\delta > 0$ such that, whenever $0 \leq h < \delta$ and $x - h \in \mathbb{S}$,

$$|f(x) - f(x-h)| < \varepsilon. \tag{2.6}$$

Right continuity at x implies $f(x) = f(x+)$ and left continuity at x implies $f(x) = f(x-)$. If $f(x) = f(x+) = f(x-)$, the function is continuous at x.

Continuity is the property of a point x, not of the function f as a whole. Despite continuity holding pointwise on $\mathbb{S}$, the property may nonetheless break down as certain points are approached.

2.8 Example Consider $f(x) = 1/x$, with $\mathbb{S} = \mathbb{T} = (0, \infty)$. For $\varepsilon > 0$,

$$|f(x+\delta) - f(x)| = \frac{\delta}{x(x+\delta)} < \varepsilon \text{ iff } \delta < \frac{\varepsilon x^2}{1 - \varepsilon x}$$

and hence the choice of δ depends on both ε and x. $f(x)$ is continuous for all $x > 0$, but not in the limit as $x \to 0$. □

The function $f : \mathbb{S} \mapsto \mathbb{T}$ is said to be *uniformly continuous* if for every $\varepsilon > 0$ $\exists$ $\delta > 0$ such that

$$|x - y| < \delta \Rightarrow |f(x) - f(y)| < \varepsilon \tag{2.7}$$

for every $x, y \in \mathbb{S}$. In **2.8** the function is not uniformly continuous, for whichever δ is chosen x can be chosen small enough to invalidate the definition. The problem arises because the set on which the function is defined is open and the boundary point is a discontinuity. Another class of cases that gives difficulty is where the domain is unbounded and continuity at x is breaking down as $x \to \infty$. However:

2.9 Theorem If a function is continuous everywhere on a compact set $\mathbb{S}$, then it is bounded and uniformly continuous on $\mathbb{S}$. □

For proof, see **5.20** and **5.21**.

Continuity is the weakest concept of smoothness of a function. So-called *Lipschitz conditions* provide a whole class of smoothness properties. A function f is said to satisfy a Lipschitz condition at a point x, with respect to a function $h : \mathbb{R}^+ \mapsto \mathbb{R}^+$ satisfying $h(d) \downarrow 0$ as $d \downarrow 0$, if for any $\delta > 0$ $\exists$ $M > 0$ such that for $y \in S(x, \delta)$,

$$|f(y) - f(x)| \le Mh(|x - y|). \tag{2.8}$$

f is said to satisfy a *uniform Lipschitz condition* if condition (2.8) holds with fixed M for all x, $y \in \mathbb{S}$. The type of smoothness imposed depends on the function h. Continuity (resp. uniform continuity) follows from the Lipschitz (resp. uniform Lipschitz) property for any choice of h. Implicit in continuity, as defined by (2.5) and (2.6), is the idea that some function $\delta(\cdot) : \mathbb{R}^+ \mapsto \mathbb{R}^+$ exists satisfying $\delta(\varepsilon) \downarrow 0$ as $\varepsilon \downarrow 0$. This is equivalent to the Lipschitz condition holding for *some* $h(\cdot)$, the case $h = \delta^{-1}$. Imposing some degree of smoothness on h—making it a positive power of the argument for example—imposes a degree of smoothness on the function, forbidding sharp 'corners'.

The next smoothness concept is undoubtedly well known to the reader, although differential calculus plays a fairly minor role in the sequel. Let a function $f : \mathbb{S} \mapsto \mathbb{T}$ be continuous at $x \in \mathbb{S}$. If the function

$$f'_+(x) = \lim_{h \downarrow 0} \left\{ \frac{f(x+h) - f(x)}{h} \right\} \tag{2.9}$$

exists, $f'_+(x)$ is called the *right-hand derivative* of f at x. The *left-hand derivative*, $f'_-(x)$, is defined correspondingly for the case $h \uparrow 0$. If $f'_+(x) = f'_-(x)$, the common value is called the *derivative* of f at x, denoted $f'(x)$ or $\mathrm{d}f/\mathrm{d}x$ and f is said to be *differentiable* at x. If $f' : \mathbb{S} \mapsto \mathbb{R}$ is a continuous function, f is said to be *continuously differentiable* on $\mathbb{S}$.

When the domain is an interval there is yet another smoothness condition. A function $f : [a, b] \mapsto \mathbb{R}$ is *of bounded variation* if

$$T = \sup \sum_{i=1}^{m} |f(x_i) - f(x_{i-1})| < \infty \tag{2.10}$$

where the supremum is taken over the finite partitions of $[a,b]$ of the form $a = x_0 < x_1 < \ldots < x_m = b$. A function that satisfies the uniform Lipschitz condition on $[a,b]$ with $h(|x-y|) = |x-y|$ is of bounded variation on $[a,b]$.

A function f is said to be non-decreasing if $f(y) \geq f(x)$ whenever $y > x$ and also non-increasing if $-f$ is non-decreasing. The term *monotone function* covers both cases.

2.10 Theorem f is of bounded variation iff there exist non-decreasing functions f_1 and f_2 such that $f = f_2 - f_1$.

Proof Define the positive and negative parts of a variable x by $x^+ = \max\{x, 0\}$ and $x^- = x^+ - x$ so that $|x| = x^+ + x^-$. For any given partition of $[a,b]$ let $p = \sum_{i=1}^{m}(f(x_i) - f(x_{i-1}))^+$ and $n = \sum_{i=1}^{m}(f(x_i) - f(x_{i-1}))^-$, so that by construction, $p - n = f(b) - f(a)$. The total variation in this case is $t = p + n$. Letting $P = \sup p$, $N = \sup n$, and $T = \sup t$, it is clear that $P \leq T$ and also that $T \leq P + N$. Note that $p \leq N + f(b) - f(a)$ for all choices of partition and hence $P \leq N + f(b) - f(a)$. By the same reasoning with signs reversed, $N \leq P + f(a) - f(b)$. Hence,

$$P - N = f(b) - f(a). \tag{2.11}$$

Further, $T \geq p + n = 2p + f(a) - f(b)$ for every choice of partition and hence $T \geq 2P + f(a) - f(b) = P + N$, and it follows that $T = P + N$.

To show necessity, define monotone functions $f_2(x) = P(a,x)$ and $f_1(x) = N(a,x) - f(a)$ where in this notation $P(a,b) = P$ and $N(a,b) = N$ with P and N as defined above. Applying (2.11) with b set to x gives, for $a \leq x \leq b$,

$$f_2(x) - f_1(x) = P(a,x) - N(a,x) + f(a) = f(x).$$

To show sufficiency note that if $f = f_2 - f_1$ where f_1 and f_2 are monotone non-decreasing functions then (in the obvious notation) $N_1 = N_2 = 0$ and so (2.11) implies $P_2 = f_2(b) - f_2(a)$ and $P_1 = f_1(b) - f_1(a)$. Hence,

$$T \leq T_2 + T_1 = P_2 + P_1 = f_2(b) + f_1(b) - f_2(a) - f_1(a) < \infty. \quad \blacksquare$$

2.3 Vector Sequences and Functions

A sequence $\{\boldsymbol{x}_n\}$ of real k-vectors is said to converge to a limit $\boldsymbol{x}$ if for every $\varepsilon > 0$ there is an integer N_ε for which

$$||\boldsymbol{x}_n - \boldsymbol{x}|| < \varepsilon \text{ for all } n > N_\varepsilon \tag{2.12}$$

where $|| \cdot ||$ denotes length of the vector (Euclidean norm). The sequence is called a Cauchy sequence in $\mathbb{R}^k$ iff $||\boldsymbol{x}_n - \boldsymbol{x}_m|| < \varepsilon$ whenever $n > N_\varepsilon$ and $m > N_\varepsilon$.

A function

$f: \mathbb{S} \mapsto \mathbb{T}$

where $\mathbb{S} \subseteq \mathbb{R}^k$ and $\mathbb{T} \subseteq \mathbb{R}$ associates each point of $\mathbb{S}$ with a unique point of $\mathbb{T}$. Its graph is the subset of $\mathbb{S} \times \mathbb{T}$ consisting of the $(k+1)$-vectors $\big(\boldsymbol{x}, f(\boldsymbol{x})\big)$ for each $\boldsymbol{x} \in \mathbb{S}$. f is continuous at $\boldsymbol{x} \in \mathbb{S}$ if for any $\varepsilon > 0 \ \exists \ \delta > 0$ such that

$$||\boldsymbol{b}|| < \delta \Rightarrow |f(\boldsymbol{x}+\boldsymbol{b}) - f(\boldsymbol{x})| < \varepsilon \tag{2.13}$$

whenever $\boldsymbol{x} + \boldsymbol{b} \in \mathbb{S}$. The choice of δ may here depend on $\boldsymbol{x}$. On the other hand, f is uniformly continuous on $\mathbb{S}$ if for any $\varepsilon > 0 \ \exists \ \delta > 0$ such that

$$||\boldsymbol{b}|| < \delta \Rightarrow \sup_{\boldsymbol{x} \in \mathbb{S}, \boldsymbol{x}+\boldsymbol{b} \in \mathbb{S}} |f(\boldsymbol{x}+\boldsymbol{b}) - f(\boldsymbol{x})| < \varepsilon. \tag{2.14}$$

A vector $\boldsymbol{f} = (f_1, \ldots, f_m)'$ of functions of $\boldsymbol{x}$ is called, simply enough, a vector function.[1] Continuity concepts apply element-wise to $\boldsymbol{f}$ in the obvious way. The function

$\boldsymbol{f} : \mathbb{S} \mapsto \mathbb{S}, \mathbb{S} \subseteq \mathbb{R}^k$

is said to be one-to-one if there exists a vector function $\boldsymbol{f}^{-1} : \mathbb{S} \mapsto \mathbb{S}$ such that $\boldsymbol{f}^{-1}(\boldsymbol{f}(\boldsymbol{x})) = \boldsymbol{x}$ for each $\boldsymbol{x} \in \mathbb{S}$. An example of a 1–1 continuous function is the affine transformation[2]

$$\boldsymbol{f}(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b}$$

for constants $\boldsymbol{b}$ $(k \times 1)$ and $\boldsymbol{A}$ $(k \times k)$ with $|\boldsymbol{A}| \neq 0$, having inverse $\boldsymbol{f}^{-1}(\boldsymbol{y}) = \boldsymbol{A}^{-1}$ $(\boldsymbol{y} - \boldsymbol{b})$. In most other cases the function $\boldsymbol{f}^{-1}$ does not possess a closed form, but there is a generalization of **2.7**, as follows.

2.11 Theorem Iff $\boldsymbol{f} : \mathbb{S} \mapsto \mathbb{T}$ is continuous, where $\mathbb{S} \subseteq \mathbb{R}^k$ and $\mathbb{T} \subseteq \mathbb{R}^m$, $\boldsymbol{f}^{-1}(A)$ is open in $\mathbb{S}$ when A is open in $\mathbb{T}$ and $\boldsymbol{f}^{-1}(A)$ is closed in $\mathbb{S}$ when A closed in $\mathbb{T}$. □

[1] The prime symbol $'$ here denotes transposition. $\boldsymbol{f}$ is a column vector, written as a row for notational convenience.

[2] An affine transformation is a linear transformation $\boldsymbol{x} \mapsto \boldsymbol{A}\boldsymbol{x}$ followed by a translation, addition of a constant vector $\boldsymbol{b}$. By an accepted abuse of terminology such transformations tend to be referred to as 'linear'.

2.4 Sequences of Functions

Let $f_n : \Omega \mapsto \mathbb{T}$, $\mathbb{T} \subseteq \mathbb{R}$, be a function, where in this case Ω may be an arbitrary set, not necessarily a subset of $\mathbb{R}$. Let $\{f_n, n \in \mathbb{N}\}$ be a sequence of such functions. If there exists f such that, for each $\omega \in \Omega$ and $\varepsilon > 0$, $\exists\, N_{\varepsilon\omega}$ such that $|f_n(\omega) - f(\omega)| < \varepsilon$ when $n > N_{\varepsilon\omega}$, then f_n is said to converge to f, *pointwise* on Ω. As for real sequences the notations $f_n \to f$, $f_n \uparrow f$, or $f_n \downarrow f$ can be used as appropriate, for general or monotone convergence, where in the latter case the monotonicity must apply for every $\omega \in \Omega$.

This is a relatively weak notion of convergence, for it does not rule out the possibility that the convergence is breaking down at certain points of Ω. The following example is related to **2.8** above.

2.12 Example Let $f_n(x) = n/(nx + 1)$, $x \in (0, \infty)$. The pointwise limit of $f_n(x)$ on $(0, \infty)$ is $1/x$. But

$$\left| f_n(x) - \frac{1}{x} \right| = \frac{1}{x(nx+1)}$$

and $1/\left(x(N_{\varepsilon x}x + 1)\right) < \varepsilon$ only for $N_{\varepsilon x} > (1/\varepsilon x - 1)/x$. Thus for given ε, $N_{\varepsilon x} \to \infty$ as $x \to 0$ and it is not possible to put an upper bound on $N_{\varepsilon x}$ such that $|f_n(x) - 1/x| < \varepsilon$, $n \geq N_{ex}$, for every $x > 0$. □

To rule out cases of this type requires the stronger notion of *uniform convergence*. If there exists a function f such that, for each $\varepsilon > 0$ there exists N_ε such that

$$\sup_{\omega \in \Omega} |f_n(\omega) - f(\omega)| < \varepsilon \text{ when } n > N_\varepsilon$$

f_n is said to converge to f uniformly on Ω.

2.5 Summability and Order Relations

The sum of the terms of a real sequence $\{x_n\}_1^\infty$ is called a *series*, written $\sum_{n=1}^\infty x_n$ (or just $\sum x_n$). The terms of the real sequence $\{\sum_{m=1}^n x_m,\ n \in \mathbb{N}\}$ are called the *partial sums* of the series. The series is said to converge if the partial sums converge to a finite limit. A series is said to *converge absolutely* if the monotone sequence $\{\sum_{m=1}^n |x_m|,\ n \in \mathbb{N}\}$ converges.

2.13 Example Consider the geometric series, $\sum_{j=1}^\infty x^j$. This converges to $1/(1-x)$ when $|x| < 1$ and also converges absolutely. It oscillates between cluster points 0 and 1 for $x = -1$ and for other values of x it diverges. □

As a preliminary, recall the fundamental *triangle inequality* for real numbers x and y which says that $|x + y| \leq |x| + |y|$. This holds as an equality when x and y have the same sign and serves merely to indicate the state of affairs when the signs are different. By iteration it extends to sums of any finite number of terms.

2.14 Theorem If a series converges absolutely, then it converges.

Proof The sequence $\{\sum_{m=1}^{n} |x_m|, n \in \mathbb{N}\}$ is monotone and either diverges to $+\infty$ or converges to a finite limit. In the latter case the Cauchy criterion implies that $|x_n| + \ldots + |x_{n+m}| \to 0$ as m and n tend to infinity. Since $|x_n| + \ldots + |x_{n+m}| \geq |x_n + \ldots + x_{n+m}|$ by the triangle inequality, convergence of $\{\sum_{m=1}^{n} x_m, n \in \mathbb{N}\}$ follows by the same criterion. ∎

An alternative terminology speaks of *summability*. A real sequence $\{x_n\}_1^\infty$ is said to be summable if the series $\sum x_n$ converges and absolutely summable if $\{|x_n|\}_1^\infty$ is summable. Any absolutely summable sequence is summable by **2.14** and any summable sequence must be converging to zero. Convergence to zero does not imply summability (see **2.17** below, for example), but convergence of the *tail sums* to zero is necessary and sufficient.

2.15 Theorem Iff $\{x_n\}_1^\infty$ is summable, $\sum_{m=n}^{\infty} x_m \to 0$ as $n \to \infty$.

Proof For necessity, write $|\sum_{m=1}^{\infty} x_m| \leq |\sum_{m=1}^{n-1} x_m| + |\sum_{m=n}^{\infty} x_m|$. Since for any $\varepsilon > 0$ there exists N such that $|\sum_{m=n}^{\infty} x_m| < \varepsilon$ for $n \geq N$, it follows that $|\sum_{m=1}^{\infty} x_m| \leq |\sum_{m=1}^{N-1} x_m| + \varepsilon < \infty$. Conversely, assume summability. Let $A = \sum_{n=1}^{\infty} x_n$ so that $|A| < \infty$ and then note that $|\sum_{m=n}^{\infty} x_m| = |A - \sum_{m=1}^{n-1} x_m| \to 0$ as $n \to \infty$. ∎

Consider the sequence constructed as the averages of the coordinates of another sequence, say $\bar{x}_n = n^{-1} \sum_{m=1}^{n} x_m$ for $n = 1, 2, 3, \ldots$. The sequence $\{\bar{x}_n\}_1^\infty$ often converges in cases where $\{x_n\}_1^\infty$ does not, a property called *Cesàro-summability*. Thus the sequence $\{1, 0, 1, 0, \ldots\}$ does not converge but, as is easily verified, the Cesàro sum is $\frac{1}{2}$ in this terminology. Cesàro's convergence concept was defined for partial sum sequences and properly specifies the convergence of the sequence $\{\bar{y}_n\}$ where $y_n = \sum_{m=0}^{n-1} x_m$, which is equivalent to the convergence of the sequence $\{\sum_{m=0}^{n-1} (1 - m/n)x_m\}_{n=1}^\infty$ (see e.g. Zygmund [196], page 76). However, the designation 'Cesàro sum' is also used for the limiting average in the context of sequences not explicitly constructed as partial sums and it is used here in this broader sense. A useful fact in this context is the following.

2.16 Theorem If $x_n \to x$ then $\bar{x}_n \to x$.

Proof Given $\varepsilon > 0$, choose k so that $|x - x_m| < \varepsilon/2$ for $m > k$. Let $M = \max_{1 \le m \le k} |x_m|$ and then for $n > 2kM/\varepsilon$,

$$\begin{aligned}\left| x - \frac{1}{n}\sum_{m=1}^{n} x_m \right| &\le \frac{1}{n}\sum_{m=1}^{k} |x_m| + \frac{1}{n}\sum_{m=k+1}^{n} |x - x_m| \\ &\le \frac{kM}{n} + \frac{(n-k)\varepsilon}{2n} < \varepsilon. \quad \blacksquare\end{aligned}$$

Various notations are used to indicate the relationships between rates of divergence or convergence of different sequences. If $\{x_n\}_1^\infty$ is any real sequence, $\{a_n\}_1^\infty$ a sequence of positive real numbers, and there exists a constant $B < \infty$ such that $|x_n|/a_n \le B$ for all n, x_n is said to be (at most) of the order of magnitude of a_n written $x_n = O(a_n)$ ('big Oh'). If $\{x_n/a_n\}$ converges to zero, write $x_n = o(a_n)$ ('little Oh') and say that x_n is of strictly smaller order of magnitude than a_n. Here a_n can be increasing or decreasing, so this notation can be used to express an upper bound either on the rate of growth of a divergent sequence or on the rate of convergence of a sequence to zero. Here are some rules for manipulation of $O(\cdot)$, whose proof follows from the definition. If $x_n = O(n^\alpha)$ and $y_n = O(n^\beta)$, then

$$x_n + y_n = O(n^{\max\{\alpha,\beta\}}) \tag{2.15}$$

$$x_n y_n = O(n^{\alpha+\beta}) \tag{2.16}$$

$$x_n^\beta = O(n^{\alpha\beta}), \text{whenever } x_n^\beta \text{ is defined.} \tag{2.17}$$

When $x_n \ge 0$ an alternative notation is $x_n \ll a_n$, which means that there is a constant, $0 < B < \infty$, such that $x_n \le Ba_n$ for all n. This may be more convenient in algebraic manipulations.

The notation $x_n \sim a_n$ is commonly used to mean that $x_n/a_n \to 1$. Also the notation $x_n \simeq a_n$ is used in this book to indicate that there exist $N < \infty$ and finite constants $A > 0$ and $B \ge A$, such that $\inf_{n \ge N}(|x_n|/a_n) \ge A$ and $\sup_{n \ge N}(|x_n|/a_n) \le B$. This says that x_n and a_n grow ultimately at the same rate and is different from the relation $x_n = O(a_n)$, since the latter does not exclude $|x_n|/a_n \to 0$.

2.17 Theorem If $\{x_n\}$ is a real positive sequence and $x_n \simeq n^a$,

(i) if $\alpha > -1$ then $\sum_{m=1}^{n} x_m \simeq n^{1+\alpha}$;

(ii) if $\alpha = -1$ then $\sum_{m=1}^{n} x_m \simeq \log n$;

(iii) if $\alpha < -1$ then $\sum_{m=1}^{\infty} x_m < \infty$ and $\sum_{m=n}^{\infty} x_m = O(n^{1+a})$.

Proof The arguments are by integral approximation. If $\alpha > -1$,

$$\frac{n^{1+\alpha}-1}{1+\alpha} = \int_1^n m^\alpha \mathrm{d}m = \sum_{m=1}^{n-1} \int_m^{m+1} x^\alpha \mathrm{d}x$$

where either $m^\alpha \leq \int_m^{m+1} x^\alpha \mathrm{d}x \leq (m+1)^\alpha$ for $\alpha \geq 0$ or $(m+1)^\alpha \leq \int_m^{m+1} x^\alpha \mathrm{d}x \leq m^\alpha$ for $-1 < \alpha < 0$, so that

$$\left| \frac{1}{n^{1+\alpha}} \sum_{m=1}^{n} m^\alpha - \frac{1-n^{-1-\alpha}}{1+\alpha} \right| \leq \frac{\max\{1, n^\alpha\}}{n^{1+\alpha}} = o(1). \tag{2.18}$$

By assumption and the definition of '$\simeq$' there exist $N \geq 1$ and constants $A > 0$ and $B \geq A$ such that $An^\alpha \leq x_n \leq Bn^\alpha$ for $n \geq N$ and hence $A \sum_{m=N}^n m^\alpha \leq \sum_{m=N}^n x_m \leq B \sum_{m=N}^n m^\alpha$. Since the sum of the terms from 1 to $N-1$ is finite, their omission cannot change the conclusions. This proves (i). The case $\alpha = -1$ is similar given the fundamental result

$$\log n = \int_1^n m^{-1} \mathrm{d}m = \sum_{m=1}^n m^{-1} + C + O(n^{-1})$$

where $C \approx 0.5772$ is Euler's constant. This proves (ii). Finally for $\alpha < -1$,

$$\frac{-1}{1+\alpha} n^{1+\alpha} = \int_n^\infty m^\alpha \mathrm{d}m = \sum_{m=n}^\infty \int_m^{m+1} x^\alpha \mathrm{d}x$$

and similarly to (2.18)

$$\left| \frac{1}{n^{1+\alpha}} \sum_{m=n}^\infty m^\alpha + \frac{1}{1+\alpha} \right| \leq \frac{1}{n}.$$

Similar reasoning now gives part (iii). ∎

There is however a more delicate summability condition than **2.17**(iii).

2.18 Theorem If $x_n \simeq 1/\big(n(\log n)^{1+\delta}\big)$ with $\delta > 0$, then $\sum_{n=1}^\infty x_n < \infty$. If $\delta = 0$, then $\sum_{m=1}^n x_m \simeq \log\log n$.

Proof Similar to **2.17**, using for $n > m > 1$ the equalities

$$\int_m^n \frac{1}{x(\log x)^{1+\delta}} \mathrm{d}x = \frac{1}{\delta}\Big((\log m)^{-\delta} - (\log n)^{-\delta} \Big) \tag{2.19}$$

and

$$\int_m^n \frac{1}{x\log x}\mathrm{d}x = \log\log n - \log\log m. \tag{2.20}$$

(See [85], 2.721.) ∎

Another result related to **2.17** is the following.

2.19 Theorem For $b \geq a$,

$$\sum_{k=1}^{j-1} k^a (j-k)^b \simeq \begin{cases} j^{a+b+1}, & a > -1 \\ j^b \log j, & a = -1 \\ j^b, & a < -1. \end{cases}$$

Proof In the case $a > -1$,

$$\begin{aligned} \sum_{k=1}^{j-1} k^a (j-k)^b &= j^{a+b+1} \sum_{k=1}^{j-1} \left(\frac{k}{j}\right)^a \left(1-\frac{k}{j}\right)^b j^{-1} \\ &\simeq j^{a+b+1} \int_0^1 \xi^a (1-\xi)^b \mathrm{d}\xi \\ &= j^{a+b+1} B(a+1, b+1) \end{aligned}$$

where $B(x,y) = \Gamma(x)\Gamma(y)/\Gamma(x+y)$ denotes the Beta function. Otherwise,

$$\sum_{k=1}^{j-1} k^a (j-k)^b = j^b \sum_{k=1}^{j-1} k^a \left(1-\frac{k}{j}\right)^b \tag{2.21}$$

where $\sum_{k=1}^{j-1} k^a \simeq \log j$ if $a = -1$ by **2.17**(ii) and $\sum_{k=1}^{\infty} k^a < \infty$ if $a < -1$ by **2.17**(iii). In each case the second factor in the right-hand-side terms of (2.21) is made as close to 1 as desired by taking j large enough. ∎

2.6 Inequalities

Closely related to theorems relating to the orders of sums are inequalities that supply bounds for certain sums in terms of others. The concept of a *convex function* is fundamental.

2.20 Definition Let $I \subseteq \mathbb{R}$ be any interval. A function $\phi : I \mapsto \mathbb{R}$ is said to be convex on I if

$$\phi\big((1-\lambda)x+\lambda y\big) \le (1-\lambda)\phi(x)+\lambda\phi(y) \tag{2.22}$$

for all $x, y \in I$ and $\lambda \in [0,1]$. ϕ is *concave* on I if $-\phi$ is convex on I. □

Since the definition can be iterated, a direct implication for convex ϕ is that for real numbers $x_1, \dots, x_n$ from I and weights $a_i \ge 0$ such that $\sum_{i=1}^n a_i = 1$,

$$\phi\Big(\sum_{i=1}^n a_i x_i\Big) \le \sum_{i=1}^n a_i \phi(x_i). \tag{2.23}$$

A natural application is to the power function $|\cdot|^p$, which with $p \ge 1$ is convex everywhere on $\mathbb{R}$. Thus, putting $a_i = 1/n$ for each i in (2.23) yields, for $p \ge 1$ and arbitrary real numbers x_i,

$$\left|\frac{1}{n}\sum_{i=1}^n x_i\right|^p \le \frac{1}{n}\sum_{i=1}^n |x_i|^p. \tag{2.24}$$

This relation is a key ingredient of what is known as the c_r *inequality.*

2.21 Theorem For any collection of real numbers $x_1, \dots, x_n$ and $r > 0$,

$$\left|\sum_{i=1}^n x_i\right|^r \le c_r \sum_{i=1}^n |x_i|^r \tag{2.25}$$

where $c_r = 1$ when $r \le 1$ and $c_r = n^{r-1}$ when $r \ge 1$.

Proof Since $\left|\sum_{i=1}^m x_i\right|^r \le \left(\sum_{i=1}^n |x_i|\right)^r$ there is no loss of generality in letting the x_i be non-negative. For the case $0 < r \le 1$, define $z_i = x_i/(\sum_{j=1}^n x_j)$. Since $0 \le z_i \le 1$ for each i, $z_i^r \ge z_i$ and hence since $\sum_{i=1}^m z_i = 1$, $\sum_i^m z_i^r \ge 1$ and (2.25) follows. For $r > 1$, (2.25) is found from (2.24) with $p = r$ after rearranging. ∎

The following is *Hölder's inequality.*

2.22 Theorem For arbitrary collections of real numbers $x_1, \dots, x_n$ and $y_1, \dots, y_n$ and $p \ge 1$,

$$\sum_{i=1}^n |x_i y_i| \le \Big(\sum_{i=1}^n |x_i|^p\Big)^{1/p} \Big(\sum_{i=1}^n |y_i|^q\Big)^{1/q} \tag{2.26}$$

for $q = \dfrac{p}{p-1}$. □

The possibility that one or both sides of this inequality is infinite is not ruled out. The case $p = 1$ and $q = \infty$ is immediate with the second majorant factor taking the value $\max_{1 \le i \le n} |y_i|$. The case $p > 1$ requires a small lemma, as follows.

2.23 Lemma For $p > 1$, $q = p/(p-1)$ and any pair of real numbers a and b,

$$|ab| \le \frac{|a|^p}{p} + \frac{|b|^q}{q}. \tag{2.27}$$

Proof If either a or b are zero this is trivial. Otherwise let $s = p\log|a|$ and $t = q\log|b|$. Inverting these relations gives $|a| = \mathrm{e}^{s/p}$, $|b| = \mathrm{e}^{t/q}$, $|ab| = \mathrm{e}^{s/p+t/q}$. Noting $1/p + 1/q = 1$, the lemma now follows from the fact that the exponential function is convex. ∎

Proof of 2.22. In (2.27) let $a_i = x_i/\big(\sum_{i=1}^n |x_i|^p\big)^{1/p}$ and $b_i = y_i/\big(\sum_{i=1}^n |y_i|^q\big)^{1/q}$ and observe that

$$\sum_{i=1}^n |a_i b_i| \le \sum_{i=1}^n \frac{|a_i|^p}{p} + \sum_{i=1}^n \frac{|b_i|^p}{q} = \frac{1}{p} + \frac{1}{q} = 1. \quad ∎ \tag{2.28}$$

The case $p = q = 2$ is frequently used in applications and in this form the result is known as the *Cauchy–Schwarz inequality*. Probabilistic versions of these relations are given in §9.7. Also notice that (2.24) follows on setting $y_i = 1/n$. A further implication of (2.28) is that

$$-1 \le \sum_{i=1}^n a_i b_i \le 1 \tag{2.29}$$

which for $p = q = 2$ says that the sum of products of pairs of real numbers is bounded absolutely by the product of their Euclidean norms. This reflects the familiar fact from descriptive statistics that correlation coefficients must fall in the interval $[-1, 1]$.

Hölder's inequality is used to show another useful relation, *Minkowski's inequality.*

2.24 Theorem For arbitrary collections of real numbers $x_1, \ldots, x_n$ and $y_1, \ldots, y_n$ and $p \ge 1$,

$$\left(\sum_{i=1}^n |x_i + y_i|^p\right)^{1/p} \le \left(\sum_{i=1}^n |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^n |y_i|^p\right)^{1/p}. \tag{2.30}$$

Proof If $p = 1$ this follows directly by the triangle inequality. For $p > 1$ write

$$\begin{aligned}\sum_{i=1}^{n} |x_i + y_i|^p &= \sum_{i=1}^{n} |x_i + y_i||x_i + y_i|^{p-1} \\ &\leq \sum_{i=1}^{n} |x_i||x_i + y_i|^{p-1} + \sum_{i=1}^{n} |y_i||x_i + y_i|^{p-1} \\ &\leq \left(\left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p} + \left(\sum_{i=1}^{n} |y_i|^p\right)^{1/p}\right)\left(\sum_{i=1}^{n} |x_i + y_i|^p\right)^{1-1/p}\end{aligned}$$

where the first inequality is by the triangle inequality, and for the second one Hölder's inequality has been applied to each right-hand side term. Cancellation and rearrangement now gives (2.30). ∎

The following identity is known as Abel's partial summation formula. It proves useful in a variety of contexts for verifying summability.

2.25 Theorem Given arbitrary sequences $a_0, \ldots, a_n$ and $b_0, \ldots, b_n$,

$$\sum_{t=1}^{n} a_t(b_t - b_{t-1}) = \sum_{t=1}^{n} (a_{t-1} - a_t)b_{t-1} + a_n b_n - a_0 b_0. \tag{2.31}$$

Proof Write the sum on the left-hand side as $A - B$ where $A = a_1 b_1 + \cdots + a_n b_n$ and $B = a_1 b_0 + \cdots + a_n b_{n-1}$. Then, the sum of n terms on the right-hand side is the difference of $A + a_0 b_0 - a_n b_n$ and B. ∎

An application that arises more than once in the theory of dependent processes is the following. (See Chapters 17 and 31.)

2.26 Theorem If $\{x_m, m \geq 1\}$ is a monotone decreasing positive real sequence then

$$\sum_{m=1}^{\infty} (x_{m-1} - x_m)^{1/2} \leq K\left(\sum_{m=1}^{\infty} x_m \log^2 m\right)^{1/2}. \tag{2.32}$$

where $K < \infty$.

Proof The theorem is true if the majorant of (2.32) is infinite, so assume that x_m tends to zero fast enough that the majorant series converges. Define a sequence $\{c_m, m \geq 0\}$ by setting $c_0 = 0$, $c_1 = 1$, and $c_m = 1/(m \log^2 m)$ for $m > 1$. This sequence is summable according to **2.18**. Multiply the terms in the minorant of (2.32) by $1 = c_m^{1/2} c_m^{-1/2}$ and apply the Cauchy–Schwarz inequality (**2.22** with $p = 2$) to obtain

$$\sum_{m=1}^{\infty} (x_{m-1} - x_m)^{1/2} \le \left(\sum_{m=1}^{\infty} c_m \right)^{1/2} \left(\sum_{m=1}^{\infty} (x_{m-1} - x_m) c_m^{-1} \right)^{1/2}. \tag{2.33}$$

Apply **2.25** with $b = x$ and $a = c^{-1}$ to the majorant of (2.33), reversing the signs on each side of (2.31) and letting $n \to \infty$. Note that for $m > 1$,

$$c_m^{-1} - c_{m-1}^{-1} = \log^2 m + \frac{2(m-1)\log m^*}{m^*} \tag{2.34}$$

where $m - 1 \le m^* \le m$ by the mean value theorem. Since the second term of (2.34) increases more slowly than the first, there exists K_1 large enough that

$$\sum_{m=1}^{\infty} x_m (c_m^{-1} - c_{m-1}^{-1}) \le K_1 \sum_{m=1}^{\infty} x_m \log^2 m$$

showing that $K = \left(K_1 \sum_{m=1}^{\infty} c_m \right)^{1/2} < \infty$ satisfies inequality (2.32). ∎

2.7 Regular Variation

It is common practice to express the rate of convergence to zero of a positive real sequence in terms of the summability of the coordinates raised to a given power. The following device due to Karamata [110] allows some further refinement of summability conditions. Let $U(\nu)$ be a positive function of ν. If $U(\nu x)/U(\nu) \to x^\rho$ as $\nu \to \infty$ (0) for $x > 0$ and $-\infty < \rho < +\infty$, U is said to be *regularly varying at infinity (zero).* If a positive function $L(\nu)$ has the property $L(\nu x)/L(\nu) \to 1$ for $x > 0$ as $\nu \to \infty$ (0), it is said to be *slowly varying at infinity* (*zero*). Evidently, any regularly varying function can be expressed in the form $U(\nu) = \nu^\rho L(\nu)$, where $L(\nu)$ is slowly varying. While the definition allows ν to be a real variable, in the cases of interest $\nu = n$ for $n \in \mathbb{N}$, with U and L having the interpretation of positive sequences.

2.27 Example $(\log \nu)^\alpha$ is slowly varying at infinity for any α. □

On the theory of regular variation see Feller ([74]), or Loève ([121]). The important property is the following, of which the proof is direct from the definitions.

2.28 Theorem If L is slowly varying at infinity, then for any $\delta > 0$ there exists $N \geq 1$ such that

$$v^{-\delta} < L(v) < v^{\delta}, \text{ all } v > N. \quad \square \tag{2.35}$$

The following corollary of **2.17** shows how the notion of a convergent power series can be refined by allowing for the presence of a slowly varying function.

2.29 Corollary If $x_n = O(n^{\alpha}L(n))$ then $\sum_{n=1}^{\infty} x_n < \infty$ for all $\alpha < -1$ and all functions $L(n)$ that are slowly varying at infinity. □

On the other hand, the presence of a slowly varying component can affect the summability of a sequence, as has already been demonstrated in Theorem **2.18**.

The following key result is after Feller ([74], p. 275).

2.30 Theorem If a positive monotone function $U(v)$ satisfies

$$\frac{U(vx)}{U(v)} \to \Psi(x) \text{ as } v \to \infty, x \in \mathbb{R}^{+} \tag{2.36}$$

where $0 < \Psi(x) < \infty$ and Ψ is continuous, then $\Psi(x) = x^{\rho}$ for some $\rho \in \mathbb{R}$. □

The proof requires a lemma that has use in other contexts.

2.31 Lemma If $\psi(x) > 0$ is continuous and $\psi(x_1x_2) = \psi(x_1)\psi(x_2)$ for $x_1, x_2 \in \mathbb{R}^{+}$ then $\psi(x) = x^{\rho}$ for $\rho \in \mathbb{R}$.

Proof For $z = \log x$ define $\phi(z) = \psi(\mathrm{e}^{z})$, so that $\phi(z_1 + z_2) = \phi(z_1)\phi(z_2)$. Let ρ be defined by the equality $\phi(1) = \mathrm{e}^{\rho}$ and so define $v(z) = \mathrm{e}^{-\rho z}\phi(z)$, so that $v(1) = 1$ and $v(z_1 + z_2) = v(z_1)v(z_2)$. For integer m, $v(m) = v(1)^{m} = 1$ and similarly for integer n, $v(m/n)^{n} = v(m) = 1$ and hence $v(m/n) = 1$. Since $v(z) = 1$ for all rational z it follows that $\phi(z) = \mathrm{e}^{\rho z}$ and so $\psi(x) = \mathrm{e}^{\rho \log x} = x^{\rho}$, for $x \in D$ where D is dense in $\mathbb{R}^{+}$. The property extends to $x \in \mathbb{R}^{+}$ by continuity. ∎

Proof of 2.30 $\Psi(x_1x_2) = \Psi(x_1)\Psi(x_2)$, since by (2.36)

$$\frac{U(vx_1x_2)}{U(v)} = \frac{U(vx_1x_2)}{U(vx_2)}\frac{U(vx_2)}{U(v)} \to \Psi(x_1)\Psi(x_2).$$

The theorem follows by **2.31**. ∎

To the extent that (2.36) is a general property, monotone functions are as a rule regularly varying.

2.32 Theorem The derivative of a monotone regularly varying function is regularly varying at ∞.

Proof Given $U(v) = v^{\rho} L(v)$, write

$$U'(v) = \rho v^{\rho-1} L(v) + v^{\rho} L'(v) = v^{\rho-1}\big(\rho L(v) + vL'(v)\big). \tag{2.37}$$

If $L'(v) \to 0$ there is no more to show, so assume $\liminf_v L'(v) > 0$. Then

$$\frac{d}{dv}\left(\frac{L(vx)}{L(v)}\right) = \frac{L'(v)}{L(v)}\left(\frac{L'(vx)}{L'(v)} - \frac{L(vx)}{L(v)}\right) \to 0 \tag{2.38}$$

which implies $L'(vx)/L'(v) \to 1$. Thus, as $v \to \infty$,

$$\frac{U'(vx)}{U'(v)} = x^{\rho-1}\frac{\rho L(vx) + vxL'(vx)}{\rho L(v) + vL'(v)} \to x^{\rho}. \quad \blacksquare \tag{2.39}$$

The behaviour of integrals of regularly varying functions parallels **2.17** in an obvious way.

2.33 Theorem Let U be regularly varying at ∞ with parameter ρ.

(i) If $\rho \geq -1$,

$$\frac{\int_0^v U(t)\mathrm{d}t}{vU(v)} \to \frac{1}{\rho+1} \text{ as } v \to \infty. \tag{2.40}$$

(ii) If $\rho < -1$,

$$\frac{\int_v^\infty U(t)\mathrm{d}t}{vU(v)} \to -\frac{1}{\rho+1} \text{ as } v \to \infty. \tag{2.41}$$

Proof For case (i), since $U(v) = v^{\rho} L(v)$,

$$\begin{aligned}\int_0^v U(t)\mathrm{d}t &= v\int_0^1 U(sv)\mathrm{d}s \\ &= v^{1+\rho} L(v)\int_0^1 s^{\rho}\frac{L(sv)}{L(v)}\mathrm{d}s\end{aligned}$$

and (2.40) follows since

$$\int_0^1 s^\rho \frac{L(sv)}{L(v)}\mathrm{d}s \to \int_0^1 s^\rho \mathrm{d}s = \frac{1}{\rho+1} \text{ as } v \to \infty.$$

In the case $\rho = -1$ the ratio is understood to diverge. In case (ii),

$$\int_v^\infty U(t)\mathrm{d}t = v^{1+\rho}L(v)\int_1^\infty s^\rho \frac{L(sv)}{L(v)}\mathrm{d}s$$

and

$$\int_1^\infty s^\rho \frac{L(sv)}{L(v)}\mathrm{d}s \to \int_1^\infty s^\rho \mathrm{d}s = -\frac{1}{\rho+1}. \quad \blacksquare$$

2.8 Arrays

Arguments concerning stochastic convergence often involve a double-indexing of elements. An *array* is a mapping whose domain is the Cartesian product of countable, linearly ordered sets, such as $\mathbb{N}\times\mathbb{N}$ or $\mathbb{Z}\times\mathbb{N}$, or a subset thereof. A real double array, in particular, is a double-indexed collection of numbers, or alternatively a sequence whose members are real sequences. Notations used include $\{\{x_{nt}, t \in \mathbb{Z}\}, n \in \mathbb{N}\}$, and just $\{x_{nt}\}$ when the context is clear.

A collection of finite sequences $\{\{x_{nt}, t = 1, \ldots, k_n\}, n \in \mathbb{N}\}$, where $k_n \uparrow \infty$ as $n \to \infty$, is called a *triangular array.* As an example, consider array elements of the form $x_{nt} = y_t/n$, where $\{y_t, t = 1, \ldots, n\}$ is a real sequence. The question of whether the series $\{\sum_{t=1}^n x_{nt}, n \in \mathbb{N}\}$ converges is equivalent to that of the Cesàro convergence of the original sequence; however, the array formulation is frequently the more convenient. The following is *Toeplitz's lemma.*

2.34 Lemma Suppose $\{y_n\}$ is a real sequence and $y_n \to y$. If $\{\{x_{nt}, t = 1, \ldots, k_n\}$, $n \in \mathbb{N}\}$ is a triangular array such that

(a) $x_{nt} \to 0$ as $n \to \infty$ for each fixed t

(b) $\lim_{n\to\infty} \sum_{t=1}^{k_n} |x_{nt}| \leq C < \infty$

(c) $\lim_{n\to\infty} \sum_{t=1}^{k_n} x_{nt} = 1$

then $\sum_{t=1}^{k_n} x_{nt}y_t \to y$. For $y = 0$, (c) can be omitted.

Proof By assumption on $\{y_n\}$, for any $\varepsilon > 0 \;\exists\; N_\varepsilon \geq 1$ such that for $n > N_\varepsilon$, $|y_n - y| < \varepsilon/C$. Hence by (c) and then (b) and the triangle inequality,

$$\begin{aligned}\lim_{n\to\infty}\left|\sum_{t=1}^{k_n} x_{nt}y_t - y\right| &= \lim_{n\to\infty}\left|\sum_{t=1}^{k_n} x_{nt}(y_t - y)\right| \\ &\leq \lim_{n\to\infty}\left|\sum_{t=1}^{N_\varepsilon} x_{nt}(y_t - y)\right| + \varepsilon = \varepsilon \qquad (2.42)\end{aligned}$$

in view of (a). This completes the proof, since ε is arbitrary. ∎

An array $\{x_{nt}\}$ that satisfies the conditions of the lemma is defined by $x_{nt} = (\sum_{s=1}^n y_s)^{-1}y_t$, where $\{y_t\}$ is a positive sequence and $\sum_{s=1}^n y_s \to \infty$. Putting $y_t = b_t - b_{t-1}$, an alternative form of the same case is $x_{nt} = (b_t - b_{t-1})/b_n$ with $b_0 = 0$ and $b_n \to \infty$. In this form **2.34** is known as Cesàro's lemma ([190], 12.6).

A leading application of **2.34** is to prove the following, known as *Kronecker's lemma*, a fundamental tool of limit theory.

2.35 Lemma Let $\{x_t\}_1^\infty$ be a sequence of real numbers and $\{a_t\}_1^\infty$ a positive monotone sequence with $a_t \uparrow \infty$. If $\sum_{t=1}^n x_t/a_t \to C < \infty$ as $n \to \infty$, then

$$\frac{1}{a_n}\sum_{t=1}^n x_t \to 0. \qquad (2.43)$$

Proof Defining $c_0 = 0$ and $c_n = \sum_{t=1}^n x_t/a_t$ for $n \in \mathbb{N}$, note that $x_t = a_t(c_t - c_{t-1})$, $t = 1,\ldots,n$. Define $a_0 = 0$ and $b_t = a_t - a_{t-1}$ for $t = 1,\ldots,n$, so that $a_n = \sum_{t=1}^n b_t$. Now apply the Abelian summation formula (**2.25**) to obtain

$$\begin{aligned}\frac{1}{a_n}\sum_{t=1}^n x_t &= \frac{1}{a_n}\sum_{t=1}^n a_t(c_t - c_{t-1}) \\ &= c_n - \frac{1}{a_n}\sum_{t=1}^n b_t c_{t-1} \to C - C = 0 \qquad (2.44)\end{aligned}$$

where the convergence is by **2.34** setting $x_{nt} = b_t/a_n$. ∎

The notion of array convergence extends the familiar sequence concept. Consider for full generality an array of subsequences, a collection $\{\{x_{mn_k}, k \in \mathbb{N}\}, m \in \mathbb{N}\}$, where $\{n_k, k \in \mathbb{N}\}$ is an increasing sequence of positive integers. If the limit $x_m = \lim_{k\to\infty} x_{mn_k}$ exists for each $m \in \mathbb{N}$ then the array is convergent and its limit is the infinite sequence $\{x_m, m \in \mathbb{N}\}$. Whether *this* sequence converges is a separate question from whether it exists at all.

Suppose the array is bounded, in the sense that $\sup_{k,m} |x_{mn_k}| \leq B < \infty$. By **2.2** there exists for each m at least one cluster point, say x_m, of the inner sequence $\{x_{mn_k}, k \in \mathbb{N}\}$. An important question in several contexts is this: is it valid to say that the array as a whole has a cluster point?

2.36 Theorem Corresponding to any bounded array $\{\{x_{mn_k}, k \in \mathbb{N}\}, m \in \mathbb{N}\}$, there exists a sequence $\{x_m\}$, the limit of the array $\{\{x_{mn^*_k}, k \in \mathbb{N}\}, m \in \mathbb{N}\}$ as $k \to \infty$, where $\{n^*_k\}$ is the same subsequence of $\{n_k\}$ for each m.

Proof This is by construction of the required subsequence. Begin with a convergent subsequence for $m = 1$; let $\{n^1_k\}$ be a subsequence of $\{n_k\}$ such that $x_{1,n^1_k} \to x_1$. Next, consider the sequence $\{x_{2,n^1_k}\}$. Like $\{x_{2,n_k}\}$, this is on the bounded interval $(-B, B)$ and so contains a convergent subsequence. Let the indices of this latter subsequence, drawn from the members of $\{n^1_k\}$, be denoted $\{n^2_k\}$ and note that $x_{1,n^2_k} \to x_1$ as well as $x_{2,n^2_k} \to x_2$. Proceeding in the same way for each m generates an array $\{\{n^m_k, k \in \mathbb{N}\}, m \in \mathbb{N}\}$, having the property that $\{x_{i,n^m_k}, k \in \mathbb{N}\}$ is a convergent sequence for $1 \leq i \leq m$.

Now consider the sequence $\{n^k_k, k \in \mathbb{N}\}$; in other words, take the first member of $\{n^1_k\}$, the second member of $\{n^2_k\}$, and so on. For each m, this sequence is a subsequence of $\{n^m_k\}$ from the mth point of the sequence onwards and hence the sequence $\{x_{m,n^k_k}, k \geq m\}$ is convergent. This means that the sequence $\{x_{m,n^k_k}, k \in \mathbb{N}\}$ is convergent, so setting $\{n^*_k\} = \{n^k_k\}$ satisfies the requirement of the theorem. ∎

This is called the 'diagonal method'. The elements n^k_k may be thought of as the diagonal elements of the square matrix (of infinite order) whose rows contain the sequences $\{n^m_k\}$, each a subsequence of the row above it. This theorem holds independently of the nature of the elements $\{x_{mn}\}$.

A simple application is to sequences of random vectors, which can be thought of a points in Euclidean k-space. A sequence in $\mathbb{R}^k$ is an array that is finite in one direction. To prove the Bolzano–Weierstrass theorem for $\mathbb{R}^k$, for example, the diagonal method is used to exhibit a single convergent subsequence. More generally, any space of points on which convergent sequences are defined is covered by the theorem. Applications arise in §6.5, §12.4, §23.5, and elsewhere.

3
Measure

3.1 Measure Spaces

A measure is a set function, a mapping that associates a (possibly extended) real number with a set. Commonplace examples of measures include the lengths, areas, and volumes of geometrical figures, but wholly abstract sets can be 'measured' in an analogous way. Here's the formal definition.

3.1 Definition Given a class $\mathcal{F}$ of subsets of a set Ω, a measure $\mu : \mathcal{F} \mapsto \bar{\mathbb{R}}$ is a function having the following properties.

(a) $\mu(A) \geq 0$, all $A \in \mathcal{F}$.

(b) $\mu(\varnothing) = 0$.

(c) For a countable collection $\{A_j \in \mathcal{F}, j \in \mathbb{N}\}$ with $A_j \cap A_{j'} = \varnothing$ for $j \neq j'$ and $\bigcup_j A_j \in \mathcal{F}$,

$$\mu\Big(\bigcup_j A_j\Big) = \sum_j \mu(A_j). \quad \square \tag{3.1}$$

Condition (a) is optional and set functions taking either sign may be referred to as measures (see e.g. §4.4), but non-negativity is desirable for present purposes. The particular cases at issue in this book are of course the probabilities of random events in a sample space Ω and this application of the theory is the subject of Chapter 7.

A *measurable space* is a pair $(\Omega, \mathcal{F})$ where Ω is any collection of objects and $\mathcal{F}$ is a σ-field of subsets of Ω. When $(\Omega, \mathcal{F})$ is a measurable space, the triple $(\Omega, \mathcal{F}, \mu)$ is called a *measure space.* More than one measure can be associated with the measurable space $(\Omega, \mathcal{F})$, hence the distinction between measure space and measurable space is important.

Condition **3.1**(c) is called *countable additivity.* If a set function has the property

$$\mu(A \cup B) = \mu(A) + \mu(B) \tag{3.2}$$

for each disjoint pair A and B, a property that extends by iteration to finite collections $A_1, \ldots, A_n$, it is said to be *finitely additive.* In **3.1** $\mathcal{F}$ could be a field, but

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0003

the possibility of extending the properties of μ to the corresponding σ-field, by allowing additivity over countable collections, is an essential feature of a measure.

The measure μ is said to be *finite* if $\mu(\Omega) < \infty$. Also, if $\Omega = \bigcup_j \Omega_j$ where $\{\Omega_j\}$ is a countable collection of $\mathcal{F}$-sets and $\mu(\Omega_j) < \infty$ for each j, μ is said to be *σ-finite*. In particular, if there is a collection $\mathcal{S}$ such that $\mathcal{F} = \sigma(\mathcal{S})$ and $\Omega_j \in \mathcal{S}$ for each j, μ is said to be σ-finite *on* $\mathcal{S}$ (rather than on $\mathcal{F}$). If $\mathcal{F}_A = \{A \cap B : B \in \mathcal{F}\}$ for some $A \in \mathcal{F}$, $(A, \mathcal{F}_A)$ is a measurable space and $(A, \mathcal{F}_A, \mu)$ is a measure space called the *restriction* of $(\Omega, \mathcal{F}, \mu)$ to A. If in this case $\mu(A^c) = 0$ (equivalent to $\mu(A) = \mu(\Omega)$ when $\mu(\Omega) < \infty$) A is called a *support* of the measure. When A supports Ω, the sets of $\mathcal{F}_A$ have the same measures as the corresponding ones of $\mathcal{F}$. A point $\omega \in \Omega$ with the property $\mu(\{\omega\}) > 0$ is called an *atom* of the measure.

3.2 Example The case closest to everyday intuition is *Lebesgue measure m* on the measurable space $(\mathbb{R}, \mathcal{B})$, where $\mathcal{B}$ is the Borel field of $\mathbb{R}$. Generalizing the notion of length in geometry, Lebesgue measure assigns $m([a,b]) = b - a$ to an interval $[a,b]$. Lebesgue measure is atomless (see **3.22** below), every point of the line taking measure 0, hence $m((a,b)) = b - a$ also. $m(\mathbb{R}) = \infty$, but letting $((a,b], \mathcal{B}_{(a,b]}, m)$ denote the restriction of $(\mathbb{R}, \mathcal{B}, m)$ to a finite interval, m is a finite measure on $(a,b]$. Since $\mathbb{R}$ can be partitioned into a countable collection of finite intervals, m is σ-finite. □

Additivity is an intuitively plausible property of Lebesgue measure, thinking in terms of measuring the total length of a collection of disjoint intervals.

3.3 Example In the measurable space $(\mathbb{R}^2, \mathcal{B}^2)$ (see **1.24**) the Lebesgue measure of the rectangle $[a_1, b_1] \times [a_2, b_2]$ is $m = (b_1 - a_1)(b_2 - a_2)$. Simply enough, this is the area of the rectangle and is the same whether the defining intervals are open or closed. This is an example of a *product measure*, to be further discussed in §3.5. □

3.4 Example Consider the set of integers $\mathbb{N}$ and the set $\mathcal{B}_{\mathbb{N}}$ of subsets of $\mathbb{N}$. *Counting measure* defined on a set of $\mathcal{B}_{\mathbb{N}}$ is simply the number of elements contained in the set. It is evidently a σ-finite measure. □

Some additional properties of measures may be deduced from Definition **3.1**.

3.5 Theorem For arbitrary $\mathcal{F}$-sets A, B and $\{A_j, j \in \mathbb{N}\}$,

(i) $A \subseteq B \Rightarrow \mu(A) \leq \mu(B)$ (monotonicity)

(ii) $\mu(A \cup B) + \mu(A \cap B) = \mu(A) + \mu(B)$

(iii) $\mu\left(\bigcup_j A_j\right) \leq \sum_j \mu(A_j)$ (countable subadditivity).

Proof To show (i) note that A and $B - A$ are disjoint sets whose union is B, by hypothesis, and use **3.1**(a) and **3.1**(c). To show (ii), use $A \cup B = A \cup (B - A)$ and $B = (A \cap B) \cup (B - A)$, where again the sets in each union are disjoint. The result follows on application of **3.1**(c). To show (iii), define $B_1 = A_1$ and $B_n = A_n - \bigcup_{j=1}^{n-1} A_j$. Note that the sets B_n are disjoint, that $B_n \subseteq A_n$, and that $\bigcup_{j=1}^{\infty} B_j = \bigcup_{j=1}^{\infty} A$. Hence,

$$\mu\Big(\bigcup_{j=1}^{\infty} A_j\Big) = \mu\Big(\bigcup_{j=1}^{\infty} B_j\Big) = \sum_{j=1}^{\infty} \mu(B_j) \le \sum_{j=1}^{\infty} \mu(A_j) \tag{3.3}$$

where the inequality follows from (i). ∎

This proof illustrates a standard technique of measure theory, of converting a sequence of sets into a disjoint sequence having the same union by taking differences. This trick will become familiar in numerous later applications.

The idea behind **3.5**(ii) can be extended to give an expression for the measure of any finite union. This is the *inclusion–exclusion formula.*

3.6 Theorem

$$\mu\Big(\bigcup_{j=1}^{n} A_j\Big) = \sum_{j=1}^{n} \mu(A_j) - \sum_{k \ne j} \mu(A_j \cap A_k) + \sum_{k \ne j \ne l} \mu(A_j \cap A_k \cap A_l) - \dots \pm \mu(A_1 \cap A_2 \cap \dots \cap A_n) \tag{3.4}$$

where there are $2^n - 1$ terms in the sum in total and the sign of the last term is negative if n is even and positive if n is odd.

Proof This by induction from **3.5**(ii), substituting for the terms on the right-hand side of

$$\mu\Big(\bigcup_{j=1}^{n} A_j\Big) = \mu(A_n) + \mu\Big(\bigcup_{j=1}^{n-1} A_j\Big) - \mu\Big(\bigcup_{j=1}^{n-1} A_j \cap A_n\Big) \tag{3.5}$$

repeatedly, for $n - 1, n - 2, \dots, 1$. ∎

Let $\{A_n, n \in \mathbb{N}\}$ be a monotone sequence of $\mathcal{F}$-sets with limit $A \in \mathcal{F}$. A set function μ is said to be *continuous* if $\mu(A_n) \to \mu(A)$.

3.7 Theorem A finite measure is continuous.

Proof First let $\{A_n\}$ be increasing, with $A_{n-1} \subseteq A_n$ and $A = \bigcup_{n=1}^{\infty} A_n$. The sequence $\{B_j, j \in \mathbb{N}\}$, where $B_1 = A_1$ and $B_j = A_j - A_{j-1}$ for $j > 1$, is disjoint by construction with $B_j \in \mathcal{F}$, $A_n = \bigcup_{j=1}^{n} B_j$, and

$$\mu(A_n) = \sum_{j=1}^{n} \mu(B_j). \tag{3.6}$$

The real sequence $\{\mu(A_n)\}$ is therefore monotone and converges since it is bounded above by $\mu(\Omega) < \infty$. Countable additivity implies $\sum_{j=1}^{\infty} \mu(B_j) = \mu(\bigcup_{j=1}^{\infty} B_j) = \mu(A)$. Alternatively, let $\{A_n\}$ be decreasing, with $A_{n-1} \supseteq A_n$ and $A = \bigcap_{n=1}^{\infty} A_n$. Consider the increasing sequence $\{A_n^c\}$, determine $\mu(A^c)$ by the same argument, and use finite additivity to conclude that $\mu(A) = \mu(\Omega) - \mu(A^c)$ is the limit of $\mu(A_n) = \mu(\Omega) - \mu(A_n^c)$. ∎

The finiteness of the measure is needed for the second part of the argument, but the result that $\mu(A_n) \to \mu(A)$ when $A_n \uparrow A$ actually holds generally, not excluding the case $\mu(A) = \infty$. Theorem **3.7** has a partial converse:

3.8 Theorem A non-negative set function μ which is finitely additive and continuous is countably additive.

Proof Let $\{B_n\}$ be a countable, disjoint sequence. If $A_n = \bigcup_{j=1}^{n} B_j$, the sequence $\{A_n\}$ is increasing, $B_n \cap A_{n-1} = \varnothing$ and so $\mu(A_n) = \mu(B_n) + \mu(A_{n-1})$ for every n, by finite additivity. Given non-negativity, it follows by induction that $\{\mu(A_n)\}$ is monotone. If $A = \bigcup_{j=1}^{\infty} B_j$, $\mu(A) = \sum_{j=1}^{\infty} \mu(B_j)$, whereas continuity implies that $\mu(A) = \mu(\bigcup_{j=1}^{\infty} B_j)$. ∎

The following inequalities hold for general set sequences, which are not necessarily monotone.

3.9 Theorem For any sequence of $\mathcal{F}$-sets $\{A_n\}$ and finite measure μ,

(i) $\limsup_n \mu(A_n) \leq \mu(\limsup_n A_n)$

(ii) $\liminf_n \mu(A_n) \geq \mu(\liminf_n A_n)$.

Proof If the sequence is monotone, either increasing or decreasing with limit $A \in \mathcal{F}$, the relations both reduce to $\mu(A) = \mu(A)$ by **3.7**. Otherwise, define the non-increasing sequence $B_n = \bigcup_{m=n}^{\infty} A_n$ where $B_\infty = \limsup_n A_n$. Since $\mu(A_n) \leq \mu(B_n)$ for any choice of n,

$$\limsup_n \mu(A_n) = \inf_n \sup_{m\geq n} \mu(A_m) \leq \inf_n \sup_{m\geq n} \mu(B_m) \leq \inf_n \mu(B_n) = \mu(B_\infty)$$

proving (i). Similarly, let $C_n = \bigcap_{m=n}^{\infty} A_n$ so that $C_\infty = \liminf_n A_n$ and $\mu(A_n) \geq \mu(C_n)$ for any choice of n. Then,

$$\liminf_n \mu(A_n) = \sup_n \inf_{m\geq n} \mu(A_m) \geq \sup_n \inf_{m\geq n} \mu(C_m) \geq \sup_n \mu(C_n) = \mu(C_\infty)$$

which proves (ii). ∎

The following corollary of **3.9** is immediate.

3.10 Corollary If the sequence $\{A_n\}$ converges to a limit A, $\mu(A_n) \to \mu(A)$. □

Arguments in the theory of integration often turn on the notion of a 'negligible' set. In a measure space $(\Omega, \mathcal{F}, \mu)$, a *set of measure zero* is (simply enough) a set $M \in \mathcal{F}$ with $\mu(M) = 0$. Lest it be thought that such a set must have a finite or at most a countable number of points, the following classic example is illustrative.

3.11 Example The Cantor set is defined by recursively removing portions of the unit interval, as follows. Divide the interval $[0,1]$ into three equal parts and discard the open middle interval $(\frac{1}{3}, \frac{2}{3})$, leaving the set $C_1 = [0, \frac{1}{3}] \cup [\frac{2}{3}, 1]$. Now remove the open middle third of each of these intervals to give $C_2 = [0, \frac{1}{9}] \cup [\frac{2}{9}, \frac{1}{3}] \cup [\frac{2}{3}, \frac{7}{9}] \cup [\frac{8}{9}, 1]$. Define C_n for $n = 3, 4, 5, \ldots$ by the same scheme and the Cantor set is $C = \bigcap_{j=1}^{\infty} C_j$. The lengths of the removed segments are respectively $\frac{1}{3}$, $2(\frac{1}{3})^2$, $4(\frac{1}{3})^3$, $8(\frac{1}{3})^4$, etc., which sum to $\frac{1}{2}\sum_{n=1}^{\infty}(\frac{2}{3})^n = 1$. Hence, the complementary set C has Lebesgue measure zero. □

The endpoints of the retained intervals making up C in **3.11** are multiples of integer powers of $\frac{1}{3}$. These points are countable but since the intervals have length zero they form a dense subset of C. Albeit the union of a countable collection of intervals of length zero, C is nonetheless an uncountable set. Viewing elements of the unit interval in the base-3 (ternary) representation of the reals (i.e. the case of (1.15) with $m = 1$, $p = 0$, and $D = 3$), the elements of C are all those whose expansions contain only the digits 0 and 2. The middle-third operations remove all cases with 1 appearing in the ternary expansion but leave the remainder. The elements of C can therefore be paired with the points of $[0,1]$, with the latter expressed in binary representation ($m = 1$, $p = 0$, $D = 2$), by replacing the 2s in their expansions by 1s.

A condition or restriction on the elements of Ω is said to occur *almost everywhere* (a.e.) if it holds on a set E and $\Omega - E$ has measure zero. If more than one measure is assigned to the same space, it may be necessary to indicate which measure the statement applies to, by writing a.e.$[\mu]$ or a.e.$[\nu]$ as the case may be.

3.12 Theorem

(i) If M and N are $\mathcal{F}$-sets, M has measure 0 and $N \subseteq M$, then N has measure 0.

(ii) If $\{M_j\}$ is a countable sequence with $\mu(M_j) = 0$, $\forall j$, then $\mu(\bigcup_j M_j) = 0$.

(iii) If $\{E_j\}$ is a countable sequence with $\mu(E_j^c) = 0$, $\forall j$, then $\mu\big((\bigcap_j E_j)^c\big) = 0$.

Proof (i) is an application of monotonicity; (ii) is a consequence of countable additivity; and (iii) follows likewise, using the second de Morgan law. ∎

Sections §3.2 and §3.3 below consider the issue of the measurability of the sets in a given space. The essential conclusion of the extension theorem is that if the sets of a given collection of suitable type are measurable then the sets of the σ-field generated by that collection are also measurable. For many purposes this fact is sufficient, but there may be sets outside the σ-field that can be shown in other ways to be measurable. For example, if $\mu(A) = \mu(B)$ it would seem reasonable to assign $\mu(E) = \mu(A)$ whenever $A \subset E \subset B$. This is equivalent to assigning measure 0 to any subset of a set of measure 0.

3.13 Example It is shown in §3.3 that non-measurable subsets of $\mathbb{R}$ can be constructed. In the plane $\mathbb{R}^2$ (see **1.24**) consider the set $\{x\} \times A$ where $\{x\}$ is a singleton set whose Lebesgue measure is $m(\{x\}) = 0$ and A is non-measurable. In this case the Lebesgue measure of the set $\{x\} \times A$ is undefined. However, let E be any interval such that $A \subset E$. As in Example **3.3** the Lebesgue measure of $\{x\} \times E$ is $m(E)$ multiplied by 0, which is 0. It therefore appears reasonable to assign measure 0 to $\{x\} \times A$. □

Including such sets is known as completing a measure space. The space $(\Omega, \mathcal{F}, \mu)$ is said to be *complete* if for any $E \in \mathcal{F}$ with $\mu(E) = 0$, all subsets of E are also in $\mathcal{F}$ and likewise all sets F such that $\mu(E \triangle F) = 0$. According to the following result, every measure space can be completed without changing any conclusions except in respect of these negligible sets.

3.14 Theorem Given any measure space $(\Omega, \mathcal{F}, \mu)$, there exists a complete measure space $(\Omega, \mathcal{F}^{\mu}, \bar{\mu})$, called the *completion* of $(\Omega, \mathcal{F}, \mu)$, such that $\mathcal{F} \subseteq \mathcal{F}^{\mu}$ and $\mu(E) = \bar{\mu}(E)$ for all $E \in \mathcal{F}$. □

Notice that the completion of a space is defined with respect to a particular measure. The measurable space $(\Omega, \mathcal{F})$ has a different completion for each measure that can be defined on it. However, $\mathcal{F}^U = \bigcap_\mu \mathcal{F}^\mu$ where the intersection is taken over all measures defined on the space defines the *universally measurable* sets of Ω.

Proof of 3.14 Let N^μ denote the collection of all subsets of $\mathcal{F}$-sets of μ-measure 0 and let

$$\mathcal{F}^\mu = \{F \subseteq \Omega : E \triangle F \in N^\mu \text{ for some } E \in \mathcal{F}\}. \tag{3.7}$$

If $\mu(E) = 0$, any set $F \subset E$ satisfies the criterion of (3.7) and so is in $\mathcal{F}^\mu$ as the definition requires. For $F \in \mathcal{F}^\mu$, let $\bar{\mu}(F) = \mu(E)$ where E is any $\mathcal{F}$-set satisfying $E \triangle F \in N^\mu$. To show that the choice of E is immaterial, let E_1 and E_2 be two such sets and note that

$$\mu(E_1 \triangle E_2) = \mu\big((F \triangle E_1) \triangle (F \triangle E_2)\big) = 0. \tag{3.8}$$

(Use a Venn diagram as in Figure 1.1 on page 5 to partition $\Omega \subseteq \mathbb{R}^2$ and show that the sets in question match.) Since $\mu(E_1 \cup E_2) = \mu(E_1 \cap E_2) + \mu(E_1 \triangle E_2)$, it follows from (3.8) that

$$\mu(E_1 \cap E_2) \geq \mu(E_i) \geq \mu(E_1 \cap E_2) \tag{3.9}$$

for $i = 1$ and 2, or $\mu(E_1) = \mu(E_2)$. Hence, the measure is unique. When $F \in \mathcal{F}$ choose $E = F$, since $F \triangle F = \emptyset \in N^\mu$, confirming that the measures agree on $\mathcal{F}$.

It remains to be shown that $\mathcal{F}^\mu$ is a σ-field containing $\mathcal{F}$. Choosing $E = F$ in (3.7) for $F \in \mathcal{F}$ shows $\mathcal{F} \subseteq \mathcal{F}^\mu$. If $F \in \mathcal{F}^\mu$, then $E \triangle F \in N^\mu$ for $E \in \mathcal{F}$ and hence $E^c \triangle E \triangle F \in N^\mu$ where $E^c \in \mathcal{F}$ and so $F^c \in \mathcal{F}^\mu$. And finally if $F_j \in \mathcal{F}^\mu$ for $j \in \mathbb{N}$, there exist $E_j \in \mathcal{F}$ for $j \in \mathbb{N}$ such that $E_j \triangle F_j \in N^\mu$. Hence

$$\Big(\bigcup_j E_j\Big) \triangle \Big(\bigcup_j F_j\Big) \subseteq \bigcup_j (E_j \triangle F_j) \in N^\mu \tag{3.10}$$

by **3.12**(ii), which means that $\bigcup_j F_j \in \mathcal{F}^\mu$, completing the proof. The inclusion in (3.10) follows since an element of the left-hand set belongs either to one of the E_j and none of the F_j, or to one of the F_j and none of the E_j. The right-hand set, on the other hand, may contain elements belonging to both an E_j and a $F_{j'}$ for $j' \neq j$. ∎

3.2 The Extension Theorem

In the definition of a measurable space, why could not $\mathcal{F}$ simply be the set of *all* subsets, the power set of Ω? The problem is to find a consistent method of assigning a measure to every set. This is straightforward when the space has a finite number of elements, but not in an infinite space where there is no way, even conceptually, to assign specific measures to individual sets. The problem of measurability is to show that measures can be assigned to collections of sets too large for the use of constructive methods without running into inconsistencies. This problem can be solved for σ-fields, which are sufficiently rich collections to cope with most situations arising in probability.

Begin by assigning a measure, to be denoted μ_0, to the members of some basic collection $\mathcal{C}$ for which this can feasibly be done. For example, Lebesgue measure on $\mathbb{R}$ assigns to each interval $(a, b]$ the measure $b - a$. Provided $\mathcal{C}$ is rich enough to allow μ_0 to be uniquely defined by it, reasoning from the properties of μ_0 permits extension to all the sets of interest. A collection $\mathcal{C} \subseteq \mathcal{F}$ is called a *determining class* for $(\Omega, \mathcal{F})$ if, whenever μ and ν are measures on $\mathcal{F}$, $\mu(A) = \nu(A)$ for all $A \in \mathcal{C}$ implies that $\mu = \nu$.

Given $\mathcal{C}$ there must be a way to assign μ_0-values to sets derived from $\mathcal{C}$ by operations such as union, intersection, complementation, and difference. For disjoint sets A and B, $\mu_0(A \cup B) = \mu_0(A) + \mu_0(B)$ by finite additivity. There are the rules $\mu_0(A \cap B) = \mu_0(A) + \mu_0(B) - \mu_0(A \cup B)$, $\mu_0(A - B) = \mu_0(A) - \mu_0(B)$ for $B \subseteq A$ including the case $A = \Omega$; and so forth. When such assignments are possible for any pair of sets whose measures are themselves known, the measure is thereby extended to a wider class of sets, to be denoted $\mathcal{S}$. Often $\mathcal{S}$ and $\mathcal{C}$ are the same collection, but in any event $\mathcal{S}$ must be closed under various finite set operations and at least be a semi-ring. In the applications $\mathcal{S}$ is typically either a field (algebra) or a semi-algebra. Example **1.18** is a good case to keep in mind.

However, $\mathcal{S}$ cannot be a σ-field since at most a finite number of operations are permitted to determine $\mu_0(A)$ for any $A \in \mathcal{S}$. At this point, consider the opposite question and ask why $\mathcal{S}$ might not itself be a rich enough collection? In fact, sets of interest frequently arise which $\mathcal{S}$ cannot contain. **3.22** below illustrates the necessity of being able to go to the limit and consider sets that are expressible only as countably infinite unions or intersections of $\mathcal{C}$-sets. Extending to the sets $\mathcal{F} = \sigma(\mathcal{S})$ proves indispensable.

The desired result is Carathéodory's extension theorem. This falls into two parts establishing, respectively, existence of the measure (**3.15**) and its uniqueness (**3.20**).

3.15 Theorem (extension, existence) Let $\mathcal{S}$ be a semi-ring and let $\mu_0 : \mathcal{S} \mapsto \bar{\mathbb{R}}^+$ be a measure on $\mathcal{S}$. If $\mathcal{F} = \sigma(\mathcal{S})$, there exists a measure μ on $(\Omega, \mathcal{F})$ such that $\mu(E) = \mu_0(E)$ for each $E \in \mathcal{S}$. □

Although the proof of the theorem is rather lengthy and some of the details are fiddly, the basic idea is simple. Take a set $A \subseteq \Omega$ to which a measure $\mu(A)$ is to be assigned. If $A \in \mathcal{S}$, $\mu(A) = \mu_0(A)$. If $A \notin \mathcal{S}$, consider choosing a finite or countable covering for A from members of $\mathcal{S}$; that is, a selection of sets $E_j \in \mathcal{S}, j = 1, 2, 3, \dots$ such that $A \subseteq \bigcup_j E_j$. The object is to find as 'economical' a covering as possible in the sense that $\sum_j \mu_0(E_j)$ is as small as possible. The *outer measure* of A is

$$\mu^*(A) = \inf \sum_j \mu_0(E_j) \tag{3.11}$$

where the infimum is taken over all finite and countable coverings of A by $\mathcal{S}$-sets. If no such covering exists, set $\mu^*(A) = \infty$. Clearly, $\mu^*(A) = \mu_0(A)$ for each $A \in \mathcal{S}$. μ^* is called the outer measure because for any eligible definition of $\mu(A)$,

$$\mu^*(A) \geq \sum_j \mu(E_j) \geq \mu\left(\bigcup_j E_j\right) \geq \mu(A), \text{ for } E_j \in \mathcal{S}. \tag{3.12}$$

The first inequality here is by the stipulation that $\mu(E_j) = \mu_0(E_j)$ for $E_j \in \mathcal{S}$ in the case where a covering exists, in which case this is an equality by (3.11), or else the majorant side is infinite. The second and third follow by countable subadditivity and monotonicity respectively, because μ is a measure.

Next, consider trying to construct a minimal covering for A^c and so define the inner measure of A. In the case where $\Omega \in \mathcal{S}$ and $\mu(\Omega) = \mu_0(\Omega) < \infty$, the inner measure may be defined as

$$\mu_*(A) = \mu(\Omega) - \mu^*(A^c).$$

Since $\mu^*(A^c) \geq \mu(A^c)$ by (3.12), it must be the case that for any eligible definition of μ,

$$\mu(A) = \mu(\Omega) - \mu(A^c) \geq \mu_*(A). \tag{3.13}$$

Then, if $\mu^*(A) = \mu_*(A)$ it would make sense to call this common value the measure of A and say that A is measurable. In fact the measurability criterion employed is more stringent than this, but also more robust. A set $A \subseteq \Omega$ is said to be *measurable* if for any $B \subseteq \Omega$,

$$\mu^*(A \cap B) + \mu^*(A^c \cap B) = \mu^*(B). \tag{3.14}$$

This yields $\mu^*(A) = \mu_*(A)$ as the special case on putting $B = \Omega$, with $\mu(\Omega) < \infty$, but the criterion remains valid even if $\mu(\Omega) = \infty$.

Let $\mathcal{M}$ denote the collection of all measurable sets, those subsets of Ω satisfying (3.14). Since $\mu^*(A) = \mu_0(A)$ for $A \in \mathcal{S}$ and $\mu_0(\emptyset) = 0$, putting $A = \emptyset$ in (3.14) gives the trivial equality $\mu^*(B) = \mu^*(B)$. Hence $\emptyset \in \mathcal{M}$ and since the definition implies that $A^c \in \mathcal{M}$ if $A \in \mathcal{M}$, $\Omega \in \mathcal{M}$ too.

The next steps are to determine what properties the set function $\mu^* : \mathcal{M} \mapsto \bar{\mathbb{R}}$ shares with a measure. Clearly,

$$\mu^*(A) \geq 0 \text{ for all } A \subseteq \Omega. \tag{3.15}$$

Another property that follows directly from the definition of μ^* is monotonicity:

$$A_1 \subseteq A_2 \Rightarrow \mu^*(A_1) \leq \mu^*(A_2), \text{ for } A_1, A_2 \subseteq \Omega. \tag{3.16}$$

The goal is to show that countable additivity also holds for μ^* in respect of $\mathcal{M}$-sets, but it proves convenient to begin by establishing countable *sub*additivity.

3.16 Lemma If $\{A_j, j \in \mathbb{N}\}$ is any sequence of subsets of Ω, then

$$\mu^*\Big(\bigcup_j A_j\Big) \leq \sum_j \mu^*(A_j). \tag{3.17}$$

Proof Assume $\mu^*(A_j) < \infty$ for each j. (If not, the result is trivial.) For each j let $\{E_{jk}\}$ denote a countable covering of A_j by $\mathcal{S}$-sets satisfying

$$\sum_k \mu_0(E_{jk}) < \mu^*(A_j) + 2^{-j}\varepsilon$$

for any $\varepsilon > 0$. Such a collection always exists by the definition of μ^*. Since $\bigcup_j A_j \subseteq \bigcup_{j,k} E_{jk}$ and $\sum_{j=1}^{\infty} 2^{-j} = 1$ it follows by definition that

$$\mu^*\Big(\bigcup_j A_j\Big) \leq \sum_{j,k} \mu_0(E_{jk}) < \sum_j \mu^*(A_j) + \varepsilon. \tag{3.18}$$

(3.17) now follows since ε is arbitrary and the last inequality is strict. ∎

The following is an immediate consequence of the lemma. Since (3.14) defines measurability and $\mu^*(A \cap B) + \mu^*(A^c \cap B) \geq \mu^*(B)$ follows by subadditivity as in (3.17), the reverse inequality must also hold.

3.17 Corollary $A \in \mathcal{M}$ if, for any $B \subseteq \Omega$

$$\mu^*(A \cap B) + \mu^*(A^c \cap B) \leq \mu^*(B). \quad \square \tag{3.19}$$

The following lemma is central to the proof of the extension theorem. It yields countable additivity as a corollary, but also has a wider purpose.

3.18 Lemma $\mathcal{M}$ is a monotone class.

Proof Letting $\{A_n, n \in \mathbb{N}\}$ be an increasing sequence of $\mathcal{M}$-sets converging to $A = \bigcup_n A_n$, the object is to show $A \in \mathcal{M}$. For $n > 1$ and $E \subseteq \Omega$, the definition of an $\mathcal{M}$-set gives

$$\begin{aligned}\mu^*\big(A_n \cap E\big) &= \mu^*(A_{n-1} \cap (A_n \cap E)\big) + \mu^*\big(A^c_{n-1} \cap (A_n \cap E)\big) \\ &= \mu^*(A_{n-1} \cap E) + \mu^*(B_n \cap E) \qquad (3.20)\end{aligned}$$

where $B_n = A_n - A_{n-1}$ and the sequence $\{B_n\}$ is disjoint. Put $A_0 = \varnothing$ so that $\mu^*(A_0 \cap E) = 0$; then by induction on (3.20),

$$\mu^*(A_n \cap E) = \sum_{j=1}^{n} \mu^*(B_j \cap E) \qquad (3.21)$$

holds for every n. The right-hand side of (3.21) for $n \in \mathbb{N}$ is a monotone real sequence and $\mu^*(A_n \cap E) \to \mu^*(A \cap E)$ as $n \to \infty$. Now, since $A_n \in \mathcal{M}$,

$$\begin{aligned}\mu^*(E) &= \mu^*(A_n \cap E) + \mu^*(A^c_n \cap E) \\ &\geq \mu^*(A_n \cap E) + \mu^*(A^c \cap E) \qquad (3.22)\end{aligned}$$

using the monotonicity of μ^* and the fact that $A^c \subseteq A^c_n$. Taking the limit, the foregoing argument gives

$$\mu^*(E) \geq \mu^*(A \cap E) + \mu^*(A^c \cap E) \qquad (3.23)$$

so that $A \in \mathcal{M}$ by **3.17**. For the case of a decreasing sequence, simply move to the complements and argue as above. ∎

Notice how **3.17** was needed in the proof of **3.18**. Showing that (3.14) holds for the limit, given (3.17), is done by showing (3.23). Additivity rather than subadditivity is the fundamental property of a measure, but since $\{B_j\}$ is a disjoint sequence, countable additivity emerges as a by-product of the lemma, as the following corollary shows.

3.19 Corollary If $\{B_j\}$ is a disjoint sequence of $\mathcal{M}$-sets,

$$\mu^*\Big(\bigcup_j B_j\Big) = \sum_j \mu^*(B_j). \qquad (3.24)$$

Proof This is immediate on putting $E = \Omega$ in (3.21) and letting $n \to \infty$, noting $\bigcup_j B_j = A$. ∎

Proof of 3.15 Equations (3.15) and (3.24) show that μ^* is a measure for the elements of $\mathcal{M}$. If it can be shown that $\mathcal{F} \subseteq \mathcal{M}$, setting $\mu(A) = \mu^*(A)$ for all $A \in \mathcal{F}$ will satisfy the existence criteria of the theorem.

The first step is to show that $\mathcal{S} \subseteq \mathcal{M}$ or, by **3.17**, that $A \in \mathcal{S}$ implies

$$\mu^*(E \cap A) + \mu^*(E \cap A^c) \leq \mu^*(E) \tag{3.25}$$

for any $E \subseteq \Omega$. Let $\{A_j \in \mathcal{S}\}$ denote a finite or countable covering of E such that

$$\sum_j \mu_0(A_j) < \mu^*(E) + \varepsilon$$

for $\varepsilon > 0$. If no such covering exists, $\mu^*(E) = \infty$ by definition and (3.25) holds trivially. Note that

$$E \cap A \subseteq \bigcup_j (A_j \cap A) \tag{3.26}$$

and since $\mathcal{S}$ is a semi-ring the sets $A_j \cap A$ are in $\mathcal{S}$. Similarly,

$$E \cap A^c \subseteq \bigcup_j (A_j \cap A^c) \tag{3.27}$$

and, by set algebra and the definition of a semi-ring,

$$A_j \cap A^c = A_j - (A_j \cap A) = \bigcup_k C_{jk} \tag{3.28}$$

where the C_{jk} are a finite collection of $\mathcal{S}$-sets, disjoint with each other and also with $A_j \cap A$. Now, applying **3.16** and combining (3.26), (3.27), (3.28), and the fact that $\mu^*(B) = \mu_0(B)$ for $B \in \mathcal{S}$,

$$\begin{aligned} \mu^*(E \cap A) + \mu^*(E \cap A^c) &\leq \sum_j \mu_0(A_j \cap A) + \sum_j \sum_k \mu_0(C_{jk}) \\ &= \sum_j \mu_0(A_j) < \mu^*(E) + \varepsilon \end{aligned} \tag{3.29}$$

where the equality follows from (3.28) because μ_0 is finitely additive and $A_j \cap A$ and the C_{jk} are mutually disjoint. Since ε is arbitrary, (3.25) follows.

The next step is to show that $\mathcal{M}$ is a σ-field. It is sufficient to show that $\mathcal{M}$ is a field, because **3.18** implies it is also a σ-field by **1.25**. $\Omega \in \mathcal{M}$ and $\mathcal{M}$ is closed under complementation, so it remains to be shown that unions of $\mathcal{M}$-sets are in $\mathcal{M}$. Suppose that A_1 and A_2 are $\mathcal{M}$-sets and $E \subseteq \Omega$. Then

$$\begin{aligned}\mu^*(E) &= \mu^*(A_1 \cap E) + \mu^*(A_1^c \cap E) \\ &= \mu^*(A_2 \cap A_1 \cap E) + \mu^*(A_2^c \cap A_1 \cap E) \\ &\qquad\qquad + \mu^*(A_2 \cap A_1^c \cap E) + \mu^*(A_2^c \cap A_1^c \cap E) \\ &\geq \mu^*(A_2 \cap A_1 \cap E) \\ &\quad + \mu^*\big((A_2^c \cap A_1 \cap E) \cup (A_2 \cap A_1^c \cap E) \cup (A_2^c \cap A_1^c \cap E)\big) \\ &= \mu^*\big((A_2 \cap A_1) \cap E\big) + \mu^*\big((A_2 \cap A_1)^c \cap E\big) \qquad (3.30)\end{aligned}$$

where the inequality is by subadditivity and the rest is set algebra. By **3.17** this is sufficient for $A_1 \cap A_2 \in \mathcal{M}$ and hence also for $A_1 \cup A_2 \in \mathcal{M}$, using closure under complementation.

It follows that $\mathcal{M}$ is a σ-field containing $\mathcal{S}$ and since $\mathcal{F}$ is the smallest such σ-field, $\mathcal{F} \subseteq \mathcal{M}$ as required. ∎

Notice that (3.30) was got by using (3.14) as the relation defining measurability. The proof does not go through using $\mu^*(A) = \mu_*(A)$ as the definition.

The style of this argument shows some important things about the role of $\mathcal{S}$. Any set that has no covering by $\mathcal{S}$-sets is assigned the measure ∞, so for finite measures it is a requisite that $\Omega \subseteq \bigcup_j E_j$ for a finite or countable collection $\{E_j \in \mathcal{S}\}$. The measure of a union of $\mathcal{S}$-sets must be able to approximate the measure of any $\mathcal{F}$-set arbitrarily well and the basic content of the theorem is to establish that a semi-ring has this property.

To complete the demonstration of the extension, there remains the question of uniqueness. To get this result σ-finiteness must be imposed, which was not needed for existence.

3.20 Theorem (extension, uniqueness) Let μ and μ' denote measures on a space $(\Omega, \mathcal{F})$, where $\mathcal{F} = \sigma(\mathcal{S})$ and $\mathcal{S}$ is a semi-ring. If the measures are σ-finite on $\mathcal{S}$ and $\mu(E) = \mu'(E)$ for all $E \in \mathcal{S}$, then $\mu(E) = \mu'(E)$ for all $E \in \mathcal{F}$. □

It is perhaps worth remarking at this point that σ-finiteness is an attribute that arises only occasionally in the applications studied in the sequel. The chief concern is going to be with finite measures, specifically probability measures having the memorable property $\mu(\Omega) = 1$. The σ-finite case is included in this treatment for completeness and because the issues have some interest for their own sake; it is of course the key feature of Lebesgue measure on $\mathbb{R}$. The reader may nonetheless choose to skim at first reading those parts of proofs and examples that deal with it, such as the second part of the following.

Proof of 3.20 The theorem is proved for the case of finite measures by an application of the π-λ theorem. Define $\mathcal{A} = \{E \in \mathcal{F} : \mu(E) = \mu'(E)\}$. Then $\mathcal{S} \subseteq \mathcal{A}$ by hypothesis. If $\mathcal{S}$ is a semi-ring, it is also a π-system. By **1.29**, the proof is completed if $\mathcal{A}$ is a λ-system and hence contains $\sigma(\mathcal{S})$.

When the measure is finite, $\Omega \in \mathcal{A}$ and condition **1.27**(a) holds. Additivity implies that, for $A \in \mathcal{A}$,

$$\mu(A^c) = \mu(\Omega) - \mu(A) = \mu'(\Omega) - \mu'(A) = \mu'(A^c), \tag{3.31}$$

so that **1.27**(b) holds. Lastly, let $\{A_j\}$ be a disjoint sequence in $\mathcal{A}$. By countable additivity

$$\mu\Bigl(\bigcup_{j=1}^{\infty} A_j\Bigr) = \sum_{j=1}^{\infty} \mu(A_j) = \sum_{j=1}^{\infty} \mu'(A_j) = \mu'\Bigl(\bigcup_{j=1}^{\infty} A_j\Bigr) \tag{3.32}$$

and **1.27**(c) holds. It follows by **1.27** and **1.29** that $\mathcal{F} = \sigma(\mathcal{S}) \subseteq \mathcal{A}$.

Now consider the σ-finite case. Let $\Omega = \bigcup_j B_j$ where $B_j \in \mathcal{S}$ and $\mu(B_j) = \mu'(B_j) < \infty$. $\mathcal{F}_j = \{B_j \cap A : A \in \mathcal{F}\}$ is a σ-field, so that the $(B_j, \mathcal{F}_j)$ are measurable spaces on which μ and μ' are finite measures agreeing on $\mathcal{S} \cap \mathcal{B}_j$. The preceding argument showed that, for $A \in \mathcal{F}$, $\mu(B_j \cap A) = \mu'(B_j \cap A)$ only if μ and μ' are the same measure.

Consider the following recursion. By **3.5**(ii),

$$\mu\bigl(A \cap (B_1 \cup B_2)\bigr) = \mu(A \cap B_1) + \mu(A \cap B_2) - \mu(A \cap B_1 \cap B_2). \tag{3.33}$$

Letting $C_n = \bigcup_{j=1}^{n} B_j$ the same relation yields

$$\mu(A \cap C_n) = \mu(A \cap B_n) + \mu(A \cap C_{n-1}) - \mu(A \cap B_n \cap C_{n-1}). \tag{3.34}$$

The terms involving C_{n-1} on the right-hand side can be solved backwards to yield an expression for $\mu(A \cap C_n)$, as a sum of terms having the general form

$$\mu(A \cap B_{j_1} \cap B_{j_2} \cap B_{j_3} \cap \ldots) = \mu(D \cap B_j) < \infty \tag{3.35}$$

for *some* j, say $j = j_1$, in which case $D = A \cap B_{j_2} \cap B_{j_3} \cap \ldots \in \mathcal{F}$. Since $\mu(D \cap B_j) = \mu'(D \cap B_j)$ for all $D \in \mathcal{F}$ by the preceding argument, it follows that in (3.34)

$$\mu(A \cap C_n) = \mu'(A \cap C_n). \tag{3.36}$$

This holds for any n. Since $C_n \to \Omega$ as $n \to \infty$,

$$\mu(A) = \mu'(A) \tag{3.37}$$

where the two sides of the equality are either finite and equal or both equal to $+\infty$. This completes the proof, since A is arbitrary. ∎

3.21 Example Let $\mathcal{M}$ denote the subsets of $\mathbb{R}$ which are measurable according to (3.14) when μ^* is the outer measure defined on the half-open intervals, whose measures μ_0 are taken equal to their lengths. This defines Lebesgue measure m. These sets form a semi-ring by **1.18**, a countable collection of them covers $\mathbb{R}$, and the extension theorem shows that, given m is a σ-finite measure, $\mathcal{M}$ contains the Borel field on $\mathbb{R}$ (see **1.21**), so $(\mathbb{R}, \mathcal{B}, m)$ is a measure space. It can be shown (although this demonstration is omitted) that all the Lebesgue-measurable sets *not* in $\mathcal{B}$ are subsets of $\mathcal{B}$-sets of measure 0. □

The following is a basic property of Lebesgue measure. Notice the need to deal with a countable intersection of intervals to determine so simple a thing as the measure of a point.

3.22 Theorem Any countable set from $\mathbb{R}$ has Lebesgue measure 0.

Proof The measure of a point $\{x\}$ is zero, since for $x \in \mathbb{R}$,

$$\{x\} = \bigcap_{n=1}^{\infty} (x - 1/n, x] \in \mathcal{B} \tag{3.38}$$

and by continuity of the measure, $m(\{x\}) = \lim_{n\to\infty} 1/n = 0$. The result follows by **3.12**(ii). ∎

3.3 Non-measurability

To give the ideas of the last section their true force it needs to be shown that $\mathcal{M} \subset 2^{\Omega}$ is possible, in other words that Ω can contain non-measurable subsets. In this section, such a set is constructed in the half-open unit interval $(0, 1]$. This is a standard counterexample from Lebesgue theory.

For $x, y \in (0, 1]$, define the operator

$$y \dot{+} x = \begin{cases} y + x, & y + x \leq 1 \\ y + x - 1, & y + x > 1. \end{cases} \tag{3.39}$$

This is *addition modulo 1.* Imagine the unit interval mapped onto a circle, like a clock face with 0 at the top. $y\dotplus x$ is the point obtained by moving a hand clockwise through an angle of $2\pi x$ from an initial point y on the circumference. For each set $A \subseteq (0,1]$ and $x \in (0,1]$, define the set

$$A\dotplus x = \{y\dotplus x : y \in A\}. \tag{3.40}$$

3.23 Theorem If A is Lebesgue-measurable so is $A\dotplus x$ and $m(A\dotplus x) = m(A)$, for any x.

Proof For $(a,b] \subseteq (0,1]$, $m((a+x,b+x]) = b-a = m((a,b])$, for any real x such that $a+x > 0$ and $b+x \leq 1$. The property extends to finite unions of intervals translated by x. If A is any Lebesgue-measurable subset of $(0,1]$ and $A+x \subseteq (0,1]$ where $A+x = \{y+x : y \in A\}$, the construction of the extension similarly implies that $A+x$ is measurable and $m(A) = m(A+x)$.

Now let $A_1 = A \cap (0,1-x]$ and $A_2 = A \cap (1-x,1]$. Then $m(A_1+x) = m(A_1)$ and $m(A_2+x-1) = m(A_2)$, where the sets on the left-hand sides of these equalities are in each case contained in $(0,1]$. A_1+x and A_2+x-1 are disjoint sets whose union is $A\dotplus x$ and hence

$$\begin{aligned} m(A\dotplus x) &= m(A_1+x) + m(A_2+x-1) \\ &= m(A_1) + m(A_2) = m(A). \quad \blacksquare \end{aligned} \tag{3.41}$$

Define a relation for points of $(0,1]$ by letting xRy if $y = x\dotplus r$ for $r \in \mathbb{Q}$; that is, xRy if y is separated from x by a rational distance along the circle. R is an equivalence relation. Next, define the equivalence classes

$$E^x = \{y : y = x\dotplus r, r \in \mathbb{Q}\} \tag{3.42}$$

where the sets of the collection $\{E^x, x \in (0,1]\}$ are either identical or disjoint. Since every x is a rational distance from *some* other point of the interval, these sets cover $(0,1]$. A collection formed by choosing just one of each of the identical sets and discarding the duplicates is therefore a partition of $(0,1]$. Write this as $\{E^x, x \in C\}$, where C denotes the residual set of indices.

Another example may help the reader to visualize these sets. In the set of integers, the set of even integers is an equivalence class and can be defined as E^0, the set of integers which differ from 0 by an even integer. Of course $E^0 = E^2 = E^4 = \ldots = E^{2n}$, for any $n \in \mathbb{Z}$. The set of odd integers can be defined similarly as E^1, the set of integers differing by an even integer from 1. $E^1 = E^3 = \ldots = E^{2_n+1}$ for any $n \in \mathbb{Z}$. Discarding the redundant members of the collection $\{E^x, x \in \mathbb{Z}\}$ leaves just the collection $\{E^0, E^1\}$ to define a partition of $\mathbb{Z}$.

Now construct a set H by taking an element from E^x for each $x \in C$.

3.24 Theorem H is not Lebesgue-measurable.

Proof Consider the countable collection $\{H \dot{+} r, r \in \mathbb{Q}\}$. This collection is shown to be a partition of $(0, 1]$. To show disjointness, argue by contradiction. Suppose $z \in H \dot{+} r_1$ and $z \in H \dot{+} r_2$ for $r_1 \neq r_2$. This means there are points $h_1, h_2 \in H$, such that

$$h_1 \dot{+} r_1 = z = h_2 \dot{+} r_2. \tag{3.43}$$

If $r_1 \neq r_2$ then $h_1 = h_2$ is not possible, but if $h_1 \neq h_2$ then h_1 and h_2 belong to different equivalence classes by construction of H and cannot be a rational distance $|r_1 - r_2|$ apart; hence no z satisfying (3.43) exists. On the other hand, let $H^* = \bigcup_r (H \dot{+} r)$ and consider any point $x \in (0, 1]$. x belongs to one of the equivalence classes and hence is within a rational distance of some element of H; but H^* contains all the points that are a rational distance r from a point of H, for some r and hence $x \in H^*$ and it follows that $(0, 1] \subseteq H^*$.

Suppose $m(H)$ exists. Then by **3.23**, $m(H \dot{+} r) = m(H)$ for all r and either $m(H) = 0$ or $m(H) > 0$. The first case is ruled out by **3.12**(ii) since $m(H^*) \geq m((0, 1]) = 1$. However, if $m(H) > 0$, then countable additivity gives $m(H^*) = \sum_r m(H \dot{+} r) = \infty$, which is also impossible. It follows that $m(H)$ does not exist. ∎

The problem is that a countable collection of disjoint sets has been constructed, each having the same measure (were this measure defined) whose union covers the unit interval. These sets cannot have measure zero but nor can they have positive measure, hence the contradiction.

However, the definition of H is not without controversy since the set of equivalence classes is uncountable. It is not possible to devise *even in principle* constructive rules for selecting the set C and elements from E^x for each $x \in C$. Naïve set theory is based on the intuitive notion of what a 'set' is and makes no attempt to avoid paradoxical conclusions by specifying rules for the definition and construction of sets. Axiomatic set theory does this, but it turns out that the proposition that sets like H exist cannot be deduced as a consequence of the fundamental axioms. It must be asserted as an additional axiom, the so-called *axiom of choice.* If the validity of the axiom of choice is questioned, the counterexample fails. The question to be decided is: can a mathematical object exist that cannot be constructed even in imagination? Suffice it to say that most mathematicians think it can.

Sets like H do not belong to $\mathcal{B}_{(0,1]} = \{B \cap (0, 1], B \in \mathcal{B}\}$. All the sets of $\mathcal{B}_{(0,1]}$ are Lebesgue-measurable (see **3.21** and restrict m to $(0, 1]$ as in **3.2**), but while it might appear to the uninitiated that every conceivable set of points of the line must be constructible by countably infinite sequences of operations on intervals, this is very far from the truth. The Borel field of $[0, 1]$ might appear to be a big collection but it is nonetheless dwarfed by its complement in $2^{[0,1]}$. The possible ways of forming subsets from an uncountable number of elements defy the imagination, but what can be said, thankfully, is that the sets imagination can easily encompass are also those most likely to arise in practical applications of probability. Sticking with Borel sets avoids measurability difficulties on the line, but the example illustrates the need for caution. In less familiar situations (such as will arise in Part VI) measurability can fail in superficially plausible cases.

However, if measurability is ever in doubt remember that outer measure μ^* is well defined for all subsets of Ω and coincides with μ whenever the latter is defined. There are situations in which measurability problems are best dealt with by working explicitly with outer measure and forgetting about them.

3.4 Product Spaces

If $(\Omega, \mathcal{F})$ and $(\Xi, \mathcal{G})$ are two measurable spaces, let

$$\Omega \times \Xi = \{(\omega, \xi) : \omega \in \Omega, \xi \in \Xi\} \tag{3.44}$$

be the Cartesian product of Ω and Ξ and define $\mathcal{F} \otimes \mathcal{G} = \sigma(\mathcal{R}_{\mathcal{F}\mathcal{G}})$, where

$$\mathcal{R}_{\mathcal{F}\mathcal{G}} = \{F \times G : F \in \mathcal{F}, G \in \mathcal{G}\}. \tag{3.45}$$

The space $(\Omega \times \Xi, \mathcal{F} \otimes \mathcal{G})$ is called a *product space* and $(\Omega, \mathcal{F})$ and $(\Xi, \mathcal{G})$ are the *factor spaces,* or *coordinate spaces,* of the product. The elements of the collection $\mathcal{R}_{\mathcal{F}\mathcal{G}}$ are called the *measurable rectangles.* The rectangles of the Euclidean plane $\mathbb{R} \times \mathbb{R} = \mathbb{R}^2$ (products of intervals) are a familiar case.

3.25 Example Consider the two-element sets $A = \{\omega_1, \omega_2\} \in \mathcal{F}$ and $B = \{\xi_1, \xi_2\} \in \mathcal{G}$. Then $A \times B = \{(\omega_1, \xi_1), (\omega_1, \xi_2), (\omega_2, \xi_1), (\omega_2, \xi_2)\}$ is a rectangle. Other rectangles include the singletons $\{(\omega_i, \xi_j)\}$ and also $\{(\omega_i, \xi_1), (\omega_i, \xi_2)\}$ and $\{(\omega_1, \xi_j), (\omega_2, \xi_j)\}$ for $i, j = 1$ and 2. The sets $\{(\omega_1, \xi_1), (\omega_2, \xi_2)\}$ and $\{(\omega_1, \xi_2), (\omega_2, \xi_1)\}$ are not rectangles. However, they are unions of rectangles and so are elements of $\mathcal{F} \otimes \mathcal{G}$. □

Here are two important pieces of terminology. If $E \subseteq \Omega \times \Xi$, the set $\pi_\Omega(E) = \{\omega : (\omega, \xi) \in E\}$ is called the *projection* of E onto Ω. Also, if $A \subseteq \Omega$ the *inverse projection* of A is the set

$$\pi_\Omega^{-1}(A) = A \times \Xi = \{(\omega, \xi) : \omega \in A, \xi \in \Xi\}. \tag{3.46}$$

$A \times \Xi$ is also called a *cylinder set* in $\Omega \times \Xi$, with *base* A. The latter terminology is natural if you think about the case $\Omega = \mathbb{R}^2$ and $\Xi = \mathbb{R}$. Cylinder sets with bases in $\mathcal{F}$ and $\mathcal{G}$ are elements of $\mathcal{R}_{\mathcal{F}\mathcal{G}}$. It might appear that if $E \in \mathcal{F} \otimes \mathcal{G}$ then $\pi_\Omega(E)$ should be an $\mathcal{F}$-set, but this is *not* necessarily the case. $\pi_\Omega(E)^c \neq \pi_\Omega(E^c)$ in general (see **1.3**) so the collection $\mathcal{C}$ of projections of $\mathcal{F} \otimes \mathcal{G}$-sets onto Ω is not closed under complementation. However, notice that $A = \pi_\Omega(A \times \Xi)$ so that $\mathcal{F} \subseteq \mathcal{C}$.

The main task of this section is to establish a pair of results required in the construction of measures on product spaces.

3.26 Theorem If $\mathcal{C}$ and $\mathcal{D}$ are semi-rings of subsets of Ω and Ξ respectively then

$$\mathcal{R}_{\mathcal{C}\mathcal{D}} = \{C \times D : C \in \mathcal{C}, D \in \mathcal{D}\}$$

is a semi-ring of $\Omega \times \Xi$.

Proof There are three conditions from **1.16** to be established. First, $\mathcal{R}_{\mathcal{C}\mathcal{D}}$ clearly contains $\emptyset$. Second, consider $C_1, C_2 \in \mathcal{C}$ and $D_1, D_2 \in \mathcal{D}$. $C_1 \cap C_2 \in \mathcal{C}$ and $D_1 \cap D_2 \in \mathcal{D}$ and as a matter of definition,

$$\begin{aligned}(C_1 \times D_1) \cap (C_2 \times D_2) &= \{\omega \in C_1, \xi \in D_1\} \cap \{\omega \in C_2, \xi \in D_2\} \\ &= \{\omega \in C_1 \cap C_2,\ \xi \in D_1 \cap D_2\} \\ &= (C_1 \cap C_2) \times (D_1 \cap D_2) \\ &\in \mathcal{R}_{\mathcal{C}\mathcal{D}}.\end{aligned} \tag{3.47}$$

Third, assume that $C_1 \times D_1 \subseteq C_2 \times D_2$ and by a similar argument

$$\begin{aligned}(C_2 \times D_2) - (C_1 \times D_1) &= \{(\omega \in C_2, \xi \in D_2) : \text{either } \omega \notin C_1 \text{ or } \xi \notin D_1\} \\ &= \big((C_2 - C_1) \times D_1\big) \cup \big(C_1 \times (D_2 - D_1)\big) \\ &\quad \cup \big((C_2 - C_1) \times (D_2 - D_1)\big)\end{aligned} \tag{3.48}$$

where the sets in the union on the right-hand side are disjoint. By hypothesis, the sets $C_2 - C_1$ and $D_2 - D_1$ are finite disjoint unions of $\mathcal{C}$-sets and $\mathcal{D}$-sets respectively, say $(C'_1, \ldots, C'_n)$ and $(D'_1, \ldots, D'_m)$. The product of a finite disjoint union of sets is a disjoint union of products; for example,

$$\left(\bigcup_{j=1}^n C'_j\right) \times D_1 = \left\{(\omega, \xi) : \omega \in \left(\bigcup_{j=1}^n C'_j\right), \xi \in D_1\right\} = \bigcup_{j=1}^n (C'_j \times D_1). \tag{3.49}$$

Extending the same type of argument, write

$$(C_2 - C_1) \times (D_2 - D_1) = \left(\bigcup_{j=1}^{n}(C'_j \times D_1)\right) \cup \left(\bigcup_{k=1}^{m}(C_1 \times D'_k)\right) \cup \left(\bigcup_{j,k}(C'_j \times D'_k)\right). \tag{3.50}$$

All of the product sets in this union are disjoint (i.e., a pair (ω, ξ) can appear in at most one of them) and all are in $\mathcal{R}_{\mathcal{CD}}$. This completes the proof. ∎

The second theorem leads to the useful result that to extend a measure on a product space it suffices to assign measures to the elements of $\mathcal{R}_{\mathcal{CD}}$, where $\mathcal{C}$ and $\mathcal{D}$ are suitable classes of the factor spaces.

3.27 Theorem If $\mathcal{F} = \sigma(\mathcal{C})$ and $\mathcal{G} = \sigma(\mathcal{D})$ where $\mathcal{C}$ and $\mathcal{D}$ are semi-rings of subsets of Ω and Ξ respectively, then $\mathcal{F} \otimes \mathcal{G} = \mathcal{C} \otimes \mathcal{D}$.

Proof It is clear that $\mathcal{R}_{\mathcal{CD}} \subseteq \mathcal{R}_{\mathcal{FG}}$ and hence that $\mathcal{C} \otimes \mathcal{D} \subseteq \mathcal{F} \otimes \mathcal{G}$. To show the converse, consider the collection of inverse projections,

$$\mathcal{S}_{\mathcal{F}} = \left\{\pi_{\Omega}^{-1}(F), F \in \mathcal{F}\right\} \subseteq \mathcal{R}_{\mathcal{FG}}.$$

It can easily be verified that $\mathcal{S}_{\mathcal{F}}$ is a σ-field of $\Omega \times \Xi$ and is in fact the smallest σ-field containing the collection $\mathcal{S}_{\mathcal{C}} = \left\{\pi_{\Omega}^{-1}(C), C \in \mathcal{C}\right\} \subseteq \mathcal{C} \otimes \mathcal{D}$. $\mathcal{S}_{\mathcal{C}}$ is a π-system and since $\mathcal{C} \otimes \mathcal{D}$ is a σ-field and hence a λ-system, it follows by the π-λ theorem (**1.29**) that $\mathcal{S}_{\mathcal{F}} = \sigma(\mathcal{S}_{\mathcal{C}}) \subseteq \mathcal{C} \otimes \mathcal{D}$. Exactly same conclusion holds for $\mathcal{S}_{\mathcal{G}}$, the corresponding collection for $\mathcal{G}$. Every element of $\mathcal{R}_{\mathcal{FG}}$ is the intersection of an element from $\mathcal{S}_{\mathcal{F}}$ and one from $\mathcal{S}_{G}$ and it follows that $\mathcal{R}_{\mathcal{FG}} \subseteq \mathcal{C} \otimes \mathcal{D}$. But $\mathcal{R}_{\mathcal{FG}}$ is a π-system by **3.26** and hence a further application of **1.29** gives $\mathcal{F} \otimes \mathcal{G} \subseteq \mathcal{C} \otimes \mathcal{D}$. ∎

The notion of a product extends beyond pairs to triples and general n-tuples. A separate theory is not needed, at least for finite n, because results can be obtained by recursion. Neither of the last two theorems precludes the factors being themselves product spaces. If $(\Psi, \mathcal{H})$ is a third measurable space then trivially $\Omega \times \Xi \times \Psi = (\Omega \times \Xi) \times \Psi$ and so on, up to any finite order.

3.5 Measurable Transformations

Consider measurable spaces $(\Omega, \mathcal{F})$ and $(\Xi, \mathcal{G})$ in a different context, as domain and codomain of a mapping

$$T : \Omega \mapsto \Xi.$$

T is said to be $\mathcal{F}/\mathcal{G}$-measurable if $T^{-1}(B) \in \mathcal{F}$ for all $B \in \mathcal{G}$. The idea is that a measure μ defined on $(\Omega, \mathcal{F})$ can be mapped into $(\Xi, \mathcal{G})$, every set $B \in \mathcal{G}$ being assigned a measure $\nu(B) = \mu(T^{-1}(B))$. One example just encountered is the projection mapping, whose inverse defined in (3.46) takes each $\mathcal{F}$-set A into a measurable rectangle.

Corresponding to a measurable transformation there is always a transformed measure, in the following sense.

3.28 Theorem Let μ be a measure on $(\Omega, \mathcal{F})$ and $T : \Omega \mapsto \Xi$ a measurable transformation. Then μT^{-1} is a measure on $(\Xi, \mathcal{G})$ where

$$\mu T^{-1}(B) = \mu\big(T^{-1}(B)\big), \text{ each } B \in \mathcal{G}. \tag{3.51}$$

Proof Check conditions **3.1**(a)–(c). Clearly, $\mu T^{-1}(A) \geq 0$ for all $A \in \mathcal{B}_T$. Since $T^{-1}(\Xi) = \Omega$ holds by definition, $T^{-1}(\varnothing) = \varnothing$ by **1.2**(iii) and so $\mu T^{-1}(\varnothing) = \mu\big(T^{-1}(\varnothing)\big) = \mu(\varnothing) = 0$. For countable additivity it must be shown that

$$\mu T^{-1}\Big(\bigcup_j B_j\Big) = \sum_j \mu T^{-1}(B_j) \tag{3.52}$$

for a disjoint collection $B_1, B_2, \ldots \in \Xi$. Letting $B_j' = T^{-1}(B_j)$, **1.2**(iii) shows that the B_j' are disjoint and **1.2**(ii) that $T^{-1}\big(\bigcup_j B_j\big) = \bigcup_j B_j'$. Equation (3.52) therefore becomes

$$\mu\Big(\bigcup_j B_j'\Big) = \sum_j \mu(B_j') \tag{3.53}$$

for disjoint sets B_j', which holds because μ is a measure. ∎

The main result on general transformations is the following.

3.29 Theorem Suppose $T^{-1}(B) \in \mathcal{F}$ for each $B \in \mathcal{D}$, where $\mathcal{D}$ is an arbitrary class of sets and $\mathcal{G} = \sigma(\mathcal{D})$. Then the transformation T is $\mathcal{F}/\mathcal{G}$ -measurable.

Proof By **1.2**(ii) and (iii), if $T^{-1}(B_j) \in \mathcal{F}$ for $j \in \mathbb{N}$, then $T^{-1}(\bigcup_j B_j) = \bigcup_j T^{-1}(B_j) \in \mathcal{F}$ and if $T^{-1}(B) \in \mathcal{F}$ then $T^{-1}(B^c) = T^{-1}(B)^c \in \mathcal{F}$. It follows that the class of sets

$$\mathcal{A} = \big\{B : T^{-1}(B) \in \mathcal{F}\big\}$$

is a σ-field. Since $\mathcal{D} \subseteq \mathcal{A}$, $\mathcal{G} \subseteq \mathcal{A}$, being the smallest σ-field containing $\mathcal{D}$ by definition. ∎

This result is easily iterated. If $(\Psi, \mathcal{H})$ is another measurable space and $U : \Xi \mapsto \Psi$ is a $\mathcal{G}/\mathcal{H}$-measurable transformation, then

$$U \circ T : \Omega \mapsto \Psi$$

is $\mathcal{F}/\mathcal{H}$-measurable, since for $C \in \mathcal{H}$, $U^{-1}(C) \in \mathcal{G}$ and hence

$$(U \circ T)^{-1}(C) = T^{-1}\big(U^{-1}(C)\big) \in \mathcal{F}. \tag{3.54}$$

An important special case is the following. $T : \Omega \leftrightarrow \Xi$ is called a *measurable isomorphism* if it is 1–1 onto and both T and T^{-1} are measurable. The measurable spaces $(\Omega, \mathcal{F})$ and $(\Xi, \mathcal{G})$ are said to be *isomorphic* if such a mapping between them exists. The implication is that measure-theoretic discussions can be conducted equivalently in either $(\Omega, \mathcal{F})$ or $(\Xi, \mathcal{G})$. This might appear related to the homeomorphic property of real-valued functions and a homeomorphism is indeed measurably isomorphic. But there is no implication the other way.

3.30 Example Consider $g : [0,1] \mapsto [0,1]$, defined by

$$g(x) = \begin{cases} x + \frac{1}{2}, & 0 \le x \le \frac{1}{2} \\ x - \frac{1}{2}, & \frac{1}{2} < x \le 1. \end{cases} \tag{3.55}$$

Note that g is discontinuous, but is 1–1 onto and of bounded variation. Hence, it is $\mathcal{B}_{[0,1]}/\mathcal{B}_{[0,1]}$-measurable by **3.39** below and $g^{-1} = g$. □

The class of measurable transformations most often encountered is where the codomain is $(\mathbb{R}, \mathcal{B})$, $\mathcal{B}$ being the linear Borel field. In this case the usual terminology is ‘function’ and the notation f is preferred to T. A function may also have the extended real line $(\bar{\mathbb{R}}, \bar{\mathcal{B}})$ as codomain. The measurability criteria are as follows.

3.31 Theorem

(i) A function $f : \Omega \mapsto \mathbb{R}$ for which $\{\omega : f(\omega) \le x\} \in \mathcal{F}$ for each $x \in \mathbb{Q}$ is $\mathcal{F}/\mathcal{B}$-measurable. So is a function for which $\{\omega : f(\omega) < x\} \in \mathcal{F}$ for each $x \in \mathbb{Q}$.

(ii) A function $f : \Omega \mapsto \bar{\mathbb{R}}$ for which $\{\omega : f(\omega) \le x\} \in \mathcal{F}$ for each $x \in \mathbb{Q} \cup \{+\infty\} \cup \{-\infty\}$ is $\mathcal{F}/\bar{\mathcal{B}}$-measurable.

Proof For case (i), the sets $\{\omega : f(\omega) \le x\}$ are of the form $f^{-1}(B), B \in \mathcal{C}$ where $\mathcal{C}$ is defined in **1.21**. Since $\mathcal{B} = \sigma(\mathcal{C})$, the theorem follows by **3.29**. The other collection indicated also generates $\mathcal{B}$ and the same argument applies. The extension to case (ii) is equally straightforward. ∎

Note that the requirements $x \in \mathbb{Q}$ are the minimum necessary, but in most cases the conditions will be satisfied for $x \in \mathbb{R}$, which is of course sufficient. The basic properties of measurable functions follow directly.

3.32 Theorem

(i) If f is measurable, so are $c + f$ and cf where c is any constant.

(ii) If f and g are measurable, so is $f + g$.

Proof If $f \le x$, then $f + c \le x + c \in \mathbb{R}$ so that $f + c$ is measurable by **3.31**. Also, for $x \in \mathbb{R}$,

$$\{\omega : cf(\omega) \le x\} = \begin{cases} \{\omega : f(\omega) \le x/c\}, & c > 0 \\ \{\omega : f(\omega) < x/c\}^c, & c < 0 \\ \Omega, & c = 0 \text{ and } x \ge 0 \\ \varnothing, & c = 0 \text{ and } x < 0 \end{cases} \tag{3.56}$$

where for each of the cases on the right-hand side and each $x/c \in \mathbb{R}$ the sets are in $\mathcal{F}$, proving part (i).

If and only if $f + g < x$, there exist $r \in \mathbb{Q}$ such that $f < r < x - g$ (see **1.10**). It follows that

$$\{\omega : f(\omega) + g(\omega) < x\} = \bigcup_{r \in \mathbb{Q}} \{\omega : f(\omega) < r\} \cap \{\omega : g(\omega) < x - r\}. \tag{3.57}$$

The countable union of $\mathcal{F}$-sets on the right-hand side is an $\mathcal{F}$-set and since this holds for every x, part (ii) also follows by **3.31**(i), where in this case it is convenient to generate $\mathcal{B}$ from the open half-lines. ∎

Combining parts (i) and (ii) shows that if $f_1, \ldots, f_n$ are measurable functions so is $\sum_{j=1}^{n} c_j f_j$, where the c_j are constant coefficients.

The measurability of suprema, infima, and limits of sequences of measurable functions is important in many applications, especially the derivation of integrals in Chapter 4. These are the main cases involving the extended line, because of the possibility that sequences in $\mathbb{R}$ are diverging. Such limits lying in $\bar{\mathbb{R}}$ are called extended functions.

3.33 Theorem Let $\{f_n\}$ be a sequence of $\mathcal{F}/\mathcal{B}$-measurable functions. Then $\inf_n f_n$, $\sup_n f_n$, $\liminf_n f_n$, and $\limsup_n f_n$ are $\mathcal{F}/\bar{\mathcal{B}}$-measurable.

Proof For any $x \in \bar{\mathbb{R}}$, $\{\omega : f_n(\omega) \leq x\} \in \mathcal{F}$ for each n by assumption. Hence

$$\left\{\omega : \sup_n f_n(\omega) \leq x\right\} = \bigcap_{n=1}^{\infty} \{\omega : f_n(\omega) \leq x\} \in \mathcal{F} \tag{3.58}$$

so that $\sup_n f_n$ is measurable by **3.31**(ii). Since $\inf_n f_n = -\sup_n(-f_n)$,

$$\begin{aligned} \left\{\omega : \inf_n f_n(\omega) < x\right\} &= \left\{\omega : \sup_n(-f_n(\omega)) > -x\right\} \\ &= \left\{\omega : \sup_n(-f_n(w)) \leq -x\right\}^c \\ &= \left(\bigcap_{n=1}^{\infty} \{\omega : -f_n(\omega) \leq -x\}\right)^c \\ &= \bigcup_{n=1}^{\infty} \{\omega : f_n(\omega) < x\} \in \mathcal{F}. \end{aligned} \tag{3.59}$$

To extend this result from strong to weak inequalities, write

$$\left\{\omega : \inf_n f_n(\omega) \leq x\right\} = \bigcap_{m=1}^{\infty} \left\{\omega : \inf_n f_n(\omega) < x + 1/m\right\} \in \mathcal{F}. \tag{3.60}$$

Similarly to (3.58),

$$\left\{\omega : \sup_{k \geq n} f_k(\omega) \leq x\right\} = \bigcap_{k \geq n} \{\omega : f_n(\omega) \leq x\} \in \mathcal{F} \tag{3.61}$$

and applying (3.60) to the sequence of functions $g_n = \sup_{k \geq n} f_k$ yields

$$\left\{\omega : \limsup_n f_n(\omega) x\right\} \in \mathcal{F}. \tag{3.62}$$

In much the same way,

$$\left\{\omega : \liminf_n f_n(\omega) \leq x\right\} \in \mathcal{F}. \tag{3.63}$$

The measurability condition of **3.31** is therefore satisfied in each case. ∎

$\lim_n f_n(\omega)$ exists and is measurable whenever $\limsup_n f_n(\omega) = \liminf_n f_n(\omega)$. This equality may hold only on a subset of Ω, but f_n converges a.e. when the complement of this set has measure zero.

The *indicator function* $1_E(\omega)$ of a set $E \in \mathcal{F}$ takes the value $1_E(\omega) = 1$ when $\omega \in E$ and $1_E(\omega) = 0$ otherwise. Some authors call 1_E the *characteristic function* of E, a usage avoided here since this term has a different application in Chapter 11. It is also sometimes written as I_E or as χ_E, and when the set E is defined by an expression too elaborate for a subscript, the form $1(E)$ may also be used.

Here are some useful facts about indicator functions.

3.34 Theorem

(i) $1_E(\omega)$ is $\mathcal{F}/\mathcal{B}$ measurable iff $E \in \mathcal{F}$.

(ii) $1_{E^c}(\omega) = 1 - 1_E(\omega)$.

(iii) $1_{\cup_i E_i}(\omega) = \sup_i 1_{E_i}(\omega)$.

(iv) $1_{\cap_i E_i}(\omega) = \inf_i 1_{E_i}(\omega) = \prod_i 1_{E_i}(\omega)$.

Proof To show (i) note that, for each $B \in \mathcal{B}$

$$1_E^{-1}(B) = \begin{cases} \Omega \text{ if } 0 \in B \text{ and } 1 \in B \\ E \text{ if } 1 \in B \text{ and } 0 \notin B \\ E^c \text{ if } 0 \in B \text{ and } 1 \notin B \\ \varnothing, \text{ otherwise.} \end{cases} \tag{3.64}$$

These sets are in $\mathcal{F}$ if and only if $E \in \mathcal{F}$. The other parts of the theorem are immediate from the definition. ∎

Indicator functions are the building blocks for more elaborate functions, constructed so as to ensure measurability. A *simple function* is a $\mathcal{F}/\mathcal{B}$-measurable function $f : \Omega \mapsto \mathbb{R}$ having finite range; that is, it has the form

$$f(\omega) = \sum_{i=1}^{n} \alpha_i 1_{E_i}(\omega) = \alpha_i,\ \omega \in E_i \tag{3.65}$$

where the $\alpha_1, \ldots, \alpha_n$ are constants and the collection of $\mathcal{F}$-sets $E_1, \ldots, E_n$ is a finite partition of Ω. $\mathcal{F}/\mathcal{B}$-measurability holds because, for any $B \in \mathcal{B}$,

$$f^{-1}(B) = \bigcup_{\alpha_i \in B} E_i \in \mathcal{F}. \tag{3.66}$$

Simple functions are ubiquitous devices in measure and probability theory, because many problems can be solved for such functions rather easily and then

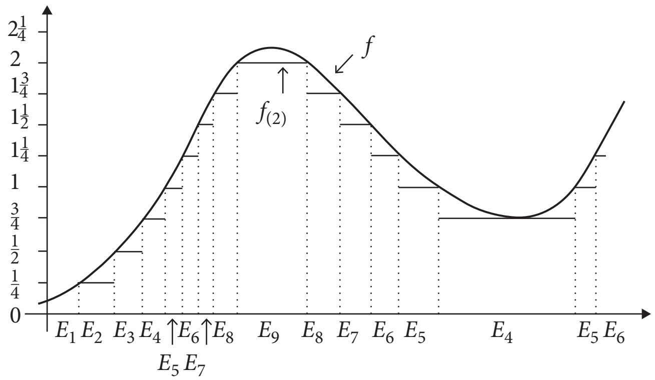

Figure 3.1

generalized to arbitrary functions by a limiting approximation argument such as the following.

3.35 Theorem If f is $\mathcal{F}/\mathcal{B}$-measurable and non-negative, there exists a monotone sequence of $\mathcal{F}/\mathcal{B}$-measurable simple functions $\{f_{(n)}, n \in \mathbb{N}\}$ such that $f_{(n)}(\omega) \uparrow f(\omega)$ for every $\omega \in \Omega$.

Proof For $i = 1, \dots, n2^n$ consider the sets $E_i = \{\omega : (i-1)/2^n \leq f(\omega) < i/2^n\}$. Augment these with the set $E_{n2^n+1} = \{\omega : f(\omega) \geq n\}$. This collection corresponds to a $n2^n + 1$-fold partition of $[0, \infty)$ into $\mathcal{B}$-sets and since f is a function, each ω maps into one and only one $f(\omega)$ and hence belongs to one and only one E_i. The collection $\{E_i\}$ therefore constitutes a partition of Ω. Since f is measurable, $E_i \in \mathcal{F}$ for each i. Define a simple function $f_{(n)}$ on the E_i by letting $\alpha_i = (i-1)/2^n$, for $i = 1, \dots, n2^n + 1$. Then $f_{(n)} \leq f$, but $f_{(n+1)}(\omega) \geq f_{(n)}(\omega)$ for every ω; incrementing n bisects each interval and if $f_{(n)}(\omega) = (i-1)/2^n$, $f_{(n+1)}(\omega)$ is equal to either $2(i-1)/2^{n+1} = f_{(n)}(\omega)$, or $(2i-1)/2^{n+1} > f_n(\omega)$. It follows that the sequence is monotone and $\lim_{n\to\infty} f_{(n)}(\omega) = f(\omega)$. This holds for each $\omega \in \Omega$. ∎

Figure 3.1 illustrates the construction for $n = 2$ and the case $\Omega = \mathbb{R}$, so that $f(\omega)$ is a function on the real line.

To extend from non-negative to general functions take the positive and negative parts. Define $f^+ = \max\{f, 0\}$ and $f^- = f^+ - f$. Both f^+ and f^- are non-negative functions. If $f^+_{(n)}$ and $f^-_{(n)}$ are the non-negative simple approximations to f^+ and f^- defined in **3.35** and $f_{(n)} = f^+_{(n)} - f^-_{(n)}$,

$$|f - f_{(n)}| \leq |f^+ - f^+_{(n)}| + |f^- - f^-_{(n)}| \to 0. \tag{3.67}$$

3.6 Borel Functions

If f is a measurable function and

$$g : \mathbb{S} \mapsto \mathbb{T};\ \mathbb{S} \subseteq \mathbb{R}, \mathbb{T} \subseteq \mathbb{R}$$

is a function of a real variable, is the composite function $g \circ f$ measurable? The answer to this question is yes if and only if g is a *Borel function*. Let $\mathcal{B}_{\mathbb{S}} = \{B \cap \mathbb{S} : B \in \mathcal{B}\}$, where $\mathcal{B}$ is the Borel field of $\mathbb{R}$. $\mathcal{B}_{\mathbb{S}}$ is a σ-field of subsets of $\mathbb{S}$ and $B \cap \mathbb{S}$ is open (closed) in the relative topology on $\mathbb{S}$ whenever B is open (closed) in $\mathbb{R}$ (see **1.30** and **1.32**). $\mathcal{B}_{\mathbb{S}}$ is called the Borel field on $\mathbb{S}$. Define $\mathcal{B}_{\mathbb{T}}$ similarly with respect to $\mathbb{T}$. Then g is called a Borel function (i.e., is Borel-measurable) if $g^{-1}(B) \in \mathcal{B}_{\mathbb{S}}$ for all sets $B \in \mathcal{B}_{\mathbb{T}}$.

3.36 Example Consider $g(x) = |x|$. g^{-1} takes each point $x \in \mathbb{R}^+$ into the points x and $-x$. For any $B \in \mathcal{B}^+$ (the restriction of $\mathcal{B}$ to $\mathbb{R}^+$) the image under g^{-1} is the set containing the points x and $-x$ for each $x \in B$ which is an element of $\mathcal{B}$. □

3.37 Example Let $g(x) = 1$ if x is rational, 0 otherwise. Note that $\mathbb{Q} \in \mathcal{B}$ (see **3.22**) and g^{-1} is defined according to (3.64) with $E = \mathbb{Q}$, so g is Borel-measurable. □

In fact, to construct a 'plausible' non-measurable function is quite difficult. The obvious case is the following.

3.38 Example Take a set $A \notin \mathcal{B}$; for example, let A be the set H defined in **3.24**. Now construct the indicator function $1_A(x) : \mathbb{R} \mapsto \{0, 1\}$. Since $1_A^{-1}(\{1\}) = A \notin \mathcal{B}$, this function is not measurable. □

Necessary conditions for Borel measurability are hard to pin down, but the following sufficient conditions are convenient.

3.39 Theorem If $g : \mathbb{S} \mapsto \mathbb{T}$ is either (i) continuous or (ii) of bounded variation, it is Borel-measurable.

Proof (i) follows immediately from **3.29** and the definition of a Borel field, since continuity implies that $g^{-1}(B)$ is open (closed) in $\mathbb{S}$ whenever B is open (closed) in $\mathbb{T}$, by **2.7**.

To prove (ii), consider first a non-decreasing function $h : \mathbb{R} \mapsto \mathbb{R}$ having the property $h(y) \leq h(x)$ when $y < x$. If $A = \{y : h(y) \leq h(x)\}$, sup $A = x$ and A is one of $(-\infty, x)$ and $(-\infty, x]$, so the condition of **3.31** is satisfied. So suppose g is non-decreasing on $\mathbb{S}$. Applying the last result to any non-decreasing h with the property $h(x) = g(x)$ for $x \in \mathbb{S}$ shows that g is Borel-measurable, because $g^{-1}(B \cap \mathbb{T}) = h^{-1}(B) \cap \mathbb{S} \in \mathcal{B}_{\mathbb{S}}$ for each $B \cap \mathbb{T} \in \mathcal{B}_{\mathbb{T}}$. Since a function of bounded variation is the difference of two non-decreasing functions by **2.10** the theorem now follows easily by **3.32**. ∎

The next result adds a further case to those of **3.32**.

3.40 Theorem If f and g are continuous and hence measurable, fg is measurable.

Proof $fg = \frac{1}{2}\left((f+g)^2 - f^2 - g^2\right)$ and the result follows on combining **3.39**(i) with **3.32**(ii). ∎

The concept of a Borel function extends naturally to Euclidean n-spaces, and indeed to mappings between spaces of different dimension. A vector function

$$\boldsymbol{g} : \mathbb{S} \to \mathbb{T};\ \mathbb{S} \subseteq \mathbb{R}^k,\ \mathbb{T} \subseteq \mathbb{R}^m$$

is Borel-measurable if $\boldsymbol{g}^{-1}(B) \in \mathcal{B}_{\mathbb{S}}$ for all $B \in \mathcal{B}_{\mathbb{T}}$, where $\mathcal{B}_{\mathbb{S}} = \{B \cap \mathbb{S} : B \in \mathcal{B}^k\}$ and $\mathcal{B}_{\mathbb{T}} = \{B \cap \mathbb{T} : B \in \mathcal{B}^m\}$.

3.41 Theorem If $\boldsymbol{g}$ is continuous, it is Borel-measurable.

Proof This is similar to **3.39**(i) where **2.11** provides the required property of continuous mappings. ∎

Finally, note the application of **3.28** to these cases.

3.42 Theorem If μ is a measure on $(\mathbb{R}^k, \mathcal{B}^k)$ and $\mathrm{g} : \mathbb{S} \mapsto \mathbb{T}$ is Borel-measurable where $\mathbb{S} \subseteq \mathbb{R}^k$ and $\mathbb{T} \subseteq \mathbb{R}^m$, μ_{g}^{-1} is a measure on $(\mathbb{T}, \mathcal{B}_{\mathbb{T}})$ where

$$\mu_{\mathrm{g}}^{-1}(B) = \mu\left(\mathrm{g}^{-1}(B)\right) \tag{3.68}$$

for each $B \in \mathcal{B}_{\mathbb{T}}$. □

3.43 Example The projection of $\mathbb{R}^k$ onto $\mathbb{R}^m$ for $m < k$. If $\boldsymbol{X}$ is $k \times 1$ with partition

$$\boldsymbol{X} = \begin{bmatrix} \boldsymbol{X}_* \\ \boldsymbol{X}_{**} \end{bmatrix}$$

where $\boldsymbol{X}_*$ is $m \times 1$ and $\boldsymbol{X}_{**}$ is $(k-m) \times 1$, let $\boldsymbol{g} : \mathbb{R}^k \mapsto \mathbb{R}^m$ be defined by

$$\boldsymbol{g}(\boldsymbol{X}) = \boldsymbol{X}_*. \tag{3.69}$$

In this case, $\mu\boldsymbol{g}^{-1}(B) = \mu\big(\boldsymbol{g}^{-1}(B)\big) = \mu(B \times \mathbb{R}^{k-m})$ for $B \in \mathbb{R}^m$. □

4
Integration

4.1 Construction of the Integral

The reader is doubtless familiar with the Riemann integral of a bounded non-negative function f on a bounded interval of the line $[a, b]$, what is usually written $\int_a^b f\mathrm{d}x$. The objects to be studied in this chapter represent a heroic generalization of the same idea. Instead of intervals of the line, the integral is defined on an arbitrary measure space.

Suppose $(\Omega, \mathcal{F}, \mu)$ is a measure space and

$$f : \Omega \mapsto \bar{\mathbb{R}}^+$$

is a $\mathcal{F}/\bar{\mathcal{B}}$-measurable function into the non-negative, extended real line. The integral of f is defined to be the real valued functional

$$\int f\mathrm{d}\mu = \sup\left\{\sum_i \Big(\inf_{\omega\in E_i} f(\omega)\Big)\mu(E_i)\right\} \tag{4.1}$$

where the supremum is taken over all finite partitions of Ω into sets $E_i \in \mathcal{F}$ and the supremum exists. If no supremum exists, the integral is assigned the value $+\infty$.[1] The integral of the function $1_A f$, where $1_A(\omega)$ is the indicator of the set $A \in \mathcal{F}$, is called the integral of f over A and written $\int_A f\mathrm{d}\mu$.

The expression in (4.1) is sometimes called the *lower integral* and denoted $\int_* f\mathrm{d}\mu$. Likewise, define the *upper integral* of f,

$$\int^* f\mathrm{d}\mu = \inf\left\{\sum_i \Big(\sup_{\omega\in E_i} f(\omega)\Big)\mu(E_i)\right\}. \tag{4.2}$$

These two constructions, approximating f from below and from above, should agree. Indeed, it is possible to show that $\int_* f\mathrm{d}\mu = \int^* f\mathrm{d}\mu$ whenever f is bounded and $\mu(\Omega) < \infty$. However, $\int^* f\mathrm{d}\mu = \infty$ if either the set $\{\omega : f(\omega) > 0\}$ has infinite measure or f is unbounded on sets of positive measure. Definition (4.1) is preferred because it can yield a finite value in these cases.

[1] The notations $\int f\mathrm{d}\mu$, $\int f\mu(\mathrm{d}\omega)$, or simply $\int f$ when the relevant measure is understood are used synonymously by different authors.

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0004

To extend the definition from non-negative functions to general functions, take positive and negative parts. If $f : \Omega \to \bar{\mathbb{R}}$ is any measurable function let $f^+ = \max\{f, 0\} \geq 0$ and $f^- = f^+ - f \geq 0$. The integral of f is defined as

$$\int f \mathrm{d}\mu = \int f^+ \mathrm{d}\mu - \int f^- \mathrm{d}\mu \tag{4.3}$$

so long as at least one of the right-hand-side integrals is finite. If both $\int f^+ \mathrm{d}\mu = \infty$ and $\int f^- \mathrm{d}\mu = \infty$ the integral is undefined; the difference of two infinities is undefined and in particular it is not zero. A function is said to be *integrable* only if its integral is both defined and finite. Noting that $|f| = f^+ + f^-$, f is integrable if and only if

$$\int |f| \mathrm{d}\mu < \infty. \tag{4.4}$$

In the discussion that follows it will generally be assumed that f is non-negative with the extension via equation (4.3), if required, taken as implicit.

4.1 Example A familiar case is the measure space $(\mathbb{R}, \mathcal{B}, m)$ where m is Lebesgue measure. The integral $\int f \mathrm{d}m$ where f is a Borel function is the *Lebesgue integral* of f. This is customarily written $\int f \mathrm{d}x$, reflecting the fact that $m\big((x, x + \mathrm{d}x]\big) = \mathrm{d}x$, even though the sets $\{E_i\}$ in (4.1) need not be intervals. A natural way to implement the formula in (4.1) for a non-negative function is to consider the monotone sequence of simple functions defined in Theorem **3.35**, where $\mu = m$, the partition $\{E_i\}$ corresponds to the level sets defined there, and by construction $\inf_{\omega \in E_i} f(\omega) = f_{(n)}(\omega)$. Letting n increase defines a monotone sequence converging to $\int f \mathrm{d}m$. The monotone convergence idea is formalized in Theorem **4.7** below. □

4.2 Example Consider a measure space $(\mathbb{R}, \mathcal{B}, \mu)$ where μ differs from m. The integral $\int f \mathrm{d}\mu$ where f is a Borel function is the *Lebesgue–Stieltjes integral*. The monotone function

$$F(x) = \mu((-\infty, x]) \tag{4.5}$$

has the property $\mu((a, b]) = F(b) - F(a)$ and the measure of the interval $(x, x + \mathrm{d}x]$ can be written $\mathrm{d}F(x)$. The notation $\int f \mathrm{d}F$ means exactly the same as $\int f \mathrm{d}\mu$, the choice between the μ and F representations being a matter of taste. See §8.2 and §9.1 for details. Similarly to Example **4.1**, the limit of the monotone sequence $\sum_i f_{(n)}(\omega)\mu(E_i)$ with $f_{(n)}$ defined in **3.35** is a natural way to define $\int f \mathrm{d}\mu$. □

4.3 Example Consider a real sequence, $\{f_i, i \in \mathbb{N}\}$ with $f_i \geq 0$. The integral of the sequence with respect to counting measure (see **3.4**) is calculated from formula (4.1) with the sets E_i corresponding to the supremum equated with the singletons of $\mathbb{N}$, the sets $\{i\}$ each having counting measure 1. The integral is merely the sum $\sum_{i=1}^{\infty} f_i$. □

The last example is particularly useful because it shows formally that sums of suitable type are integrals and share the properties of the latter.

For a contrast with these cases, consider the *Riemann–Stieltjes integral.* For an interval $[a, b]$, let a partition into subintervals be defined by a set of points $\Pi = \{x_1, \dots, x_n\}$, with $a = x_0 < x_1 < \dots < x_n = b$. Another set Π' is called a refinement of Π if $\Pi \subseteq \Pi'$. Given functions f and $\alpha : \mathbb{R} \mapsto \mathbb{R}$, let

$$S(\Pi, \alpha, f) = \sum_{i=1}^{n} f(t_i)(\alpha(x_i) - \alpha(x_{i-1})) \tag{4.6}$$

where $t_i \in [x_{i-1}, x_i]$. If there exists a number $\int_a^b f \mathrm{d}\alpha$, such that for every $\varepsilon > 0$ there is a partition Π_ε with

$$\left| S(\Pi, \alpha, f) - \int_a^b f \mathrm{d}\alpha \right| < \varepsilon$$

for all $\Pi \supseteq \Pi_\varepsilon$ and every choice of $\{t_i\}$, this is called the Riemann–Stieltjes integral of f with respect to α. When $\alpha = x$ and f is bounded this definition yields the ordinary Riemann integral, and when it exists this always agrees with the Lebesgue integral of f over $[a, b]$. Moreover, if α is an increasing function of the form in (4.5), this integral is equal to the Lebesgue–Stieltjes integral whenever it is defined. There do however exist bounded, measurable functions that are not Riemann-integrable so that even for bounded intervals the Lebesgue integral is the more inclusive concept.

4.4 Example Consider the function defined in **3.37**. Since $m(E) = 0$ for $E = \{\omega : f(\omega) > 0\}$, the Lebesgue integral unambiguously takes the value 0. The Riemann integral is undefined, since every interval $[x_{i-1}, x_i]$ contains points t_i for which $f(t_i) = 0$ and others for which $f(t_i) = 1$. □

On the other hand, the Riemann–Stieltjes integral is defined for more general classes of integrator function. In particular if f is continuous, the integral exists for α of bounded variation on $[a, b]$, not necessarily monotone as in **4.2**. These integrals therefore fall outside the class defined by (4.1) although note that when α is of bounded variation, having a representation as the difference of two nondecreasing functions by **2.10**, the Riemann–Stieltjes integral is the difference between a pair of Lebesgue–Stieltjes integrals on $[a, b]$.

The best way to understand the general integral is not to study a particular measure space such as the line, but to restrict attention initially to particular classes of function. The simplest possible case is the indicator of a set. Then, every partition $\{E_i\}$ containing A yields the same value for the sum in (4.1), which is

$$\int_A \mathrm{d}\mu = \int 1_A \mathrm{d}\mu = \mu(A) \tag{4.7}$$

for any $A \in \mathcal{F}$. If $A \notin \mathcal{F}$, the integral is undefined.

Another case of much importance is the following.

4.5 Theorem If $f = 0$ a.e.$[\mu]$, then $\int f\mathrm{d}\mu = 0$.

Proof The theorem says there exists $C \in \mathcal{F}$ with $\mu(C) = 1$, such that $f(\omega) = 0$ for $\omega \in C$. For any partition $\{E_1, \ldots, E_n\}$ let $E_i' = E_i \cap C$ and $E_i'' = E_i - E_i'$. By additivity of μ,

$$\begin{aligned} \sum_i \Big(\inf_{\omega\in E_i} f(\omega)\Big)\mu(E_i) &= \sum_i \Big(\inf_{\omega\in E_i'} f(\omega)\Big)\mu(E_i') + \sum_i \Big(\inf_{\omega\in E_i''} f(\omega)\Big)\mu(E_i'') \\ &= 0 \end{aligned} \tag{4.8}$$

where the first sum of terms disappears because $f(\omega) = 0$ and the second disappears by **3.12**(i) since $\mu(E_i'') \le \mu(C^c) = 0$ for each i. ∎

A class of functions for which evaluation of the integral is simple, as their name suggests, is the non-negative simple functions.

4.6 Theorem Let $\varphi(\omega) = \sum_{i=1}^n \alpha_i 1_{E_i}(\omega)$, where $\alpha_i \ge 0$ for $i = 1, \ldots, n$ and $E_1, \ldots, E_n \in \mathcal{F}$ is a partition of Ω. Then

$$\int \varphi \mathrm{d}\mu = \sum_{i=1}^n \alpha_i \mu(E_i). \tag{4.9}$$

Proof For an arbitrary finite partition $A_1, \ldots, A_m$ of Ω, define $\beta_j = \inf_{\omega\in A_j} \varphi(\omega)$. Note that $\beta_j = \alpha_{i(j)}$ where $i(j)$ denotes the value of i at which α_i is minimized, over those cases where $A_j \cap E_i$ is nonempty. Using additivity of μ,

$$\begin{aligned} \sum_{j=1}^m \beta_j \mu(A_j) &= \sum_{i=1}^n \sum_{j=1}^m \alpha_{i(j)} \mu(A_j \cap E_i) \\ &\le \sum_{i=1}^n \alpha_i \sum_{j=1}^m \mu(A_j \cap E_i) \\ &= \sum_{i=1}^n \alpha_i \mu(E_i) \end{aligned} \tag{4.10}$$

where the inequality uses the fact that $\alpha_{i(j)} \leq \alpha_i$ in every case j where $\mu(A_j \cap E_i) > 0$. The theorem follows according to (4.1) since (4.10) holds as an equality for the case $m = n$ and $A_i = E_i$, $i = 1, \ldots, n$. ∎

Thus, for functions with finite range the integral is the sum of the possible values of f, weighted by the measures of the sets on which those values hold. Look at Figure 3.1 (page 76). The Lebesgue integral of the approximating function $f_{(2)}$ in the figure is the sum of the areas of the rectangular regions. In the Lebesgue–Stieltjes integral with respect to some measure μ the widths of the sets E_i are replaced in the calculation by their measures $\mu(E_i)$.

The next challenge is to find a way to construct the integrals of arbitrary non-negative functions, and hence of general functions via (4.3). The *monotone convergence theorem* is the cornerstone of integration theory, both providing the main step in the construction of the general integral and also spawning a range of useful corollaries.

4.7 Theorem (monotone convergence) If $\{f_n\}$ is a non-decreasing sequence of measurable non-negative functions, with $f_n(\omega) \uparrow f(\omega)$ for each $\omega \in \Omega$,

$$\lim_{n\to\infty} \int f_n \mathrm{d}\mu = \int f \mathrm{d}\mu \tag{4.11}$$

where by implication the two sides of (4.11) are either both infinite, or finite and equal.

Proof If g is a function such that $0 \leq f(\omega) \leq g(\omega)$ for each $\omega \in \Omega$, it is immediate from (4.1) that $0 \leq \int f \mathrm{d}\mu \leq \int g \mathrm{d}\mu$. Hence, $\int f_n \mathrm{d}\mu$ is a non-decreasing sequence bounded above by $\int f \mathrm{d}\mu$ and having a limit, $\lim_{n\to\infty} \int f_n \mathrm{d}\mu \leq \int f \mathrm{d}\mu$. To complete the proof, it suffices to show that $\lim_{n\to\infty} \int f_n \mathrm{d}\mu \geq \int f \mathrm{d}\mu$.

For an arbitrary partition $\{A_i \in \mathcal{F}\}$ of Ω let $\beta_i = \inf_{\omega \in A_i} f(\omega)$. For $k > 1$ define

$$b_i = \begin{cases} k, & \beta_i = \infty \\ (1 - 1/k)\beta_i, & 0 < \beta_i < \infty \\ 0, & \beta_i = 0 \end{cases} \tag{4.12}$$

so that either $b_i = \beta_i = 0$ or $b_i < \beta_i$. Choose a constant c so that $\sum_i \beta_i \mu(A_i) > c$ and then choose k large enough that $\sum_i b_i \mu(A_i) > c$ also. By choice of $\{A_i\}$, $\int f \mathrm{d}\mu - \sum_i \beta_i \mu(A_i) \geq 0$ can be made arbitrarily small and hence c can be chosen large enough that $\int f \mathrm{d}\mu - c > 0$ is arbitrarily small. The proof is completed by showing that $\lim_{n\to\infty} \int f_n \mathrm{d}\mu > c$. To do this, partition A_i into $A_{ni} = A_i \cap \{\omega : f_n(\omega) \geq b_i\} \in \mathcal{F}$ and $A_i - A_{ni}$, so that

$$\sum_i b_i\mu(A_{ni}) \leq \sum_i \Big(\inf_{\omega\in A_{ni}} f_n(\omega)\Big)\mu(A_{ni}) + \sum_i \Big(\inf_{\omega\in A_i - A_{ni}} f_n(\omega)\Big)\mu(A_i - A_{ni})$$
$$\leq \int f_n \mathrm{d}\mu \tag{4.13}$$

where the first inequality is by construction and the second by (4.1). For any $\omega \in A_i$, since $b_i < f(\omega)$ unless $f(\omega) = 0$ and $f_n \uparrow f$, there exists n large enough that $b_i \leq f_n(\omega)$. Hence, $A_{ni} \uparrow A_i$ and $\mu(A_{ni}) \to \mu(A_i)$ as $n \to \infty$ (see the remark following **3.7**). Noting how c is defined, this means that with k large enough the minorant side of (4.13) strictly exceeds c in the limit. This completes the proof. ∎

The leading application of this result may be apparent. For arbitrary non-negative f, a monotone sequence of simple functions converges to f from below, by **3.35**. The integral of f is the corresponding limit of the integrals of the simple functions defined in (4.9), whose existence is assured by **4.7**.

If **3.35** shows that a non-negative measurable function is the limit of a simple sequence, **3.33** shows that every convergent simple sequence has a measurable function as its limit. The next theorem teams these results with the monotone convergence theorem and provides an alternative definition of the integral.

4.8 Theorem For any non-negative $\mathcal{F}/\mathcal{B}$-measurable function f,

$$\int f\mathrm{d}\mu = \sup_{0\leq\varphi\leq f} \int \varphi\mathrm{d}\mu \tag{4.14}$$

where φ denotes the class of simple functions. □

For bounded f and finite μ, the equality also holds in respect of the infimum over simple functions $\varphi \geq f$, in parallel with (4.2).

4.2 Properties of the Integral

A really useful feature of **4.8** is that it allows the proof of easy results for the integrals of simple functions, which are then extended to the general case by the limiting argument. The most important of the properties established in this way is linearity.

4.9 Theorem If f and g are $\mathcal{F}/\mathcal{B}$-measurable, integrable functions and a and b are constants, then $(af + bg)$ is integrable and

$$\int (af + bg)\mathrm{d}\mu = a\int f\mathrm{d}\mu + b\int g\mathrm{d}\mu. \tag{4.15}$$

Proof First let f, g and a, b all be non-negative. If $\{A_i\}$ and $\{B_j\}$ are finite partitions of Ω and $\varphi(\omega) = \sum_i \alpha_i 1_{Ai}(\omega)$ and $\gamma(\omega) = \sum_j \beta_j 1_{B_j}(\omega)$ are simple functions defined on these partitions, then

$$\begin{aligned} a\varphi(\omega) + b\gamma(\omega) &= \sum_i \sum_j (a\alpha_i + b\beta_j) 1_{A_i \cap B_j}(\omega) \\ &= b\alpha_i + a\beta_j,\ \omega \in A_i \cap B_j \end{aligned} \tag{4.16}$$

which is a simple function. Hence,

$$\begin{aligned} \int (a\varphi + b\gamma)\mathrm{d}\mu &= \sum_i \sum_j (a\alpha_i + b\beta_j)\mu(A_i \cap B_j) \\ &= a\sum_i \alpha_i \sum_j \mu(A_i \cap B_j) + b\sum_j \beta_j \sum_i \mu(A_i \cap B_j) \\ &= a\sum_i \alpha_i \mu(A_i) + b\sum_j \beta_j \mu(B_j) \\ &= a\int \varphi \mathrm{d}\mu + b\int \gamma \mathrm{d}\mu \end{aligned} \tag{4.17}$$

showing that linearity applies to simple functions. Now applying **4.8**,

$$\begin{aligned} \int (af + bg)\mathrm{d}\mu &= \sup_{\varphi \le f, \gamma \le g} \int (a\varphi + b\gamma)\mathrm{d}\mu \\ &= a\left(\sup_{\varphi \le f} \int \varphi \mathrm{d}\mu\right) + b\left(\sup_{\gamma \le g} \int \gamma \mathrm{d}\mu\right) \\ &= a\int f\mathrm{d}\mu + b\int g\mathrm{d}\mu. \end{aligned} \tag{4.18}$$

To extend the result to general functions note that since $|af + bg| \le |a| . |f| + |b| . |g|$, $af + bg$ is integrable so long as f and g are integrable and a and b are finite. Separating positive and negative parts the identity

$$af + bg = (af)^+ - (af)^- + (bg)^+ - (bg)^-$$

implies as in (4.3),

$$\int (af + bg)\mathrm{d}\mu = \int (af)^+ \mathrm{d}\mu - \int (af)^- \mathrm{d}\mu + \int (bg)^+ \mathrm{d}\mu - \int (bg)^- \mathrm{d}\mu. \tag{4.19}$$

If $a \geq 0$ then

$$\int (af)^+ \mathrm{d}\mu - \int (af)^- \mathrm{d}\mu = a\left(\int f^+ \mathrm{d}\mu - \int f^- \mathrm{d}\mu\right) = a \int f \mathrm{d}\mu$$

whereas if $a < 0$,

$$\int (af)^+ \mathrm{d}\mu - \int (af)^- \mathrm{d}\mu = |a|\left(\int f^- \mathrm{d}\mu - \int f^+ \mathrm{d}\mu\right) = |a|\left(-\int f \mathrm{d}\mu\right) = a \int f \mathrm{d}\mu.$$

The same argument applies to the terms in b and g so (4.15) holds as required. ∎

Linearity is a fundamental property that underpins most of the useful results in integration theory. One immediate implication is the following.

4.10 Theorem (modulus inequality) $\left|\int f \mathrm{d}\mu\right| \leq \int |f| \mathrm{d}\mu.$

Proof

$$\begin{aligned}\left|\int f \mathrm{d}\mu\right| = \left|\int f^+ \mathrm{d}\mu - \int f^- \mathrm{d}\mu\right| &\leq \int f^+ \mathrm{d}\mu + \int f^- \mathrm{d}\mu \\ &= \int (f^+ + f^-) \mathrm{d}\mu = \int |f| \mathrm{d}\mu. \quad ∎\end{aligned}$$

The next thing to show is the invariance of the integral to the behaviour of functions on sets of measure 0, extending the basic result of **4.5**.

4.11 Lemma Let f and g be integrable functions.

(i) If $f \leq g$ a.e.$[\mu]$ then $\int f \mathrm{d}\mu \leq \int g \mathrm{d}\mu$.

(ii) If $f = g$ a.e.$[\mu]$ then $\int f \mathrm{d}\mu = \int g \mathrm{d}\mu$.

Proof For (i), consider first the case $f = 0$. If $g \geq 0$ everywhere, $\int g \mathrm{d}\mu \geq 0$ directly from (4.1). So suppose $g \geq 0$ a.e.$[\mu]$ and define

$$h(\omega) = \begin{cases} 0, & g(\omega) \geq 0 \\ \infty, & g(\omega) < 0. \end{cases}$$

Then $h = 0$ a.e.$[\mu]$ but $g + h \geq 0$ everywhere. Applying **4.9**,

$$0 \leq \int (g + h) \mathrm{d}\mu = \int g \mathrm{d}\mu + \int h \mathrm{d}\mu = \int g \mathrm{d}\mu \tag{4.20}$$

since $\int h\mathrm{d}\mu = 0$ by **4.5**. Now replace g by $g-f$ in the last argument to show $\int(g-f)\mathrm{d}\mu \geq 0$ and hence $\int g\mathrm{d}\mu \geq \int f\mathrm{d}\mu$ by **4.9**.

To prove (ii), let $h = f - g$ so that $h = 0$ a.e.$[\mu]$ and $\int h\mathrm{d}\mu = 0$ by **4.5**. Then $\int f\mathrm{d}\mu = \int(g+h)\mathrm{d}\mu = \int g\mathrm{d}\mu + \int h\mathrm{d}\mu = \int g\mathrm{d}\mu$, where the second equality is by **4.9**. ∎

These last results permit the extension to the more commonly quoted version of the monotone convergence theorem, which is as follows.

4.12 Corollary If $f_n \geq 0$ and $f_n \uparrow f$ a.e.$[\mu]$, $\lim_{n\to\infty} \displaystyle\int f_n \mathrm{d}\mu = \int f\mathrm{d}\mu$. □

Theorem **4.9** extends by recursion from pairs to arbitrary finite sums of functions, and in particular $\int(\sum_{i=1}^n g_i)\mathrm{d}\mu = \sum_{i=1}^n \int g_i \mathrm{d}\mu$. Put $f_n = \sum_{i=1}^n g_i$ and $\int f_n \mathrm{d}\mu = \sum_{i=1}^n \int g_i \mathrm{d}\mu$, where the g_i are non-negative functions. Then, if $f_n \uparrow f = \sum_{i=1}^\infty g_i < \infty$ a.e., **4.12** also implies the following.

4.13 Corollary If $\{g_i\}$ is a sequence of non-negative functions,

$$\int \left(\sum_{i=1}^{\infty} g_i\right)\mathrm{d}\mu = \sum_{i=1}^{\infty} \int g_i \mathrm{d}\mu. \quad \square \tag{4.21}$$

By implication, the two sides of this equation are either both infinite, or finite and equal. This has a particular application to results involving σ-finite measures. To evaluate an integral $\int g\mathrm{d}\mu$ using a method that works for finite measures, extend to the σ-finite case by constructing a countable partition $\{\Omega_i\}$ of Ω, such that $\mu(\Omega_i) < \infty$ for each i. Letting $g_i = 1_{\Omega_i} g$, note that $g = \sum_i g_i$ and $\int g\mathrm{d}\mu = \sum_i \int g_i \mathrm{d}\mu$ by (4.21).

In the form of **4.12** the monotone convergence theorem has several other useful corollaries, including the so-called Fatou's lemma. If $\{f_n\}$ is a collection of non-negative functions, $\inf_n f_n$ is the function that assumes the value $\inf_n f_n(\omega)$ at each point of the domain $\omega \in \Omega$. Similarly, $\operatorname{liminf}_{n\to\infty} f_n$ is the function defined by $\sup_n \inf_{k\geq n} f_k(\omega)$, $\omega \in \Omega$.

4.14 Lemma (Fatou) If $f_n \geq 0$ a.e.$[\mu]$, then

$$\int \left(\operatorname*{liminf}_{n\to\infty} f_n\right)\mathrm{d}\mu \leq \operatorname*{liminf}_{n\to\infty} \int f_n \mathrm{d}\mu.$$

Proof Let $g_n = \inf_{k\geq n} f_k$, so that $\{g_n\}$ is a non-decreasing sequence and $g_n \uparrow g = \operatorname{liminf}_n f_n$. Since $f_n \geq g_n$, $\int f_n \mathrm{d}\mu \geq \int g_n \mathrm{d}\mu$. Letting $n \to \infty$ on both sides of the inequality, noting that $\int g_n \mathrm{d}\mu$ is monotone by **4.11**(i), gives

$$\int \left(\liminf_{n\to\infty} f_n\right) d\mu = \int g d\mu = \lim_{n\to\infty} \int g_n d\mu \le \liminf_{n\to\infty} \int f_n d\mu. \quad \blacksquare \tag{4.22}$$

4.15 Corollary (reverse Fatou) Suppose there exists $g \geq 0$ a.e.[μ] such that $\int g d\mu < \infty$ and $f_n \leq g$, all n. Then

$$\int \left(\limsup_{n\to\infty} f_n\right) d\mu \ge \limsup_{n\to\infty} \int f_n d\mu.$$

Proof Apply **4.14** to the sequence $g - f_n$ using the first equality of (2.3). ■

4.16 Theorem (dominated convergence) If $f_n \to f$ a.e.[μ] and there exists a function g such that $|f_n| \leq g$ a.e.[μ] for all n and $\int g d\mu < \infty$, then $\int f_n d\mu \to \int f d\mu$.

Proof According to **4.11**(i), $\int g d\mu < \infty$ implies $\int |f_n| d\mu < \infty$. Let $h_n = |f_n - f|$, such that $0 \leq h_n \leq 2g$ a.e.[μ] and $h_n \to 0$ a.e.[μ]. Applying **4.5** to $\liminf_n h_n$, linearity and Fatou's lemma give

$$\begin{aligned} 2\int g d\mu = \int \liminf_{n\to\infty}(2g - h_n) d\mu &\le \liminf_{n\to\infty} \int (2g - h_n) d\mu \\ &= 2\int g d\mu - \limsup_{n\to\infty} \int h_n d\mu \end{aligned} \tag{4.23}$$

where the last equality uses (2.3). Clearly, $\limsup_{n\to\infty} \int h_n d\mu = 0$. Since $\int h_n d\mu \geq 0$, the modulus inequality implies

$$\left|\lim_{n\to\infty} \int f_n d\mu - \int f d\mu\right| \le \lim_{n\to\infty} \int h_n d\mu = 0. \quad \blacksquare \tag{4.24}$$

Taking the case where the g is replaced by a finite constant produces the *bounded convergence theorem*, often more convenient:

4.17 Theorem (bounded convergence) If $f_n \to f$ a.e.[μ] and $|f_n| \leq B < \infty$ for all n, then $\lim_{n\to\infty} \int f_n d\mu \to \int f d\mu < \infty$. □

Other widely exploited implications of linearity are the extension from sums to integrals of **2.22** and **2.24**. The proofs are essentially the same.

4.18 Theorem (Hölder inequality) For integrable functions f and g and constant $p \geq 1$,

$$\int |fg|\mathrm{d}\mu \leq \left(\int |f|^p \mathrm{d}\mu\right)^{1/p} \left(\int |g|^q \mathrm{d}\mu\right)^{1/q} \text{ for } q = \frac{p}{p-1}.$$

Proof Define $\varphi = |f|/(\int |f|^p \mathrm{d}\mu)^{1/p}$ and $\gamma = |g|/(\int |g|^q \mathrm{d}\mu)^{1/q}$. It follows by linearity and Lemma **2.23** that

$$\int |\varphi\gamma|\mathrm{d}\mu \leq \frac{1}{p} + \frac{1}{q} = 1. \quad ∎$$

4.19 Theorem (Minkowski inequality) For integrable functions f and g and constant $p \geq 1$,

$$\left(\int |f+g|^p \mathrm{d}\mu\right)^{1/p} \leq \left(\int |f|^p \mathrm{d}\mu\right)^{1/p} + \left(\int |g|^p \mathrm{d}\mu\right)^{1/p}. \tag{4.25}$$

Proof Use the triangle inequality to write

$$\int |f+g|^p \mathrm{d}\mu \leq \int |f||f+g|^{p-1}\mathrm{d}\mu + \int |g||f+g|^{p-1}\mathrm{d}\mu.$$

Apply **4.18** to each term and rearrange to obtain (4.25). ∎

4.3 Product Measure and Multiple Integrals

Let $(\Omega, \mathcal{F}, \mu)$ and $(\Xi, \mathcal{G}, \nu)$ be measure spaces. In general, $(\Omega \times \Xi, \mathcal{F} \otimes \mathcal{G}, \pi)$ might also be a measure space, with π a measure on the sets of $\mathcal{F} \otimes \mathcal{G}$. In this case measures μ and ν, defined by $\mu(F) = \pi(F \times \Xi)$ and $\nu(G) = \pi(\Omega \times G)$ respectively, are called the *marginal measures* corresponding to π.

Alternatively, suppose that μ and ν are given and define the set function

$$\pi : \mathcal{R}_{\mathcal{F}\mathcal{G}} \mapsto \bar{\mathbb{R}}^+$$

where $\mathcal{R}_{\mathcal{F}\mathcal{G}}$ denotes the measurable rectangles of the space $\Omega \times \Xi$, by

$$\pi(F \times G) = \mu(F)\nu(G). \tag{4.26}$$

π is a measure on $\mathcal{R}_{\mathcal{F}\mathcal{G}}$ called the *product measure* and has an extension to $\mathcal{F} \otimes \mathcal{G}$, so that $(\Omega \times \Xi, \mathcal{F} \otimes \mathcal{G}, \pi)$ is indeed a measure space. The first step in this demonstration is to define the mapping

$$T_\omega : \Xi \mapsto \Omega \times \Xi$$

by $T_\omega(\xi) = (\omega, \xi)$ so that, for $G \in \mathcal{G}$, $T_\omega(G) = \{\omega\} \times G$. For $E \in \mathcal{F} \otimes \mathcal{G}$, let

$$E_\omega = T_\omega^{-1}(E) = \{\xi : (\omega, \xi) \in E\} \subseteq \Xi. \tag{4.27}$$

The set E_ω can be thought of as the cross-section through E at the element ω. For any countable collection of $\mathcal{F} \otimes \mathcal{G}$-sets $\{E_j, j \in \mathbb{N}\}$,

$$\left(\bigcup_j E_j\right)_\omega = \left\{\xi : (\omega, \xi) \in \bigcup_j E_j\right\} = \bigcup_j \left\{\xi : (\omega, \xi) \in E_j\right\} = \bigcup_j (E_j)_\omega. \tag{4.28}$$

For future reference, note the following.

4.20 Lemma T_ω is a $\mathcal{G}/(\mathcal{F} \otimes \mathcal{G})$-measurable mapping for each $\omega \in \Omega$.

Proof If $E = F \times G$ for $F \in \mathcal{F}$ and $G \in \mathcal{G}$ it is obvious that

$$E_\omega = \left\{ \begin{array}{c} G, \omega \in F \\ \varnothing, \omega \notin F \end{array} \right\} \in \mathcal{G}. \tag{4.29}$$

Since $\mathcal{F} \otimes \mathcal{G} = \sigma(\mathcal{R}_{\mathcal{F}\mathcal{G}})$, $E_\omega \in \mathcal{G}$ whenever $E \in \mathcal{F} \otimes \mathcal{G}$ and the lemma follows by **3.29**. ∎

The second step is to show the following.

4.21 Theorem π is a measure on $\mathcal{R}_{\mathcal{F}\mathcal{G}}$.

Proof Clearly π is non-negative and $\pi(\varnothing) = 0$, recalling that $F \times \varnothing = \varnothing \times G = \varnothing$ for any $F \in \mathcal{F}$ or $G \in \mathcal{G}$ and applying (4.26). To show countable additivity, let $\{E_j \in \mathcal{R}_{\mathcal{F}\mathcal{G}}, j \in \mathbb{N}\}$ be a disjoint collection, such that there exist sets $F_j \in \mathcal{F}$ and $G_j \in \mathcal{G}$ with $E_j = F_j \times G_j$; and also suppose $E = \bigcup_j E_j \in \mathcal{R}_{\mathcal{F}\mathcal{G}}$, such that there exist sets F and G with $E = F \times G$. Any point $(\omega, \xi) \in F \times G$ belongs to one and only one of the sets $F_j \times G_j$, so that for any $\omega \in F$, the sets of the subcollection $\{G_j\}$ for which $\omega \in F_j$ must constitute a partition of G. Hence, applying (4.28) and (4.29),

$$\begin{aligned} \nu(E_\omega) &= \nu\left(\left(\bigcup_j E_j\right)_\omega\right) = \nu\left(\bigcup_j (E_j)_\omega\right) \\ &= \nu\left(\bigcup_j \left\{ \begin{array}{c} G_j, \omega \in F_j \\ \varnothing, \omega \notin F_j \end{array} \right\}\right) = \sum_j 1_{F_j}(\omega)\nu(G_j) \end{aligned} \tag{4.30}$$

where the additivity of ν can be applied since the sets G_j appearing in this decomposition are disjoint. Since by (4.29) it is also the case that $\nu(E_\omega) = \nu(G)1_F(\omega)$,

$$\pi(E) = \mu(F)\nu(G) = \int \nu(E_\omega)\mathrm{d}\mu(\omega) = \int\Big(\sum_j 1_{F_j}(\omega)\nu(G_j)\Big)\mathrm{d}\mu(\omega)$$
$$= \sum_j \mu(F_j)\nu(G_j) = \sum_j \pi(E_j) \qquad (4.31)$$

as required, where the penultimate equality is by **4.13**. ∎

It is now straightforward to extend the measure from $\mathcal{R}_{\mathcal{F}\mathcal{G}}$ to $\mathcal{F} \otimes \mathcal{G}$.

4.22 Theorem $(\Omega \times \Xi, \mathcal{F} \otimes \mathcal{G}, \pi)$ is a measure space.

Proof $\mathcal{F}$ and $\mathcal{G}$ are σ-fields and hence semi-rings; hence $\mathcal{R}_{\mathcal{F}\mathcal{G}}$ is a semi-ring by **3.26**. The theorem follows from **4.21** and **3.15**. ∎

Iterating the preceding arguments (i.e. letting $(\Omega, \mathcal{F})$ and/or $(\Xi, \mathcal{G})$ be product spaces) allows the concept to be extended to products of higher order. In later chapters, product probability measures will embody the intuitive notion of statistical independence, although this is by no means the only application to be met. The following case has a familiar geometrical interpretation.

4.23 Example In the plane $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$, Lebesgue measure is defined for products of intervals by

$$m\big((a_1, b_1] \times (a_2, b_2]\big) = (b_1 - a_1)(b_2 - a_2). \qquad (4.32)$$

Here the measurable rectangles include the actual geometrical rectangles and $\mathcal{B}^2$, the Borel sets of the plane, is generated from these as a consequence of **3.27**. By the foregoing reasoning, $(\mathbb{R}^2, \mathcal{B}^2, m)$ is a measure space in which the measure of a set is given by its area. □

The next step must be to construct integrals of functions $f(\omega, \xi)$ on the product space. The following lemma is a natural extension of **4.20**. What might be thought of as a cross-section through the two-dimensional mapping at a point $\omega \in \Omega$ yields a function with domain Ξ.

4.24 Lemma Let $f : \Omega \times \Xi \mapsto \mathbb{R}$ be $\mathcal{F} \otimes \mathcal{G}/\mathcal{B}$-measurable. Define $f_\omega(\xi) = f(\omega, \xi)$ for fixed $\omega \in \Omega$. Then $f_\omega : \Xi \mapsto \mathbb{R}$ is $\mathcal{G}/\mathcal{B}$-measurable.

Proof Write

$$f_\omega(\xi) = f(\omega,\xi) = f\big(T_\omega(\xi)\big) = f \circ T_\omega(\xi). \tag{4.33}$$

By **4.20** and the remarks following **3.29**, the composite function $f \circ T_\omega$ is $\mathcal{G}/\mathcal{B}$-measurable. ∎

Suppose that f_ω can be integrated with respect to ν over Ξ. There are two questions of interest that arise here. First, is the resulting function, $g(\omega) = \int_\Xi f_\omega \mathrm{d}\nu$, $\mathcal{F}/\bar{\mathcal{B}}$-measurable? And second, if g is now integrated with respect to μ over Ω, what is the relationship between this integral and the integral $\int_{\Omega\times\Xi} f\mathrm{d}\pi$ over $\Omega \times \Xi$? The affirmative answer to the first of these questions and the fact that the 'iterated' integral is identical with the 'double' integral where these exist, are the most important results for product spaces, known jointly as the *Fubini theorem.* Since iterated integration is an operation taken for granted with multiple Riemann integrals, perhaps the main point needing to be stressed here is that this convenient property of product measures (and multivariate Lebesgue measure in particular) does *not* generalize to arbitrary measures on product spaces.

Initially let f be the indicator of a set $E \in \mathcal{F} \otimes \mathcal{G}$, so that f_ω is the indicator of the set E_ω defined in (4.27). Define the the function $g_E : \Omega \mapsto \bar{\mathbb{R}}^+$ by

$$g_E(\omega) = \int f_\omega \mathrm{d}\nu = \nu(E_\omega). \tag{4.34}$$

In view of **4.20**, $E_\omega \in \mathcal{G}$ and g_E is well-defined, although unless ν is a finite measure it may take its values in the extended half line, as shown.

4.25 Lemma Let μ and ν be σ-finite. For all $E \in \mathcal{F} \otimes \mathcal{G}$, g_E is $\mathcal{F}/\bar{\mathcal{B}}$-measurable and

$$\int_\Omega g_E \mathrm{d}\mu = \pi(E). \quad \Box \tag{4.35}$$

By implication, the two sides of the equality in (4.35) are either both infinite, or finite and equal.

Proof of 4.25 Assume first that the measures are finite. The theorem is proved for this case using the π-λ theorem. Let $\mathcal{A}$ denote the collection of sets E such that g_E satisfies (4.35). $\mathcal{R}_{\mathcal{F}\mathcal{G}} \subseteq \mathcal{A}$, since if $E = F \times G$ then by (4.29),

$$g_E(\omega) = \nu(G)1_F(\omega), \; F \in \mathcal{F} \tag{4.36}$$

and $\int_\Omega g_E d\mu = \mu(F)\nu(G) = \pi(E)$ as required. $\mathcal{A}$ is a λ-system. To see this note first that $\Omega \times \Xi \in \mathcal{A}$, so **1.26**(a) holds. If $E_1, E_2 \in \mathcal{A}$ and $E_1 \subset E_2$, then since $1_{E_2 - E_1} = 1_{E_2} - 1_{E_1}$,

$$\begin{aligned} g_{E_2 - E_1}(\omega) &= \int_\Xi 1_{E_2}(\omega, \xi)\mathrm{d}\nu(\xi) - \int_\Xi 1_{E_1}(\omega, \xi)\mathrm{d}\nu(\xi) \\ &= g_{E_2}(\omega) - g_{E_1}(\omega). \end{aligned} \tag{4.37}$$

This is an $\mathcal{F}/\bar{\mathcal{B}}$-measurable function by **3.32** and so, by additivity of π,

$$\int_\Omega g_{E_2 - E_1}\mathrm{d}\mu(\omega) = \pi(E_2) - \pi(E_1) = \pi(E_2 - E_1) \tag{4.38}$$

showing that $\mathcal{A}$ satisfies **1.26**(b). Finally, if A_1 and A_2 are disjoint so are $(A_1)_\omega$ and $(A_2)_\omega$ and $g_{A_1 \cup A_2}(\omega) = g_{A_1}(\omega) + g_{A_2}(\omega)$. To establish **1.26**(c), let $\{E_j \in \mathcal{A}, j \in \mathbb{N}\}$ be a monotone sequence, with $E_j \uparrow E$. Define the disjoint collection $\{A_j\}$ with $A_1 = E_1$ and $A_j = E_j - E_{j-1}, j > 1$, so that $E = \bigcup_{j=1}^\infty A_j$ and $A_j \in \mathcal{A}$ by (4.37). By countable additivity of ν,

$$g_E(\omega) = \sum_{j=1}^\infty g_{A_j}(\omega). \tag{4.39}$$

This is $\mathcal{F}/\mathcal{B}$-measurable by **3.33** and

$$\int_\Omega \left(\sum_{j=1}^\infty g_{A_j}(\omega) \right) \mathrm{d}\mu(\omega) = \sum_{j=1}^\infty \int_\Omega g_{A_j}(\omega)\mathrm{d}\mu(\omega) = \sum_{j=1}^\infty \pi(A_j) = \pi(E) \tag{4.40}$$

where the first equality is by **4.13**. This shows that $\mathcal{A}$ is a λ-system. Since $\mathcal{R}_{\mathcal{F}\mathcal{G}}$ is a semi-ring it is also a π-system and $\mathcal{F} \otimes \mathcal{G} = \sigma(\mathcal{R}_{\mathcal{F}\mathcal{G}}) \subseteq \mathcal{A}$ by **1.29**. This completes the proof for finite measures.

To extend to the σ-finite case, let $\{\Omega_i\}$ and $\{\Xi_j\}$ be countable partitions of Ω and Ξ with finite μ-measure and ν-measure respectively; then the collection $\{\Omega_i \times \Xi_j \in \mathcal{R}_{\mathcal{F}\mathcal{G}}\}$ forms a countable partition of $\Omega \times \Xi$ having finite measures, $\pi(\Omega_i \times \Xi_j) = \mu(\Omega_i)\nu(\Xi_j)$. For a set $E \in \mathcal{F} \otimes \mathcal{G}$, write $E_{ij} = E \cap (\Omega_i \times \Xi_j)$. Then, by the last argument,

$$\int_{\Omega_i} g_{E_{ij}}\mathrm{d}\mu = \pi(E_{ij}) \tag{4.41}$$

where $g_{E_{ij}} : \Omega_i \mapsto \mathbb{R}^+$ is defined by $g_{E_{ij}}(\omega) = \nu\big((E_{ij})_\omega\big)$, $\omega \in \Omega_i$. The sets E_{ij} are disjoint and $g_E(\omega) = \nu\big((\bigcup_j E_{ij})_\omega\big)$ when $\omega \in \Omega_i$, or

$$g_E(\omega) = \sum_i 1_{\Omega_j}(\omega) \sum_j g_{E_{ij}}(\omega). \tag{4.42}$$

The sum on the right need not converge and in that case $g_E(\omega) = +\infty$. However, $\mathcal{F}/\bar{\mathcal{B}}$-measurability holds by **3.32/3.33** and

$$\begin{aligned}\int_\Omega g_E \mathrm{d}\mu &= \int_\Omega \left(\sum_i 1_{\Omega_j} \sum_j g_{E_{ij}} \right) \mathrm{d}\mu \\ &= \sum_i \sum_j \int_{\Omega_i} g_{E_{ij}} \mathrm{d}\mu = \sum_i \sum_j \pi(E_{ij}) = \pi(E) \end{aligned} \tag{4.43}$$

using **4.13** and countable additivity. This completes the proof. ∎

The next step is to extend from indicator functions to non-negative functions. The following is Tonelli's theorem.

4.26 Theorem (Tonelli) Let π be a product measure with σ-finite marginal measures μ and ν and let $f : \Omega \times \Xi \mapsto \bar{\mathbb{R}}^+$ be $(\mathcal{F} \otimes \mathcal{G})/\bar{\mathcal{B}}$-measurable. Define functions $f_\omega : \Xi \mapsto \bar{\mathbb{R}}^+$ by $f_\omega(\xi) = f(\omega, \xi)$ and let $g(\omega) = \int_\Xi f_\omega \mathrm{d}\nu$. Then

(i) g is $\mathcal{F}/\bar{\mathcal{B}}$-measurable

(ii) $\int_{\Omega\times\Xi} f \mathrm{d}\pi = \int_\Omega \left(\int_\Xi f_\omega \mathrm{d}\nu \right) \mathrm{d}\mu$. □

In part (ii) it is again understood that the two sides of the equation are either finite and equal, or both infinite. Like the other results of this section, the theorem is symmetric in $(\Omega, \mathcal{F}, \mu)$ and $(\Xi, \mathcal{G}, \nu)$ and the complementary results given by interchanging the roles of the marginal spaces do not require a separate statement. The theorem holds even for measures that are not σ-finite, but this further complicates the proof.

Proof of 4.26 This is on the lines of **4.8**. For a partition $\{E_1, \ldots, E_n\}$ of $\Omega \times \Xi$ let $f = \sum_i \alpha_i 1_{E_i}$ and $g = \sum_i \alpha_i \nu((E_i)_\omega)$ by **4.6**. g is $\mathcal{F}/\bar{\mathcal{B}}$-measurable by **3.32** and **4.25** gives

$$\int_\Omega g \mathrm{d}\mu = \sum_i \alpha_i \int_\Omega \nu((E_i)_\omega) \mathrm{d}\mu = \sum_i \alpha_i \pi(E_i) = \int_{\Omega\times\Xi} f \mathrm{d}\pi \tag{4.44}$$

so that the theorem holds for simple functions. For general non-negative f, choose a monotone sequence of simple functions converging to f as in **3.35**, show measurability of g in the limit using **3.33**, and apply the monotone convergence theorem. ∎

Fubini's theorem extends the last result from non-negative f to general f. This requires the additional assumption of integrability.

4.27 Theorem (Fubini) Let π be a product measure with σ-finite marginal measures μ and ν; let $f : \Omega \times \Xi \mapsto \mathbb{R}$ be $(\mathcal{F} \otimes \mathcal{G})/\mathcal{B}$-measurable with

$$\int_{\Omega\times\Xi} |f(\omega,\xi)| \mathrm{d}\pi(\omega,\xi) < \infty. \tag{4.45}$$

Define $f_\omega : \Xi \mapsto \mathbb{R}$ by $f_\omega(\xi) = f(\omega,\xi)$; and let $g(\omega) = \int_\Xi f_\omega \mathrm{d}\nu$. Then

(i) f_ω is $\mathcal{G}/\mathcal{B}$-measurable and integrable for $\omega \in C \subseteq \Omega$, with $\mu(\Omega - C) = 0$

(ii) g is $\mathcal{F}/\mathcal{B}$-measurable and integrable on C

(iii) $\int_{\Omega\times\Xi} f(\omega,\xi)\mathrm{d}\pi(\omega,\xi) = \int_\Omega \Big(\int_\Xi f(\omega,\xi)\mathrm{d}\nu(\xi)\Big) d\mu(\omega)$.

Proof Apart from the integrability, **4.24** shows (i) and Tonelli's theorem shows (ii) and (iii) for the functions $f^+ = \max\{f,0\}$ and $f^- = f^+ - f$, where $|f| = f^+ + f^-$. But under (4.45), $|f(\omega,\xi)| < \infty$ on a set whose complement has π-measure 0. With C defined as the projection of this set onto Ω, (i), (ii), and (iii) hold for f^+ and f^-, with both sides of the equation finite in (iii). Since $f = f^+ - f^-$, (i) extends to f by **3.32** and (ii) and (iii) extend to f by **4.9**. ∎

4.4 The Radon–Nikodym Theorem

Consider finite or σ-finite measures μ and ν on a measurable space $(\Omega, \mathcal{F})$. μ is said to be *absolutely continuous* with respect to ν if for $E \in \mathcal{F}$, $\nu(E) = 0$ implies $\mu(E) = 0$. This relationship is written as $\mu \ll \nu$. If $\mu \ll \nu$ and $\nu \ll \mu$, the measures are said to be *equivalent*. If there exists a partition (A, A^c) of Ω, such that $\mu(A) = 0$ and $\nu(A^c) = 0$, then μ and ν are said to be *mutually singular*, written $\mu \perp \nu$. Mutual singularity is symmetric, such that $\mu \perp \nu$ means the same as $\nu \perp \mu$.

Now suppose that there exists a function

$$f : \Omega \mapsto \bar{\mathbb{R}}^+$$

such that $\mu(E) = \int_E f \mathrm{d}\nu$ for $E \in \mathcal{F}$. It follows directly (choose E such that $\nu(E) = 0$) that $\mu \ll \nu$. f might be thought of as the derivative of one measure with respect to the other and even written $f = \mathrm{d}\mu/\mathrm{d}\nu$. The result that absolute continuity of μ with respect to ν *implies* the existence of such a function is the Radon–Nikodym theorem. If g is another such function and $\mu(E) = \int_E g \mathrm{d}\nu$ for all $E \in \mathcal{F}$ then $\nu(f \neq g) = 0$; otherwise, at least one of the sets $E_1 = \{\omega : f(\omega) > g(\omega)\}$ and

$E_2 = \{\omega : f(\omega) < g(\omega)\}$ must contradict the definition. In this sense f is unique. It is called the Radon–Nikodym derivative of μ with respect to ν.

Not all measures have this property. However, the *Lebesgue decomposition* of a finite or σ-finite measure μ with respect to another such measure ν specifies the existence of measures μ_1 and μ_2 such that $\mu = \mu_1 + \mu_2$, $\mu_1 \perp \nu$, and $\mu_2 \ll \nu$. Of course either μ_1 or μ_2 may be zero, but the decomposition shows that in the former case μ is absolutely continuous with respect to ν and possesses a Radon–Nikodym derivative, while in the latter case μ and ν are mutually singular.

The existence of the Lebesgue decomposition is proved as Theorem **4.30**. The proof requires the concept of a *signed measure.*

4.28 Definition A signed measure on $(\Omega, \mathcal{F})$ is a set function

$$\chi : \mathcal{F} \mapsto \bar{\mathbb{R}}$$

satisfying

(a) $\chi(\emptyset) = 0$
(b) $\chi\left(\bigcup_j A_j\right) = \sum_j \chi(A_j)$ for any countable, disjoint collection $\{A_j \in \mathcal{F}\}$
(c) either $\chi < \infty$ or $\chi > -\infty$. □

For example, let μ and ν be non-negative measures on a space $(\Omega, \mathcal{F})$ with at least one of them finite. For a non-negative constant r define

$$\chi(A) = \mu(A) - r\nu(A) \tag{4.46}$$

for any $A \in \mathcal{F}$. For disjoint collection $\{A_j\}$,

$$\chi\left(\bigcup_{j=1}^{\infty} A_j\right) = \mu\left(\bigcup_{j=1}^{\infty} A_j\right) - r\nu\left(\bigcup_{j=1}^{\infty} A_j\right) = \sum_{j=1}^{\infty}(\mu(A_j) - r\nu(A_j)) \tag{4.47}$$

so that countable additivity holds.

If A is an $\mathcal{F}$-set with the property that $\chi(B) \geq 0$ for every subset $B \in \mathcal{F}$ with $B \subseteq A$, A is called a *positive set,* a *negative set* being defined in the complementary manner. A set that is both positive and negative is called a *null set.* Be careful to distinguish between positive (negative, null) sets and sets of positive measure (negative measure, measure zero). A set A has measure zero if $\mu(A) = r\nu(A)$ in (4.46), but it is not in general a null set. By the definition, any subset of a positive set is positive.

The following theorem defines the *Hahn decomposition.*

4.29 Theorem Let χ be a signed measure on a measurable space $(\Omega, \mathcal{F})$, having the property $\chi(A) < \infty$ for all $A \in \mathcal{F}$. There exists a partition of Ω into a positive set A^+ and a negative set A^-.

Proof Let $\lambda = \sup \chi(A)$ where the supremum is taken over the positive sets A of χ. Choose a sequence of positive sets $\{A_n\}$ such that $\lim_{n\to\infty} \chi(A_n) = \lambda$ and let $A^+ = \bigcup_n A_n$. To show that A^+ is also a positive set, consider any measurable $E \subseteq A^+$. Letting $B_1 = A_1$ and $B_n = A_n - A_{n-1}$ for $n > 1$, the sequence $\{B_n\}$ is disjoint, positive since $B_n \subseteq A_n$ for each n and $\bigcup_n B_n = A^+$. Likewise, if $E_n = E \cap B_n$ the sequence $\{E_n\}$ is disjoint, positive since $E_n \subseteq B_n$ and $\bigcup_n E_n = E$. Hence $\chi(E) = \sum_n \chi(E_n) \geq 0$ and since E was arbitrary, A^+ is shown to be positive. $A^+ - A_n$ being therefore positive,

$$\chi(A^+) = \chi(A_n) + \chi(A^+ - A_n) \geq \chi(A_n), \text{ all } n, \tag{4.48}$$

and hence $\chi(A^+) \geq \lambda$, implying $\chi(A^+) = \lambda$.

Now let $A^- = \Omega - A^+$ and show by contradiction that A^- has no subset E with positive measure. Suppose there exists $E \subseteq A^-$ with $\chi(E) > 0$. By construction E and A^+ are disjoint. Every subset of $A^+ \cup E$ is the disjoint union of a subset of A^+ with a subset of E. If E were a positive set so would be $A^+ \cup E$, but by definition of λ,

$$\lambda \geq \chi(A^+ \cup E) = \lambda + \chi(E) \tag{4.49}$$

which requires $\chi(E) = 0$, contradicting our assumption. Hence E cannot be a positive set. If F is a subset of E, it is also a subset of A^-. If positive, it must have zero measure by the argument just applied to E. The desired contradiction is obtained by showing that if $\chi(E) > 0$, E must have a subset F which is both positive and has positive measure.

The technique is to successively remove subsets of negative measure from E until what is left has to be a positive set and then to show that this remainder has positive measure. Let n_1 be the smallest integer such that there is a subset $E_1 \subseteq E$ with $\chi(E_1) < -1/n_1$ and define $F_1 = E - E_1$. Then let n_2 be the smallest integer such that there exists $E_2 \subseteq F_1$ with $\chi(E_2) < -1/n_2$. In general, for $k = 2, 3, \ldots$, let n_k be the smallest positive integer such that F_{k-1} has a subset E_k satisfying $\chi(E_k) < -1/n_k$ and let

$$F_k = E - \bigcup_{j=1}^{k} E_j. \tag{4.50}$$

If no such set exists for finite n_k, let $n_k = +\infty$ and $E_k = \emptyset$. The sequence $\{F_k\}$ is non-increasing and so must converge to a limit F as $k \to \infty$.

Therefore, consider writing $E = F \cup \left(\bigcup_{k=1}^{\infty} E_k \right)$ where the sets on the right-hand side are mutually disjoint and hence, by countable additivity,

$$\chi(E) = \chi(F) + \sum_{k=1}^{\infty} \chi(E_k) < \chi(F) - \sum_{k=1}^{\infty} 1/n_k. \tag{4.51}$$

Since $\chi(E) > 0$ it must be the case that $\chi(F) > 0$, but since $\chi(F) < \infty$ by assumption, it is also the case that $\sum_{k=1}^{\infty}(1/n_k) < \infty$ and hence $n_k \to \infty$ as $k \to \infty$. This means that F contains no subset with negative measure and is therefore a positive set having positive measure. ∎

For any set $B \in \mathcal{F}$, define $\chi^+(B) = \chi(A^+ \cap B)$ and $\chi^-(B) = -\chi(A^- \cap B)$, such that $\chi(B) = \chi^+(B) - \chi^-(B)$. It is easy to verify that χ^+ and χ^- are mutually singular, non-negative measures on $(\Omega, \mathcal{F})$. $\chi = \chi^+ - \chi^-$ is called the *Jordan decomposition* of a signed measure. χ^+ and χ^- are called the *upper variation* and *lower variation* of χ and the measure $|\chi| = \chi^+ + \chi^-$ is called the *total variation* of χ. The Jordan decomposition shows that all signed measures can be represented in the form of (4.46). Signed measures therefore introduce no new technical difficulties. Integrate with respect to χ by taking the difference of the integrals with respect to χ^+ and χ^-.

Now the Radon–Nikodym theorem can be proved. This is properly part (i) of the following result and it is easiest to begin with the case of finite measures and then extend the results to the σ-finite case.

4.30 Theorem Finite, non-negative measures μ and ν define a Lebesgue decomposition $\mu = \mu_1 + \mu_2$ where

(i) $\mu_2 \ll \nu$ and there exists an $\mathcal{F}/\bar{\mathcal{B}}$-measurable function $f : \Omega \mapsto \bar{\mathbb{R}}^+$ such that $\mu_2(E) = \int_E f \mathrm{d}\nu$ for all $E \in \mathcal{F}$.

(ii) $\mu_1 \perp \nu$.

Proof Let $\mathbb{G}$ denote the class of $\mathcal{F}/\mathcal{B}$-measurable functions $g : \Omega \mapsto \mathbb{R}^+$ for which $\int_E g \mathrm{d}\nu \le \mu(E)$ for all $E \in \mathcal{F}$. This set is never empty since it always contains 0. Letting $\alpha = \sup_{g \in \mathbb{G}} \int g \mathrm{d}\nu$ so that $\alpha \le \mu(\Omega) < \infty$, the object is to show that $\mathbb{G}$ contains an element attaining this supremum. To do this, first construct a sequence $\{g_n \in \mathbb{G}\}$ such that $\alpha - 1/n \le \int g_n \mathrm{d}\nu \le \alpha$. This is not necessarily monotone in the sense of **4.7**, but a sequence $\{f_n\}$ with this property is generated as follows. Put $f_1 = g_1$ and define f_n by $f_n(\omega) = \max\{f_{n-1}(\omega), g_n(\omega)\}$, so that $f_n(\omega) \ge f_{n-1}(\omega)$ for $\omega \in \Omega$. Defining the sets $A_n = \{\omega : f_{n-1}(\omega) > g_n(\omega)\}$ for $n = 2, 3, \ldots$, note that if $f_{n-1} \in \mathbb{G}$ then

$$\begin{aligned}\int_E f_n \mathrm{d}\nu &= \int_{E\cap A_n} f_{n-1}\mathrm{d}\nu + \int_{E\cap A_n^c} g_n \mathrm{d}\nu \\ &\le \mu(E\cap A_n) + \mu(E\cap A_n^c) = \mu(E) \end{aligned} \tag{4.52}$$

so that $f_n \in \mathbb{G}$. Since $f_n \uparrow f$, it follows by the monotone convergence theorem that $\int_E f_n \mathrm{d}\nu \to \int_E f\mathrm{d}\nu \le \mu(E)$ and hence $f \in \mathbb{G}$. Since $f_n \ge g_n$ so that $\alpha - 1/n \le \int g_n \mathrm{d}\nu \le \int f_n \mathrm{d}\nu \le \alpha$, $\int f\mathrm{d}\nu = \alpha$ as was required to be shown. Now define μ_2 by

$$\mu_2(E) = \int_E f\mathrm{d}\nu,\ E \in \mathcal{F}. \tag{4.53}$$

μ_2 is a non-negative measure. Properties **3.1**(a) and (b) are immediate and note that $\mu_2 \ll \nu$ by construction. To show that it is countably additive, consider for a countable collection $\{E_j \in \mathcal{F}\}$ the functions $f_j = 1_{E_j} f$ and use **4.13**. This completes the proof of (i).

To show (ii), define $\mu_1 = \mu - \mu_2$. This is non-negative by construction of f and also a measure and it remains to be shown that $\mu_1 \perp \nu$. To do this let (A_n^+, A_n^-) be a Hahn decomposition for the measure $\mu_1 - \nu/n$, for $n = 1, 2, 3, \ldots$. Then for $E \in \mathcal{F}$,

$$\begin{aligned}\mu(E\cap A_n^+) &= \mu_1(E\cap A_n^+) + \mu_2(E\cap A_n^+) \\ &\ge \frac{1}{n}\nu(E\cap A_n^+) + \int_{E\cap A_n^+} f\mathrm{d}\nu \\ &= \int_{E\cap A_n^+}\left(f + \frac{1}{n}\right)\mathrm{d}\nu \end{aligned} \tag{4.54}$$

where the inequality holds because $E \cap A_n^+$ is a positive set with respect to $\mu_1 - \nu/n$. Hence, recalling that $\mu_2(E) \le \mu(E)$ for all $E \in \mathcal{F}$ by construction of μ_2,

$$\begin{aligned}\mu(E) &= \mu(E\cap A_n^+) + \mu(E\cap A_n^-) \\ &\ge \mu(E\cap A_n^+) + \mu_2(E\cap A_n^-) \\ &\ge \int_{E\cap A_n^+}\left(f + \frac{1}{n}\right)\mathrm{d}\nu + \int_{E\cap A_n^-} f\mathrm{d}\nu \\ &= \int_E f\mathrm{d}\nu + \frac{1}{n}\nu(E\cap A_n^+). \end{aligned} \tag{4.55}$$

Note from this inequality that $f + n^{-1}1_{A_n^+} \in \mathbb{G}$, so that

$$\alpha \ge \int f\mathrm{d}\nu + \frac{1}{n}\nu(A_n^+) = \alpha + \frac{1}{n}\nu(A_n^+) \tag{4.56}$$

implying $\nu(A_n^+)/n = 0$. This holds for each $n \in \mathbb{N}$ so $\nu(A) = 0$, where $A = \bigcup_{n=1}^{\infty} A_n^+$. By **1.1**(iii), $A^c = \bigcap_{n=1}^{\infty} A_n^-$. This means that $A^c \subseteq A_n^-$ and so A^c is a negative set for $\mu_1 - \nu/n$ and $\mu_1(A^c) \leq \nu(A^c)/n$. Since this is true for every n, it follows that $\mu_1(A^c) = 0$ and so $\mu_1 \perp \nu$. ∎

It remains to extend these results to the σ-finite case.

4.31 Theorem The conclusions of Theorem **4.30** hold when μ and ν are non-negative σ-finite measures on $(\Omega, \mathcal{F})$.

Proof By σ-finiteness there exists a countable partition $\{\Omega_j\}$ of Ω, such that $\nu(\Omega_j)$ and $\mu(\Omega_j)$ are finite for each j. If $\{A_j\}$ is any collection with finite measures whose union is Ω, letting $\Omega_1 = A_1$ and $\Omega_j = A_j - A_{j-1}$ for $j > 1$ defines such a partition. If different collections with finite measures are known for ν and μ, say $\{A_{\mu j}\}$ and $\{A_{\nu j}\}$, the collection containing all the $A_{\mu j} \cap A_{\nu k}$ for $j, k \in \mathbb{N}$ is countable and of finite measure with respect to both ν and μ and after re-indexing this collection can generate $\{\Omega_j\}$.

Consider the restrictions of μ and ν to the measurable spaces $(\Omega_j, \mathcal{F}_j)$, for $j \in \mathbb{N}$, where $\mathcal{F}_j = \{E \cap \Omega_j,\ E \in \mathcal{F}\}$ and $\mu(E \cap \Omega_j) = \mu_1(E \cap \Omega_j) + \mu_2(E \cap \Omega_j) < \infty$. By countable additivity, $\mu(E) = \sum_j \mu(E \cap \Omega_j)$ with similar equalities for μ_1, μ_2, and ν, where by implication the two sides are in each case either finite and equal, or both $+\infty$. If $\nu(E \cap \Omega_j) = 0$ implies $\mu_2(E \cap \Omega_j) = 0$ for each j, then $\nu(E) = 0$ implies $\mu_2(E) = 0$ for $E \in \mathcal{F}$ and hence $\mu_2 \ll \nu$.

Theorem **4.30**(i) implies the existence of $\mathcal{F}_j/\mathcal{B}$-measurable non-negative f_j such that

$$\mu_2(E \cap \Omega_j) = \int_{E \cap \Omega_j} f_j \mathrm{d}\nu, \text{ all } E \in \mathcal{F}. \tag{4.57}$$

Define $f : \Omega \to \mathbb{R}^+$ by

$$f(\omega) = \sum_j 1_{\Omega_j}(\omega) f_j(\omega). \tag{4.58}$$

This is a function since the Ω_j are disjoint and is $\mathcal{F}/\mathcal{B}$-measurable since $f^{-1}(B) = \bigcup_j f_j^{-1}(B) = \bigcup_j E_j \in \mathcal{F}$ where $E_j \in \mathcal{F}_j$, for each $B \in \mathcal{B}$. Apply **4.13** to give

$$\begin{aligned} \mu_2(E) = \sum_j \mu_2(E \cap \Omega_j) &= \sum_j \left(\int_E 1_{\Omega_j} f_j \mathrm{d}\nu \right) \\ &= \int_E \sum_j 1_{\Omega_j} f_j \mathrm{d}\nu = \int_E f \mathrm{d}\nu. \end{aligned} \tag{4.59}$$

Similarly, let (A_j, A_j^c) define partitions of the Ω_j such that $\mu_1(A_j) = \nu(A_j^c) = 0$; then $A = \bigcup_j A_j$ and $A^c = \bigcup_j A_j^c$ are disjoint unions, $\mu_1(A) = \sum_j \mu_1(A_j) = 0$, and $\nu(A^c) = \sum_j \nu(A_j^c) = 0$. Hence $\mu_1 \perp \nu$. ∎

4.32 Example Let ν in the last result be Lebesgue measure m on $\mathbb{R}$, a σ-finite measure, and let $\mu = \mu_1 + \mu_2$ be any other measure on the line. In the majority of applications in the sequel μ is a probability measure and hence finite. Clearly, $\mu_1 \perp m$ requires that $\mu_1(E) = 0$ except when E is of Lebesgue measure 0. Absolute continuity of μ_2 with respect to m implies that any set of 'zero length', any countable collection of isolated points for example, has measure 0 under μ_2. If $\mu_1 = 0$ so that $\mu = \mu_2$ is absolutely continuous with respect to m, the integral of a measurable function g may be written as

$$\int_{-\infty}^{+\infty} g(x)\mathrm{d}\mu(x) = \int_{-\infty}^{+\infty} g(x)f(x)\mathrm{d}x \tag{4.60}$$

so that all integrals reduce to integrals with respect to Lebesgue measure. f is known as the density function of the measure μ and is an equivalent representation of μ, with the relation

$$\mu(E) = \int_E f(x)\mathrm{d}x \tag{4.61}$$

(the Lebesgue integral of f over E) holding for each $E \in \mathcal{B}$. □

4.33 Example Let ν denote counting measure on the positive integers (see **3.4** and **4.3**) so that in particular the singleton sets $\{i\}$ for $i \in \mathbb{N}$ are assigned measure 1. Let a real non-negative sequence $\{f_i, i \in \mathbb{N}\}$ define another measure μ on the integers such that $\mu(E) = \sum_{i \in E} f_i$. Both ν and μ are σ-finite and μ is a finite measure if the sequence is summable. μ is absolutely continuous with respect to ν since only empty sets have counting measure zero. Following the pattern of **4.3**, the integral of a function g (which is another sequence $\{g_i, i \in \mathbb{N}\}$) with respect to μ is the weighted sum $\sum_{i=1}^{\infty} g_i f_i$. In this case the Radon–Nikodym derivative is identical with the sequence $\{f_i\}$. □

5
Metric Spaces

5.1 Spaces

As remarked in §1.1 a space is just a set of objects, though usually distinguished in applications as the set that includes all the objects under consideration. In most applications of interest there are an infinite number of these. It is to Euclidean spaces such as the real line $\mathbb{R}$ that most attention has been directed so far.

The real line is the most familiar member of the class of *vector spaces*, also known as linear spaces. These have in common the property that two operations are defined on their elements, those of *addition* and *scalar multiplication*. $\mathbb{R}$ is trivially a vector space. In linear algebra where a vector is defined as a column of real numbers, the sum of two vectors of the same dimension is another vector whose elements are the sums of the corresponding elements. The product of a real vector and a scalar is the vector each of whose elements has been multiplied by the scalar. A vector space is closed under the operations of addition and scalar multiplication, in the sense that the operations always result in further elements of the same space. As well as real vectors with the familiar definition, vector spaces may also contain complex-valued vectors, infinite-dimensioned vectors, functions of a real variable, and other exotic cases.

Suppose that an element of a vector space, say $\boldsymbol{x}$, has associated with it a non-negative real number $||\boldsymbol{x}||$ called the *norm of* $\boldsymbol{x}$, satisfying the following properties: (i) $||\boldsymbol{x}|| \geq 0$, with equality only in the unique case called the zero element, $\mathbf{0}$; (ii) if a is a scalar and $a\boldsymbol{x}$ denotes the scalar product with $\boldsymbol{x}$, then $||a\boldsymbol{x}|| = |a|||\boldsymbol{x}||$; (iii) if $\boldsymbol{x}$ and $\boldsymbol{y}$ are two elements then $||\boldsymbol{x}+\boldsymbol{y}|| \leq ||\boldsymbol{x}|| + ||\boldsymbol{y}||$, the so-called *triangle inequality*. A space whose elements have this attribute is called a *normed space*. In $\mathbb{R}$ the norm of an element is its absolute value. In the familiar Euclidean case the norm of a vector is the square root of the sum of its squared elements and the zero vector is the vector of zeros. Notice that by considering the norm of $\boldsymbol{x}-\boldsymbol{y} = \boldsymbol{x}+(-1)\boldsymbol{y}$, the idea of a measure of distance between vectors is defined. Vectors are close to each other when the norm of their difference is small.

A *Banach space* is a normed space that is also complete. The notion of completeness of a space is examined in detail in §5.3, but in essence it is the property that a Cauchy sequence of elements always converges (in the sense that the norm of the successive differences converges to zero) to a limit point belonging to the space.

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0005

Lastly, a *Hilbert space* is a Banach space having one extra element of structure. The *inner product* $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ is a scalar defined for each pair of elements. This has the properties (i) symmetry, $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \langle \boldsymbol{y}, \boldsymbol{x} \rangle$; (ii) linearity, $\langle a\boldsymbol{x} + b\boldsymbol{y}, \boldsymbol{z} \rangle = a\langle \boldsymbol{x}, \boldsymbol{z} \rangle + b\langle \boldsymbol{y}, \boldsymbol{z} \rangle$; and (iii) positive definiteness, that is, $\langle \boldsymbol{x}, \boldsymbol{x} \rangle = ||\boldsymbol{x}||^2 \geq 0$, with equality only in the case $\boldsymbol{x} = \boldsymbol{0}$. In the familiar Euclidean case the inner product is of course the sum of products of the corresponding elements, often written $\boldsymbol{x}'\boldsymbol{y}$. The inner product allows concepts of orientation as well as distance to be defined. If $\langle \boldsymbol{x}, \boldsymbol{y} \rangle = 0$ the vectors are said to be orthogonal, or at right angles to one another. More generally, the *angle* between two vectors is a scalar $\theta \in [0, \pi]$ with the property

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = ||\boldsymbol{x}|| \, ||\boldsymbol{y}|| \cos\theta. \tag{5.1}$$

The familiar Euclidean vector spaces are Hilbert spaces, a fact reflected in their best-known properties. The relation in (5.1) is embodied in the Cauchy–Schwarz inequality and the bound on the inner product of vectors is derived in (2.29).

5.2 Distances and Metrics

Through their algebraic properties normed vector spaces are relatively familiar mathematical entities, but they form a subclass of the class of *metric spaces*. The common attribute is the idea of distance between set elements. Central to the properties of $\mathbb{R}$ studied in §1.8 was the idea that for any pair of real numbers x and y the Euclidean distance between them is the number $d_E(x,y) = |x - y| \in \mathbb{R}^+$. Generalizing this idea, a metric space is any set having such a distance measure defined for each pair of elements, even though possibly lacking the arithmetic properties that define vector spaces.

Let $\mathbb{S}$ denote such a set.

5.1 Definition A metric is a mapping $d : \mathbb{S} \times \mathbb{S} \mapsto \mathbb{R}^+$ having the properties

(a) $d(y,x) = d(x,y)$

(b) $d(x,y) = 0$ iff $x = y$

(c) $d(x,y) + d(y,z) \geq d(x,z)$ (triangle inequality). □

A metric space $(\mathbb{S}, d)$ is a set $\mathbb{S}$ paired with metric d, such that conditions (a)–(c) hold for each pair of elements of $\mathbb{S}$. If **5.1**(a) and (c) hold and $d(x,x) = 0$, but $d(x,y) = 0$ is also possible when $x \neq y$, d is called a *pseudo-metric.* A fundamental fact is that if (A, d) is a metric space and $B \subset A$, (B, d) is also a metric space. If $\mathbb{Q}$ is the set of rational numbers, $\mathbb{Q} \subset \mathbb{R}$ and $(\mathbb{Q}, d_E)$ is a metric space. Another example is the unit interval $([0,1], d_E)$.

While the Euclidean metric on $\mathbb{R}$ is the familiar case and the proof that d_E satisfies **5.1**(a)–(c) is elementary, d_E is not the only possible metric on $\mathbb{R}$.

5.2 Example For $x, y \in \mathbb{R}$ let

$$d_0(x,y) = \frac{|x-y|}{1+|x-y|}. \tag{5.2}$$

It is immediate that **5.1**(a) and (b) hold. To show (c), note that $|x-y| = d_0(x,y)/\big(1-d_0(x,y)\big)$. The inequality $a/(1-a) + b/(1-b) \geq c/(1-c)$ simplifies to $a+b \geq c+ab(2-c)$. **5.1**(c) is obtained on putting $a = d_0(x,y)$, $b = d_0(y,z)$, $c = d_0(x,z)$ and using the fact that $0 \leq d_0 \leq 1$. Unlike the Euclidean metric, d_0 is defined for x or $y = \pm\infty$. $(\bar{\mathbb{R}}, d_0)$ is a metric space on the definition, while $\bar{\mathbb{R}}$ with the Euclidean metric is not. □

In the space $\mathbb{R}^2$ a larger variety of metrics is found.

5.3 Example The Euclidean distance on $\mathbb{R}^2$ is

$$d_E(x,y) = \|x-y\| = \big[(x_1-y_1)^2 + (x_2-y_2)^2\big]^{1/2} \tag{5.3}$$

and $(\mathbb{R}^2, d_E)$ is the Euclidean plane. An alternative is the 'taxicab' metric,

$$d_T(x,y) = |x_1-y_1| + |x_2-y_2|. \tag{5.4}$$

d_E is the shortest distance between two addresses in Manhattan as the crow flies, but d_T is the shortest distance by taxi (see Figure 5.1). The reader will note that d_T and d_E are actually the cases for $p = 1$ and $p = 2$ of a sequence of metrics on $\mathbb{R}^2$. He/she is invited to supply the definition for the case $p = 3$ and so for any p. The limiting case as $p \to \infty$ is the maximum metric,

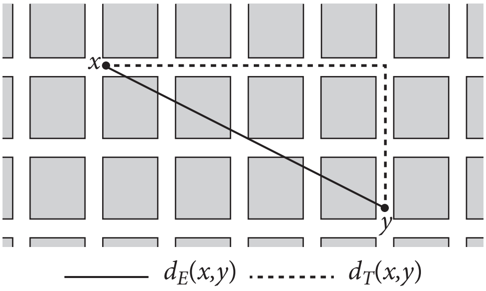

Figure 5.1

$$d_M(x,y) = \max\{|x_1 - y_1|, |x_2 - y_2|\}. \tag{5.5}$$

All these distance measures can be shown to satisfy **5.1**(a)–(c). Letting $\mathbb{R}^n = \mathbb{R} \times \mathbb{R} \times \ldots \times \mathbb{R}$ for any finite n, they can be generalized in the obvious fashion, to define metric spaces $(\mathbb{R}^n, d_E)$, $(\mathbb{R}^n, d_T)$, $(\mathbb{R}^n, d_M)$, and so forth. □

Metrics d_1 and d_2 on a space $\mathbb{S}$ are said to be *equivalent* if for each $x \in \mathbb{S}$ and $\varepsilon > 0$ there is a $\delta > 0$ such that for each $y \in \mathbb{S}$,

$$d_1(x,y) < \delta \Rightarrow d_2(x,y) < \varepsilon \tag{5.6}$$

and

$$d_2(x,y) < \delta \Rightarrow d_1(x,y) < \varepsilon. \tag{5.7}$$

The idea here is that the two metrics confer essentially the same properties on the space, apart from a possible relabelling of points and axes. A metric that is a continuous, increasing function of another metric is equivalent to it; thus, if d is any metric on $\mathbb{S}$, it is equivalent to the bounded metric $d/(1+d)$. d_E and d_0 of **5.2** are equivalent in $\mathbb{R}$, as are d_E and d_M in $\mathbb{R}^2$. On the other hand, consider for any $\mathbb{S}$ the *discrete* metric d_D, where for $x, y \in \mathbb{S}$, $d_D(x,y) = 0$ if $x = y$ and 1 otherwise. d_D is a metric, but d_D and d_E are not equivalent in $\mathbb{R}$.

In metric space theory, the properties of $\mathbb{R}$ outlined in §1.8 are revealed as a special case. Many definitions are the same, word for word, although other concepts are novel. In a metric space $(\mathbb{S}, d)$ the concept of an open neighbourhood in $\mathbb{R}$ generalizes to the *sphere* or *ball,* a set $S_d(x,\varepsilon) = \{y : y \in \mathbb{S}, d(x,y) < \varepsilon\}$ where $x \in \mathbb{S}$ and $\varepsilon > 0$. Write simply $S(x,\varepsilon)$ when the context makes clear which metric is being adopted. In $(\mathbb{R}^2, d_E)$, $S(x,\varepsilon)$ is a circle with centre at x and radius ε. In $(\mathbb{R}^2, d_T)$ it is a 'diamond' (rotated square) centred on x with ε the distance from x to the vertices. In $(\mathbb{R}^2, d_M)$ it is a regular square centred on x, with sides of 2ε. For $(\mathbb{R}^3, d_E)\ldots$ well, think about it!

An *open set* of $(\mathbb{S}, d)$ is a set $A \subseteq \mathbb{S}$ such that, for each $x \in A$, $\exists\, \delta > 0$ such that $S(x,\delta)$ is a subset of A. If metrics d_1 and d_2 are equivalent, a set is open in $(\mathbb{S}, d_1)$ iff it is open in $(\mathbb{S}, d_2)$. The theory of open sets of $\mathbb{R}$ generalizes straightforwardly. For example, the Borel field of $\mathbb{S}$ is a well-defined notion, the smallest σ-field containing the open sets of $(\mathbb{S}, d)$. Here is the general version of **1.33**.

5.4 Theorem

(i) If $\mathcal{C}$ is any collection of open sets of $(\mathbb{S}, d)$, then

$$C = \bigcup_{A \in \mathcal{C}} A \tag{5.8}$$

is open.

(ii) If A and B are open in $(\mathbb{S}, d)$, then $A \cap B$ is open.

Proof (i) If $S(x,\varepsilon) \subseteq A$ and $A \in \mathcal{C}$, then $S(x,\varepsilon) \subseteq C$. Since such a ball exists by definition for all $x \in A$ and all $A \in \mathcal{C}$, it follows that one exists for all $x \in C$.
(ii) If $S(x,\varepsilon_A)$ and $S(x,\varepsilon_B)$ are two spheres centred on x, then

$$S(x,\varepsilon_A) \cap S(x,\varepsilon_B) = S(x,\varepsilon) \tag{5.9}$$

where $\varepsilon = \min\{\varepsilon_A,\varepsilon_B\} > 0$. If $x \in A$, $\exists$ $S(x,\varepsilon_A) \subseteq A$ with $\varepsilon_A > 0$ and if $x \in B$, $\exists$ $S(x,\varepsilon_B) \subseteq B$ similarly, with $\varepsilon_B > 0$. If $x \in A \cap B$, $S(x,\varepsilon) \subseteq A \cap B$. ∎

The important thing to bear in mind is that openness is *not* preserved under arbitrary intersections.

A *closure point* of a set A is a point $x \in \mathbb{S}$ (not necessarily belonging to A) such that for any $\delta > 0$, $\exists$ $y \in A$ with $d(x,y) < \delta$. The set of closure points of A, denoted $\bar{A}$, is called the closure of A. Closure points are also called *adherent* points, 'sticking to' a set though not necessarily belonging to it. If for some $\delta > 0$ the definition of a closure point is satisfied only for $y = x$, so that $S(x,\delta) \cap A = \{x\}$, x is said to be an *isolated* point of A, while an *accumulation point* x of A is a closure point of $A - \{x\}$ and hence cannot be isolated.

A *boundary point* of A is a point $x \in \bar{A}$, such that for all $\delta > 0$, $\exists$ $z \in A^c$ with $d(x,z) < \delta$. The set of boundary points of A is denoted ∂A and $\bar{A} = A \cup \partial A$. The *interior* of A is $A^o = A - \partial A$. A *closed* set is one containing all its closure points, such that $\bar{A} = A$. An open set does not contain all of its closure points since the boundary points do not belong to the set. The empty set $\varnothing$ and the space $\mathbb{S}$ are both open *and* closed. A subset B of A is said to be *dense* in A if $B \subseteq A \subseteq \bar{B}$.

A collection of sets $\mathcal{C}$ is called a *covering* for A if $A \subseteq \bigcup_{B\in\mathcal{C}} B$, while a *subcovering* is any collection of sets belonging to $\mathcal{C}$ that also cover A. If each B is open, $\mathcal{C}$ is called an *open covering.* A set A is called *compact* if every open covering of A contains a finite subcovering. A is said to be *relatively compact* if $\bar{A}$ is compact. If $\mathbb{S}$ is itself compact, $(\mathbb{S}, d)$ is said to be a *compact space.* The remarks in §1.8 about compactness in $\mathbb{R}$ are equally relevant to the general case.

A is said to be *bounded* if $\exists$ $x \in A$ and $0 < r < \infty$, such that $A \subseteq S(x,r)$; and also *totally bounded* (or *precompact*) if for every $\varepsilon > 0$ there exists a finite collection of points $x_1,\dots,x_m$ (called an ε-net) such that the spheres $S(x_i,\varepsilon)$ for $i = 1,\dots,m$ form a covering for A. The $S(x_i,\varepsilon)$ can be replaced in this definition by their closures $\bar{S}(x_i,\varepsilon)$, noting that $\bar{S}(x_i,\varepsilon)$ is contained in $S(x_i,\varepsilon+\delta)$ for all $\delta > 0$. The points of the ε-net need not be elements of A. An attractive mental image is a region of $\mathbb{R}^2$ covered with little cocktail umbrellas of radius ε (Figure 5.2). Any set that is totally bounded is also bounded. In certain cases such as $(\mathbb{R}^n, d_E)$ the converse is also true, but this is *not* true in general.

5.5 Theorem If a set is relatively compact, it is totally bounded.

Proof Let A be relatively compact and consider the covering of $\bar{A}$ consisting of the ε-balls $S(x,\varepsilon)$ for all $x \in \bar{A}$. By the definition this contains a finite subcover

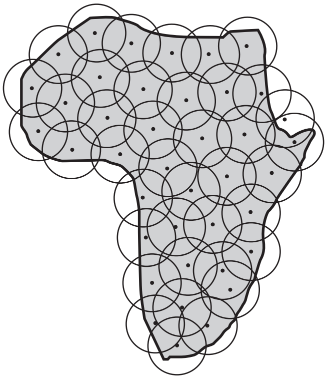

Figure 5.2

$S(x_i, \varepsilon)$, $i = 1, \ldots, m$, which also covers A. Then $\{x_1, \ldots, x_m\}$ is an ε-net for A and the theorem follows since ε is arbitrary. ∎

The converse is true only when the space is complete; see **5.13**.

5.3 Separability and Completeness

In thinking about metric spaces, it is sometimes helpful to visualize the analogue problem for $\mathbb{R}$, or at most for $\mathbb{R}^n$ with $n \leq 3$ and use one's intuitive knowledge of those cases. But this trick can be misleading if the space in question is too alien to geometrical intuition.

A metric space is said to be *separable* if it contains a countable, dense subset. Separability is one of the properties that might be considered to characterize an '$\mathbb{R}$-like' space. The rational numbers $\mathbb{Q}$ are countable and dense in $\mathbb{R}$, so $\mathbb{R}$ is separable, as is $\mathbb{R}^n$. An alternative definition of a separable metric space is a metric space for which the Lindelöf property holds (see **1.36**). This result can be given in the following form.

5.6 Theorem In a metric space $\mathbb{S}$ the following three properties are equivalent:

(a) $\mathbb{S}$ is separable.

(b) Every open set $A \subseteq \mathbb{S}$ has the representation

$$A = \bigcup_{i=1}^{\infty} B_i,\ B_i \in \mathcal{V} \tag{5.10}$$

where $\mathcal{V}$ is a countable collection of open spheres in $\mathbb{S}$.

(c) Every open cover of a set in $\mathbb{S}$ has a countable subcover. □

A collection $\mathcal{V}$ with property (b) is called a *base* of $\mathbb{S}$, so that separability is equated in this theorem with the existence of a countable base for the space. In topology this property is called *second-countability* (see §6.2). (c) is the Lindelöf property.

Proof of 5.6 First, show that (a) implies (b). Let $\mathcal{V}$ be the countable collection of spheres $\{S(x,r) : x \in D, r \in \mathbb{Q}^+\}$, where D is a countable, dense subset of $\mathbb{S}$ and $\mathbb{Q}^+$ is the set of positive rationals. If A is an open subset of $\mathbb{S}$, then for each $x \in A$, $\exists$ $\delta > 0$ such that $S(x,\delta) \subseteq A$. For any such x, choose $x_i \in D$ such that $d(x_i,x) < \delta/2$ (possible since D is dense) and then choose rational r_i to satisfy $d(x_i,x) < r_i < \delta/2$. Define $B_i = S(x_i,r_i) \in \mathcal{V}$ and observe that

$$x \in B_i \subseteq S(x,\delta) \subseteq A. \tag{5.11}$$

Since $\mathcal{V}$ as a whole is countable, the subcollection $\{B_i\}$ of all the sets that satisfy this condition for at least one $x \in A$ is also countable and clearly $A \subseteq \bigcup_i B_i \subseteq A$, so $A = \bigcup_i B_i$.

Next show that (b) implies (c). Since $\mathcal{V}$ is countable its elements may be indexed as $\{V_j, j \in \mathbb{N}\}$. If $\mathcal{C}$ is any collection of open sets covering A, choose a subcollection $\{C_j, j \in \mathbb{N}\}$, where C_j is a set from $\mathcal{C}$ which contains V_j if such exists, otherwise let $C_j = \emptyset$. There exists a covering of A by $\mathcal{V}$-sets as just shown, and each $\mathcal{V}_j$ can itself be covered by other elements of $\mathcal{V}$ with smaller radii so that by taking small enough spheres there is always an element of $\mathcal{C}$ to contain them. Thus $A \subseteq \bigcup_j C_j$ and the Lindelöf property holds.

Finally, to show that (c) implies (a), consider the open cover of $\mathbb{S}$ by the sets $\{S(x,1/n), x \in \mathbb{S}\}$. If there exists for each n a countable subcover $\{S(x_{nk},1/n), k \in \mathbb{N}\}$, for each k there must be one or more indices k' such that $d(x_{nk},x_{nk'}) < 2/n$. Since this must be true for every n, the countable set $\{x_{nk}, k \in \mathbb{N}, n \in \mathbb{N}\}$ must be dense in $\mathbb{S}$. This completes the proof. ∎

The theorem has a useful corollary.

5.7 Corollary A totally bounded space is separable. □

Another important property is that subspaces of separable spaces are separable, shown as follows.

5.8 Theorem If $(\mathbb{S},d)$ is a separable space and $A \subset \mathbb{S}$, then (A,d) is separable.

Proof Suppose D is countable and dense in $\mathbb{S}$. Construct the countable set E by taking one point from each nonempty set

$$A \cap S(y,r),\ y \in D,\ r \in \mathbb{Q}^+. \tag{5.12}$$

For any $x \in A$ and $\delta > 0$ choose $y \in D$ so that $d(x,y) < \delta/2$. For every such y, $\exists\, z \in E$ satisfying $z \in A \cap S(y,r)$ for $r < \delta/2$, so that $d(y,z) < \delta/2$. Thus

$$d(x,z) \le d(x,y) + d(y,z) < \delta \tag{5.13}$$

and since x and δ are arbitrary it follows that E is dense in A. ∎

This argument does not rule out the possibility that A and D are disjoint. The separability of the irrational numbers, $\mathbb{R} - \mathbb{Q}$, is a case in point.

On the other hand, certain conditions are incompatible with separability. A subset A of a metric space $(\mathbb{S}, d)$ is *discrete* if for each $x \in A$, $\exists\, \delta > 0$ such that $\big(S(x,\delta) - \{x\}\big) \cap A$ is empty. In other words, each element is an isolated point. The integers $\mathbb{Z}$ are a discrete set of $(\mathbb{R}, d_E)$, for example. If $\mathbb{S}$ is itself discrete, the discrete metric d_D is equivalent to d.

5.9 Theorem If a metric space contains an uncountable discrete subset, it is not separable.

Proof This is immediate from **5.6**. Let A be discrete and consider the open set $\bigcup_{x\in A} S(x,\varepsilon_x)$, where ε_x is chosen small enough that the specified spheres form a disjoint collection. This is an open cover of A and if A is uncountable it has no countable subcover. ∎

The separability question arises when defining measures on metric spaces (see Chapter 27). Unless a space is separable there is no guarantee that its Borel sets are all measurable. The space $D_{[a,b]}$ discussed below (see **5.27**) is an important example of this difficulty.

The concepts of sequence, limit, subsequence, and cluster point all extend from $\mathbb{R}$ to general metric spaces. A sequence $\{x_n\}$ of points in $(\mathbb{S}, d)$ is said to *converge* to a limit x if for all $\varepsilon > 0$ there exists $N_\varepsilon \ge 1$ such that

$$d(x_n, x) < \varepsilon \text{ for all } n > N_\varepsilon. \tag{5.14}$$

Theorems **2.2** and **2.3** extend in an obvious way, as follows.

5.10 Theorem Every sequence in a compact subset of a metric space $(\mathbb{S}, d)$ has one or more cluster points. □

5.11 Theorem If a sequence in a compact subset of a metric space $(\mathbb{S}, d)$ has a unique cluster point, then it converges. □

The notion of a Cauchy sequence also remains fundamental. A sequence $\{x_n\}$ of points in a metric space $(\mathbb{S}, d)$ is a Cauchy sequence if for all $\varepsilon > 0$, $\exists\ N_\varepsilon$ such that $d(x_n, x_m) < \varepsilon$ whenever $n > N_\varepsilon$ and $m > N_\varepsilon$. The novelty is that Cauchy sequences in a metric space do not always possess limits. It is possible that the point on which the sequence is converging lies outside the space. Consider the space $(\mathbb{Q}, d_E)$. The sequence $\{x_n\}$, where $x_n = 1 + 1/2 + 1/6 + \ldots + 1/n! \in \mathbb{Q}$, is a Cauchy sequence since $|x_{n+1} - x_n| = 1/(n+1)! \to 0$; but of course, $x_n \to \mathrm{e}$ (the base of the natural logarithms), an irrational number. A metric space $(\mathbb{S}, d)$ is said to be *complete* if it contains the limits of all Cauchy sequences defined on it. $(\mathbb{R}, d_E)$ is a complete space, while $(\mathbb{Q}, d_E)$ is not. In honour of the various distinguished Polish mathematicians who have worked on these questions the term *Polish space* is sometimes applied to a metric space that is separable and complete, although strictly this is a topological concept; a Polish space is a topological space that is metrizable and homeomorphic to a separable complete metric space. Chapter 6 gives the relevant details.

Although compactness is a primitive notion that does not require the concept of a Cauchy sequence, it can nevertheless be defined in terms of the properties of sequences, following the idea in **2.2**. This is often convenient from a practical point of view.

5.12 Theorem The following statements about a metric space $(\mathbb{S}, d)$ are equivalent:

(a) $\mathbb{S}$ is compact.

(b) Every sequence in $\mathbb{S}$ has a cluster point in $\mathbb{S}$.

(c) $\mathbb{S}$ is totally bounded and complete. □

Notice the distinction between completeness and compactness. In a complete space all Cauchy sequences converge, which says nothing about the behaviour of non-Cauchy sequences. But in a compact space, which is also totally bounded, *all* sequences contain Cauchy subsequences which converge in the space.

Proof of 5.12 This is by showing in turn that (a) implies (b), (b) implies (c), and (c) implies (a). Suppose $\mathbb{S}$ is compact. Let $\{x_n, n \in \mathbb{N}\}$ be a sequence in $\mathbb{S}$ and so define a decreasing sequence of nonempty subsets of $\mathbb{S}$ by $B_n = \{x_k : k \geq n\}$. The sets $\bar{B}_n$ are closed and the cluster points of the sequence, if any, compose the set $C = \bigcap_{n=1}^{\infty} \bar{B}_n = (\bigcup_{n=1}^{\infty} \bar{B}_n^c)^c$. If $C = \varnothing$, $\mathbb{S} = \bigcup_{n=1}^{\infty} \bar{B}_n^c$, so that the open sets $\bar{B}_n^c$ are a cover for $\mathbb{S}$ and by assumption these contain a finite subcover. This means that, for some $m < \infty$, $\mathbb{S} \subseteq \bigcup_{n=1}^{m} \bar{B}_n^c = (\bigcap_{n=1}^{m} \bar{B}_n)^c = \bar{B}_m^c$, leading to the contradiction $\bar{B}_m = \varnothing$. It follows that C must be nonempty, so (a) implies (b).

Now suppose that every sequence has a cluster point in $\mathbb{S}$. Considering the case of Cauchy sequences, it is clear that the space is complete; it remains to be shown that it is totally bounded. Suppose not: then there must exist an $\varepsilon > 0$ for which no ε-net exists; in other words, no finite n and points $\{x_1, \dots, x_n\}$ such that $d(x_j, x_k) \leq \varepsilon$ for all $j \neq k$. But letting $n \to \infty$ in this case a sequence has been found with no cluster point, which is again a contradiction. Hence, (b) implies (c).

Finally, let $\mathcal{C}$ be an arbitrary open cover of $\mathbb{S}$, assume that $\mathcal{C}$ contains no finite subcover of $\mathbb{S}$, and obtain a contradiction. Since $\mathbb{S}$ is totally bounded it must possess for each $n \geq 1$ a finite cover of the form

$$B_{ni} = S(x_{ni}, 1/2^n), i = 1, \dots, k_n. \tag{5.15}$$

Fixing n, by hypothesis it is always possible to choose an i for which B_{ni} has no finite cover by $\mathcal{C}$-sets. Call this set D_n. For $n > 1$, $\{B_{ni}\}_{i=1}^{k_n}$ is also a covering for D_{n-1} and hence D_n can be chosen so that $D_n \cap D_{n-1}$ is nonempty. Choose a sequence of points $\{x_n \in D_n, n \in \mathbb{N}\}$. Since D_n is a ball of radius $1/2^n$ containing x_n and D_{n+1} is of radius $1/2^{n+1}$ containing x_{n+1}, $d(x_n, x_{n+1}) < 3/2^n$. The triangle inequality implies that $d(x_n, x_{n+m}) < 3\sum_{i=0}^{m} 2^{-n-i} \leq 6/2^n \to 0$ as $n \to \infty$, thus $\{x_n\}$ is a Cauchy sequence and converges to a limit $x \in \mathbb{S}$, by completeness.

Choose a set $A \in \mathcal{C}$ containing x. Since A is open, $S(x, \varepsilon) \subset A$ for some $\varepsilon > 0$. Since for any n $d(x_n, x) < 6/2^n$ and x_n is in D_n which has radius $1/2^n$, choosing n large enough that $9/2^n < \varepsilon$ ensures that $D_n \subset S(x, \varepsilon)$. But this means $D_n \subset A$, which is a contradiction since D_n has no finite cover by $\mathcal{C}$-sets. Hence $\mathcal{C}$ must contain a finite subcover and (c) implies (a). ∎

In complete spaces, the set properties of relative compactness and precompactness (total boundedness) are identical. The following is the converse of **5.5**.

5.13 Corollary In a complete metric space a totally bounded set A is relatively compact.

Proof If $\mathbb{S}$ is complete, every Cauchy sequence in A has a limit in $\mathbb{S}$ and all such points are closure points of A. The subspace $(\bar{A}, d)$ is therefore a complete space. It follows from **5.12** that if A is totally bounded, $\bar{A}$ is compact. ∎

5.4 Examples

The following are cases somewhat more remote from ordinary geometric intuition than the ones looked at above.

5.14 Example In §12.3 and subsequently the infinite-dimensional Euclidean space $\mathbb{R}^\infty$ is encountered. If $x = (x_1, x_2, \ldots) \in \mathbb{R}^\infty$ and $y = (y_1, y_2, \ldots) \in \mathbb{R}^\infty$ similarly, a metric for $\mathbb{R}^\infty$ is given by

$$d_\infty(x,y) = \sum_{k=1}^{\infty} 2^{-k} d_0(x_k, y_k) \tag{5.16}$$

where d_0 is defined in (5.2). Like d_0, d_∞ is a bounded metric with $d_\infty(x,y) \le 1$ for all x and y. □

5.15 Theorem $(\mathbb{R}^\infty, d_\infty)$ is separable and complete.

Proof To show separability, consider the collection

$$\begin{aligned} A_m &= \{x = (x_1, x_2, \ldots) : x_k \text{ rational if } k \le m,\, x_k = 0 \text{ otherwise}\} \\ &\subseteq \mathbb{R}^\infty. \end{aligned} \tag{5.17}$$

A_m is countable and by **1.5** the collection $A = \bigcup_{m=1}^{\infty} A_m$ is also countable. For any $y \in \mathbb{R}^\infty$ and $\varepsilon > 0$, $\exists\, x \in A_m$ such that

$$d_\infty(x,y) \le \sum_{k=1}^{m} 2^{-k}\varepsilon + \sum_{k=m+1}^{\infty} 2^{-k} d_0(0, y_k) \le \varepsilon + 2^{-m}. \tag{5.18}$$

Since the right-hand side can be made as small as desired by choice of ε and m, y is a closure point of A. Hence, A is dense in $\mathbb{R}^\infty$.

To show completeness, suppose $\{x_n = (x_{1n}, x_{2n}, \ldots), n \in \mathbb{N}\}$ is a Cauchy sequence in $\mathbb{R}^\infty$. Since $d_0(x_{kn}, x_{km}) \le 2^k d_\infty(x_n, x_m)$ for any k, $\{x_{kn}, n \in \mathbb{N}\}$ must be a Cauchy sequence in $\mathbb{R}$. Since

$$d_\infty(x, x_n) \le \sum_{k=1}^{m} 2^{-k} d_0(x_k, x_{kn}) + 2^{-m} \tag{5.19}$$

for all m, $x_n \to x = (x_1, x_2, \ldots) \in \mathbb{R}^\infty$ iff $x_{kn} \to x_k$ for each $k = 1, 2, \ldots$. The completeness of $\mathbb{R}$ implies that $\{x_n\}$ has a limit in $\mathbb{R}^\infty$. ∎

5.16 Example Consider the 'infinite-dimensional cube', $[0,1]^\infty$; the Cartesian product of an infinite collection of unit intervals. The space $\left([0,1]^\infty, d_\infty\right)$ is separable by **5.15** and **5.8**. $[0,1]^\infty$ can be endowed with the equivalent and in this case bounded metric,

$$\rho_\infty(x,y) = \sum_{k=1}^{\infty} 2^{-k}|x_k - y_k|. \quad \square \tag{5.20}$$

In a metric space $(\mathbb{S}, d)$, where d can be assumed bounded without loss of generality, define the distance between a point $x \in \mathbb{S}$ and a subset $A \subseteq \mathbb{S}$ as $d(x,A) = \inf_{y \in A} d(x,y)$. Then for a pair of subsets A, B of $(\mathbb{S}, d)$ define the function

$$d_H : 2^{\mathbb{S}} \times 2^{\mathbb{S}} \mapsto \mathbb{R}^+$$

where $2^{\mathbb{S}}$ is the power set of $\mathbb{S}$, by

$$d_H(A,B) = \max\Big\{\sup_{x \in B} d(x,A), \sup_{y \in A} d(y,B)\Big\}. \tag{5.21}$$

$d_H(A,B)$ is called the Hausdorff distance between sets A and B.

5.17 Theorem Letting $\mathcal{H}_S$ denote the compact, nonempty subsets of $\mathbb{S}$, $(\mathcal{H}_{\mathbb{S}}, d_H)$ is a metric space.

Proof Clearly d_H satisfies **5.1**(a). It satisfies **5.1**(b) since the sets of $\mathcal{H}_{\mathbb{S}}$ are closed, although note that $d_H(A, \bar{A}) = 0$, so that d_H is only a pseudo-metric for general subsets of $\mathbb{S}$. To show **5.1**(c), for any $x \in A \in \mathcal{H}_S$ and any $z \in C \in \mathcal{H}_S$ the definition of $d(x,B)$ and the fact that d is a metric implies

$$\sup_{x \in A} d(x,B) \le \sup_{x \in A}\{d(x,z) + d(z,B)\}. \tag{5.22}$$

Since C is compact, the infimum over C of the expression in braces on the right-hand side above is attained at a point $z \in C$. Therefore,

$$\begin{aligned}\sup_{x \in A} d(x,B) &\le \sup_{x \in A}\Big\{\inf_{z \in C}\{d(x,z) + d(z,B)\}\Big\} \\ &\le \sup_{x \in A} d(x,C) + \sup_{z \in C} d(z,B). \end{aligned} \tag{5.23}$$

Similarly, $\sup_{y \in B} d(y,A) \le \sup_{z \in C} d(z,A) + \sup_{y \in B} d(y,C)$ and hence,

$$\begin{aligned} d_H(A,B) &\le \max\Big\{\sup_{x \in A} d(x,C) + \sup_{z \in C} d(z,B), \sup_{z \in C} d(z,A) + \sup_{y \in B} d(y,C)\Big\} \\ &\le d_H(B,C) + d_H(A,C). \quad \blacksquare \end{aligned} \tag{5.24}$$

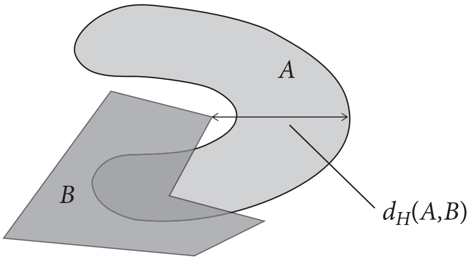

Figure 5.3

When $(\mathbb{S}, d)$ is complete, it can be shown that $(\mathcal{H}_{\mathbb{S}}, d_H)$ is also complete.

5.18 Example Let $\mathbb{S} = \mathbb{R}$ in Theorem **5.17**. The compact intervals with the Hausdorff metric define a complete metric space. Thus, $\{[0, 1 - 1/n], n \in \mathbb{N}\}$ is a Cauchy sequence which converges in the Hausdorff metric to $[0, 1]$. This is the closure of the set $[0, 1)$, the limit of the sequence under the Euclidean metric (compare **2.6**). Although $[0, 1) \notin \mathcal{H}_{\mathbb{S}}$, observe that $d_H\big([0, 1), [0, 1]\big) = 0$. □

Another case is where $\mathbb{S} = (\mathbb{R}^2, d_E)$ and $\mathcal{H}_{\mathbb{S}}$ contains the closed and bounded subsets of the Euclidean plane. To cultivate intuition about metric spaces, a useful exercise is to draw some figures on a sheet of paper and measure the Hausdorff distances between them, as in Figure 5.3. For compact A and B, $d_H(A,B) = 0$ if and only if $A = B$; compare this with another intuitive concept of the 'distance between two sets', $\inf_{x \in A, y \in B} d_E(x, y)$, which is zero if the sets touch or intersect.

5.5 Mappings on Metric Spaces

A function has been defined as a mapping which takes set elements to unique points of $\mathbb{R}$, but the term is also used where the codomain is a general metric space. Where the domain is another metric space, the results of §2.2 arise as special cases of the theory. Some of the following properties are generalizations of those given previously, while others are new. The terms mapping, transformation, etc. are again synonyms for function, but an extra usage is *functional,* which refers to the case where the domain is a space whose elements are themselves functions, with (usually) $\mathbb{R}$ as co-domain. An example is the integral defined in §4.1.

The function $f : (\mathbb{S}, d) \mapsto (\mathbb{T}, \rho)$ is said to be *continuous at x* if for all $\varepsilon > 0$ $\exists$ $\delta > 0$ such that

$$\sup_{y \in S_d(x,\delta)} \rho\big(f(y), f(x)\big) < \varepsilon. \tag{5.25}$$

Here, δ may depend on x. Another way to state the condition is that for $\varepsilon > 0$ $\exists$ $\delta > 0$ such that

$$f(S_d(x,\delta)) \subseteq S_\rho(f(x),\varepsilon) \tag{5.26}$$

where S_d and S_ρ are respectively balls in $(\mathbb{S},d)$ and $(\mathbb{T},\rho)$. Similarly, f is said to be *uniformly continuous* on a set $A \subseteq \mathbb{S}$ if for all $\varepsilon > 0$ $\exists$ $\delta > 0$ such that

$$\sup_{x\in A}\ \sup_{y\in S_d(x,\delta)\cap A} \rho(f(y),f(x)) < \varepsilon. \tag{5.27}$$

Theorem **2.7** was a special case of the following important result.

5.19 Theorem For $A \subseteq \mathbb{T}$, $f^{-1}(A)$ is open (closed) in $\mathbb{S}$ whenever A is open (closed) in $\mathbb{T}$, iff f is continuous at all points of $\mathbb{S}$.

Proof Assume A is open and let $f(x) \in A$ so that $x \in f^{-1}(A)$. Then $S_\rho(f(x),\varepsilon) \subseteq A$ for some $\varepsilon > 0$. By **1.2**(iv) and (5.26) with continuity at x,

$$S_d(x,\delta) \subseteq f^{-1}(f(S_d(x,\delta))) \subseteq f^{-1}(S_\rho(f(x),\varepsilon)) \subseteq f^{-1}(A). \tag{5.28}$$

If A is open then $\mathbb{T}-A$ is closed and $f^{-1}(\mathbb{T}-A) = \mathbb{S}-f^{-1}(A)$ by **1.2**(iii), which is closed if $f^{-1}(A)$ is open. This proves sufficiency.

To prove necessity, suppose $f^{-1}(A)$ is open in $\mathbb{S}$ whenever A is open in $\mathbb{T}$ and in particular, $f^{-1}(S_\rho(f(x),\varepsilon))$ for $\varepsilon > 0$ is open in $\mathbb{S}$. Since $x \in f^{-1}(S_\rho(f(x),\varepsilon))$, there is a $\delta > 0$ such that (5.26) holds. Use complements again for the case of closed sets. ∎

This property of inverse images under f provides an alternative characterization of continuity and in topological spaces provides the primary definition of continuity. The notion of Borel measurability discussed in §3.6 extends naturally to mappings between pairs of metric spaces and the theorem establishes that continuous transformations are Borel-measurable.

The properties of functions on compact sets are of interest in a number of contexts. The essential results are as follows.

5.20 Theorem The continuous image of a compact set is compact.

Proof It is to be shown that if $A \subseteq \mathbb{S}$ is compact and f is continuous then $f(A)$ is compact. Let $\mathcal{C}$ be an open covering of $f(A)$. Continuity of f means that the sets $f^{-1}(B)$, $B \in \mathcal{C}$ are open by **5.19** and their union covers A by **1.2**(ii). Since A is compact, these sets contain a finite subcover, say, $f^{-1}(B_1),\dots,f^{-1}(B_m)$. It follows that

$$f(A) \subseteq f\left(\bigcup_{j=1}^{m} f^{-1}(B_j)\right) = \bigcup_{j=1}^{m} f(f^{-1}(B_j)) \subseteq \bigcup_{j=1}^{m} B_j \tag{5.29}$$

where the equality is by **1.2**(i) and the second inclusion by **1.2**(v). Hence, $B_1, \dots, B_m$ is a finite subcover of $f(A)$ by $\mathcal{C}$-sets. Since $\mathcal{C}$ is arbitrary, it follows that $f(A)$ is compact. ∎

5.21 Theorem If f is continuous on a compact set, it is uniformly continuous on the set.

Proof Let $A \subseteq \mathbb{S}$ be compact. For each $x \in A$, continuity at x means that for $\varepsilon > 0$ there exists a sphere $S_d(x, r)$ (r may depend on x) such that $\rho(f(x), f(y)) < \frac{1}{2}\varepsilon$ for each $y \in S_d(x, 2r) \cap A$. These balls cover A and since A is compact they contain a finite subcover, say $S_d(x_k, r_k)$, $k = 1, \dots, m$. Let $\delta = \min_{1 \le k \le m} r_k$ and consider a pair of points $x, y \in \mathbb{S}$ such that $d(x, y) < \delta$. Now, $y \in S_d(x_k, r_k)$ for some k, so that $\rho(f(x_k), f(y)) < \frac{1}{2}\varepsilon$ and also

$$d(x_k, x) \le d(x_k, y) + d(x, y) \le r_k + \delta \le 2r_k \tag{5.30}$$

using the triangle inequality. Hence $\rho(f(x_k), f(x)) \le \frac{1}{2}\varepsilon$ and

$$\rho\big(f(x), f(y)\big) \le \rho\big(f(x), f(x_k)\big) + \rho\big(f(x_k), f(y)\big) < \varepsilon. \tag{5.31}$$

Since δ does not depend on x or y, f is uniformly continuous on A. ∎

If $f : \mathbb{S} \mapsto \mathbb{T}$ is 1–1 onto and f and f^{-1} are continuous, f is called a *homeomorphism* and $\mathbb{S}$ and $\mathbb{T}$ are said to be *homeomorphic* if such a function exists. If $\mathbb{S}$ is homeomorphic with a subset of $\mathbb{T}$, it is said to be *embedded* in $\mathbb{T}$ by f. If f also preserves distances so that $\rho(f(x), f(y)) = d(x, y)$ for each $x, y \in \mathbb{S}$, it is called an *isometry.* Metrics d_1 and d_2 in a space $\mathbb{S}$ are equivalent if and only if the identity mapping from $(\mathbb{S}, d_1)$ to $(\mathbb{S}, d_2)$ (the mapping which takes each point of $\mathbb{S}$ into itself) is a homeomorphism.

5.22 Example If d_∞ and ρ_∞ are the metrics defined in (5.16) and (5.20) respectively, the mapping $g : (\mathbb{R}^\infty, d_\infty) \to ([0,1]^\infty, \rho_\infty)$, where $g = (g_1, g_2, \dots)$ and

$$g_i(x) = \frac{1}{2} + \frac{x_i}{2(1 + |x_i|)}, i = 1, 2, \dots, \tag{5.32}$$

is a homeomorphism. □

Right and left continuity are not well-defined notions for general metric spaces, but there is a concept of continuity which is 'one-sided' with respect to the *range* of the function. A function $f: (\mathbb{S}, d) \mapsto \mathbb{R}$ is said to be *upper semicontinuous* at x if for each $\varepsilon > 0$, $\exists\ \delta > 0$ such that, for $y \in \mathbb{S}$,

$$d(x,y) < \delta \Rightarrow f(y) < f(x) + \varepsilon. \tag{5.33}$$

If $\{x_n\}$ is a sequence of points in $\mathbb{S}$ and $d(x_n, x) \to 0$, upper semicontinuity implies $\limsup_n f(x_n) \leq f(x)$. The *level sets* of the form $\{x : f(x) < \alpha\}$ are open for all $\alpha \in \mathbb{R}$ iff f is upper semicontinuous everywhere on $\mathbb{S}$. f is *lower semicontinuous* iff $-f$ is upper semicontinuous and f is continuous at x iff it is both upper and lower semicontinuous at x.

A function of a real variable is upper semicontinuous at x even if it jumps at x with $f(x) \geq \max\{f(x-), f(x+)\}$. Isolated discontinuities such as point A in Figure 5.4 are not ruled out if this inequality is satisfied. On the other hand, if the illustrated function takes its value at x_0 at point B (i.e., is right continuous at that point) then upper semi-continuity fails at x_0. While it is unintuitive that a set defined by a strict inequality should contain a boundary point, this is just what happens with the level set at x_0 if α is a point of the ordinate within the discontinuity. Thus, semicontinuity is not the same as right/left continuity except in the case of monotone functions; if f is increasing, right (left) continuity is equivalent to upper (lower) semicontinuity and the reverse holds for decreasing functions.

The concept of a Lipschitz condition generalizes to metric spaces. A function f on $(\mathbb{S}, d)$ satisfies a Lipschitz condition at $x \in \mathbb{S}$ if for $\delta > 0$, $\exists\ M > 0$ such that, for any $y \in S_d(x, \delta)$,

$$\rho\big(f(y),\ f(x)\big) \leq Mh\big(d(x,y)\big) \tag{5.34}$$

where $h(\cdot) : \mathbb{R}^+ \mapsto \mathbb{R}^+$ satisfies $h(d) \downarrow 0$ as $d \downarrow 0$. It satisfies a *uniform* Lipschitz condition if condition (5.34) holds uniformly, with fixed M, for all $x \in \mathbb{S}$. The remarks following (2.8) apply equally here. Continuity is enforced by this condition with arbitrary h and stronger smoothness conditions are obtained for special cases of h.

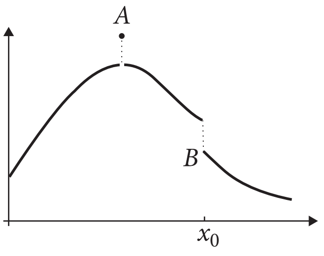

Figure 5.4

5.6 Function Spaces

The non-Euclidean metric spaces met in later chapters are mostly spaces of real functions on an interval of $\mathbb{R}$. The elements of such spaces are graphs, subsets of $\mathbb{R}^2$. However, most of the relevant theory holds for functions whose domain is any metric space $(\mathbb{S}, d)$ and it is this general case that is treated here. To interpret the results it will usually be helpful to visualize the case of $\mathbb{S} = [0, 1]$, the unit interval of the line with the Euclidean metric, which will receive most attention in the sequel. On the other hand, to focus ideas the discussion of this section is confined to the case of continuous functions as defined in §2.2, chiefly because intuitions about Euclidean space can make the properties of these objects relatively easy to appreciate. Extending to functions having discontinuities introduces some major difficulties and these cases are treated in detail in Chapter 30.

Thus, let $C_{\mathbb{S}}$ denote the set of all bounded continuous functions, $f : \mathbb{S} \mapsto \mathbb{R}$. To make $C_{\mathbb{S}}$ a metric space, define for any pair of elements f and g the function

$$d_U(f,g) = \sup_{x \in \mathbb{S}} |f(x) - g(x)|. \tag{5.35}$$

5.23 Theorem d_U is a metric.

Proof Conditions **5.1**(a) and (b) are immediate. To prove the triangle inequality write, given functions f, g and $h \in C_{\mathbb{S}}$,

$$\begin{aligned} d_U(f,h) &= \sup_{x \in \mathbb{S}} |f(x) - g(x) + g(x) - h(x)| \\ &\leq \sup_{x \in \mathbb{S}} (|f(x) - g(x)| + |g(x) - h(x)|) \\ &\leq d_U(f,g) + d_U(g,h). \quad \blacksquare \end{aligned} \tag{5.36}$$

d_U is called the *uniform* metric and $(C_{\mathbb{S}}, d_U)$ is a metric space.

An important subset of $C_{\mathbb{S}}$ is the space $U_{\mathbb{S}}$ of *uniformly* continuous functions. If $\mathbb{S}$ is compact, $C_{\mathbb{S}} = U_{\mathbb{S}}$ by **5.21**. Also if $\mathbb{S}$ is relatively compact and $\bar{\mathbb{S}}$ is its closure, every $f \in C_{\mathbb{S}}$ has a uniformly continuous restriction to $\mathbb{S}$ and every $f \in U_{\mathbb{S}}$ has a continuous extension to $\bar{\mathbb{S}}$, say $\bar{f}$, constructed by setting $\bar{f}(x) = f(x)$ for $x \in \mathbb{S}$ and $\bar{f}(x) = \lim_n f(x_n)$ for a sequence $\{x_n \in \mathbb{S}\}$ converging to x, for each $x \in \bar{\mathbb{S}} - \mathbb{S}$. Note that for any pair $f, f' \in \mathbb{S}$, $d_U(f,f') = d_U(\bar{f}, \bar{f}')$, so that the spaces $C_{\mathbb{S}}$ and $U_{\mathbb{S}}$ are isometric. There are functions that are continuous on $\mathbb{S}$ and not on $\bar{\mathbb{S}}$, but these cannot be uniformly continuous.

The following basic property of $C_{\mathbb{S}}$ holds independently of the nature of the domain $\mathbb{S}$.

5.24 Theorem $(C_{\mathbb{S}}, d_U)$ is complete.

Proof Let $\{f_n\}$ be a Cauchy sequence in $C_{\mathbb{S}}$; in other words, for $\varepsilon > 0 \ \exists \ N_\varepsilon \geq 1$ such that $d_U(f_n, f_m) \leq \varepsilon$ for $n, m > N_\varepsilon$. This means that for each $x \in \mathbb{S}$ the sequences $\{f_n(x)\}$ satisfy $|f_n(x) - f_m(x)| \leq d_U(f_n, f_m)$; these are Cauchy sequences in $\mathbb{R}$ and so have limits $f(x)$. In view of the definition of d_U, $f_n \to f$ uniformly in $\mathbb{S}$. For any $x, y \in \mathbb{S}$, the triangle inequality gives

$$|f(x) - f(y)| \leq |f(x) - f_n(x)| + |f_n(x) - f_n(y)| + |f_n(y) - f(y)|. \tag{5.37}$$

Fix $\varepsilon > 0$. Since $f_n \in C_{\mathbb{S}}$, $\exists \ \delta > 0$ such that $|f_n(x) - f_n(y)| < \frac{1}{3}\varepsilon$ if $d(x, y) < \delta$. Also, by uniform convergence $\exists \ n$ large enough that

$$\max\{|f(x) - f_n(x)|, \ |f_n(y) - f(y)|\} < \tfrac{1}{3}\varepsilon \tag{5.38}$$

so that $|f(x) - f(y)| < \varepsilon$. Hence $f \in C_{\mathbb{S}}$, which establishes that $C_{\mathbb{S}}$ is complete. ∎

Notice how this property holds by virtue of the uniform metric. It is easy to devise sequences of continuous functions converging to discontinuous limits, but none of these are Cauchy sequences. It is not possible for a continuous function to be arbitrarily close to a discontinuous function at *every* point of the domain.

A number of the results to follow call for a continuous approximating function to be exhibited, lying uniformly close to a given function in $U_{\mathbb{S}}$ but fully specified by a finite collection of numbers. This is possible when the domain is totally bounded. In the following result, rational values are specified so that the set of all possible approximating functions is countable.

5.25 Theorem Let $(\mathbb{S}, d)$ be a totally bounded metric space. For any $f \in U_{\mathbb{S}}$, there exists for any $\varepsilon > 0$ a function $g \in U_{\mathbb{S}}$, completely specified by points of the domain $x_1, \ldots, x_m$ and rational numbers $a_1, \ldots, a_m$, such that $d_U(f, g) < \varepsilon$.

Proof By total boundedness of $\mathbb{S}$, $\exists$ for $\delta > 0$ a finite δ-net $\{x_1, \ldots, x_m\}$. For each x_i let $A_i = \{x : d(x, x_i) \geq 2\delta\}$ and $B_i = \left\{x : d(x, x_i) \leq \frac{1}{2}\delta\right\}$ and define functions $g_i : \mathbb{S} \mapsto [0, 1]$ by

$$g_i(x) = \frac{d(x, A_i)}{d(x, A_i) + d(x, B_i)} \tag{5.39}$$

where $d(x,A) = \inf_{y \in A} d(x,y)$, a uniformly continuous function of x. Thus, $g_i(x) = 1$ when $d(x,x_i) \le \delta/2$ and $g_i(x) = 0$ when $d(x,x_i) \ge 2\delta$, with continuous interpolation over intermediate values. The denominator is never less than $\frac{3}{2}\delta$ and hence $g_i(x)$ is also uniformly continuous. Now define the function

$$g(x) = \frac{\sum_{i=1}^m g_i(x) a_i}{\sum_{i=1}^m g_i(x)}. \tag{5.40}$$

Since $\{x_i\}$ is a δ-net for $\mathbb{S}$ there exists for every $x \in \mathbb{S}$ *some* i such that $d(x,A_i) \ge \delta$ as well as $d(x,B_i) \le \delta$ and hence such that $g_i \ge \frac{1}{2}$. Hence, $\sum_{i=1}^m g_i(x) \ge \frac{1}{2}$ for all $x \in \mathbb{S}$ and uniform continuity extends from the g_i to g.

For arbitrary $f \in U_{\mathbb{S}}$, fix $\varepsilon > 0$ and choose δ small enough that $|f(x) - f(y)| < \frac{1}{2}\varepsilon$ when $d(x,y) < 2\delta$, for any $x,y \in \mathbb{S}$. Then, fix m large enough that $x_1, \ldots, x_m$ and $a_1, \ldots, a_m$ can be chosen with the balls $S(x_i, \delta)$ covering $\mathbb{S}$ and $|f(x_i) - a_i| < \frac{1}{2}\varepsilon$ for each i. Since $g_i(x) = 0$ if $d(x,x_i) \ge 2\delta$ it is always the case that

$$g_i(x)\,|f(x) - f(x_i)| \le \tfrac{1}{2} g_i(x)\varepsilon.$$

Hence, for each $x \in \mathbb{S}$ and each i,

$$\begin{aligned} g_i(x)\,|f(x) - a_i| &\le g_i(x)\,|f(x) - f(x_i)| + g_i(x)\,|f(x_i) - a_i| \\ &< g_i(x)\varepsilon \end{aligned} \tag{5.41}$$

and

$$\begin{aligned} d_U(f,g) &= \sup_{x \in \mathbb{S}} |f(x) - g(x)| \\ &\le \sup_{x \in \mathbb{S}} \left\{ \frac{\sum_{i=1}^m g_i(x)\,|f(x) - a_i|}{\sum_{i=1}^m g_i(x)} \right\} < \varepsilon. \quad \blacksquare \end{aligned} \tag{5.42}$$

The next result makes use of this approximation theorem and is fundamental. It shows (recalling the earlier discussion of separability) that spaces of continuous functions are not such alien objects from an analytic point of view as they might at first appear, at least when the domain is totally bounded.

5.26 Theorem

(i) If $(\mathbb{S}, d)$ is totally bounded then $(U_{\mathbb{S}}, d_U)$ is separable.

(ii) If $(\mathbb{S}, d)$ is compact then $(C_{\mathbb{S}}, d_U)$ is separable.

Proof It is enough to prove part (i), since for part (ii), $C_{\mathbb{S}} = U_{\mathbb{S}}$ by **5.21** and the same conclusion follows.

Fix m and suitable points $x_1, \dots, x_m$ of $\mathbb{S}$ so as to define a countable family of functions $A_m = \{g_{mk}, k \in \mathbb{N}\}$, where the g_{mk} are defined as in **5.25** and the index k enumerates the countable collection of sets of m rationals. For each $\varepsilon > 0$, there exists m large enough that, for each $f \in U_{\mathbb{S}}$, $d_U(f, g_{mk}) < \varepsilon$ for some k. By **1.5**, $A = \lim_{m\to\infty} A_m$ is countable and there exists $g_k \in A$ such that $d_U(f, g_k) < \varepsilon$ for every $\varepsilon > 0$. This means that A is dense in $U_{\mathbb{S}}$. ∎

To show that these properties cannot be relied on to hold under more general circumstances, here is a nonseparable function space.

5.27 Example For $\mathbb{S}$ an interval of the real line $[a, b]$, consider the metric space $(D_{[a,b]}, d_U)$ of real, bounded càdlàg functions of a real variable. 'Càdlàg' is a colourful French acronym (continue à droite, limite à gauche) to describe functions of a real variable that may have discontinuities but are right continuous at every point, with the image of every decreasing sequence in $[a, b]$ containing its limit point; in other words, there is a limit point to the left of every point. Of course, $C_{[a,b]} \subseteq D_{[a,b]}$. Functions with completely arbitrary discontinuities form a larger class still, but one that for most purposes is too unstructured to permit a useful theory.

To show that $(D_{[a,b]}, d_U)$ is not separable, consider the subset with elements

$$f_\theta(t) = \begin{cases} 0, & t < \theta \\ 1, & t \geq \theta, \end{cases} \quad \theta \in [a, b]. \tag{5.43}$$

This set is uncountable, containing as many elements as there are points in $[a, b]$. But $d_U(f_\theta, f'_\theta) = 1$ for $\theta \neq \theta'$, so it is also discrete. Hence $(D_{[a,b]}, d_U)$ is not separable by **5.9**. □

Let A denote a collection of functions $f : (\mathbb{S}, d) \mapsto (\mathbb{T}, \rho)$. A is said to be *equicontinuous* at $x \in \mathbb{S}$ if $\forall\, \varepsilon > 0\ \exists\, \delta > 0$ such that

$$\sup_{f\in A}\ \sup_{y\in S_d(x,\delta)} \rho\big(f(y), f(x)\big) < \varepsilon. \tag{5.44}$$

A is also said to be *uniformly equicontinuous* if $\forall\, \varepsilon > 0\ \exists\, \delta > 0$ such that

$$\sup_{f\in A}\ \sup_{x\in\mathbb{S}}\ \sup_{y\in S_d(x,\delta)} \rho\big(f(y), f(x)\big) < \varepsilon. \tag{5.45}$$

Equicontinuity is the property of a set of continuous functions (or as it may be, uniformly continuous functions) which forbids limit points of the set to be points of discontinuity. In the case when $A \subseteq C_{\mathbb{S}}$ (or $A \subseteq U_{\mathbb{S}}$) but A is not (uniformly) equicontinuous, the possibility that $\bar{A} \nsubseteq C_{\mathbb{S}}$ ($\bar{A} \nsubseteq U_{\mathbb{S}}$) cannot be ruled out.

An important class of applications is to countable A. Restricting attention to the case $A = \{f_n, n \in \mathbb{N}\}$, it may be that only the behaviour in the tail of the sequence matters, in which case the following concept is the relevant one. A sequence of functions $\{f_n, n \in \mathbb{N}\}$ is said to be *asymptotically equicontinuous* at x if $\forall\ \varepsilon > 0\ \exists\ \delta > 0$ such that

$$\limsup_{n\to\infty}\left\{\sup_{y\in S_d(x,\delta)} \rho\big(f_n(y), f_n(x)\big)\right\} < \varepsilon \tag{5.46}$$

and also *asymptotically uniformly equicontinuous* if $\forall\ \varepsilon > 0\ \exists\ \delta > 0$ such that

$$\limsup_{n\to\infty}\left\{\sup_{x\in\mathbb{S}}\ \sup_{y\in S_d(x,\delta)} \rho\big(f_n(y), f_n(x)\big)\right\} < \varepsilon. \tag{5.47}$$

If the functions f_n are indeed continuous for all n then '$\limsup_{n\to\infty}$' can be replaced by '$\sup_n$' in (5.46) and similarly for (5.47) when all the f_n are uniformly continuous. In these circumstances, the qualifier 'asymptotic' can be dropped.

The main result on equicontinuous sets is the Arzelà–Ascoli theorem. This designation covers a number of closely related results, but the following version, which is the one appropriate to subsequent developments, identifies equicontinuity as the property of a set of bounded real-valued functions on a totally bounded domain which converts boundedness into *total* boundedness.

5.28 Theorem (Arzelà–Ascoli) Let $(\mathbb{S}, d)$ be a totally bounded metric space. A set $A \subseteq C_{\mathbb{S}}$ is relatively compact under d_U iff it is bounded and uniformly equicontinuous. □

It is convenient for this result to define the *modulus of continuity* for an element of C. For a fixed positive constant δ let this be the functional

$$w(\cdot, \delta) : (C_{\mathbb{S}}, d_U) \mapsto (\mathbb{R}^+, d_E)$$

where for $f \in C_{\mathbb{S}}$,

$$w(f, \delta) = \sup_{x\in\mathbb{S}}\ \sup_{y\in S_d(x,\delta)} |f(y) - f(x)|. \tag{5.48}$$

A key property of w is that it is continuous in its argument, in the sense of the following lemma. If two functions are close in the uniform metric, then so are their moduli of continuity.

5.29 Lemma For $f,g \in C_{\mathbb{S}}$ and any $\delta > 0$,

$$|w(f,\delta) - w(g,\delta)| \leq 2d_U(f,g). \tag{5.49}$$

Proof Let x_f, y_f denote the solutions to (5.48) in the sense that

$$w(f,\delta) = |f(y_f) - f(x_f)|. \tag{5.50}$$

Note that $d(y_f, x_f) < \delta$ by definition and hence

$$|g(y_f) - g(x_f)| \leq w(g,\delta). \tag{5.51}$$

It follows by (5.50), the triangle inequality, and (5.51) that

$$\begin{aligned} w(f,\delta) &\leq |g(y_f) - g(x_f)| + |f(y_f) - g(y_f)| + |g(x_f) - f(x_f)| \\ &\leq w(g,\delta) + 2d_U(f,g). \end{aligned} \tag{5.52}$$

However, the same argument with f and g interchanged gives

$$w(g,\delta) \leq w(f,\delta) + 2d_U(f,g). \tag{5.53}$$

Putting together (5.52) and (5.53) yields (5.49). ∎

Proof of 5.28 Since $C_{\mathbb{S}}$ is complete, total boundedness of A is equivalent to relative compactness by **5.5** and **5.13**. To prove 'if' assume boundedness and equicontinuity and construct a finite ε-net for A. Boundedness of A under the uniform metric means that there exist finite real numbers U and L such that

$$L \leq \inf_{f \in A, x \in \mathbb{S}} f(x) \leq \sup_{f \in A, x \in \mathbb{S}} f(x) \leq U. \tag{5.54}$$

Let $\{x_1, \ldots, x_m\}$ be a δ-net for $\mathbb{S}$ and construct the finite family

$$D_m = \{g_k \in A, k = 1, \ldots, (\nu + 1)^m\} \tag{5.55}$$

according to the recipe of **5.25**, with the constants $a_1, \ldots, a_m$ drawn from the finite set

$$\{L + (U - L)u/\nu,\ u = 0, \ldots, \nu,\ \nu > (U - L)/\varepsilon\}.$$

This set contains $\nu + 1$ real values between U and L less than ε apart, so that D_m has $(\nu + 1)^m$ members, as indicated. Since the assumptions imply $A \subseteq U_{\mathbb{S}}$ it follows by **5.25** that for every $f \in A$ there exists $g_k \in D_m$ with $d_U(f, g_k) < \varepsilon$. This shows that D_m is a ε-net for A and A is totally bounded.

To prove 'only if', suppose A is relatively compact and hence totally bounded by **5.5**. Trivially, total boundedness implies boundedness. To show uniform equicontinuity, consider for $\varepsilon > 0$ the set

$$B_k(\varepsilon) = \{f : w(f, 1/k) < \varepsilon\} \tag{5.56}$$

where w is defined in (5.48). According to (5.45), uniform equicontinuity of A is the condition that for any $\varepsilon > 0$ there exists k large enough that $\bar{A} \subseteq B_k(\varepsilon)$. $B_k(\varepsilon)$ is the inverse image under $w(\cdot, \delta)$ of the half-line $[0, \varepsilon)$ which is open in $\mathbb{R}^+$. Hence, since $w(., \delta)$ is continuous by **5.29** the set $B_k(\varepsilon)$ is open by **5.19**. By definition of $C_{\mathbb{S}}$, $w(f, 1/k) \to 0$ as $k \to \infty$ for each $f \in C_{\mathbb{S}}$. In other words, w converges to 0 pointwise on $C_{\mathbb{S}}$, which implies that the collection $\{B_k(\varepsilon), k \in \mathbb{N}\}$ is an open covering for $C_{\mathbb{S}}$ and hence for $\bar{A}$. But by hypothesis $\bar{A}$ is compact. Every such covering of $\bar{A}$ has a finite subcover and so $\bar{A} \subseteq B_k(\varepsilon)$ for finite k, as required. ∎

6
Topology

6.1 Topological Spaces

Metric spaces form a subclass of a larger class of mathematical objects called topological spaces. These need not have a distance defined upon them but the concepts of open set, neighbourhood, and continuous mapping are still well defined. Even though only metric spaces are encountered in the sequel (principally in Part VI), much of the reasoning is essentially topological in character. An appreciation of the topological underpinnings is essential for getting to grips with the theory of weak convergence. The present chapter aims to review the main ideas. For additional details and results Kelley [112] and Willard [189] may be found useful.

6.1 Definition A *topological space* $(\mathbb{X}, \tau)$ is a set $\mathbb{X}$ on which is defined a *topology*, a class of subsets τ called *open sets* having the following properties:

(a) $\mathbb{X} \in \tau, \varnothing \in \tau$

(b) If $\mathcal{C} \subseteq \tau$, then $\bigcup_{O \in \mathcal{C}} O \in \tau$

(c) If $O_1 \in \tau, O_2 \in \tau$, then $O_1 \cap O_2 \in \tau$. □

These three conditions *define* an open set, so that openness becomes a primitive concept of which the notion of ε-spheres around points is only one characterization. A metric induces a topology on a space because it is one way (though not the only way) of defining what an open set is and all metric spaces are also topological spaces. On the other hand, some topological spaces may be made into metric spaces by defining a metric on them under which sets of τ are open in the sense defined in §5.2. Such spaces are called *metrizable.*

A subset of a topological space $(\mathbb{X}, \tau)$ has a topology naturally induced on it by the parent space. If $A \subset \mathbb{X}$, the collection $\tau_A = \{A \cap O : O \in \tau\}$ is called the *relative topology* for A. (A, τ_A) would normally be referred to as a *subspace* of $\mathbb{X}$. If two topologies τ_1 and τ_2 are defined on a space and $\tau_1 \subset \tau_2$, then τ_1 is said to be *coarser*, or *weaker*, than τ_2, whereas τ_2 is *finer (stronger)* than τ_1. In particular, the power set of $\mathbb{X}$ is a topology, called the *discrete* topology, whereas $\{\varnothing, \mathbb{X}\}$ is called the *trivial* topology. Two metrics define the same topology on a space if and only if they are

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0006

equivalent. If two points are close in one space, their images in the other space must be correspondingly close.

If a set O is open, its complement O^c in $\mathbb{X}$ is said to be closed. The *closure* $\bar{A}$ of an arbitrary set $A \subseteq \mathbb{X}$ is the intersection of all the closed sets containing A. As for metric spaces, a set $A \subseteq B$ for $B \subseteq \mathbb{X}$ is said to be *dense* in B if $B \subseteq \bar{A}$.

6.2 Theorem The intersection of any collection of closed sets is closed. $\mathbb{X}$ and $\varnothing$ are both open and closed. □

However, an arbitrary union of closed sets need not be closed, just as an arbitrary intersection of open sets need not be open.

For given $\mathbb{X}$ a collection $\mathcal{V}_x$ of open sets is called a *base* for the point $x \in \mathbb{X}$ if for every open O containing x there is a set $B \in \mathcal{V}_x$ such that $x \in B$ and $B \subset O$. This is the generalization to topological spaces of the idea of a system of neighbourhoods or spheres in a metric space. A *base for the topology* τ on $\mathbb{X}$ is a collection $\mathcal{V}$ of sets such that, for every $O \in \tau$ and every $x \in O$, there exists $B \in \mathcal{V}$ such that $x \in B \subset O$. The definition implies that any open set can be expressed as the union of sets from the base of the topology; a topology may be defined for a space by specifying a base collection and letting the open sets be defined as the unions and finite intersections of the base sets. In the case of $\mathbb{R}$, for example, the open intervals form a base.

6.3 Theorem A collection $\mathcal{V}$ is a base for a topology τ on $\mathbb{X}$ iff

(a) $\bigcup_{B \in \mathcal{V}} B = \mathbb{X}$

(b) $\forall\, B_1, B_2 \in \mathcal{V}$ and $x \in B_1 \cap B_2$, $\exists\, B_3 \in \mathcal{V}$ such that $x \in B_3 \subset B_1 \cap B_2$.

Proof Necessity of these conditions follows from the definitions of base and open set. For sufficiency, define a collection τ in terms of the base $\mathcal{V}$, as follows:

$$O \in \tau \text{ iff, for each } x \in O, \exists\, B \in \mathcal{V} \text{ such that } x \in B \subset O. \tag{6.1}$$

$\varnothing$ satisfies the condition in (6.1) and $\mathbb{X}$ satisfies it given condition (a) of the theorem. If $\mathcal{C}$ is a collection of τ-sets, $\bigcup_{O \in \mathcal{C}} O \in \tau$ since (6.1) holds in this case in respect of a base set B corresponding to any set in $\mathcal{C}$ which contains x. Also, if $O_1, O_2 \in \tau$ and $x \in O_1 \cap O_2$ then, letting B_1 and B_2 be the base sets specified in (6.1) in respect of x and O_1 and O_2 respectively, condition (b) implies that $x \in B_3 \subset O_1 \cap O_2$, which shows that τ is closed under finite intersections. Hence, τ is a topology for $\mathbb{X}$. ■

The concept of base allows further notions familiar from metric spaces to be defined for topological spaces. The *closure points* of a set A in a topological space

$(\mathbb{X},\tau)$ are the points $x \in \mathbb{X}$ such that every set in the base of x contains a point of A. An important exercise is to show that x is a closure point of A if and only if x is in the closure of A. The *accumulation points* of A in $(\mathbb{X},\tau)$ are the points $x \in \mathbb{X}$ such that every set in the base of x contains a point of A other than x.

Two other concepts familiar from metric spaces are convergence and cluster points of a sequence. A sequence $\{x_n\}$ of points in a topological space is said to *converge* to x if for every open set O containing x, $\exists\ N \geq 1$ such that $x_n \in O$ for all $n \geq N$. And x is called a *cluster point* of $\{x_n\}$ if for every open O containing x and every $N \geq 1$, $x_n \in O$ for some $n \geq N$. In general topological spaces the notion of a convergent sequence is inadequate for characterizing basic properties such as the continuity of mappings and is augmented by the concepts of *net* (in the topological sense of the word) and *filter*. However, these extensions are not as a rule required in the context of metric spaces (see e.g. [112] ch. 2L, [189] ch. 4).

6.2 Countability and Compactness

The *countability axioms* provide one classification of topological spaces according, roughly speaking, to their degree of structure and amenability to the methods of analysis. A topological space is said to satisfy the first axiom of countability (to be *first-countable*) if every point of the space has a countable base. It satisfies the second axiom of countability (is *second-countable*) if the space as a whole has a countable base. Every metric space is first-countable in view of the existence of the countable base composed of open spheres, $S(x, 1/n)$ for each x. More generally, sequences in first-countable spaces tend to behave in a similar manner to those in metric spaces, as the following theorem illustrates.

6.4 Theorem In a first-countable space, x is a cluster point of a sequence $\{x_n, n \in \mathbb{N}\}$ iff there is a subsequence $\{x_{n_k}, k \in \mathbb{N}\}$ converging to x.

Proof Sufficiency is immediate. For necessity, the definition of a cluster point implies that $\exists\ n \geq N$ such that $x_n \in O$ for every open O containing x and every $N \geq 1$. Let the countable base of x be the collection $\mathcal{V}_x = \{B_i, i \in \mathbb{N}\}$ and choose a monotone sequence of base sets $\{A_k, k \in \mathbb{N}\}$ containing x (and hence nonempty) with $A_1 = B_1$ and $A_k \subset A_{k-1} \cap B_k$ for $k = 2, 3, \ldots$. This is always possible by **6.3**. Since x is a cluster point, an infinite subsequence can be constructed by taking x_{n_k} as the next member of the sequence contained in A_k, for $k = 1, 2, \ldots$. For every open set O containing x, $\exists\ N \geq 1$ such that $x_{n_k} \in A_k \subseteq O$, for all $k \geq N$ and hence $x_{n_k} \to x$ as $k \to \infty$, as required. ∎

The point of quoting a result such as this has less to do with demonstrating a new property than with highlighting the need for caution in assuming properties taken for granted in metric spaces. While the intuition derived from $\mathbb{R}$-like situations might suggest that the existence of a cluster point and a convergent subsequence amount to the same thing, this need not be the case without first-countability.

A topological space $\mathbb{X}$ is said to be *separable* if it contains a countable dense subset, and to have the *Lindelöf property* if every open cover of $\mathbb{X}$ has a countable subcover. That second-countable spaces are separable follows directly on taking a point from each set in a countable base and verifying that these points are dense in the space. The converse is not generally true, but it is true for metric spaces, where separability, second countability, and Lindelöf are all equivalent to one another. This is just what was shown in **5.6**. More generally:

6.5 Theorem A second-countable space is both separable and Lindelöf.

Proof The proof of separability is in the text above. To prove the Lindelöf property, let $\mathcal{C}$ be an open cover of $\mathbb{X}$, such that $\bigcup_{A\in\mathcal{C}} A = \mathbb{X}$. For each $A \in \mathcal{C}$ and $x \in A$, there is a base set B_i such that $x \in B_i \subset A$. Since $\bigcup_{i=1}^{\infty} B_i = \mathbb{X}$, a countable subcollection A_i, $i = 1,2,\ldots$ can be chosen such that $B_i \subset A_i$ for each i and hence $\bigcup_{i=1}^{\infty} A_i = \mathbb{X}$. ∎

A topological space is said to be *compact* if every covering of the space by open sets has a finite subcovering. It is said to be *countably compact* if each countable covering has a finite subcovering. And it is said to be *sequentially compact* if each sequence on the space has a convergent subsequence. Sometimes, compactness is more conveniently characterized in terms of the complements. The complements of an open cover of $\mathbb{X}$ are a collection of closed sets whose intersection is empty, by **1.1**(iii). If and only if $\mathbb{X}$ is compact, every such collection must have a finite subcollection with empty intersection. An equivalent way to state this proposition is in terms of the converse implication. A collection of sets is said to have the *finite intersection property* if *no* finite subcollection has an empty intersection. If a collection of closed sets has empty intersection then their complements cover $\mathbb{X}$ and if a finite subcollection of closed sets has empty intersection then their complements are a finite subcover of $\mathbb{X}$. Thus, another way to characterize compactness is as follows.

6.6 Theorem $\mathbb{X}$ is compact (countably compact) iff no collection (countable collection) of closed sets having the finite intersection property has an empty intersection. □

The following pair of theorems summarize important relationships between the different varieties of compactness.

6.7 Theorem A first-countable space $\mathbb{X}$ is countably compact iff it is sequentially compact.

Proof Let the space be countably compact. Let $\{x_n, n \in \mathbb{N}\}$ be a sequence in $\mathbb{X}$ and define the sets $B_n = \{x_n, x_{n+1}, \dots\}$, $n = 1, 2, \dots$. The collection of closed sets $\{\bar{B}_n, n \in \mathbb{N}\}$ clearly possesses the finite intersection property and hence $\bigcap_n \bar{B}_n$ is nonempty by **6.6**, which is another way of saying that $\{x_n\}$ has a cluster point. Since the sequence is arbitrary, sequential compactness follows by **6.4**. This proves necessity.

For sufficiency, **6.4** implies that under sequential compactness all sequences in $\mathbb{X}$ have a cluster point. Let $\{C_i, i \in \mathbb{N}\}$ be a countable collection of closed sets having the finite intersection property such that $A_n = \bigcap_{i=1}^n C_i \neq \emptyset$, for every finite n. Consider a sequence $\{x_n\}$ chosen such that $x_n \in A_n$ and note since $\{A_n\}$ is monotone that $x_n \in A_m$ for all $n \geq m$; or in other words, A_m contains the sequence $\{x_n, n \geq m\}$. Since $\{x_n\}$ has a cluster point x and A_m is closed, $x \in A_m$. This is true for every $m \in \mathbb{N}$, so that $\bigcap_{i=1}^\infty C_i$ is nonempty and $\mathbb{X}$ is countably compact by **6.6**. ∎

6.8 Theorem A metric space $(\mathbb{S}, d)$ is countably compact iff it is compact.

Proof Sufficiency is immediate. For necessity, first show that if $\mathbb{S}$ is countably compact then it is separable. A metric space is first-countable, hence countable compactness implies sequential compactness (**6.7**), which in turn implies that every sequence in $\mathbb{S}$ has a cluster point (**6.4**). This must mean that for any $\varepsilon > 0$ there exists a finite ε-net $\{x_1, \dots, x_m\}$ such that for all $x \in \mathbb{S}$, $d(x, x_k) < \varepsilon$ for some $k \in \{1, \dots, m\}$. If this were not the case, an infinite sequence $\{x_n\}$ could be constructed with $d(x_n, x_{n'}) \geq \varepsilon$ for $n \neq n'$, contradicting the existence of a cluster point. It follows that for each $n \in \mathbb{N}$ there is a collection of countable sets $A_1, \dots, A_n$ such that, for every $x \in \mathbb{S}$, for some n there is a point $y \in A_n$ for which $d(x, y) < 2^{-n}$. The set $D = \bigcup_{n=1}^\infty A_n$ is countable by **1.5** and dense in $\mathbb{S}$ and hence $\mathbb{S}$ is separable.

Separability in a metric space is equivalent by **5.6** to the Lindelöf property that every open cover of $\mathbb{S}$ has a countable subcover, whereas countable compactness implies that this countable subcover has a finite subcover in its turn. Compactness is therefore proved. ∎

Like separability and compactness, the notion of a continuous mapping may be defined in terms of a distance measure, but is really topological in character. In a pair of topological spaces $\mathbb{X}$ and $\mathbb{Y}$, the mapping $f : \mathbb{X} \mapsto \mathbb{Y}$ is said to be *continuous* if $f^{-1}(B)$ is open in $\mathbb{X}$ when B is open in $\mathbb{Y}$ and closed in $\mathbb{X}$ when B is closed in $\mathbb{Y}$. That in metric spaces this definition is equivalent to the more familiar one in terms of ε- and δ-neighbourhoods follows from **5.19**. The concepts of homeomorphism and embedding, as mappings that are respectively onto or into and 1–1 continuous with continuous inverse, remain well defined. The following theorem gives two important properties of continuous maps.

6.9 Theorem Suppose there exists a continuous mapping f from a topological space $\mathbb{X}$ *onto* another space $\mathbb{Y}$.

(i) If $\mathbb{X}$ is separable, $\mathbb{Y}$ is separable.

(ii) If $\mathbb{X}$ is compact, $\mathbb{Y}$ is compact.

Proof (i) The problem is to exhibit a countable, dense subset of $\mathbb{Y}$. Consider $f(D)$ where D is dense in $\mathbb{X}$. If $\overline{f(D)}$ is the closure of $f(D)$, the inverse image $f^{-1}\big(\overline{f(D)}\big)$ is closed by continuity of f and contains $f^{-1}\big(f(D)\big)$ and hence also contains D by **1.2**(iv). Since $\bar{D}$ is the smallest closed set containing D and $\mathbb{X} \subseteq \bar{D}$, it follows that $\mathbb{X} \subseteq f^{-1}\big(\overline{f(D)}\big)$. But since the mapping is onto, $\mathbb{Y} = f(\mathbb{X}) \subseteq f\big(f^{-1}\big(\overline{f(D)}\big)\big) \subseteq \overline{f(D)}$, where the last inclusion is by **1.2**(v). $f(D)$ is therefore dense in $\mathbb{Y}$ as required. $f(D)$ is countable if D is countable and the conclusion follows directly.

(ii) Let $\mathcal{C}$ be an open cover of $\mathbb{Y}$. Then $\{f^{-1}(B) : B \in \mathcal{C}\}$ must be an open cover of $\mathbb{X}$ by the definition. The compactness of $\mathbb{X}$ means that it contains a finite subcover, say $f^{-1}(B_1), \ldots, f^{-1}(B_n)$ such that

$$\mathbb{Y} = f(\mathbb{X}) = f\Big(\bigcup_{j=1}^{n} f^{-1}(B_j)\Big) = f\Big(f^{-1}\Big(\bigcup_{j=1}^{n} B_j\Big)\Big) \subseteq \bigcup_{j=1}^{n} B_j \tag{6.2}$$

where the third equality uses **1.2**(ii) and the inclusion, **1.2**(v). Hence $\mathcal{C}$ contains a finite subcover. ∎

Note the importance of the stipulation ‘onto’ in both these results. The extension of (ii) to the case of compact subsets of $\mathbb{X}$ and $\mathbb{Y}$ is obvious and can be supplied by the reader.

Completeness, unlike separability, compactness, and continuity, is *not* a topological property. To define a Cauchy sequence it is necessary to have the concept of a distance between points. In this connection recall the concept of a Polish space that was explained in §5.3. One of the advantages of defining a metric on a space is that the relatively weak notion of completeness provides some of the essential features of compactness in a wider class than the compact spaces.

6.3 Separation Properties

Another classification of topological spaces is provided by the *separation axioms*, which in one sense are more primitive than the countability axioms. They are indicators of the richness of a topology, in the sense of being able to distinguish between different points of the space. From one point of view, they could be said to define a hierarchy of resemblances between topological spaces and metric spaces. Don't confuse separation with separability, which is a different concept altogether. A topological space $\mathbb{X}$ is said to be:

- a T_1-space, iff $\forall\, x, y \in \mathbb{X}$ with $x \neq y$ $\exists$ an open set containing x but not y and also an open set containing y but not x;
- a *Hausdorff* (*or* T_2-) space, iff $\forall\, x, y \in \mathbb{X}$ with $x \neq y$ $\exists$ disjoint open sets O_1 and O_2 in $\mathbb{X}$ with $x \in O_1$ and $y \in O_2$;
- a *regular* space iff for each closed set C and $x \notin C$ $\exists$ disjoint open sets O_1 and O_2 with $x \in O_1$ and $C \subset O_2$;
- a *normal* space iff, given disjoint closed sets C_1 and C_2, $\exists$ disjoint open sets O_1 and O_2 such that $C_1 \subset O_1$ and $C_2 \subset O_2$.

A regular T_1-space is called a T_3-space and a normal T_1-space is called a T_4-space.

In a T_1-space, the singleton sets $\{x\}$ are always closed. In this case $y \in \{x\}^c$ whenever $y \neq x$, where $\{x\}^c$ is the complement of a closed set and hence open. Conversely, if the T_1 property holds, every $y \neq x$ is contained in an open set not containing x and the union of all these sets, also open by **6.1**(b), is $\{x\}^c$. It is easy to see that T_4 implies T_3 implies T_2 implies T_1, although the reverse implications do not hold and without the T_1 property, normality need not imply regularity. Metric spaces are always T_4, for there is no difficulty in constructing the sets specified in the definition out of unions of ε-spheres.

Here are the important links between separation, compactness, countability, and metrizability. The first two results are of general interest but will not be exploited directly in this book, so the proofs are omitted. The proof of **6.12** needs some as yet undefined concepts and is postponed to §6.6 below, see page 142.

6.10 Theorem A regular Lindelöf space is normal. □

6.11 Theorem A compact Hausdorff space is T_4. □

6.12 Theorem (Urysohn metrization) A second-countable T_4-space is metrizable. □

In fact, the conditions of the last theorem can be weakened with T_4 replaced by T_3 in view of **6.10**, since as already shown a second-countable space is Lindelöf (**6.5**).

The properties of functions from $\mathbb{X}$ to the real line play an important role in defining the separation properties of a space. The key to these results is the remarkable Urysohn's lemma.

6.13 Lemma A topological space $\mathbb{X}$ is normal iff for any pair A and B of disjoint closed subsets there exists a continuous function $f: \mathbb{X} \to [0,1]$ such that $f(A) = 0$ and $f(B) = 1$.

The function f is called a *separating function.*

Proof This is by construction of the required function. Recall that the dyadic rationals $\mathbb{D}$ are dense in $[0,1]$. There exists a system of open sets $\{U_r, r \in \mathbb{D}\}$ with the properties

$$A \subset U_r \tag{6.3}$$

$$\bar{U}_r \cap B = \varnothing \tag{6.4}$$

$$\bar{U}_s \subseteq U_r \text{ for } r > s. \tag{6.5}$$

Normality implies the existence of an open set $U_{1/2}$ (say) such that $U_{1/2}$ contains A and $(\overline{U_{1/2}})^c$ contains B. The same story can be told with $U^c_{1/2}$ replacing B in the role of C_2 to define $U_{1/4}$ and then again with $\overline{U_{1/2}}$ replacing A in the role of C_1 to define $U_{3/4}$. The argument extends by induction to generate sets $\{U_{m/2^n}, m = 1, \ldots, 2^n - 1\}$ for any $n \in \mathbb{N}$ and the collection $\{U_r, r \in \mathbb{D}\}$ is obtained on letting $n \to \infty$. It is easy to verify conditions (6.3)–(6.5) for this collection. Figure 6.1 illustrates the construction for $n = 3$ when A and B are regions of the plane. One must imagine countably many more 'layers of the onion' in the limiting case.

Now define $f: \mathbb{X} \to [0,1]$ by

$$f(x) = \begin{cases} \inf \{r \in \mathbb{D} : x \in U_r\}, & x \in \bigcup_{r\in\mathbb{D}} U_r \\ 1, & x \in \mathbb{X} - \bigcup_{r\in\mathbb{D}} U_r \end{cases} \tag{6.6}$$

where, in particular, $f(x) = 1$ for $x \in B$. Because of the monotone property (6.5),

$$\begin{aligned} \{x : f(x) < \alpha\} &= \{x : \inf \{r \in \mathbb{D} : x \in U_r\} < \alpha\} \\ &= \{x : \exists\, r < \alpha \text{ such that } x \in U_r\} \\ &= \bigcup_{r<\alpha} U_r \end{aligned} \tag{6.7}$$

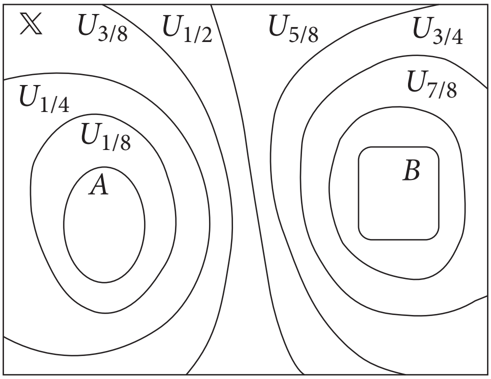

Figure 6.1

for any $\alpha \in (0,1]$, and this set is open. On the other hand, because $\mathbb{D}$ is dense in $[0,1]$, for $\beta \in [0,1)$

$$\begin{aligned}\{x : f(x) \le \beta\} &= \{x : \inf\{r \in \mathbb{D} : x \in U_r\} \le \beta\} \\ &= \{x : x \in U_r \ \forall\, r > \beta\} \\ &= \bigcap_{r>\beta} U_r \\ &= \bigcap_{r>\beta} \bar{U}_r \qquad (6.8)\end{aligned}$$

and this set is closed. Here, the final equality must hold to reconcile the following two facts: first that $U_r \subseteq \bar{U}_r$ and second that for all $r > \beta$ there exists (since $\mathbb{D}$ is dense) $s \in \mathbb{D}$ with $r > s > \beta$ and $\bar{U}_s \subseteq U_r$ by (6.5). This shows that for $0 \le \beta < \alpha \le 1$, the set

$$\{x : \beta < f(x) < \alpha\} = \{x : f(x) < \alpha\} \cap \{x : f(x) \le \beta\}^c \qquad (6.9)$$

is open, being the intersection of open sets. Since every open set of $[0,1]$ is a union of disjoint open intervals (see **1.34**) it follows that $f^{-1}(A)$ is open in $\mathbb{X}$ whenever A is open in $[0,1]$ and accordingly f is continuous. It is immediate that $f(A) = 0$ and $f(B) = 1$ as required and necessity is proved.

Sufficiency is simply a matter, given the existence of f with the indicated properties, of citing the two sets $f^{-1}\big([0,\frac{1}{2})\big)$ and $f^{-1}\big((\frac{1}{2},1]\big)$, that are open in $\mathbb{X}$, disjoint, and contain A and B respectively, so that $\mathbb{X}$ is normal. ∎

It is delightful the way this theorem conjures a continuous function out of thin air! It shows that the properties of real-valued functions provide a legitimate means of classifying the separation properties of the space.

In metric spaces, separating functions are obtained by a simple direct construction. If A and B are closed and disjoint subsets of a metric space $(\mathbb{S}, d)$, the normality property implies the existence of $\delta > 0$ such that $\inf_{x\in A, y\in B} d(x,y) \geq \delta$. The required function is

$$f(x) = \frac{d(x,A)}{d(x,A) + d(x,B)} \tag{6.10}$$

where $d(x,A) = \inf_{y\in A} d(x,y)$ and $d(x,B)$ is defined similarly. The continuity of f follows since $d(x,A)$ and $d(x,B)$ are continuous in x and the denominator in (6.10) is bounded below by δ. A similar construction was used in the proof of **5.25**.

The regularity property can be strengthened by requiring the existence of separating functions for closed sets C and points x. A topological space $\mathbb{X}$ is said to be *completely regular* if, for all closed $C \subseteq \mathbb{X}$ and points $x \notin C$, $\exists$ a continuous function $f : \mathbb{X} \mapsto [0,1]$ with $f(C) = 0$ and $f(x) = 1$. A completely regular T_1-space is called a *Tychonoff space* or $T_{3\frac{1}{2}}$-space. As the tongue-in-cheek terminology suggests, a T_4-space is $T_{3\frac{1}{2}}$ (this is immediate from Urysohn's lemma) and a $T_{3\frac{1}{2}}$-space is clearly T_3, although the reverse implications do not hold. Being T_4, metric spaces are always $T_{3\frac{1}{2}}$.

6.4 Weak Topologies

Now, go the other way; instead of using a topology to define a class of real functions, use a class of functions to define a topology. Let $\mathbb{X}$ be a space and $\mathbb{F}$ a class of functions $f : \mathbb{X} \mapsto \mathbb{Y}_f$ where the codomains $\mathbb{Y}_f$ are topological spaces. The *weak topology* induced by $\mathbb{F}$ on $\mathbb{X}$ is the weakest topology under which every $f \in \mathbb{F}$ is continuous. Recall, continuity means that $f^{-1}(A)$ is open in $\mathbb{X}$ whenever A is open in $\mathbb{Y}_f$. The weak topology can also be called the topology generated by the base sets $\mathcal{V}$ consisting of the inverse images of the open sets of the $\mathbb{Y}_f$ under $f \in \mathbb{F}$, together with the finite intersections of these sets. The inverse images themselves are called a *sub-base* for the topology, meaning that the sets of the topology can be generated from them by operations of union *and* finite intersection.

Enlarging $\mathbb{F}$ (potentially) increases the number of sets in this base to get a stronger topology, while contracting $\mathbb{F}$ likewise gives a weaker topology. With a given $\mathbb{F}$, any topology stronger than the weak topology contains a richer collection of open sets so the elements of $\mathbb{F}$ must retain their continuity in this case, but weakening the topology further must by definition force some $f \in \mathbb{F}$ to be discontinuous. That is to say, if $B = f^{-1}(A)$ for open $A \subseteq \mathbb{Y}_f$ does not belong to the topology, it is not an open set and so f is not continuous by definition.

The class of cases in which $\mathbb{Y}_f = \mathbb{R}$ for each f suggests using the concept of a weak topology to investigate the structure of a space. One way to represent the richness of a given topology τ on $\mathbb{X}$ is to ask whether τ contains, or is contained in, the weak topology generated by a particular collection of bounded real-valued functions on $\mathbb{X}$. For example, complete regularity is the minimal condition which makes the sort of construction in **6.13** feasible. According to the next result, this is sufficient to allow the topology to be completely characterized in terms of bounded, continuous real-valued functions on the space.

6.14 Theorem If a topological space $(\mathbb{X}, \tau)$ is completely regular, the topology τ is the weak topology induced by the set $\mathbb{F}$ of the separating functions.

Proof Let $\mathcal{V}$ denote the collection of inverse images of open sets of $\mathbb{R}$ under the functions of $\mathbb{F}$. Also, let T denote the weak topology induced by $\mathbb{F}$ such that the $\mathcal{V}$-sets together with their finite intersections form a base for T. To show that $T = \tau$, for any $x \in \mathbb{X}$ let $O \in \tau$ be an open set containing x. Then O^c is closed and by complete regularity there exists $f \in \mathbb{F}$ taking values in $[0,1]$ with $f(x) = 1$ and $f(O^c) = 0$. The set $(\frac{1}{2}, 1]$ is open in $[0,1]$; therefore $B = f^{-1}\big((\frac{1}{2}, 1]\big)$ is an open set containing x and disjoint with O^c, so that $B \subset O$. Since this holds for every such O, x has a base $\mathcal{V}_x$ consisting of inverse images of open sets under functions from $\mathbb{F}$. Since x is arbitrary the collection $\mathcal{V} = \{\mathcal{V}_x, x \in \mathbb{X}\}$ forms a base for τ. It follows that $\tau \subseteq T$.

On the other hand, T is by definition the weakest topology under which every $f \in \mathbb{F}$ is continuous. Since $f \in \mathbb{F}$ is a separating function and continuous under τ, it also follows that $T \subseteq \tau$. ∎

6.5 The Topology of Product Spaces

Let $\mathbb{X}$ and $\mathbb{Y}$ be a pair of topological spaces and consider the product space $\mathbb{X} \times \mathbb{Y}$. The plane $\mathbb{R} \times \mathbb{R}$ and subsets thereof are the natural examples for appreciating the properties of product spaces, although it is a useful exercise to think up more exotic cases. An example always given in textbooks of topology is $\mathbb{C} \times \mathbb{C}$ where $\mathbb{C}$ is the unit circle; this space has the topology of the torus (doughnut).

Let the ordered pair (x,y) be the generic element of $\mathbb{X} \times \mathbb{Y}$. The *coordinate projections* are the mappings $\pi_{\mathbb{X}} : \mathbb{X} \times \mathbb{Y} \mapsto \mathbb{X}$ and $\pi_{\mathbb{Y}} : \mathbb{X} \times \mathbb{Y} \mapsto \mathbb{Y}$, defined by

$$\pi_{\mathbb{X}}(x,y) = x \tag{6.11}$$

$$\pi_{\mathbb{Y}}(x,y) = y. \tag{6.12}$$

If $\mathbb{X}$ and $\mathbb{Y}$ are topological spaces, the coordinate projections can be used to generate a new topology on the product space. The *product topology* on $\mathbb{X}\times\mathbb{Y}$ is the weak topology induced by the coordinate projections.

The underlying idea here is very simple. If $A\subseteq\mathbb{X}$ and $B\subseteq\mathbb{Y}$ are open sets, the set $A\times B=(A\times\mathbb{Y})\cap(\mathbb{X}\times B)$, where $A\times\mathbb{Y}=\pi_{\mathbb{X}}^{-1}(A)$ and $\mathbb{X}\times B=\pi_{\mathbb{Y}}^{-1}(B)$, will be regarded as open in $\mathbb{X}\times\mathbb{Y}$ and is called an *open rectangle* of $\mathbb{X}\times\mathbb{Y}$. The product topology on $\mathbb{X}\times\mathbb{Y}$ is the topology having the open rectangles as a base. Equivalently, it is the weakest topology under which the coordinate projections are continuous.

If the factors are metric spaces $(\mathbb{X},d_{\mathbb{X}})$ and $(\mathbb{Y},d_{\mathbb{Y}})$, several metrics can be constructed to induce the product topology on $\mathbb{X}\times\mathbb{Y}$, including

$$\rho\big((x_1,y_1),(x_2,y_2)\big)=\max\big\{d_{\mathbb{X}}(x_1,x_2),d_{\mathbb{Y}}(y_1,y_2)\big\}\tag{6.13}$$

and

$$\rho'\big((x_1,y_1),(x_2,y_2)\big)=d_{\mathbb{X}}(x_1,x_2)+d_{\mathbb{Y}}(y_1,y_2).\tag{6.14}$$

An open sphere in the space $(\mathbb{X}\times\mathbb{Y},\rho)$ where ρ is the metric in (6.13) also happens to be an open rectangle, for

$$S_\rho\big((x,y),\delta\big)=S_{d_{\mathbb{X}}}(x,\delta)\times S_{d_{\mathbb{Y}}}(y,\delta).\tag{6.15}$$

Of course, this is not true for every metric.

Since either $\mathbb{X}$ or $\mathbb{Y}$ may be a product space, the generalization of these results from two to any finite number of factors is straightforward. The generic element of the space $\prod_{i=1}^n\mathbb{X}_i$ is the n-tuple $(x_1,\dots,x_n : x_i\in\mathbb{X}_i)$ and so on. However, to deal with infinite collections of factor spaces it is necessary to approach the product from a slightly different viewpoint. Let A denote an arbitrary index set and $\{\mathbb{X}_\alpha,\alpha\in A\}$ a collection of spaces indexed by A. The existence of the nonempty Cartesian product $\mathbb{X}^A=\prod_{\alpha\in A}\mathbb{X}_\alpha$ whose elements are collections drawn one from each member of the possibly uncountable collection $\{\mathbb{X}_\alpha\}$ is validated by the axiom of choice (see page 67). It is found most helpful to think of $\mathbb{X}^A$ as the collection of all the mappings $x : A\mapsto\bigcup_{\alpha\in A}\mathbb{X}_\alpha$ such that $x(\alpha)\in\mathbb{X}_\alpha$ for each $\alpha\in A$. This definition contains that given in §1.1 as a special case, but is fundamentally more general in character.

The coordinate projections are the mappings $\pi_\alpha : \mathbb{X}^A\mapsto\mathbb{X}_\alpha$ with

$$\pi_\alpha(x)=x(\alpha)\tag{6.16}$$

but can also be defined as the images under x of the points $\alpha \in A$. Thus, a *point* in the product space is a *mapping*, the one that generates the coordinate projections when it is evaluated at points of the domain A. In the case of a finite product, A can be the integers $1, \ldots, n$. In a countable case such as $A = \mathbb{N}$ or a set equipotent with $\mathbb{N}$, x should be called an infinite sequence, an element of $\mathbb{X}^\infty$ (say). A familiar uncountable example is provided by a class of real-valued functions $x : \mathbb{R} \mapsto \mathbb{R}$, so that $A = \mathbb{R}$. In this case, x associates each point $\alpha \in \mathbb{R}$ with a real number $x(\alpha)$ and defines an element of the product $\mathbb{R}^{\mathbb{R}}$.

The product topology is now generalized as follows. Let $\{\mathbb{X}_\alpha, \alpha \in A\}$ be an arbitrary collection of topological spaces. The *Tychonoff topology* (product topology) on the space $\mathbb{X}^A$ has as base the finite-dimensional open rectangles, sets of the form $\prod_{\alpha \in A} O_\alpha$, where the $O_\alpha \subseteq \mathbb{X}_\alpha$ are open sets and $O_\alpha = \mathbb{X}_\alpha$ except for at most a finite number of coordinates. These basic sets can be written as the intersections of finite collections of cylinders, say

$$B = \pi_{\alpha_1}^{-1}(O_{\alpha_1}) \cap \ldots \cap \pi_{\alpha_m}^{-1}(O_{\alpha_m}) \tag{6.17}$$

for indices $\alpha_1, \ldots, \alpha_m \in A$.

Let τ be a topology on $\mathbb{X}^A$ under which the coordinate projections are continuous. If O_α is open in $\mathbb{X}_\alpha$, $\pi_\alpha^{-1}(O_\alpha) \in \tau$ and hence τ contains the Tychonoff topology. Since this is true for any such τ, the Tychonoff topology can be characterized as the weak topology generated by the coordinate projections. The sets $\pi_\alpha^{-1}(O_\alpha)$ form the sub-base for the topology, whose *finite* intersections yield the base sets.

Something to keep in mind in these infinite product spaces is that if any of the sets $\mathbb{X}_\alpha$ are empty then $\mathbb{X}^A$ is empty. Some results are true only for nonempty spaces so for full rigour the stipulation that elements exist is desirable.

6.15 Example The space (C, d_U) examined in §5.6 is an uncountable product space having the Tychonoff topology; the uniform metric is the generalization of the maximum metric ρ of (6.13). Continuous functions are regarded as close to one another under d_U only if they are close at every point of the domain. The subsequent usefulness of this characterization of (C, d_U) stems mainly from the fact that the coordinate projections are known to be continuous. □

The two essential theorems on product spaces extend separability and compactness from the factor spaces to the product. The following theorem has a generalization to uncountable products that is not pursued here, since this is harder to prove and the countable case is sufficient for present purposes.

6.16 Theorem Finite or countable product spaces are separable under the product topology iff the factor spaces are separable.

Proof The proof for finite products is an easy implication of the countable case, hence consider $\mathbb{X}^\infty = \prod_{i=1}^\infty \mathbb{X}_i$. Let $D_i = \{d_{i1}, d_{i2}, \dots\} \subseteq \mathbb{X}_i$ be a countable dense set for each i and construct a set $D \subseteq \mathbb{X}^\infty$ by defining

$$F_m = \prod_{i=1}^{m} D_i \times \prod_{i=m+1}^{\infty} \{d_{i1}\} \tag{6.18}$$

for $m = 1, 2, \dots$ and then letting $D = \bigcup_{m=1}^\infty F_m$. F_m is equipotent with the set of m-tuples formed from the elements of the countable $D_1, \dots, D_m$ and is countable by induction from **1.4**. Hence D is countable, as a countable union of countable sets.

To show D is dense in $\mathbb{X}^\infty$, let $B = \prod_{i=1}^\infty O_i$ be a nonempty basic set, with O_i open in $\mathbb{X}_i$ and $O_i = \mathbb{X}_i$ except for a finite number of coordinates. Choose m such that $O_i = \mathbb{X}_i$ for $i > m$ and then

$$B \cap F_m = \prod_{i=1}^{m} (O_i \cap D_i) \times \prod_{i=m+1}^{\infty} \{d_{i1}\} \neq \emptyset \tag{6.19}$$

recalling that the dense property implies $O_i \cap D_i \neq \emptyset$, for $i = 1, \dots, m$. Since $B \cap F_m \subseteq B \cap D$, it follows that B contains a point of D; and since B is an arbitrary basic set, D is dense in $\mathbb{X}^\infty$, as required. ∎

One of the most powerful and important results in topology is *Tychonoff's theorem,* which states that arbitrary products of compact topological spaces are also compact under the product topology. It suffices here to prove the result for countable products of metric spaces and this case can be dealt with using a more elementary and familiar line of argument. It is not necessary to specify the metrics involved since the spaces need to be metric solely to exploit the equivalence of compactness and sequential compactness.

6.17 Theorem A finite or countable product of separable metric spaces $(\mathbb{X}_i, d_i)$ is compact under the product topology iff the factor spaces are compact.

Proof As before, the finite case follows easily from the countable case, so assume $\mathbb{X}^\infty = \prod_{i=1}^\infty \mathbb{X}_i$ where the $\mathbb{X}_i$ are separable spaces. In a metric space, which is first countable, compactness implies separability and is equivalent to sequential compactness by **6.8** and **6.7**. Since $\mathbb{X}_i$ is sequentially compact and first-countable, every sequence $\{x_{in}, n \in \mathbb{N}\}$ in $\mathbb{X}_i$ has a cluster point x_i in the space (**6.4**). Applying the diagonal argument of **2.36** there exists a single subsequence of integers, $\{n_k, k \in \mathbb{N}\}$, such that $x_{in_k} \to x_i$ for every i. Consider the subsequence $\{x_{n_k}, k \in \mathbb{N}\}$ in $\mathbb{X}^\infty$ where $x_{n_k} = (x_{1n_k}, x_{2n_k}, \dots)$. In the product topology, $x_{n_k} \to x = (x_1, x_2, \dots)$

iff $x_{in_k} \to x_i$ for every i, which proves that $\mathbb{X}^\infty$ is sequentially compact. $\mathbb{X}^\infty$ can be endowed with the metric $\rho_\infty = \sum_{i=1}^{\infty} d_i/2^i$ which induces the product topology. $\mathbb{X}^\infty$ is separable by **6.16** and sequential compactness is equivalent to compactness by **6.8** and **6.7**, as above. This proves sufficiency.

Necessity follows from **6.9**(ii), by continuity of the projections as before. ■

6.18 Example The space $\mathbb{R}^\infty$ (see **5.14**) is endowed with the Tychonoff topology, taking as the base sets of a point x the collection

$$V(x;k,\varepsilon) = \{y : |x_i - y_i| < \varepsilon, i - 1, \ldots, k\};\ k \in \mathbb{N},\ \varepsilon > 0. \qquad (6.20)$$

A point in $\mathbb{R}^\infty$ is close to x in this topology if many of its coordinates are close to those of x; another point is closer if either more coordinates are within ε of each other, or the same coordinates are closer than ε, or both. The metric d_∞ defined in (5.16) induces the topology of (6.20). If $\{x_n\}$ is a sequence in $\mathbb{X}$, $d_\infty(x_n, x) \to 0$ iff $\forall\ \varepsilon, k\ \exists\ N \geq 1$ such that $x_n \in V(x;k,\varepsilon)$ for all $n \geq N$. $\mathbb{R}^\infty$ is already known to be separable under d_∞ (**5.15**) but now this can be deduced as a purely topological property since $\mathbb{R}^\infty$ inherits separability from $\mathbb{R}$ by **6.16**. □

The infinite cube $[0,1]^\infty$ shares the topology (6.20) with $\mathbb{R}^\infty$ and is a compact space by **6.17.** This can be shown by assigning the Euclidean metric to the factor spaces $[0,1]$. The trick of metrizing a space to establish a topological property is frequently useful and is exploited in the next section.

6.6 Embedding and Metrization

Let $\mathbb{X}$ be a topological space and $\mathbb{F}$ a class of functions $f : \mathbb{X} \mapsto \mathbb{Y}_f$. The *evaluation map* $e : \mathbb{X} \mapsto \prod_{f \in \mathbb{F}} \mathbb{Y}_f$ is the mapping defined by

$$e(x)_f = f(x). \qquad (6.21)$$

The class $\mathbb{F}$ may be quite general, but if it were finite $e(x)$ might be thought of as the vector whose elements are the $f(x), f \in \mathbb{F}$. (6.21) could also be written $\pi_f \circ e = f$ where π_f is the coordinate projection. A minor complication arises because f need not be *onto* $\mathbb{Y}_f$ and $e(\mathbb{X}) \subset \prod_{f \in \mathbb{F}} \mathbb{Y}_f$ is possible. If A is a set of points in $\mathbb{Y}_f$, the inverse projection $\pi_f^{-1}(A)$ may contain points not in $e(\mathbb{X})$. Therefore the inverse of A under f might be expressed in terms of e as

$$f^{-1}(A) = (\pi_f \circ e)^{-1}(A) = e^{-1}\big(\pi_f^{-1}(A) \cap e(\mathbb{X})\big) \subseteq \mathbb{X}. \qquad (6.22)$$

The importance of this concept stems from the fact that under the right conditions the evaluation map *embeds* $\mathbb{X}$ in the product space generated by it. It would be homeomorphic to it in the case $e(\mathbb{X}) = \prod_{f \in \mathbb{F}} \mathbb{Y}_f$.

6.19 Theorem Suppose the class $\mathbb{F}$ separates points of $\mathbb{X}$, meaning that $f(x) \neq f(y)$ for some $f \in \mathbb{F}$ whenever x and y are distinct points of $\mathbb{X}$. If $\mathbb{X}$ is endowed with the weak topology induced by $\mathbb{F}$, the evaluation map defines an embedding of $\mathbb{X}$ into $\prod_f \mathbb{Y}_f$.

Proof It has to be shown that e is a 1–1 mapping from $\mathbb{X}$ onto a subset of $\prod_f \mathbb{Y}_f$, which is continuous with continuous inverse. Since $\mathbb{F}$ separates points of $\mathbb{X}$ e is 1–1, since $e(x) \neq e(y)$ whenever $f(x) \neq f(y)$ for some $f \in \mathbb{F}$. To show continuity of e note first that $f^{-1}(A)$ is open in $\mathbb{X}$ whenever A is open in $\mathbb{Y}_f$ under the weak topology, and sets of the form $\pi_f^{-1}(A)$ are likewise open in $\prod_f \mathbb{Y}_f$ with the product topology, since the projections are continuous. But $e^{-1}(\pi_f^{-1}(A)) = (\pi_f \circ e)^{-1}(A) = f^{-1}(A)$, so the inverse images under e of sets of the form $\pi_f^{-1}(A)$, $A \subseteq \mathbb{Y}_f$, are open. Since inverse images preserve unions and intersections (see **1.2**), the same property extends first to the base sets of $\prod_f \mathbb{Y}_f$ which are finite intersections of these inverse projections under the product topology and thence to all the open sets of $\prod_f \mathbb{Y}_f$. Therefore e is continuous.

e^{-1} is continuous if $e(B)$ is open in $e(\mathbb{X})$ whenever B is open in $\mathbb{X}$. Let B be a set of the form $f^{-1}(A)$, where A is open in $\mathbb{Y}_f$. Since $\mathbb{F}$ defines the topology on $\mathbb{X}$ this set is open and the finite intersections of such sets form a base for $\mathbb{X}$ by assumption. Since e is 1–1 and e^{-1} a mapping, it will suffice to verify that their images under e are open in $e(\mathbb{X})$. Noting that B is a set of the type shown in (6.22), $e(B) = \pi_f^{-1}(A) \cap e(\mathbb{X})$, but since $\pi_f^{-1}(A)$ is open, $e(B)$ is open in $e(\mathbb{X})$ as required. ∎

The following is what is for present purposes the most important case of *Urysohn's embedding theorem.*

6.20 Theorem A second-countable T_4-space $(\mathbb{X}, \tau)$ can be embedded in $[0,1]^\infty$. □

The proof requires the sufficiency part of the following lemma.

6.21 Lemma Let $x \in \mathbb{X}$ and let $O \subseteq \mathbb{X}$ be any open set containing x. Iff $\mathbb{X}$ is a regular space, there exists an open set U with $x \in \bar{U} \subset O$.

Proof Let $\mathbb{X}$ be regular. If O is open and $x \in O$, there exist disjoint open sets U and C such that $x \in U$ and $O^c \subset C$ and hence $C^c \subset O$. Since $U \subseteq C^c$ by the disjointness and C^c is closed, $\bar{U} \subseteq C^c \subset O$ and sufficiency is proved. To prove necessity, suppose $x \in U$ and $\bar{U} \subset O$. O^c is a closed set not containing x and $O^c \subset \bar{U}^c$ where U and $\bar{U}^c$ are disjoint open sets. Hence $\mathbb{X}$ is regular. ∎

Proof of 6.20 Let $\mathcal{V}$ be countable base for τ. Since the space is T_4 it is T_3 and hence regular. For any $x \in \mathbb{X}$ and $B \in \mathcal{V}$ containing x, there is by **6.21** a $U \in \tau$ such that $x \in \bar{U} \subset B$ and also, by definition of a base, $\exists\, A \in \mathcal{V}$ with $x \in A \subset U \subset \bar{U}$ and hence $x \in \bar{A} \subset \bar{U} \subset B$. ($\bar{A}$ is the smallest closed set containing A, note.) Let

$$\mathcal{A} = \{(A,B) : A \in \mathcal{V}, B \in \mathcal{V}; \bar{A} \subset B\} \tag{6.23}$$

denote the collection of all such pairs. Since $\mathcal{V}$ is countable, $\mathcal{A}$ is countable and so its elements can be labelled $(A,B)_i = (A_i, B_i)$, $i = 1, 2, \ldots$. Every $x \in \mathbb{X}$ lies in A_i for some $i \in \mathbb{N}$.

Since the space is normal, there exists by Urysohn's lemma a separating function $f_i : \mathbb{X} \mapsto [0,1]$ for each element of $\mathcal{A}$, such that $f_i(\bar{A}_i) = 1$ and $f_i(B_i^c) = 0$. For each $x \in \mathbb{X}$ and closed set C such that $x \notin C$, choose (A_i, B_i) such that $x \in \bar{A}_i \subset B_i \subset C^c$ and then $f_i(x) = 1$ and $f_i(C) = 0$. These separating functions form a countable class $\mathbb{F}$, a subset of $e(\mathbb{X})$. Since the space is T_1, C can be a point $\{y\}$ so that $\mathbb{F}$ separates points. And since the space is $T_{3\frac{1}{2}}$ and hence completely regular, τ is the weak topology induced by $\mathbb{F}$, by **6.14**. It follows by **6.19** that the evaluation map for $\mathbb{F}$ embeds $\mathbb{X}$ into $[0,1]^\infty$. ∎

Recall that $[0,1]^\infty$ endowed with the metric ρ_∞ defined in (5.20) is a compact metric space. It follows that $e(\mathbb{X})$, which is homeomorphic to $\mathbb{X}$ under the evaluation mapping by $\mathbb{F}$, is a totally bounded metric space. It further follows that $(\mathbb{X}, \tau)$ is metrizable, since among other possibilities, it can be endowed with the metric under which the distance between points x and y of $\mathbb{X}$ is taken to be $\rho_\infty\big(e(x), e(y)\big)$. This proves the Urysohn metrization theorem, **6.12**.

The topology induced by this metric on $[0,1]^\infty$ is the Tychonoff topology. A base for a point $p = (p_1, p_2, \ldots) \in [0,1]^\infty$ in this topology is provided by sets of the form

$$V(p; k, \varepsilon) = \{q \in [0,1]^\infty : |p_i - q_i| < \varepsilon,\ i = 1, \ldots, k\} \tag{6.24}$$

for some finite k and $0 < \varepsilon < 1$, which is the same as (6.20). The topology induced on $\mathbb{X}$ by the embedding is accordingly generated by the base sets

$$V(x; k, \varepsilon) = \{y \in \mathbb{X} : |f_i(x) - f_i(y)| < \varepsilon,\ i = 1, \ldots, k\} \tag{6.25}$$

which can be recognized as finite intersections of the inverse images, under functions from $\mathbb{F}$, of ε-neighbourhoods of $\mathbb{R}$; this is indeed the weak topology induced by $\mathbb{F}$. This further serves to remind us of the close link between product topologies and weak topologies.

Since metric spaces are T_4, separable metric spaces can be embedded in $[0,1]^\infty$ by **6.20**. In this case the motivation is not metrization, but usually compactification; that is, to show that separable spaces can be topologized as totally bounded spaces. Both metrization and compactification are techniques with important applications in the theory of weak convergence to be studied in Chapters 27 and 29. Although the following theorem is a straightforward corollary of **6.20**, the result is of sufficient interest to deserve its own proof; the main interest is to see how in metric spaces there always exists a ready-made collection of functions to define the weak topology.

6.22 Theorem A separable metric space $(\mathbb{S},d)$ is homeomorphic to a subset of $[0,1]^\infty$.

Proof Let $d_0 = d/(1+d)$, which satisfies $0 \le d_0 \le 1$ and is equivalent to d so that $(\mathbb{S},d_0)$ is homeomorphic to $(\mathbb{S},d)$. By separability there exists a countable set of points $\{z_i, i \in \mathbb{N}\}$ which is dense in $\mathbb{S}$. Let a countable family of functions be defined by $f_i(x) = d_0(x,z_i)$, $i = 1,2,\dots$ and define an evaluation map $h : \mathbb{S} \mapsto [0,1]^\infty$ by

$$h(x) = \big(d_0(x,z_1), d_0(x,z_2), \dots\big). \tag{6.26}$$

h is an embedding in $([0,1]^\infty, \rho_\infty)$ where $\rho_\infty(h,g) = \sum_{k=1}^\infty |h_k - g_k|/2^k$. If $\{x_n\}$ is a sequence in $\mathbb{S}$ converging to x, then for each k, $d_0(x_n,z_k) \to d_0(x,z_k)$. Accordingly, $\forall\ k,\varepsilon\ \exists\ N \ge 1$ such that $x_n \in V(x;k,\varepsilon)$ for all $n \ge N$, $\rho_\infty(h(x_n),h(x)) \to 0$ and h is continuous at x. On the other hand, if $x_n \nrightarrow x$, there exists $\varepsilon > 0$ such that $\forall\ N \ge 1$, $d_0(x_n,x) \ge \varepsilon$ for some $n \ge N$. Since $\{z_k\}$ is dense in $\mathbb{S}$, there is a k for which $d_0(x_n,z_k) \ge \frac{3}{4}\varepsilon$ and $d_0(x,z_k) < \frac{1}{4}\varepsilon$, so that $|d_0(x_n,z_k) - d_0(x,z_k)| \ge \frac{1}{2}\varepsilon$ and hence

$$\rho_\infty\big(h(x_n),h(x)\big) \ge \varepsilon/2^{k+1}. \tag{6.27}$$

Since this holds for some $n \ge N$ for every $N \ge 1$, it holds for infinitely many n and $h(x_n) \nrightarrow h(x)$, showing that $h(x_n) \to h(x)$ if and only if $x_n \to x$. This is the property of a 1–1 continuous function with continuous inverse. ∎

But note too the alternative approach of transforming the distance functions into separating functions as in (6.10) and applying **6.20**.

PART II

PROBABILITY

7
Probability Spaces

7.1 Probability Measures

A *random experiment* is an action or observation whose outcome is uncertain in advance of its occurrence. Tosses of a coin, spins of a roulette wheel and observations of the price of a stock are familiar examples. A *probability space*, the triple $(\Omega, \mathcal{F}, P)$, is a measure space that here plays the role of a mathematical model of a random experiment. Ω is known by convention as the *sample space* and is the set of all the possible outcomes of the experiment, called the *random elements*, individually denoted ω. The collection $\mathcal{F}$ of *random events* is a σ-field of subsets of Ω, the event $A \in \mathcal{F}$ being said to have occurred if the outcome of the experiment is an element of A.

The measure P assigned to the elements of $\mathcal{F}$, with $P(A)$ being the *probability* of random event A, is formally defined as follows.

7.1 Definition A probability measure (p.m.) on a measurable space $(\Omega, \mathcal{F})$ is a set function $P : \mathcal{F} \mapsto [0, 1]$ satisfying the axioms of probability:

(a) $P(A) \geq 0$, for all $A \in \mathcal{F}$

(b) $P(\Omega) = 1$

(c) Countable additivity: for a disjoint collection $\{A_j \in \mathcal{F}, j \in \mathbb{N}\}$,

$$P\Bigl(\bigcup_j A_j\Bigr) = \sum_j P(A_j). \quad \square \tag{7.1}$$

It is only by property (b) that P is distinguished from the general definition of a measure in **3.1**.

Under the frequentist interpretation of probability, $P(A)$ is the limiting case of the proportion of a long run of repeated independent experiments in which the outcome is in A. Alternatively, probability may be viewed as a subjective notion with $P(A)$ said to represent an observer's degree of belief that A will occur in the next experiment. However, for present purposes the interpretation given to the probabilities has no relevance. The theory stands or falls by its mathematical consistency alone, although it is then up to the practitioner to decide whether the results accord with intuition and are useful in the analysis of real-world problems.

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0007

Additional properties of P follow from the axioms.

7.2 Theorem If A, B, and $\{A_j, j \in \mathbb{N}\}$ are arbitrary $\mathcal{F}$-sets, then

(i) $P(A) \le 1$
(ii) $P(A^c) = 1 - P(A)$
(iii) $P(\varnothing) = 0$
(iv) $A \subseteq B \Rightarrow P(A) \le P(B)$ (monotonicity)
(v) $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
(vi) $P\left(\bigcup_j A_j\right) \le \sum_j P(A_j)$ (countable subadditivity)
(vii) $A_j \uparrow A$ or $A_j \downarrow A \Rightarrow P(A_j) \to P(A)$ (continuity). □

Most of these are properties of measures in general. The complementation property (ii) is special to P, although an analogous condition holds for any finite measure, with $P(\Omega)$ replacing 1 in the formula. (iii) confirms P is a measure, on the definition.

Proof of 7.2 Applying **7.1**(a), (b), and (c)

$$P(A) + P(A^c) = P(A \cup A^c) = P(\Omega) = 1 \tag{7.2}$$

from which follow (i) and (ii) and also (iii) on setting $A = \Omega$. (iv)–(vi) follow by **3.5** and (vii) by **3.7**. ∎

To create a probability space, probabilities are assigned to a basic class of events $\mathcal{C}$, according to a hypothesis about the mechanisms underlying the random outcome. For example, in coin- or die-tossing experiments the usual hypothesis is of a fair coin or die and hence of equally likely outcomes. Then, provided $\mathcal{C}$ is rich enough to be a determining class for the space, $(\Omega, \mathcal{F}, P)$ exists by **3.15** (extension theorem) where $\mathcal{F} = \sigma(\mathcal{C})$.

7.3 Example Let $\mathcal{B}_{[0,1]} = \{B \cap [0,1], B \in \mathcal{B}\}$ where $\mathcal{B}$ is the Borel field of $\mathbb{R}$. Then $\left([0,1], \mathcal{B}_{[0,1]}, m\right)$ is a probability space where m is Lebesgue measure, since $m([0,1]) = 1$. The random elements of this space are real numbers between 0 and 1 and a drawing from the distribution is called a random variable. It is said to be distributed uniformly on the unit interval. The inclusion or exclusion of the endpoints is optional, remembering that $m([0,1]) = m((0,1)) = 1$. □

The *atoms* of a p.m. are the outcomes (singleton sets of Ω) that have positive probability. The following is true for finite measures generally but has special importance in the theory of distributions.

7.4 Theorem The atoms of a p.m. are at most countable.

Proof Let ω_1 be an atom satisfying $P(\{\omega_1\}) \geq P(\{\omega\})$ for all $\omega \in \Omega$, let ω_2 satisfy $P(\{\omega_2\}) \geq P(\{\omega\})$ for all $\omega \in \Omega - \{\omega_1\}$, and so forth, to generate a sequence with

$$P(\{\omega_1\}) \geq P(\{\omega_2\}) \geq P(\{\omega_3\}) \geq \ldots. \tag{7.3}$$

The partial sums $\sum_{i=1}^{n} P(\{\omega_1\})$ form a monotone sequence that cannot exceed $P(\Omega) = 1$ and therefore converges by **2.1**, implying by **2.15** that $\lim_{n\to\infty} P(\{\omega_n\}) = 0$. All points with positive probability are therefore in the countable set $\{\omega_i, i \in \mathbb{N}\}$. ∎

Suppose a random experiment represented by the space $(\Omega, \mathcal{F}, P)$ is modified so as to confine the possible outcomes to a subset of the sample space, say $\Lambda \subset \Omega$. Imagine switching from playing roulette with a wheel having a zero slot to one without, for example. The restricted probability space is derived as follows. Let $\mathcal{F}_A$ denote the collection $\{E \cap \Lambda, E \in \mathcal{F}\}$. $\mathcal{F}_\Lambda$ is a σ-field (compare **1.23**) and is called the *trace* of $\mathcal{F}$ on Λ. Defining $P_\Lambda(E) = P(E)/P(\Lambda)$ for $E \in \mathcal{F}_\Lambda$, P_Λ can be verified to be a p.m. The triple $(\Lambda, \mathcal{F}_\Lambda, P_\Lambda)$ is called the trace of $(\Omega, \mathcal{F}, P)$ on Λ. This is similar to the restriction of a measure space to a subspace, except that the measure is renormalized so that it remains a p.m.

In everyday language events are said to be 'impossible' or 'certain'. If such events are nonetheless elements of $\mathcal{F}$ and hence technically random, the idea that they will occur or not occur with 'certainty' is expressed by assigning them probabilities of one or zero. The usage of the term 'certain' here is deliberately loose, as the quotation marks suggest. To say an event cannot occur because it has probability zero is different from saying it cannot occur because the outcomes it contains are not elements of Ω. Similarly, to say an event has probability 1 is different from saying it is the event Ω. In technical discussion there is a nice distinction between *sure*, which means the latter, and *almost sure*, which means the former. An event E is said to occur *almost surely* (a.s.), or equivalently, *with probability one* (w.p.1) if $M = \Omega - E$ has probability measure zero. This terminology is synonymous with *almost everywhere* (a.e.) in the measure-theoretic context. When there is ambiguity about the p.m. being considered, the notation 'a.s.[P]' may be used.

7.2 Conditional Probability

A central issue of probability is the treatment of relationships. When random experiments generate a multi-dimensional outcome (e.g. a poker deal generates several different hands) questions always arise about relationships between the different aspects of the experiment. The natural way to pose such questions is: "if I observe only one facet of the outcome, does this change the probabilities I should assign to what is unobserved?" (Skilled poker players know the answer to this question, of course.)

The idea underlying conditional probability is that some but not all aspects of a random experiment have been observed. By eliminating those outcomes that are incompatible with partial knowledge, it is necessary to consider only a part of the sample space. In $(\Omega, \mathcal{F}, P)$ let partial information about the outcome take the form 'the event A has occurred', where $A \in \mathcal{F}$. How should this knowledge change the probabilities attached to other events? Since the outcomes in A^c are ruled out, the sample space is reduced from Ω to A. To generate probabilities on this restricted space, define the *conditional probability* of an event B as $P(B|A) = P(A \cap B)/P(A)$ for $A, B \in \mathcal{F}$, $P(A) > 0$. $P(\cdot|A)$ satisfies the probability axioms as long as P does and $P(A) > 0$. In particular, $P(A|A) = 1$ and $P(B^c|A) = 1 - P(B|A)$, since $B \cap A$ and $B^c \cap A$ are disjoint and their union is A. The space $(A, \mathcal{F}_A, P_A)$, the trace of the set A on $(\Omega, \mathcal{F}, P)$, models the random experiment from the point of view of an observer who knows that $\omega \in A$. Events A and B are said to be *dependent* when $P(B|A) \neq P(B)$.

In certain respects the conditioning concept seems a little improper. A context in which the components of the random outcome are revealed sequentially to an observer might appear relevant only to a subjective interpretation of probability and lead a sceptical reader to call the neutrality of the mathematical theory into question. A random event is after all random and has no business defining a probability space. In practice, the applications of conditional probability in limit theory are usually quite remote from any considerations of subjectivity, but there is a serious point here, which is the difficulty of constructing a rigorous theory after moving beyond the restricted goal of predicting random outcomes a priori.

The way to overcome improprieties of this kind and obtain a much more powerful theory into the bargain is to condition on a *class* of events, a sub-σ-field of $\mathcal{F}$. Given an event $B \in \mathcal{F}$, let the set function

$$P(B|\mathcal{G}) : \mathcal{G} \mapsto [0,1]$$

represent the *contingent* probability to be assigned to B after drawing an event A from $\mathcal{G} \subseteq \mathcal{F}$. Think of $\mathcal{G}$ as an information set in the sense that for each $A \in \mathcal{G}$ an observer knows whether or not the outcome is in A. Since the elements of the domain are random events, $P(B|\mathcal{G})$ is itself a random outcome (a random variable, in the terminology of Chapter 8) derived from the restricted probability space $(\Omega, \mathcal{G}, P)$. This space is a model of the action of an observer possessing information $\mathcal{G}$ who assigns the conditional probability $P(B|A)$ to B when he observes the occurrence of A viewed from the standpoint of *another* observer who has no prior information. $\mathcal{G}$ is a σ-field, because if an outcome is in A it is not in A^c and if it is known whether or not it falls in A_j for each $j = 1, 2, 3, \ldots$ it is known whether or not it is in $\bigcup_j A_j$. The more sets there are in $\mathcal{G}$ the larger the volume of information, all the way from the trivial set $\mathcal{T} = (\Omega, \varnothing)$ (complete ignorance, with $P(B|\mathcal{T}) = P(B)$ a.s.) to the set $\mathcal{F}$ itself, which corresponds to almost sure knowledge of the outcome. In the latter case, $P(B|\mathcal{F}) = 1$ a.s. if $\omega \in B$ and 0 otherwise. If you know whether or not $\omega \in A$ for every $A \in \mathcal{F}$ you effectively know ω.

Incidentally, if there is a subset $N \subset \Omega$ such that either N or N^c is contained in every $\mathcal{F}$-set, the elements of N cease to be distinguishable as different outcomes. An equivalent model of the random experiment is obtained by redefining Ω to have N itself as an element, replacing its individual members.

7.3 Independence

A pair of events $A, B \in \mathcal{F}$ is said to be independent if

$$P(A \cap B) = P(A)P(B) \tag{7.4}$$

or, equivalently, given $P(A) > 0$, if $P(B|A) = P(B)$. If in a collection of events $\mathcal{C}$ (7.4) holds for every pair of distinct sets A and B from the collection, $\mathcal{C}$ is said to be *pairwise* independent. In addition, $\mathcal{C}$ is said to be *totally* independent if for every subset $\mathcal{I} \subseteq \mathcal{C}$ containing two or more events,

$$P\Bigl(\bigcap_{A \in \mathcal{I}} A\Bigr) = \prod_{A \in \mathcal{I}} P(A). \tag{7.5}$$

This is a stronger condition than pairwise independence. Suppose $\mathcal{C}$ consists of sets A, B, and C. Knowing that B has occurred may not influence the probability attached to A and similarly for C; but the joint occurrence of B and C may nonetheless imply something about A. Pairwise independence implies that $P(A \cap B) = P(A)P(B)$, $P(A \cap C) = P(A)P(C)$, and $P(B \cap C) = P(B)P(C)$, but total independence would also require $P(A \cap B \cap C) = P(A)P(B)P(C)$.

Here are some useful results concerning independence.

7.5 Theorem Let $\mathcal{C}$ be a totally independent collection satisfying (7.5) for each subset $\mathcal{I} \subseteq \mathcal{C}$. The collection $\mathcal{C}'$ containing A and A^c for each $A \in \mathcal{C}$ is totally independent.

Proof It is sufficient to prove that the independence of A and B implies that of A^c and B, for B can denote any arbitrary intersection of sets from the collection and (7.5) will be satisfied, for either A or A^c. This is certainly true, since if $P(A \cap B) = P(A)P(B)$, then

$$\begin{aligned} P(A^c \cap B) &= P(A^c \cap B) + P(A \cap B) - P(A)P(B) \\ &= P(B) - P(A)P(B) = P(A^c)P(B). \quad \blacksquare \end{aligned} \tag{7.6}$$

7.6 Theorem Let $\mathcal{C}$ be a totally independent collection and let $\mathcal{B}$ be a countable disjoint collection. If the collections consisting of $B_j \in \mathcal{B}$ for $j = 1, 2, \ldots$ and the sets of $\mathcal{C}$ are totally independent for each j, then the collection consisting of $B = \bigcup_j B_j$ and $\mathcal{C}$ is also totally independent.

Proof Let $\mathcal{I}$ be any subset of $\mathcal{C}$. Using the disjointness of the sets of B and countable additivity,

$$\begin{aligned} P\Big(B \cap \bigcap_{A \in \mathcal{I}} A\Big) &= P\Big(\bigcup_j B_j \cap \bigcap_{A \in \mathcal{I}} A\Big) = \sum_j P\Big(B_j \cap \bigcap_{A \in \mathcal{I}} A\Big) \\ &= \sum_j P(B_j) P\Big(\bigcap_{A \in \mathcal{I}} A\Big) = P(B) \prod_{A \in \mathcal{I}} P(A). \quad \blacksquare \end{aligned} \tag{7.7}$$

Collections of events $\mathcal{C} \subset \mathcal{F}$ and $\mathcal{D} \subset \mathcal{F}$ are said to have the *independence property* if $P(C \cap D) = P(C)P(D)$ for all pairs $C \in \mathcal{C}$ and $D \in \mathcal{D}$. Recall that a π-system is a collection that is closed under the operation of pairwise intersection.

7.7 Theorem If $\mathcal{C}$ and $\mathcal{D}$ are π-systems having the independence property, the generated σ-fields $\sigma(\mathcal{C})$ and $\sigma(\mathcal{D})$ also have the independence property.

Proof Without loss of generality augment both collections with Ω. Given $D \in \mathcal{D}$, let $\mathcal{A}$ denote the collection of sets A having the property $P(A \cap D) = P(A)P(D)$ for each $A \in \mathcal{A}$. If $A \in \mathcal{A}$ then $A^c \in \mathcal{A}$, by (7.6). Also, if $A_1, A_2, \ldots$ is a countable disjoint sequence of $\mathcal{A}$-sets then

$$P\Big(\bigcup_j A_j \cap D\Big) = P\Big(\bigcup_j A_j\Big) P(D) \tag{7.8}$$

follows as a case of **7.6**. Therefore, $\mathcal{A}$ is a λ-system. It follows by **1.29** that if $\mathcal{C} \subseteq \mathcal{A}$, $\sigma(C) \subseteq \mathcal{A}$. Since $D \in \mathcal{D}$ was arbitrary, it follows that $\sigma(\mathcal{C})$ and $\mathcal{D}$ have the independence property. Repeat the argument with the roles of $\mathcal{C}$ and $\mathcal{D}$ reversed to complete the proof. $\blacksquare$

Iterating this result shows that if π-systems $\mathcal{C}_1, \ldots, \mathcal{C}_n$ have the independence property, in the sense that any $C_j \in \mathcal{C}_j$ for $j = 1, \ldots, n$ are totally independent, the property extends to the generated σ-fields $\sigma(\mathcal{C}_1), \ldots, \sigma(\mathcal{C}_n)$.

7.4 Product Spaces

Questions of dependence and independence arise when multiple random experiments run in parallel and product spaces play a natural role in the analysis of these issues. Let $(\Omega \times \Xi, \mathcal{F} \otimes \mathcal{G}, P)$ be a probability space where $\mathcal{F} \otimes \mathcal{G}$ is the σ-field generated by the measurable rectangles of $\Omega \times \Xi$ and $P(\Omega \times \Xi) = 1$. The random outcome is a pair (ω, ξ). This is no more than a case of the general theory of §7.1 (where the nature of ω is unspecified) except that it becomes possible to ask questions about the part of the outcome represented by ω or ξ alone. $P_\Omega(F) = P(F \times \Xi)$ for $F \in \mathcal{F}$ and $P_\Xi(G) = P(\Omega \times G)$ for $G \in \mathcal{G}$ are called the *marginal probabilities*. $(\Omega, \mathcal{F}, P_\Omega)$ and $(\Xi, \mathcal{G}, P_\Xi)$ are probability spaces representing an incompletely observed random experiment, with ω or ξ, respectively, being the only things observed in an experiment generating (ω, ξ).

On the other hand, suppose ξ is observed and subsequently the 'experiment' of observing ω performed. Knowing ξ means that for each event $\Omega \times G$ it is known whether or not it contains (ω, ξ). The conditional probabilities generated by this two-stage experiment can be written by a slight abuse of notation as $P(F|\mathcal{G})$, although strictly speaking the relevant events are the cylinders $F \times \Xi$ and the elements of the conditioning σ-field are $\Omega \times G$ for $G \in \mathcal{G}$. It would perhaps be completely proper to write something like $P(F \times \Xi | \Omega \times \mathcal{G})$.

In this context, product measure assumes a special role as the model of independence. In $(\Omega \times \Xi, \mathcal{F} \otimes \mathcal{G}, P)$, the coordinate spaces are said to be independent when

$$P(F \times G) = P_\Omega(F) P_\Xi(G) \tag{7.9}$$

for each $F \in \mathcal{F}$ and $G \in \mathcal{G}$. Unity of the notation is preserved since $F \times G = (F \times \Xi) \cap (\Omega \times G)$. This might also be written $P(F \times \Xi | \Omega \times G) = P_\Omega(F)$, or with a further slight abuse of notation $P(F|G) = P_\Omega(F)$, for any pair $F \in \mathcal{F}$ and $G \in \mathcal{G}$. Independence means that knowing ξ does not affect the probabilities assigned to sets of $\mathcal{F}$. Since the measurable rectangles are a determining class for the space, the p.m. P is entirely determined by the marginal measures.

8
Random Variables

8.1 Measures on the Line

Let $(\Omega, \mathcal{F}, P)$ be a probability space. A real *random variable* (r.v.) is an $\mathcal{F}/\mathcal{B}$-measurable function $X : \Omega \mapsto \mathbb{R}$. That is to say, $X(\omega)$ induces an inverse mapping from $\mathcal{B}$ to $\mathcal{F}$ such that $X^{-1}(B) \in \mathcal{F}$ for every $B \in \mathcal{B}$, where $\mathcal{B}$ is the linear Borel field. The term '$\mathcal{F}$-measurable r.v.' may be used when the role of $\mathcal{B}$ is understood. The symbol μ will generally be used to denote a p.m. on the line, reserving P for the p.m. on the underlying space. 'Real' here implies that X is a point on the line. Complex-valued random variables also arise and this case is dealt with in §11.2.

Random variables therefore live in the space $(\mathbb{R}, \mathcal{B}, \mu)$ where μ is the derived measure such that $\mu(B) = P\big(X^{-1}(B)\big) = P(X \in B)$. The term *distribution* is synonymous with measure in this context. The properties of r.v.s are special cases of the results in Chapter 3 and in particular the contents of §3.6 should be reviewed in conjunction with this chapter. If $g : \mathbb{R} \mapsto \mathbb{R}$ is a Borel function then $g \circ X(\omega) = g(X(\omega))$ is also a r.v., having derived p.m. μg^{-1} according to **3.28**.

If there is a set $S \in \mathcal{B}$ having the property $\mu(S) = 1$, the trace of $(\mathbb{R}, \mathcal{B}, \mu)$ on S is equivalent to the original space in the sense that the same measure is assigned to B and to $B \cap S$, for each $B \in \mathcal{B}$. Which space to work with is basically a matter of technical convenience. If X is a r.v. it may be more satisfactory to say that the Borel function X^2 is a r.v. distributed on $\mathbb{R}^+$ than that it is distributed on $\mathbb{R}$ but supported on $\mathbb{R}^+$, in the sense of taking values in $\mathbb{R}^+$ almost surely. One could substitute for $(\mathbb{R}, \mathcal{B}, \mu)$ the extended space $(\bar{\mathbb{R}}, \bar{\mathcal{B}}, \mu)$ (see **1.22**), but note that assigning a positive probability to infinity does not lead to meaningful results. Random variables must be finite with probability 1. Thus $(\mathbb{R}, \mathcal{B}, \mu)$, the trace of $(\bar{\mathbb{R}}, \bar{\mathcal{B}}, \mu)$ on $\mathbb{R}$, is equivalent to it for nearly all purposes. However, while it is always finite a.s., a r.v. is not necessarily *bounded* a.s.; there may exist no constant B such that $|X(\omega)| \leq B$ for all $\omega \in C$ with $P(\Omega - C) = 0$. The *essential supremum* of X is

$$\operatorname{ess\,sup} X = \inf\big\{x : P(|X| > x) = 0\big\} \tag{8.1}$$

and this may be either a finite number or $+\infty$.

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0008

8.2 Distribution Functions

The *cumulative distribution function* (c.d.f.) of X is the function $F : \bar{\mathbb{R}} \mapsto [0,1]$ where

$$F(x) = \mu\big((-\infty, x]\big) = P(X \le x),\ x \in \bar{\mathbb{R}}. \tag{8.2}$$

The domain is $\bar{\mathbb{R}}$ since it is natural to assign the values 0 and 1 to $F(-\infty)$ and $F(+\infty)$ respectively. No other values are possible so there is no contradiction in confining attention to just the points of $\mathbb{R}$. To specify a distribution for X it is sufficient to assign a functional form for F; μ and F are equivalent representations of the distribution, each useful for different purposes. To represent $\mu(A)$ in terms of F for a set A much more complicated than an interval would be cumbersome, but on the other hand, the graph of F is an appealing way to display the characteristics of the distribution.

To see how probabilities are assigned to sets using F, start with the half-open interval $(x,y]$ for $x < y$. This is the intersection of the half-lines $(-\infty, y]$ and $(-\infty, x]^c = (x, +\infty)$. Let $A = (-\infty, x]$ and $B = (-\infty, y]$, so that $\mu(A) = F(x)$ and $\mu(B) = F(y)$. Then, since A and B^c are disjoint,

$$\begin{aligned}\mu\big((x,y]\big) &= \mu(A^c \cap B) = 1 - \mu(A \cup B^c)\\ &= 1 - \big(\mu(A) + 1 - \mu(B)\big) = \mu(B) - \mu(A) = F(y) - F(x).\end{aligned} \tag{8.3}$$

The half-open intervals form a semi-ring (see **1.18**) and from the results of §3.2 the measure extends uniquely to the sets of $\mathcal{B}$.

As an example of the extension, try calculating $\mu(\{x\}) = P(X = x)$ for $x \in \mathbb{R}$ (compare **3.22**). Putting $x = y$ in (8.3) does not yield this result since $A \cap A^c = \varnothing$, not $\{x\}$. The singleton $\{x\}$ is the intersection of $(-\infty, x]$ and $[x, +\infty) = (-\infty, x)^c$, but then there is no obvious way to find the probability for the open interval $(-\infty, x) = (-\infty, x] - \{x\}$. The solution to the problem is to consider the monotone sequence of half-lines $(-\infty, x - 1/n]$ for $n \in \mathbb{N}$. Since $(x - 1/n, x] = (-\infty, x - 1/n]^c \cap (-\infty, x]$, $\mu((x - 1/n, x]) = F(x) - F(x - 1/n)$, according to (8.3). Since $\{x\} = \bigcap_{n=1}^{\infty}(x - 1/n, x]$, $\{x\} \in \mathcal{B}$ and $\mu(\{x\}) = F(x) - F(x-)$ where $F(x-)$ is the left limit of F at x. $F(x)$ exceeds $F(x-)$ (i.e., F jumps) at the points x where $\mu(\{x\}) > 0$, known as the atoms of the distribution. By the same kind of reasoning, $\mu((x,y)) = F(y-) - F(x)$ and $\mu([x,y)) = F(y-) - F(x-)$. Measures of open intervals are the same as those of closed intervals unless the endpoints are atoms of the distribution.

Certain characteristics imposed on the c.d.f. by its definition in terms of a measure were implicit in the above conclusions. The next three theorems establish these properties.

8.1 Theorem F is non-negative and non-decreasing with $F(-\infty) = 0$ and $F(+\infty) = 1$ and is increasing at $x \in \mathbb{R}$ iff every open neighbourhood $S(x, \varepsilon)$ of x has positive measure.

Proof These are all direct consequences of the definition. Non-negativity is from (8.2) and monotonicity from **7.2**(iv). F is increasing at x if $F(x + \varepsilon) > F(x - \varepsilon)$ for each $\varepsilon > 0$. To show the asserted sufficiency, for each such ε,

$$F(x + \varepsilon) - F(x - \varepsilon) \geq F\big((x + \varepsilon)-\big) - F(x - \varepsilon) = \mu\big(S(x, \varepsilon)\big). \tag{8.4}$$

For the necessity, suppose $\mu(S(x, \varepsilon)) = 0$ and note that, by monotonicity of F,

$$\mu\big(S(x, \varepsilon)\big) = F\big((x + \varepsilon)-\big) - F(x - \varepsilon) \geq F(x + \varepsilon/2) - F(x - \varepsilon/2). \quad \blacksquare \tag{8.5}$$

The support of μ is the collection of points on which F increases. Its complement in $\mathbb{R}$, the largest set of zero measure, consists of points lying in open neighbourhoods of zero measure, on which F does not increase and hence must be open. The support of μ is accordingly a closed set. Singletons are either atoms of the distribution (jump points of F) or have zero measure, whether or not in the support.

8.2 Theorem F is right-continuous everywhere.

Proof For $x \in \mathbb{R}$ and $n \geq 1$, additivity of the p.m. implies

$$\mu\big((-\infty, x + 1/n]\big) = \mu\big((-\infty, x]\big) + \mu\big((x, x + 1/n]\big). \tag{8.6}$$

The half-open intervals $(x, x + 1/n]$ do not contain x and converge to $\emptyset$ as $n \to \infty$. Hence, $\mu((x, x + 1/n]) \to 0$ by continuity of the measure. It follows that for any $\varepsilon > 0$ there exists N_ε such that for $n \geq N_\varepsilon$,

$$F(x + 1/n) - F(x) < \varepsilon \tag{8.7}$$

and so $F(x+) = F(x)$, proving the theorem since x was arbitrary. $\blacksquare$

If $F(x)$ had been defined as $\mu((-\infty, x))$, similar arguments would show that it was left-continuous in that case.

8.3 Theorem F has the decomposition

$$F(x) = F'(x) + F''(x) \tag{8.8}$$

where $F'(x)$ is a right-continuous step function with at most a countable number of jumps and $F''(x)$ is everywhere continuous.

Proof By **7.4** the jump points of F are at most countable. Letting $\{x_1, x_2, \ldots\}$ denote these points

$$F'(x) = \sum_{x_i \leq x} \big(F(x_i) - F(x_i-)\big) \tag{8.9}$$

is a step function with jumps at the points x_i and $F''(x) = F(x) - F'(x)$ has $F(x_i-) = F(x_i)$ at each x_i and is continuous everywhere. ∎

Figure 8.1 illustrates the decomposition.

This is not the only decomposition of F. The Lebesgue decomposition of μ with respect to Lebesgue measure on $\mathbb{R}$ (see **4.32**) is $\mu = \mu_1 + \mu_2$ where μ_1 is singular with respect to m (is positive only on a set of Lebesgue measure 0) and μ_2 is absolutely continuous with respect to Lebesgue measure. Recall that $\mu_2(A) = \int_A f(x)\mathrm{d}x$ for $A \in \mathcal{B}$, where f is the associated Radon–Nikodym derivative (density function). If F is decomposed in the same way such that $F_i(x) = \mu_i((-\infty, x])$ for $i = 1$ and 2 then $F_2(x) = \int_{-\infty}^{x} f(\xi)\mathrm{d}\xi$ implying that $f(x) = \mathrm{d}F_2/\mathrm{d}\xi|_{\xi=x}$. This must hold for almost all x (Lebesgue measure). F_2 is called an *absolutely continuous* function, meaning that it is differentiable almost everywhere on its domain.

$F_1 \geq F'$ since F_1 may increase on a set of Lebesgue measure 0. Such sets can be uncountable and hence larger than the set of atoms. It is customary to summarize these relations by decomposing F'' into two additive components, the absolutely continuous part F_2 and a component $F_3 = F'' - F_2 = F_1 - F'$. F_3 is continuous and also singular, meaning that it is constant except on a set of zero Lebesgue measure.

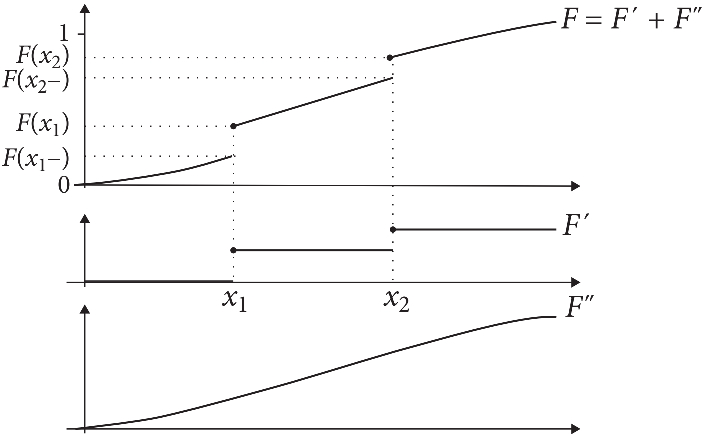

Figure 8.1

In most cases F_3 can be neglected since virtually all those distributions whose functional form is known are either defined by a density function or consist entirely of atoms. There can be few plausible real-world distributions where $F = F_3$. However, the following is a classic instance.

8.4 Example Consider the sets C_n for $n = 1, 2, 3, \ldots$ defined in **3.11**. For any n, define the distribution that is uniform on the support C_n such that the c.d.f. $F^{(n)}$ (say) has $F^{(n)}(0) = 0$ and $F^{(n)}(1) = 1$, is constant over any interval of the line not included in C_n, and increases linearly otherwise. Hence, the distribution is continuous. The case $F = \lim_{n\to\infty} F^{(n)}$ is a uniform distribution supported on the Cantor set C. This is a well-defined distribution by Theorem **3.7** since the sets composing C are countable intersections of intervals and hence in $\mathcal{B}$. However it is singular with respect to m, increasing only on a set of Lebesgue measure zero. It is not discrete since there are no point masses, but nor is it absolutely continuous. In this case, $F = F_3$. □

The collection of half-lines with rational endpoints generates $\mathcal{B}$ (**1.21**) and should be a determining class for measures on $(\mathbb{R}, \mathcal{B})$. The following theorem establishes the fact that a c.d.f. defined on a dense subset of $\mathbb{R}$ is a unique representation of μ.

8.5 Theorem Let μ be a finite measure on $(\mathbb{R}, \mathcal{B})$ and D a dense subset of $\mathbb{R}$. The function G defined by

$$G(x) = \begin{cases} F(x) = \mu((-\infty, x]), & x \in D \\ F(x+), & x \in \mathbb{R} - D \end{cases} \tag{8.10}$$

is identical with F.

Proof By definition, $\mathbb{R} \subseteq \bar{D}$ and the points of $\mathbb{R} - D$ are all closure points of D. For each $x \in \mathbb{R}$, not excluding points in $\mathbb{R} - D$, there is a sequence of points in D converging to x (e.g. choose a point from $S(x, 1/n) \cap D$ for $n \in \mathbb{N}$). Since F is right-continuous everywhere on $\mathbb{R}$, $\mu((-\infty, x]) = F(x+)$ for each $x \in \mathbb{R} - D$. ∎

Likewise, every F corresponds to some μ as well as every μ to an F.

8.6 Theorem Let $F: \bar{\mathbb{R}} \to [0, 1]$ be a non-negative, non-decreasing, right-continuous function, with $F(-\infty) = 0$ and $F(+\infty) = 1$. There exists a unique p.m. μ on $(\mathbb{R}, \mathcal{B})$ such that $F(x) = \mu((-\infty, x])$ for all $x \in \mathbb{R}$. □

Right-continuity, as noted above, corresponds to the convention of defining F by (8.2). If instead the definition took the form $F(x) = \mu((-\infty, x))$, a left-continuous non-decreasing F would represent a p.m.

Proof of 8.6 Consider the function $\phi : [0,1] \mapsto \bar{\mathbb{R}}$, defined by

$$\phi(u) = \inf\{x : u \leq F(x)\}. \tag{8.11}$$

ϕ can be thought of as the inverse of F; $\phi(0) = -\infty$, $\phi(1) = +\infty$ and since F is non-decreasing and right-continuous, ϕ is non-decreasing and left-continuous; ϕ is therefore Borel-measurable by **3.39**(ii). According to **3.28**, a measure on $(\mathbb{R}, \mathcal{B})$ may be defined by $m\phi^{-1}(B)$ for each $B \in \mathcal{B}$, where m is Lebesgue measure on the Borel sets of $[0,1]$.

In particular, consider the class $\mathcal{C}$ of the half-open intervals $(a,b]$ for all $a, b \in \mathbb{R}$ with $a < b$. This is a semi-ring by **1.18** and $\sigma(\mathcal{C}) = \mathcal{B}$ by **1.21**. Note that

$$\phi^{-1}\big((a,b]\big) = \big\{u : \inf\{x : u \leq F(x)\} \in (a,b]\big\} = (F(a), F(b)]. \tag{8.12}$$

For each of these sets define the measure

$$\mu\big((a,b]\big) = m\phi^{-1}\big((a,b]\big)\big) = F(b) - F(a). \tag{8.13}$$

The fact that this is a measure follows from the argument of the preceding paragraph. $\mathcal{C}$ is a determining class for $(\mathbb{R}, \mathcal{B})$ and the measure has an extension by **3.15**. It is a p.m. since $\mu(\mathbb{R}) = 1$ and is unique by **3.20**. ∎

The neat construction used in this proof has other applications in the theory of random variables. It reappears in §8.3 and in more elaborate form in §23.2. The

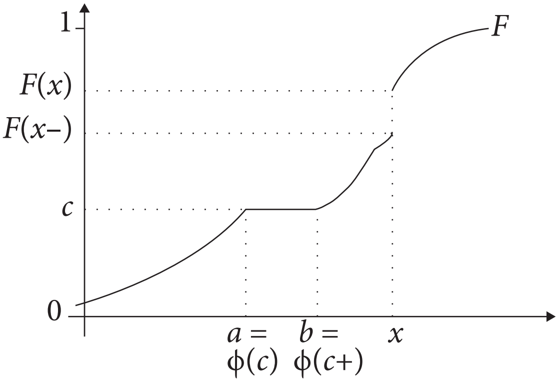

Figure 8.2

graph of ϕ is found by rotating and reflecting the graph of F, sketched in Figure 8.2; to see the former with the usual coordinates, turn the page on its side and view in a mirror. If F has a discontinuity at x, then $\phi = x$ on the interval $(F(x-), F(x)]$ and $\phi^{-1}(\{x\}) = (F(x-), F(x)]$. Thus,

$$\mu\big(\{x\}\big) = m\big((F(x-), F(x)]\big) = F(x) - F(x-)$$

as required. On the other hand, if an interval $(a, b]$ has measure 0 under F, F is constant on this interval and ϕ has a discontinuity at $F(a) = F(b) = c$ (say). Note that $\phi^{-1}((a, b]) = \{c\}$ so that $\mu((a, b]) = m(c) = 0$, as required.

8.3 Examples

The closed-form distributions used to model random phenomena are in most cases either *discrete* or *continuous*. A discrete distribution has $F'' = 0$ in the decomposition of **8.3**, assigning zero probability to all but a finite or countable set of atoms, while in a continuous distribution F is absolutely continuous, with $F_1 = 0$ in the Lebesgue decomposition of the c.d.f. A mixed continuous–discrete distribution such as the stylized example sketched in Figure 8.1 would arise in practice only as a convolution (see (11.1)) of two or more distinct functional forms.

Well-known discrete cases include the following.

8.7 Example A *two-point* distribution assigns outcomes to points u and v on the line with respective probabilities p and $q = 1 - p$. The best-known case is the *Bernoulli* distribution with $u = 1$ and $v = 0$. Think of this as a mapping from any probability space containing two elements such as 'Success' and 'Failure', 'Yes' and 'No', etc. The case with $u = 1$, $v = -1$, and $p = \frac{1}{2}$ is known as the *Rademacher* distribution. A two-point distribution having a mean of zero is completely specified by the point values, say u and $-v$ for $u, v > 0$, noting that $P(X = u) = v/(u + v)$ and $P(X = -v) = u/(u + v)$ uniquely fix $\mathrm{E}(X) = 0$. □

8.8 Example The *binomial* distribution with parameters n and p is the distribution of the number of ones obtained in n independent drawings from the Bernoulli distribution, with $n + 1$ distinct outcomes. The probability function is

$$P(X = x) = \binom{n}{x} p^x (1 - p)^{n-x}, \; x = 0, \dots, n \tag{8.14}$$

where

$$\binom{n}{x} = \frac{n!}{x!(n - x)!}$$

is the number of different ways that x ones and $n-x$ zeros can arise in n Bernoulli draws, disregarding order. $\sum_{x=0}^{n} P(X=x) = (p+(1-p))^n = 1$ by the binomial theorem. □

8.9 Example The limiting case as $n \to \infty$ of **8.8** with $p = \lambda/n$ is the *Poisson* distribution having probability function

$$P(X=x) = \frac{1}{x!}\mathrm{e}^{-\lambda}\lambda^x,\ x = 0,1,2,\ldots. \tag{8.15}$$

This is a discrete distribution with a countably infinite set of outcomes. The formula is easily obtained by noting that $n^{-x}n!/(n-x)! \to 1$ and $(1-\lambda/n)^{n-x} \to \mathrm{e}^{-\lambda}$ when $n \to \infty$ with x fixed. $\sum_{x=0}^{\infty} P(X=x) = 1$ follows by definition of e^{λ}. □

Continuous cases are distinguished by the existence of the derivative $f = \mathrm{d}F/\mathrm{d}x$ a.e.$[m]$ on $\mathbb{R}$ called the *probability density function* (p.d.f.) of the p.m. According to the Radon–Nikodym theorem the p.d.f. has the property

$$\mu(E) = \int_E f(x)\mathrm{d}x \tag{8.16}$$

for each $E \in \mathcal{B}$. When it exists in closed form, as in the following examples, the p.d.f. is the usual means of characterizing a distribution.

8.10 Example For the *uniform* distribution on the unit interval (see **7.3**) the c.d.f. is

$$F(x) = \begin{cases} 0, & x < 0 \\ x, & 0 \le x \le 1 \\ 1, & x > 1. \end{cases} \tag{8.17}$$

The p.d.f. is constant at 1 on the interval, but is undefined at 0 and 1. This r.v. is denoted $X \sim_{\mathrm{d}} \mathrm{U}[0,1]$. □

Matching Lebesgue measure on the unit interval, the uniform distribution has a neat application in the computer simulation of random variables. U[0, 1] random variables (strictly, pseudo-random variables) are easy to generate on a computer. There are various computational tricks for forming a string of digits in any desired base (digital or binary) that appear randomly chosen, such as dividing one large number by another and taking the remainder. Place a decimal point at the front of the string to specify a number in the unit interval, as in (1.14) with $m = 1$ and $p = 0$. Such a number can be indistinguishable from a U[0, 1] drawing that has been truncated to the available word length of the computer and long sequences of such numbers generated recursively by a suitable algorithm can appear serially independent. (See e.g. [119].)

See Example **24.1** for a well-known application of this technique. More generally, consider Figure 8.2 and the discussion following Theorem **8.6**. Let the generated uniform variate X define a point on the vertical axis between 0 and 1, and read off from the c.d.f. the corresponding point of the abscissa, that is, $\phi(X)$. In case $X = c$ in the plot, note that $\phi(X)$ is the infimum of the possible values. It is easy to appreciate that $\phi(X)$ is a drawing from the distribution with c.d.f. F. In principle, any distribution whatsoever may be generated from $\mathrm{U}[0,1]$ drawings and a formula or tabulation for F. In addition to its computational role this construction is also the key to Skorokhod's characterization of convergence in distribution, to be shown in **23.6**.

8.11 Example The *exponential* distribution on $[0,\infty)$ with rate λ has c.d.f. $F(x) = 1 - \mathrm{e}^{-\lambda x}$ and the p.d.f. is therefore

$$f(x;\lambda) = \lambda \mathrm{e}^{-\lambda x}. \quad \square \tag{8.18}$$

8.12 Example The family of *gamma* distributions, denoted gamma(p,λ), have the p.d.f.

$$f(x;p,\lambda) = \frac{\lambda}{\Gamma(p)}\mathrm{e}^{-\lambda x}(\lambda x)^{p-1},\ 0 \le x < +\infty. \tag{8.19}$$

where $p > 0$ is the shape parameter and $\lambda > 0$ the scale parameter. □

The gamma(p,λ) for integer p is the distribution of the sum of p independently drawn exponential r.v.s with rate λ so that in particular (8.19) with $p = 1$ matches (8.18). $\Gamma(p) = \int_0^\infty \xi^{p-1}\mathrm{e}^{-\xi}\mathrm{d}\xi$ is the gamma function whose properties include $\Gamma(\frac{1}{2}) = \sqrt{\pi}$, $\Gamma(1) = \Gamma(2) = 1$, and the recurrence relation $\Gamma(p) = (p-1)\Gamma(p-1) = \cdots = (p-1)!$ for $p > 1$.

For reasons doubtless familiar to the reader, the Gaussian distribution and those derived from it are ubiquitous in probability.

8.13 Example The *standard normal* distribution, also known as Gaussian, has p.d.f.

$$f(x) = \frac{1}{\sqrt{2\pi}}\mathrm{e}^{-x^2/2},\ -\infty < x < +\infty \tag{8.20}$$

whose graph is the well-known bell-shaped curve with mode at 0. □

Since integrals involving (8.20) do not generally have closed forms, the Gaussian formula is not the easiest to calculate with. The following result provides a useful approximation for tail probabilities.

8.14 Theorem If X is standard Gaussian,

$$P(X \geq c) = \frac{1}{c\sqrt{2\pi}} \mathrm{e}^{-c^2/2}(1 + O(c^{-2})) \tag{8.21}$$

and for $c \geq 1/\sqrt{2\pi} \approx 0.4$,

$$P(X \geq c) \leq \mathrm{e}^{-c^2/2}. \tag{8.22}$$

Proof For (8.21) see for example [1], where Formula 26.2.12 gives the full asymptotic expansion. To show (8.22) by a simple calculation with confirmation that the correction term in (8.21) is negative for large enough c, note that

$$\begin{aligned} P(X \geq c) &= \frac{1}{\sqrt{2\pi}} \int_c^\infty \mathrm{e}^{-x^2/2} \mathrm{d}x \\ &\leq \frac{1}{\sqrt{2\pi}} \int_c^\infty \frac{x}{c} \mathrm{e}^{-x^2/2} \mathrm{d}x \\ &= \frac{1}{c\sqrt{2\pi}} \mathrm{e}^{-c^2/2} \end{aligned} \tag{8.23}$$

where the first equality is by definition and the second one evaluates the integral. Inequality (8.22) follows directly given the choice of c. ∎

8.15 Example The standard *Cauchy* distribution has p.d.f.

$$f(x) = \frac{1}{\pi(1 + x^2)}, \quad -\infty < x < +\infty. \quad \square \tag{8.24}$$

Like the Gaussian, the Cauchy density is symmetric with mode at 0. It is possible to show that this is the distribution of the ratio of two independently drawn standard normals.

A useful trick is to be able to derive the distribution of $\varphi(X)$ from that of X when φ is a function of a suitable type. In particular the inverse function φ^{-1} needs to be differentiable with the derivative never equal to zero and hence never changing its sign.

8.16 Theorem Let $\varphi : \mathbb{S} \mapsto \mathbb{T}$ be a 1–1 function onto $\mathbb{T}$, where $\mathbb{S}$ and $\mathbb{T}$ are open subsets of $\mathbb{R}$. Let $\psi = \varphi^{-1}$ be continuously differentiable with $\mathrm{d}\psi/\mathrm{d}y \neq 0$ for all $y \in \mathbb{T}$. If X is continuously distributed with p.d.f. f and $Y = \varphi(X)$, then Y is continuously distributed with p.d.f.

$$g(y) = f(\psi(y))\left|\frac{\mathrm{d}\psi}{\mathrm{d}y}\right|. \tag{8.25}$$

Proof Start from the relation $P(Y \leq y) = P(\varphi(X) \leq y) = P(X \leq \psi(y))$. Letting F and G be the c.d.f.s of X and Y respectively, if φ is an increasing function of x this relation has the form $G(y) = F(\psi(y))$ and differentiating G using the chain rule gives

$$\mathrm{d}G = g(y)\mathrm{d}y = f(\psi(y))\frac{\mathrm{d}\psi}{\mathrm{d}y}\mathrm{d}y. \tag{8.26}$$

If φ is a decreasing function of x, then $\mathrm{d}\psi/\mathrm{d}y < 0$ and formula (8.26) fails since G must be non-decreasing in its argument by definition, with $g(y) \geq 0$. To have G increase with $y = \varphi(x)$ necessitates having x decrease. However, the effect of replacing $\mathrm{d}x$ with $-\mathrm{d}x$ is achieved in the formula by replacing $\mathrm{d}\psi$ with $-\mathrm{d}\psi$. Formula (8.25) therefore applies in either case. ∎

This result illustrates **3.28** but in most other cases it is a great deal harder than this to derive a closed form for a transformed distribution. A leading application where the assumptions apply is to generate families of distributions with alternative location and scale parameters.

8.17 Example Generalize the uniform distribution (**8.10**) from $[0, 1]$ to an arbitrary interval $[a, b]$. The transformation is linear with

$$X = a + (b - a)Z \tag{8.27}$$

where Z is a U$[0, 1]$ drawing, so that $f(x) = (b - a)^{-1}$ on (a, b) by (8.25). The c.d.f. is defined on $[a, b]$ by

$$F_X(x) = (x - a)/(b - a). \tag{8.28}$$

Membership of the uniform family is denoted by $X \sim_{\mathrm{d}} \mathrm{U}[a, b]$. □

8.18 Example Linear transformations of the form $X = \mu + \sigma Z$ with $\sigma > 0$ where Z is standard normal (**8.13**) generate the Gaussian family of distributions denoted by $X \sim_{\mathrm{d}} \mathrm{N}(\mu, \sigma^2)$. $\psi = (x - \mu)/\sigma$ in (8.25) and the p.d.f.s have the form

$$f_X(x; \mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma}\mathrm{e}^{-(x-\mu)^2/2\sigma^2}, \ -\infty < x < +\infty. \tag{8.29}$$

The location parameter μ and scale parameter σ^2 have better-known designations as the first two *moments* of the distribution, see **9.6** and **9.14**. □

8.19 Example The family of Cauchy distributions denoted $\mathrm{C}(\nu,\delta)$ is generated from the standard Cauchy (**8.15**) by linear transformations $X = \nu + \delta Z, \delta > 0$. The family of p.d.f.s. with location parameter ν and scale parameter δ take the form

$$f_X(x;\nu,\delta) = \frac{1}{\pi\delta\left(1+\left(\frac{x-\nu}{\delta}\right)^2\right)}, -\infty < x < +\infty. \quad \square \tag{8.30}$$

Consider the square of a standard Gaussian r.v. with $\mu = 0$ and $\sigma = 1$. Since the transformation is not monotone **8.16** cannot be used to determine the density, but consider the 'half-normal' density

$$f_{|X|}(x) = \begin{cases} 2f_X(x), & x \geq 0 \\ 0, & x < 0 \end{cases} \tag{8.31}$$

where f_X is given by (8.20). This is the p.d.f. of $|X|$, the absolute value of a standard Gaussian variable. The transformation $\varphi(|x|) = x^2$ is 1−1 with inverse $\sqrt{x}$. Applying (8.25) to (8.31), the p.d.f. of $Y = X^2$ is therefore

$$f_Y(y) = \frac{1}{\sqrt{2\pi}} \mathrm{e}^{-y/2} y^{-1/2}, \; 0 < y < \infty. \tag{8.32}$$

8.20 Example The family of *chi-squared* distributions with degrees of freedom ν, written $\chi^2(\nu)$, are the distributions of the sum of the squares of ν independently drawn standard normals. Equation (8.32) is the p.d.f. for the case $\chi^2(1)$. Remarkably, the $\chi^2(\nu)$ is the member of the gamma(p,λ) family (**8.12**) with $\lambda = \frac{1}{2}$ and $\nu = 2p$; for the case $\nu = 1$, compare (8.32) with (8.19). □

8.21 Example The family of *Student's t* distributions with degrees-of-freedom parameter ν, denoted $t(\nu)$, has p.d.f.

$$f(x;\nu) = \frac{\Gamma((\nu+1)/2)}{\sqrt{\nu\pi}\Gamma(\nu/2)}\left(1+\frac{x^2}{\nu}\right)^{-(\nu+1)/2}, \; -\infty < x < +\infty. \tag{8.33}$$

The $t(\nu)$ distributions are the distributions of the ratio of a $\mathrm{N}(0,1)$ r.v. to the square root of an independently drawn $\chi^2(\nu)$ divided by ν. It is easily verified from (8.33) that $t(1)$ is identical to the standard Cauchy and also (noting that $\Gamma(n+z)/(\Gamma(n)n^z) \to 1$ as $n \to \infty$ for $z > 0$) that the $t(\infty)$ is identical to the standard Gaussian. □

8.4 Multivariate Distributions

In Euclidean k-space $\mathbb{R}^k$, the k-dimensional Borel field $\mathcal{B}^k$ is $\sigma(\mathcal{R}^k)$ where $\mathcal{R}^k$ denotes the measurable rectangles of $\mathbb{R}^k$, the sets of the form $B_1 \times B_2 \times \ldots \times B_k$ where $B_i \in \mathcal{B}$ for $i = 1, \ldots, k$. In a space $(\Omega, \mathcal{F}, P)$ a random vector $\boldsymbol{X} = (X_1, X_2, \ldots, X_k)'$ is a measurable mapping

$$\boldsymbol{X} : \Omega \to \mathbb{R}^k.$$

If μ is the derived measure such that $\mu(A) = P(E)$ for $A = \boldsymbol{X}(E) \in \mathcal{B}^k$ and $E \in \mathcal{F}$, the multivariate c.d.f. $F : \bar{\mathbb{R}}^k \to [0,1]$ is defined for $(x_1, \ldots, x_k)' = \boldsymbol{x}$ by

$$F(\boldsymbol{x}) = \mu\big((-\infty, x_1] \times \ldots \times (-\infty, x_k]\big). \tag{8.34}$$

The extension proceeds much like the scalar case.

8.22 Example Consider the random pair (X, Y). Let $F(x,y) = \mu((-\infty, x] \times (-\infty, y])$. The measure of the half-open rectangle $(x, x + \Delta x] \times (y, y + \Delta y]$ is

$$\Delta F(x,y) = F(x + \Delta x, y + \Delta y) - F(x + \Delta x, y) - F(x, y + \Delta y) + F(x,y). \tag{8.35}$$

To show this, consider the four disjoint sets of $\mathbb{R}^2$ illustrated in Figure 8.3:

$$\begin{aligned} A &= (x, x+\Delta x] \times (y, y+\Delta y] \quad & B &= (-\infty, x] \times (y, y+\Delta y] \\ C &= (x, x+\Delta x] \times (-\infty, y] \quad & D &= (-\infty, x] \times (-\infty, y]. \end{aligned}$$

A is the set whose probability is sought. Since $P(A \cup B \cup C \cup D) = F(x + \Delta x, y + \Delta y)$, $P(B \cup D) = F(x, y + \Delta y)$, $P(C \cup D) = F(x + \Delta x, y)$, and $P(D) = F(x,y)$, the result is immediate from the probability axioms. □

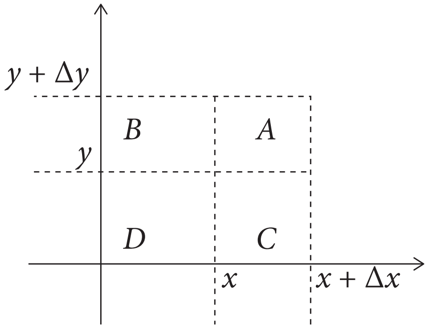

Figure 8.3

Extending the approach of **8.22** inductively, the measure of the k-dimensional rectangle $\Pi_{i=1}^{k}(x_i, x_i + \Delta_i]$ can be shown to be

$$\Delta F(x_1, \dots, x_k) = \sum_j (\pm F_j) \tag{8.36}$$

where the sum on the right has 2^k terms and the F_j are the values of F at each of the vertices of the k-dimensional rectangle extending from $(x_1, \dots, x_k)'$ with sides of length Δx_i, $i = 1, \dots, k$. The sign pattern depends on k; if k is odd, the F_j having as arguments even numbers of upper vertices (points of the form $x_i + \Delta x_i$) take negative signs and the others positive; while if k is even, the F_j with odd numbers of upper vertices as arguments are negative. Generalizing the monotonicity of the univariate c.d.f., F must satisfy the condition that $\Delta F(x_1, \dots, x_k)$ be non-negative for every choice of $(x_1, \dots, x_k)' \in \mathbb{R}^k$ and $(\Delta x_1, \dots, \Delta x_k)' \in (\mathbb{R}^k)^+$. Applying **3.26** inductively shows that the class of k-dimensional half-open rectangles is a semi-ring, so that the measure defined by F extends to the sets of $\mathcal{B}^k$, hence $(\mathbb{R}^k, \mathcal{B}^k, \mu)$ is a probability space derived from $(\Omega, \mathcal{F}, P)$.

If the distribution is continuous with p.d.f. f, Fubini's theorem gives

$$\begin{aligned} F(x_1, \dots, x_k) &= \int_{(-\infty, x_1] \times \dots \times (-\infty, x_k]} f(\xi_1, \dots, \xi_k) \mathrm{d}\xi_1 \dots \mathrm{d}\xi_k \\ &= \int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_k} f(\xi_1, \dots, \xi_k) \mathrm{d}\xi_1 \dots \mathrm{d}\xi_k. \end{aligned} \tag{8.37}$$

A compact notation is helpful for writing expressions of the form (8.37). Defining $\boldsymbol{\xi} = (\xi_1, \dots, \xi_k)'$ and letting $\mathrm{d}\boldsymbol{\xi}$ denote the corresponding column k-vector of differentials, also let $\mathrm{d}\hat{\boldsymbol{\xi}} = \mathrm{diag}(\mathrm{d}\boldsymbol{\xi})$ denote the $k \times k$ matrix having these elements on the diagonal and zeros elsewhere. Note that the determinant of this matrix is $|\mathrm{d}\hat{\boldsymbol{\xi}}| = \mathrm{d}\xi_1 \dots \mathrm{d}\xi_k$. Also letting $\int_{-\infty}^{\boldsymbol{x}}$ stand for $\int_{-\infty}^{x_1} \dots \int_{-\infty}^{x_k}$ where $\boldsymbol{x} = (x_1, \dots, x_k)'$, a compact representation of the multiple integral in (8.37) is

$$F(\boldsymbol{x}) = \int_{-\infty}^{\boldsymbol{x}} f(\boldsymbol{\xi}) |\mathrm{d}\hat{\boldsymbol{\xi}}|. \tag{8.38}$$

With this notation it is straightforward to develop the generalization of Theorem **8.16**. The problem is to obtain the joint distribution of the vector $\boldsymbol{Y} = \boldsymbol{\varphi}(\boldsymbol{X})$ where $\boldsymbol{\varphi} : \mathbb{S} \mapsto \mathbb{T}$ ($\mathbb{S}$ and $\mathbb{T}$ open subsets of $\mathbb{R}^k$) is a *diffeomorphism* (also, *coordinate transformation).* This is a function that is 1−1 onto and continuously differentiable with continuously differentiable inverse, to be denoted $\boldsymbol{\psi} = \boldsymbol{\varphi}^{-1}$. If $\boldsymbol{y} = (y_1, \dots, y_k)'$, $\boldsymbol{\zeta}$ $(k \times 1)$ is the corresponding argument, and $\partial \boldsymbol{\psi} / \partial \boldsymbol{\zeta}'$ $(k \times k)$ is the Jacobian matrix of the inverse transformation whose (i,j)th element is $\partial \psi_j / \partial \zeta_i$

for $i, j = 1, \ldots, k$, the definition requires that the Jacobian determinant $J(\boldsymbol{\zeta}) = |\partial\boldsymbol{\psi}/\partial\boldsymbol{\zeta}'|$ is nonzero at all $\boldsymbol{\zeta} \in \mathbb{T}$ meaning in particular that it does not change sign.

The c.d.f. of the transformed vector is

$$G(\boldsymbol{y}) = \int_{-\infty}^{\boldsymbol{y}} g(\boldsymbol{\zeta})|\mathrm{d}\hat{\boldsymbol{\zeta}}|. \tag{8.39}$$

The question at issue is the functional form of the derived probability density function g. The relationship between the differentials has the form

$$\mathrm{d}\boldsymbol{\xi} = (\partial\boldsymbol{\psi}/\partial\boldsymbol{\zeta}')\mathrm{d}\boldsymbol{\zeta} \quad (k \times 1). \tag{8.40}$$

However, these equations can equally well be written $\mathrm{d}\hat{\boldsymbol{\xi}} = (\partial\boldsymbol{\psi}/\partial\boldsymbol{\zeta}')\mathrm{d}\hat{\boldsymbol{\zeta}}\,(k \times k)$, of which (8.40) represents merely the sum of the columns of the diagonal matrix $\mathrm{d}\hat{\boldsymbol{\xi}}$. Taking the determinants of each side gives

$$|\mathrm{d}\hat{\boldsymbol{\xi}}| = |(\partial\boldsymbol{\psi}/\partial\boldsymbol{\zeta}')\mathrm{d}\hat{\boldsymbol{\zeta}}| = J(\boldsymbol{\zeta})|\mathrm{d}\hat{\boldsymbol{\zeta}}|. \tag{8.41}$$

Thus, assuming $J > 0$ the calculation reduces to

$$\begin{aligned} \mathrm{d}G(\boldsymbol{\zeta}) &= \mathrm{d}F(\boldsymbol{\psi}(\boldsymbol{\zeta})) \\ &= f(\boldsymbol{\psi}(\boldsymbol{\zeta}))|\mathrm{d}\widehat{\boldsymbol{\psi}(\boldsymbol{\zeta})}| \\ &= f(\boldsymbol{\psi}(\boldsymbol{\zeta}))J(\boldsymbol{\zeta})|\mathrm{d}\hat{\boldsymbol{\zeta}}| \end{aligned} \tag{8.42}$$

where the last equality applies (8.41). This gives a formula for the derived density in the case $J > 0$.

However, since the solution requires $|\mathrm{d}\hat{\boldsymbol{\zeta}}| > 0$ this does not work if $J < 0$, which can happen if elements of $\boldsymbol{\varphi}$ are decreasing in their arguments; compare the univariate case in **8.16**. If an element of $\mathrm{d}\boldsymbol{\xi}$ in (8.40) must be negative to have $\mathrm{d}G > 0$ in (8.42), the equivalent effect is obtained by changing the sign of the corresponding row of $\partial\boldsymbol{\psi}/\partial\boldsymbol{\zeta}'$. Such changes necessarily render $J > 0$ and so are equivalent to replacing J by the absolute value $|J|$. The required counterpart of **8.16** is therefore the following.

8.23 Theorem If f is the p.d.f. of $\boldsymbol{X}$ and $\boldsymbol{Y} = \boldsymbol{\varphi}(\boldsymbol{X})$ where $\boldsymbol{\varphi}$ is a diffeomorphism and $\boldsymbol{\psi} = \boldsymbol{\varphi}^{-1}$, the p.d.f. of $\boldsymbol{Y}$ is

$$g(\boldsymbol{\zeta}) = f(\boldsymbol{\psi}(\boldsymbol{\zeta}))|J| \tag{8.43}$$

where $J = \det(\partial\boldsymbol{\psi}/\partial\boldsymbol{\zeta}')$. □

8.24 Example Letting f denote the standard Gaussian p.d.f. (see **8.13**), consider

$$\phi(\boldsymbol{z}) = \prod_{i=1}^{k} f(z_i) = (2\pi)^{-k/2}\exp\{-\tfrac{1}{2}\boldsymbol{z}'\boldsymbol{z}\}. \tag{8.44}$$

This is the p.d.f. of the standard Gaussian k-vector $\boldsymbol{Z} = (Z_1, \ldots, Z_k)'$. Consider the affine transformation

$$\boldsymbol{X} = \boldsymbol{A}\boldsymbol{Z} + \boldsymbol{\mu} \tag{8.45}$$

where $\boldsymbol{A}$ ($k \times k$ nonsingular) and $\boldsymbol{\mu}$ ($k \times 1$) are constants. This is $1-1$ continuous with inverse $\boldsymbol{Z} = \boldsymbol{A}^{-1}(\boldsymbol{X} - \boldsymbol{\mu})$ having Jacobian determinant $J = |\boldsymbol{A}^{-1}| \; = 1/|\boldsymbol{A}|$. Define $\boldsymbol{\Sigma} = \boldsymbol{A}\boldsymbol{A}'$ such that $(\boldsymbol{A}^{-1})'\boldsymbol{A}^{-1} = (\boldsymbol{A}\boldsymbol{A}')^{-1} = \boldsymbol{\Sigma}^{-1}$ and $||\boldsymbol{A}|^{-1}| = |\boldsymbol{\Sigma}|^{-1/2}$, the positive square root being understood. Applying **8.23** produces

$$\begin{aligned} f(\boldsymbol{x}) &= \phi\big(\boldsymbol{A}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\big)\Big|\frac{1}{|\boldsymbol{A}|}\Big| \\ &= (2\pi)^{-k/2}||\boldsymbol{A}^{-1}||\exp\{-\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'(\boldsymbol{A}^{-1})'\boldsymbol{A}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\} \\ &= (2\pi)^{-k/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\{-\tfrac{1}{2}(\boldsymbol{x} - \boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x} - \boldsymbol{\mu})\}. \end{aligned} \tag{8.46}$$

This is the multivariate normal p.d.f., depending on parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Every such distribution is generated by an affine transform applied to the standard normal vector $\boldsymbol{Z}$. Membership of the multivariate normal family is denoted $\boldsymbol{X} \sim_{\mathrm{d}} \mathrm{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$. □

8.5 Independent Random Variables

Suppose that out of a pair of r.v.s (X, Y) on $(\mathbb{R}^2, \mathcal{B}^2, \mu)$ the goal is exclusively to predict X. In this situation the events of interest are the cylinder sets in $\mathbb{R}^2$ having the form $B \times \mathbb{R}$, $B \in \mathcal{B}$. The marginal distribution of X is defined by $(\mathbb{R}, \mathcal{B}, \mu_X)$ where

$$\mu_X(A) = \mu(A \times \mathbb{R}) \tag{8.47}$$

for $A \in \mathcal{B}$. The associated marginal c.d.f. is $F_X(x) = F(x, +\infty)$.

The notion of independence defined in §7.4 specializes in the following way. X and Y are called independent r.v.s iff

$$\mu(A \times B) = \mu_X(A)\mu_Y(B) \tag{8.48}$$

for all pairs of events $A, B \in \mathcal{B}$, where μ_X is defined by (8.47) and μ_Y is analogous. Equivalently, μ is the product measure generated by μ_X and μ_Y.

8.25 Theorem X and Y are independent iff for each $x, y \in \mathbb{R}$

$$F(x,y) = F_X(x)F_Y(y). \tag{8.49}$$

If the distribution is continuous the p.d.f. factorizes as

$$f(x,y) = f_X(x)f_Y(y). \tag{8.50}$$

Proof Obviously (8.49) is true only if μ satisfies (8.48). The problem is to show that the former condition is also sufficient. Consider the half-open rectangles,

$$\mathcal{C} = \{(a,b] \times (c,d],\ a,c \in \bar{\mathbb{R}},\ b,d \in \mathbb{R}\}. \tag{8.51}$$

If and only if (8.49) holds,

$$\begin{aligned}\mu((a,b] \times (c,d]) &= F(b,d) - F(b,c) - F(a,d) + F(a,c) \\ &= (F_X(b) - F_X(a))(F_Y(d) - F_Y(c)) \\ &= \mu_X((a,b])\mu_Y((c,d])\end{aligned} \tag{8.52}$$

where the first equality is by **8.22**. $\mathcal{C}$ in (8.51) is a determining class for $(\mathbb{R}^2, \mathcal{B}^2)$ and μ is defined by the extension of the measure satisfying (8.52) for ordinary rectangles or, equivalently, satisfying (8.49). The extension theorem **3.20** (uniqueness part) shows that this is identical to the product measure satisfying (8.48). The extension to p.d.f.s follows directly from the definition. ∎

With more than two variables there are alternative independence concepts (compare §7.3). Variables $X_1, \ldots, X_k$ distributed on the space $(\mathbb{R}^k, \mathcal{B}^k, \mu)$ are said to be *totally independent* if

$$\mu\left(\prod_{i=1}^k A_i\right) = \prod_{i=1}^k \mu_{X_i}(A_i) \tag{8.53}$$

for all k-tuples of events $A_1, \ldots, A_k \in \mathcal{B}$. By contrast, pairwise independence can hold between each pair X_i, X_j without implying total independence of the set. Another way to think of total independence is in terms of a partitioning of a vector $\boldsymbol{X} = (X_1, \ldots, X_k)'$ into subvectors $\boldsymbol{X}_1$ $(j \times 1)$ and $\boldsymbol{X}_2$ $((k-j) \times 1)$ for $0 < j < k$. Under total independence, the measure of $\boldsymbol{X}$ is always expressible as the product measure of the two subvectors, under all orderings and partitionings of the elements.

9
Expectations

9.1 Averages and Integrals

When it exists the *expectation*, or *expected value*, or *mean*, of a r.v. $X(\omega)$ in a probability space $(\Omega, \mathcal{F}, P)$ is the integral

$$\mathrm{E}(X) = \int_{\Omega} X(\omega)\mathrm{d}P(\omega). \tag{9.1}$$

$\mathrm{E}(X)$ measures the central tendency of the distribution of X. It is sometimes identified with the limiting value of the sample average of realized values x_t drawn in n identical random experiments,

$$\bar{x}_n = \frac{1}{n}\sum_{t=1}^{n} x_t \tag{9.2}$$

as n becomes large. However, the validity of this hypothesis depends on the method of repeating the experiment. See Part IV for the details but suffice it to say at this point that the equivalence certainly holds if $\mathrm{E}(X)$ exists and the random experiments are independent of one another.

The connection is most evident for simple random variables. If $X = \sum_j x_j 1_{E_j}$ where the $\{E_j\}$ are a partition of Ω, then by **4.6**

$$\mathrm{E}(X) = \sum_j x_j P(E_j). \tag{9.3}$$

When the probabilities are interpreted as relative frequencies of the events $E_j = \{\omega : X(\omega) = x_j\}$ in a large number of drawings from the distribution, (9.2) with large n should approximate (9.3). The values x_j will appear in the sum in a proportion approaching their probability of occurrence with n large.

$\mathrm{E}(X)$ has a dual characterization, as an abstract integral on the parent probability space and as a Lebesgue–Stieltjes integral on the line, under the derived distribution. It is equally correct to write either (9.1) or

$$\mathrm{E}(X) = \int_{-\infty}^{+\infty} x\mathrm{d}F(x). \tag{9.4}$$

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0009

Which of these representations is adopted is mainly a matter of convenience. If $1_A(\omega)$ is the indicator function of a set $A \in \mathcal{F}$, then

$$\mathrm{E}(1_A X) = \int_A X(\omega)\mathrm{d}P(\omega) = \int_{X(A)} x\mathrm{d}F(x) \tag{9.5}$$

where $X(A) \in \mathcal{B}$ is the image of A under X. Here the abstract integral is obviously the more direct and simple representation, but by the same token the Stieltjes form is the natural way to represent integration over a set in $\mathcal{B}$.

If the distribution is discrete, X is a simple function and the formula in (9.3) applies directly. Under the derived distribution,

$$\mathrm{E}(X) = \sum_j x_j \mu(\{x_j\}) \tag{9.6}$$

where $x_j, j = 1, 2, \ldots$ are the atoms of the distribution.

9.2 Applications

9.1 Example If X has a two-point distribution (**8.7**) on points u, v with respective probabilities p and $1 - p$ then $\mathrm{E}(X) = pu + (1 - p)v$. The Bernoulli r.v. with $u = 1$ and $v = 0$ has $\mathrm{E}(X) = p$. The Rademacher r.v. with $u = 1$, $v = -1$, and $p = \frac{1}{2}$ has $\mathrm{E}(X) = 0$. □

9.2 Example If X is binomial(n,p) (**8.8**),

$$\mathrm{E}(X) = \sum_{x=1}^{n} x\binom{n}{x}p^x(1-p)^{n-x} = np\sum_{x=1}^{n}\binom{n-1}{x-1}p^{x-1}(1-p)^{n-x} = np.$$

A quicker route to this formula is to combine the fact that X is the sum of n independent Bernoulli drawings with mean p with the linearity of the integral. □

9.3 Example If X is Poisson (**8.9**),

$$\mathrm{E}(X) = \mathrm{e}^{-\lambda}\sum_{x=1}^{\infty} x\frac{\lambda^x}{x!} = \mathrm{e}^{-\lambda}\lambda\sum_{x=1}^{\infty}\frac{\lambda^{x-1}}{(x-1)!} = \lambda. \tag{9.7}$$

Alternatively note that λ is the mean of the binomial with $p = \lambda/n$ from **9.2** and take the limit as $n \to \infty$. □

In continuous distributions the Lebesgue–Stieltjes integral of x coincides with the integral with respect to ordinary Lebesgue measure of the function $xf(x)$.

9.4 Example For the uniform distribution on the interval $[a,b]$ (**8.17**),

$$E(X) = \frac{1}{b-a}\int_a^b x\mathrm{d}x = \frac{1}{2}(a+b). \quad \square \tag{9.8}$$

9.5 Example For the exponential distribution with rate λ (**8.11**),

$$E(X) = \lambda\int_0^\infty xe^{-\lambda x}\mathrm{d}x = \frac{1}{\lambda} \tag{9.9}$$

(see [85], 3.351). It follows that the gamma(p,λ) distribution (**8.12**) being the sum of p independent exponentials has

$$E(X) = \frac{p}{\lambda}. \quad \square \tag{9.10}$$

9.6 Example For the Gaussian family (**8.24**),

$$E(X) = \frac{1}{\sqrt{2\pi\sigma}}\int_{-\infty}^\infty xe^{-(x-\mu)^2/2\sigma^2}\mathrm{d}x = \mu. \tag{9.11}$$

With a change of variable from x to $z = x - \mu$, this follows since the density integrates to 1 and $ze^{-z^2/2\sigma^2}$ integrates to zero, being an odd function. □

In a mixed continuous–discrete distribution with atoms $x_1, x_2, \ldots$, there is the decomposition $F = F_1 + F_2$ where $F_1(x) = \sum_{x_j \le x}\mu_1(\{x_j\})$ and $F_2(x)$ is absolutely continuous with derivative $f_2(x)$. Then

$$E(X) = \sum_j x_j\mu_1(\{x_j\}) + \int xf_2(x)\mathrm{d}x. \tag{9.12}$$

The set of atoms has Lebesgue measure zero in $\mathbb{R}$, so there is no need to exclude these from the integral on the right-hand side of (9.12).

Some random variables do not have an expectation.

9.7 Example Recall the condition for integrability in (4.4) and note that for the Cauchy distribution (**8.15**),

$$\mathrm{E}(|X|1_{\{|X|\le a\}}) = \frac{1}{\pi}\int_{-a}^{+a}\frac{|x|}{1+x^2}\mathrm{d}x = \frac{\log(a^2+1)}{\pi} \to \infty \text{ as } a \to \infty. \quad \square \qquad (9.13)$$

In this case E(X) is undefined since it has the representation $\infty - \infty$ which is not a number. □

9.3 Expectations of Functions of X

If X is a r.v. on the probability space $(\mathbb{R}, \mathcal{B}, \mu)$ and $g : \mathbb{R} \mapsto \mathbb{R}$ is a Borel function, $g \circ X = g(X)$ is a r.v. on the space $(\mathbb{R}, \mathcal{B}, \mu g^{-1})$, as noted in §8.1. This leads to the following dual characterization of the expectation of a function.

9.8 Theorem If g is a Borel function.

$$\mathrm{E}\big(g(X)\big) = \int g(x)\mathrm{d}\mu(x) = \int y\mathrm{d}\mu g^{-1}(y). \qquad (9.14)$$

Proof Define a sequence of simple functions $Z_{(n)} : \mathbb{R}^+ \mapsto \mathbb{R}^+$ by

$$Z_{(n)}(x) = \sum_{i=1}^{m}\Big(\frac{i-1}{2^n}\Big)1_{B_i}(x) \qquad (9.15)$$

where $m = n2^n + 1$ and $B_i = \big[2^{-n}(i-1), 2^{-n}i\big)$ for $i = 1, \ldots, m$. Then, $Z_{(n)}(x) \uparrow x$ for $x \ge 0$, by arguments paralleling **3.35**. According to **3.28** $(\mathbb{R}, \mathcal{B}, \mu g^{-1})$ is a measure space where $\mu g^{-1}(B) = \mu\big(g^{-1}(B)\big)$ for $B \in \mathcal{B}$ and so by the monotone convergence theorem,

$$\int Z_{(n)}(y)\mathrm{d}\mu g^{-1}(y) = \sum_{i=1}^{m}\Big(\frac{i-1}{2^n}\Big)\mu g^{-1}(B_i) \to \int y\mathrm{d}\mu g^{-1}(y). \qquad (9.16)$$

Consider first the case of non-negative g. Let $1_B(x)$ be the indicator of the set $B \in \mathcal{B}$ and then if g is Borel, so is the composite function

$$(1_B \circ g)(x) = \left\{\begin{array}{ll} 1, & g(x) \in B \\ 0, & g(x) \notin B \end{array}\right\} = 1_{g^{-1}(B)}(x). \qquad (9.17)$$

Hence, consider the simple function

$$(Z_{(n)}\circ g)(x)=\sum_{i=1}^{m}\Big(\frac{i-1}{2^n}\Big)(1_{B_i}\circ g)(x)=\sum_{i=1}^{m}\Big(\frac{i-1}{2^n}\Big)1_{g^{-1}(B_i)}(x). \tag{9.18}$$

By monotone convergence, $Z_{(n)}\circ g\uparrow g$ and $\mathrm{E}(Z_{(n)}\circ g)\to \mathrm{E}(g)=\int g\mathrm{d}\mu$. However,

$$\begin{aligned}\mathrm{E}(Z_{(n)}\circ g)&=\sum_{i=1}^{m}\Big(\frac{i-1}{2^n}\Big)\mu\big(g^{-1}(B_i)\big)\\&=\sum_{i=1}^{m}\Big(\frac{i-1}{2^n}\Big)\mu g^{-1}(B_i)=\int Z_{(n)}(y)\mathrm{d}\mu g^{-1}(y)\end{aligned} \tag{9.19}$$

and (9.14) follows from (9.16).

To extend the result to general g consider the non-negative functions $g^{+}=\max\{g,0\}$ and $g^{-}=g^{+}-g$. It is immediate that

$$\mathrm{E}(Z_{(n)}\circ g^{+})-\mathrm{E}(Z_{(n)}\circ g^{-})\to \mathrm{E}(g^{+})-\mathrm{E}(g^{-})=\mathrm{E}(g) \tag{9.20}$$

so consider each component of this limit separately.

$$\begin{aligned}\mathrm{E}(Z_{(n)}\circ g^{+})&=\sum_{i=1}^{m}\Big(\frac{i-1}{2^n}\Big)\mu\big((g^{+})^{-1}(B_i)\big)\\&=\sum_{i=1}^{m}\Big(\frac{i-1}{2^n}\Big)\mu\big(g^{-1}(B_i)\big)\to\int_0^{\infty}y\mathrm{d}\mu g^{-1}(y)\end{aligned} \tag{9.21}$$

where the second equality holds because $(g^{+})^{-1}(B_i)=g^{-1}(B_i)$ for $i\geq 2$ since the elements of B_i are all positive for these cases, whereas for $i=1$ the term disappears. Similarly, $-Z_n(x)\downarrow x$ for $x<0$ and

$$\begin{aligned}-\mathrm{E}(Z_{(n)}\circ g^{-})&=-\sum_{i-1}^{m}\Big(\frac{i-1}{2^n}\Big)\mu\big((g^{-})^{-1}(B_i)\big)\\&=-\sum_{i=1}^{m}\Big(\frac{i-1}{2^n}\Big)\mu\big(g^{-1}(B_i^{-})\big)\to\int_{-\infty}^{0}y\mathrm{d}\mu g^{-1}(y)\end{aligned} \tag{9.22}$$

where $B_i^{-}=\big(-2^{-n}i,-2^{-n}(i-1)\big]$ and $(g^{-})^{-1}(B_i)=g^{-1}(B_i^{-})$ for $i\geq 2$. Hence

$$\mathrm{E}(Z_{(n)}\circ g^{+})-\mathrm{E}(Z_{(n)}\circ g^{-})\to\int_{-\infty}^{\infty}y\mathrm{d}\mu g^{-1}(y) \tag{9.23}$$

and the theorem follows in view of (9.20). ∎

9.4 Moments

The quantities $\mathrm{E}(X^k)$ for integer $k \geq 1$ are called the moments of the distribution of X, while more usefully the *central moments* for $k > 1$ are defined by

$$\mathrm{E}(X - \mathrm{E}(X))^k = \sum_{j=0}^{k} \binom{k}{j} \mathrm{E}(X^j)(-\mathrm{E}(X))^{k-j}. \tag{9.24}$$

The *variance* ($k = 2$), *skewness* ($k = 3$), and *kurtosis* ($k = 4$) are the familiar cases. The first of these arises most frequently in the sequel, with the decomposition

$$\mathrm{Var}(X) = \mathrm{E}(X - \mathrm{E}(X))^2 = \mathrm{E}(X^2) - \mathrm{E}(X)^2.$$

9.9 Theorem When the distribution is symmetric with $P\big(X - \mathrm{E}(X) \in A\big) = P\big(\mathrm{E}(X) - X \in A\big)$ for each $A \in \mathcal{B}$, the odd-order central moments are all zero.

Proof Let $\mathrm{E}(X) = 0$ without loss of generality. Write $X = |X|Z$ where

$$Z = \begin{cases} 1, & X > 0 \\ -1, & X < 0 \\ 0, & X = 0. \end{cases}$$

By assumption $P(|X|Z \in A) = P(|X|(-Z) \in A)$ for every $A \in \mathcal{B}$ and it follows that $|X|$ and Z are independently distributed. Also the case $A = [0, \infty)$ shows that $P(Z > 0) = P(Z < 0)$. Therefore $\mathrm{E}(Z^{2k+1}) = 0$ for $k = 0, 1, 2, \ldots$ and the proof is completed by noting $\mathrm{E}(X^{2k+1}) = \mathrm{E}|X|^{2k+1}\mathrm{E}(Z^{2k+1})$. ∎

9.10 Example In the two-point class (see **8.7**, **9.1**) the k^{th} central moment is

$$\mathrm{E}(X - \mathrm{E}(X))^k = (u - v)^k p(1-p)^k + (v - u)^k p^k (1-p).$$

The Bernoulli r.v. has $\mathrm{E}(X^2) = p$ and hence $\mathrm{Var}(X) = p(1-p)$. The distributions are symmetric about the mean in the cases with $p = \frac{1}{2}$, the Rademacher r.v. in particular being symmetric about zero with all even-order moments equal to 1.

9.11 Example For the binomial distribution, linearity gives $\mathrm{Var}(X) = np(1-p)$ similarly to **9.2**. For the case $p = \lambda/n$ the formula becomes $\lambda(1 - \lambda/n)$ so that for the Poisson distribution $\mathrm{Var}(X) = \lambda$ by the same reasoning as in **9.3**. □

9.12 Example The U$[a,b]$ distribution (**8.17**, **9.4**) has

$$\mathrm{E}(X^2) = \frac{1}{b-a}\int_a^b x^2 \mathrm{d}x = \frac{b^3-a^3}{3(b-a)}$$

and hence

$$\mathrm{Var}(X) = \frac{b^3-a^3}{3(b-a)} - \frac{(a+b)^2}{4} = \frac{(a-b)^2}{12}. \quad \Box$$

9.13 Example The exponential distribution (**9.5**) has, by [85], 3.351 and (9.9),

$$\mathrm{Var}(X) = \lambda\int_0^\infty x^2 \mathrm{e}^{-\lambda x}\mathrm{d}x - \frac{1}{\lambda^2} = \frac{1}{\lambda^2} \tag{9.25}$$

and hence the gamma(p,λ) has

$$\mathrm{Var}(X) = \frac{p}{\lambda^2}. \quad \Box \tag{9.26}$$

9.14 Example For the Gaussian case (**8.18**), the formula

$$\mathrm{E}(X-\mu)^k = \begin{cases} \dfrac{k!\sigma^k}{2^{k/2}(k/2)!}, & k \text{ even} \\ 0, & k \text{ odd} \end{cases} \tag{9.27}$$

may be derived after some manipulation from equation (11.22) below. All the finite-order moments exist, with $\mathrm{Var}(X) = \sigma^2$ and $\mathrm{E}(X-\mu)^4 = 3\sigma^4$ being the familiar cases. □

The existence of a moment of given order requires the existence of the corresponding *absolute moment*. If $\mathrm{E}|X|^p < \infty$ for any real $p > 0$, X may be said to belong to the set L_p (of functions Lebesgue-integrable to order p), or otherwise to be L_p-*bounded*. Taking the corresponding root of the p^{th} absolute moment is convenient for purposes of comparison (see **9.30**) and for $X \in L_p$ the L_p-*norm* of X is defined as

$$\|X\|_p = (\mathrm{E}|X|^p)^{1/p}. \tag{9.28}$$

9.15 Example For the Gaussian case (**8.18**) the k^{th} central absolute moments for $k = 1,3,5,\ldots$ are

$$\mathrm{E}|X-\mu|^k = \frac{2}{\sqrt{2\pi}\sigma}\int_0^\infty x^k \mathrm{e}^{-x^2/2\sigma^2}\mathrm{d}x = \sqrt{\frac{2^k}{\pi}}\sigma^k\Gamma((k+1)/2) \tag{9.29}$$

by (8.31) and [85], 3.462.9, with the even-order cases given by (9.27). □

While the Gaussian distribution possesses all finite-order moments, being supported on the whole of $\mathbb{R}$, its p-norms are not uniformly bounded. The formulae in (9.27) and (9.29) eventually diverge with k. If $\|X\|_p$ has a finite limit as $p \to \infty$, it coincides with the essential supremum of X, so that a random variable belonging to L_∞ is bounded almost surely.

9.5 Theorems for the Probabilist's Toolbox

The following inequalities for expected values are exploited in the proof of innumerable theorems in probability. The first is better known as *Chebyshev's inequality* for the special case $p = 2$, but generally as *Markov's inequality*.

9.16 Theorem For $\varepsilon > 0$ and $p > 0$,

$$P(|X| \geq \varepsilon) \leq \frac{\mathrm{E}|X|^p}{\varepsilon^p}. \tag{9.30}$$

Proof $\varepsilon^p P(|X| \geq \varepsilon) = \varepsilon^p \int_{|x|\geq\varepsilon} \mathrm{d}F(x) \leq \int_{|x|\geq\varepsilon} |x|^p \mathrm{d}F(x) \leq \mathrm{E}|X|^p$. ∎

This inequality does not bind unless $\mathrm{E}|X|^p/\varepsilon^p < 1$, but it shows that if $\mathrm{E}|X|^p < \infty$, the tail probabilities converge to zero at the rate ε^{-p} as $\varepsilon \to \infty$. The order of L_p-boundedness measures the tendency of a distribution to generate outliers. The Markov inequality is a special case of (at least) two more general inequalities, as follows.

9.17 Corollary For any event $A \in \mathcal{F}$,

$$\varepsilon^p \int_{A\cap\{|X|\geq\varepsilon\}} \mathrm{d}P \leq \int_A |X|^p \mathrm{d}P. \quad \square \tag{9.31}$$

Equivalently, $P\big(\{\omega : |X(\omega)| \geq \varepsilon\} \cap A\big) \leq \mathrm{E}(1_A|X|^p)/\varepsilon^p$.

9.18 Corollary Let $g : \mathbb{R} \mapsto \mathbb{R}^+$ be an integrable function with the property that $x \geq a$ implies $g(x) \geq g(a) > 0$, for a given constant a. Then

$$P(X \geq a) \leq \frac{\mathrm{E}(g(X))}{g(a)}. \tag{9.32}$$

Proof $g(a)P(X \geq a) = g(a)\int_{x \geq a} \mathrm{d}F(x) \leq \int_{x \geq a} g(x)\mathrm{d}F(x) \leq \mathrm{E}(g(X))$. ∎

An increasing function has the requisite property for all $a > 0$.

While linearity is a fundamental property of the expectation, *Jensen's inequality* is the key fact about its behaviour under nonlinear transformations.

9.19 Theorem (Jensen) If a Borel function ϕ is convex on an interval I containing the support of an integrable r.v. X where $\phi(X)$ is also integrable,

$$\phi\big(\mathrm{E}(X)\big) \leq \mathrm{E}\big(\phi(X)\big). \tag{9.33}$$

For a concave function the reverse inequality holds. □

See **2.20** for the definition of a convex function. A proof of **9.19** is obtained from the following non-probabilistic lemma. Let I° denote the interior of an interval I.

9.20 Lemma If ϕ is convex there exists a function $A(x)$ such that, for all $x \in I^{\circ}$ and $y \in I$,

$$A(x)(y - x) \leq \phi(y) - \phi(x). \tag{9.34}$$

Proof A convex function possesses right and left derivatives at all points of I°. To show this suppose first that $y > x$. Rearrange inequality (2.22) after substituting $h = y - x$ to give

$$\frac{\phi(x + \lambda h) - \phi(x)}{\lambda h} \leq \frac{\phi(x + h) - \phi(x)}{h},\ \lambda \in (0, 1]. \tag{9.35}$$

It follows that the sequence $\{n(\phi(x + 1/n) - \phi(x)), n \in \mathbb{N}\}$ is monotone decreasing and has a limit $\phi'_{+}(x)$ as $n \to \infty$. Taking the limit of (9.35) as $\lambda \downarrow 0$ gives

$$\phi'_{+}(x)(y - x) \leq \phi(y) - \phi(x). \tag{9.36}$$

Now suppose $y < x$ in (2.22) and let $h = x - y$. Reversing the inequality in (9.35) gives

$$\frac{\phi(x) - \phi(x - \lambda h)}{\lambda h} \geq \frac{\phi(x) - \phi(x - h)}{h},\ \lambda \in (0, 1]$$

which shows $\phi'_-(x)$ to exist as the limit of the increasing sequence $\{n(\phi(x-1/n)-\phi(x)),\ n\in\mathbb{N}\}$. The counterpart of (9.36) is

$$\phi'_-(x)(y-x)\le\phi(y)-\phi(x). \tag{9.37}$$

The lemma follows by either of (9.36) and (9.37). ∎

Note that $\phi'_-(x)\le\phi'_+(x)$ in all cases although which of the minorant sides of (9.36) and (9.37) dominates depends on the sign of $y-x$.

Proof of 9.19 Set $x=\mathrm{E}(X)$ and $y=X$ in (9.34) to give

$$A\big(\mathrm{E}(X)\big)\big(X-\mathrm{E}(X)\big)\le\phi(X)-\phi\big(\mathrm{E}(X)\big). \tag{9.38}$$

Taking expectations of both sides gives inequality (9.33), since the left-hand side has expectation zero. ∎

The intuition here is easily grasped by thinking about a two-point r.v. taking values x_1 with probability p and x_2 with probability $1-p$. A convex ϕ is illustrated in Figure 9.1. $\mathrm{E}(X)=px_1+(1-p)x_2$, whereas $\mathrm{E}\big(\phi(X)\big)=p\phi(x_1)+(1-p)\phi(x_2)$. This point is mapped from $\mathrm{E}(X)$ onto the vertical axis by the chord joining x_1 and x_2 on ϕ, while $\phi\big(\mathrm{E}(X)\big)$ is mapped from the same point by ϕ itself.

Next, here is an alternative approach to bounding tail probabilities that yields the Markov inequality as a corollary.

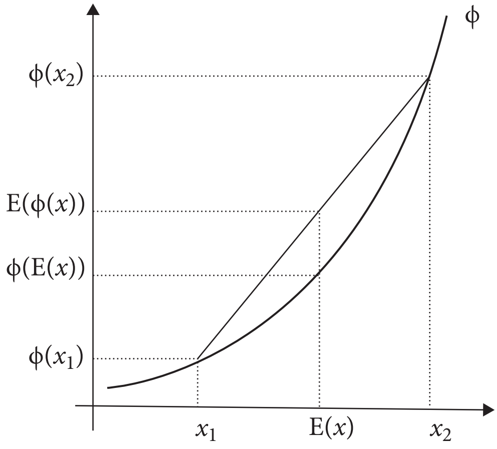

Figure 9.1

9.21 Theorem Let X be a non-negative r.v. For $r > 0$ and $x > 0$,

$$\mathrm{E}\big(X^r 1_{\{X\le x\}}\big) = r\int_0^x \xi^{r-1}P(X > \xi)\mathrm{d}\xi - x^r P(X > x). \tag{9.39}$$

Proof Let F be the c.d.f. of X. Integration by parts gives

$$\begin{aligned} r\int_0^x \xi^{r-1}(1 - F(\xi))\mathrm{d}\xi - \int_0^x \xi^r \mathrm{d}F(\xi) &= \big[\xi^r(1 - F(\xi))\big]_0^x \\ &= x^r(1 - F(x)). \quad \blacksquare \end{aligned} \tag{9.40}$$

The case where $x = \infty$ is worthy of a separate statement as follows.

9.22 Corollary If X is a non-negative r.v. and $r > 0$,

$$\mathrm{E}(X^r) = r\int_0^\infty \xi^{r-1}P(X > \xi)\mathrm{d}\xi. \tag{9.41}$$

Proof Using $x^r = r\int_0^x \xi^{r-1}d\xi$ write (9.40) as

$$\begin{aligned} \int_0^x \xi^r \mathrm{d}F(\xi) &= r\int_0^x \xi^{r-1}(1 - F(\xi))\mathrm{d}\xi - x^r(1 - F(x)) \\ &= r\int_0^x \xi^{r-1}(F(x) - F(\xi))\mathrm{d}\xi \end{aligned} \tag{9.42}$$

and let $x \to \infty$. ∎

If the left-hand side of (9.42) diverges so does the right and in this sense the theorem is true whether or not $\mathrm{E}(X^r)$ is finite.

9.23 Corollary If X is non-negative and integrable,

$$\int_\varepsilon^\infty x\mathrm{d}F = \varepsilon P(X \ge \varepsilon) + \int_\varepsilon^\infty P(X > x)\mathrm{d}x. \tag{9.43}$$

Proof Apply **9.22** with $r = 1$ to the r.v. $1_{\{X\ge\varepsilon\}}X$. Note that if $0 \le x \le \varepsilon$ the event $\{1_{\{X\ge\varepsilon\}}X > x\}$ occurs if and only if $\{X \ge \varepsilon\}$ occurs. Hence,

$$\begin{aligned}\int_{\varepsilon}^{\infty} x\mathrm{d}F &= \int_0^{\infty} P(1_{\{X\geq\varepsilon\}}X > x)\mathrm{d}x \\ &= \int_0^{\varepsilon} P(1_{\{X\geq\varepsilon\}}X > x)\mathrm{d}x + \int_{\varepsilon}^{\infty} P(X > x)\mathrm{d}x \\ &= P(X \geq \varepsilon)\int_0^{\varepsilon} \mathrm{d}x + \int_{\varepsilon}^{\infty} P(X > x)\mathrm{d}x. \quad \blacksquare \end{aligned} \tag{9.44}$$

Not only does (9.43) give the Markov inequality on replacing non-negative X by $|X|^p$ for $p > 0$ and arbitrary X, but the error in the Markov estimate of the tail probability is neatly quantified. Noting that $P(|X| \geq \varepsilon) = P(|X|^p \geq \varepsilon^p)$,

$$\begin{aligned}\varepsilon^p P(|X| \geq \varepsilon) &= \int_{\varepsilon^p}^{\infty} |x|^p \mathrm{d}F - \int_{\varepsilon^p}^{\infty} P(|X|^p > x)\mathrm{d}x \\ &= \mathrm{E}|X|^p - \int_0^{\varepsilon^p} |x|^p \mathrm{d}F - \int_{\varepsilon^p}^{\infty} P(|X|^p > x)\mathrm{d}x \end{aligned} \tag{9.45}$$

where both the subtracted terms on the right-hand side are non-negative.

9.6 Multivariate Distributions

From one point of view, the integral of a function of two or more random variables presents no special problems. For example, if

$$g : \mathbb{R}^2 \mapsto \mathbb{R}$$

is Borel-measurable, meaning in this case that $g^{-1}(B) \in \mathcal{B}^2$ for every $B \in \mathcal{B}$, then $h(\omega) = g\big(X(\omega), Y(\omega)\big)$ is just an $\mathcal{F}/\mathcal{B}$-measurable r.v. and

$$\mathrm{E}\big(g(X,Y)\big) = \int_{\Omega} h(\omega)\mathrm{d}P(\omega) \tag{9.46}$$

is its expectation, which involves no new ideas apart from the particular way in which the r.v. $h(\omega)$ happens to be defined.

Alternatively, the Lebesgue–Stieltjes form is

$$\mathrm{E}\big(g(X,Y)\big) = \int_{\mathbb{R}^2} g(x,y)\mathrm{d}F(x,y) \tag{9.47}$$

where $\mathrm{d}F(x,y)$ is to be thought of as the limiting case of $\Delta F(x,y)$ defined in (8.35) as the rectangle tends to the differential of area. When the distribution is continuous, the integral is the ordinary integral of $g(x,y)f(x,y)$ with respect to Lebesgue product measure. According to Fubini's theorem, it is equivalent to an iterated integral and may be written

$$\mathrm{E}\big(g(X,Y)\big) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} g(x,y)f(x,y)\mathrm{d}y\mathrm{d}x. \tag{9.48}$$

But caution must be exercised with formula (9.47) because this is *not* a double integral in general. The abstract notation of (9.46) is often preferable because it avoids these ambiguities.

In spite of these caveats, the expectation of a function of (say) X alone can in every case be constructed with respect to either the marginal distribution or the joint distribution.

9.24 Theorem $\mathrm{E}\big(g(X)\big) = \int_{\mathbb{R}^2} g(x)\mathrm{d}F(x,y) = \int_{\mathbb{R}} g(x)\mathrm{d}F_X(x)$.

Proof Define a function

$g^* : \mathbb{R}^2 \mapsto \mathbb{R}$

by setting $g^*(x,y) = g(x)$ for all $y \in \mathbb{R}$. $g^{*-1}(B)$ is a cylinder in $\mathbb{R}^2$ with base $g^{-1}(B) \in \mathcal{B}$ for $B \in \mathcal{B}$ and g^* is $\mathcal{B}^2/\mathcal{B}$-measurable. For non-negative g, let

$$g^*_{(n)} = \sum_{i=1}^{m}\Big(\frac{i-1}{2^n}\Big)1_{E_i} \tag{9.49}$$

where $m = n2^n + 1$ and $E_i = \big\{(x,y) : 2^{-n}(i-1) \le g^*(x,y) < 2^{-n}i\big\} \in \mathcal{B}^2$. Since $E_i = A_i \times \mathbb{R}$ where $A_i = \{x : 2^{-n}(i-1) \le g(x) < 2^{-n}i\}$ and $\mu_X(A_i) = \mu(E_i)$, $g_{(n)}$ can be defined as in formula (9.49) but with A_i replacing E_i and

$$\mathrm{E}\big(g^*_{(n)}\big) = \sum_{i=1}^{m}\Big(\frac{i-1}{2^n}\Big)\mu(E_i) = \sum_{i=1}^{m}\Big(\frac{i-1}{2^n}\Big)\mu_x(A_i) = \mathrm{E}\big(g_{(n)}\big). \tag{9.50}$$

By the monotone convergence theorem the left- and right-hand members of (9.50) converge to $\mathrm{E}(g^*) = \int g^*(x,y)\mathrm{d}F(x,y)$ and $\mathrm{E}(g) = \int g(x)\mathrm{d}F_X(x)$ respectively. Extend from non-negative to general g similarly to **9.8** to complete the proof. ∎

The means and variances of X and Y are the leading cases of this result. There are also cross moments and in particular the *covariance* of X and Y is

$$\mathrm{Cov}(X, Y) = \mathrm{E}\big(X - \mathrm{E}(X)\big)\big(Y - \mathrm{E}(Y)\big) = \mathrm{E}(XY) - \mathrm{E}(X)\mathrm{E}(Y). \tag{9.51}$$

Fubini's theorem suggests a characterization of pairwise independence.

9.25 Theorem If X and Y are independent r.v.s, $\mathrm{Cov}\big(\phi(X), \psi(Y)\big) = 0$ for all pairs of integrable Borel functions ϕ and ψ.

Proof Fubini's theorem gives

$$\begin{aligned}\mathrm{E}\big(\phi(X)\psi(Y)\big) &= \int_{\mathbb{R}^2} \phi(x)\psi(y)\mathrm{d}F(x,y) \\ &= \int_{\mathbb{R}} \phi(x)\mathrm{d}F_X(x) \int_{\mathbb{R}} \psi(y)\mathrm{d}F_Y(y) \\ &= \mathrm{E}\big(\phi(X)\big)\mathrm{E}\big(\psi(Y)\big). \quad \blacksquare \end{aligned} \tag{9.52}$$

The condition is actually sufficient as well as necessary for independence, although this cannot be shown using the present approach; see **10.26** below.

Extending from the bivariate to the general k-dimensional case adds nothing of substance to the above and is mainly a matter of appropriate notation. If $\boldsymbol{X}$ is a random k-vector,

$$\mathrm{E}(\boldsymbol{X}) = \int \boldsymbol{x}\mathrm{d}F(\boldsymbol{x}) \tag{9.53}$$

denotes the k-vector of expectations, $\mathrm{E}(X_i)$ for $i = 1, \dots, k$. The variance of a scalar r.v. generalizes to the covariance matrix of a random vector. The $k \times k$ matrix

$$\boldsymbol{XX}' = \begin{bmatrix} X_1^2 & X_1X_2 & \dots & X_1X_k \\ X_2X_1 & X_2^2 & & \vdots \\ \vdots & \ddots & \ddots & \vdots \\ X_kX_1 & \dots & \dots & X_k^2 \end{bmatrix} \tag{9.54}$$

is called the outer product of $\boldsymbol{X}$ and $\mathrm{E}(\boldsymbol{XX}')$ is the $k \times k$ positive semi-definite matrix whose elements are the expectations of the elements of $\boldsymbol{XX}'$. The *covariance matrix* of $\boldsymbol{X}$ is

$$\mathrm{Var}(\boldsymbol{X}) = \mathrm{E}\big(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})\big)\big(\boldsymbol{X} - \mathrm{E}(\boldsymbol{X})\big)' = \mathrm{E}(\boldsymbol{XX}') - \mathrm{E}(\boldsymbol{X})\mathrm{E}(\boldsymbol{X})'. \tag{9.55}$$

That $\text{Var}(\boldsymbol{X})$ is positive semi-definite generalizes the non-negative property of a scalar variance. It is of full rank (notwithstanding that $\boldsymbol{X}\boldsymbol{X}'$ has rank 1) unless an element of $\boldsymbol{X}$ is an exact linear function of the remainder. The following generalizes **4.9**, the proof being essentially an exercise interpreting the matrix formulae.

9.26 Theorem If $\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{X} + \boldsymbol{c}$ where $\boldsymbol{X}$ is a k-vector with $\text{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ and $\text{Var}(\boldsymbol{X}) = \boldsymbol{\Sigma}$ and $\boldsymbol{B}$ and $\boldsymbol{c}$ are respectively an $m \times k$ constant matrix and a constant m-vector, then

(i) $\text{E}(\boldsymbol{Y}) = \boldsymbol{B}\boldsymbol{\mu} + \boldsymbol{c}$

(ii) $\text{Var}(\boldsymbol{Y}) = \boldsymbol{B}\boldsymbol{\Sigma}\boldsymbol{B}'$. □

If $m > k$ then $\text{Var}(\boldsymbol{Y})$ is singular, having rank k.

9.27 Example If a random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_k)'$ is standard Gaussian (**8.24**) it is easy to verify, applying **9.25**, that $\text{E}(\boldsymbol{Z}) = \boldsymbol{0}$ and $\text{E}(\boldsymbol{Z}\boldsymbol{Z}') = \boldsymbol{I}_k$. Applying **9.26** to the transformation in (8.45) produces $\text{E}(\boldsymbol{X}) = \boldsymbol{\mu}$ and

$$\text{Var}(\boldsymbol{X}) = \text{E}(\boldsymbol{X} - \boldsymbol{\mu})(\boldsymbol{X} - \boldsymbol{\mu})' = \text{E}(\boldsymbol{A}\boldsymbol{Z}\boldsymbol{Z}'\boldsymbol{A}') = \boldsymbol{A}\text{E}(\boldsymbol{Z}\boldsymbol{Z}')\boldsymbol{A}' = \boldsymbol{A}\boldsymbol{A}' = \boldsymbol{\Sigma}. \qquad \square$$

A very useful fact following from the form of the multivariate Gaussian p.d.f. in (8.46) is that the distribution is completely specified by the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. Just as all the higher moments are functions of σ^2 according to formula (9.27), so the multivariate distribution is completely determined by the means, variances, and covariances. The Gaussian is the only distribution having this characteristic.

9.7 More Theorems for the Toolbox

The following collection of theorems, together with the Jensen and Markov inequalities of §9.5, constitute the basic toolbox for the proof of results in probability. The student will find that it will suffice to have his/her thumb in these pages to be able to follow a gratifyingly large number of the arguments to be encountered in subsequent chapters.

9.28 Theorem (Cauchy–Schwarz inequality)

$$\text{E}(XY)^2 \leq \text{E}(X^2)\text{E}(Y^2) \tag{9.56}$$

with equality attained when $Y = cX$, for a constant c.

Proof By linearity of the integral,

$$\mathrm{E}(aX+Y)^2 = a^2\mathrm{E}(X^2) + 2a\mathrm{E}(XY) + \mathrm{E}(Y^2) \geq 0$$

for any constant a. (9.56) follows on setting $a = -\mathrm{E}(XY)/\mathrm{E}(X^2)$ and holds as an equality if and only if $aX + Y = 0$. ∎

The *correlation coefficient* $r_{xy} = \mathrm{Cov}(X,Y)/\big(\mathrm{Var}(X)\mathrm{Var}(Y)\big)^{1/2}$ accordingly lies in the interval $[-1,+1]$. Compare this fact with the formula for sums in (2.29). By considering the case with variables $|X|^r$ and $|Y|^r$ for $r > 0$ the result may be generalized as

$$\|XY\|_r \leq \|X\|_{2r}\|Y\|_{2r}. \tag{9.57}$$

The Cauchy–Schwarz inequality is the special case with $p = 2$ of the following, whose proof has already been given for general measures as **4.18**.

9.29 Theorem (Hölder inequality) For any $p \geq 1$,

$$\mathrm{E}|XY| \leq \|X\|_p\|Y\|_q,\ q = p/(p-1).\quad \square \tag{9.58}$$

For the case $p = 1$ the inequality reduces to $\mathrm{E}|XY| \leq \mathrm{E}|X|$ ess sup Y, which is valid since $Y \leq$ ess sup Y a.s. by definition.

The following important result is the Liapunov inequality, also known as the norm inequality. It is proved here as a corollary of **9.29** but is also obtainable as a corollary of the Jensen inequality.

9.30 Theorem (Liapunov inequality) If $r > p > 0$ then $\|X\|_r \geq \|X\|_p$.

Proof Write (9.58) as $\mathrm{E}|ZY| \leq \|Z\|_s\|Y\|_{s/(s-1)}$ and let $Z = |X|^p$, $Y = 1$, and $s = r/p$ so that $\mathrm{E}(|X|^p) \leq \||X|^p\|_s$. Taking the p^{th} root of each side gives

$$\|X\|_p = \mathrm{E}(|X|^p)^{1/p} \leq \big((\mathrm{E}(|X|^{ps})\big)^{1/ps} = \|X\|_r. \quad ∎ \tag{9.59}$$

Complementing these results involving products of random variables are inequalities relating to sums. The following has already been proved for general measures as **4.19**.

9.31 Theorem (Minkowski inequality) For $r \geq 1$, $\|X+Y\|_r \leq \|X\|_r + \|Y\|_r$. □

By recursive application to the sum of m variables the Minkowski inequality generalizes directly to

$$\left\|\sum_{i=1}^{m} X_i\right\|_r \leq \sum_{i=1}^{m} \|X_i\|_r \tag{9.60}$$

for $r \geq 1$. For the case of an infinite series,

$$\left\|\sum_{i=1}^{\infty} X_i\right\|_r \leq \sum_{i=1}^{m} \|X_i\|_r + \left\|\sum_{i=m+1}^{\infty} X_i\right\|_r \tag{9.61}$$

and if $\left\|\sum_{i=m+1}^{\infty} X_i\right\|_r \to 0$ as $m \to \infty$ it is permissible to conclude that

$$\left\|\sum_{i=1}^{\infty} X_i\right\|_r \leq \sum_{i=1}^{\infty} \|X_i\|_r \tag{9.62}$$

while not ruling out the possibility that the right-hand side is infinite.

9.32 Theorem (Loève's c_r inequality) For $r > 0$,

$$\mathrm{E}\left|\sum_{i=1}^{m} X_i\right|^r \leq c_r \sum_{i=1}^{m} \mathrm{E}|X_i|^r \tag{9.63}$$

where $c_r = 1$ when $r \leq 1$ and $c_r = m^{r-1}$ when $r \geq 1$.

Proof Set the arguments of inequality (2.25) to random variables $X_1, \ldots, X_n$ and take expectations of both sides. ∎

The following inequality has more specialized applications but is relevant to arguments concerning uniform integrability in Chapters 12 and 17.

9.33 Theorem If X, Y, and Z are non-negative r.v.s satisfying $X \leq a(Y + Z)$ a.s. for a constant $a > 0$, then, for any constant $M > 0$,

$$\mathrm{E}(1_{\{X>M\}}X) \leq 2a\big(\mathrm{E}(1_{\{Y>M/2a\}}Y) + \mathrm{E}(1_{\{Z>M/2a\}}Z)\big). \tag{9.64}$$

Proof If the almost sure inequality

$$1_{\{X>M\}}X \leq 2a\big(1_{\{Y>M/2a\}}Y + 1_{\{Z>M/2a\}}Z\big) \text{ a.s.} \tag{9.65}$$

is proved the theorem follows on taking expectations. $1_{\{X>M\}}X$ is the r.v. that is equal to X if $X > M$ and 0 otherwise. If $X \leq M$ (9.65) is immediate. At least one of the inequalities $Y \geq X/2a$ and $Z \geq X/2a$ must hold and if $X > M$, (9.65) is no less obviously true. ∎

9.8 Random Variables Depending on a Parameter

Let $G(\omega,\theta) : \Omega \times \Theta \mapsto \mathbb{R}, \Theta \subseteq \mathbb{R}$ denote a random function of a real variable θ, or in other words a family of random variables indexed on points of the real line. The following results due to Cramér ([36]) are easy consequences of the dominated convergence theorem.

9.34 Theorem Suppose that for each $\omega \in C$ with $P(C) = 1$, $G(\omega,\theta)$ is continuous at a point θ_0 and $|G(\omega,\theta)| < Y(\omega)$ for each θ in an open neighbourhood N_0 of θ_0 where $\mathrm{E}(Y) < \infty$. Then

$$\lim_{\theta\to\theta_0} \mathrm{E}\big(G(\theta)\big) = \mathrm{E}\big(G(\theta_0)\big). \tag{9.66}$$

Proof Passage to a limit θ_0 through a continuum of points in Θ, as indicated in (9.66), is implied by the convergence of a countable sequence in Θ. Let $\{\theta_\nu, \nu \in \mathbb{N}\}$ be such a sequence in N_0 converging to θ_0. Putting $G_\nu(\omega) = G(\omega,\theta_\nu)$ defines a countable sequence of r.v.s and $\limsup_\nu G_\nu(\omega)$ and $\liminf_\nu G_\nu(\omega)$ are r.v.s by **3.33**. By continuity, they are equal to each other and to $G(\omega,\theta_0)$ for $\omega \in C$; in other words, $G(\theta_\nu) \to G(\theta_0)$ a.s. The result follows from the dominated convergence theorem. ∎

9.35 Theorem If for each $\omega \in C$ with $P(C) = 1$, $(\mathrm{d}G/\mathrm{d}\theta)(\omega)$ exists at a point θ_0 and

$$\left|\frac{G(\omega,\theta_0+h) - G(\omega,\theta_0)}{h}\right| < Y_1(\omega)$$

for $0 < h \leq h_1$, where $\mathrm{E}(Y_1) < \infty$ and h_1 is independent of ω, then

$$\mathrm{E}\left(\left.\frac{\mathrm{d}G}{\mathrm{d}\theta}\right|_{\theta=\theta_0}\right) = \left.\frac{\mathrm{d}}{\mathrm{d}\theta}\mathrm{E}(G)\right|_{\theta=\theta_0}. \tag{9.67}$$

Proof The argument goes like the preceding one, by considering a real sequence $\{h_\nu\}$ tending to zero through positive values and hence the sequence of r.v.s $\{H_\nu\}$ where $H_\nu = \big(G(\theta_0 + h_\nu) - G(\theta_0)\big)/h_\nu$, whose limit $H = H(\theta_0)$ exists by assumption. ∎

This is a case of 'differentiation under the integral sign'. The fact that the expected value of a derivative can be evaluated as the derivative of the expected value plays an important role in various arguments involving the limiting distribution of estimators.

The same sort of results hold for integrals. Fubini's theorem provides the extension to general double integrals. The following result for Riemann integrals on intervals of the line is no more than a special case of Fubini, but it is useful to note the requisite assumptions in common notation with the above.

9.36 Theorem Suppose that for each $\omega \in C$ with $P(C) = 1$, $G(\omega, \theta)$ is continuous on a finite open interval (a, b) and $|G(\omega, \theta)| < Y_2(\omega)$ for $a < \theta < b$ where $\mathrm{E}(Y_2) < \infty$. Then

$$\mathrm{E}\left(\int_a^b G(\theta)\mathrm{d}\theta\right) = \int_a^b \mathrm{E}\big(G(\theta)\big)\mathrm{d}\theta. \tag{9.68}$$

If $\int_0^\infty |G(\omega, \theta)|\mathrm{d}\theta < Y_3(\omega)$ for $\omega \in C$ and $\mathrm{E}(Y_3) < \infty$, (9.68) holds for either or both of $a = -\infty$ and $b = +\infty$.

Proof For the case of finite a and b, consider $H(\omega, t) = \int_a^t G(\omega, \theta)\mathrm{d}\theta$. This has the properties $|H(\omega, t)| < (b - a)Y_2(\omega)$ and $|(\mathrm{d}H/\mathrm{d}t)(\omega)| = |G(\omega, t)| < Y_2(w)$, for each $t \in (a, b)$. Hence, $\mathrm{E}(H(t))$ exists for each t and by **9.35**,

$$\frac{\mathrm{d}}{\mathrm{d}t}\mathrm{E}\big(H(t)\big) = \mathrm{E}\left(\frac{\mathrm{d}H}{\mathrm{d}t}\right) = \mathrm{E}\big(G(t)\big). \tag{9.69}$$

$\mathrm{E}\big(G(t)\big)$ is continuous on (a, b) by the a.s. continuity of G and **9.34** and hence

$$\Delta(t) = \int_a^t \mathrm{E}\big(G(\theta)\big)\mathrm{d}\theta - \mathrm{E}\big(H(t)\big) \tag{9.70}$$

is differentiable on (a, b) and $\mathrm{d}\Delta/\mathrm{d}t = 0$ at each point by (9.69). Since $H(\omega, a) = 0$ for $\omega \in C$ by definition, implying $\Delta(a) = 0$, it follows that $\Delta(b) = 0$. This is equivalent to (9.68), completing the proof for finite a and b.

Otherwise, under the stated integrability condition on $G(\omega, \theta)$, $\int_{-\infty}^\infty G(\omega, \theta)\mathrm{d}\theta$ exists and is finite on C. Hence $H(\omega, t) = \int_{-\infty}^t G(\omega, \theta)\mathrm{d}\theta$ is well defined and has an expectation for all $t \in \mathbb{R}$ and the argument above goes through with $a = -\infty$ and/or $b = \infty$. ∎

10
Conditioning

10.1 Conditioning in Product Measures

It is difficult to do justice to conditioning at an elementary level. Without resort to some measure-theoretic insights, one can get only so far with the theory before running into problems. There are nonetheless some relatively simple results that apply to a restricted (albeit important) class of distributions. There is something to be said for introducing the topic by way of this 'naïve' approach and so demonstrating the difficulties that arise, before going on to see how they can be resolved.

In the bivariate context, the natural question to pose is usually: "if we know $X = x$, what is the best predictor of Y?" For a random real pair $\{X, Y\}$ on $(\Omega, \mathcal{F}, P)$ a class of conditional distribution functions for Y might be defined (see §7.2). For any $A \in \mathcal{B}$ such that $P(X \in A) > 0$, let

$$F(y|X \in A) = \frac{P(X \in A, Y \leq y)}{P(X \in A)}. \tag{10.1}$$

This corresponds to the idea of working in the trace of $(\Omega, \mathcal{F}, P)$ with respect to A, once A is known to have occurred. Proceeding in this way, a theory of conditioning for random variables based on the c.d.f. might be constructed. The conditional distribution function $F(y|x)$, when it exists, might be defined as a mapping from $\mathbb{R}^2$ to $\mathbb{R}$ which for fixed $x \in \mathbb{R}$ is a non-decreasing, right-continuous function of y with $F(-\infty|x) = 0$ and $F(+\infty|x) = 1$ and for fixed $y \in \mathbb{R}$ and any $A \in \mathcal{B}$ satisfies the equation

$$P(X \in A, Y \leq y) = \int_A F(y|x) \mathrm{d}F_X(x). \tag{10.2}$$

(Compare Rao [155], §2a.8.) Think of the graph of $F(y|x)$ in y-space as the profile of a 'slice' through the surface of the joint distribution function, parallel to the y-axis, at the point x.

However, much care is needed in interpreting this construction. Unlike the ordinary c.d.f. it does not represent a probability in general. Trying to interpret it as $P(Y \leq y|X = x)$ runs into difficulty over the case $P(X = x) = 0$, as when the distribution of X is continuous. Since the integral of $F(y|x)$ over a set in the

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0010

marginal distribution of X yields a probability, as in (10.2), it might even be treated as a type of density function. Taking $A = \{X \le x\}$ leads to the representation

$$F(x,y) = \int_{-\infty}^{x} F(y|\xi)\mathrm{d}F_X(\xi) = \int_{-\infty}^{x}\int_{-\infty}^{y} \mathrm{d}F(v|\xi)\mathrm{d}F_X(\xi). \tag{10.3}$$

Since $F(x,y)$ is an integral over $\mathbb{R}^2$, Fubini's theorem implies that $F(y|x)$ is well defined *only* when the integrals in (10.3) are with respect to a product measure.

If X and Y are independent then unambiguously, but not very usefully, $F(y|x) = F_Y(y)$. $F(y|x)$ is also well defined for continuous distributions. Let S_X denote the support of X (the set on which $f_X > 0$). The conditional p.d.f. is

$$f(y|x) = \frac{f(x,y)}{f_X(x)},\ x \in S_X \tag{10.4}$$

where $f_X(x)$ is the marginal density of X. For $A \in \mathcal{B} \cap \mathcal{S}_X$ it is valid to write

$$P(X \in A, Y \le y) = \int_A \int_{-\infty}^{y} f(v|x)f_X(x)\mathrm{d}x\mathrm{d}v = \int_A F(y|x)f_X(x)\mathrm{d}x \tag{10.5}$$

where

$$F(y|x) = \int_{-\infty}^{y} f(v|x)\mathrm{d}v. \tag{10.6}$$

The second equality of (10.5) follows by Fubini's theorem, since the function $f(x,y)$ is integrated with respect to Lebesgue product measure. However, (10.6) appears to exist by a trick, rather than to have a firm relationship with intuition. The problem is working with the trace $(\Lambda, \mathcal{F}_\Lambda, P_\Lambda)$ when $\Lambda = \{\omega : X(\omega) = x\}$ and $P(\Lambda) = 0$, because then $P_\Lambda = P/P(\Lambda)$ is undefined. It is not clear what it means to "consider the case when $\{X = x\}$ has occurred" when this event fails to occur almost surely.

Except in special cases such as the above, the factorization $\mathrm{d}F(y,x) = \mathrm{d}F(y|x)$ $\mathrm{d}F_X(x)$ is not legitimate, but with this very important caveat the mean and other moments of the conditional distribution can be defined. The *conditional expectation* of a measurable function $g(X, Y)$, given $X = x$, can be defined as

$$\mathrm{E}\big(g(X,Y)|x\big) = \int_{-\infty}^{+\infty} g(x,y)\mathrm{d}F(y|x), \tag{10.7}$$

also written as $\mathrm{E}(g(X,Y)|X = x)$. The simplest case is where $g(X,Y)$ is just Y. $\mathrm{E}(Y|x)$ is to be understood in terms of the attempt of an observer to predict Y

after the realization of X has been observed. When X and Y are independent, $\mathrm{E}(Y|x) = \mathrm{E}(Y)$ where $\mathrm{E}(Y)$ is the ordinary expectation of Y, also called the marginal or unconditional expectation. In this case, the knowledge that $X = x$ is no help in predicting Y.

10.1 Example These concepts apply to the bivariate Gaussian distribution. From (8.46), the density is

$$
\begin{aligned}
&f(x,y)\\
&= \frac{1}{2\pi\begin{vmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{12} & \sigma_{22}\end{vmatrix}^{1/2}} \exp\left\{-\tfrac{1}{2}[x-\mu_1 \; y-\mu_2]\begin{bmatrix}\sigma_{11} & \sigma_{12}\\ \sigma_{12} & \sigma_{22}\end{bmatrix}^{-1}\begin{bmatrix}x-\mu_1\\ y-\mu_2\end{bmatrix}\right\}\\
&= \frac{1}{2\pi(\sigma_{11}\sigma_{22}-\sigma_{12}^2)^{1/2}} \exp\left\{-\frac{\left(y-\mu_2-\frac{\sigma_{12}}{\sigma_{11}}(x-\mu_1)\right)^2}{2\left(\sigma_{22}-\frac{\sigma_{12}^2}{\sigma_{11}}\right)} - \frac{(x-\mu_1)^2}{2\sigma_{11}}\right\} \qquad (10.8)
\end{aligned}
$$

where the last equality is got by completing the square in the exponent. Evidently, $f(x,y) = f(y|x)f_X(x)$ where

$$
f(y|x) = \frac{1}{\sqrt{2\pi}\left(\sigma_{22}-\frac{\sigma_{12}^2}{\sigma_{11}}\right)^{1/2}} \exp\left\{-\frac{\left((y-\mu_2)-\frac{\sigma_{12}}{\sigma_{11}}(x-\mu_1)\right)^2}{2\left(\sigma_{22}-\frac{\sigma_{12}^2}{\sigma_{11}}\right)}\right\} \qquad (10.9)
$$

and

$$
f_X(x) = \frac{1}{\sqrt{2\pi\sigma_{11}}}\exp\left\{-\frac{(x-\mu_1)^2}{2\sigma_{11}}\right\}. \qquad (10.10)
$$

Thus,

$$
\mathrm{E}(Y|x) = \mu_2 + \frac{\sigma_{12}}{\sigma_{22}}(x-\mu_1) \qquad (10.11)
$$

and

$$
\mathrm{Var}(Y|x) = \sigma_{22} - \frac{\sigma_{12}^2}{\sigma_{11}}. \qquad (10.12)
$$

If $\sigma_{12} = 0$, $f(y|x)$ reduces to $f_Y(y)$, so that the joint density is the product of the marginals and x and y are independent. □

10.2 Conditioning on a Sigma Field

In view of the limitations of working directly with the distribution of (X, Y) consider the approach introduced in §7.2 of representing partial knowledge of the distribution of Y by specifying a σ-field of events $\mathcal{G} \subseteq \mathcal{F}$ such that, for each $G \in \mathcal{G}$, an observer knows whether or not the realized outcome belongs to G.

The idea of knowing the value of a random variable is captured by the concept of sub-σ-field measurability. A random variable $X(\omega) : \Omega \mapsto \mathbb{R}$ is said to be measurable with respect to a σ-field $\mathcal{G} \subset \mathcal{F}$ if

$$X^{-1}(B) = \{\omega : X(\omega) \in B\} \in \mathcal{G}, \text{ for all } B \in \mathcal{B}. \tag{10.13}$$

The implication of the condition $\mathcal{G} \subset \mathcal{F}$ is that the r.v. X is not a complete representation of the random outcome ω. Denote by $\sigma(X)$ the intersection of all σ-fields with respect to which X is measurable, called the σ-field generated by X. If given the distribution of a random pair $(X(\omega), Y(\omega))$ it is known that $X = x$, to determine whether of not an event $G \in \sigma(X)$ has occurred is a matter of checking whether the set $X(G)$ contains x. The image of each $G \in \sigma(X)$ under the mapping

$$\big(X(\omega), Y(\omega)\big) : \Omega \mapsto \mathbb{R}^2$$

is a cylinder set in $\mathbb{R}^2$ and the p.m. defined on $\sigma(X)$ is the marginal distribution of X.

10.2 Example The knowledge that $x_1 \le X \le x_2$ is represented by

$$\mathcal{H} = \sigma\big(\{-\infty, x] : x < x_1\}, \{[x, \infty) : x > x_2\}, \mathbb{R}\big).$$

For every element of this σ-field, it is known whether or not X belongs to the set. Also, it contains all the sets about which this is known. The closer together x_1 and x_2 are, the more sets there are in $\mathcal{H}$. When $x_1 = x_2$, $\mathcal{H} = \sigma(X)$ and at the other extreme when $x_1 = -\infty$, $x_2 = +\infty$, $\mathcal{H} = \mathcal{T} = \{\emptyset, \mathbb{R}\}$. □

The relationships between transformations and sub-σ-field measurability are summarized in the next theorem, of which the first part is an easy consequence of the definitions but the second is trickier. If two random variables are measurable with respect to the same sub-σ-field, the implication is that they contain the same information; knowledge of one is equivalent to knowledge of the other. If g denotes the transformation in question, this means that *every* Borel set is the image of a Borel set under g^{-1}. This is a stronger condition than measurability and requires that g be an isomorphism. It suffices for g to be a homeomorphism, although this is not necessary as was shown in **3.30**.

10.3 Theorem Let X be a r.v. on the space $(\mathbb{S},\mathcal{B}_{\mathbb{S}},\mu)$ and let $Y=g(X)$ where $g:\mathbb{S}\mapsto\mathbb{T}$ is a Borel function, with $\mathbb{S}\subseteq\mathbb{R}$ and $\mathbb{T}\subseteq\mathbb{R}$.

(i) $\sigma(Y)\subseteq\sigma(X)$.

(ii) $\sigma(Y)=\sigma(X)$ iff g is a Borel-measurable isomorphism.

Proof Each $B\in\mathcal{B}_{\mathbb{T}}$ has an image in $\mathcal{B}_{\mathbb{S}}$ under g^{-1}, which in turn has an image in $\sigma(X)$ under X^{-1}. This proves (i). To prove (ii), define a class of subsets of $\mathbb{S}$, $\mathcal{C}=\{g^{-1}(B):B\in\mathcal{B}_{\mathbb{T}}\}$. To every $A\subseteq\mathbb{S}$ there corresponds (since g is a mapping) a set $B\subseteq\mathbb{T}$ such that $A=g^{-1}(B)$. Making this substitution gives

$$\mathcal{B}_{\mathbb{S}}\subseteq\{A:g(A)\in\mathcal{B}_{\mathbb{T}}\}=\{g^{-1}(B):g(g^{-1}(B))\in\mathcal{B}_{\mathbb{T}}\}=\mathcal{C}\qquad(10.14)$$

where the inclusion is by measurability of g^{-1} and the equality is because $g(g^{-1}(B))=B$ for any $B\subseteq\mathbb{T}$, since g is 1–1 onto. It follows from (10.14) that

$$\{X^{-1}(A)\subseteq\Omega:A\in\mathcal{B}_{\mathbb{S}}\}\subseteq\{X^{-1}(g^{-1}(B))\subseteq\Omega:B\in\mathcal{B}_{\mathbb{T}}\}.\qquad(10.15)$$

If Y is $\mathcal{G}$-measurable for some σ-field $\mathcal{G}\subseteq\mathcal{F}$ (such that $\mathcal{G}$ contains the sets of the right-hand member of (10.15)), then X is also $\mathcal{G}$-measurable. In particular, $\sigma(X)\subseteq\sigma(Y)$. Part (i) then implies $\sigma(X)=\sigma(Y)$, proving sufficiency of the conditions.

To show the necessity, suppose first that g is not 1–1 and $g(x_1)=g(x_2)=y$ (say) for $x_1\neq x_2$. The sets $\{x_1\}$ and $\{x_2\}$ are elements of $\mathcal{B}_{\mathbb{S}}$ but not of $\mathcal{C}$ which contains only $g^{-1}(\{y\})=\{x_1\}\cup\{x_2\}$. Hence, $\mathcal{B}_{\mathbb{S}}\nsubseteq\mathcal{C}$ and $\exists$ a $\mathcal{B}_{\mathbb{S}}$-set A for which there is no $\mathcal{B}_{\mathbb{T}}$-set B having the property $g^{-1}(B)=A$. This implies that $X^{-1}(A)\in\sigma(X)$ but $\notin\sigma(Y)$, so that $\sigma(Y)\subset\sigma(X)$. Therefore, g is 1–1. If g^{-1} is not Borel-measurable, then by definition $\exists\,A=g^{-1}(B)\in\mathcal{B}_{\mathbb{S}}$ such that $g(A)=B\notin\mathcal{B}_{\mathbb{T}}$ and hence $A\notin\mathcal{C}$; and again, $\mathcal{B}_{\mathbb{S}}\nsubseteq\mathcal{C}$ so that $\sigma(Y)\subset\sigma(X)$ by the same argument. This completes the proof of necessity. ∎

Briefly note the generalization of these results to the vector case. A random vector $\boldsymbol{X}(\omega):\Omega\mapsto\mathbb{R}^k$ is measurable with respect to $\mathcal{G}\subseteq\mathcal{F}$ if

$$\boldsymbol{X}^{-1}(B)=\{\omega:\boldsymbol{X}(\omega)\in B\}\in\mathcal{G},\ \forall B\in\mathcal{B}^k.\qquad(10.16)$$

If $\sigma(\boldsymbol{X})$ is the σ-field generated by $\boldsymbol{X}$ there is the following result.

10.4 Theorem Let $\boldsymbol{X}$ be a random vector on the probability space $(\mathbb{S},\mathcal{B}^k_{\mathbb{S}},\mu)$ where $\mathbb{S}\subseteq\mathbb{R}^k$ and $\mathcal{B}^k_{\mathbb{S}}=\{B\cap\mathbb{S}:B\in\mathcal{B}^k\}$ and consider a Borel function

$$\boldsymbol{Y}=\boldsymbol{g}(\boldsymbol{X}):\mathbb{S}\mapsto\mathbb{T},\ \mathbb{T}\subseteq\mathbb{R}^m.\qquad(10.17)$$

(i) $\sigma(\boldsymbol{Y}) \subseteq \sigma(\boldsymbol{X})$.

(ii) If $m = k$ and $\boldsymbol{g}$ is 1–1 with Borel-measurable inverse, then $\sigma(\boldsymbol{Y}) = \sigma(\boldsymbol{X})$.

Proof This follows the proof of **10.3** almost word for word, with the substitutions of $\mathcal{B}_{\mathbb{S}}^k$ and $\mathcal{B}_{\mathbb{T}}^k$ for $\mathcal{B}_{\mathbb{S}}$ and $\mathcal{B}_{\mathbb{T}}$, $\boldsymbol{X}$ and $\boldsymbol{Y}$ for X and Y, and so forth. ∎

10.3 Conditional Expectations

Let Y be an integrable r.v. on $(\Omega, \mathcal{F}, P)$ and $\mathcal{G}$ a σ-field contained in $\mathcal{F}$. The term *conditional expectation* and symbol $\mathrm{E}(Y|\mathcal{G})$ can be used to refer to any integrable, $\mathcal{G}$-measurable r.v. having the property

$$\int_G \mathrm{E}(Y|\mathcal{G})\mathrm{d}P = \int_G Y\mathrm{d}P = \mathrm{E}(Y|G)P(G), \text{ all } G \in \mathcal{G}. \tag{10.18}$$

Intuitively, $\mathrm{E}(Y|\mathcal{G})(\omega)$ represents the prediction of $Y(\omega)$ made by an observer having information $\mathcal{G}$, when the outcome ω is realized. The second equality of (10.18) defines the constant $\mathrm{E}(Y|G)$, when $P(G) > 0$, as the expected value of Y in the trace of $(\Omega, \mathcal{F}, P)$ on G (see §7.1). The two extreme cases are $\mathrm{E}(Y|\mathcal{F}) = Y$ a.s. and $\mathrm{E}(Y|\mathcal{T}) = \mathrm{E}(Y)$ a.s., where $\mathcal{T}$ denotes the trivial σ-field with elements $\{\Omega, \emptyset\}$. Note that $\Omega \in \mathcal{G}$ so integrability of Y is necessary for the existence of $\mathrm{E}(Y|\mathcal{G})$.

The conditional expectation is a slightly bizarre construction, not only a r.v., but evidently not even an integral. To demonstrate that an object satisfying (10.18) actually does exist, consider initially the case $Y \geq 0$ and define

$$\nu(G) = \int_G Y\mathrm{d}P. \tag{10.19}$$

10.5 Theorem ν is a measure and is absolutely continuous with respect to P.

Proof Clearly, $\nu(G) \geq 0$ and $P(G) = 0$ implies $\nu(G) = 0$. To show countable additivity, let $\{G_j\}$ be a disjoint sequence. Then

$$\nu\left(\bigcup_j G_j\right) = \int_{\bigcup_j G_j} Y\mathrm{d}P = \sum_j \int_{G_j} Y\mathrm{d}P = \sum_j \nu(G_j) \tag{10.20}$$

where the second equality holds under disjointness. ∎

The implication of (10.18) for non-negative Y turns out to be that $\mathrm{E}(Y|\mathcal{G})$ is the Radon–Nikodym derivative of ν with respect to P. The extension from non-negative to general Y is easy, since $Y = Y^+ - Y^-$ where Y^+ and Y^- are non-negative and, from (10.18), $\mathrm{E}(Y|\mathcal{G}) = \mathrm{E}(Y^+|\mathcal{G}) - \mathrm{E}(Y^-|\mathcal{G})$ where both of the right-hand r.v.s are Radon–Nikodym derivatives.

The Radon–Nikodym theorem therefore establishes the existence of $\mathrm{E}(Y|\mathcal{G})$; at any rate, it establishes the existence of at least one r.v. satisfying (10.18). It does not guarantee that there is only one such r.v. and in the event of non-uniqueness one speaks of the different *versions* of $\mathrm{E}(Y|\mathcal{G})$. The possibility of multiple versions is rarely of practical concern since **10.5** provides the assurance that they are all equal to one another a.s.$[P]$, but it does make it necessary to qualify any statement about conditional expectations with the tag 'a.s.', to indicate that there may be sets of measure zero on which the assertions do not apply.

In the bivariate context $\mathrm{E}(Y|\sigma(X))$, written as $\mathrm{E}(Y|X)$ when the context is clear, is interpreted as the prediction of Y made by observers who observe X. This notion is related to (10.7) by thinking of $\mathrm{E}(Y|x)$ as a drawing from the distribution of $\mathrm{E}(Y|X)$.

10.6 Example In place of (10.11), write

$$\mathrm{E}(Y|X)(\omega) = \mu_2 + \frac{\sigma_{12}}{\sigma_{11}}(X(\omega) - \mu_1) \tag{10.21}$$

which as a function of $X(\omega)$ is a r.v. derived from the marginal distribution of X. $\mathrm{E}(Y|X)$ is Gaussian with mean μ_2 and variance $\sigma_{12}^2/\sigma_{11}$. □

Making $\mathrm{E}(Y|x)$ a point in a probability space on $\mathbb{R}$ circumvents the difficulty encountered previously with conditioning on events of probability 0. The construction is valid for all distributions and it is possible to define $\mathrm{E}(Y|G)$ when $P(G) = 0$. What is required is to exhibit a decreasing sequence $\{G_n \in \mathcal{G}\}$ with $P(G_n) > 0$ for every n and $G_n \downarrow G$, such that the real sequence $\{\mathrm{E}(Y|G_n)\}$ converges. This is why (10.7) works for continuous distributions. Take $G_n = [x, x + 1/n] \times \mathbb{R} \in \sigma(X)$ so that $G = \{x\} \times \mathbb{R}$. Using (10.4) in (10.18),

$$\mathrm{E}(Y|G_n) = \frac{\displaystyle\int_{-\infty}^{+\infty}\int_{x}^{x+1/n} yf(\xi, y)\mathrm{d}\xi\mathrm{d}y}{\displaystyle\int_{-\infty}^{+\infty}\int_{x}^{x+1/n} f(\xi, y)\mathrm{d}\xi\mathrm{d}y} \to \int_{-\infty}^{+\infty} yf(y|x)\mathrm{d}y = \mathrm{E}(Y|x) \tag{10.22}$$

as $n \to \infty$. Fubini's theorem allows these double integrals to be evaluated one dimension at a time. The limits with respect to n are taken inside the integrals with respect to y.

Conditional probability can sometimes generate paradoxical results, as the following case demonstrates.

10.7 Example Let Y be a drawing from the space $\big([0,1],\mathcal{B}_{[0,1]},m\big)$ where m is Lebesgue measure. Let $\mathcal{G}\subset\mathcal{B}_{[0,1]}$ denote the σ-field generated from the singletons $\{y\}$, $y\in[0,1]$. All countable unions of singletons have measure 0, while all the complements have measure 1. Since for each $G\in\mathcal{G}$ either $P(G)=0$ or $P(G)=1$, setting $\mathrm{E}(Y|\mathcal{G})=\mathrm{E}(Y)=\frac{1}{2}$ a.s. satisfies (10.18), yielding $0=0$ in the first case and $\frac{1}{2}=\frac{1}{2}$ in the second. However, since $\{y\}\in\mathcal{G}$, if it is known whether or not $y\in G$ for each $G\in\mathcal{G}$ then y is known. Since $\mathcal{G}$ contains knowledge of the outcome it should be the case that $\mathrm{E}(Y|\mathcal{G})=Y$ a.s. □

The mathematics are unambiguous but there is evidently some difficulty with the idea that a sub-σ-field must always represent partial knowledge. The mathematical model may sometimes part company with intuition and generate paradoxical results. Whether it is the model or the intuition that fails is a nice point for debate.

10.4 Some Theorems on Conditional Expectations

The key step in many arguments involving conditional expectations is the fundamental *law of iterated expectations*, often abbreviated to LIE in the sequel.

10.8 Theorem

$$\mathrm{E}\big(\mathrm{E}(Y|\mathcal{G})\big)=\mathrm{E}(Y). \tag{10.23}$$

Proof Immediate from (10.18), setting $G=\Omega$. ■

It is often found that the easiest way to calculate an expected value is to nest a known conditional mean inside the unconditional mean and apply the LIE.

The intuitive idea that conditioning variables can be held 'as if constant' under the conditional distribution is confirmed by the following pair of results.

10.9 Theorem If X is integrable and $\mathcal{G}$-measurable, then $\mathrm{E}(X|\mathcal{G})=X$ a.s.

Proof Since X is $\mathcal{G}$-measurable, $E^+=\big\{\omega:X(\omega)>\mathrm{E}(X|\mathcal{G})(\omega)\big\}\in\mathcal{G}$. If $P(E^+)>0$, then

$$\int_{E^+}X\mathrm{d}P-\int_{E^+}\mathrm{E}(X|\mathcal{G})\mathrm{d}P=\int_{E^+}\big(X-\mathrm{E}(X|\mathcal{G})\big)\mathrm{d}P>0. \tag{10.24}$$

This contradicts (10.18), so $P(E^+) = 0$. By the same argument, $P(E^-) = 0$ where $E^- = \{\omega : X(\omega) < \mathrm{E}(X|\mathcal{G})(\omega)\} \in \mathcal{G}$. ∎

The second result shows how the linearity property of expected values is preserved. Of course, linearity in the conventional sense of

$$\mathrm{E}(aX + bY|\mathcal{G}) = a\mathrm{E}(X|\mathcal{G}) + b\mathrm{E}(Y|\mathcal{G}) \text{ a.s.} \tag{10.25}$$

is a direct consequence of the definition in (10.18). However, $\mathcal{G}$-measurable random variables also behave like constants under the conditional distribution, in the following way.

10.10 Theorem If Y is $\mathcal{F}$-measurable and integrable, X is $\mathcal{G}$-measurable for $\mathcal{G} \subseteq \mathcal{F}$, and $\mathrm{E}|XY| < \infty$, then $\mathrm{E}(XY|\mathcal{G}) = X\mathrm{E}(Y|\mathcal{G})$ a.s.

Proof By definition, the theorem follows if

$$\int_G X\mathrm{E}(Y|\mathcal{G})\mathrm{d}P = \int_G XY\mathrm{d}P \text{ for all } G \in \mathcal{G}. \tag{10.26}$$

Let $X_{(n)} = \sum_{i=1}^n \alpha_i 1_{E_i}$ be a $\mathcal{G}$-measurable simple r.v. with $E_1, \ldots, E_n$ a partition of Ω and $E_i \in \mathcal{G}$ for each i. (10.26) holds for $X = X_{(n)}$ since, for all $G \in \mathcal{G}$,

$$\begin{aligned}\int_G X_{(n)}\mathrm{E}(Y|\mathcal{G})\mathrm{d}P &= \sum_{i=1}^n \alpha_i \int_{G\cap E_i} \mathrm{E}(Y|\mathcal{G})\mathrm{d}P \\ &= \sum_{i=1}^n \alpha_i \int_{G\cap E_i} Y\mathrm{d}P = \int_G X_{(n)} Y\mathrm{d}P\end{aligned} \tag{10.27}$$

noting $G \cap E_i \in \mathcal{G}$ when $G \in \mathcal{G}$ and $E_i \in \mathcal{G}$.

For the case $X \geq 0$ a.s. let $\{X_{(n)}\}$ be a monotone sequence of simple $\mathcal{G}$-measurable functions converging to X as in **3.35**. Then $X_{(n)}Y \to XY$ a.s. and $|X_{(n)}Y| \leq |XY|$, where $\mathrm{E}|XY| < \infty$ by assumption. Similarly, $X_{(n)}\mathrm{E}(Y|\mathcal{G}) \to X\mathrm{E}(Y|\mathcal{G})$ a.s. and

$$\begin{aligned}\mathrm{E}|X_{(n)}\mathrm{E}(Y|\mathcal{G})| = \mathrm{E}|\mathrm{E}(X_{(n)}Y|\mathcal{G})| &\leq \mathrm{E}\big(\mathrm{E}(|X_{(n)}Y||\mathcal{G})\big) \\ &= \mathrm{E}|X_{(n)}Y| \leq \mathrm{E}|XY| < \infty\end{aligned} \tag{10.28}$$

where the first equality is (10.27), the inequality is the conditional modulus inequality shown in **10.15** below, and the second equality is the law of iterated expectations (LIE). It follows by the dominated convergence theorem that

$$\int_G X_{(n)}\mathrm{E}(Y|\mathcal{G})\mathrm{d}P \to \int_G X\mathrm{E}(Y|\mathcal{G})\mathrm{d}P$$

and so (10.26) holds for non-negative X. The extension to general $\mathcal{G}$-measurable X is got by putting

$$X = X^+ - X^- \tag{10.29}$$

where $X^+ = \max\{X, 0\} \geq 0$ and $X^- \geq 0$. Using (10.25) and the result for non-negative X yields

$$\begin{aligned} \mathrm{E}(YX|\mathcal{G}) = \mathrm{E}(YX^+ - YX^-|\mathcal{G}) &= \mathrm{E}(YX^+|\mathcal{G}) - \mathrm{E}(YX^-|\mathcal{G}) \\ &= (X^+ - X^-)\mathrm{E}(Y|\mathcal{G}) \\ &= X\mathrm{E}(Y|\mathcal{G}) \text{ a.s.} \quad \blacksquare \end{aligned} \tag{10.30}$$

X does not need to be integrable for the linearity result to hold, but the following is an important application to integrable X.

10.11 Theorem If Y is $\mathcal{F}$-measurable and integrable and $\mathrm{E}(Y|\mathcal{G}) = \mathrm{E}(Y)$ for $\mathcal{G} \subseteq \mathcal{F}$ then $\mathrm{Cov}(X, Y) = 0$ for integrable, $\mathcal{G}$-measurable X.

Proof From **10.8** and **10.10**

$$\mathrm{E}(XY) = \mathrm{E}\big(\mathrm{E}(XY|\mathcal{G})\big) = \mathrm{E}\big(X\mathrm{E}(Y|\mathcal{G})\big). \tag{10.31}$$

If $\mathrm{E}(Y|\mathcal{G}) = \mathrm{E}(Y)$ a.s. (a constant) then $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$. $\blacksquare$

In general $\mathrm{Cov}(X, Y)$ is defined only for *square* integrable r.v.s X and Y, but $\mathrm{Cov}(X, Y) = 0$, equivalently $\mathrm{E}(XY) = \mathrm{E}(X)\mathrm{E}(Y)$, is a property that an integrable pair can satisfy.

The following is the result that justifies the characterization of the conditional mean as the optimal predictor of Y given partial information. 'Optimal' is seen to have the specific connotation of minimizing the mean of the squared prediction errors.

10.12 Theorem Let $\hat{Y}$ denote any $\mathcal{G}$-measurable approximation to Y. Then

$$||Y - \mathrm{E}(Y|\mathcal{G})||_2 \leq ||Y - \hat{Y}||_2. \tag{10.32}$$

Proof

$$\begin{aligned} (Y - \hat{Y})^2 &= \big(Y - \mathrm{E}(Y|\mathcal{G}) + \mathrm{E}(Y|\mathcal{G}) - \hat{Y}\big)^2 \\ &= \big(Y - \mathrm{E}(Y|\mathcal{G})\big)^2 + \big(\mathrm{E}(Y|\mathcal{G}) - \hat{Y}\big)^2 + 2\big(Y - \mathrm{E}(Y|\mathcal{G})\big)\big(\mathrm{E}(Y|\mathcal{G}) - \hat{Y}\big) \end{aligned}$$

and hence

$$\mathrm{E}\big((Y - \hat{Y})^2|\mathcal{G}\big) = \mathrm{E}\big((Y - \mathrm{E}(Y|\mathcal{G}))^2|\mathcal{G}\big) + \big(\mathrm{E}(Y|\mathcal{G}) - \hat{Y}\big)^2 \text{ a.s.} \tag{10.33}$$

noting that the conditional expectation of the cross-product disappears by definition of $\mathrm{E}(Y|\mathcal{G})$ and **10.10**. The proof is completed by taking unconditional expectations through (10.33) and using the LIE. ∎

The next result is known as the *law of total variance.*

10.13 Theorem $\mathrm{Var}(Y) = \mathrm{E}\big(\mathrm{Var}(Y|\mathcal{G})\big) + \mathrm{Var}\big(\mathrm{E}(Y|\mathcal{G})\big)$.

Proof

$$\begin{aligned}\mathrm{Var}(Y) &= \mathrm{E}\big((Y-\mathrm{E}(Y))^2\big)\\ &= \mathrm{E}\big((Y-\mathrm{E}(Y|\mathcal{G})+\mathrm{E}(Y|\mathcal{G})-\mathrm{E}(Y))^2\big)\\ &= \mathrm{E}\big(\mathrm{E}((Y-\mathrm{E}(Y|\mathcal{G}))^2|\mathcal{G})\big)+\mathrm{E}\big((\mathrm{E}(Y|\mathcal{G})-\mathrm{E}(Y))^2\big)\\ &\qquad +2\mathrm{E}\big((\mathrm{E}(Y|\mathcal{G})-\mathrm{E}(Y))(Y-\mathrm{E}(Y|\mathcal{G}))\big)\end{aligned}$$

and by the LIE,

$$\begin{aligned}&\mathrm{E}\big((\mathrm{E}(Y|\mathcal{G})-\mathrm{E}(Y))(Y-\mathrm{E}(Y|\mathcal{G}))\big)\\ &\qquad= \mathrm{E}\big((\mathrm{E}(Y|\mathcal{G})-\mathrm{E}(Y))\mathrm{E}((Y-\mathrm{E}(Y|\mathcal{G}))|\mathcal{G})\big)=0\end{aligned}$$

since

$$\mathrm{E}\big((Y-\mathrm{E}(Y|\mathcal{G}))|\mathcal{G}\big)=0 \text{ a.s.} \quad ∎$$

The foregoing are results that have no counterpart in ordinary integration theory, but in many respects the conditional expectation behaves like a 'real' expectation, apart from the standard caveat that different behaviour is possible on sets of measure zero. The following are conditional versions of various results in Chapters 4 and 9. The first extends **4.7** and **4.16**.

10.14 Lemma

(i) If $X = 0$ a.s. then $\mathrm{E}(X|\mathcal{G}) = 0$ a.s.

(ii) If $X \le Y$ a.s. then $\mathrm{E}(X|\mathcal{G}) \le \mathrm{E}(Y|\mathcal{G})$ a.s.

(iii) If $X = Y$ a.s. then $\mathrm{E}(X|\mathcal{G}) = \mathrm{E}(Y|\mathcal{G})$ a.s.

Proof (i) follows directly from (10.18). To prove (ii) note that the hypothesis (10.18) and **4.11**(i) together imply

$$\int_G \mathrm{E}(X|\mathcal{G})\mathrm{d}P = \int_G X\mathrm{d}P \le \int_G Y\mathrm{d}P = \int_G \mathrm{E}(Y|\mathcal{G})\mathrm{d}P$$

for all $G \in \mathcal{G}$. Since $A = \{\omega : \mathrm{E}(X|\mathcal{G})(\omega) > \mathrm{E}(Y|\mathcal{G})(\omega)\} \in \mathcal{G}$, it follows that $P(A) = 0$. The proof of (iii) uses **4.11**(ii) and is otherwise identical to that of (ii). ∎

10.15 Theorem (conditional modulus inequality)

$$\left|\mathrm{E}(Y|\mathcal{G})\right| \le \mathrm{E}\big(|Y||\mathcal{G}\big) \text{ a.s.}$$

Proof $|Y| = Y^+ + Y^-$ where Y^+ and Y^- are defined in (10.29). These are non-negative r.v.s so that $\mathrm{E}(Y^+|\mathcal{G}) \ge 0$ a.s. and $\mathrm{E}(Y^-|\mathcal{G}) \ge 0$ a.s. by **10.14** (i) and (ii). For $\omega \in C$ with $P(C) = 1$,

$$\begin{aligned}\left|\mathrm{E}(Y^+ - Y^-|\mathcal{G})(\omega)\right| &= \left|\mathrm{E}(Y^+|\mathcal{G})(\omega) - \mathrm{E}(Y^-|\mathcal{G})(\omega)\right| \\ &\le \mathrm{E}(Y^+|\mathcal{G})(\omega) + \mathrm{E}(Y^-|\mathcal{G})(\omega) \\ &= \mathrm{E}(Y^+ + Y^-|\mathcal{G})(\omega) \qquad (10.34)\end{aligned}$$

where both the equalities are by (10.25). ∎

10.16 Theorem (conditional monotone convergence) If $Y_n \le Y$ and $Y_n \uparrow Y$ a.s. then $\mathrm{E}(Y_n|\mathcal{G}) \uparrow \mathrm{E}(Y|\mathcal{G})$ a.s.

Proof Consider the monotone sequence $Z_n = Y_n - Y$. Since $Z_n \le 0$ and $Z_{n+1} \ge Z_n$, **10.14** (ii) implies that the sequence $\{\mathrm{E}(Z_n|\mathcal{G})\}$ is negative and non-decreasing a.s. and hence converges a.s. By the reverse Fatou lemma **4.15**,

$$\begin{aligned}\int_G \Big(\limsup_{n\to\infty} \mathrm{E}(Z_n|\mathcal{G})\Big)\mathrm{d}P &\ge \limsup_{n\to\infty} \int_G \mathrm{E}(Z_n|\mathcal{G})\mathrm{d}P \\ &= \limsup_{n\to\infty} \int_G Z_n \mathrm{d}P = 0 \qquad (10.35)\end{aligned}$$

for $G \in \mathcal{G}$, the first equality being by (10.18) and the second by regular monotone convergence (**4.7**). Choose $G = \{\omega : \limsup_n \mathrm{E}(Z_n|\mathcal{G})(\omega) < 0\}$, which is in $\mathcal{G}$ by **3.33** and (10.35) implies that $P(G) = 0$. It follows that

$$\lim_{n\to\infty} \mathrm{E}(Z_n|\mathcal{G}) = 0 \text{ a.s.} \quad ∎ \qquad (10.36)$$

10.17 Theorem (conditional Fatou's lemma) If $Y_n \ge 0$ a.s. then

$$\liminf_{n\to\infty} \mathrm{E}(Y_n|\mathcal{G}) \ge \mathrm{E}\big(\liminf_{n\to\infty} Y_n|\mathcal{G}\big) \text{ a.s.} \qquad (10.37)$$

Proof Put $Y'_n = \inf_{k\geq n} Y_k$ so that Y'_n is non-decreasing and converges to $Y' = \liminf_n Y_n$. Then $\mathrm{E}(Y'_n|\mathcal{G}) \to \mathrm{E}(Y'|\mathcal{G})$ by **10.16**. $Y_n \geq Y'_n$ and hence $\mathrm{E}(Y_n|\mathcal{G}) \geq \mathrm{E}(Y'_n|\mathcal{G})$ a.s. by **10.14**(ii). The theorem follows on letting $n \to \infty$. ∎

Extending the various other corollaries, such as the dominated convergence theorem, follows the pattern of the last results and is left to the reader.

10.18 Theorem (conditional Markov inequality)

$$P(\{|Y| \geq \varepsilon\}|\mathcal{G}) \leq \frac{\mathrm{E}(|Y|^p|\mathcal{G})}{\varepsilon^p} \text{ a.s.}$$

Proof By Corollary **9.17**,

$$\varepsilon^p \int_G 1_{\{|Y|\geq\varepsilon\}} \mathrm{d}P \leq \int_G |Y|^p \mathrm{d}P,\ G \in \mathcal{F}. \tag{10.38}$$

For any $G \in \mathcal{G}$,

$$\int_G P(\{|Y| \geq \varepsilon\}|\mathcal{G})\mathrm{d}P = \int_G 1_{\{|Y|\geq\varepsilon\}} \mathrm{d}P \tag{10.39}$$

by (10.18) and likewise

$$\int_G \mathrm{E}(|Y|^p|\mathcal{G})\mathrm{d}P = \int_G |Y|^p \mathrm{d}P. \tag{10.40}$$

Substituting (10.39) and (10.40) into (10.38), it follows that

$$\int_G \big(\varepsilon^p P(\{|Y| \leq \varepsilon\}|\mathcal{G}) - \mathrm{E}(|Y|^p|\mathcal{G})\big)\mathrm{d}P \leq 0. \tag{10.41}$$

The integrand of (10.41) is a $\mathcal{G}$-measurable r.v. If $G \in \mathcal{G}$ denotes the set on which it is positive, if follows from (10.41) that $P(G) = 0$. ∎

10.19 Theorem (conditional Jensen inequality) Let a Borel function ϕ be convex on an interval I containing the support of a $\mathcal{F}$-measurable r.v. Y where Y and $\phi(Y)$ are integrable. Then

$$\phi\big(\mathrm{E}(Y|\mathcal{G})\big) \leq \mathrm{E}\big(\phi(Y)|\mathcal{G}\big) \text{ a.s.} \tag{10.42}$$

Proof The proof applies **9.20**. Setting $x = \mathrm{E}(Y|\mathcal{G})$ and $y = Y$ in (9.34)

$$A\big(\mathrm{E}(Y|\mathcal{G})\big)(Y - \mathrm{E}(Y|\mathcal{G})) \leq \phi(Y) - \phi\big(\mathrm{E}(Y|\mathcal{G})\big) \text{ a.s.} \tag{10.43}$$

However, unlike $A(\mathrm{E}(Y))$, $A(\mathrm{E}(Y|\mathcal{G}))$ is a random variable. It is not certain that the left-hand side of (10.43) is integrable, so the proof cannot proceed exactly like that of **9.19**. The extra trick is to replace Y with $1_E Y$, where $E = \{\omega : \mathrm{E}(Y|\mathcal{G})(\omega) \le B\}$ for $B < \infty$. $\mathrm{E}(Y|\mathcal{G})$ and hence also 1_E are $\mathcal{G}$-measurable random variables, so $\mathrm{E}(1_E Y|\mathcal{G}) = 1_E \mathrm{E}(Y|\mathcal{G})$ a.s. by **10.10** and

$$\begin{aligned}\mathrm{E}\big(\phi(1_E Y)|\mathcal{G}\big) &= \mathrm{E}\big(\phi(Y)1_E + \phi(0)1_{E^c}|\mathcal{G}\big) \\ &= 1_E\mathrm{E}\big(\phi(Y)|\mathcal{G}\big) + (1 - 1_E)\phi(0) \text{ a.s.}\end{aligned} \tag{10.44}$$

Thus, instead of (10.43), consider

$$A\big(\mathrm{E}(1_E Y|\mathcal{G})\big)\big(1_E Y - 1_E\mathrm{E}(Y|\mathcal{G})\big) \le \phi(1_E Y) - \phi\big(1_E\mathrm{E}(Y|\mathcal{G})\big) \text{ a.s.} \tag{10.45}$$

The majorant side of (10.45) is integrable given that $\phi(Y)$ is integrable and hence so is the minorant side. Application of **10.9** and **10.10** establishes that the conditional expectation of the latter term is zero almost surely so given (10.44),

$$\phi\big(1_E\mathrm{E}(Y|\mathcal{G})\big) \le 1_E\mathrm{E}\big(\phi(Y)|\mathcal{G}\big) + (1 - 1_E)\phi(0) \text{ a.s.} \tag{10.46}$$

Finally, let $B \to \infty$ so that $1_E \to 1$ to complete the proof. ∎

The following is a simple application of the last result which will have use subsequently.

10.20 Theorem Let X be $\mathcal{G}$-measurable and L_r-bounded for $r \ge 1$. If Y is $\mathcal{F}$-measurable, $X + Y$ is L_r-bounded and $\mathrm{E}(Y|\mathcal{G}) = 0$ a.s. then

$$\mathrm{E}|X + Y|^r \ge \mathrm{E}|X|^r. \tag{10.47}$$

Proof Take expectations and apply the LIE to the inequality

$$\mathrm{E}\big(|X + Y|^r|\mathcal{G}\big) \ge \big|\mathrm{E}(X + Y|\mathcal{G})\big|^r = |X|^r \text{ a.s.} \quad ∎ \tag{10.48}$$

The final step is to generalize the results of §9.8. It suffices to illustrate with the case of differentiation under the conditional expectation.

10.21 Theorem Let a function $G(\omega, \theta)$ satisfy the conditions of **9.35**. Then

$$\mathrm{E}\Big(\frac{\mathrm{d}G}{\mathrm{d}\theta}\Big|_{\theta=\theta_0}\Big|\mathcal{G}\Big) = \frac{\mathrm{dE}(G|\mathcal{G})}{\mathrm{d}\theta}\Big|_{\theta=\theta_0} \text{ a.s.} \tag{10.49}$$

Proof Take a countable sequence $\{h_\nu, \nu \in \mathbb{N}\}$ with $h_\nu \to 0$ as $\nu \to \infty$. By linearity of the conditional expectation,

$$\mathrm{E}\Big(\frac{G(\theta_0 + h_\nu) - G(\theta_0)}{h_\nu}\Big|\mathcal{G}\Big) = \frac{\mathrm{E}\big(G(\theta_0 + h_\nu)\mathcal{G}\big) - \mathrm{E}\big(G(\theta_0)|\mathcal{G}\big)}{h_\nu} \text{ a.s.} \tag{10.50}$$

If $C_\nu \in \mathcal{G}$ is the set on which the equality in (10.50) holds, with $P(C_\nu) = 1$, the two sequences agree in the limit on the set $\bigcap_\nu C_\nu$ and $P(\bigcap_\nu C_\nu) = 1$ by **3.12**(iii). The left-hand side of (10.50) converges a.s. to the left-hand side of (10.49) by assumption, applying the conditional version of the dominated convergence theorem. Since whenever it exists the a.s. limit of the right-hand side of (10.50) is the right-hand side of (10.49) by definition, the theorem follows. ∎

10.5 Relationships between Sub-σ-fields

The sub-σ-fields $\mathcal{G}_1 \subseteq \mathcal{F}$ and $\mathcal{G}_2 \subseteq \mathcal{F}$ are said to be independent if

$$P(G_1 \cap G_2) = P(G_1)P(G_2) \tag{10.51}$$

for every pair of events $G_1 \in \mathcal{G}_1$ and $G_2 \in \mathcal{G}_2$. If Y is measurable on $\mathcal{G}_1$ it is also measurable on any collection containing $\mathcal{G}_1$ and on $\mathcal{F}$ in particular. Theorems **10.10** and **10.11** cover cases where Y as well as X is measurable on a sub-σ-field.

10.22 Theorem Random variables X and Y are independent iff $\sigma(X)$ and $\sigma(Y)$ are independent.

Proof Under the inverse mapping in (10.13), $G_1 \in \sigma(X)$ if and only if $B_1 = X(G_1) \in \mathcal{B}$ with a corresponding condition for $\sigma(Y)$. It follows that (10.51) holds for each $G_1 \in \sigma(X)$, $G_2 \in \sigma(Y)$ iff $P(X \in B_1, Y \in B_2) = P(X \in B_1)P(Y \in B_2)$ for each $B_1 = X(G_1)$, $B_2 = Y(G_2)$. The 'only if' of the theorem then follows directly from the definition of $\sigma(X)$. The 'if' follows, given (8.48), from the fact that every $B_i \in \mathcal{B}$ has an inverse image in any sub-σ-field on which a r.v. is measurable. ∎

The 'only if' in the first line of this proof is essential. Independence of the sub-σ-fields always implies independence of X and Y, but the converse holds only for the infimal cases, $\sigma(X)$ and $\sigma(Y)$.

10.23 Theorem Let Y be integrable and measurable on $\mathcal{G}_1$. Then $\mathrm{E}(Y|\mathcal{G}) = \mathrm{E}(Y)$ a.s. for all $\mathcal{G}$ independent of $\mathcal{G}_1$.

Proof Define the simple $\mathcal{G}_1$-measurable r.v.s $Y_{(n)} = \sum_{i=1}^n \gamma_i 1_{G_{1i}}$ on a partition $G_{11}, \ldots, G_{1n}$ of Ω where $G_{1i} \in \mathcal{G}_1$, each i, with $Y_{(n)} \uparrow Y$ as in **3.35**. Then

$$\int_G \mathrm{E}(Y_{(n)}|\mathcal{G})\mathrm{d}P = \sum_{i=1}^n \gamma_i \int_G \mathrm{E}(1_{G_{1i}}|\mathcal{G})\mathrm{d}P = \sum_{i=1}^n \gamma_i P(G_{1i} \cap G)$$
$$= P(G)\sum_{i=1}^n \gamma_i P(G_{1i}) = P(G)\mathrm{E}(Y_{(n)}) \text{ for all } G \in \mathcal{G}. \qquad (10.52)$$

$\mathrm{E}(Y_{(n)}) \to \mathrm{E}(Y)$ by the monotone convergence theorem. $\mathrm{E}(Y_{(n)}|\mathcal{G})$ is not a simple function, but $\mathrm{E}(Y_{(n)}|\mathcal{G}) \uparrow \mathrm{E}(Y|\mathcal{G})$ a.s. by **10.16** and

$$\int_G \mathrm{E}(Y_{(n)}|\mathcal{G})\mathrm{d}P \to \int_G \mathrm{E}(Y|\mathcal{G})\mathrm{d}P \qquad (10.53)$$

by regular monotone convergence. Hence for $\mathcal{G}_1$-measurable Y,

$$\int_G \mathrm{E}(Y|\mathcal{G})\mathrm{d}P = P(G)\mathrm{E}(Y) \text{ for all } G \in \mathcal{G}. \qquad (10.54)$$

From the second equality of (10.18) it follows that $\mathrm{E}(Y|G) = \mathrm{E}(Y)$ for all $G \in \mathcal{G}$, which proves the theorem. ∎

10.24 Corollary If X and Y are independent, then $\mathrm{E}(Y|X) = \mathrm{E}(Y)$.

Proof Direct from **10.22** and **10.23** putting $\mathcal{G} = \sigma(X)$ and $\mathcal{G}_1 = \sigma(Y)$. ∎

10.25 Corollary A pair of σ-fields $\mathcal{G}_1 \subset \mathcal{F}$ and $\mathcal{G}_2 \subset \mathcal{F}$ are independent iff $\mathrm{Cov}(X, Y) = 0$ for every pair of integrable r.v.s X and Y such that X is measurable on $\mathcal{G}_1$ and Y is measurable on $\mathcal{G}_2$.

Proof By **10.23** independence implies the condition of **10.11** is satisfied for $\mathcal{G} = \mathcal{G}_1$, proving 'only if'. To prove 'if', consider $X = 1_{G_1}$ for $G_1 \in \mathcal{G}_1$ and $Y = 1_{G_2}$ for $G_2 \in \mathcal{G}_2$. X is $\mathcal{G}_1$-measurable and Y is $\mathcal{G}_2$-measurable. For this case,

$$\mathrm{Cov}(X, Y) = P(G_1 \cap G_2) - P(G_1)P(G_2). \qquad (10.55)$$

$\mathrm{Cov}(X, Y) = 0$ for every such pair implies $\mathcal{G}_1$ and $\mathcal{G}_2$ are independent by (10.51). ∎

10.26 Corollary Random variables X and Y are independent iff Cov $\big(\phi(X), \psi(Y)\big) = 0$ for every pair of integrable Borel functions ϕ and ψ.

Proof By **10.3**(i), $\phi(X)$ is measurable with respect to $\sigma(X)$ for all ϕ and $\psi(Y)$ is $\sigma(Y)$-measurable for all ψ. If and only if all these pairs are uncorrelated, it follows by **10.25** that $\sigma(X)$ and $\sigma(Y)$ are independent sub-σ-fields. The result then follows by **10.22**. ∎

An alternative proof of the necessity part is given in **9.25**.

The next result generalizes the law of iterated expectations to sub-σ-fields. σ-fields $\mathcal{G}_1$ and $\mathcal{G}_2$ are said to be *nested* if $\mathcal{G}_1 \subseteq \mathcal{G}_2$.

10.27 Theorem If $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \mathcal{F}$, then for $\mathcal{F}$-measurable Y,

(i) $\mathrm{E}\big(\mathrm{E}(Y|\mathcal{G}_2)|\mathcal{G}_1\big) = \mathrm{E}(Y|\mathcal{G}_1)$ a.s.

(ii) $\mathrm{E}\big(\mathrm{E}(Y|\mathcal{G}_1)|\mathcal{G}_2\big) = \mathrm{E}(Y|\mathcal{G}_1)$ a.s.

Proof By (10.18),

$$\int_G \mathrm{E}\big(\mathrm{E}(Y|\mathcal{G}_2)|\mathcal{G}_1\big)\mathrm{d}P = \int_G \mathrm{E}(Y|\mathcal{G}_2)\mathrm{d}P \text{ for all } G \in \mathcal{G}_1. \tag{10.56}$$

But, since $G \in \mathcal{G}_1$ implies $G \in \mathcal{G}_2$, (10.18) and (10.56) imply that

$$\int_G \mathrm{E}\big(\mathrm{E}(Y|\mathcal{G}_2)|\mathcal{G}_1\big)\mathrm{d}P = \int_G Y\mathrm{d}P \text{ for all } G \in \mathcal{G}_1 \tag{10.57}$$

so that $\mathrm{E}(\mathrm{E}(Y|\mathcal{G}_2)|\mathcal{G}_1)$ is a version of $\mathrm{E}(Y|\mathcal{G}_1)$, proving (i). Part (ii) is by **10.9**, since $\mathrm{E}(Y|\mathcal{G}_1)$ is a $\mathcal{G}_2$-measurable r.v. ∎

A simple application of the theorem is to a three-variable distribution. If $(X(\omega)$, $Y(\omega)$, $Z(\omega))$ is a random point in $\mathbb{R}^3$, measurable on $\mathcal{F}$, let $\sigma(Z)$ and $\sigma(Y,Z)$ be the infimal σ-fields on which Z and (Y,Z) respectively are measurable and $\sigma(Z) \subseteq \sigma(Y,Z) \subseteq \mathcal{F}$. Unifying notation by writing $\mathrm{E}(Y|Z) = \mathrm{E}(Y|\sigma(Z))$ and $\mathrm{E}(Y|X,Z) = \mathrm{E}(Y|\sigma(X,Z))$, **10.27** implies that

$$\mathrm{E}\big(\mathrm{E}(Y|X,Z)|Z\big) = \mathrm{E}\big(\mathrm{E}(Y|Z)|X,Z\big) = \mathrm{E}(Y|Z). \tag{10.58}$$

The next results derive from the conditional Jensen inequality.

10.28 Theorem Let Y be a $\mathcal{F}$-measurable r.v. and $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \mathcal{F}$. If $\phi(\cdot)$ is convex,

$$\mathrm{E}\big(\phi\big(\mathrm{E}(Y|\mathcal{G}_2)\big)\big) \geq \mathrm{E}\big(\phi\big(\mathrm{E}(Y|\mathcal{G}_1)\big)\big). \tag{10.59}$$

Proof Applying **10.19** to the $\mathcal{G}_2$-measurable r.v. $\mathrm{E}(Y|\mathcal{G}_2)$ gives

$$\mathrm{E}\big(\phi(\mathrm{E}(Y|\mathcal{G}_2))|\mathcal{G}_1\big) \geq \phi\big(\mathrm{E}(\mathrm{E}(Y|\mathcal{G}_2)|\mathcal{G}_1)\big) = \phi\big(\mathrm{E}(Y|\mathcal{G}_1)\big) \text{ a.s.} \tag{10.60}$$

where the a.s. equality is by **10.27**(i). The theorem follows on taking unconditional expectations and using the LIE. ∎

The application of interest here is the comparison of absolute moments. Since $|x|^p$ is convex for $p \geq 1$, the absolute moments of $\mathrm{E}(Y|\mathcal{G}_2)$ exceed those of $\mathrm{E}(Y|\mathcal{G}_1)$ when $\mathcal{G}_1 \subseteq \mathcal{G}_2$. In particular,

$$\mathrm{E}\big(\mathrm{E}(Y|\mathcal{G}_2)^2\big) \geq \mathrm{E}\big(\mathrm{E}(Y|\mathcal{G}_1)^2\big). \tag{10.61}$$

Since $\mathrm{E}(Y|\mathcal{G}_1)$ and $\mathrm{E}(Y|\mathcal{G}_2)$ both have means of $\mathrm{E}(Y)$, (10.61) implies $\mathrm{Var}\big(\mathrm{E}(Y|\mathcal{G}_2)\big) \geq \mathrm{Var}\big(\mathrm{E}(Y|\mathcal{G}_1)\big)$. Also, noting that $\mathrm{E}\big(Y\mathrm{E}(Y|\mathcal{G}_i)\big) = \mathrm{E}\big(\mathrm{E}(Y|\mathcal{G}_i)^2\big)$,

$$\mathrm{E}\Big(\big(Y - \mathrm{E}(Y|\mathcal{G}_i)\big)^2\Big) = \mathrm{E}(Y^2) - \mathrm{E}\big(\mathrm{E}(Y|\mathcal{G}_i)^2\big)$$

for $i = 1$ or 2 and so an equivalent inequality is

$$\mathrm{E}\Big(\big(Y - \mathrm{E}(Y|\mathcal{G}_2)\big)^2\Big) \leq \mathrm{E}\Big(\big(Y - \mathrm{E}(Y|\mathcal{G}_1)\big)^2\Big). \tag{10.62}$$

The interpretation is simple. $\mathcal{G}_1$ represents a smaller information set than $\mathcal{G}_2$ and if one predictor is based on more information than another, it exhibits more variation and the prediction error accordingly less variation. The extreme cases are $\mathrm{E}(Y|\mathcal{F}) = Y$ and $\mathrm{E}(Y|\mathcal{T}) = \mathrm{E}(Y)$, with variances of $\mathrm{Var}(Y)$ and zero respectively. This generalizes a fundamental inequality, that a variance is non-negative, to the partial information case.

While (10.61) generalizes from the square to any convex function, (10.62) does not. However, there is the following norm inequality for prediction errors.

10.29 Theorem If Y is $\mathcal{F}$-measurable and $\mathcal{G}_1 \subseteq \mathcal{G}_2 \subseteq \mathcal{F}$ then for $p \geq 1$,

$$\|Y - \mathrm{E}(Y|\mathcal{G}_2)\|_p \leq 2\|Y - \mathrm{E}(Y|\mathcal{G}_1)\|_p. \tag{10.63}$$

Proof Let $\eta = Y - \mathrm{E}(Y|\mathcal{G}_1)$. Then by **10.27**(ii),

$$\begin{aligned}\eta - \mathrm{E}(\eta|\mathcal{G}_2) &= Y - \mathrm{E}(Y|\mathcal{G}_1) - \mathrm{E}(Y|\mathcal{G}_2) + \mathrm{E}\big(\mathrm{E}(Y|\mathcal{G}_1)|\mathcal{G}_2\big) \\ &= Y - \mathrm{E}(Y|\mathcal{G}_2).\end{aligned} \tag{10.64}$$

The theorem now follows, since

$$\|\eta - \mathrm{E}(\eta|\mathcal{G}_2)\|_p \le \|\eta\|_p + \|\mathrm{E}(\eta|\mathcal{G}_2)\|_p \le 2\,\|\eta\|_p \tag{10.65}$$

by, respectively, the Minkowski and conditional Jensen inequalities and the LIE. ∎

10.6 Conditional Distributions

The conditional probability of an event $A \in \mathcal{F}$ can evidently be defined as $P(A|\mathcal{G}) = \mathrm{E}(1_A|\mathcal{G})$, where $1_A(\omega)$ is the indicator function of A. But is it therefore meaningful to speak of a conditional distribution on $(\Omega,\mathcal{F})$ which assigns probabilities $P(A|\mathcal{G})$ to each $A \in \mathcal{F}$? There are two ways to approach this question.

First, observe straightforwardly that conditional probabilities satisfy the axioms of probability except on sets of probability 0 and, in this sense, satisfactorily mimic the properties of true probabilities, just as was found for the expectations. Thus:

10.30 Theorem

(i) $P(A|\mathcal{G}) \ge 0$ a.s., all $A \in \mathcal{F}$.
(ii) $P(\Omega|\mathcal{G}) = 1$ a.s.
(iii) For a countable collection of disjoint sets $A_j \in \mathcal{F}$,

$$P\Big(\bigcup_j A_j \Big| \mathcal{G}\Big) = \sum_j P(A_j|\mathcal{G}) \text{ a.s.} \tag{10.66}$$

Proof To prove (i), suppose $\exists\; G \in \mathcal{G}$ with $P(G) > 0$ and $P(A|\mathcal{G})(\omega) < 0$, for all $\omega \in G$. Then, by (10.18),

$$\int_{G\cap A} \mathrm{d}P = \int_G P(A|\mathcal{G})\mathrm{d}P < 0 \tag{10.67}$$

which is a contradiction, since the left-hand member is a probability. To prove (ii), note that $P(\Omega|\mathcal{G})$ is $\mathcal{G}$-measurable and let $G^+ \in \mathcal{G}$ denote the set of ω such that $P(\Omega|\mathcal{G})(\omega) > 1$. Suppose $P(G^+) > 0$. Then, since $G^+ \cap \Omega = G^+$,

$$P(G^+) = \int_{G^+} \mathrm{d}P < \int_{G^+} P(\Omega|\mathcal{G})\mathrm{d}P = \int_{G^+\cap\Omega} \mathrm{d}P = P(G^+). \tag{10.68}$$

This is a contradiction, implying that $P(G^+) = 0$. Repeating the argument for a set G^- on which $P(\Omega|\mathcal{G})(\omega) < 1$ shows that $P(G^-) = 0$. For (iii) (10.18) gives, for any $G \in \mathcal{G}$,

$$\int_G P\Big(\bigcup_j A_j \Big| \mathcal{G}\Big)\mathrm{d}P = \int_{G\cap\left(\bigcup_j A_j\right)} \mathrm{d}P = \int_{\bigcup_j (G\cap A_j)} \mathrm{d}P = \sum_j \int_{G\cap A_j} \mathrm{d}P \tag{10.69}$$

since the sets $G \cap A_j$ are disjoint if this is true of the A_j. By definition there exists a version of $P(A_j|\mathcal{G})$ such that $\forall\, G \in \mathcal{G}$,

$$\int_{G\cap A_j} \mathrm{d}P = \int_G P(A_j|\mathcal{G})\mathrm{d}P \tag{10.70}$$

and hence

$$\int_G P\Big(\bigcup_j A_j \Big| \mathcal{G}\Big)\mathrm{d}P = \sum_j \int_G P(A_j|\mathcal{G})\mathrm{d}P = \int_G \Big(\sum_j P(A_j|\mathcal{G})\Big)\mathrm{d}P. \tag{10.71}$$

The left- and right-hand members of (10.71) define the same measure on $\mathcal{G}$ (see **10.5**) and hence $P(\bigcup_j A_j|\mathcal{G}) = \sum_j P(A_j|\mathcal{G})$ a.s. by the Radon–Nikodym theorem. ∎

However, there is also a more exacting criterion to consider. Does there exist, for fixed ω, a p.m. μ_ω on $(\Omega, \mathcal{F})$ that satisfies

$$\mu_\omega(A) = P(A|\mathcal{G})(\omega), \text{ each } A \in \mathcal{F} \tag{10.72}$$

for all $\omega \in C$, where $P(C) = 1$? If this condition holds, the fact that conditional expectations and probabilities behave like regular expectations and probabilities requires no separate proof, since the properties hold for μ_ω. If a family of p.m.s $\{\mu_\omega, \omega \in \Omega\}$ satisfying (10.72) does exist, it is said to define a *regular* conditional probability on $\mathcal{G}$.

However, the existence of regular conditioning is *not* guaranteed in every case and counterexamples have been constructed (see e.g. Doob [60], 623–4). The problem is this. In (10.66), there is allowed to exist for a given collection $\mathcal{A} = \{A_j \in \mathcal{F}\}$ an exceptional set, say $C_\mathcal{A}$ with $P(C_\mathcal{A}) = 0$, on which the equality fails. This in itself does not violate (10.72), but the set $C_\mathcal{A}$ is specific to $\mathcal{A}$ and since there is typically an uncountable number of countable subsets $\mathcal{A} \subseteq \mathcal{F}$, it cannot be guaranteed that $P(\bigcup_\mathcal{A} C_\mathcal{A}) = 0$, as would be required for μ_ω both to be a p.m. and to satisfy (10.72).

This is not a particularly serious problem because the existence of the family $\{\mu_\omega\}$ has not been critical to the development of conditioning theory up to this point, but for certain purposes it is useful to know that p.m.s on the line *do* admit regular conditional distributions. This is shown as follows.

10.31 Theorem Given a probability space $(\Omega,\mathcal{F},P)$ and a sub-σ-field $\mathcal{G}\subset\mathcal{F}$, a r.v. Y has a regular conditional distribution defined by

$$F_Y(y|\mathcal{G})(\omega)=P\big((-\infty,y]|\mathcal{G}\big)(\omega),\ y\in\mathbb{R} \tag{10.73}$$

for $\omega\in C$ with $P(C)=1$, where $F_Y(\cdot|\mathcal{G})(\omega)$ is a c.d.f. for all $\omega\in\Omega$.

Proof Write $F^*_\omega(y)$ to denote a version of $P((-\infty,y]|\mathcal{G})(\omega)$. Let M_{ij} denote the set of ω such that $F^*_\omega(r_i)>F^*_\omega(r_j)$ for r_i, $r_j\in\mathbb{Q}$ with $r_i<r_j$. Similarly, let R_i denote the set of ω on which $\lim_{n\to\infty}F^*_\omega(r_i+1/n)\neq F^*_\omega(r_i)$, $r_i\in\mathbb{Q}$. Finally, let L denote the set of those ω for which $F^*_\omega(+\infty)\neq 1$ and $F^*_\omega(-\infty)\neq 0$. Then $C=(\bigcup_{ij}M_{ij})^c\cap(\bigcup_i R_i)^c\cap L^c$ is the set of ω on which $F^*_\omega(y)$ is monotone and right-continuous at all rational points of the line, with $F^*_\omega(+\infty)=1$ and $F^*_\omega(-\infty)=0$. For $y\in\mathbb{R}$ let

$$F_Y(\cdot|\mathcal{G})(\omega)=\begin{cases}\left.\begin{cases}F^*_\omega(y),\ y\in\mathbb{Q}\\ F^*_\omega(y+),\ y\in\mathbb{R}-\mathbb{Q}\end{cases}\right\}, & \omega\in C\\ \qquad G(y), & \text{otherwise}\end{cases} \tag{10.74}$$

where G is an arbitrary c.d.f. In view of **10.30**, $P(M_{ij})=0$ for each pair i,j, $P(R_i)=0$ for each i, and $P(L)=0$. (If need be, work in the completion of the space to define these probabilities.) Since this collection is countable, $P(C)=1$, and in view of **8.5**, $F_Y(\cdot|\mathcal{G})(\omega)$ is a c.d.f. which satisfies (10.73), as it was required to show. ∎

It is straightforward, at least in principle, to generalize this argument to multivariate distributions.

For $B\in\mathcal{B}$ it is possible to write

$$\mathrm{E}(1_B|\mathcal{G})(\omega)=\int_B \mathrm{d}F_Y(y|\mathcal{G})(\omega)\ \text{a.s.} \tag{10.75}$$

and the standard argument by way of simple functions and monotone convergence goes full circle to the representation

$$\mathrm{E}(Y|\mathcal{G})(\omega)=\int_{-\infty}^{+\infty} y\mathrm{d}F_Y(y|\mathcal{G})(\omega)\ \text{a.s.} \tag{10.76}$$

If $\mathcal{G}=\sigma(X)$ there are constructions to parallel those of §10.1. Since no restriction had to be placed on the distribution to obtain this result, there is evidently a way around the difficulties associated with the earlier definitions.

However, $F_Y(\cdot|\mathcal{G})(\omega)$ is something of a novelty, a c.d.f. that is a random element from a probability space. Intuitively, it must be understood as representing the subjective distribution of $Y(\omega)$ in the mind of the observer who knows whether or not $\omega \in G$ for each $G \in \mathcal{G}$. The particular case $F_Y(\cdot|\mathcal{G})(\omega)$ is the one of interest to the statistical modeller when the outcome ω is realized. Many random variables may be generated from the elements of $(\Omega, \mathcal{F}, P)$; not only the outcome itself—in the bivariate case the pair $Y(\omega)$, $X(\omega)$—but also variables such as $\mathrm{E}(Y|X)(\omega)$ and the quantiles of $F_Y(\cdot|X)(\omega)$. All these have to be thought of as different aspects of the same random experiment.

Let X and Y be r.v.s and $\mathcal{G}$ a sub-σ-field with $\mathcal{G} \subseteq \mathcal{H}_X = \sigma(X)$ and $\mathcal{G} \subseteq \mathcal{H}_Y = \sigma(Y)$. X and Y are said to be *independent conditional on* $\mathcal{G}$ if

$$F_{XY}(x,y|\mathcal{G}) = F_X(x|\mathcal{G})F_Y(y|\mathcal{G}) \text{ a.s.} \tag{10.77}$$

This condition implies, for example, that $\mathrm{E}(XY|\mathcal{G}) = \mathrm{E}(X|\mathcal{G})\mathrm{E}(Y|\mathcal{G})$ a.s. Let $\mu_\omega = \mu(\cdot,\omega)$ be the conditional measure such that

$$\mu_\omega\big(\{X \in (-\infty,x], Y \in (-\infty,y]\}\big) = F_{XY}(x,y|\mathcal{G})(\omega).$$

With ω fixed this is a regular p.m. by (the bivariate generalization of) **10.31** and $\mu_\omega(A \cap B) = \mu_\omega(A)\mu_\omega(B)$ for each $A \in \mathcal{H}_X$ and $B \in \mathcal{H}_Y$ by **10.22**. In this sense, the sub-σ-fields $\mathcal{H}_X$ and $\mathcal{H}_Y$ can be called conditionally independent.

10.32 Theorem If X and Y are independent conditional on $\mathcal{G}$, then

$$\mathrm{E}(Y|\mathcal{H}_X) = \mathrm{E}(Y|\mathcal{G}) \text{ a.s.} \tag{10.78}$$

Proof By independence of $\mathcal{H}_X$ and $\mathcal{H}_Y$ under μ_ω

$$\int_A \mathrm{E}(Y|\mathcal{H}_X)d\mu_\omega = \int_A Y d\mu_\omega = \mu_\omega(A)\int Y d\mu_\omega,\ A \in \mathcal{H}_X. \tag{10.79}$$

This is equivalent to

$$\mathrm{E}(1_A\mathrm{E}(Y|\mathcal{H}_X)|\mathcal{G})(\omega) = \mathrm{E}(1_A Y|\mathcal{G})(\omega) = \mathrm{E}(1_A\mathrm{E}(Y|\mathcal{G})|\mathcal{G})(\omega) \text{ a.s.}[P] \tag{10.80}$$

where the first equality also follows from **10.27**(i) and **10.10**. Integrating over Ω with respect to P, noting $\Omega \in \mathcal{G}$, using **4.11**(ii) and the LIE, leads to

$$\int_A \mathrm{E}(Y|\mathcal{H}_X)\mathrm{d}P = \int_A Y\mathrm{d}P = \int_A \mathrm{E}(Y|\mathcal{G})\mathrm{d}P,\ A \in \mathcal{H}_X. \tag{10.81}$$

This shows $\mathrm{E}(Y|\mathcal{H}_X)$ is a version of $\mathrm{E}(Y|\mathcal{G})$, completing the proof. ∎

Thus, while $\mathrm{E}(Y|\mathcal{H}_X)$ is in principle $\mathcal{H}_X$-measurable, it is in fact almost surely $[P]$ equal to a $\mathcal{G}$-measurable r.v. Needless to say, the whole argument is symmetric in X and $\mathcal{H}_Y$.

The idea captured here is that for an observer who possesses the information in $\mathcal{G}$ (knows whether $\omega \in G$ for each $G \in \mathcal{G}$), observing X does not yield any additional information that improves his prediction of Y, and vice versa. This need not be true for an observer who does not possess prior information. Equation (10.78) shows that the predictors of Y based on the smaller information set $\mathcal{G}$ and the larger information set $\mathcal{H}_X$ are the same a.s.$[P]$, although this does not imply $\mathrm{E}(Y|\mathcal{H}_X) = \mathrm{E}(Y)$ a.s., so that X and Y are not independent in the ordinary sense.

11
Characteristic Functions

11.1 The Distribution of Sums of Random Variables

Let an independent pair of r.v.s X and Y have marginal c.d.f.s F_X and F_Y. The c.d.f of the sum $W = X + Y$ is given by the *convolution* of F_X and F_Y, the function

$$F_X * F_Y(w) = \int_{-\infty}^{+\infty} F_X(w-y)\mathrm{d}F_Y(y). \tag{11.1}$$

11.1 Theorem If r.v.s X and Y are independent, then

$$F_X * F_Y(w) = P(X + Y \leq w) = F_Y * F_X(w). \tag{11.2}$$

Proof Let $1_w(x,y)$ be the indicator function of the set $\{x,y : x \leq w - y\}$, so that $P(X + Y \leq w) = \mathrm{E}\big(1_\omega(X,Y)\big)$. By independence $F(x,y) = F_X(x)F_Y(y)$, so this is

$$\begin{aligned}\int_{\mathbb{R}^2} 1_w(x,y)\mathrm{d}F(x,y) &= \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{+\infty} 1_w(x,y)\mathrm{d}F_X(x)\right)\mathrm{d}F_Y(y)\\ &= \int_{-\infty}^{+\infty}\left(\int_{-\infty}^{w-y} \mathrm{d}F_X(x)\right)\mathrm{d}F_Y(y)\\ &= \int_{-\infty}^{+\infty} F_X(w-y)\mathrm{d}F_Y(y) \qquad (11.3)\end{aligned}$$

where the first equality is by Fubini's theorem. This establishes the first equality in (11.2). Reversing the roles of X and Y in (11.3) establishes the second. ∎

For continuous distributions, the convolution $f = f_X * f_Y$ of p.d.f.s f_X and f_Y is

$$f(w) = \int_{-\infty}^{+\infty} f_X(w-y)f_Y(y)\mathrm{d}y \tag{11.4}$$

such that $\int_{-\infty}^{w} f(\xi)\mathrm{d}\xi = F(w)$.

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0011

11.2 Example Let X and Y be independent drawings from the uniform distribution on $[0,1]$, so that $f_X(x) = 1_{[0,1]}(x)$. Applying (11.4) gives

$$f_{X+Y}(w) = \int_0^1 1_{[w-1,w]}\mathrm{d}y. \tag{11.5}$$

It is easily verified that the graph of this function forms an isosceles triangle with base $[0,2]$ and height 1. □

This is the most direct result on the distribution of sums, but the formulae generated by applying the rule recursively are not easy to handle and other approaches are preferred. The *moment-generating function* (m.g.f.) of X, when it exists, is

$$M_X(t) = \mathrm{E}(\mathrm{e}^{tX}) = \int \mathrm{e}^{tx}\mathrm{d}F(x),\ t \in \mathbb{R}. \tag{11.6}$$

(Integrals are taken over $(-\infty, +\infty)$ unless otherwise indicated.) For future reference, this is a good point at which to recall the binomial expansion

$$\mathrm{e}^x = \lim_{n\to\infty}\left(1 + \frac{x}{n}\right)^n = 1 + x + \frac{x^2}{2} + \frac{x^3}{6} + \ldots + \frac{x^n}{n!} + \ldots. \tag{11.7}$$

If X and Y are independent,

$$\begin{aligned} M_{X+Y}(t) &= \int\int \mathrm{e}^{t(x+y)}\mathrm{d}F_X(x)\mathrm{d}F_Y(y) \\ &= \int \mathrm{e}^{tx}\mathrm{d}F_X(x)\int \mathrm{e}^{ty}\mathrm{d}F_Y(y) \\ &= M_X(t)M_Y(t). \end{aligned} \tag{11.8}$$

This suggests a simple approach to analysing the distribution of independent sums. The difficulty is that the method is not universal since the m.g.f. is not defined for every distribution. Considering the series expansion of e^{tX} as in (11.7), all the moments of X must evidently exist for M_X to be well defined.

The solution to this problem is to replace the variable t by it, where 'i' is the imaginary number, $\sqrt{-1}$. The *characteristic function* (ch.f.) of X is defined as

$$\phi_X(t) = \mathrm{E}(\mathrm{e}^{\mathrm{i}tX}) = \int \mathrm{e}^{\mathrm{i}tX}\mathrm{d}F(x). \tag{11.9}$$

11.2 Complex Numbers

A complex number is $z = a + \mathrm{i}b$, where a and b are real numbers and $\mathrm{i} = \sqrt{-1}$, the 'i' standing for 'imaginary' since, as everyone knows, negative real numbers do not have square roots. a and b are called the real and imaginary parts of z, denoted $a = \mathrm{Re}(z)$ and $b = \mathrm{Im}(z)$. The space of complex numbers is represented by the symbol $\mathbb{C}$, to complement $\mathbb{R}$ where the latter stands for that subset of $\mathbb{C}$ for which $b = 0$. Complex arithmetic is mainly a matter of carrying i as an algebraic unknown and replacing i^2 by -1, i^3 by $-\mathrm{i}$, i^4 by 1, etc. wherever these appear in an expression.

Conventionally, $z \in \mathbb{C}$ is represented as a point in the plane with Cartesian coordinates a and b, the so-called *Argand diagram*. The *modulus* or absolute value of z is then its Euclidean distance from the origin in the Argand plane. Defining the *complex conjugate* of $z \in \mathbb{C}$ to be the number $\bar{z} = a - \mathrm{i}b$, the modulus is therefore the real number

$$|z| = (z\bar{z})^{1/2} = (a^2 + b^2)^{1/2}. \tag{11.10}$$

In polar coordinates, *Euler's formula* defines the complex exponential

$$\mathrm{e}^{\mathrm{i}\theta} = \cos\theta + \mathrm{i}\sin\theta \tag{11.11}$$

where $\theta \in (-\pi, \pi]$ is called the *argument* of the number, being the angle of the hypotenuse of the right-angle triangle with sides $\sin\theta$ and $\cos\theta$ in the Argand plane.[1] The simplest way to confirm the truth of this famous formula is by substituting the power series expansions

$$\cos\theta = 1 - \frac{\theta^2}{2!} + \frac{\theta^4}{4!} - \frac{\theta^6}{6!} + \dots, \quad \sin\theta = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \dots$$

to compare with (11.7). A basic trigonometric identity gives

$$|\mathrm{e}^{\mathrm{i}\theta}| = (\cos^2\theta + \sin^2\theta)^{1/2} = 1 \tag{11.12}$$

for any θ and therefore $z = |z|\mathrm{e}^{\mathrm{i}\theta}$ where $\mathrm{Re}(z) = |z|\cos\theta$ and $\mathrm{Im}(z) = |z|\sin\theta$. Also note that (11.12) gives

$$|\mathrm{e}^{z}| = |\mathrm{e}^{\mathrm{Re}(z)+\mathrm{i}\,\mathrm{Im}(z)}| = \mathrm{e}^{\mathrm{Re}(z)}|\mathrm{e}^{\mathrm{i}\,\mathrm{Im}(z)}| = \mathrm{e}^{\mathrm{Re}(z)}. \tag{11.13}$$

Going the other way, the number $z = a + \mathrm{i}b$ expressed in polar coordinates is $z = |z|\mathrm{e}^{\mathrm{i}\mathrm{Arg}(z)}$ where

[1] The range of $\mathrm{e}^{\mathrm{i}\theta}$ is the unit circle in the Argand plane. The effect of setting the domain of θ to $[0, 2\pi)$ (say) instead of $(-\pi, \pi]$ is to relocate the start point on the circle.

$$\operatorname{Arg}(z) = \begin{cases} \arctan(b/a), & a \geq 0 \\ \arctan(b/a) + \pi \operatorname{sgn}(b), & a < 0. \end{cases}$$

To take the logarithm of a complex number appears straightforward in polar coordinates, since $\log(re^{i\theta}) = \log r + i\theta$. However, $e^{i\theta} = e^{i(\theta+2k\pi)}$ for every integer k and hence the logarithm is not uniquely defined. Restricting attention to $k = 0$, or equivalently restricting θ to the interval $(-\pi, \pi]$, defines the *principal value* of the logarithm which for most purposes is what is wanted.

The integration of complex-valued functions, and in particular the formation of expected values of complex-valued random variables, is in most cases a straightforward application of linearity. If X and Y are real r.v.s, $Z = X + iY$ is a complex-valued r.v. whose distribution is defined in the obvious way by a bivariate c.d.f. $F(x, y)$, and in particular,

$$E(Z) = E(X) + iE(Y). \tag{11.14}$$

Since $(x^2 + y^2)^{1/2} \leq |x| + |y|$ for real variables x and y, $E|Z| \leq E|X| + E|Y|$ and therefore integrability of X and Y is sufficient for the integrability of Z. Many of the standard properties of expectations extend to the complex case in a straightforward way. One result that does need proof is the generalization of the modulus inequality. This is proved for general measures for direct comparison with **4.10**, although it has the immediate corollary

$$|E(Z)| \leq E|Z|. \tag{11.15}$$

11.3 Theorem If $f = g + ih$ is a complex-valued integrable function,

$$\left|\int f d\mu\right| \leq \int |f| d\mu.$$

Proof This is proved first for simple functions. Let

$$f_{(n)} = g_{(n)} + ih_{(n)} = \sum_{j=1}^{n} a_j 1_{E_j} + i\sum_{j=1}^{n} b_j 1_{E_j} \tag{11.16}$$

where the a_j and b_j are real non-negative constants and the $E_j \in \mathcal{F}$ constitute a partition of Ω. Write $\mu_j = \mu(E_j)$. Then

$$\begin{aligned} \left|\int f_{(n)} d\mu\right|^2 &= \Big(\sum_j a_j\mu_j\Big)^2 + \Big(\sum_j b_j\mu_j\Big)^2 \\ &= \sum_j (a_j^2 + b_j^2)\mu_j^2 + \sum_j \sum_{k\neq j} (a_j a_k + b_j b_k)\mu_j\mu_k \end{aligned} \tag{11.17}$$

whereas

$$\begin{aligned}\left(\int |f_{(n)}|\mathrm{d}\mu\right)^2 &= \left(\sum_j (a_j^2+b_j^2)^{1/2}\mu_j\right)^2 \\ &= \sum_j (a_j^2+b_j^2)\mu_j^2 + \sum_j \sum_{k\neq j} (a_j^2+b_j^2)^{1/2}(a_k^2+b_k^2)^{1/2}\mu_j\mu_k. \qquad (11.18)\end{aligned}$$

The modulus inequality holds if

$$\begin{aligned}0 &\le \left(\int |f_{(n)}|\mathrm{d}\mu\right)^2 - \left|\int f_{(n)}\mathrm{d}\mu\right|^2 \\ &= \sum_j \sum_{k\neq j} \left((a_j^2+b_j^2)^{1/2}(a_k^2+b_k^2)^{1/2} - (a_ja_k+b_jb_k)\right)\mu_j\mu_k. \qquad (11.19)\end{aligned}$$

The coefficients of $\mu_j\mu_k$ in this expression are the differences of pairs of non-negative terms and these differences are non-negative iff the differences of the squares are non-negative. As required,

$$\begin{aligned}(a_j^2+b_j^2)(a_k^2+b_k^2) - (a_ja_k+b_jb_k)^2 &= a_j^2b_k^2 + a_k^2b_j^2 - 2a_ja_kb_jb_k \\ &= (a_jb_k - a_kb_j)^2 \ge 0. \qquad (11.20)\end{aligned}$$

This result extends to any complex-valued function having non-negative real and imaginary parts by letting $g_{(n)} \uparrow g$ and $h_{(n)} \uparrow h$, using **3.35** and invoking the monotone convergence theorem. To extend to general integrable r.v.s, split g and h into positive and negative parts so that $f = f^+ - f^-$, where $f^+ = g^+ + \mathrm{i}h^+$ and $f^- = g^- + \mathrm{i}h^-$, with both parts non-negative in each case. The proof is then completed by noting

$$\left|\int (f^+ - f^-)\mathrm{d}\mu\right| \le \left|\int (g^+ + g^- + \mathrm{i}(h^+ + h^-))\mathrm{d}\mu\right|. \quad ∎ \qquad (11.21)$$

Before going further, it might be of interest to reflect on the reasons why complex-valued functions like (11.9) are useful in distribution theory. After all, the best-known application of complex numbers in mathematics, extracting the roots of polynomials, does not feature significantly here. The reason might be best appreciated from the discussions in §27.2 and §29.1 where in the context of general metric spaces it is shown that the key tools for analysing the weak convergence of distributions are the expectations of uniformly bounded functions. For distributions on the real line, in particular, it turns out that the characteristic function supplies the essential ingredients of a bounded function pairing uniquely with any probability measure while satisfying a neat multiplicative rule for independent sums. To fulfil this role it has been lifted conveniently ready-made off the shelf labelled 'Fourier analysis'. The fact that it is complex-valued is really incidental but the whole Fourier box of tricks, with the Euler identity at its heart, is just too useful and powerful to pass up.

11.3 The Theory of Characteristic Functions

The fact that the characteristic function $\phi_X(t)$ is defined for any distribution follows from the fact that $|\mathrm{e}^{\mathrm{i}tx}| = 1$ for all x. Since $\mathrm{E}(|\mathrm{e}^{\mathrm{i}tX}|) = 1$, $\mathrm{E}(\mathrm{e}^{\mathrm{i}tX})$ is finite regardless of the distribution of X. The real and imaginary parts of $\phi_X(t)$ are respectively $\mathrm{E}(\cos tX)$ and $\mathrm{E}(\sin tX)$.

The connection linking the ch.f. of a distribution and its integer moments is demonstrated through the following result.

11.4 Theorem If $\mathrm{E}|X|^k < \infty$, then

$$\left.\frac{\mathrm{d}^k\phi_X(t)}{\mathrm{d}t^k}\right|_{t=0} = \mathrm{i}^k\mathrm{E}(X^k). \tag{11.22}$$

Proof Consider the case $k = 1$. Since $\mathrm{d}(\mathrm{e}^{\mathrm{i}tx})/\mathrm{d}t = \mathrm{i}x\mathrm{e}^{\mathrm{i}tx}$, applying **9.35** gives

$$\frac{\mathrm{d}\phi_X(t)}{\mathrm{d}t} = \mathrm{E}\big(\mathrm{i}X\mathrm{e}^{\mathrm{i}tX}\big). \tag{11.23}$$

Since $|\mathrm{i}x\mathrm{e}^{\mathrm{i}tx}| = |x|$ the expectation in (11.23) exists if $E|X| < \infty$. To complete the proof, apply the same argument inductively to the functions $(\mathrm{i}x)^{k-1}e^{\mathrm{i}tx}$ for $k = 2, 3, \ldots$. ∎

Two properties of the characteristic function will be much exploited. First, for a pair of constants a and b,

$$\phi_{aX+b}(t) = \mathrm{E}\big(\mathrm{e}^{\mathrm{i}t(aX+b)}\big) = \mathrm{e}^{\mathrm{i}bt}\phi_X(at). \tag{11.24}$$

Second, there is the counterpart of the factorization in (11.8). For a pair of independent random variables X and Y,

$$\begin{aligned}\phi_{X+Y}(t) &= \int\int \mathrm{e}^{\mathrm{i}t(x+y)}\mathrm{d}F_X(x)\mathrm{d}F_Y(y)\\ &= \int \mathrm{e}^{\mathrm{i}tx}\mathrm{d}F_X(x)\int \mathrm{e}^{\mathrm{i}ty}\mathrm{d}F_Y(y)\\ &= \phi_X(t)\phi_Y(t). \end{aligned} \tag{11.25}$$

An interesting case of (11.25) is $Y = -X'$, where X' is an independent drawing from the distribution of X. The distribution of $X - X'$ is the same as that of $X' - X$ and hence this r.v. is symmetric about 0. The ch.f. of $X - X'$ is real-valued in view of the fact that $\phi_X(-t) = \mathrm{E}(\mathrm{e}^{-\mathrm{i}tX}) = \overline{\phi_X(t)}$ and hence

$$\phi_X(t)\phi_X(-t) = |\phi_X(t)|^2. \tag{11.26}$$

It can be verified from (11.22) that if the ch.f. is real all the existing odd-order moments are zero. This is the trademark of a symmetric distribution (see **9.9**). Considering more generally a sum $S_n = \sum_{i=1}^n X_i$ where $\{X_1, \ldots, X_n\}$ are a totally independent collection, recursive application of (11.25) yields

$$\phi_{S_n}(t) = \prod_{i=1}^n \phi_{X_i}(t). \tag{11.27}$$

The distribution of S_n can therefore be found given a formula linking the ch.f. with the relevant c.d.f., or, where appropriate, p.d.f.

The Taylor expansion of the ch.f. to k^{th} order with remainder provides the following useful approximation theorem, to be exploited in the sequel in proofs of convergence in distribution.

11.5 Theorem If $\mathrm{E}|X|^k < \infty$, then

$$\left|\phi_X(t) - \sum_{j=0}^k \frac{(\mathrm{i}t)^j \mathrm{E}(X^j)}{j!}\right| \le \mathrm{E}\left(\min\left\{\frac{2|tX|^k}{k!}, \frac{|tX|^{k+1}}{(k+1)!}\right\}\right). \tag{11.28}$$

Proof Integration by parts of the product $\mathrm{e}^{\mathrm{i}s}(x-s)^{j+1}/(j+1)$ for $j \ge 0$ provides the recursive relation

$$\int_0^x \mathrm{e}^{\mathrm{i}s}(x-s)^j \mathrm{d}s = \frac{x^{j+1}}{j+1} + \frac{\mathrm{i}}{j+1}\int_0^x \mathrm{e}^{\mathrm{i}s}(x-s)^{j+1}\mathrm{d}s \tag{11.29}$$

where with $j = 0$ the left-hand side is $(\mathrm{e}^{\mathrm{i}x} - 1)/\mathrm{i}$. Iterating the formula k times therefore yields

$$\mathrm{e}^{\mathrm{i}x} = \sum_{j=0}^k \frac{(\mathrm{i}x)^j}{j!} + \frac{\mathrm{i}^{k+1}}{k!}\int_0^x \mathrm{e}^{\mathrm{i}s}(x-s)^k \mathrm{d}s \tag{11.30}$$

where the final term can be recognized as a form of the Taylor expansion remainder (compare (11.7)). However, (11.29) can also be written in the form

$$\int_0^x (\mathrm{e}^{\mathrm{i}s} - 1)(x-s)^j \mathrm{d}s = \frac{\mathrm{i}}{j+1}\int_0^x \mathrm{e}^{\mathrm{i}s}(x-s)^{j+1}\mathrm{d}s.$$

Hence, setting $j = k-1$, an alternative form for the expansion is

$$\mathrm{e}^{\mathrm{i}x} = \sum_{j=0}^{k}\frac{(\mathrm{i}x)^j}{j!} + \frac{\mathrm{i}^k}{(k-1)!}\int_0^x (\mathrm{e}^{\mathrm{i}s} - 1)(x-s)^{k-1}\mathrm{d}s. \tag{11.31}$$

It follows that

$$\left|\mathrm{e}^{\mathrm{i}x} - \sum_{j=0}^{k}\frac{(\mathrm{i}x)^j}{j!}\right| \le \min\left\{\frac{2|x|^k}{k!}, \frac{|x|^{k+1}}{(k+1)!}\right\} \tag{11.32}$$

where the terms whose minimum appears on the majorant side are, respectively, bounds on the moduli of the remainders in (11.31) and (11.30). Substituting tX for x, taking expectations through and applying the modulus inequality gives the result. ∎

There is no need for $\mathrm{E}|X|^{k+1}$ to exist for this theorem to hold. Think of it as giving the best approximation regardless of whether $|tX|$ is large or small. To interpret the expectation on the right-hand side of (11.28) note that for any pair of non-negative, measurable functions g and h,

$$\mathrm{E}\big(\min\{g(X), h(X)\}\big) = \inf_{A\in\mathcal{B}} \mathrm{E}\big(g(X)1_A + h(X)1_{A^c}\big). \tag{11.33}$$

The infimal set A in (11.33) contains the points $x \in \mathbb{R}$ at which $g(x) \le h(x)$. In particular, for any $\varepsilon \ge 0$ the set $A = \{|X| > \varepsilon\}$ belongs to the class over which the infimum in (11.33) is taken and

$$\begin{aligned}\mathrm{E}\left(\min\left\{\frac{2|tX|^k}{k!}, \frac{|tX|^{k+1}}{(k+1)!}\right\}\right) &\le \mathrm{E}\left(\frac{2|tX|^k}{k!}1_{\{|X|>\varepsilon\}}\right) + \mathrm{E}\left(\frac{|tX|^{k+1}}{(k+1)!}1_{\{|X|\le\varepsilon\}}\right)\\ &\le \begin{cases}\dfrac{2|t|^k}{k!}\mathrm{E}(|X|^k 1_{\{|X|>\varepsilon\}}) + \dfrac{|t|^{k+1}}{(k+1)!}\varepsilon^{k+1}\\[2ex] \dfrac{2|t|^k}{k!}\mathrm{E}(|X|^k 1_{\{|X|>\varepsilon\}}) + \dfrac{|t|^{k+1}}{(k+1)!}\mathrm{E}|X|^k\varepsilon.\end{cases}\end{aligned} \tag{11.34}$$

The second alternative on the right is obtained in view of the fact that

$$\mathrm{E}(|X|^{k+1}1_{\{|X|\le\varepsilon\}}) = \mathrm{E}(|X|^k|X|1_{\{|X|\le\varepsilon\}}) \le \mathrm{E}|X|^k\varepsilon.$$

Both of these versions of the bound on the truncation error prove useful subsequently.

11.4 Examples

Consider how the transformation of (11.9) applies to the special distributions listed in §8.3.

11.6 Example For the Bernoulli distribution (**8.7**),

$$\phi_X(t;p) = pe^{it} + (1-p)e^0 = 1 + p(e^{it} - 1). \quad \square \tag{11.35}$$

11.7 Example For the binomial distribution (**8.8**), recalling that this is the sum of n independent Bernoulli draws,

$$\phi_X(t;p,n) = (1 + p(e^{it} - 1))^n. \quad \square \tag{11.36}$$

11.8 Example For the Poisson distribution (**8.9**),

$$\phi_X(t;\lambda) = e^{-\lambda} \sum_{x=0}^{\infty} \frac{1}{x!} \lambda^x e^{itx} = e^{\lambda(e^{it}-1)}. \tag{11.37}$$

Note that putting $p = \lambda/n$ in (11.36) and taking the limit as in (11.7) also yields the formula. □

11.9 Example The Uniform[0,1] distribution (**8.10**) has

$$\phi_X(t) = \int_0^1 e^{itx} dx = \frac{e^{it} - 1}{it}. \quad \square$$

11.10 Example The exponential distribution (**8.11**) has

$$\phi_X(t;\lambda) = \lambda \int_0^{\infty} e^{(it-\lambda)x} dx = \frac{\lambda}{\lambda - it}. \quad \square$$

11.11 Example The gamma(p,λ) distribution (**8.12**), recalling it to be the sum of p independent exponential drawings, has

$$\phi_X(t;p,\lambda) = \frac{\lambda^p}{\Gamma(p)} \int_0^{\infty} e^{(it-\lambda)x} x^{p-1} dx = \left(\frac{\lambda}{\lambda - it}\right)^p. \quad \square \tag{11.38}$$

11.12 Example In the standard Gaussian case (**8.13**),

$$\phi_X(t) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{+\infty} e^{itx - x^2/2} dx. \tag{11.39}$$

Completing the square yields $\mathrm{i}tx - x^2/2 = -(x-\mathrm{i}t)^2/2 - t^2/2$ and hence

$$\phi_X(t) = \mathrm{e}^{-t^2/2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{+\infty}\mathrm{e}^{-(x-\mathrm{i}t)^2/2}\mathrm{d}x = \mathrm{e}^{-t^2/2}. \tag{11.40}$$

Note that by Cauchy's integral theorem the integral in the middle member of (11.40) has the value $\sqrt{2\pi}$ for any choice of t. Accordingly, consider $X = \sigma Z + \mu$, whose p.d.f. is given by (8.29). Using (11.24) yields

$$\phi_X(t;\mu,\sigma) = \mathrm{e}^{\mathrm{i}\mu t - \sigma^2 t^2/2}. \tag{11.41}$$

Equation (11.22) can be used to verify the moment formulae given in **9.6** and **9.14**. With $\mu = 0$ the ch.f. is real, reflecting the symmetry of the distribution. □

11.13 Example The Cauchy distribution (**8.15**) has no integer moments. The ch.f. is

$$\int_{-\infty}^{\infty}\frac{\mathrm{e}^{\mathrm{i}tx}}{\pi(1+x^2)}\mathrm{d}x = \int_{-\infty}^{\infty}\frac{\cos(tx)}{\pi(1+x^2)}\mathrm{d}x + \mathrm{i}\int_{-\infty}^{\infty}\frac{\sin(tx)}{\pi(1+x^2)}\mathrm{d}x = \mathrm{e}^{-|t|}.$$

The imaginary term vanishes because sin is an odd function; for the solution of the real part see for example [85], 3.723. This formula is not differentiable at $t = 0$, just as (11.22) predicts. By (11.24) the ch.f. for the Cauchy family (**8.19**) is

$$\phi_X(t;\nu,\delta) = \mathrm{e}^{\mathrm{i}\nu - \delta|t|}. \quad \square \tag{11.42}$$

11.14 Example The degenerate distribution, defined as the case $X = a$ with probability 1 where a is a constant, has

$$\phi_X(t) = \mathrm{e}^{\mathrm{i}a}. \quad \square \tag{11.43}$$

The ch.f. is also defined for multivariate distributions. For a random vector $\boldsymbol{X}$ $(m \times 1)$ the ch.f. is

$$\phi_X(\boldsymbol{t}) = \mathrm{E}\big(\exp\{\mathrm{i}\boldsymbol{t}'\boldsymbol{X}\}\big) \tag{11.44}$$

where $\boldsymbol{t}$ is a m-vector of arguments. This case is important not least because of the ease with which, by the generalization of (11.24), the ch.f. can be derived for an affine transformation of the vector. Let $\boldsymbol{Y} = \boldsymbol{B}\boldsymbol{X} + \boldsymbol{d}$ $(k \times 1)$, where $\boldsymbol{B}$ $(k \times m)$ and $\boldsymbol{d}$ $(k \times 1)$ are constants, and then

$$\begin{aligned}\phi_Y(\boldsymbol{t}) &= \mathrm{E}\big(\exp\{\mathrm{i}\boldsymbol{t}'\boldsymbol{Y}\}\big)\\ &= \exp\{\mathrm{i}\boldsymbol{t}'\boldsymbol{d}\}\mathrm{E}\big(\exp\{\mathrm{i}\boldsymbol{t}'\boldsymbol{BX}\}\big) = \exp\{\mathrm{i}\boldsymbol{t}'\boldsymbol{d}\}\phi_X(\boldsymbol{B}'\boldsymbol{t}). \quad \square \qquad (11.45)\end{aligned}$$

11.15 Example Let $\boldsymbol{X}$ $(m \times 1)$ be multivariate normal with p.d.f. as in (8.46). The ch.f. is

$$\begin{aligned}\phi_X(\boldsymbol{t};\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \frac{1}{(2\pi)^{m/2}|\boldsymbol{\Sigma}|^{1/2}}\int_{-\infty}^{+\infty}\cdots\int_{-\infty}^{+\infty}\exp\{\mathrm{i}\boldsymbol{t}'\boldsymbol{x} - \tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\}\mathrm{d}\boldsymbol{x}\\ &= \exp\{\mathrm{i}\boldsymbol{t}'\boldsymbol{\mu} - \tfrac{1}{2}\boldsymbol{t}'\boldsymbol{\Sigma}\mathbf{t}\}. \qquad (11.46)\end{aligned}$$

The second equality is obtained as before by completing the square:

$$\mathrm{i}\boldsymbol{t}'\boldsymbol{x} - \tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}) = \mathrm{i}\boldsymbol{t}'\boldsymbol{\mu} - \tfrac{1}{2}\boldsymbol{t}'\boldsymbol{\Sigma}\mathbf{t} - \tfrac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu}-\mathrm{i}\boldsymbol{\Sigma}\mathbf{t})'\boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu}-\mathrm{i}\boldsymbol{\Sigma}\mathbf{t})$$

where it can be shown similarly to (11.40) that the exponential of the last term integrates over $\mathbb{R}^m$ to $(2\pi)^{m/2}|\boldsymbol{\Sigma}|^{1/2}$. □

11.5 Infinite Divisibility

A distribution F is called *infinitely divisible* if for every $n \in \mathbb{N}$ there exists a distribution F_n such that F has a representation as the n-fold convolution

$$F = F_n * F_n * \ldots * F_n. \qquad (11.47)$$

As noted in §11.1 this is the distribution of the n-fold sum of independent terms having distribution F_n, and in view of (11.27) infinite divisibility implies a corresponding multiplicative rule for the ch.f.s. Since the distribution of sums of independent random variables provides the chief motivation for the study of characteristic functions, there is a natural interest in whether such sums give rise to a particular form of ch.f. Specifically, if $X_1,\ldots,X_n$ are an i.i.d. collection with identical ch.f.s ϕ_n, the ch.f. of $X_1 + \ldots + X_n$ must obey the relation $\phi = (\phi_n)^n$, for any $n \in \mathbb{N}$. If a ch.f. ϕ is linked to a ϕ_n by this identity for any n and this is the case even as $n \to \infty$, it is called infinitely divisible and clearly a special functional form is implied.

The best-known case is the normal distribution, noting from (11.41) that

$$\phi(t;\mu/n,\sigma/\sqrt{n})^n = \left(\mathrm{e}^{\mathrm{i}n^{-1}\mu t - n^{-1}\sigma^2t^2/2}\right)^n = \phi(t;\mu,\sigma).$$

Likewise, it can be verified that the Poisson ch.f. (11.37) has the property

$$\phi(t;\lambda/n)^n = \phi(t;\lambda).$$

The Cauchy family (11.42) has

$$\phi(t;\nu/n,\delta/n)^n = \phi(t;\nu,\delta)$$

and the gamma family (11.38) has

$$\phi(t;p/n,\lambda)^n = \phi(t;p,\lambda).$$

Also, don't overlook the degenerate distribution (11.43), noting that in that case,

$$\phi(t;a/n)^n = \phi(t;a).$$

These are the only infinitely divisible cases among the examples listed in §8.3. The binomial distribution does have the property that the ch.f. (11.36) is the n-fold product of Bernoulli ch.f.s (11.35), but this is not a case of infinite divisibility because n is a parameter of the distribution. There are no distributions having the property that a sum of (say) m drawings is binomial(p,n) for any choice of m.

These cases of infinite divisibility are special in the sense that ϕ_n is linked to ϕ only through the rescaling of one or other parameter of the distribution by a function of n. Otherwise, the ch.f.s look alike. Natural questions to ask are: can a more general class of distributions play the role of ϕ_n? (asymptotically the answer to this is certainly yes), and what form must ϕ generally take to be infinitely divisible? To investigate this latter question, since it is easier to work with sums than with products a useful step is to take logarithms. If ϕ is infinitely divisible then $\phi(t) \neq 0$ for all t and there must exist a sequence of ch.f.s $\{\phi_n, n \in \mathbb{N}\}$ such that

$$\log\phi(t) = \lim_{n\to\infty} n\log\phi_n(t). \tag{11.48}$$

To investigate this limit consider that when n is large, $\log\phi_n$ is necessarily close to zero and hence close to $\phi_n - 1$. The terms of the sequence have the approximate form

$$\log\phi(t) = n\log\phi_n(t) \approx n(\phi_n(t) - 1) \tag{11.49}$$

where the approximation is closer as n is larger. Letting F_n be the c.d.f. corresponding to ϕ_n, assume initially that the variance is finite. It is customary to decompose the formula in (11.49) as

$$n(\phi_n(t)-1) = n\int_{-\infty}^{\infty}(\mathrm{e}^{itx}-1)\mathrm{d}F_n(x)$$
$$= \mathrm{i}t\gamma_n + \int_{-\infty}^{\infty}(\mathrm{e}^{itx}-1-\mathrm{i}tx)\frac{1}{x^2}\mathrm{d}K_n(x)$$

where

$$\gamma_n = n\int_{-\infty}^{\infty}\xi\mathrm{d}F_n(\xi),\quad K_n(x) = n\int_{-\infty}^{x}\xi^2\mathrm{d}F_n(\xi).$$

Taking this representation to the limit as $n \to \infty$ gives the so-called canonical representation of the infinitely divisible ch.f.,

$$\log\phi(t) = \mathrm{i}t\gamma + \int_{-\infty}^{\infty}(\mathrm{e}^{\mathrm{i}tx}-1-\mathrm{i}tx)\frac{1}{x^2}\mathrm{d}K(x) \tag{11.50}$$

where $\gamma = \lim_{n\to\infty}\gamma_n$ and $K = \lim_{n\to\infty}K_n$. Equation (11.50) is known as *Kolmogorov's formula.* A noteworthy feature is that the integrand has the form $-\frac{1}{2}t^2 + O(x)$ as $x \to 0$.

Consider the application of this formula to the normal distribution. To have a finite limit requires setting F_n to be the c.d.f. of $X_1/\sqrt{n}$. If

$$\mathrm{d}F_n(x) = \frac{\sqrt{n}}{\sqrt{2\pi\sigma}}\exp\left\{-\frac{n(x-\mu)^2}{2\sigma^2}\right\}\mathrm{d}x \tag{11.51}$$

then

$$K_n(+\infty) = n\int_{-\infty}^{\infty}\xi^2\mathrm{d}F_n(\xi) = n\frac{\sigma^2}{n} = \sigma^2 \tag{11.52}$$

and similarly $\gamma_n = \mu$, both formulae not depending on n and therefore holding in the limit. However, in view of (11.51) K_n is converging as $n \to \infty$ to a step at zero. Assuming right-continuity, this means $K(0-) = 0$ and $K(0+) = \sigma^2$. Formula (11.50) therefore reduces, as required by (11.41), to

$$\log\phi(t) = \mathrm{i}t\mu - \tfrac{1}{2}\sigma^2t^2.$$

However, formulation (11.50) does not work for distributions such as the Cauchy which are without a finite variance, since in this case K_n is undefined. A version of the canonical form to accommodate distributions without moments incorporates a squashing function. The so-called *Lévy–Khinchine formula* is

$$\log\phi(t) = \mathrm{i}t\gamma + \int_{-\infty}^{\infty}\left(\mathrm{e}^{\mathrm{i}tx}-1-\frac{\mathrm{i}tx}{1+x^2}\right)\frac{1+x^2}{x^2}\mathrm{d}G(x) \tag{11.53}$$

where now

$$\gamma = \lim_{n\to\infty} n \int_{-\infty}^{\infty} \frac{\xi}{1+\xi^2} \mathrm{d}F_n(\xi) \tag{11.54}$$

and

$$G(x) = \lim_{n\to\infty} n \int_{-\infty}^{x} \frac{\xi^2}{1+\xi^2} \mathrm{d}F_n(\xi). \tag{11.55}$$

A key feature is that G is a nondecreasing function of bounded variation with $G(-\infty) = 0$ and $G(+\infty) < \infty$, whether or not the random variables possess a variance. It can also be verified that

$$\left(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{1+x^2}\right)\frac{1+x^2}{x^2} = -\frac{t^2}{2} + O(x). \tag{11.56}$$

An alternative form of (11.53), known as Lévy's formula, separates out the positive and negative parts of the distribution after centring by γ. Thus,

$$\begin{aligned}\log\phi(t) = \mathrm{i}t\gamma - \frac{\sigma^2 t^2}{2} &+ \int_{-\infty}^{0} \left(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{1+x^2}\right)\mathrm{d}M(x) \\ &+ \int_{0}^{\infty} \left(\mathrm{e}^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{1+x^2}\right)\mathrm{d}N(x)\end{aligned} \tag{11.57}$$

where $\sigma^2 = G(0+) - G(0-)$,

$$M(x) = \int_{-\infty}^{x} \frac{1+\xi^2}{\xi^2} \mathrm{d}G(\xi) = \lim_{n\to\infty} nF_n(x),\ x < 0, \tag{11.58}$$

and

$$N(x) = -\int_{x}^{\infty} \frac{1+\xi^2}{\xi^2} \mathrm{d}G(\xi) = \lim_{n\to\infty} n\big(F_n(x) - 1\big),\ x > 0. \tag{11.59}$$

In cases where the variance exists, $M = N = 0$ since nF_n reduces to the step at 0. Otherwise, $\sigma^2 = 0$. Take care to note how $N(x)$ is made negative so as to be nondecreasing in its argument, approaching 0 as $x \to \infty$ just as $M(x) \to 0$ as $x \to -\infty$.

An insightful way to interpret the formula in (11.57) is to note that the integrands, apart from the centring terms, are the logarithms of Poisson ch.f.s (compare **11.8**). The formula decomposes the distribution of an infinitely divisible r.v. as the sum of a Gaussian component with variance σ^2 and the aggregate of innumerable Poisson jumps of different magnitudes, depending on the integrator functions M and N whose form captures the tail behaviour of the distribution according to (11.58) and (11.59). If G has a jump at 0 so that $\sigma^2 > 0$, this means according to formula (11.55) that the contributions of the distributions F_n are to a significant degree infinitesimal, being concentrated around 0. If the contributions from outside the infinitesimal region are negligible such that M and N are zero, then of course (11.57) reduces to the (logarithm of the) Gaussian ch.f. Which of these cases arises is critical to the central limit behaviour of a normalized sum of independent r.v.s, a question to be explored in §23.6 and subsequently.

11.6 The Inversion Theorem

Paired with (11.9) is a unique inverse transformation so that the ch.f. and c.d.f. are fully equivalent representations of the distribution. The chief step in the proof of this proposition is the construction of the inverse transformation, which is shown as follows.

11.16 Lemma If $\phi(t) = \int \mathrm{e}^{\mathrm{i}tx}\mathrm{d}F(x)$ then for any pair a and b of continuity points of F with $a < b$,

$$F(b) - F(a) = \frac{1}{2\pi}\lim_{T\to\infty}\int_{-T}^{T}\frac{\mathrm{e}^{-\mathrm{i}ta} - \mathrm{e}^{-\mathrm{i}tb}}{\mathrm{i}t}\phi(t)\mathrm{d}t. \quad \Box \tag{11.60}$$

It can be verified from the series expansion of e^x in (11.7) that as $t \to 0$,

$$\frac{\mathrm{e}^{-\mathrm{i}ta} - \mathrm{e}^{-\mathrm{i}tb}}{\mathrm{i}t} \to b - a. \tag{11.61}$$

The integral in (11.60) is therefore well defined in spite of including the point $t = 0$. It is nonetheless necessary to avoid writing (11.60) as

$$F(b) - F(a) = \frac{1}{2\pi}\int_{-\infty}^{+\infty}\frac{\mathrm{e}^{-\mathrm{i}ta} - \mathrm{e}^{-\mathrm{i}tb}}{\mathrm{i}t}\phi(t)\mathrm{d}t \tag{11.62}$$

because the Lebesgue integral on the right may not exist. For example, suppose the random variable is degenerate at the point 0; this means that $\phi(t) = \mathrm{e}^{\mathrm{i}t.0} = 1$ for every t, and for $a < b$,

$$\frac{1}{2\pi}\int_{-T}^{T}\left|\frac{\mathrm{e}^{-\mathrm{i}ta} - \mathrm{e}^{-\mathrm{i}tb}}{\mathrm{i}t}\right|\mathrm{d}t \gg \int_{1}^{T}\frac{1}{t}\mathrm{d}t \sim \log T. \tag{11.63}$$

The criterion for Lebesgue integrability over $(-\infty, +\infty)$ fails. However, the limit in (11.60) does exist, as the proof reveals.

Proof of 11.16 After substituting for ϕ in (11.60), interchange the order of integration by applying **9.36**, whose continuity and a.s. boundedness requirements are certainly satisfied, to give

$$\begin{aligned}\int_{-T}^{T}\frac{\mathrm{e}^{-\mathrm{i}ta}-\mathrm{e}^{-\mathrm{i}tb}}{2\pi\mathrm{i}t}\phi(t)\mathrm{d}t &= \int_{-T}^{T}\frac{\mathrm{e}^{-\mathrm{i}ta}-\mathrm{e}^{-\mathrm{i}tb}}{2\pi\mathrm{i}t}\left(\int_{-\infty}^{\infty}\mathrm{e}^{\mathrm{i}tx}\mathrm{d}F(x)\right)\mathrm{d}t\\ &= \int_{-\infty}^{+\infty}\left(\int_{-T}^{T}\frac{\mathrm{e}^{\mathrm{i}t(x-a)}-\mathrm{e}^{\mathrm{i}t(x-b)}}{2\pi\mathrm{i}t}\mathrm{d}t\right)\mathrm{d}F(x).\end{aligned}\tag{11.64}$$

Applying the Euler identity (11.11), noting that the cosine is an even function so that the terms containing cosines (which are also the imaginary terms) vanish in the integral, gives

$$\int_{-T}^{T}\frac{\mathrm{e}^{\mathrm{i}t(x-a)}-\mathrm{e}^{\mathrm{i}t(x-b)}}{2\pi\mathrm{i}t}\mathrm{d}t = \int_{0}^{T}\frac{\sin t(x-a)}{\pi t}\mathrm{d}t - \int_{0}^{T}\frac{\sin t(x-b)}{\pi t}\mathrm{d}t. \tag{11.65}$$

The limit as $T \to \infty$ of this expression is obtained from the standard formula (see [85], 3.721)

$$\int_{0}^{\infty}\frac{\sin\alpha t}{t}\mathrm{d}t = \begin{cases}\pi/2, & \alpha > 0\\ 0, & \alpha = 0\\ -\pi/2, & \alpha < 0.\end{cases}\tag{11.66}$$

Substituting into (11.65) yields the result

$$\int_{-\infty}^{+\infty}\frac{\mathrm{e}^{\mathrm{i}t(x-a)}-\mathrm{e}^{\mathrm{i}t(x-b)}}{2\pi\mathrm{i}t}\mathrm{d}t = \begin{cases}0, & x < a \text{ or } x > b\\ \frac{1}{2}, & x = a \text{ or } x = b\\ 1, & a < x < b.\end{cases}\tag{11.67}$$

Letting $T \to \infty$ in (11.64) and applying the bounded convergence theorem now gives

$$\begin{aligned}\lim_{T\to\infty}\int_{-T}^{T}\frac{\mathrm{e}^{-\mathrm{i}ta}-\mathrm{e}^{-\mathrm{i}tb}}{2\pi\mathrm{i}t}\phi(t)\mathrm{d}t &= \int_{-\infty}^{+\infty}\left(\tfrac{1}{2}1_{\{a\}}+\tfrac{1}{2}1_{\{b\}}+1_{(a,b)}\right)\mathrm{d}F(x)\\ &= \tfrac{1}{2}\big(F(b)+F(b-)-F(a)-F(a-)\big).\end{aligned}\tag{11.68}$$

This reduces to $F(b) - F(a)$ when a and b are continuity points of F. ∎

If the distribution is absolutely continuous with respect to Lebesgue measure the inversion theory is simplified. The p.d.f. is the Fourier transform of the ch.f. and $\phi(t)$ is absolutely integrable, as can be verified for Examples **11.12**, **11.13**, and **11.9**, although **11.8** provides a counterexample. Using (11.60) and also (11.61) yields the inverse relation connecting the density and characteristic functions as

$$\begin{aligned} f(a) &= \lim_{h\to 0}\frac{F(a+h)-F(a)}{h} \\ &= \lim_{h\to 0}\lim_{T\to\infty}\int_{-T}^{T}\mathrm{e}^{-\mathrm{i}ta}\frac{1-\mathrm{e}^{-\mathrm{i}th}}{2\pi\mathrm{i}th}\phi(t)\mathrm{d}t \\ &= \frac{1}{2\pi}\int_{-\infty}^{\infty}\mathrm{e}^{-\mathrm{i}ta}\phi(t)\mathrm{d}t. \end{aligned} \tag{11.69}$$

The multivariate generalization of (11.60) is as follows, although the proof will not be given explicitly since it follows the lines of Lemma **11.16** closely and the main generalizations are in the notation.

11.17 Lemma If $\phi_{X_1,\dots,X_k}(t_1,\dots,t_k) = \mathrm{E}\big(\exp\{\mathrm{i}\sum_{j=1}^{k} t_jX_j\}\big)$, then

$$\Delta F(x_1,\dots,x_k) = \left(\frac{1}{2\pi}\right)^k \lim_{T\to\infty}\int_{-T}^{T}\cdots\int_{-T}^{T}\left(\prod_{j=1}^{k}\frac{\mathrm{e}^{-\mathrm{i}t_jx_j}-\mathrm{e}^{-\mathrm{i}t_j(x_j+\Delta x_j)}}{\mathrm{i}t_j}\right) \times \phi_{X_1,\dots,X_k}(t_1,\dots,t_k)\mathrm{d}t_1\dots\mathrm{d}t_k, \tag{11.70}$$

where $\Delta F(x_1,\dots,x_k)$ is defined in (8.36) and the vertices of the rectangle based at the point $x_1,\dots,x_k$, with sides $\Delta x_j > 0$, are all continuity points of F. □

Lemmas **11.16** and **11.17** are the basic ingredient of the result that primarily motivates the interest in characteristic functions, the so-called inversion theorem.

11.18 Theorem Distributions having the same ch.f. are the same.

Proof It suffices to give the proof for the univariate case. By (11.60), the c.d.f.s of the two distributions are the same at every point which is a continuity point of both c.d.f.s. Since the set of jump points of each c.d.f. is countable by **8.3**, their union is countable and it follows by **1.39** that the set of continuity points is dense in $\mathbb{R}$. It then follows by **8.5** that the c.d.f.s are the same everywhere. ∎

A simple application of the inversion theorem is to provide a proof of a well-known result, that affine functions of Gaussian vectors are also Gaussian.

11.19 Example Let $\boldsymbol{X} \sim_{\mathrm{d}} \mathrm{N}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ $(m\times 1)$ and $\boldsymbol{Y} = \boldsymbol{BX}+\boldsymbol{d}$ $(n\times 1)$ where $\boldsymbol{B}$ $(n\times m)$ and $\boldsymbol{d}$ $(n\times 1)$ are constants. Then, by (11.45),

$$\phi_Y(\boldsymbol{t}) = \exp\{\mathrm{i}\boldsymbol{t}'\boldsymbol{d}\}\mathrm{E}\big(\exp\{\mathrm{i}\boldsymbol{t}'\boldsymbol{BX}\}\big) = \exp\big\{\mathrm{i}\boldsymbol{t}'(\mathbf{B}\boldsymbol{\mu}+\boldsymbol{d}) - \tfrac{1}{2}\boldsymbol{t}'\mathbf{B}\boldsymbol{\Sigma}\mathbf{B}'\boldsymbol{t}\big\}. \tag{11.71}$$

If rank($\mathbf{B\Sigma B}'$) $= n$ (implying $n \leq m$), **11.18** implies that $\boldsymbol{Y}$ has p.d.f.

$$f(y) = \frac{\exp\{-\frac{1}{2}(\boldsymbol{y} - \mathbf{B}\boldsymbol{\mu} - \boldsymbol{d})'(\mathbf{B\Sigma B}')^{-1}(\boldsymbol{y} - \mathbf{B}\boldsymbol{\mu} - \boldsymbol{d})\}}{(2\pi)^{n/2}|\mathbf{B\Sigma B}'|^{1/2}}. \tag{11.72}$$

Compare this formula with (8.46). If rank ($\mathbf{B\Sigma B}'$) $< n$, (11.71) remains valid although (11.72) is not. But by the same arguments, every linear combination $\boldsymbol{c}'\boldsymbol{Y}$ where $\boldsymbol{c}$ is $p \times 1$ is either scalar Gaussian with variance $\boldsymbol{c}'\mathbf{B\Sigma B}'\boldsymbol{c}$, or the constant $\boldsymbol{c}'\boldsymbol{d}$, corresponding to the cases $\boldsymbol{B}'\boldsymbol{c} \neq \boldsymbol{0}$ and $\boldsymbol{B}'\boldsymbol{c} = \boldsymbol{0}$ respectively. In this case $\boldsymbol{Y}$ is said to have a singular Gaussian distribution. □

11.7 The Conditional Characteristic Function

Let Y be a $\mathcal{F}$-measurable r.v. and let $\mathcal{G} \subset \mathcal{F}$. The conditional ch.f. of $Y|\mathcal{G}$, $\phi_{Y|\mathcal{G}}(t)$, is for each t a random variable having the property

$$\int_G \phi_{Y|\mathcal{G}}(t)\mathrm{d}P = \int_G \mathrm{e}^{\mathrm{i}tY}\mathrm{d}P, \text{ all } G \in \mathcal{G}. \tag{11.73}$$

The conditional ch.f. shares the properties of the regular ch.f. whenever the theory of conditional expectations parallels that of ordinary expectations according to the results of Chapter 10. Its real and imaginary parts are, respectively, the $\mathcal{G}$-measurable random variables $\mathrm{E}(\cos tY|\mathcal{G})$ and $\mathrm{E}(\sin tY|\mathcal{G})$. It can be expanded as in **11.5** in terms of the existing conditional moments. If X is $\mathcal{G}$-measurable, the conditional ch.f. of $X + Y$ is $\phi_{X+Y|\mathcal{G}}(t) = e^{\mathrm{i}tX}\mathrm{E}(e^{\mathrm{i}tY}|\mathcal{G})$ by **10.10**. And if Y is $\mathcal{G}_1$-measurable and $\mathcal{G}$ and $\mathcal{G}_1$ are independent sub-σ-fields, then $\phi_{Y|\mathcal{G}}(t) = \phi_Y(t)$ a.s.

The conditional ch.f. is used to prove a useful inequality due to von Bahr and Eséen ([184]). Start with a technical lemma that appears obscure at first sight but turns out to have useful applications.

11.20 Lemma If $\mathrm{E}|Z|^r < \infty$ for $0 < r < 2$, then

$$\mathrm{E}|Z|^r = K(r)\int_{-\infty}^{+\infty} \frac{1 - \mathrm{Re}\big(\phi_Z(t)\big)}{|t|^{1+r}}\mathrm{d}t \tag{11.74}$$

where $K(r) = \left(\int_{-\infty}^{+\infty}(1 - \cos u)/|u|^{1+r}\mathrm{d}u\right)^{-1}$. □

It can be shown that $K(r)$ has the solution $\Gamma(1 + r)\sin(\pi r/2)/\pi$, which is positive for $0 < r < 2$.

Proof of 11.20 The identity for real z,

$$|z|^r = K(r)\int_{-\infty}^{+\infty}\frac{1-\cos zt}{|t|^{1+r}}\mathrm{d}t, \tag{11.75}$$

is easily obtained by a change of variable in the integral on the right. The lemma follows on applying **9.36** and noting that $\mathrm{Re}\big(\phi_Z(t)\big) = \mathrm{E}(\cos tZ)$. ∎

Equality (11.74) also holds, for $\omega \in C$ with $P(C) = 1$, if $\mathrm{E}(|Z|^r|\mathcal{G})(\omega)$ and $\phi_{Z|\mathcal{G}}(t)(\omega)$ are substituted for $\mathrm{E}|Z|^r$ and $\phi_Z(t)$. In other words, the conditional r^{th} moment and conditional ch.f. are almost surely related by the same formula. Therefore consider L_r-bounded r.v.s Z and X, where Z is $\mathcal{F}$-measurable and X is $\mathcal{G}$-measurable for $\mathcal{G} \subset \mathcal{F}$.

First, suppose that $\phi_{Z|\mathcal{G}}(t)$ is a real r.v. almost surely. Then for each $\omega \in \Omega$,

$$\begin{aligned}1-\mathrm{Re}\big(\phi_{X+Z|\mathcal{G}}(t)\big)(\omega) &= 1-\mathrm{Re}\big(\mathrm{e}^{\mathrm{i}tX}\phi_{Z|\mathcal{G}}(t)\big)(\omega)\\ &= 1-\big(\cos tX(\omega)\big)\phi_{Z|\mathcal{G}}(t)(\omega)\\ &\le \big(1-\cos tX(\omega)\big)+\big(1-\phi_{Z|\mathcal{G}}(t)(\omega)\big),\end{aligned} \tag{11.76}$$

the difference between the last two members being $\big(1-\cos tX(\omega)\big)$ $\big(1-\phi_{Z|\mathcal{G}}(t)(\omega)\big)$, which is non-negative for all ω. Hence, for $0 < r < 2$,

$$\begin{aligned}\mathrm{E}(|X+Z|^r|\mathcal{G}) &= K(r)\int_{-\infty}^{+\infty}\frac{1-\mathrm{Re}\big(\phi_{X+Z|\mathcal{G}}(t)\big)}{|t|^{1+r}}\mathrm{d}t\\ &\le K(r)\int_{-\infty}^{+\infty}\frac{1-\cos tX}{|t|^{1+r}}\mathrm{d}t + K(r)\int_{-\infty}^{+\infty}\frac{1-\phi_{Z|\mathcal{G}}(t)}{|t|^{1+r}}\mathrm{d}t\\ &= |X|^r + \mathrm{E}(|Z|^r|\mathcal{G}),\ \text{a.s.,}\end{aligned} \tag{11.77}$$

and taking expectations through yields

$$\mathrm{E}|X+Z|^r \le \mathrm{E}|X|^r + \mathrm{E}|Z|^r. \tag{11.78}$$

For the case $0 < r \le 1$ this inequality holds by the c_r inequality for general Z and X, so it is the case $1 < r < 2$ that is of special interest here.

Generalizing from the remarks following (11.25), the condition that $\phi_{Z|\mathcal{G}}(t)$ be real almost surely is fulfilled by letting $Z = Y - Y'$ where Y and Y' are identically distributed and independent, conditional on $\mathcal{G}$. Note that if $\mathcal{H} = \sigma(Y)$, then

$$\mathrm{E}(Y'|\mathcal{H}) = \mathrm{E}(Y'|\mathcal{G})\ \text{a.s.} \tag{11.79}$$

by **10.32**. Identical conditional distributions mean simply that $F_Y(\cdot|\mathcal{G}) = F_{Y'}(\cdot|\mathcal{G})$ a.s. and equivalently that $\phi_{Y|\mathcal{G}}(t) = \phi_{Y'|\mathcal{G}}(t)$ a.s. Hence

$$\begin{aligned}\phi_{Y-Y'|\mathcal{G}}(t) &= \mathrm{E}(\mathrm{e}^{\mathrm{i}tY}\mathrm{e}^{-\mathrm{i}tY'}|\mathcal{G}) \\ &= \mathrm{E}(\mathrm{e}^{\mathrm{i}tY}|\mathcal{G})\mathrm{E}(\mathrm{e}^{-\mathrm{i}tY}|\mathcal{G}) \\ &= |\phi_{Y|\mathcal{G}}(t)|^2 \text{ a.s.,} \qquad (11.80)\end{aligned}$$

where the right-hand side is a real r.v. Now, for each $\omega \in \Omega$ the following identity can be verified:

$$2\big(1 - \mathrm{Re}(\phi_{Y|\mathcal{G}}(t)(\omega))\big) = 1 - |\phi_{Y|\mathcal{G}}(t)(\omega)|^2 + |1 - \phi_{Y|\mathcal{G}}(t)(\omega)|^2. \qquad (11.81)$$

Applying (11.80) and **11.20** and taking expectations, this yields the inequality

$$2\mathrm{E}|Y|^r \geq \mathrm{E}|Y - Y'|^r,\ 0 < r < 2, \qquad (11.82)$$

noting that the difference between the two sides here is the non-negative function of r, $K(r)\int_{-\infty}^{+\infty} \mathrm{E}|1 - \phi_{Y|\mathcal{G}}(t)|^2/|t|^{1+r}\mathrm{d}t$.

These arguments lead us to the following conclusion.

11.21 Theorem Suppose $\mathrm{E}(Y|\mathcal{G}) = 0$ a.s. and X is $\mathcal{G}$-measurable where $\mathcal{G} \subseteq \mathcal{H} = \sigma(Y)$ and both variables are L_r-bounded. Then

$$\mathrm{E}|X + Y|^r \leq \mathrm{E}|X|^r + 2\mathrm{E}|Y|^r,\ 0 \leq r \leq 2. \qquad (11.83)$$

Proof Let Y' be independent and distributed identically with Y, conditional on $\mathcal{G}$. Applying (11.79), these conditions jointly imply $\mathrm{E}(Y'|\mathcal{H}) = \mathrm{E}(Y'|\mathcal{G}) = \mathrm{E}(Y|\mathcal{G}) = 0$. Noting that $X + Y$ is $\mathcal{H}$-measurable, it follows by **10.20** (in applying this result be careful to note that $\mathcal{H}$ plays the role of the sub-σ-field here) that

$$\mathrm{E}|X + Y|^r \leq \mathrm{E}|X + (Y - Y')|^r. \qquad (11.84)$$

The conclusion for $1 < r < 2$ now follows on applying (11.78) for the case $Z = Y - Y'$ and then (11.82). The inequality holds for $0 < r \leq 1$ by the c_r inequality and for $r = 2$ from elementary considerations since $\mathrm{E}(YX) = 0$. In these latter cases the factor 2 in (11.83) can be omitted. ∎

This result can be iterated, given a sequence of r.v.s measurable on an increasing sequence of σ-fields. An easy application is to independent zero-mean r.v.s $X_1, \ldots, X_n$, for which the condition $\mathrm{E}\big(X_t|\sigma(X_1, \ldots, X_{t-1})\big) = 0$ holds for $t = 2, \ldots, n$. Letting $S_n = \sum_{t=1}^{n} X_t$ and $\sigma(S_n) = \sigma(X_1, \ldots, X_n)$, **11.21** yields for $0 \le r \le 2$,

$$\mathrm{E}|S_n|^r \le \mathrm{E}|S_{n-1}|^r + 2\mathrm{E}|X_n|^r \le \ldots \le 2\sum_{t=1}^{n} \mathrm{E}|X_t|^r. \tag{11.85}$$

If the series in the majorant converges, this inequality remains valid as $n \to \infty$. It may be contrasted for sharpness with the c_r inequality for general X_t, (9.63). In this case 2 must be replaced by n^{r-1} for $1 < r \le 2$, which is of no use for large n.

PART III

THEORY OF STOCHASTIC PROCESSES

12
Stochastic Processes

12.1 Basic Ideas and Terminology

Let $(\Omega, \mathcal{F}, P)$ be a probability space, let $\mathbb{T}$ be any set, and let $\mathbb{R}^{\mathbb{T}}$ be the product space generated by taking a copy of $\mathbb{R}$ for each element of $\mathbb{T}$. Then, a *stochastic process* is a measurable mapping $x : \Omega \mapsto \mathbb{R}^{\mathbb{T}}$, where

$$x(\omega) = \{X_\tau(\omega), \tau \in \mathbb{T}\}. \tag{12.1}$$

$\mathbb{T}$ is called the *index set* and the r.v. $X_\tau(\omega)$ is called a *coordinate* of the process. A stochastic process can also be characterized as a mapping from $\Omega \times \mathbb{T}$ to $\mathbb{R}$. However, the significant feature of the definition given is the requirement of *joint* measurability of the coordinates. Something more is implied than having $X_\tau(\omega)$ a measurable r.v. for each τ.

Here, $\mathbb{T}$ is an arbitrary set which in principle need not even be ordered, although linear ordering characterizes the important cases. A familiar example is $\mathbb{T} = \{1, \dots, k\}$, where x is a random k-vector. Another important case of $\mathbb{T}$ is an interval of $\mathbb{R}$, such that $x(\omega)$ is a function of a real variable and $\mathbb{R}^{\mathbb{T}}$ the space of random functions. And when $\mathbb{T}$ is a countable subset of $\mathbb{R}$, $x = \{X_\tau(\omega), \tau \in \mathbb{T}\}$ defines a *stochastic sequence* (equivalently, *random sequence*). Thus, a stochastic sequence is a stochastic process whose index set is countable and linearly ordered. When the X_τ represent random observations equally spaced in time, no relevant information is lost by assigning a linear ordering through $\mathbb{N}$ or $\mathbb{Z}$, indicated by the notations $\{X_\tau(\omega)\}_1^\infty$ and $\{X_\tau(\omega)\}_{-\infty}^\infty$. The definition does not rule out $\mathbb{T}$ containing information about distances between the sequence coordinates, as when the observations are irregularly spaced in time with τ a real number representing elapsed time from a chosen origin. A case of this kind is the Poisson process described in §13.2.

Familiarly, a *time series* is a time-ordered sequence of observations, although the term may extend to unobserved or hypothetical variables, such as the errors in a regression model. Time-series coordinates are labelled t. If a *sample* is defined as a time series of finite length n (or more generally, a collection of such series for different variables), it is convenient to assume that samples are embedded in infinite sequences of 'potential' observations. Various mathematical functions of sample observations, *statistics* or *estimators*, will also be well known to the

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0012

reader, characteristically involving a summation of terms over the coordinates. The sample moments of a time series, regression coefficients, and log-likelihood functions and their derivatives are standard examples. By letting n take the values $1,2,3,\ldots$, these functions of n observations generate *derived* sequences. The notion of a sequence in this case comes from the idea of analysing samples of progressively increasing size. The mathematical theory often does not distinguish between the types of sequence under consideration. Some definitions and results apply generally, but a clue to the usual application will be given by the choice of index symbol, t or n as the case may be.

A leading case which does not fall under the definition of a sequence is where $\mathbb{T}$ is partially ordered. When there are two dimensions to the observations, as in a panel data set having both a time dimension and a dimension over agents, x may be called a *random field.* Such cases are not treated explicitly here, although in many applications one dimension is regarded as fixed and the sequence notion is adequate for asymptotic analysis. However, cases where $\mathbb{T}$ is either the product set $\mathbb{Z} \times \mathbb{N}$, or a subset thereof, are often met below in a different context. A *triangular stochastic array* is a doubly indexed collection of random variables,

$$\begin{bmatrix} X_{11} & X_{21} & X_{31} & \cdots \\ X_{12} & X_{22} & X_{32} & \cdots \\ \vdots & \vdots & \vdots & \\ X_{1k_1} & \vdots & \vdots & \\ & X_{2k_2} & \vdots & \\ & & X_{3k_3} & \\ & & & \ddots \end{bmatrix}, \tag{12.2}$$

compactly written as $\{\{X_{nm}\}_{m=1}^{k_n}\}_{n=1}^{\infty}$, where $\{k_n\}_{n=1}^{\infty}$ is some increasing integer sequence. Array notation is called for when the points of a sample are subjected to scale transformations or the like, depending on the complete sample. A standard example is $\{\{X_{nt}\}_{t=1}^{n}\}_{n=1}^{\infty}$, where $X_{nt} = X_t/s_n$ and $s_n^2 = \sum_{t=1}^{n} \mathrm{Var}(X_t)$, or some similar function of the sample moments from 1 to n.

12.2 Convergence of Stochastic Sequences

Consider the functional expression $\{X_n(\omega)\}_1^{\infty}$ for a random sequence on the space $(\Omega,\mathcal{F},P)$. When evaluated at a point $\omega \in \Omega$ this denotes a *realization* of the sequence, the actual collection of real numbers generated when the outcome ω is drawn. It is natural to consider in the spirit of ordinary analysis whether this sequence converges to a limit, say $X(\omega)$. If this is the case for every $\omega \in \Omega$ we would say that $X_n \to X$ surely (or elementwise) where, if X_n is an $\mathcal{F}/\mathcal{B}$-measurable r.v. for each n, then so is X, by **3.33**.

However, except by direct construction it is usually difficult to establish in terms of a given collection of distributional properties that a stochastic sequence converges surely to a limit. A much more useful notion (because more easily shown) is *almost sure convergence*. Let $C \subseteq \Omega$ be the set of outcomes such that, for every $\omega \in C$, $X_n(\omega) \to X(\omega)$ as $n \to \infty$. If $P(C) = 1$, the sequence is said to converge almost surely, or equivalently, *with probability one*. The notations $X_n \to_{\text{a.s.}} X$ or $X_n \to X$ a.s. and also a.s. $\lim X_n = X$ are all used to denote almost sure convergence. A similar concept, convergence almost everywhere (a.e.), was invoked in connection with the properties of integrals in §4.2. For many purposes, almost sure convergence can be thought of as yielding the same implications as sure convergence in probabilistic arguments.

However, attaching probabilities to the convergent set is not the only way in which stochastic convergence can be understood. Associated with any stochastic sequence are various non-stochastic sequences of variables and functions describing aspects of its behaviour, moments being the obvious case. Convergence of the stochastic sequence may be defined in terms of the ordinary convergence of an associated sequence. If the sequence $\{\mathrm{E}(X_n - X)^2\}_1^\infty$ converges to zero there is clearly a sense in which X_n converges to X; this is called *convergence in mean square* or equivalently, convergence in L_2-norm. Or suppose that for any $\varepsilon > 0$, the probabilities of the events $\{\omega : |X_n(\omega) - X(\omega)| < \varepsilon\} \in \mathcal{F}$ form a real sequence converging to 1. This is another distinct convergence concept, so-called *convergence in probability*. In neither case is there any obvious way to attach a probability to the convergent set; this can even be zero! These issues are studied in Part IV.

Another convergence concept relates to the sequence of marginal p.m.s of the coordinates, $\{\mu_n\}_1^\infty$, or equivalently the marginal c.d.f.s, $\{F_n\}_1^\infty$. Here the sequences in question are of the form $\{\mu_n(A)\}_1^\infty$ for various sets $A \in \mathcal{B}$ or alternatively $\{F_n(x)\}_1^\infty$ for various $x \in \mathbb{R}$. In the latter case, pointwise convergence on $\mathbb{R}$ is a possibility but this is a relatively strong notion. It is sufficient for a theory of the limiting distribution if convergence is confined just to the continuity points of the limiting function F, or equivalently to the sequences $\{\mu_n(A)\}$ for which $\mu(\delta A) = 0$. This condition is referred to as the *weak convergence* of the distributions and forms the subject of Part V.

12.3 The Probability Model

Some very important ideas are implicit in the notion of a stochastic sequence. Given the equipotency of $\mathbb{N}$ and $\mathbb{Z}$, it suffices to consider the random element $\{X_\tau(\omega)\}_1^\infty$ for $\omega \in \Omega$, mapping from a point of Ω to a point in infinite-dimensional Euclidean space $\mathbb{R}^\infty$. From a probabilistic point of view, the entire infinite sequence corresponds to a *single* outcome ω of the underlying abstract probability space. In principle, a sampling exercise in this framework is the random drawing of a point in $\mathbb{R}^{\infty}$, called a realization or *sample path* of the random sequence; in practice only a finite segment of this sequence is observed, but the key idea is that a random experiment consists of drawing a complete realization. *Repeated sampling* means observing the same finite segment (relative to the origin of the index set) of different realizations, *not* different segments of the same realization.

The reason for this characterization of the random experiment will become clear in the sequel. Meanwhile, a simple example may place the slightly outlandish notion of an infinite-dimensioned random element into perspective, by showing the close correspondence between a random sequence and a probability space of familiar type.

12.1 Example Consider a repeated game of coin tossing, generating a random sequence of heads and tails. If the game continues for ever it will generate a sequence of infinite length. Let 1 represent a head and 0 a tail, to define a random sequence of 1s and 0s. Such a sequence corresponds to the binary digits in the base 2 representation of a real number. According to equation (1.15) with $D = 2$, $p = 0$, and $M = 1$ there is a one-to-one correspondence between infinite sequences of coin tosses and points on the unit interval. If the coin toss sequence begins $110100100011\ldots$ (say), just append a leading decimal point to obtain the real number $0.110100100011\ldots \in [0,1]$ (approximately 0.821 in decimal).

On this basis, the fundamental space $(\Omega, \mathcal{F})$ for the coin tossing experiment can be chosen as $\big([0,1), \mathcal{B}_{[0,1)}\big)$. The form of P can be deduced in an elementary way from the stipulation that $P(\text{heads}) = P(\text{tails}) = 0.5$ (i.e. the coin is fair) and successive tosses are independent. For example, the events {tails on first toss} and {heads on first toss} are the images of the sets $[0, 0.5)$ and $[0.5, 1)$ respectively, whose measures must accordingly be 0.5 each. More generally, the probability that the first n tosses in a sequence yields a given configuration of heads and tails out of the 2^n possible ones is equal in every case to $1/2^n$, so that each sequence is (in an appropriate limiting sense) 'equally likely'. The corresponding sets in $[0,1]$ of the binary expansions with the identical pattern of 0s and 1s in the first n positions occupy intervals all of width precisely $1/2^n$ in the unit interval. The conclusion is that the probability measure of any interval is equal to its width. This is nothing but Lebesgue measure on the unit interval. The upper boundary point can be included after observing that the sequence containing only heads, yielding $0.11111\ldots$ with 1 recurring, is identical with 1.0. □

This example can be elaborated from binary sequences to sequences of real variables without too much difficulty. There is an intimate connection between infinite random sequences and continuous probability distributions on the line, and understanding one class of problem is frequently an aid to understanding the other. The question often posed about the probability of some sequence predicted in advance being realized, say an infinite run of heads or a perpetual alternation of heads and tails, is precisely answered. In either the decimal or binary expansions, all the numbers whose digit sequences either terminate or, beyond some point, are found to cycle perpetually through a finite sequence belong to the set of rational numbers. Since the rationals have Lebesgue measure zero in the space of the reals, the probability of *any* such sequence occurring is zero.

Another well-known conundrum concerns the troupe of monkeys equipped with typewriters who, it is claimed, will eventually type out the complete works of Shakespeare. This event will occur with probability 1. For the sake of argument assume that a single monkey types into a word processor and his ASCII-encoded output takes the form of a string of bits (binary digits). Suppose Shakespeare's encoded complete works occupy k bits, equivalent to $k/5$ characters allowing for a 32-character keyboard (upper-case only, but including some punctuation marks). This string is one of the 2^k possible strings of k bits. Assuming that each such string is equally likely to arise in $k/5$ random key presses, the probability that the monkey will type Shakespeare without an error *from scratch* is exactly 2^{-k}. However, the probability that the second string of 2^k bits it produces is the right one, given that the first one is wrong, is $(1-2^{-k})2^{-k}$ when the strings are independent. In general, the probability that the monkey types Shakespeare correctly on the $(m+1)^{\text{th}}$ independent attempt, given that the first m attempts were failures, is $(1-2^{-k})^m 2^{-k}$. These events are disjoint and summing their probabilities over all $m \geq 0$ yields

$$P(\text{monkey types Shakespeare eventually}) = 1.$$

In the meantime, of course, the industrious primate has produced much of the rest of world literature, not to mention a good many telephone books. It is also advisable to estimate the length of time it will take for the desired text to appear, which requires a further calculation. The average waiting time, expressed in units of the time taken to type k bits, is $2^{-k}\sum_{m=1}^{\infty} m(1-2^{-k})^m = 2^k - 1$. If we scale down our ambitions and decide to be content with just 'TO BE OR NOT TO BE' ($5 \times 18 = 90$ bits) and the monkey takes 1 minute over each attempt, we shall wait on average 2.3×10^{21} years. So the Complete Works don't really bear thinking about.

What has been shown is that almost every infinite string of bits contains every finite string somewhere in its length; but also, that the mathematical concept of 'almost surely' has no difficulty in coinciding with an everyday notion indistinguishable from 'never'. The example is frivolous, but it is useful to be reminded occasionally that limit theory deals in large numbers. A sense of perspective is always desirable in evaluating the claims of the theory.

In the theory of stochastic processes, the first technical challenge is to handle distributions on $\mathbb{R}^\infty$. To construct the Borel field $\mathcal{B}^\infty$ of events on $\mathbb{R}^\infty$, implicitly endow $\mathbb{R}^\infty$ with the Tychonoff, or product, topology. It is not essential to have absorbed the theory of §6.5 to make sense of the discussion that follows, but it may help to glance at Example **6.18** to see what this assumption implies.

Given a process $x = \{X_t\}_1^\infty$, write

$$\pi_k(x) = (X_1, \dots, X_k) : \mathbb{R}^\infty \mapsto \mathbb{R}^k \tag{12.3}$$

for each $k \in \mathbb{N}$, to denote the k-dimensional coordinate projection. Let $\mathcal{C}$ denote the collection of *finite-dimensional cylinder sets* of $\mathbb{R}^\infty$, the sets

$$C = \left\{x \in \mathbb{R}^\infty : \pi_k(x) \in E, E \in \mathcal{B}^k, k \in \mathbb{N}\right\}. \tag{12.4}$$

In other words, elements of $\mathcal{C}$ have the form

$$C = \pi_k^{-1}(E) \tag{12.5}$$

for some $E \in \mathcal{B}^k$ and some finite k. There is no loss of generality in considering projections onto coordinates $1, \dots, k$, since any finite-dimensional cylinder can be embedded in a cylinder of the form $\pi_k^{-1}(E)$, $E \in \mathcal{B}^k$, where k is just the largest of the restricted coordinates. The distinguishing feature of an element of $\mathcal{C}$ is that at most a finite number of its coordinates are restricted.

12.2 Theorem $\mathcal{C}$ is a field.

Proof First, the complement in $\mathbb{R}^\infty$ of a set C defined by (12.4) is

$$C^c = \left\{x \in \mathbb{R}^\infty : \pi_k(x) \in E^c, E \in \mathcal{B}^k\right\} = \pi_k^{-1}(E^c) \tag{12.6}$$

which is another element of $\mathcal{C}$, that is, $C^c \in \mathcal{C}$. Second, consider the union of sets $C = \pi_k^{-1}(E) \in \mathcal{C}$ and $C' = \pi_k^{-1}(E') \in \mathcal{C}$, for $E, E' \in \mathcal{B}^k$. $C \cup C'$ is given by (12.4) with E replaced by $E \cup E'$ and hence $C \cup C' \in \mathcal{C}$. Third, if $E \in \mathcal{B}^k$ and $E' \in \mathcal{B}^m$ for $m > k$, since $E \times \mathbb{R}^{m-k} \in \mathcal{B}^m$ and $\pi_k^{-1}(E) = \pi_m^{-1}(E \times \mathbb{R}^{m-k}) \in \mathcal{C}$ the argument of the second case applies. ∎

It is not easy to imagine sets in arbitrary numbers of dimensions, but good visual intuition is provided by thinking about one-dimensional cylinders in $\mathbb{R}^3$. Letting (x, y, z) denote the coordinate directions, the one-dimensional cylinder generated by an interval of the x axis is a region of 3-space bounded by two infinite planes at right angles to the x axis. See Figure 12.1 for a cut-away representation. A union of

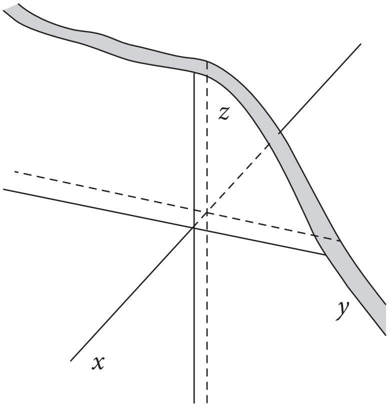

Figure 12.1

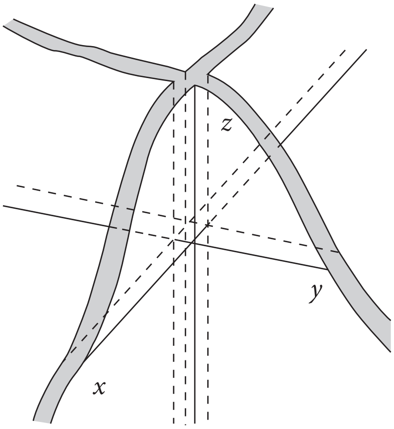

Figure 12.2

x-cylinders is another x-cylinder, a collection of parallel 'walls'. But the union and intersection of an x-cylinder with a y-cylinder are two-dimensional cylinder sets, a 'cross' and a 'column' respectively (see Figure 12.2).

These examples show that the collection of cylinder sets in $\mathbb{R}^k$ for fixed k is *not* a field; the intersection of three mutually orthogonal 'walls' in $\mathbb{R}^3$ is a bounded 'cube', not a cylinder set. The set of finite-dimensional cylinders is not closed under the operations of union and complementation (and hence intersection) *except* in an infinite-dimensional space. This fact is critical in considering $\sigma(\mathcal{C})$, the class obtained by adding the countable unions to $\mathcal{C}$. By the last-mentioned property of unions, $\sigma(\mathcal{C})$ includes sets of the form (12.4) with k tending to infinity.

12.3 Theorem $\sigma(\mathcal{C}) = \mathcal{B}^\infty$, the Borel field of sets in $\mathbb{R}^\infty$ with the Tychonoff topology. □

This is something taken for granted in the usual applications. Recalling that the Borel field of a space is the smallest σ-field containing the open sets, **12.3** is true by definition since $\mathcal{C}$ is a sub-base for the product topology (see §6.5) and all the open sets of $\mathbb{R}^\infty$ are generated by unions and finite intersections of $\mathcal{C}$-sets. To avoid explicit topological considerations, the reader may like to think of **12.3** as providing the definition of $\mathcal{B}^\infty$.

One straightforward implication, since the coordinate projections are continuous mappings and hence measurable, is that, given a distribution on $(\mathbb{R}^\infty, \mathcal{B}^\infty)$, finite collections of sequence coordinates can always be treated as random vectors. But, while this is obviously a condition that will need to be satisfied, the real problem runs the other way. The only *practical* method of defining distributions for sequences is to assign probabilities to finite collections of coordinates, after the manner of §8.4. The serious question is whether this can be done in a consistent manner, so that in particular there exists a p.m. on $(\mathbb{R}^\infty, \mathcal{B}^\infty)$ that corresponds to a given set of the finite-dimensional distributions. The affirmative answer to this question is the famous Kolmogorov consistency theorem.

12.4 The Consistency Theorem

The goal is to construct a p.m. on $(\mathbb{R}^\infty, \mathcal{B}^\infty)$ and, following the approach of §3.2, the plausible first step in this direction is to assign probabilities to elements of $\mathcal{C}$. Let μ_k denote a p.m. on the space $(\mathbb{R}^k, \mathcal{B}^k)$ for $k = 1, 2, 3, \ldots$. This family of measures satisfies the *consistency property* if

$$\mu_k(E) = \mu_m(E \times \mathbb{R}^{m-k}) \tag{12.7}$$

for $E \in \mathcal{B}^k$ and all $m > k > 0$. In other words, any k-dimensional distribution can be obtained from an m-dimensional distribution with $m > k$, by the usual operation of marginalization.

The consistency theorem actually generalizes to stochastic processes with uncountable index sets $\mathbb{T}$ (see **27.10**) but it is sufficient for present purposes to consider the countable case. Although priority is attributed to P. J. Daniell, the following result adapted from [114], III.4 is generally known as Kolmogorov's consistency theorem.

12.4 Theorem Suppose there exists a family of p.m.s $\{\mu_k\}$ that satisfy consistency condition (12.7). Then there exists a stochastic sequence $x = \{X_t, t \in \mathbb{N}\}$ in a probability space $(\mathbb{R}^\infty, \mathcal{B}^\infty, \mu)$ such that μ_k is the p.m. of the finite vector of coordinate functions $(X_1, \ldots, X_k)$. □

The candidate measure for x is defined for sets in $\mathcal{C}$ by

$$\mu(C) = \mu_k(E) \tag{12.8}$$

where C and E are related by (12.4). The problem is to show that μ is a p.m. on $\mathcal{C}$. If this is the case then, since $\mathcal{C}$ is a field and $\mathcal{B}^\infty = \sigma(\mathcal{C})$, the extension theorem (**3.15** + **3.20**) establishes the existence of a unique measure on $(\mathbb{R}^\infty, \mathcal{B}^\infty)$ which agrees with μ for all $C \in \mathcal{C}$. The theorem has a simple but important corollary.

12.5 Corollary $\mathcal{C}$ is a determining class for $(\mathbb{R}^\infty, \mathcal{B}^\infty)$. □

In other words, if μ and ν are two measures on $(\mathbb{R}^\infty, \mathcal{B}^\infty)$ and $\mu_k = \nu_k$ for every finite k, then $\mu = \nu$.

To prove the consistency theorem requires a technical lemma that is quite intuitive but whose proof is beyond us at this stage. It will be proved in a more general context as **29.14** (see page 654).

12.6 Lemma For every $E \in \mathcal{B}^k$ and $\delta > 0$ there exists K, a compact subset of E, such that $\mu_k(E - K) < \delta$. □

In other words, a p.m. on the space $(\mathbb{R}^k, \mathcal{B}^k)$ has nearly all of its mass confined to a compact set; this implies in particular the proposition asserted in §8.1, that random variables are finite almost surely.

Proof of 12.4 This is by verifying that μ of (12.8) satisfies the probability axioms with respect to elements of $\mathcal{C}$. When $E = \mathbb{R}^k$, $C = \mathbb{R}^\infty$ so that the first two probability axioms, **7.1**(a) and (b), are certainly satisfied. To establish finite additivity, suppose there are $\mathcal{C}$-sets $C = \pi_k^{-1}(E)$ and $C' = \pi_m^{-1}(E')$ for $E \in \mathcal{B}^m$ and $m \geq k$. If C and C' are disjoint,

$$\begin{aligned}\mu(C) + \mu(C') &= \mu_k(E) + \mu_m(E') \\ &= \mu_m(E \times \mathbb{R}^{m-k}) + \mu_m(E') \\ &= \mu_m(E \times \mathbb{R}^{m-k} \cup E') = \mu(C \cup C')\end{aligned} \tag{12.9}$$

where the second equality applies the consistency condition (12.7) and the third uses the fact that $E \times \mathbb{R}^{m-k}$ and E' are disjoint if C and C' are.

The remaining, relatively tricky, step is to extend finite additivity to countable additivity. This is done by proving continuity, which is an equivalent property according to **3.8**. If and only if the measure is continuous, a monotone sequence $\{C_j \in \mathcal{C}\}$ such that $C_j \downarrow C$ or $C_j \uparrow C$ has the property, $\mu(C_j) \to \mu(C)$. Since $C_j \uparrow C$ implies $C_j^c \downarrow C^c$ where $\mu(C^c) = 1 - \mu(C)$, it is sufficient to consider the decreasing case. By considering the sequence $C_j - C$ there is also no loss of generality in setting $C = \emptyset$, so that continuity implies $\mu(C_j) \to 0$. To prove continuity it is then sufficient to show that if $\mu(C_j) > \varepsilon$ for every j, for some $\varepsilon > 0$, then C cannot be empty.

If $C_j \in \mathcal{C}$ for some $j \geq 1$, then $\mu(C_j) = \mu_{k(j)}(E_j)$ for some set $E_j \in \mathcal{B}^{k(j)}$ where $k(j)$ is the dimension of the cylinder C_j. By consistency, $\mu_{k(j)}(E_j) = \mu_m(E_j \times \mathbb{R}^{m-k(j)})$ for any $m > k(j)$, so there is no loss of generality in assuming that $k(1) \leq k(2) \leq \ldots \leq k(j) \leq \ldots$. Therefore define sets $E_i^* \in \mathcal{B}^{k(j)}$, $i = 1, \ldots, j$ by setting $E_j^* = E_j$ and

$$E_i^* = E_i \times \mathbb{R}^{k(j)-k(i)}, i = 1, \ldots, j-1. \tag{12.10}$$

Since $\{C_i\}_{i=1}^j$ is a decreasing sequence, so is the sequence of $\mathcal{B}^{k(j)}$-sets $\{E_i^*\}_{i=1}^j$ for each $j \geq 1$.

Consider any fixed j. There exists, by **12.6**, a compact set $K_j \subseteq E_j$ such that

$$\mu_{k(j)}(E_j - K_j) < \varepsilon/2^{j+1}. \tag{12.11}$$

Define the sets $K_i^* = K_i \times \mathbb{R}^{k(j)-k(i)} \in \mathcal{B}^{k(j)}$ by analogy with the E_i^* and so define

$$F_j = \bigcap_{i=1}^{j} K_i^* \in \mathcal{B}^{k(j)}. \tag{12.12}$$

$F_j \subseteq E_j$ and hence $D_j \subseteq C_j$ where

$$D_j = \pi_{k(j)}^{-1}(F_j) \in \mathcal{C}. \tag{12.13}$$

Applying **1.1**(iii) and then **1.1**(i), observe that

$$E_j - F_j = E_j \cap \left(\bigcup_{i=1}^{j} K_i^{*c} \right) = \bigcup_{i=1}^{j} (E_j - K_i^*) \subseteq \bigcup_{i=1}^{j} (E_i^* - K_i^*). \tag{12.14}$$

The inclusion is because the sequence $\{E_i^*\}_{i=1}^j$ is decreasing, and it implies that

$$\mu_{k(j)}(E_j - F_j) \leq \sum_{i=1}^{j} \mu_{k(j)}(E_i^* - K_i^*) = \sum_{i=1}^{j} \mu_{k(i)}(E_i - K_i) < \varepsilon/2. \tag{12.15}$$

The first inequality here is from (12.14) by finite subadditivity, which follows from finite additivity as a case of **3.5**(iii). The equality applies consistency and the second inequality applies the summation of 2^{-i-1}.

Following these preliminaries, the main step of the argument is show that if $\mu(C_j) > \varepsilon$ for each $j \geq 1$ then C is nonempty. Since $E_j - F_j$ and F_j are disjoint and $\mu_{k(j)}(E_j) = \mu(C_j)$, it then follows from (12.15) that $\mu(D_j) = \mu_{k(j)}(F_j) > \varepsilon/2$ and accordingly that D_j is nonempty. Let $\{x(j), j \in \mathbb{N}\}$ denote a sequence of elements of $\mathbb{R}^\infty$ where $x(j) \in D_j$ for each j and hence

$$\big(X_1(j), \ldots, X_{k(j)}(j)\big) = \pi_{k(j)}\big(x(j)\big) \in F_j. \tag{12.16}$$

Fix $m \geq 1$ and write

$$\big(X_1(j), \ldots, X_{k(m)}(j)\big) = \pi_{k(m)}\big(x(j)\big) \in K_m \tag{12.17}$$

for $j \geq m$, where the inclusion is by (12.12), noting that K_m is compact. Letting j increase, (12.17) defines a sequence of $k(m)$-dimensional vectors in a compact set which accordingly has a cluster point. A set in $\mathbb{R}^{k(m)}$ is compact iff each of the coordinate sets in $\mathbb{R}$ is compact and the bounded scalar sequences are $\{X_i(j), j \geq m\}$ for $i = 1, \ldots, k(m)$. Each of these has a cluster point X_i^* and using the diagonal method (**2.36**) a single subsequence $\{j_n, n \in \mathbb{N}\}$ can be constructed, with the property that $X_i(j_n) \to X_i^*$ as $n \to \infty$ for each $i = 1, \ldots, k(m)$ where $(X_1^*, \ldots, X_{k(m)}^*) \in K_m \subseteq E_m$. This is true for every $m \in \mathbb{N}$. Therefore, consider the element $x^* \in \mathbb{R}^\infty$ that is defined by $\pi_{k(m)}(x^*) = (X_1^*, \ldots, X_{k(m)}^*)$ for each m. Since $\pi_{k(m)}(x^*) \in E_m$ it follows that $x^* \in C_m$ for each m and hence that $x^* \in \bigcap_{m=1}^\infty C_m = C$. ∎

This theorem shows that if a p.m. satisfying (12.7) can be assigned to the finite-dimensional distributions of a sequence x, then x is a random element of a probability space $(\mathbb{R}^\infty, \mathcal{B}^\infty, \mu)$. If $(\mathbb{R}^\infty, \mathcal{B}^\infty, \mu)$ is derived from an abstract probability space $(\Omega, \mathcal{F}, P)$, x is shown $\mathcal{F}/\mathcal{B}^\infty$-measurable if $x^{-1}(E) \in \mathcal{F}$ for each event $E \in \mathcal{B}^\infty$. This statement implies the coordinates X_t are $\mathcal{F}/\mathcal{B}$-measurable r.v.s for each t, but it also implies a great deal more than this, since it is possible to assign measures to events involving countably many sequence coordinates.

12.5 Uniform and Limiting Properties

Much the largest part of stochastic process theory has to do with the joint distribution of sets of coordinates under the general heading of *dependence.* Before getting into these topics, the rest of this chapter deals with various issues relating exclusively to the marginal distributions of the coordinates. Of special interest are conditions that limit the random behaviour of a sequence as the index tends to infinity. The concept of a *uniform* condition on the marginal distributions often plays a key role. Thus, a collection of r.v.s $\{X_\tau, \tau \in \mathbb{T}\}$ is said to be *uniformly bounded in probability* if for any $\varepsilon > 0$ there exists $B_\varepsilon < \infty$ such that

$$\sup_\tau P(|X_\tau| > B_\varepsilon) < \varepsilon. \tag{12.18}$$

It is also said to be *uniformly L_p-bounded* for $p > 0$ if

$$\sup_\tau \|X_\tau\|_p \leq B < \infty. \tag{12.19}$$

For the case $p = \infty$, (12.19) reduces to the condition $\sup_\tau |X_\tau| < \infty$ a.s. In this case the sequence is said to be uniformly bounded a.s., or equivalently, to have a finite sup-norm. For the case $p = 1$, $\sup_\tau \mathrm{E}|X_\tau| < \infty$ and one might think it correct to refer to this property as 'uniform integrability'. Unfortunately, this term is already in use for a different concept (see the next section) and so must be avoided here. Speak of 'uniform L_1-boundedness' in this context.

To interpret these conditions, recall that in mathematics a property is said to hold uniformly if it holds for all members of a class of objects, *including* the limits of any convergent sequences in the class. Consider the case where the collection in question is itself a sequence, with $\mathbb{T} = \mathbb{N}$ and $\tau = t$. Random variables are finite with probability 1, and for each finite $t \in \mathbb{N}$, $P(|X_t| \geq B_{\varepsilon t}) < \varepsilon$ always holds for *some* $B_{\varepsilon t} < \infty$, for any $\varepsilon > 0$. The point of a uniform bound is to ensure that the constants $B_{\varepsilon t}$ are not having to get larger as t increases. 'Bounded uniformly in t' is a different and stronger notion than 'bounded for all $t \in \mathbb{N}$' because, for example, the supremum of the set $\{\|X_t\|_p\}_1^\infty$ may lie outside the set. If $\|X_t\|_p \leq B_t < \infty$ for every t, the sequence is L_p-bounded, but it is not uniformly L_p-bounded unless the possibility that $B_t \to \infty$ as $t \to \infty$ (or $-\infty$) is also ruled out. The statement '$\|X_t\|_p \leq B$, $t \in \mathbb{N}$', where B is the same finite constant for all t, is equivalent to '$\sup_t \|X_t\|_p \leq B$', because the former condition must extend to any limit of the sequence $\{\|X_t\|_p\}$. However, the 'sup' notation is less ambiguous and a good habit to adopt.

The relationships between integrability conditions studied in §9.5 and §9.7 can be used here to establish a hierarchy of boundedness conditions. Uniform L_r-boundedness implies uniform L_p-boundedness for $r > p > 0$, by Liapunov's inequality (**9.30**). Also, uniform L_p-boundedness for any $p > 0$ implies uniform boundedness in probability. The Markov inequality (**9.16**) gives

$$P(|X_t| \geq B) \leq \|X_t\|_p^p / B^p, \tag{12.20}$$

so that, for given $\varepsilon > 0$, (12.18) holds for $B_\varepsilon > \|X_t\|_p/\varepsilon^{1/p}$. By a mild abuse of terminology the case of (12.18) is sometimes called uniform L_0-boundedness.

A standard shorthand due to Mann and Wald ([132]) for the maximum rate of increase or minimum rate of decrease of a stochastic sequence uses the notion of uniform boundedness in probability to extend the 'big Oh' and 'little Oh' notation for ordinary real sequences (see §2.5). If for $\varepsilon > 0$ there exists $B_\varepsilon < \infty$ such that the stochastic sequence $\{X_n\}_1^\infty$ satisfies $\sup_n P(|X_n| > B_\varepsilon) < \varepsilon$, write $X_n = O_p(1)$. If $\{Y_n\}_1^\infty$ is another sequence, either stochastic or nonstochastic, and $X_n/Y_n = O_p(1)$, then write $X_n = O_p(Y_n)$, or in words, 'X_n is at most of order Y_n in probability'. If $P(|X_n| > \varepsilon) \to 0$ as $n \to \infty$, then write $X_n = o_p(1)$ and more generally $X_n = o_p(Y_n)$ when $X_n/Y_n = o_p(1)$, or in words, 'X_n is of order less than Y_n in probability'.

The main use of these notations is in manipulating small-order terms in an expression without specifying them explicitly. Usually, Y_n is a positive or negative power of n. To say that $X_n = o_p(1)$ is equivalent to saying that X_n converges in probability to zero, following the terminology of §12.2. Sometimes $X_n = O_p(1)$ is defined by the condition that for each $\varepsilon > 0$ there exists $B_\varepsilon < \infty$ and an integer $N_\varepsilon \geq 1$ such that $P(|X_n| > B_\varepsilon) < \varepsilon$ for all $n \geq N_\varepsilon$. But X_n is finite almost surely and there necessarily exists (by **12.6**) a constant $B'_\varepsilon < \infty$, possibly larger than B_ε, such that $P(|X_n| > B'_\varepsilon) < \varepsilon$ for $1 \leq n < N_\varepsilon$. For all practical purposes, the formulations are equivalent.

12.6 Uniform Integrability

If a r.v. X is integrable, the contributions to the integral of extreme X values must be negligible. In other words, if $\mathrm{E}|X| < \infty$,

$$\mathrm{E}\big(|X|1_{\{|X|>M\}}\big) \to 0 \text{ as } M \to \infty. \tag{12.21}$$

However, it is possible to construct uniformly L_1-bounded sequences $\{X_n\}$ that fail to satisfy (12.21) in the limit.

12.7 Example Define a stochastic sequence as follows: for $n = 1, 2, 3, \ldots$ let $X_n = 0$ with probability $1 - 1/n$ and $X_n = n$ with probability $1/n$. Note that $\mathrm{E}(|X_n|) = n/n = 1$ for every n and hence the sequence is uniformly L_1-bounded. But to have

$$\lim_{M\to\infty} \mathrm{E}\big(|X_n|1_{\{|X_n|>M\}}\big) = 0 \tag{12.22}$$

uniformly in n requires that for each $\varepsilon > 0$ there exists M_ε such that

$$\mathrm{E}\big(|X_n|1_{\{|X_n|>M\}}\big) < \varepsilon$$

for all $M > M_\varepsilon$, uniformly in n. Clearly, this condition fails for $\varepsilon < 1$ in view of the cases $n > M_\varepsilon$. □

Something very strange is going on in this example. Although $\mathrm{E}(X_n) = 1$ for any n, $X_n = 0$ with probability that approaches 1 as $n \to \infty$. To be precise, $X_n \to_{\text{a.s.}} 0$ (see **19.16**). The intuitive concept of expectation appears to fail when faced with r.v.s taking values approaching infinity with probabilities approaching zero.

The *uniform integrability* condition rules out this type of perverse behaviour in a sequence. Speaking generally, the collection $\{X_\tau, \tau \in \mathbb{T}\}$ is said to be uniformly integrable if

$$\lim_{M\to\infty} \sup_{\tau\in\mathbb{T}} \mathrm{E}\big(|X_\tau| 1_{\{|X_\tau>M\}}\big) = 0. \tag{12.23}$$

In applications the collection in question is usually either a sequence or an array. In the latter case, uniform integrability of $\{X_{nt}\}$ (say) is defined by taking the supremum with respect to both t and n.

The following is a collection of theorems on uniform integrability that will find frequent application later on; **12.8** in particular provides insight into why this concept is so important, since the last example shows that the conclusion does not generally hold without uniform integrability.

12.8 Theorem Let $\{X_n\}_n^\infty$ be a uniformly integrable sequence. If $X_n \to_{\text{a.s.}} X$, then $\mathrm{E}(X_n) \to \mathrm{E}(X)$.

Proof Note that

$$\begin{aligned}\mathrm{E}(|X_n|) &= \mathrm{E}(|X_n| 1_{\{|X_n|\le M\}}) + \mathrm{E}(|X_n| 1_{\{|X_n|>M\}}) \\ &\le M + \mathrm{E}(|X_n| 1_{\{|X_n|>M\}}).\end{aligned} \tag{12.24}$$

By assumption, the second term on the right can be made uniformly small by choosing M large enough, and it follows that $\mathrm{E}|X_n|$ is uniformly bounded. Fatou's lemma implies that $\mathrm{E}|X| < \infty$ and $\mathrm{E}(X)$ exists. Define $Y_n = |X_n - X|$, so that $Y_n \to 0$ a.s. Since $Y_n \le |X_n| + |X|$ by the triangle inequality, **9.33** gives

$$\mathrm{E}(Y_n 1_{\{Y_n>M\}}) \le 2\mathrm{E}(|X_n| 1_{\{|X_n|>M/2\}}) + 2\mathrm{E}(|X| 1_{\{|X|>M/2\}}). \tag{12.25}$$

The second right-hand-side term goes to zero as $M \to \infty$, so $\{Y_n\}$ is uniformly integrable if $\{X_n\}$ is.

$$\mathrm{E}(Y_n) = \mathrm{E}(Y_n 1_{\{Y_n\le M\}}) + \mathrm{E}(Y_n 1_{\{Y_n>M\}}) \tag{12.26}$$

and by the bounded convergence theorem there exists, for any $\varepsilon > 0$, N_ε such that $\mathrm{E}(Y_n 1_{\{Y_n \le M\}}) < \varepsilon/2$ for $n > N_\varepsilon$, for $M < \infty$. M can be chosen large enough that $\mathrm{E}(Y_n 1_{\{Y_n > M\}}) < \varepsilon/2$ uniformly in n, so that $\mathrm{E}(Y_n) < \varepsilon$ for $n > N_\varepsilon$, or, since ε is arbitrary, $\mathrm{E}(Y_n) \to 0$. But

$$\mathrm{E}(Y_n) = \mathrm{E}(|X_n - X|) \ge |\mathrm{E}(X_n) - \mathrm{E}(X)| \tag{12.27}$$

by the modulus inequality and the theorem follows. ∎

Next, here is an alternative form for the condition in (12.23) that is often more convenient for establishing uniform integrability. Consider the following preliminary lemma which makes use of the first Borel–Cantelli lemma, a fundamental result which is to be proved in §19.1.

12.9 Lemma Let X be an $\mathcal{F}$-measurable, integrable r.v. For any $\varepsilon > 0 \;\exists\; \delta > 0$ such that for $E \in \mathcal{F}$,

$$P(E) < \delta \Rightarrow \mathrm{E}(|X|1_E) < \varepsilon. \tag{12.28}$$

Proof This is by contradiction. First note that if $P(E) = 0$ then $1_E = |X|1_E = 0$ a.s. and so $\mathrm{E}(|X|1_E) = 0$ by **4.5**. If the lemma is false, there exists a sequence $\{E_n, n \ge 1\}$ such that for any n, $P(E_n) < 2^{-n}$ and $\mathrm{E}(|X|1_{E_n}) \ge \varepsilon_0 > 0$. However, if $\sum_{n=1}^{\infty} P(E_n) < \infty$ then $P(\limsup_n E_n) = 0$ by **19.2**(i) (compare (19.3)). By the reverse Fatou's lemma (**4.15**) which applies since $\mathrm{E}(|X|) < \infty$, it follows that

$$\varepsilon_0 \le \limsup_n \mathrm{E}(|X|1_{E_n}) \le \mathrm{E}(|X|1_{\limsup_n E_n}) = 0 \tag{12.29}$$

which is the required contradiction. ∎

12.10 Theorem A collection $\{X_\tau, \tau \in \mathbb{T}\}$ of r.v.s in a probability space $(\Omega, \mathcal{F}, P)$ is uniformly integrable iff it is uniformly L_1-bounded and satisfies the following condition: $\forall\, \varepsilon > 0 \;\exists\; \delta > 0$ such that for $E \in \mathcal{F}$,

$$P(E) < \delta \Rightarrow \sup_{\tau \in \mathbb{T}}\{\mathrm{E}(|X_\tau|1_E)\} < \varepsilon. \tag{12.30}$$

Proof To show sufficiency, fix $\varepsilon > 0$ and $\tau \in \mathbb{T}$. By L_{1}-boundedness and the Markov inequality for $p = 1$,

$$MP(|X_\tau| > M) \le \mathrm{E}|X_\tau| < \infty \tag{12.31}$$

and hence with M large enough, $P(|X_\tau| > M) < \delta$ for any $\delta > 0$. By **12.9**, δ can be chosen to satisfy (12.28) for all X_τ and $E \in \mathcal{F}$ which implies in particular that $\mathrm{E}(|X_\tau|1_{\{|X_\tau|>M\}}) < \varepsilon$ is true. Since τ is arbitrary this means

$$\sup_{\tau\in\mathbb{T}}\{\mathrm{E}(|X_\tau|1_{\{|X_\tau|>M\}}\} < \varepsilon \tag{12.32}$$

and (12.23) follows since ε is arbitrary.

To show necessity note that, for any $E \in \mathcal{F}$ and $\tau \in \mathbb{T}$,

$$\begin{aligned}\mathrm{E}(|X_\tau|1_E) &= \mathrm{E}(|X_\tau|1_E 1_{\{|X_\tau|\le M\}}) + \mathrm{E}(|X_\tau|1_E 1_{\{|X_\tau|>M\}})\\ &\le MP(E) + \mathrm{E}(|X_\tau|1_{\{|X_\tau|>M\}}).\end{aligned} \tag{12.33}$$

Consider the suprema with respect to τ of each side of this inequality. For $\varepsilon > 0$, (12.23) implies there exists $M < \infty$ such that

$$\sup_{\tau}\{\mathrm{E}(|X_\tau|1_E)\} < MP(E) + \tfrac{1}{2}\varepsilon. \tag{12.34}$$

Uniform L_1-boundedness now follows on setting $E = \Omega$ and (12.30) also follows with $\delta < \varepsilon/2M$. ∎

Another way to express condition (12.30) is to say that the measures $\nu_\tau(E) = \int_E |X_\tau|\mathrm{d}P$ must be absolutely continuous with respect to P, uniformly in τ.

A condition sufficient for uniform integrability is the existence of a dominating function for the collection.

12.11 Theorem Suppose there exists a non-negative r.v. Y such that $\mathrm{E}(Y) < \infty$ and $|X_\tau(\omega)| \le Y(\omega)$, $\omega \in \Omega, \forall\, \tau \in \mathbb{T}$. Then, $\{X_\tau, \tau \in \mathbb{T}\}$ is uniformly integrable.

Proof Fix $\varepsilon > 0$. Choose M large enough that $P(Y \ge M) < \delta$ for δ small enough that $\mathrm{E}(Y1_{\{Y\ge M\}}) < \varepsilon$, which is possible by **12.9**. For $\tau \in \mathbb{T}$, $P(|X_\tau| > M) \le P(Y > M) < \delta$, and also

$$\mathrm{E}(|X_\tau|1_{\{|X_\tau|>M\}}) \le \mathrm{E}(|X_\tau|1_{\{Y>M\}}) \le \mathrm{E}(|Y|1_{\{Y>M\}}) < \varepsilon.$$

Since τ is arbitrary, the collection $\{X_\tau, \tau \in \mathbb{T}\}$ satisfies the sufficient condition of **12.10**. ∎

An obvious candidate for the dominating function in this result is to set $Y(\omega) = \sup_\tau |X_t(\omega)|$ for $\omega \in \Omega$.

Finally, here is a result that shows why the uniform boundedness of moments of a given order may be important.

12.12 Theorem If

$$\mathrm{E}|X_\tau|^{1+\theta} < \infty \tag{12.35}$$

for $\theta > 0$, then $\lim_{M\to\infty} \mathrm{E}(|X_\tau| 1_{\{|X_\tau|>M\}}) = 0$.

Proof Note that

$$\begin{aligned} \mathrm{E}|X_\tau|^{1+\theta} &\geq \mathrm{E}(|X_\tau|^{1+\theta} 1_{\{|X_\tau|>M\}}) \\ &\geq M^\theta \mathrm{E}(|X_\tau| 1_{\{|X_\tau|>M\}}) \end{aligned} \tag{12.36}$$

for any $\theta > 0$. The result follows on letting $M \to \infty$, since the majorant side of (12.36) is finite by (12.35). ■

Example **12.7** illustrated the fact that uniform L_1-boundedness is not sufficient for uniform integrability, but **12.12** shows that uniform $L_{1+\theta}$-boundedness is sufficient, for any $\theta > 0$. Adding this result to those of §12.5 establishes the hierarchy of uniform conditions summarized in the following theorem.

12.13 Theorem

Uniform boundedness a.s. $\Rightarrow$ uniform L_p-boundedness, $p > 1$
$\Rightarrow$ uniform integrability
$\Rightarrow$ uniform L_p-boundedness, $0 < p \leq 1$
$\Rightarrow$ uniform boundedness in probability.
None of the reverse implications hold. □

One further fact that will prove useful in the sequel is that collections of conditional expectations of an L_1-bounded r.v. are uniformly integrable. This is shown as follows.

12.14 Theorem Let X be a r.v. in a probability space $(\Omega, \mathcal{F}, P)$ and let $\{\mathcal{G}_\tau, \tau \in \mathbb{T}\}$ denote a class of sub-σ-fields of $\mathcal{F}$. If $\mathrm{E}|X| < \infty$ then the collection $\{\mathrm{E}(X|\mathcal{G}_\tau), \tau \in \mathbb{T}\}$ is uniformly integrable.

Proof Given $\varepsilon > 0$, choose $\delta > 0$ small enough that $P(E) < \delta$ implies $\mathrm{E}(|X| 1_E) < \varepsilon$ for any $E \in \mathcal{F}$, as is possible by **12.9**. Also choose $M > 0$ large enough that $\mathrm{E}|X| < \delta M$. For $\tau \in \mathbb{T}$ define the event $E_\tau = \{|\mathrm{E}(X|\mathcal{G}_\tau)| \geq M\} \in \mathcal{G}_\tau$. The Markov inequality, **10.15**, and the LIE give

$$MP(E_\tau) \le \mathrm{E}|\mathrm{E}(X|\mathcal{G}_\tau)| \le \mathrm{E}\big(\mathrm{E}(|X||\mathcal{G}_\tau)\big) = \mathrm{E}|X|. \tag{12.37}$$

It follows by (12.37) that $P(E_\tau) < \delta$. But in this case, noting the r.v. 1_{E_τ} is $\mathcal{G}_\tau$-measurable,

$$\mathrm{E}(|\mathrm{E}(X|\mathcal{G}_\tau)|1_{E_\tau}) \le \mathrm{E}(\mathrm{E}(|X||\mathcal{G}_\tau)1_{E_\tau}) = \mathrm{E}(|X|1_{E_\tau}) < \varepsilon.$$

Since τ is arbitrary the theorem follows by **12.10**. ∎

A point to be aware of here is that the r.v. $\mathrm{E}(X|\mathcal{G}_\tau)$ may not be uniquely defined, as pointed out on page 196. However, different versions are equal a.s. and hence these calculations hold equally for whichever version may be specified.

13
Time Series Models

The present chapter represents something of a digression from the main theme of this part of the book. Following some general remarks and definitions, its object is to review some stochastic process properties as they are manifested in a number of time series models familiar to econometricians. Sections §13.2–§13.4 might be skirted on first reading, but should justification for them be needed it is that this material is all to be put to use in various ways in later chapters.

13.1 Independence and Stationarity

Independence and stationarity are the best-known restrictions on the behaviour of a sequence but also the most stringent, from the point of view of describing economic time series. The emphasis in this book is most often on finding ways to relax these conditions but they remain important because of the many classic theorems in probability and limit theory that are founded on them. Statements of general definitions and results usually specify the case of the one-sided sequence $\{X_t\}_1^\infty$. While there is no difficulty in extending the concepts to the case $\{X_t\}_{-\infty}^{\infty}$ this is generally left implicit unless the mapping from $\mathbb{Z}$ plays a specific role in the argument. Think of $\{X_t\}_1^\infty$ as a point in $\mathbb{R}^\infty$, the Cartesian product of an infinite collection of copies of the real line. Since $\mathbb{N}$ and $\mathbb{Z}$ are equipotent and connected by a 1–1 mapping (see §1.3) there is no loss of generality in using the notation $\mathbb{R}^\infty$ to refer also to the doubly infinite product space.

In the context of time-ordered observations, the degree to which the random variations of sequence coordinates are related to those of their neighbours is sometimes called the *memory* of a sequence; think in terms of the amount of information contained in the current state of the sequence about its previous states. A sequence with no memory is a rather special kind of object, because the ordering ceases to have significance. It is like the outcome of a collection of independent random experiments conducted in parallel and indexed arbitrarily. When a time ordering does nominally exist, such a sequence is called *serially independent*.

There is a useful analogy between studying the relationships within a sequence and comparing different sequences. Generalizing the theory of §8.5, the pair of sequences

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0013

$$\big\{\{X_t\}_1^\infty, \{Y_t\}_1^\infty\big\} \in \mathbb{R}^\infty \times \mathbb{R}^\infty$$

are independent of one another if, for all $E_1, E_2 \in \mathcal{B}^\infty$,

$$P\big(\{X_t\}_1^\infty \in E_1, \{Y_t\}_1^\infty \in E_2\big) = P\big(\{X_t\}_1^\infty \in E_1\big)P\big(\{Y_t\}_1^\infty \in E_2\big). \tag{13.1}$$

Accordingly, a sequence $\{X_t\}_1^\infty$ is serially independent if it is independent of $\{X_{t+k}\}_1^\infty$ for all $k > 0$. This is equivalent to saying that every finite collection of sequence coordinates is totally independent.

A condition closely related to but weaker than independence is *exchangeability.* An exchangeable finite sequence is one whose joint distribution is invariant to arbitrary permutations of the coordinates. In an infinite sequence exchangeability means that the distribution is invariant under arbitrary finite sets of permutations.

If the marginal distribution of X_t is the same for any t, the sequence $\{X_t\}$ is said to be *identically distributed.* A stochastic sequence that is both serially independent and identically distributed, written for brevity as i.i.d., is like an arbitrarily indexed random sample drawn from some underlying population. An i.i.d. sequence is exchangeable but exchangeability does not rule out dependence unrelated to relative positions in the sequence. Thus, a sequence of the form $\{X_t + Z\}$ where $\{X_t\}$ is i.i.d. and Z is any random variable is not i.i.d., but it is exchangeable.

Serial independence is the simplest and strongest assumption that can be made about memory. Similarly, looking at the distribution of the sequence as a whole, the simplest treatment is to assume that the joint distribution of the coordinates is invariant with respect to the time index. A random sequence is called *strictly stationary* if the sequences $\{X_t\}_{t=1}^\infty$ and $\{X_{t+k}\}_{t=1}^\infty$ have the same joint distribution, for every $k > 0$. In practical terms, this means that the joint distributions of a pair of finite sequence segments $\{X_t, \ldots, X_{t+m}\}$ and $\{X_{t+k}, \ldots, X_{t+k+m}\}$ match, for any choices of t, m, and k. Note that 'identically distributed' is a different concept from stationary, for stationarity also restricts the joint distribution of neighbours in the sequence. However, an i.i.d. sequence is strictly stationary.

Subject to the existence of particular moments, less restrictive versions of the stationarity condition are also commonly employed. Letting $\mu_t = \mathrm{E}(X_t)$ and $\gamma_{kt} = \mathrm{Cov}(X_t, X_{t+k})$, consider those cases in which the sequence $\{\mu_t\}_{t=1}^\infty$ and also the array $\{\{\gamma_{kt}\}_{k=0}^\infty\}_{t=1}^\infty$ are well defined. If $\mu_t = \mu$ for all t the sequence is said to be *mean stationary*. If a mean stationary sequence has $\gamma_{kt} = \gamma_k$ where $\{\gamma_k\}_0^\infty$ is a sequence of constants, it is called *covariance stationary*, or *wide sense* stationary.[1] An implication of stationarity is that the location of the origin, $t = 1$, becomes

[1] The term 'weakly stationary' is also used in this context but is best avoided in view of the various other technical usages of the term; weak dependence, weak convergence, etc.

arbitrary. It is natural to define the doubly infinite sequence $\{X_t\}_{t=-\infty}^{\infty}$ and in this context $\mathrm{Cov}(X_t, X_{t+k}) = \mathrm{Cov}(X_{t-k}, X_t)$ for any t and k, so $\gamma_k = \gamma_{-k}$.

Consider the following clutch of examples, which includes both stationary and nonstationary cases.

13.1 Example Let $\{\varepsilon_t\}_1^{\infty}$ be an i.i.d. sequence with $\varepsilon_t \sim_{\mathrm{d}} \mathrm{N}(0,\sigma^2)$ and let

$$X_t = \begin{cases} \varepsilon_t - 1, & t \text{ odd} \\ \varepsilon_t + 1, & t \text{ even.} \end{cases} \tag{13.2}$$

Then $\{X_t\}_1^{\infty}$ is nonstationary but the processes $\{X_t^2\}_1^{\infty}$ and $\{X_t + X_{t+1}\}_1^{\infty}$ are both stationary. □

In this example the symmetry of the distribution of ε_t about zero is crucial. The r.v.s $1 - \varepsilon_t$ and $1 + \varepsilon_t$ have the same distribution and their squares match those of X_t. Also, $X_t + X_{t+1} = \varepsilon_t + \varepsilon_{t+1}$ which is a so-called moving average process (see §13.3) with $\mu = 0$, $\gamma_0 = 2\sigma^2$, $\gamma_1 = \sigma^2$, and $\gamma_k = 0$ for $k > 1$. That these sequences are both wide-sense stationary and strictly stationary follows from the Gaussian characteristics noted following Example **9.27**.

13.2 Example Let $\{X_t\}_{-\infty}^{\infty}$ be a stationary sequence with autocovariance sequence $\{\gamma_m\}_0^{\infty}$. The sequence $\{X_t + X_0\}_{-\infty}^{\infty}$ is nonstationary with autocovariances given by the array $\left\{\{2\gamma_m + \gamma_{|t+m|} + \gamma_{|t-m|}\}_{m=0}^{\infty}\right\}_{t=-\infty}^{\infty}$. □

13.3 Example Let $X \sim_{\mathrm{d}} \mathrm{N}(0,\sigma^2)$ and for $t \geq 1$

$$X_t = \begin{cases} -X, & t \text{ odd} \\ X, & t \text{ even.} \end{cases}$$

$\{X_t\}_1^{\infty}$ is a stationary sequence, and in particular $\mathrm{E}(X_t) = 0$ and $\mathrm{Cov}(X_t, X_{t+k}) = \sigma^2$ when k is even and $-\sigma^2$ when k is odd, independent of t. □

13.4 Example Let $\{\varepsilon_t\}_{-\infty}^{\infty}$ be i.i.d. with $\mathrm{E}(\varepsilon_t) = 0$ and $\mathrm{E}(\varepsilon_t^2) = \sigma^2$ and let the autoregressive process $\{X_t\}_1^{\infty}$ be generated by the difference equation

$$X_t = \gamma + \alpha X_{t-1} + \varepsilon_t \tag{13.3}$$

where $|\alpha| < 1$ and X_0 is given. If $X_0 = 0$ or other arbitrary value then the process is nonstationary. However, if X_0 is an independently drawn r.v. with $\mathrm{E}(X_0) = \gamma/(1-\alpha)$ and $\mathrm{Var}(X_0) = \sigma^2/(1-\alpha^2)$ then $\{X_t\}_1^{\infty}$ is a covariance stationary process having $\mathrm{E}(X_t) = \mathrm{E}(X_0)$ and $\mathrm{Var}(X_t) = \mathrm{Var}(X_0)$ for every $t \geq 1$. It is easy to see that the stationary case is observationally equivalent to the one in which (13.3) holds for $-\infty < t < \infty$. □

The nonstationarity featured in these cases is all of a local sort, either eliminated by local averaging (**13.1**) or due to dependence on initial conditions and disappearing at long range (**13.2** and **13.4**). **13.3** shows that a periodic pattern does not contradict stationarity, provided it does not depend systematically on the time index.

Local stationarity can be a strong assumption, particularly for the description of empirical time series where features like seasonal patterns are commonly found. If sequences $\{X_t\}_{t=1}^{\infty}$ and $\{X_{t+k}\}_{t=1}^{\infty}$ have the same distribution for *some* (not necessarily every) $k > 0$, it follows that $\{X_{t+2k}\}_{t=1}^{\infty}$ has the same distribution as $\{X_{t+k}\}_{t=1}^{\infty}$ and the same property extends to every integer multiple of k. Such a sequence accordingly has stationary characteristics, thinking in terms of the distributions of successive blocks of k coordinates. This idea retains force even as $k \to \infty$. Consider a finite sequence of length n divided into $[n^{\alpha}]$ blocks of length $[n^{1-\alpha}]$ plus any remainder, for some α between 0 and 1. (Note, $[x]$ here denotes the largest integer not exceeding x.) As $n \to \infty$, the number of blocks as well as their extent is going to infinity and the stationarity (or otherwise) of the sequence of blocks in the limit is clearly an issue. Important applications of these ideas arise in Parts V and VI below.

For this reason it is useful to distinguish between nonstationarity that might be eliminated by local averaging of the coordinates and 'global' nonstationarity involving features such as persistent trends in the moments. It is convenient to formulate a definition embodying this concept in terms of moments. Thus, a zero-mean sequence will be said to be *globally covariance stationary* if the autocovariance sequences $\{\gamma_{mt}\}_{t=1}^{\infty}$ are Cesàro-summable for each $m \geq 0$, where the Cesàro sum is strictly positive in the case of the variances ($m = 0$). The following are a pair of contrasting counterexamples.

13.5 Example A sequence with $\gamma_{0t} \sim t^{\beta}$ is globally nonstationary for any $\beta \neq 0$. □

13.6 Example Consider the integer sequence beginning

$$1, 2, 1, 1, 2, 2, 2, 2, 1, 1, 1, 1, 1, 1, 1, 1, \ldots;$$

i.e. the value changes at points $t = 2^k, k = 0, 1, 2, 3, \ldots$. The Cesàro sum of this sequence fails to converge as $n \to \infty$. It fluctuates eventually between the points

5/3 at $n = 2^k$, k odd and 4/3 at $n = 2^k$, k even. A stochastic sequence having a variance sequence of this form is globally nonstationary. □

13.2 The Poisson Process

The Poisson process is an entity of special importance in statistical modelling. This is a sequence of times at which random events occur, where the intervals elapsing between one event and the next are exponentially distributed (see **8.11**) with the rate λ depending on the units of time measurement. This distribution characterizes successive events occurring independently at random times, such as telephone calls arriving at a switchboard, emails arriving at a mail server, etc. For clarity let these events be referred to as 'arrivals' and let 'date' denote a time of arrival in the general sense.

With time measured from the initial date $t = 0$, let $T_n \geq 0$ be the random variable denoting the date at which the n^{th} arrival occurs. Then

$$N(t) = \sum_{n\geq 1} 1_{\{t\geq T_n\}} \in \mathbb{N} \tag{13.4}$$

is the integer-valued process counting the number of arrivals that have occurred by date $t \in \mathbb{R}^+$. In the Poisson process the number of arrivals occurring in the time interval $[0, t]$ is Poisson-distributed with parameter λt. In other words,

$$P(N(t) = n) = \frac{\mathrm{e}^{-\lambda t}(\lambda t)^n}{n!}. \tag{13.5}$$

This may be understood by recalling how the Poisson is the limit of the binomial as the number of trials increases while the probability of 'outcome = success' simultaneously declines. The number of telephones connected to an exchange is large, while the probability that a call is made from any one of them within a given short time period is small. The Poisson captures the limiting case as these numbers go respectively to infinity and zero.

However, another way to view the same phenomenon is as the sequence $\{W_n, n \geq 1\}$ of exponentially distributed waiting times, labelled with the counting index of integers. This sequence is both independent and stationary. Since λ is the mean number of arrivals in a unit time period according to (9.7), $1/\lambda$ from (9.9) is necessarily the average time elapsing between arrivals. The exponential distribution arises in this context due to the property known as 'memorylessness'. The distribution of the waiting time to the next arrival does not depend on the time elapsed since the last arrival, as must be the case if the arrivals are independent. Thus, with $s+t$ denoting the time predicted to elapse from the standpoint of elapsed time s since T_{n-1},

$$\begin{aligned}P(W_n > s+t|W_n > s) &= \frac{P(W_n > s+t)}{P(W_n > s)}\\ &= \frac{\mathrm{e}^{-\lambda(s+t)}}{\mathrm{e}^{-\lambda s}}\\ &= \mathrm{e}^{-\lambda t} = P(W_n > t),\end{aligned}$$

where the first equality is due to the fact that $\{W_n > t+s\} \cap \{W_n > s\} = \{W_n > t+s\}$.

In view of **8.12**, the time it takes for n arrivals to occur, in other words the sum of n independent exponential drawings, is gamma(n,λ)-distributed. A further implication is that the joint distribution of arrival dates over a given interval of time is uniform. The precise statement of this property is as follows.

13.7 Theorem Let the dates of arrivals $1,\ldots,n$ of a Poisson process be expressed as fractions of the date of the $n+1^{\text{th}}$ arrival. The joint distribution of these n points of $[0,1]$ matches that of n independent drawings from $\mathrm{U}[0,1]$ sorted smallest to largest.

Proof Let $W_1,\ldots,W_n$ denote independent exponential random variates, with rate $\lambda = 1$ without loss of generality since this is only a matter of setting the unit of time measurement. Then let $\Gamma_k = \sum_{i=1}^{k} W_i$ for $k = 1,\ldots,n+1$ denote the date of the k^{th} arrival. Consider the joint probability of the incremental events $\{\Gamma_1 \in \mathrm{d}x_1, \Gamma_2 \in \mathrm{d}x_2, \ldots, \Gamma_{n+1} \in \mathrm{d}x_{n+1}\}$, where the $\mathrm{d}x_k$ denote increments of time beginning at date x_k where the most recent previous occurrence was at date x_{k-1} and $0 \le x_1 \le x_2 \le \cdots \le x_{n+1}$. According to **8.11**,

$$P(\Gamma_k \in \mathrm{d}x_k) = P(W_k + x_{k-1} \in \mathrm{d}x_k) = \mathrm{e}^{-(x_k - x_{k-1})}\mathrm{d}x_k$$

for $k = 1,\ldots,n+1$. Hence in view of the independence,

$$\begin{aligned}&P(\Gamma_1 \in \mathrm{d}x_1, \Gamma_2 \in \mathrm{d}x_2, \ldots, \Gamma_{n+1} \in \mathrm{d}x_{n+1})\\ &\qquad = P(W_1 \in \mathrm{d}x_1, W_2 + x_1 \in \mathrm{d}x_2, \ldots, W_{n+1} + x_n \in \mathrm{d}x_{n+1})\\ &\qquad = \mathrm{e}^{-x_1}\mathrm{e}^{-(x_2-x_1)} \ldots \mathrm{e}^{-(x_{n+1}-x_n)}\mathrm{d}x_1 \cdots \mathrm{d}x_{n+1}\\ &\qquad = \mathrm{e}^{-x_{n+1}}\mathrm{d}x_1 \cdots \mathrm{d}x_{n+1}.\end{aligned}$$

The Γ_k are gamma$(k,1)$-distributed with densities $f(x_k;k,1) = \mathrm{e}^{-x_k}x_k^{k-1}/(k-1)!$ according to **8.12** and the remark following. The probabilities conditioned on the terminal state $\Gamma_{n+1} = x_{n+1}$ are found by converting the densities to conditional densities in the formula, by dividing by $\mathrm{e}^{-x_{n+1}}x_{n+1}^n/n!$ and setting $P(\Gamma_{n+1} \in x_n + \mathrm{d}x_{n+1}|\Gamma_{n+1} = x_{n+1}) = 1$. Thus,

$$\begin{aligned} P(\Gamma_1 \in \mathrm{d}x_1, \dots, \Gamma_n \in \mathrm{d}x_n|\Gamma_{n+1} = x_{n+1}) &= \frac{\mathrm{e}^{-x_{n+1}}\mathrm{d}x_1 \cdots \mathrm{d}x_n}{\mathrm{e}^{-x_{n+1}}x_{n+1}^n/n!} \\ &= n!x_{n+1}^{-n}\mathrm{d}x_1 \cdots \mathrm{d}x_n. \end{aligned}$$

Letting $y_k = x_k/x_{n+1} \in [0,1]$ for $k = 1, \dots, n$ with $0 \le y_1 \le \cdots \le y_n \le 1$,

$$P(\Gamma_1/\Gamma_{n+1} \in \mathrm{d}y_1, \dots, \Gamma_n/\Gamma_{n+1} \in \mathrm{d}y_n|\Gamma_{n+1} = x_{n+1}) = n!\mathrm{d}y_1 \cdots \mathrm{d}y_n. \tag{13.6}$$

This is also the unconditional probability since it does not depend on x_{n+1}.

Compare this formula with the distribution of n independent U$[0,1]$ r.v.s $U_1, \dots, U_n$,

$$P(U_1 \in \mathrm{d}y_1, \dots, U_n \in \mathrm{d}y_n) = \mathrm{d}y_1 \cdots \mathrm{d}y_n. \tag{13.7}$$

Let $U_{(1)} \le \cdots \le U_{(n)}$ denote the n order statistics. There are $n!$ ways for the events having the probability in (13.7) to give rise to the event $\{U_{(1)} \in \mathrm{d}y_1, \dots, U_{(n)} \in \mathrm{d}y_n\}$, so

$$P(U_{(1)} \in \mathrm{d}y_1, \dots, U_{(n)} \in \mathrm{d}y_n) = n!\mathrm{d}y_1 \cdots \mathrm{d}y_n. \tag{13.8}$$

Comparison of (13.6) and (13.8) completes the proof. ∎

13.3 Linear Processes

Consider a sequence $\{X_t\}_{-\infty}^{\infty}$ whose coordinates are $X_t = \sum_{j=-q}^{q} \theta_j U_{t-j}$ for $0 \le q \le \infty$, where $\{U_t\}_{-\infty}^{\infty}$ is a doubly infinite i.i.d. sequence with mean 0 and variance $\sigma^2 < \infty$ and $\{\theta_j\}_{-q}^{q}$ a sequence of constants with $\theta_0 = 1$. This is a *moving average* process of order q (MA(q)). The U_t are called the 'shocks', or 'innovations', of the process. The distribution of X_t may be visualized by consideration of the ch.f.

$$\phi_{qt}(\lambda) = \prod_{j=-q}^{q} \phi_{U_{t-j}}(\theta_j\lambda). \tag{13.9}$$

If $\phi_{qt}(\lambda) \to \phi_t(\lambda)$ (pointwise in $\mathbb{R}$) as $q \to \infty$ where $\phi_t(\lambda)$ is a ch.f. and continuous at $\lambda = 0$, the MA(∞) process

$$X_t = \sum_{j=-\infty}^{\infty} \theta_j U_{t-j} \tag{13.10}$$

exists as the weak limit of the sequence of MA(q)s.[2]

13.8 Definition A *linear* process is one having representation (13.10) for some collection of constant weights $\{\theta_j\}_{j=-\infty}^{\infty}$ and i.i.d. shocks $\{U_t\}_{t=-\infty}^{\infty}$. □

Linear processes are distinguished by the fact that their dependence structure depends entirely on the sequence of coefficients. (13.10) is of course the fully general case and may be specialized by requiring all but a finite set of the weights to be zero.

The existence of the MA(∞) imposes certain conditions on the coefficient sequence, at a minimum that $|\theta_j| \to 0$ as $|j| \to \infty$. Under the *square summability* condition $\sum_{j=-\infty}^{\infty} \theta_j^2 < \infty$ the process is strictly stationary and covariance stationary, with $\mathrm{E}(X_t) = 0$ and $\mathrm{E}(X_t^2) = \sigma^2 \sum_{j=-\infty}^{\infty} \theta_j^2$. Then the autocovariance sequence $\{\gamma_m\}_0^{\infty}$ has

$$\gamma_m = \mathrm{E}(X_t X_{t+m}) = \sigma^2 \sum_{j=-\infty}^{\infty} \theta_j \theta_{j+m} \tag{13.11}$$

for every t, with $\gamma_{-m} = \gamma_m$.

In econometrics the notion of two-sided time dependence is rarely appropriate and 'causal' processes depending only on present and past shocks are the norm, which means setting $\theta_j = 0$ for $j < 0$. However, the nominally two-sided representation has a number of convenient features. A key question is whether the autocovariances form a summable sequence, the attribute of a stochastic process sometimes called *short memory* and equivalently *weak dependence*. A useful fact is that summability of the lag coefficients is equivalent to the summability of the autocovariances, since

$$\sum_{m=-\infty}^{\infty} \gamma_m = \sigma^2 \sum_{m=-\infty}^{\infty} \sum_{j=-\infty}^{\infty} \theta_j \theta_{j+m} = \sigma^2 \left(\sum_{j=-\infty}^{\infty} \theta_j \right)^2 . \tag{13.12}$$

Summability further implies that

$$\mathrm{E}\left(\sum_{t=1}^{n} X_t \right)^2 = n\gamma_0 + 2 \sum_{m=1}^{n-1} (n-m)\gamma_m = O(n).$$

[2] Here 'weak' is used in the sense of the weak convergence of distributions. See §22.4 for further details.

Since the standard deviation of the partial sum grows like $n^{1/2}$ this is known as the 'square root rule' of short memory processes.

To explore this class of processes further it is useful to introduce lag operator notation, as follows.

13.9 Definition If $\{U_t\}_1^n$ is a sequence consecutively indexed by integers $t = 1, \ldots, n$ the operator L denotes the backshift or lag operation such that $LU_t = U_{t-1}$. □

L is incorporated into algebraic manipulations using the power notation L^j, where $L^j U_t = U_{t-j}$ and j can take either sign so that both back and forward shifts are represented. It is now readily seen on letting

$$\theta(L) = \sum_{j=0}^{\infty} \theta_j L^j \tag{13.13}$$

that $X_t = \theta(L)U_t$ is an equivalent representation of the causal form of (13.10). $\theta(L)$ is called a lag polynomial.

Models constructed from a small set of parameters have a natural appeal in applications.

13.10 Example The ARMA(p,q) (autoregressive-moving average) class has $\theta(L) = \beta(L)/\alpha(L)$ so that $\alpha(L)X_t = \beta(L)U_t$, where $\alpha(L)$ and $\beta(L)$ are lag polynomials of finite orders p and q respectively and the roots of the polynomial $\alpha(z)$ for $z \in \mathbb{C}$ fall strictly outside the unit circle in the Argand plane. The special feature of the ARMA class is that the process is covariance stationary and the lag coefficients converge to zero geometrically, with $\theta_j = o(j^{-\delta})$ for every finite $\delta > 0$. □

Example **13.4** is the AR(1) case of **13.10** with $p = 1$, $q = 0$, and $\alpha(L) = 1 - \alpha L$ so that $\theta_j = \alpha^j$. Stationarity implies that ARMA processes can be defined for $-\infty < t < \infty$ and no initial condition need be specified. To introduce nonstationarity into the ARMA framework the natural extension is to allow an autoregressive root of unity.

13.11 Example The unit root process $\Delta X_t = U_t$ where $\Delta = 1 - L$ produces an unweighted cumulation of the increments. Such processes require an initial condition to be specified, for example $X_0 = 0$. □

The distributions of stochastic processes generated like **13.11** are analysed in §13.4.

Most parametric cases arising in applications belong to the ARMA class, but here is one that does not.

13.12 Example Consider the process $X_t = \theta(L)U_t$ where the coefficients of (13.13) depend on a parameter d,

$$\theta_j = \frac{\Gamma(j+d)}{\Gamma(d)\Gamma(j+1)} = O(j^{d-1}).$$

According to the generalized binomial theorem this model has representation $X_t = (1-L)^{-d}U_t$. The case with $d > 0$ is called a *fractionally integrated* process. For $d < \frac{1}{2}$ the moving average coefficients are square-summable, hence the process is covariance stationary in this range. For any $d > 0$ the autocovariances are non-summable with $\gamma_j = O(j^{2d-1})$, so the process does not obey the square root rule and is said to have *long memory*. With $d = 1$ it is undefined since $\theta_j = 1$ for every j but is identical with **13.11** if an initial date is specified. If $d < 0$ the process is stationary and short memory but $\gamma_j < 0$ for $j > 0$ and $\sum_{j=0}^{\infty}\gamma_j = 0$. This is the 'anti-persistent' case. With $d = -1$ it is MA(1), the simple difference of the innovations. □

The autoregressive representation of **13.10** poses an interesting question about the *invertibility* of a MA process. Does there always exist $\alpha(L)$ such that $X_t = \theta(L)U_t$ has representation $\alpha(L)X_t = U_t$? Equally, the reverse question can be posed: does an autoregressive relation have a moving average solution?

Assuming that the relations are causal so that $\alpha_j = \theta_j = 0$ for all $j < 0$ it is convenient to write the polynomials as $\theta(z)$ and $\alpha(z)$ with argument $|z| \leq 1$ in place of the operator. The inversion relationship has the form

$$\alpha(z)\theta(z) = 1. \tag{13.14}$$

Notably the analysis is symmetric and holds equally with the roles of $\theta(z)$ and $\alpha(z)$ interchanged. Assuming $\theta_0 = \alpha_0 = 1$ the solution must take the form that the coefficients of z^j in the product are zero for all $j > 0$. Matching up the powers of z in (13.14) after multiplying out implies the relations

$$\sum_{j=0}^{k}\alpha_j\theta_{k-j} = 0,\ k = 1,2,3,\ldots. \tag{13.15}$$

Equations (13.15) have recursive solutions for (say) the θ_k of the form

$$\theta_k = -\sum_{j=1}^{k}\alpha_j\theta_{k-j},\ k = 1,2,3,\ldots. \tag{13.16}$$

Thus, given $\alpha(z)$ the companion polynomial $\theta(z)$ always exists. The question of interest is whether the corresponding lag polynomial validly defines a stochastic process, which must involve the coefficients declining to zero at some rate as the lag increases. Certainly a sufficient condition is that the moving average coefficients are summable; in other words that $|\theta(1)| < \infty$ which according to (13.12) implies weak dependence. Suppose that $\alpha(z)$ is a polynomial of order p. Summing both sides of (13.16) it can be verified that

$$\sum_{i=0}^{k}\theta_i = 1 - \sum_{j=1}^{\min\{k,p\}}\alpha_j\sum_{i=0}^{k-j}\theta_i.$$

Letting $k \to \infty$ gives the limiting case,

$$\theta(1) = 1 - \sum_{j=1}^{p}\alpha_j\theta(1) = \frac{1}{\alpha(1)}.$$

Since in this argument p is arbitrary the general rule follows, with an obvious symmetry:

13.13 Theorem $0 < \alpha(1) < \infty$ iff $0 < \theta(1) < \infty$. □

Since the product of these sums is unity, note the implication that the lag weights θ_j for $j \geq 1$ have a positive sum if the α_j for $j \geq 1$ have a negative sum and conversely.

The simplest counterexample is where $\alpha_1 = -1$ and $\alpha_j = 0$ for $j > 1$. In this case solution (13.16) has $\theta_k = 1$ for every $k \geq 1$ which does not define a stochastic process. This is the case of a unit root $\Delta X_t = U_t$ (see Example **13.11**) which is non-stationary and undefined without a finite starting point. Note the symmetry in the conditions, with $\alpha(1) = 0$ implying $\theta(1) = \infty$. It is equally the case that the 'over-differenced' process $X_t = \Delta U_t$ does not have an autoregressive representation.

Summability of the coefficients is a sufficient condition for a moving average process to be well defined, but not a necessary one. As noted, square-summability is sufficient for a process to be covariance stationary. Consider Example **13.12** where $\alpha(L) = (1-L)^d$ for $0 < d < \frac{1}{2}$. Hence $\alpha(1) = 0$. In this case, $\alpha_j < 0$ for $j > 0$ and $|\alpha_j| \simeq j^a$ where $a = -d-1$. For (13.15) to hold in the limit implies $\theta_j \simeq j^b$ where $\sum_{j=1}^{k} j^{\alpha}(k-j)^b \to 0$ as $k \to \infty$, and according to Theorem **2.19** this happens whenever $b < -1 - a = d$. In fact the coefficients of the inverse polynomial $(1-L)^{-d}$ are $\theta_j \simeq j^{d-1}$, hence square summable although non-summable.

A process is called Gaussian if every finite collection of coordinates has a multivariate normal distribution. If the innovations are i.i.d. $\mathrm{N}(0,\sigma^2)$ then under stationarity the linear process X_t is also Gaussian. This is a consequence of two facts; that $X_{kt} = \sum_{j=0}^{k} \theta_j U_{t-k}$ is Gaussian for any finite k (see Example **8.24**) and that under square-summability

$$\mathrm{E}(X_t - X_{kt})^2 = \sigma^2 \sum_{j=k+1}^{\infty} \theta_j^2 \to 0$$

as $k \to \infty$. X_t is the mean square limit of X_{kt} and hence also the limit in distribution.

A remarkable fact is that every stationary Gaussian process has a MA(∞) representation. This is a consequence of the well-known *Wold decomposition* ([193]) which is as follows.

13.14 Theorem A zero mean covariance stationary process $\{X_t\}_{-\infty}^{\infty}$ has representation

$$X_t = \sum_{j=0}^{\infty} \theta_j \varepsilon_{t-j} + v_t \qquad (13.17)$$

where $\sum_{j=0}^{\infty} \theta_j^2 < \infty$, $\{\varepsilon_t\}_{-\infty}^{\infty}$ is serially uncorrelated, $\mathrm{E}(v_t \varepsilon_{t-j}) = 0$ for all $j \geq 1$, and there exist constants α_j such that

$$\mathrm{E}\left(v_t - \sum_{j=1}^{\infty} \alpha_j v_{t-j}\right)^2 = 0. \quad \Box \qquad (13.18)$$

Note that (13.18) has the implication that v_t is perfectly predictable given the history of the sequence up to date $t-1$. Accordingly, $\{v_t\}$ is called the deterministic component of the process.

Proof of 13.14 For finite k and for $j = 1, \ldots, k$ let $W_{jt} = X_{t-j} - \sum_{m=1}^{j-1} \beta_{mj} X_{t-m}$ where the β_{mj} minimize $\mathrm{E}(X_{t-j} - \sum_{m=1}^{j-1} b_{mj} X_{t-m})^2$ with respect to $b_{1j}, \ldots, b_{j-1,j}$. The variables $W_{1t}, \ldots, W_{kt}$ are orthogonal by construction. Defining

$$\gamma_j = \frac{\mathrm{E}(W_{jt} X_t)}{\mathrm{E}(W_{jt}^2)}, j = 1, \ldots, k$$

let $\varepsilon_{kt} = X_t - \sum_{j=1}^{k} \gamma_j W_{jt}$. Since the γ_j are least squares coefficients the sequence $\mathrm{E}(\varepsilon_{kt}^2) = \mathrm{E}(X_t^2) - \sum_{j=1}^{k} \gamma_j^2 \mathrm{E}(W_{jt}^2)$ is non-increasing as $k \to \infty$ and bounded below. Therefore the sequence $\{\varepsilon_{kt}\}_{k=1}^{\infty}$ has a mean square limit, a r.v. ε_t such that $\mathrm{E}(\varepsilon_{kt} - \varepsilon_t)^2 \to 0$ and

$$X_t = \sum_{j=1}^{\infty} \gamma_j W_{jt} + \varepsilon_t = \sum_{j=1}^{\infty} \alpha_j X_{t-j} + \varepsilon_t \qquad (13.19)$$

where $\alpha_j = \gamma_j - \sum_{m=j+1}^{\infty} \gamma_m \beta_{jm}$. Since the expected sum of squares is minimized, $\mathrm{E}(\varepsilon_t X_{t-j}) = 0$ and hence also $\mathrm{E}(\varepsilon_t \varepsilon_{t-j}) = 0$ for each $j > 0$.

In view of this last property, regressing X_t onto ε_{t-j} for $j = 0, \ldots, k$ yields

$$X_t = \sum_{j=0}^{k} \theta_j \varepsilon_{t-j} + v_{kt} \tag{13.20}$$

where $\theta_j = \mathrm{E}(X_t \varepsilon_{t-j})/\mathrm{E}(\varepsilon_{t-j}^2)$ and $\mathrm{E}(\varepsilon_{t-j} v_{kt}) = 0$ for each j. Letting $k \to \infty$ similarly to above yields (13.17) where v_t is the mean square limit of v_{kt}. These coefficients must satisfy $\sum_{j=0}^{\infty} \theta_j^2 < \infty$ since $\mathrm{E}(X_t^2) < \infty$ by assumption.

Now substitute for X_{t-j} from (13.17) into (13.19), for $j \geq 0$. After rearrangement this produces, noting $\theta_0 = 1$,

$$\begin{aligned} \varepsilon_t &= \varepsilon_t + \sum_{j=1}^{\infty} \theta_j \varepsilon_{t-j} + v_t - \sum_{j=1}^{\infty} \alpha_j \left(\varepsilon_{t-j} + \sum_{m=1}^{\infty} \theta_m \varepsilon_{t-j-m} + v_{t-j} \right) \\ &= \varepsilon_t + \sum_{j=1}^{\infty} \delta_j \varepsilon_{t-j} + v_t - \sum_{j=1}^{\infty} \alpha_j v_{t-j} \end{aligned} \tag{13.21}$$

where $\delta_j = \theta_j - \sum_{m=1}^{j} \alpha_m \theta_{j-m}$. Squaring each side of (13.21) and taking expectations forces the conclusions $\delta_1 = \delta_2 = \cdots = 0$ and (13.18). ∎

In the case where X_t is a Gaussian process the ε_t are also jointly Gaussian by construction. Their serial independence then follows from the uncorrelatedness and hence a true linear representation exists. However, this is not true of other distributions. The Wold theorem shows that it is always possible in a covariance stationary process to construct a moving average representation with uncorrelated shocks, but this simply means that any dependence in the latter must be manifested in other ways; dependence in volatility is the classic case in applications, see for example §18.6. The key point to take away is that although it has a critical implication for the case of Gaussianity, the Wold representation is not in general a linear process in the sense of Definition **13.8**.

13.4 Random Walks

Stationary and possibly independent series are the main building blocks of time series modelling. If observed series do not themselves have these attributes, it may be that they are found in suitable decompositions of the observed series and it turns out that *differencing* is one of the most important of these decompositions.

That is, the series of period-to-period changes may be modelled as stationary and possibly independent. The class of processes that are formed as the cumulation of a random sequence from some given starting point has great importance in the analysis of time series; such empirical processes have a multitude of applications and Part VI is wholly devoted to their study.

A *random walk* is commonly defined as the cumulation of an independent, identically and symmetrically distributed zero-mean sequence. In the framework of §13.3 this generation mechanism is characterized as a unit autoregressive root as in **13.11**. In the econometric modelling context the sequence of differences is often treated as locally dependent and perhaps also locally nonstationary, but the simple case of independent increments exhibits properties of general importance in limit theory. Let $W_0 = 0$ and for $n = 1, 2, 3, \dots,$ define the process

$$W_n = W_{n-1} + U_n = \sum_{t=1}^{n} U_t.$$

Specifically, consider the case where the increments have the Rademacher distribution, with $U_n = -1$ with probability $\frac{1}{2}$ and otherwise $+1$, so that $\mathrm{E}(U_n) = 0$ and $\mathrm{E}(U_n^2) = 1$. These values are shared by all odd-order moments and all even-order moments, respectively. It is easily determined that $\mathrm{E}(W_n) = 0$ for each n, but also that $\mathrm{E}(W_n^2) = n$. The random walk is a globally nonstationary process with the variance increasing linearly with time. It exhibits the square root rule of evolution in the sense that that $W_n/\sqrt{n}$ is a process having mean zero and variance one, for every n. It is, however, dependent on past values at long range, with $\mathrm{E}(W_n W_{n+m}) = n$ for all $m \geq 0$. It is symmetrically distributed, with $-W_n$ having the same distribution as W_n and for $t \leq n$, $W_n - W_t$ has the same distribution as W_{n-t}. Given enough steps, this process must eventually visit every finite integer of the real line. While this particular process is integer-valued, a real-valued case that shares the same symmetry properties has $U_t \sim_{\mathrm{d}} N(0, \sigma^2)$ for variance $\sigma^2 < \infty$.

The random walk provokes additional interesting questions. For example, what is the probability that it will have attained a given extreme point, after a specified number of steps? It turns out that to calculate the distribution of the random variable $\max_{1\leq t\leq n} W_t$ is a pleasingly straightforward exercise, using what is known as the *reflection principle*. Thus, for any $x > 0$ consider $P(\max_{1\leq t\leq n} W_t > x)$. Define $\tau = \min\{t : W_t > x\}$, the time at which the process first exceeds the specified value. Hence, define a new process

$$W'_t = \begin{cases} W_t & t \leq \tau, \\ W_\tau - (W_t - W_\tau) & t > \tau. \end{cases} \tag{13.22}$$

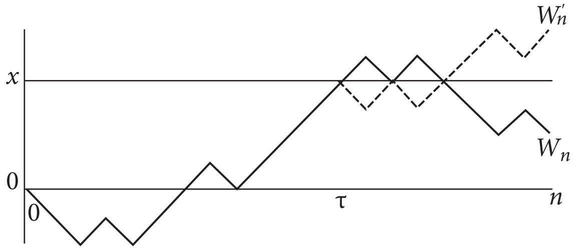

Figure 13.1

The key point to notice is that processes $\{W_n\}$ and $\{W_n'\}$ have the same distribution. From time $\tau + 1$ onwards the signs of the increments are reversed, but by symmetry this sequence has the same probability of arising as the original one; Figure 13.1 illustrates the integer-valued case.

The next thing to notice is that

$$P\big(\max_{1\le t\le n} W_t > x\big) = P(\tau \le n). \tag{13.23}$$

In other words, if $W_\tau > x$ for some $\tau \le n$ it *must* be the case that $\max_{1\le t\le n} W_t > x$. On the other hand, if $\tau > n$, it *cannot* be the case that $\max_{1\le t\le n} W_t > x$. Further, defining $\tau' = \min\{t : W_t' > x\}$, note that $\tau' = \tau$ according to the construction of (13.22). Hence,

$$\begin{aligned} P(\tau \le n) &= P(\tau \le n, W_n \le x) + P(\tau \le n, W_n > x) \\ &= P(\tau' \le n, W_n' \le x) + P(\tau \le n, W_n > x) \\ &= P(\tau \le n, W_n \ge x) + P(\tau \le n, W_n > x) \\ &= 2P(W_n > x) + P(W_n = x). \end{aligned} \tag{13.24}$$

The third equality in (13.24) holds because when $n \ge \tau$, $W_n' \le x$ if and only if $W_n \ge x$, by (13.22). The final equality then holds because $\{W_n \ge x\} \subset \{\tau \le n\}$ by definition of τ.

This argument is valid whenever the distributions of U_t and $-U_t$ are the same and hence W_t' in (13.22) is distributed like W_t. Putting together (13.23) with (13.24) gives the following result.

13.15 Theorem If $\{W_n, n \ge 1\}$ is a random walk having symmetrically distributed independent increments with zero mean, then for $x > 0$,

$$P\big(\max_{1\le t\le n} W_t > x\big) = 2P(W_n > x) + P(W_n = x). \quad \square \tag{13.25}$$

Since $\min_{1\le t\le n} W_t = \max_{1\le t\le n}(-W_t)$, by symmetry the argument applies equivalently to show

$$P\big(\min_{1\le t\le n} W_t < -x\big) = 2P(W_n < -x) + P(W_n = -x), \qquad (13.26)$$

and since the events specified in (13.25) and (13.26) are mutually exclusive, the probabilities can be summed to give the following.

13.16 Corollary Under the conditions of **13.15**,

$$P\big(\max_{1\le t\le n} |W_t| > x\big) = 2P(|W_n| > x) + P(|W_n| = x). \quad \square$$

These results might appear as nothing more that a nice demonstration for its own sake, but in fact they have a much wider application when combined with the notion of an invariance principle, to be shown in §28.4 and subsequently.

14
Dependence

14.1 Shift Transformations

In a probability space $(\Omega, \mathcal{F}, P)$ consider a 1–1 measurable mapping $T : \Omega \mapsto \Omega$ (onto). If $\omega \in \Omega$, $T\omega$ is a second outcome identified by the application of transformation T to ω. If $E \in \mathcal{F}$ is a random event, TE is the event that is defined by applying transformation T to each element of E. T is called *measure-preserving* if $P(TE) = P(E)$ for all $E \in \mathcal{F}$.

Thus if ω represents the drawing of an infinite sequence, T generates a second sequence from the first. In particular the *shift transformation* for a sequence $\{X_t(\omega)\}_1^\infty$ is defined for each t by

$$X_t(T\omega) = X_{t+1}(\omega). \tag{14.1}$$

Here, T takes each outcome ω into the outcome under which the realized value of X occurring in period t now occurs in period $t + 1$, for every t. In effect, each coordinate of the sequence from $t = 2$ onwards is relabelled with the previous period's index. More generally write $X_t(T^k\omega) = X_{t+k}(\omega)$, the relationship between points in the sequence k periods apart becoming a characteristic of the transformation T^k. Since X_t is a r.v. for all t, both the shift transformation and its inverse T^{-1}, the *backshift* transformation, must be measurable. In this framework, a stationary sequence is simply one for which the shift transformation is measure-preserving. If for a collection of coordinates $t_1, \ldots, t_m$ for any $m \geq 1$,

$$E = \{\omega : X_{t_1}(\omega), \ldots, X_{t_m}(\omega) \in B\} \in \mathcal{F}$$

for $B \in \mathcal{B}^m$ and

$$\begin{aligned} TE &= \{\omega : X_{t_1}(T\omega), \ldots, X_{t_m}(T\omega) \in B\} \\ &= \{\omega : X_{t_1+1}(\omega), \ldots, X_{t_m+1}(\omega) \in B\}, \end{aligned}$$

then a stationary sequence has the property $P(TE) = P(E)$ for all such E. Clearly, if this is true for T then it is true for T^k for any k.

Notice how taken together the single r.v. $X_1(\omega) : \Omega \mapsto \mathbb{R}$ and the shift transformation T generate a complete description of the sequence $\{X_t(\omega)\}_1^\infty$. This can

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0014

be seen as follows. Given $X_1(\omega)$, apply the transformation T to ω and obtain $X_2 = X_1(T\omega)$. Doing this for each $\omega \in \Omega$ defines the mapping $X_2(\omega) : \Omega \mapsto \mathbb{R}$. Then, $X_3 = X_2(T\omega)$ and iterating the procedure generates as many points in the sequence as required.

14.1 Example Consider Example **12.1**. Let $\{X_t(\omega)\}_1^\infty$ be the sequence of coin tosses (1 for heads, 0 for tails) beginning 110100100011.... Somewhere on the interval $[0,1]$ of reals in binary representation there is also a sequence $\{X_t(\omega')\}_1^\infty$ beginning 10100100011..., identical to the sequence indexed by ω apart from the dropping of the initial digit and the backshift of the remainder by one position. Likewise there is another sequence $\{X_t(\omega'')\}_1^\infty$, a backshifted version of $\{X_t(\omega')\}_1^\infty$ beginning 0100100011...; and so forth. Defining the transformation T by $T^{-1}\omega = \omega'$, $T^{-1}\omega' = \omega''$, etc., the sequence $\{X_t(\omega)\}_1^\infty$ can be constructed as the sequence $\{X_1(T^{1-t}\omega)\}_1^\infty$; that is, the sequence of first members of the sequences found by iterating the transformation, in this case beginning $1, 1, 0, \ldots$. □

This device reveals, among other things, the complex structure of the probability space being postulated. To each point $\omega \in \Omega$ there must correspond a countably infinite set of points $T^k\omega \in \Omega$, which reproduce the same sequence apart from the absolute date associated with X_1. The intertemporal properties of a sequence can then be treated as a comparison of two sequences, the original and the sequence lagged k periods.

Econometricians attempt to make inferences about economic behaviour from recorded economic data. In time-series analysis, the sample available is usually a *single* realization of a random sequence, economic history as it actually occurred. Because only one world is ever observed it is easy to make the mistake of looking on the observed sequence as the whole sample space, whereas it is really only one of the many possible realizations the workings of historical chance might have generated. Indeed, in the probability model the whole economic universe is the counterpart of a single random outcome ω; there is an important sense in which the time series analyst is a statistician who draws inferences from single data points! But although a single realization of the sequence can be treated as a mapping from a single ω, it is linked to a countably infinite set of ωs corresponding to the leads and lags of the sequence. The statistical analysis of time series in effect is posing the question: is this set really rich enough to allow inferences to be made about P from a single realization?

14.2 Invariant Events

The amount of dependence in a sequence is the chief factor determining how informative a realization of given length can be about the distribution that generated it.

At one extreme, the i.i.d. sequence is equivalent to a true random sample. The classical theorems of statistics can be applied to this type of distribution. At the other extreme, it is easy to specify sequences for which a single realization can never reveal the parameters of the distribution, even in the limit as its length tends to infinity. The key issue is whether averaging operations applied to sequences have useful limiting properties; for example, can parameters of the generation process be consistently estimated in this way?

To clarify these issues, imagine the repeated sampling of random sequences $\{X_t(\omega)\}_1^\infty$. In other words, imagine being given a function $X_1(\cdot)$ and transformation T, making a random drawing ω from Ω, and constructing the corresponding sequence. Example **14.1** illustrates the procedure. Imagine repeating this procedure N times with independent drawings $\omega_i \in \Omega$ for $i = 1, \dots, N$ and constructing the average of these realizations for some fixed date t_0. The average $\bar{X}_{N,t_0} = N^{-1}\sum_{i=1}^N X_{t_0}(\omega_i)$ is called an *ensemble average*. Figure 14.1 illustrates this procedure, showing a sample of three realizations of the sequence. The ensemble average is the average of the points falling on the vertical line labelled t_0.

Contrast the ensemble average with the *time average* of length n for some given $\omega_i \in \Omega$, $\bar{X}_n(\omega_i) = n^{-1}\sum_{t=1}^n X_t(\omega_i)$. It is clear that the limits of the time average and the ensemble average as n and N respectively go to infinity are not in general the same. As the average of independent drawings the ensemble average should (see e.g. **24.5**) tend to the marginal expectation $\mathrm{E}(X_{t_0})$. However, the time average will only do this in special cases. If the sequence is nonstationary $\mathrm{E}(X_{t_0})$ may depend upon t_0; but even assuming stationarity it is still possible that different realizations of the sequence depend on random effects that are common to all t so that a time average can never 'average them away', although the ensemble average can do so.

In a probability space $(\Omega, \mathcal{F}, P)$ an event $E \in \mathcal{F}$ is said to be *invariant* under a transformation T if $P(TE \triangle E) = 0$. The criterion for invariance is sometimes given

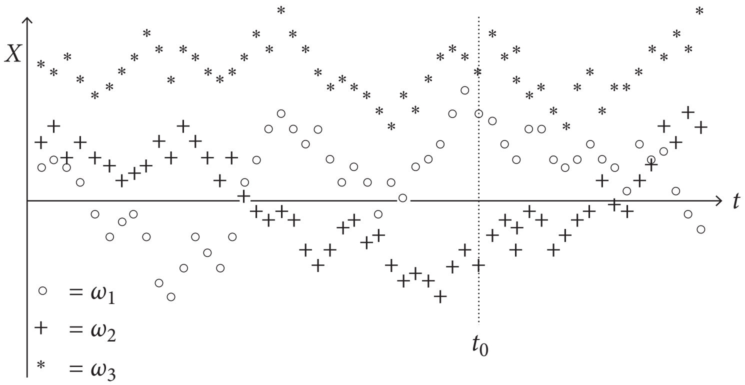

Figure 14.1

as $TE = E$, but allowing the two events to differ by a set of measure zero does not change anything important in the theory. The set of events in $\mathcal{F}$ that are invariant under the shift transformation is denoted $\mathcal{I}$.

14.2 Theorem $\mathcal{I}$ is a σ-field.

Proof Since T is onto, Ω is clearly invariant. Since T is also 1–1,

$$TE^c \triangle E^c = (TE)^c \triangle E^c = TE \triangle E \tag{14.2}$$

by definition. And given $\{E_n \in \mathcal{I}, n \in \mathbb{N}\}$,

$$P(TE_n \triangle E_n) = P(TE_n - E_n) + P(E_n - TE_n) = 0 \tag{14.3}$$

for each n and also

$$T\Big(\bigcup_n E_n\Big) \triangle \bigcup_n E_n = \bigcup_n TE_n \triangle \bigcup_n E_n, \tag{14.4}$$

using **1.2**(i). By **1.1**(i) and then **1.1**(iii),

$$\bigcup_n TE_n - \bigcup_n E_n = \bigcup_n \Big(TE_n \cap \bigcap_m E_m^c\Big) \subseteq \bigcup_n (TE_n - E_n) \tag{14.5}$$

and similarly,

$$\bigcup_n E_n - \bigcup_n TE_n \subseteq \bigcup_n (E_n - TE_n). \tag{14.6}$$

The conclusion $P\big(T(\bigcup_n E_n) \triangle (\bigcup_n E_n)\big) = 0$ now follows by (14.3) and **3.12**(ii), completing the proof. ∎

An *invariant random variable* is one that is $\mathcal{I}/\mathcal{B}$-measurable. An invariant r.v. $Z(\omega)$ has the property that $Z(T\omega) = Z(\omega)$ a.s. and an $\mathcal{I}/\mathcal{B}$-measurable sequence $\{Z_t(\omega)\}_1^\infty$ is trivial in the sense that $Z_t(\omega) = Z_1(\omega)$ a.s. for every t. The invariant events and associated r.v.s constitute those aspects of the probability model that do not alter with the passage of time.

14.3 Example Consider the sequence $\{X_t(\omega)\}$ where $X_t(\omega) = Y_t(\omega) + Z(\omega)$, $\{Y_t(\omega)\}$ being a random sequence and $Z(\omega)$ an r.v. An example of an invariant event is $E = \{\omega : Z(\omega) \le z, Y_t(\omega) \in \mathbb{R}\}$. Clearly E and TE are the same event since Z is the only thing subject to a condition. □

Figure 14.1 illustrates this case. If $\{Y_t(\omega)\}$ is a zero-mean stationary sequence, the figure illustrates the cases $Z(\omega_1) = Z(\omega_2) = 0$ and $Z(\omega_3) > 0$. Even if $\mathrm{E}(Z) = 0$, the influence of $Z(\omega)$ in the time average is not averaged out in the limit as it will be from the ensemble average.

The behaviour of the time average as $n \to \infty$ is summarized by the following fundamental result from Doob ([60]: th. X.2.1).

14.4 Theorem Let a stationary sequence $\{X_t(\omega)\}_1^\infty$ be defined by a measurable mapping $X_1(\omega)$ and measure-preserving shift transformation T, such that $X_t(\omega) = X_1(T^{t-1}\omega)$ and let $S_n(\omega) = \sum_{t=1}^n X_t(\omega)$. If $\mathrm{E}|X_1| < \infty$,

$$\lim_{n\to\infty} S_n(\omega)/n = \mathrm{E}(X_1|\mathcal{I})(\omega) \text{ a.s.} \quad \square \tag{14.7}$$

In words, the limiting case of the time average can be identified with the mean of the distribution conditional on the σ-field of invariant events.

The proof of **14.4** requires a technical lemma.

14.5 Lemma For any real β, let $\Lambda(\beta) = \{\omega : \sup_n (S_n(\omega) - n\beta) > 0\}$. Then for any set $M \in \mathcal{I}$,

$$\int_{M\cap\Lambda(\beta)} X_1(\omega)\mathrm{d}P(\omega) \geq \beta P(M \cap \Lambda(\beta)). \tag{14.8}$$

Proof This is shown for the case $\beta = 0$. To generalize consider the sequence $\{X_t - \beta\}$, which is stationary if $\{X_t\}$ is stationary. Write Λ for $\Lambda(0)$ and let $\Lambda_j = \{\omega : \max_{1\leq k\leq j} S_k(\omega) > 0\}$, the set of outcomes for which the partial sum is positive at least once by time j. Note, the sequence $\{\Lambda_j\}$ is monotone and $\Lambda_j \uparrow \Lambda$ as $j \to \infty$. Also let

$$N_{nj} = T^{-j}\Lambda_{n-j} = \left\{\omega : \max_{1\leq k\leq n-j} S_k(T^j\omega) > 0\right\}. \tag{14.9}$$

Since

$$S_k(T^j\omega) = \sum_{t=1}^{k} X_t(T^j\omega) = \sum_{t=j+1}^{j+k} X_t(\omega) = S_{j+k}(\omega) - S_j(\omega) \tag{14.10}$$

and $S_0 = 0$,

$$N_{nj} = \left\{\omega : \max_{j+1\leq k\leq n} (S_k(\omega) - S_j(\omega)) > 0\right\}, 0 \leq j \leq n-1. \tag{14.11}$$

This is the set of outcomes for which the partial sums of the coordinates from $j+1$ to n are positive at least once. The inequality

$$\sum_{j=0}^{n-1} X_{j+1}(\omega)1_{N_{nj}}(\omega) \geq 0, \text{ all } \omega \in \Omega \tag{14.12}$$

is explained below. Integrating this sum over the invariant set M gives

$$\begin{aligned} 0 &\leq \sum_{j=0}^{n-1} \int_{M\cap N_{nj}} X_{j+1}(\omega)\mathrm{d}P(\omega) \\ &= \sum_{j=0}^{n-1} \int_{M\cap\Lambda_{n-j}} X_{j+1}(T^{-j}\omega)\mathrm{d}P(\omega) \\ &= \sum_{j=0}^{n-1} \int_{M\cap\Lambda_{n-j}} X_1(\omega)\mathrm{d}P(\omega) = \sum_{j=1}^{n} \int_{M\cap\Lambda_j} X_1(\omega)\mathrm{d}P(\omega). \end{aligned} \tag{14.13}$$

Here, the first equality uses the fact that $\Lambda_{n-j} = \{\omega : T^{-j}\omega \in N_{nj}\}$ and the measure-preserving property of T and the third is by reversing the order of summation. The dominated convergence theorem applied to $\{X_1 1_{M\cap\Lambda_j}\}$, with $|X_1|$ as the dominating function, yields

$$\int_{M\cap\Lambda_j} X_1(\omega)\mathrm{d}P(\omega) \to \int_{M\cap\Lambda} X_1(\omega)\mathrm{d}P(\omega). \tag{14.14}$$

This limit is equal to the Cesàro limit by **2.16**, so that, as required,

$$\int_{M\cap\Lambda} X_1(\omega)\mathrm{d}P(\omega) = \lim_{n\to\infty} \frac{1}{n}\sum_{j=1}^{n} \int_{M\cap\Lambda_j} X_1(\omega)\mathrm{d}P(\omega) \geq 0. \quad \blacksquare \tag{14.15}$$

The inequality in (14.12) is not self-evident, but is justified as follows. The expression on the left is the sum containing only those $X_t(\omega)$ having the property that in realization ω, the partial sums from the point t onwards are positive at least once, otherwise the t^{th} contribution to the sum is 0. The sum includes only X_t lying in segments of the sequence over which S_k increases, so that their net contribution must be positive. It would be zero only in the case $X_t \leq 0$ for $1 \leq t \leq n$. Figure 14.2 depicts a realization. '∘' shows values of $S_k(\omega)$ for $k = 1, \dots, n$, so the $X_t(\omega)$ are the vertical separations between successive '∘'s. '+' shows the running sum of the terms of (14.12). The coordinates where the X_t are to be omitted from (14.12) are arrowed, the criterion being that there is no '∘' to the right which exceeds the current '∘'.

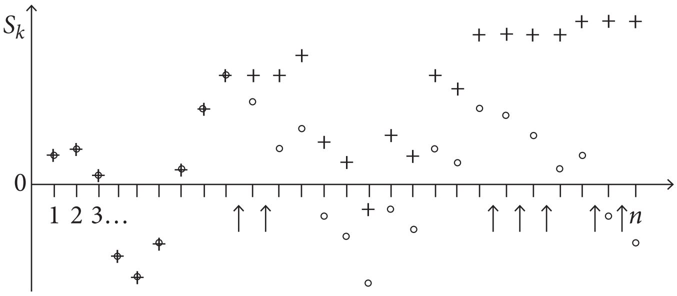

Figure 14.2

Proof of 14.4 The first step is to show that the sequence $\{S_n(\omega)/n\}_1^\infty$ converges almost surely to an invariant r.v. $S(\omega)$. Consider $\bar{S}(\omega) = \text{limsup}_n\, S_n(\omega)/n$.

$$\bar{S}(T\omega) = \lim_{m\to\infty} \sup_{n\geq m}\left\{\frac{S_{n+1}(\omega)}{n+1}\left(\frac{n+1}{n}\right) - \frac{X_1(\omega)}{n}\right\} = \bar{S}(\omega) \tag{14.16}$$

so that $\bar{S}(\omega)$ is invariant, the same being true for $\underline{S}\,(\omega) = \text{liminf}_n\, S_n(\omega)/n$. Hence, the event $M(\alpha,\beta) = \{\omega :\ \underline{S}\,(\omega) < \alpha < \beta < \bar{S}(\omega)\}$ is invariant. Since $\sup_n S_n(\omega)/n \geq \bar{S}(\omega)$, $M(\alpha,\beta) \subseteq \Lambda(\beta)$ where $\Lambda(\beta)$ is defined in the statement of **14.5**. Hence, putting $M = M(\alpha,\beta)$ in the lemma gives

$$\int_{M(\alpha,\beta)} X_1(\omega)\mathrm{d}P(\omega) \geq \beta P\big(M(\alpha,\beta)\big). \tag{14.17}$$

Now reprise the argument of **14.5** with $X_t(\omega)$ replaced by $-X_t(\omega)$ and β by $-\alpha$. Note that $M(\alpha,\beta) \subseteq \{\omega : \sup_n\big(-S_n(\omega)/n\big) > -\alpha\}$ and hence that

$$\int_{M(\alpha,\beta)} X_1(\omega)\mathrm{d}P(\omega) \leq \alpha P\big(M(\alpha,\beta)\big). \tag{14.18}$$

Since the left-hand sides of (14.18) and (14.17) are equal and $\alpha < \beta$, it follows that $P\big(M(\alpha,\beta)\big) = 0$; that is, $\bar{S}(\omega) = \underline{S}(\omega) = S(\omega)$ with probability 1.

This completes the first stage of the proof. It is now required to show that $S = \mathrm{E}(X_1|\mathcal{I})$ a.s. According to equation (10.18) this means that

$$\int_M X_1(\omega)\mathrm{d}P(\omega) = \int_M S(\omega)\mathrm{d}P(\omega),\ \text{each } M \in \mathcal{I}. \tag{14.19}$$

Since M is invariant,

$$\int_M X_1(\omega)\mathrm{d}P(\omega) = \frac{1}{n}\sum_{t=1}^{n}\int_{T^{-t}M} X_t(\omega)\mathrm{d}P(\omega) = \int_M \Big(\frac{1}{n}S_n(\omega)\Big)\mathrm{d}P(\omega) \tag{14.20}$$

and the issue hinges on the convergence of the right-hand member of (14.20) to $\mathrm{E}(S1_M)$. Since the sequence $\{X_t\}$ is stationary and integrable it is also uniformly integrable and the same is true of the sequence $\{Y_t\}$ where $Y_t = X_t 1_M$ and $M \in \mathcal{I}$. For $\varepsilon > 0$ it is possible by **12.10** to choose an event $E \in \mathcal{F}$ with $P(E) < \delta$, such that $\sup_t \mathrm{E}(|Y_t| 1_E) < \varepsilon$. For the same E, the triangle inequality gives

$$\int_E \left|\frac{1}{n}\sum_{t=1}^{n} Y_t\right| \mathrm{d}P \le \frac{1}{n}\sum_{t=1}^{n}\Big(\int_E |Y_t|\mathrm{d}P\Big) < \varepsilon. \tag{14.21}$$

By the same argument, also using stationarity and integrability of Y_t,

$$\int \left|\frac{1}{n}\sum_{t=1}^{n} Y_t\right| \mathrm{d}P \le \frac{1}{n}\sum_{t=1}^{n}\int |Y_t|\mathrm{d}P = \mathrm{E}|Y_1| < \infty. \tag{14.22}$$

Hence by **12.10** the sequence $\{n^{-1}\sum_{t=1}^{n} Y_t\}$ is also uniformly integrable, where $n^{-1}\sum_{t=1}^{n} Y_t = n^{-1}S_n 1_M$. If $n^{-1}S_n \to S$ a.s. it is clear that $n^{-1}S_n 1_M \to S1_M$ a.s., so by **12.8**,

$$\int_M \Big(\frac{1}{n}S_n(\omega)\Big)\mathrm{d}P(\omega) \to \int_M S(\omega)\mathrm{d}P(\omega). \tag{14.23}$$

Since in (14.20) n is arbitrary, (14.23) and (14.20) together give (14.19) and the proof is complete. ∎

14.3 Ergodicity and Mixing

The property of a stationary sequence that ensures the time average and the ensemble average have the same almost sure limit is *ergodicity*, which is defined in terms of the probability of invariant events. A measure-preserving transformation T is ergodic if either $P(E) = 1$ or $P(E) = 0$ for all $E \in \mathcal{I}$ where $\mathcal{I}$ is the σ-field of invariant events under T. A stationary sequence $\{X_t(\omega)\}_{-\infty}^{+\infty}$ is said to be ergodic if shift transformation T is measure-preserving and ergodic and $X_t(\omega) = X_1(T^{t-1}\omega)$ for every t.[1] Events that are invariant under ergodic transformations either occur

[1] Some authors such as Doob use the term *metrically transitive* for ergodic.

almost surely, or do not occur almost surely. In the case of **14.3** for example, Z must be a constant almost surely.

Intuitively, stationarity and ergodicity together are seen to be sufficient conditions for time averages and ensemble averages to converge to the same limit. Stationarity implies that, for example, $\mu = \mathrm{E}\{X_1(\omega)\}$ is the mean not just of X_1 but of any member of the sequence. The existence of random events that are invariant under the shift transformation means that there are regions of the sample space that a particular infinite realization of the sequence will never visit. If $P(TE \triangle E) = 0$ where the event E represents the sequence taking values in a particular region, an infinite realization never visits that region if E^c occurs. Ruling out invariant events other than the trivial ones having probabilities of either 0 or 1 ensures that a sequence eventually visits all parts of the space, with probability 1. In this case, time averaging and ensemble averaging are equivalent operations in the limit.

The following corollary, known as the ergodic theorem, is the main reason for interest in Theorem **14.4**.

14.6 Theorem Let $\{X_t(\omega)\}_1^\infty$ be a stationary, ergodic, integrable sequence. Then

$$\lim_{n\to\infty} S_n(\omega)/n = \mathrm{E}(X_1) \text{ a.s.} \tag{14.24}$$

Proof This is immediate from **14.4**, since by ergodicity, $\mathrm{E}(X_1|\mathcal{I}) = \mathrm{E}(X_1)$ a.s. ∎

In an ergodic sequence, the information contained in $\mathcal{I}$ is trivial and conditioning on events of probability zero or one is a trivial operation almost surely.

The ergodic theorem is an example of a law of large numbers, the first of several such theorems to be studied in later chapters. Unlike most of the subsequent examples this one is for stationary sequences. Its practical applications in econometrics are limited by the fact that the stationarity assumption is often inappropriate, but it is of much theoretical interest because ergodicity is a very mild constraint on the dependence.

A transformation that is measure-preserving eventually mixes up the outcomes in a non-invariant event A with those in A^c. The measure-preserving property rules out mapping sets into proper subsets of themselves, so $TA \cap A^c$ is nonempty. Repeated iterations of the transformation generate a sequence of sets $\{T^k A\}$ containing different mixtures of the elements of A and A^c. A positive dependence of B on A implies a negative dependence of B on A^c; that is, if $P(A \cap B) > P(A)P(B)$ then $P(A^c \cap B) = P(B) - P(A \cap B) < P(B) - P(A)P(B) = P(A^c)P(B)$. Intuition suggests that the average dependence of B on mixtures of A and A^c should tend to zero as the mixing-up proceeds. In fact, ergodicity can be characterized in just this kind of way, as the following theorem shows.

14.7 Theorem A measure-preserving shift transformation T is ergodic iff for any pair of events $A, B \in \mathcal{F}$,

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} P(T^k A \cap B) = P(A)P(B). \tag{14.25}$$

Proof To show 'only if' let A be an invariant event and $B = A$. Then $P(T^k A \cap B) = P(A)$ for all k and hence the left-hand side of (14.25) is equal to $P(A)$ for all k. This gives $P(A) = P(A)^2$, implying $P(A) = 0$ or 1.

To show 'if', for arbitrary $A \in \mathcal{F}$ apply the ergodic theorem to the indicator functions of the sets $T^k A$, where T is measure-preserving and ergodic, to give

$$\lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} 1_{T^k A}(\omega) = P(A) \text{ a.s.} \tag{14.26}$$

But for any $B \in \mathcal{F}$,

$$\begin{aligned}\int\left|\frac{1}{n}\sum_{k=1}^{n} 1_{T^k A}(\omega) - P(A)\right|dP(\omega) &\geq \int_B\left|\frac{1}{n}\sum_{k=1}^{n} 1_{T^k A}(\omega) - P(A)\right|dP(\omega)\\ &\geq \left|\int_B\left(\frac{1}{n}\sum_{k=1}^{n} 1_{T^k A}(\omega) - P(A)\right)dP(\omega)\right|\\ &= \left|\frac{1}{n}\sum_{k=1}^{n} P(T^k A \cap B) - P(A)P(B)\right|.\end{aligned} \tag{14.27}$$

The sequence whose absolute value is the integrand in the left-hand member of (14.27) converges almost surely to zero as $n \to \infty$ by (14.26); it is bounded absolutely by $1 + P(A)$ uniformly in n, so is clearly uniformly integrable. Hence, the left-hand member of (14.27) converges to zero by **12.8** and the theorem follows. ∎

Following from this result, ergodicity of a stationary sequence is often associated with convergence to zero in Cesàro sum of the autocovariances. Indeed, in a Gaussian sequence the conditions are equivalent.

14.8 Corollary If $\{X_t(\omega)\}_1^\infty$ is a stationary, ergodic, square-integrable sequence then

$$\frac{1}{n}\sum_{k=1}^{n} \mathrm{Cov}(X_1, X_k) \to 0 \text{ as } n \to \infty. \tag{14.28}$$

Proof Setting $B = A$ and defining a real sequence by the indicators of T^kA, $X_k(\omega) = 1_{T^kA}(\omega)$, (14.25) is equivalent to (14.28). First extend this result to sequences of simple r.v.s. Let $X_1(\omega) = \sum_i \alpha_i 1_{A_i}(\omega)$ so that $X_2(\omega) = X_1(T\omega) = \sum_i \alpha_i 1_{T^{-1}A_i}(\omega)$. The main point to be established is that the difference between X_2 and a simple r.v. can be ignored in integration. In other words, the sets $T^{-1}A_i$ must form a partition of Ω apart possibly from sets of measure 0. Since T is measure-preserving, $P(\bigcup_i T^{-1}A_i) = P\big(T^{-1}(\bigcup_i A_i)\big) = P(\bigcup_i A_i) = 1$ using **1.2**(ii) and hence $P(\Omega - \bigcup_i T^{-1}A_i) = 0$. And since $\sum_i P(T^{-1}A_i) = \sum_i P(A_i) = 1$, additivity of the measure implies that the collection $\{T^{-1}A_i\}$ is also disjoint apart from possible sets of measure 0, verifying the required property.

This argument extends by induction to X_k for any $k \in \mathbb{N}$. Hence,

$$\begin{aligned} \mathrm{E}(X_1X_k) &= \mathrm{E}\bigg(\sum_i\sum_j \alpha_i\alpha_j 1_{A_i \cap T^{-k}A_j(\omega)}\bigg) \\ &= \sum_i\sum_j \alpha_i\alpha_j P(A_i \cap T^{-k}A_j) \end{aligned} \tag{14.29}$$

(the sum being absolutely convergent by assumption) and by **14.7**,

$$\begin{aligned} \lim_{n\to\infty} \frac{1}{n}\sum_{k=1}^{n} \mathrm{E}(X_1X_k) &= \sum_i\sum_j \alpha_i\alpha_j \bigg(\lim_{n\to\infty}\frac{1}{n}\sum_{k=1}^{n} P(A_i \cap T^kA_j)\bigg) \\ &= \sum_i\sum_j \alpha_i\alpha_j P(A_i)P(A_j) = \mathrm{E}(X_1)^2 \end{aligned} \tag{14.30}$$

where $\mathrm{E}(X_1)^2 = \mathrm{E}(X_1)\mathrm{E}(X_k)$ for any k, by stationarity. The theorem extends to general sequences by the usual application of **3.35** and the monotone convergence theorem. ∎

An important fact is that in stationary sequences, ergodicity is preserved under measurable transformations including those involving arbitrary orders of lead and lag.

14.9 Theorem Let $x = \{X_t\}_{-\infty}^{\infty}$ be a stationary ergodic sequence and $\phi : \mathbb{R}^\infty \mapsto \mathbb{R}$ a measurable function. The sequence $y = \{Y_t\}_{-\infty}^{\infty}$ where

$$Y_t = \phi(\dots, X_{t-1}, X_t, X_{t+1}, \dots)$$

is also stationary and ergodic.

Proof Let $E_y \in \mathcal{F}$ be defined, as is possible under measurability, by

$$E_y = \{\omega : y(\omega) = (\dots, \phi(x)_{t-1}, \phi(x)_t, \phi(x)_{t+1}, \dots) \in B\}$$

for some $B \in \mathcal{B}^\infty$. Implicitly, also defined for the same B is the set $E_x \in \mathcal{F}$ such that

$$E_x = \{\omega : (\dots, \phi_{t-1}(x(\omega)), \phi_t(x(\omega)), \phi_{t+1}(x(\omega)), \dots) \in B\}.$$

By construction, $P(E_y \triangle E_x) = 0$. Since x is stationary, $P(TE_x) = P(E_x)$ from which it follows that $P(TE_y) = P(E_y)$, so y is also stationary. Also, if $P(E_x) = 0$ then $P(E_y) = 0$ and if $P(E_x) = 1$ then $P(E_y) = 1$. Suppose $E_y \in \mathcal{I}$ so that $P(TE_y \triangle E_y) = 0$. Since $Y_{t+k} = \phi(\dots, X_{t+k-1}, X_{t+k}, X_{t+k+1}, \dots)$ for any k and T is measure preserving, $P(TE_y \triangle TE_x) = 0$ also holds and hence $P(TE_x \triangle E_x) = 0$ and $E_x \in \mathcal{I}$. Since E_y was an arbitrary element of $\mathcal{I}$ it follows that if x is ergodic so is y. ∎

Note that the transformation ϕ preserves stationarity because, crucially, it does not depend on any fixed dates, t. Cumulation is a transformation of a stationary sequence yielding a nonstationary sequence as in Example **13.11** and §13.4, which might appear as a counterexample to this result, but to construct a random walk requires specifying an initial date $t = 0$. Setting the initial date at $-\infty$ does not result in a measurable transformation.

Theorems **14.7** and **14.8** might appear to suggest that ergodicity implies some form of asymptotic independence of the sequence, since one condition under which (14.28) certainly holds is where $\mathrm{Cov}(X_1, X_k) \to 0$ as $k \to \infty$. But this is not so. The following example illustrates nicely what ergodicity implies and does not imply.

14.10 Example Let the probability space $(\Omega, \mathcal{F}, P)$ be defined by $\Omega = \{0, 1\}$ with $\mathcal{F} = (\{\emptyset\}, \{0\}, \{1\}, \{0, 1\})$ and $P(\omega) = 0.5$ for $\omega = 0$ and $\omega = 1$. Let T be the transformation that sets $T0 = 1$ and $T1 = 0$. In this setup a random sequence $\{X_t(\omega)\}$ may be defined by letting $X_1(\omega) = \omega$ and generating the sequence by iterating T. These sequences always consist of alternating 0s and 1s, but the initial value is randomly chosen with equal probabilities. Now, T is measure-preserving; the invariant events are Ω and $\emptyset$, both trivial, so the sequence is ergodic. It can be verified that $\lim_{n\to\infty} n^{-1} \sum_{k=1}^n P(T^k A \cap B) = P(A)P(B)$ for every pair $A, B \in \mathcal{F}$. For instance, let $A = B = \{1\}$ and then

$$P(T^k A \cap B) = \begin{cases} P(B) = 0.5, & k \text{ even} \\ P(\{\emptyset\}) = 0, & k \text{ odd} \end{cases}$$

so that the limit is indeed 0.25 as required. The ergodic theorem holds since the time average of the sequence always converges to 0.5 which is the same as the ensemble mean of $X_1(\omega)$. □

In this example X_t is perfectly predictable from X_1, for any t. This shows that ergodicity does not imply independence between different parts of the sequence, even as the time separation increases. By contrast, a *mixing* sequence has this property. A measure-preserving, ergodic shift transformation T is said to be mixing if, for each $A, B \in \mathcal{F}$,

$$\lim_{k\to\infty} P(T^k A \cap B) = P(A)P(B). \tag{14.31}$$

The stationary sequence $\{X_t\}_1^\infty$ is said to be mixing if $X_t(\omega) = X_1(T^{t-1}\omega)$ for each t where T is a mixing transformation.

Compare this condition with (14.25). Cesàro convergence of the sequence $\{P(T^k A \cap B), k \in \mathbb{N}\}$ has been replaced by actual convergence. To obtain a sound intuition about mixing transformations, one cannot do better than reflect on the following oft-quoted example, originally due to Halmos ([89]).

14.11 Example Consider a dry martini initially poured as a layer of vermouth (10% of the volume) on top of the gin (90%). Let G denote the gin and F an arbitrary small region of the fluid, so that $F \cap G$ is the gin contained in F. If $P(\cdot)$ denotes the volume of a region as a proportion of the whole, $P(G) = 0.9$ and $P(F \cap G)/P(F)$, the proportion of gin in F, is initially either 0 or 1. Let T denote the operation of stirring the martini with a swizzle stick, so that $P(T^k F \cap G)/P(F)$ is the proportion of gin in F after k stirs. Assuming the fluid is incompressible, stirring is a measure-preserving transformation in that $P(T^k F) = P(F)$ for all k. If the stirring mixes the martini the proportion of gin in $T^k F$, which is $P(T^k F \cap G)/P(F)$, should tend to $P(G)$ so that each region F of the martini eventually contains 90% gin. □

This is precisely condition (14.31). Repeated applications of a mixing transformation to an event A should eventually mix outcomes in A and A^c so thoroughly that for large enough k the composition of $T^k A$ gives no clues about the original A. Mixing in a real sequence implies that events such as $A = \{\omega : X_t(\omega) \leq a\}$ and $T^k A = \{\omega : X_{t+k}(\omega) \leq a\}$ are becoming independent as k increases. It is immediate, or virtually so, that for stationary mixing sequences the result of **14.8** can be strengthened to $\text{Cov}(X_1, X_k) \to 0$ as $n \to \infty$.

14.4 Sub-σ-fields and Regularity

Here is an alternative approach to studying dependence which considers the collection of sub-σ-fields of events generated by a stochastic sequence. This theory is fundamental to nearly everything done subsequently, particularly because unlike the ergodic theory of the preceding sections, it generalizes beyond the measure-preserving (stationary) case.

Consider a doubly infinite sequence $\{X_t, t \in \mathbb{Z}\}$ (not necessarily stationary) and define the family of sub-σ-fields $\{\mathcal{F}_s^t, s \leq t\}$ where $\mathcal{F}_s^t = \sigma(X_s, \dots, X_t)$ is the smallest σ-field on which the sequence coordinates from dates s to t are measurable. The sets of $\mathcal{F}_s^t$ can be visualized as the inverse images of $(t-s)$-dimensional cylinder sets in $\mathcal{B}^\infty$; compare the discussion of §12.3, recalling that $\mathbb{N}$ and $\mathbb{Z}$ are equipotent. One or other of the bounds can tend to infinity and a particularly important sub-family is the increasing sequence $\{\mathcal{F}_{-\infty}^t, t \in \mathbb{Z}\}$. Think of this as representing the information contained in the sequence up to date t. The σ-field on which the sequence as a whole is measurable is the limiting case $\mathcal{F}_{-\infty}^{+\infty} = \bigvee_t \mathcal{F}_{-\infty}^t$. In cases where the underlying probability model concerns just the sequence $\{X_t\}$, $\mathcal{F}_{-\infty}^{+\infty}$ is identified with $\mathcal{F}$.

Another interesting object is the *remote* σ-field (or tail σ-field), $\mathcal{F}_{-\infty} = \bigcap_t \mathcal{F}_{-\infty}^t$. (As a countable intersection of σ-fields this *is* a σ-field; see page 20.) This σ-field contains events about which something can be learned by observing *any* coordinate of the sequence and it might plausibly be supposed that these events occurred at time $-\infty$, the 'remote' past when the initial conditions for the sequence were set. However, note that the set may be generated in other ways, such as $\bigcap_t \mathcal{F}_t^{+\infty}$, or $\bigcap_t \mathcal{F}_{-t}^t$.

One of the ways to characterize independence in a sequence is to say that any pair of non-overlapping sub-σ-fields, of the form $\mathcal{F}_{t_2}^{t_1}$ and $\mathcal{F}_{t_4}^{t_3}$ where $t_1 \geq t_2 > t_3 \geq t_4$, are independent (see §10.5). One of the most famous early results in the theory of stochastic processes is Kolmogorov's *zero–one law* for independent sequences. This theorem is usually given for the case of a sequence $\{X_t\}_1^\infty$ and the remote σ-field is defined in this case as $\bigcap_{t=1}^\infty \mathcal{F}_t^\infty$.

14.12 Theorem (zero–one law) If the sequence $\{X_t\}_1^\infty$ is independent, every remote event is trivial, having either probability 0 or probability 1.

Proof Let A be a remote event so that $A \in \mathcal{F}^\infty = \bigcap_{t=1}^\infty \mathcal{F}_t^\infty$ and let $\mathcal{G}$ be the collection of events B having the property $P(A \cap B) = P(A)P(B)$. Serial independence implies that for every t, $\mathcal{F}^\infty$ and $\mathcal{F}_1^t$ are independent sub-σ-fields so that $\mathcal{F}_1^t \subseteq \mathcal{G}$. Let $B \in \mathcal{F}_1^t$ for some t and $B' \in \mathcal{F}_1^{t'}$ for some $t' \leq t$. Since the sequence $\{\mathcal{F}_1^t\}$ is nested both B and B' are elements of $\mathcal{F}_1^t$, which is a σ-field and hence $B \cap B' \in \mathcal{F}_1^t$. Therefore $B \cap B' \in \mathcal{G}$, with the same conclusion holding in the obvious way if $t \leq t'$.

It follows by definition that $\mathcal{G}$ is a π-system (see **1.26**). Since by construction $\{A\}$ and $\mathcal{G}$ are collections with the independence property where $\{A\}$ (the singleton with member A) is also trivially a π-system, $\{A\}$ and $\sigma(\mathcal{G})$ have the independence property by **7.7**. However, since $\bigcup_{t=1}^{\infty}\mathcal{F}_1^t \subseteq \mathcal{G}$, $\mathcal{F} = \bigvee_{t=1}^{\infty}\mathcal{F}_1^t \subseteq \sigma(\mathcal{G})$ which means since $\mathcal{F}^{\infty} \subseteq \mathcal{F}$ that $A \in \sigma(\mathcal{G})$. A is therefore independent of itself, with the property $P(A) = P(A \cap A) = P(A)^2$. This is possible only if either $P(A) = 0$ or $P(A) = 1$. ■

The zero–one law shows that for an independent sequence there are no events, other than trivial ones, that can be relevant to all sequence coordinates. However, independence is clearly not a necessary condition for the zero–one property. The interesting problem is to identify the wider class of sequences that possess it.

A sequence $\{X_t\}_{-\infty}^{+\infty}$ is said to be *regular*, or *mixing*, if every remote event has probability 0 or 1. Regularity is the term adopted by Ibragimov and Linnik [105], to whom the basics of this theory are due. In a suitably unified framework regularity is essentially equivalent to the mixing concept defined in §14.3. The following theorem says that in a regular sequence, remote events must be independent of all events in $\mathcal{F}$. Note that trivial events are independent of themselves, on the definition.

14.13 Theorem ([105], th. 17.1.1) $\{X_t\}_{-\infty}^{+\infty}$ is regular iff for every $B \in \mathcal{F}$,

$$\sup_{A\in\mathcal{F}_{-\infty}^t} |P(A \cap B) - P(A)P(B)| \to 0 \text{ as } t \to -\infty. \tag{14.32}$$

Proof To prove 'if', suppose $\exists\, E \in \mathcal{F}_{-\infty}$ with $0 < P(E) < 1$ so that $\{X_t\}_{-\infty}^{+\infty}$ is not regular. Then $E \in \mathcal{F}_{-\infty}^t$ for every t and so

$$\sup_{A\in\mathcal{F}_{-\infty}^t} |P(A \cap E) - P(A)P(E)| \geq P(E) - P(E)^2 > 0 \tag{14.33}$$

which contradicts (14.32).

To prove 'only if', assume regularity and define random variables $\zeta = 1_A - P(A)$, $A \in \mathcal{F}_{-\infty}^t$ and $\eta = 1_B - P(B)$, $B \in \mathcal{F}$ such that $P(A \cap B) - P(A)P(B) = \mathrm{E}(\zeta\eta)$. Then, by the the law of iterated expectations and the Cauchy–Schwarz inequality,

$$|\mathrm{E}(\zeta\eta)| = |\mathrm{E}\big(\zeta\mathrm{E}(\eta|\mathcal{F}_{-\infty}^t)\big)| \leq \|\zeta\|_2\|\mathrm{E}(\eta|\mathcal{F}_{-\infty}^t)\|_2. \tag{14.34}$$

Note that $\|\zeta\|_2 \leq 1$. Showing $\|\mathrm{E}(\eta|\mathcal{F}_{-\infty}^t)\|_2 \to 0$ as $t \to -\infty$ completes the proof since A is an arbitrary element of $\mathcal{F}_{-\infty}^t$.

Consider the sequence $\{\mathrm{E}(1_B|\mathcal{F}_{-\infty}^{t})(\omega)\}_{-\infty}^{+\infty}$. For any $\omega \in \Omega$,

$$\mathrm{E}(1_B|\mathcal{F}_{-\infty}^{t})(\omega) \to \mathrm{E}(1_B|\mathcal{F}_{-\infty})(\omega) \text{ as } t \to -\infty \tag{14.35}$$

where for $E \in \mathcal{F}_{-\infty}$, by equation (10.18) and the zero–one property,

$$\int_E \mathrm{E}(1_B|\mathcal{F}_{-\infty})(\omega)\mathrm{d}P(\omega) = P(E \cap B) = \begin{cases} P(B), & P(E) = 1 \\ 0, & P(E) = 0. \end{cases} \tag{14.36}$$

It is clear that setting $\mathrm{E}(1_B|\mathcal{F}_{-\infty})(\omega) = P(B)$ a.s. agrees with the definition, so $\mathrm{E}(1_B|\mathcal{F}_{-\infty}^{t}) \to P(B)$ a.s. or equivalently,

$$\left(\mathrm{E}(1_B|\mathcal{F}_{-\infty}^{t}) - P(B)\right)^2 \to 0 \text{ a.s.} \tag{14.37}$$

$|1_B(\omega) - P(B)| < 1$ for all $\omega \in \Omega$ and $\left(\mathrm{E}(1_B|\mathcal{F}_{-\infty}^{t})(\omega) - P(B)\right)^2$ is similarly bounded, uniformly in t. Uniform integrability of the sequence can therefore be assumed and it follows from **12.8** that

$$||\mathrm{E}(\eta|\mathcal{F}_{-\infty}^{t})||_2 = ||\mathrm{E}(1_B|\mathcal{F}_{-\infty}^{t})(\omega) - P(B)||_2 \to 0 \text{ as } t \to -\infty \tag{14.38}$$

as required. ∎

In this theorem, it is less the existence of the limit than the passage to the limit, the fact that the supremum in (14.32) can be made small by choosing $-t$ large, that gives the result practical significance. When the only remote events are trivial, the dependence of X_{t+k} on events in $\mathcal{F}_{-\infty}^{k}$ for fixed k must eventually decline as t increases. The zero–one law is an instant corollary of the necessity part of the theorem, since an independent sequence would certainly satisfy (14.32).

There is an obvious connection between the properties of invariance and remoteness.

14.14 Theorem If T is a measure-preserving shift transformation and $TA = A$ then $A \in \mathcal{F}_{-\infty}$.

Proof If $A \in \mathcal{F}_{-\infty}^{t}$, $TA \in \mathcal{F}_{-\infty}^{t+1}$ and $T^{-1}A \in \mathcal{F}_{-\infty}^{t-1}$. If $TA = A$, $T^{-1}A = A$ and it follows immediately that $A \in (\bigcap_{s=t}^{\infty} \mathcal{F}_{-\infty}^{s}) \cap (\bigcap_{s=-\infty}^{t} \mathcal{F}_{-\infty}^{s}) = \mathcal{F}_{-\infty}$. ∎

The last result of this section establishes formally the relationship between regularity and ergodicity that has been implicit in the foregoing discussion.

14.15 Theorem If a stationary sequence $x(\omega) = \{X_t(\omega)\}_{-\infty}^{+\infty}$ is regular (mixing) it is also ergodic.

Proof Consider the set $A \in \mathcal{F}_{-\infty}^{+\infty}$, being the inverse image of $x(A) \in \mathcal{B}^\infty$. A is the limit of a monotone decreasing sequence $\{A_t, t \geq 1\}$ where $A_t \in \mathcal{F}_{-t}^{t}$ is the inverse image under x of the $(2t+1)$-dimensional cylinder set whose base is the product of the coordinate $\mathcal{B}$-sets for coordinates $-t, \dots, t$ of $x(A)$. Observe that $A_t \subseteq A_{t-1} \in \mathcal{F}_{1-t}^{t-1}$ since the image under x of the latter set has coordinates t and $-t$ unrestricted. $P(A_t) \to P(A)$ by continuity of P.

Let A be invariant. Using stationarity of x and the measure-preserving property of T gives for any k,

$$P(A_t \cap A) = P(A_t \cap T^{-k}A) = P(T^k A_t \cap A). \tag{14.39}$$

Taking k arbitrarily large, regularity implies by (14.31) that $P(A_t \cap A) = P(A_t)P(A)$. Letting $t \to \infty$ then yields $P(A) = P(A)^2$, so that $P(A) = 0$ or 1, as required. ∎

14.5 Strong and Uniform Mixing

The defect of mixing (regularity) as an operational concept is that remote events are of less interest than arbitrary events that happen to be widely separated in time. The extra ingredient needed for a workable theory is the concept of dependence between pairs of sub-σ-fields of events. There are several ways to characterize such dependence but the following are the concepts most commonly exploited in limit theory with econometrics applications.

Let $(\Omega, \mathcal{F}, P)$ be a probability space and let $\mathcal{G}$ and $\mathcal{H}$ be sub-σ-fields of $\mathcal{F}$; then

$$\alpha(\mathcal{G}, \mathcal{H}) = \sup_{G \in \mathcal{G}, H \in \mathcal{H}} |P(G \cap H) - P(G)P(H)| \tag{14.40}$$

is known as the *strong mixing* coefficient and

$$\phi(\mathcal{G}, \mathcal{H}) = \sup_{G \in \mathcal{G}, H \in \mathcal{H}; P(G) > 0} |P(H|G) - P(H)| \tag{14.41}$$

as the *uniform mixing* coefficient. These are alternative measures of the dependence between sub-σ-fields $\mathcal{G}$ and $\mathcal{H}$.

If $\mathcal{G}$ and $\mathcal{H}$ are independent then $\alpha(\mathcal{G}, \mathcal{H}) = 0$ and $\phi(\mathcal{G}, \mathcal{H}) = 0$ and the converse is also true in the case of uniform mixing, although *not* for strong mixing. At first sight there may not appear to be much to choose between the definitions, but since they are set up in terms of suprema of the dependence measures over the sets of events in question, it is the extreme (and possibly anomalous) cases that define the characteristics of the mixing coefficients. The strong mixing concept is in fact weaker than the uniform concept. Since

$$|P(G\cap H)-P(G)P(H)|\le|P(H|G)-P(H)|\le\phi(\mathcal{G},\mathcal{H}) \tag{14.42}$$

for all $G\in\mathcal{G}$ and $H\in\mathcal{H}$ it is clear that $\alpha(\mathcal{G},\mathcal{H})\le\phi(\mathcal{G},\mathcal{H})$. However, the following example shows how the two concepts differ more crucially.

14.16 Example Suppose that, for a sequence of sub-σ-fields $\{\mathcal{G}_m\}_1^\infty$ and a sub-σ-field $\mathcal{H}$, $\alpha(\mathcal{G}_m,\mathcal{H})\to 0$ as $m\to\infty$. This condition is compatible with the existence of sets $G_m\in\mathcal{G}_m$ and $H\in\mathcal{H}$ with the properties $P(G_m)=1/m$ and $P(G_m\cap H)=a/m$ for $a\ne P(H)$. But $\phi(\mathcal{G}_m,\mathcal{H})\ge|P(H|G_m)-P(H)|=|a-P(H)|$ for every $m\ge 1$, showing that $\mathcal{G}_m$ and $\mathcal{H}$ are not independent in the limit. □

Evidently, the strong mixing characterization of 'independence' does not rule out the possibility of dependence between negligible sets.

α and ϕ are not the only mixing coefficients that can be defined and others that have appeared in the literature include

$$\beta(\mathcal{G},\mathcal{H})=\sup_{H\in\mathcal{H}}\mathrm{E}|P(H|\mathcal{G})-P(H)| \tag{14.43}$$

and

$$\rho(\mathcal{G},\mathcal{H})=\sup\frac{|\mathrm{E}(\xi\eta)|}{\|\xi\|_2\|\eta\|_2} \tag{14.44}$$

where the latter supremum is taken with respect to all square-integrable, zero-mean, $\mathcal{G}$-measurable r.v.s ξ and $\mathcal{H}$-measurable r.v.s η. To compare these alternatives, first let $\zeta(\omega)=P(H|\mathcal{G})(\omega)-P(H)$ so that

$$\beta(\mathcal{G},\mathcal{H})=\sup_{H\in\mathcal{H}}\int|\zeta|\mathrm{d}P\ge\sup_{G\in\mathcal{G},H\in\mathcal{H}}\int_G|\zeta|\mathrm{d}P\ge\sup_{G\in\mathcal{G},H\in\mathcal{H}}\left|\int_G\zeta\mathrm{d}P\right|=\alpha(\mathcal{G},\mathcal{H}) \tag{14.45}$$

using (10.18). Moreover, since for any sets $G\in\mathcal{G}$ and $H\in\mathcal{H}$, $\xi(\omega)=1_G(\omega)-P(G)$ and $\eta(\omega)=1_H(\omega)-P(H)$ are members of the set over which $\rho(\mathcal{G},\mathcal{H})$ is defined and $\mathrm{E}(\xi\eta)=P(G\cap H)-P(G)P(H)$ while $|\xi|\le 1$ and $|\eta|\le 1$ for these cases, it is also clear that $\rho\ge\alpha$. Thus, notwithstanding its designation, α-mixing is the weakest of these four 'strong' variants, although it is of course stronger than ordinary regularity characterized by trivial remote events. Also, $\beta \le \phi$ by an immediate corollary of the following result.

14.17 Theorem $|P(H|\mathcal{G}) - P(H)| \le \phi(\mathcal{G},\mathcal{H})$ a.s., for all $H \in \mathcal{H}$. □

The main step in the proof of **14.17** is the following lemma.

14.18 Lemma Let X be an almost surely bounded, $\mathcal{G}$-measurable r.v. Then

$$\sup_{G\in\mathcal{G},P(G)>0} \frac{1}{P(G)}\left|\int_G X\mathrm{d}P\right| = \operatorname{ess\,sup} X. \tag{14.46}$$

Proof According to (8.1), $|\int_G X\mathrm{d}P| \le P(G)\operatorname{ess\,sup} X$ for any set G in the designated class. For any $\varepsilon > 0$, consider the sets

$$\begin{aligned} G^+ &= \{\omega : X(\omega) \ge \operatorname{ess\,sup} X - \varepsilon\} \\ G^- &= \{\omega : -X(\omega) \ge \operatorname{ess\,sup} X - \varepsilon\}. \end{aligned}$$

Both these sets belong to $\mathcal{G}$ and at least one of them is nonempty and has positive probability. Letting

$$G^* = \begin{cases} G^+, & P(G^+) \ge P(G^-) \\ G^-, & \text{otherwise,} \end{cases}$$

$$\frac{1}{P(G^*)}\left|\int_{G^*} X\mathrm{d}P\right| \ge (\text{ess sup}X) - \varepsilon \tag{14.47}$$

and (14.46) follows on letting ε approach 0. ∎

Proof of 14.17 Put $X = P(H|\mathcal{G}) - P(H)$ in the lemma, noting that this is a $\mathcal{G}$-measurable r.v. lying between $+1$ and -1. Then by (10.18), $P(G)^{-1}|\int_G X\mathrm{d}P| = |P(H|G) - P(H)|$ for any $G \in \mathcal{G}$. Hence the lemma together with (14.41) implies that

$$\begin{aligned} \phi(\mathcal{G},\mathcal{H}) &\ge \text{ ess sup}\{P(H|\mathcal{G}) - P(H)\} \\ &\ge |P(H|\mathcal{G}) - P(H)| \end{aligned} \tag{14.48}$$

for any $H \in \mathcal{H}$, with probability 1. ∎

15
Mixing

15.1 Mixing Sequences of Random Variables

For a sequence $\{X_t(\omega)\}_{-\infty}^{\infty}$ let $\mathcal{F}_{-\infty}^{t} = \sigma(\ldots, X_{t-2}, X_{t-1}, X_t)$ as in §14.4 and similarly define $\mathcal{F}_{t+m}^{\infty} = \sigma(X_{t+m}, X_{t+m+1}, X_{t+m+2}, \ldots)$. The sequence is said to be *α-mixing* (or strong mixing) if $\lim_{m\to\infty} \alpha_m = 0$ where

$$\alpha_m = \sup_t \alpha(\mathcal{F}_{-\infty}^{t}, \mathcal{F}_{t+m}^{\infty}) \tag{15.1}$$

and α is defined in (14.40). It is said to be *ϕ-mixing* (or uniform mixing) if $\lim_{m\to\infty} \phi_m = 0$, where

$$\phi_m = \sup_t \phi(\mathcal{F}_{-\infty}^{t}, \mathcal{F}_{t+m}^{\infty}) \tag{15.2}$$

and ϕ is defined in (14.41). ϕ-mixing implies α-mixing as noted in §14.5, while the converse does not hold. Another difference is that ϕ-mixing is not time-reversible; in other words, it is not necessarily the case that $\sup_t \phi(\mathcal{F}_{t+m}^{\infty}, \mathcal{F}_{-\infty}^{t}) = \sup_t \phi(\mathcal{F}_{-\infty}^{t}, \mathcal{F}_{t+m}^{\infty})$. By contrast, α-mixing *is* time-reversible. If the sequence $\{X_t\}_{-\infty}^{+\infty}$ is α-mixing, so is the sequence $\{Y_t\}_{-\infty}^{+\infty}$ where $Y_t = X_{-t}$.

$\{X_t(\omega)\}_{-\infty}^{\infty}$ is also said to be *absolutely regular* if $\lim_{m\to\infty} \beta_m = 0$ where

$$\beta_m = \sup_t \beta(\mathcal{F}_{-\infty}^{t}, \mathcal{F}_{t+m}^{\infty}) \tag{15.3}$$

and β is defined in (14.43). According to the results in §14.5, absolute regularity is a condition intermediate between strong mixing and uniform mixing. On the other hand, if $\{X_t\}_{-\infty}^{+\infty}$ is a stationary, L_2-bounded sequence and $\mathcal{F}_{-\infty}^{0} = \sigma(\ldots, X_{-1}, X_0)$ and $\mathcal{F}_{m}^{+\infty} = \sigma(X_m, X_{m+1}, \ldots)$, the sequence is said to be *completely regular* if $\rho_m = \rho(\mathcal{F}_{-\infty}^{0}, \mathcal{F}_{m}^{+\infty}) \to 0$ where ρ is defined in (14.44). In stationary Gaussian sequences, complete regularity is equivalent to strong mixing. Kolmogorov and Rozanov [115] show that in this case

$$\alpha_m \le \rho_m \le 2\pi\alpha_m. \tag{15.4}$$

In a completely regular sequence, the autocovariances $\gamma_j = \mathrm{E}(X_t X_{t-j})$ must tend to 0 as $j \to \infty$. A sufficient condition for complete regularity can be expressed in

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.

DOI: 10.1093/oso/9780192844507.003.0015

terms of the spectral density function. When it exists, the *spectral density* $f(\lambda)$ is the Fourier transform of the autocovariance function, that is to say,

$$f(\lambda) = \frac{1}{2\pi} \sum_{j=-\infty}^{\infty} \gamma_j e^{i\lambda j},\ \lambda \in [-\pi, \pi]. \tag{15.5}$$

The theorem of Kolmogorov and Rozanov leads to the result proved by Ibragimov and Linnik ([105] th. 17.3.3) that a stationary Gaussian sequence is strong mixing when $f(\lambda)$ exists and is continuous and strictly positive everywhere on $[-\pi, \pi]$.

This topic is something of a terminological minefield. 'Regularity' is an undescriptive term and there does not seem to be unanimity among authors with regard to usage, complete regularity and absolute regularity sometimes being used synonymously. Nor is the list of mixing concepts given here by any means exhaustive. Fortunately, this confusion can be avoided by sticking with the strong and uniform cases. There are some applications in which absolute regularity provides just the right condition but these will not arise in the sequel. Incidentally, the term 'weak mixing' might be thought appropriate as a synonym for regularity, but should be avoided as there is a risk of confusion with *weak dependence*, a term used, often somewhat imprecisely, to refer to sequences having summable covariances. *Strongly dependent* sequences may be stationary and mixing, but their covariances are non-summable. ('Weak' implies *less* dependence than 'strong' in this instance, not more!)

Confining attention now to the strong and uniform mixing definitions, measures of the dependence in a sequence can be based in various ways on the rate at which the mixing coefficients α_m or ϕ_m tend to zero. To avoid repetition the discussion will focus on strong mixing but the following remarks apply equally to the uniform mixing case, on substituting ϕ for α throughout. Since the collections $\mathcal{F}_{-\infty}^{t}$ and $\mathcal{F}_{t+m}^{\infty}$ are respectively non-decreasing in t and non-increasing in t and m, the sequence $\{\alpha_m\}$ is monotone.

The rate of convergence is often quantified by a summability criterion, that for some number $\varphi > 0$, $\alpha_m \to 0$ sufficiently fast that

$$\sum_{m=1}^{\infty} \alpha_m^{1/\varphi} < \infty. \tag{15.6}$$

The term *size* has been coined to describe the rate of convergence of the mixing numbers, although different definitions have been used by different authors and the terminology should be used with caution. One possibility is to say that the sequence is of size $-\varphi$ if the mixing numbers satisfy (15.6). However, the commonest usage (see for example [185]) is to say that a sequence is α-mixing of size $-\varphi_0$ if $\alpha_m = O(m^{-\varphi})$ for some $\varphi > \varphi_0$.[1] It is clear that such sequences are summable when raised to the power of $1/\varphi_0$, so that this concept of size is stronger than the summability concept. One temptation to be avoided is to define the size as '$-\varphi$, where φ is the largest constant such that the $\alpha_m^{1/\varphi}$ are summable'; for no such number may exist. A case often encountered is where $\alpha_m = O(\mathrm{e}^{-\rho m})$ for $\rho > 0$, meaning in effect that $\alpha_m = o(m^{-\varphi})$ for all $\varphi < \infty$. Infinite mixing size is also called geometric mixing.

Since mixing is not so much a property of the sequence $\{X_t\}$ as of the sequences of σ-fields generated by $\{X_t\}$, it holds for any random variables measurable on those σ-fields, such as measurable transformations of X_t. More generally, there is the following implication.

15.1 Theorem Let $Y_t = g(X_t, X_{t-1}, \ldots, X_{t-\tau})$ be a measurable function, for finite τ. If X_t is α-mixing (ϕ-mixing) of size $-\varphi$, then Y_t is also.

Proof Let $\mathcal{G}_{-\infty}^t = \sigma(\ldots, Y_{t-1}, Y_t)$ and $\mathcal{G}_{t+m}^\infty = \sigma(Y_{t+m}, Y_{t+m+1}, \ldots)$. Since Y_t is measurable on any σ-field on which each of $X_t, X_{t-1}, \ldots, X_{t-\tau}$ is measurable, $\mathcal{G}_{-\infty}^t \subseteq \mathcal{F}_{-\infty}^t$ and $\mathcal{G}_{t+m}^\infty \subseteq \mathcal{F}_{t+m-\tau}^\infty$. Let $\alpha_{Y,m} = \sup_t \alpha(\mathcal{G}_{-\infty}^t, \mathcal{G}_{t+m}^\infty)$ and it follows that $\alpha_{Y,m} \le \alpha_{m-\tau}$ for $m \ge \tau$. With τ finite, $\alpha_{m-\tau} = O(m^{-\varphi})$ if $\alpha_m = O(m^{-\varphi})$ and the conclusion follows. The same argument follows word for word with 'ϕ' replacing 'α'. ∎

15.2 Mixing Inequalities

Strong and uniform mixing are restrictions on the complete joint distribution of the sequence, but to make practical use of the concepts one must know what they imply about particular measures of dependence. This section establishes a set of fundamental moment inequalities for mixing processes. The main results bound the m-step-ahead predictions, $\mathrm{E}(X_{t+m}|\mathcal{F}_{-\infty}^t)$. Mixing implies that predictors of the future path of a sequence based on the history of events to date, looking further and further forward, will eventually fail to improve on the predictor based solely on the distribution of the sequence as a whole, $\mathrm{E}(X_{t+m})$. The r.v. $\mathrm{E}(X_{t+m}|\mathcal{F}_{-\infty}^t) - \mathrm{E}(X_{t+m})$ is tending to zero as m increases. Following Ibragimov ([103]) consider the convergence of this sequence in L_p-norm.

15.2 Theorem For $r \ge p \ge 1$ and with α_m defined in (15.1),

$$\|\mathrm{E}(X_{t+m}|\mathcal{F}_{-\infty}^t) - \mathrm{E}(X_{t+m})\|_p \le 2(2^{1/p} + 1)\alpha_m^{1/p-1/r}\|X_{t+m}\|_r. \tag{15.7}$$

[1] Defining size as a negative number is conventional but slightly unfortunate. The use of terms such as 'large size' to mean a slow (or rapid?) rate of mixing can obviously lead to confusion and is best avoided.

Proof To simplify notation, substitute X for X_{t+m}, $\mathcal{G}$ for $\mathcal{F}_{-\infty}^{t}$, $\mathcal{H}$ for $\mathcal{F}_{t+m}^{\infty}$, and α for α_m. It will be understood that X is an $\mathcal{H}$-measurable random variable where $\mathcal{G}, \mathcal{H} \subseteq \mathcal{F}$. The proof is in two stages, first to establish the result for $|X| \le M_X < \infty$ a.s. and then to extend it to the case where X is L_r-bounded for finite r. Define the $\mathcal{G}$-measurable r.v.

$$\eta = \operatorname{sgn}\big(\mathrm{E}(X|\mathcal{G}) - \mathrm{E}(X)\big) = \begin{cases} 1, & \mathrm{E}(X|\mathcal{G}) \ge \mathrm{E}(X) \\ -1, & \text{otherwise.} \end{cases} \tag{15.8}$$

Using **10.8** and **10.10**,

$$\begin{aligned} \mathrm{E}|\mathrm{E}(X|\mathcal{G}) - \mathrm{E}(X)| &= \mathrm{E}(\eta(\mathrm{E}(X|\mathcal{G}) - \mathrm{E}(X))) \\ &= \mathrm{Cov}(\eta, X) = |\mathrm{Cov}(\eta, X)|. \end{aligned} \tag{15.9}$$

Next define $\xi = \operatorname{sgn}(\mathrm{E}(Y|\mathcal{H}) - \mathrm{E}(Y))$ which is $\mathcal{H}$-measurable. If Y is any $\mathcal{G}$-measurable r.v., similar arguments give

$$\begin{aligned} |\mathrm{Cov}(X, Y)| &= |\mathrm{E}(X(\mathrm{E}(Y|\mathcal{H}) - \mathrm{E}(Y)))| \\ &\le \mathrm{E}(|X||(\mathrm{E}(Y|\mathcal{H}) - \mathrm{E}(Y))|) \\ &\le M_X \mathrm{E}|\mathrm{E}(Y|\mathcal{H}) - \mathrm{E}(Y)| \\ &\le M_X |\mathrm{Cov}(\xi, Y)| \end{aligned} \tag{15.10}$$

where the first inequality is the modulus inequality. Choosing $Y = \eta$ and putting (15.9) and (15.10) together gives

$$\mathrm{E}|\mathrm{E}(X|\mathcal{G}) - \mathrm{E}(X)| \le M_X |\mathrm{Cov}(\eta, \xi)|. \tag{15.11}$$

ξ and η are simple random variables taking two distinct values each. Defining the sets $A^+ = \{\eta = 1\}$, $A^- = \{\eta = -1\}$, $B^+ = \{\xi = 1\}$, and $B^- = \{\xi = -1\}$,

$$\begin{aligned} |\mathrm{Cov}(\eta, \xi)| &= |\mathrm{E}(\xi\eta) - \mathrm{E}(\xi)\mathrm{E}(\eta)| \\ &= \big|P(A^+ \cap B^+) + P(A^- \cap B^-) - P(A^+ \cap B^-) - (P(A^- \cap B^+) \\ &\qquad - (P(A^+) - P(A^-))(P(B^+) - P(B^-))\big| \\ &\le 4\alpha \end{aligned}$$

and (15.11) gives

$$\mathrm{E}|\mathrm{E}(X|\mathcal{G}) - \mathrm{E}(X)| \le 4M_X\alpha.$$

Since by assumption $|\mathrm{E}(X|\mathcal{G}) - \mathrm{E}(X)| \le 2M_X$ a.s. it further follows that for $p \ge 1$,

$$\mathrm{E}\left|\frac{\mathrm{E}(X|\mathcal{G}) - \mathrm{E}(X)}{2M_X}\right|^p \le \frac{\mathrm{E}|\mathrm{E}(X|\mathcal{G}) - \mathrm{E}(X)|}{2M_X} \le 2\alpha$$

and hence

$$\|\mathrm{E}(X|\mathcal{G}) - \mathrm{E}(X)\|_p \le 2M_X(2\alpha)^{1/p}. \tag{15.12}$$

This completes the first part of the proof. The next step is to let X be L_r-bounded. Choose a finite positive M_X and define $X_1 = 1_{\{|X| \le M_X\}}X$ and $X_2 = X - X_1$. By the Minkowski inequality and (15.12),

$$\begin{aligned}\|\mathrm{E}(X|\mathcal{G}) - \mathrm{E}(X)\|_p &\le \|\mathrm{E}(X_1|\mathcal{G}) - \mathrm{E}(X_1)\|_p + \|\mathrm{E}(X_2|\mathcal{G}) - \mathrm{E}(X_2)\|_p \\ &\le 2M_X(2\alpha)^{1/p} + 2\|X_2\|_p \end{aligned} \tag{15.13}$$

and the problem is to bound the second right-hand-side member. However, note that $|X_2/M_X|^p \le |X/M_X|^r$ when $r \ge p$. Taking expectations and rearranging the inequality gives

$$\|X_2\|_p \le M_X^{1-r/p}\|X\|_r^{r/p} \tag{15.14}$$

and so

$$\|\mathrm{E}(X|\mathcal{G}) - \mathrm{E}(X)\|_p \le 2M_X(2\alpha)^{1/p} + 2M_X^{1-r/p}\|X\|_r^{r/p}. \tag{15.15}$$

Finally, choose $M_X = \|X\|_r\alpha^{-1/r}$. Simplifying then yields

$$\|\mathrm{E}(X|\mathcal{G}) - \mathrm{E}(X)\|_p \le 2(2^{1/p} + 1)\alpha^{1/p-1/r}\|X\|_r$$

which with the original notation restored is (15.7). ∎

There is an easy corollary bounding the autocovariances of the sequence.

15.3 Corollary For $p > 1$ and $r \ge p/(p-1)$

$$|\mathrm{Cov}(X_t, X_{t+m})| \le 2(2^{1-1/p} + 1)\alpha_m^{1-1/p-1/r}\|X_t\|_p\|X_{t+m}\|_r. \tag{15.16}$$

Proof

$$\begin{aligned}|\mathrm{Cov}(X_tX_{t+m})| &= |\mathrm{E}(X_tX_{t+m}) - \mathrm{E}(X_t)\mathrm{E}(X_{t+m})| \\ &= \left|\mathrm{E}\big(X_t\big(\mathrm{E}(X_{t+m}|\mathcal{F}_{-\infty}^t) - \mathrm{E}(X_{t+m})\big)\big)\right| \\ &\le \|X_t\|_p\|\mathrm{E}(X_{t+m}|\mathcal{F}_{-\infty}^t) - \mathrm{E}(X_{t+m})\|_{p/(p-1)} \\ &\le 2(2^{1-1/p} + 1)\|X_t\|_p\|X_{t+m}\|_r\alpha_m^{1-1/p-1/r} \end{aligned} \tag{15.17}$$

where the second equality is by the LIE (**10.8**) and **10.10** noting that X_t is $\mathcal{F}_t$-measurable, the first inequality is the Hölder inequality, and the second inequality is by **15.2**. ∎

The counterpart of **15.2** for the uniform mixing case is given by Serfling ([165] th. 2.2).

15.4 Theorem For $r \geq p \geq 1$,

$$\|\mathrm{E}(X_{t+m}|\mathcal{F}_{-\infty}^{t}) - \mathrm{E}(X_{t+m})\|_p \leq 2\phi_m^{1-1/r}\|X_{t+m}\|_r \tag{15.18}$$

where ϕ_m is defined in (15.2).

Proof The result is trivial for $r = p = 1$ so assume $r > 1$. The strategy is to prove the result initially for a sequence of simple r.v.s. Let $X_{t+m} = \sum_{i=1}^{k} x_i 1_{A_i}$, $A_i \in \mathcal{F}_{t+m}^{\infty} = \sigma(X_{t+m}, X_{t+m+1}, \ldots)$, where the sets $A_1, \ldots, A_k$ partition Ω. For some $\omega \in \Omega$ consider the random element $\mathrm{E}(X_{t+m}|\mathcal{F}_{-\infty}^{t})(\omega)$, although for clarity of notation the dependence on ω is not indicated. Letting $q = r/(r-1)$,

$$\begin{aligned}
&|\mathrm{E}(X_{t+m}|\mathcal{F}_{-\infty}^{t}) - \mathrm{E}(X_{t+m})|^r \\
&\quad = \left|\sum_i x_i\big(P(A_i|\mathcal{F}_{-\infty}^{t}) - P(A_i)\big)\right|^r \\
&\quad \leq \left(\sum_i |x_i|\;|P(A_i|\mathcal{F}_{-\infty}^{t}) - P(A_i)|\right)^r \\
&\quad = \left(\sum_i |x_i|\left|P(A_i|\mathcal{F}_{-\infty}^{t}) - P(A_i)\right|^{1/r}\left|P(A_i|\mathcal{F}_{-\infty}^{t}) - P(A_i)\right|^{1/q}\right)^r \\
&\quad \leq \left(\sum_i |x_i|^r |P(A_i|\mathcal{F}_{-\infty}^{t}) - P(A_i)|\right)\left(\sum_i |P(A_i|\mathcal{F}_{-\infty}^{t}) - P(A_i)|\right)^{r/q} \\
&\quad \leq \big(\mathrm{E}(|X_{t+m}|^r|\mathcal{F}_{-\infty}^{t}) + \mathrm{E}|X_{t+m}|^r\big)\left(\sum_i |P(A_i|\mathcal{F}_{-\infty}^{t}) - P(A_i)|\right)^{r/q}.
\end{aligned} \tag{15.19}$$

The second inequality here is by **2.22**. Since the A_i are disjoint, $P(A_i \cup A_{i'}|\mathcal{F}_{-\infty}^{t}) = P(A_i|\mathcal{F}_{-\infty}^{t}) + P(A_{i'}|\mathcal{F}_{-\infty}^{t})$ a.s. and $P(A_i \cup A_{i'}) = P(A_i) + P(A_{i'})$ for $i \neq i'$. Let A^+ denote the union of all those A_i for which $P(A_i|\mathcal{F}_{-\infty}^{t}) - P(A_i) \geq 0$ and A^- the complement of A^+ on Ω. Then,

$$\sum_i |P(A_i|\mathcal{F}_{-\infty}^{t}) - P(A_i)| = |P(A^+|\mathcal{F}_{-\infty}^{t}) - P(A^+)| + |P(A^-|\mathcal{F}_{-\infty}^{t}) - P(A^-)| \tag{15.20}$$

and by **14.17** the inequalities

$$|P(A^+|\mathcal{F}_{-\infty}^t) - P(A^+)| \le \phi_m$$
$$|P(A^-|\mathcal{F}_{-\infty}^t) - P(A^-)| \le \phi_m$$

hold with probability 1. Substituting into (15.19) gives

$$|\mathrm{E}(X_{t+m}|\mathcal{F}_{-\infty}^t) - \mathrm{E}(X_{t+m})|^r \le \big(\mathrm{E}(|X_{t+m}|^r|\mathcal{F}_{-\infty}^t) + \mathrm{E}|X_{t+m}|^r\big)(2\phi_m)^{r/q} \text{ a.s.} \quad (15.21)$$

Taking expectations and using the LIE then gives

$$\mathrm{E}\big|\mathrm{E}(X_{t+m}|\mathcal{F}_{-\infty}^t) - \mathrm{E}(X_{t+m})\big|^r \le 2\mathrm{E}|X_{t+m}|^r(2\phi_m)^{r/q} \quad (15.22)$$

and raising both sides to the power $1/r$ gives

$$||\mathrm{E}(X_{t+m}|\mathcal{F}_{-\infty}^t) - \mathrm{E}(X_{t+m})||_r \le 2||X_{t+m}||_r\phi_m^{1-1/r}. \quad (15.23)$$

Inequality (15.18) follows by Liapunov's inequality.

The result extends from simple to general r.v.s using the construction of **3.35**. Given a monotone sequence of simple functions, say $\{X_{t+m,(k)}, k \in \mathbb{N}\}$ converging a.s. to an L_r-bounded limit X_{t+m}, the associated sequence of conditional expectations in (15.22) converges by **10.16** and the expected values on each side of (15.22) converge by **4.16** with dominating function $|X_{t+m}|^r$. ∎

The counterpart of **15.3** is obtained similarly.

15.5 Corollary For $r \ge 1$,

$$|\mathrm{Cov}(X_{t+m}, X_t)| \le 2\phi_m^{1/r}||X_t||_r||X_{t+m}||_{r/(r-1)} \quad (15.24)$$

where, if $r = 1$, replace $||X_{t+m}||_{r/(r-1)}$ by $||X_{t+m}||_\infty = \mathrm{ess\,sup}\, X_{t+m}$.

Proof

$$\begin{aligned}|\mathrm{Cov}(X_{t+m}X_t)| &\le ||X_t||_r||\mathrm{E}(X_{t+m}|\mathcal{F}_{-\infty}^t) - \mathrm{E}(X_{t+m})||_{r/(r-1)} \\ &\le 2\phi_m^{1/r}||X_t||_r||X_{t+m}||_{r/(r-1)} \quad (15.25)\end{aligned}$$

where the first inequality corresponds to the first of (15.17) and the second one is by (15.23). ∎

These results tell us a good deal about the behaviour of mixing sequences. A fundamental property is *mean reversion*. The mean deviation sequence $\{X_t - \mathrm{E}(X_t)\}$ must change sign frequently when the rate of mixing is high. If the sequence exhibits persistent behaviour with $X_t - \mathrm{E}(X_t)$ tending to have the same sign for a large number of successive periods, then $|\mathrm{E}(X_{t+m}|\mathcal{F}_{-\infty}^t) - \mathrm{E}(X_{t+m})|$ would likewise tend to be large for large m. If this quantity is small the sign of the mean deviation m periods hence is unpredictable, indicating that it changes frequently.

But while mixing implies mean reversion, mean reversions need not imply mixing. Theorems **15.2** and **15.4** isolate the properties of greatest importance, but not the only ones. A sequence having the property $\|\mathrm{Var}(X_{t+m}|\mathcal{F}_{-\infty}^t) - \mathrm{Var}(X_{t+m})\|_p > 0$ is called *conditionally heteroscedastic*. Mixing also requires this sequence of norms to converge as $m \to \infty$ and similarly for other integrable functions of X_{t+m}.

15.3 Mixing in Linear Processes

An issue of obvious interest is the application of the mixing concept to the linear processes reviewed in §13.3. Whether sequences of the form

$$X_t = \sum_{j=0}^{q} \theta_j Z_{t-j},\ 0 \le q \le \infty \tag{15.26}$$

are mixing depends on the properties of the i.i.d. innovations $\{Z_t\}$ and the sequence $\{\theta_j\}$. Several authors have investigated this question, see *inter alia* Ibragimov and Linnik ([105]), Chanda ([31]), Gorodetskii ([84]), Withers ([191]), Pham and Tran ([142]), and Athreya and Pantula ([12], [13]).

Mixing is an asymptotic property and when $q < \infty$ the sequence is mixing infinitely fast. This case is called *q-dependence*. The difficulties only arise in the cases with $q = \infty$. If $\{Z_t\}$ is i.i.d. with mean 0 and variance σ^2, X_t is stationary and has spectral density function

$$f(\lambda) = \frac{\sigma^2}{2\pi}\left|\sum_{j=0}^{\infty} \theta_j e^{i\lambda j}\right|^2. \tag{15.27}$$

The theorem of Ibragimov and Linnik cited in §15.1 yields the condition $\sum_{j=0}^{\infty} |\theta_j| < \infty$ as sufficient for strong mixing in the Gaussian case. However, allowing more general distributions for the innovations yields surprising results. Contrary to what might be supposed, having the θ_j tend to zero even at an exponential rate is not sufficient by itself for strong mixing. Here is a simple counterexample due to Andrews [4]. Recall that the first-order autoregressive process $X_t = \rho X_{t-1} + Z_t, |\rho| < 1$ has the MA(∞) form with $\theta_j = \rho^j, j = 0, 1, 2, \ldots$.

15.6 Example Let $\{Z_t\}_0^\infty$ be an independent sequence of Bernoulli r.v.s with $P(Z_t = 1) = P(Z_t = 0) = \frac{1}{2}$. Let $X_0 = Z_0$ and

$$X_t = \tfrac{1}{2}X_{t-1} + Z_t = \sum_{j=0}^{t} 2^{-j} Z_{t-j},\ t = 1, 2, 3, \ldots. \tag{15.28}$$

The term

$$\sum_{j=0}^{t} 2^{-j} Z_{t-j} = 2^{-t} \sum_{k=0}^{t} 2^k Z_k \tag{15.29}$$

belongs for each t to the set of dyadic rationals

$$W_t = \{k/2^t, k = 0, 1, 2, \ldots, 2^{t+1} - 1\}.$$

Each element of W_t corresponds to one of the 2^{t+1} possible drawings $\{Z_0, \ldots, Z_t\}$ and arises in every case with probability 2^{-t-1}. Iff $Z_0 = 0$,

$$X_t \in B_t = \{k/2^t, k = 0, 2, 4, \ldots, 2(2^t - 1)\},$$

whereas iff $Z_0 = 1$,

$$X_t \in W_t - B_t = \{k/2^t, k = 1, 3, 5, \ldots 2^{t+1} - 1\}.$$

It follows that $\{X_0 = 1\} \cap \{X_t \in B_t\} = \emptyset$ and therefore $P(X_t \in B_t) = P(X_0 = 0) = \frac{1}{2}$, for every finite t. Hence for every finite m,

$$\alpha_m \geq |P(\{X_0 = 1\} \cap \{X_m \in B_m\}) - P(X_0 = 1)P(X_m \in B_m)| = \tfrac{1}{4} \tag{15.30}$$

which contradicts $\alpha_m \to 0$. □

Since the process starts at $t = 0$ in this case it is not stationary, but the example is easily generalized to a wider class of processes, as in the following result from [4].

15.7 Theorem Let $\{Z_t\}_{-\infty}^\infty$ be an independent sequence of Bernoulli r.v.s taking values 1 and 0 with fixed probabilities p and $1 - p$. If $X_t = \rho X_{t-1} + Z_t$ for $\rho \in (0, \frac{1}{2}]$, $\{X_t\}_{-\infty}^\infty$ is not strong mixing. □

Note, the condition on ρ is purely to expedite the argument. The theorem surely holds for other values of ρ, although this cannot be proved by the present approach.

Proof of 15.7 Write $X_{t+s} = \rho^s X_t + X_{t,s}$ where

$$X_{t,s} = \sum_{j=0}^{s-1} \rho^j Z_{t+s-j}. \tag{15.31}$$

The support of $X_{t,s}$ is finite for finite s, having at most 2^s distinct members. Call this set W_s, so that $W_1 = (0,1)$, $W_2 = (0,1,\rho,1+\rho)$, and so on. In general, W_{s+1} is obtained from W_s by adding ρ^s to each of its elements and forming the union of these elements with those of W_s; formally,

$$W_{s+1} = W_s \cup \{w + \rho^s : w \in W_s\}, s = 2,3,\dots. \tag{15.32}$$

For given s denote the distinct elements of W_s by w_j, ordered by magnitude with $w_1 < \dots < w_J$, for $J \le 2^s$.

Now suppose that $X_t \in (0,\rho)$ so that $\rho^s X_t \in (0, \rho^{s+1})$. This means that X_{t+s} assumes a value lying strictly between w_j and $w_j + \rho^{s+1}$, for some j. Defining events $A = \{X_t \in (0,\rho)\}$ and $B_s = \{X_{t+s} \in \bigcup_{j=1}^J (w_j, w_j + \rho^{s+1})\}$, $P(B_s|A) = 1$ for any s however large. To see that $P(A) > 0$, consider the case $Z_t = Z_{t-1} = Z_{t-2} = 0$ and $Z_{t-3} = 1$ and note that

$$\sum_{j=3}^{\infty} \rho^j Z_{t-j} \le \sum_{j=3}^{\infty} \rho^j = \frac{\rho^3}{1-\rho} < \rho \tag{15.33}$$

for $\rho \in (0, \frac{1}{2}]$. So, unless $P(B_s) = 1$, strong mixing is contradicted.

The proof is completed by showing that the set $D = \{X_t \in [\rho, 1]\}$ has positive probability and is disjoint with B_s. D occurs when $Z_t = 0$ and $Z_{t-1} = 1$, since then, for $\rho \in (0, \frac{1}{2}]$,

$$\rho \le \sum_{j=1}^{\infty} \rho^j Z_{t-j} \le \sum_{j=1}^{\infty} \rho^j = \frac{\rho}{1-\rho} \le 1 \tag{15.34}$$

and hence $P(D) > 0$. Suppose that

$$\min_{j\ge 1}\{w_{j+1} - w_j\} \ge \rho^{s-1}. \tag{15.35}$$

Then, if D occurs,

$$w_j + \rho^{s+1} \le w_j + \rho^s X_t < w_j + \rho^{s-1} \le w_{j+1}. \tag{15.36}$$

Hence, $X_{t+s} = w_j + \rho^s X_t \notin \bigcup_{j=1}^J (w_j, w_j + \rho^{s+1})$, or in other words, $B_s \cap D = \varnothing$.

The assertion in (15.35) is certainly true when $s = 1$, so consider the following inductive argument. Suppose the distance between two points in W_s is at least ρ^{s-1}. Then by (15.32), the smallest distance between two points of W_{s+1} cannot be less than the smaller of ρ^s and $\rho^{s-1} - \rho^s$. But when $\rho \in (0, \frac{1}{2})$, $\rho^s \leq \frac{1}{2}\rho^{s-1}$, which implies $\rho^{s-1} - \rho^s \geq \rho^s$. It follows that (15.35) holds for every s. ∎

These results may appear surprising when one thinks of the rate at which ρ^s approaches 0 with s; but if so, this is because one is unconsciously thinking about the problem of predicting gross features of the distribution of X_{t+s} from time t, things like $P(X_{t+s} \leq x|A)$ for example. The notable feature of the sets B_s is their irrelevance to such concerns, at least for large s. What has been shown is that from a practical viewpoint the mixing concept has some undesirable features. The requirement of a decline of dependence is imposed over *all* events, whereas in practice it may be perfectly compatible with the goals of the analysis to tolerate certain uninteresting events such as the B_s defined above remaining dependent on the initial conditions.

The next section derives some sufficient conditions for strong mixing and it turns out that certain smoothness conditions on the marginal distributions of the increments will be enough to rule out this kind of counterexample. But now consider uniform mixing. The following example is similar to Athreya and Pantula [12].

15.8 Example Consider an AR(1) process with i.i.d. increments,

$$X_t = \rho X_{t-1} + Z_t,\ 0 < \rho < 1 \tag{15.37}$$

in which the marginal distribution of Z_t has unbounded support. $\{X_t\}$ is not uniform mixing. Write $Y_m = \sum_{j=0}^{m-1} \rho^j Z_{m-j}$ so that $X_m = \rho^m X_0 + Y_m$. For $\delta > 0$ choose a constant $M > 0$ to satisfy

$$P(Y_m \leq -M) < \delta. \tag{15.38}$$

Then consider the events

$$A = \{\rho^m X_0 \geq L + M\} \in \mathcal{F}_{-\infty}^0$$
$$B = \{X_m \leq L\} \in \mathcal{F}_m^{+\infty}$$

where $L > 0$ is chosen large enough that $P(B) \geq 1 - \delta$. Let $p_K = P(Z_0 < K)$, for any constant K. Since Z_0 has unbounded support, either $p_K < 1$ for every $K > 0$ or, at worst, this holds after substituting $\{-Z_t\}$ for $\{Z_t\}$ and hence $\{-X_t\}$ for $\{X_t\}$. $p_K < 1$ for all K implies according to (15.37) that $P(X_0 < 0) < 1$ and hence by stationarity that $P(X_{-1} < 0) < 1$ also. Since

$$\{X_0 < K\} \subseteq \{Z_0 < K\} \cup (\{Z_0 \geq K\} \cap \{X_{-1} < 0\}),$$

independence of the $\{Z_t\}$ implies that

$$P(X_0 < K) \leq p_K + (1 - p_K)P(X_{-1} < 0) < 1. \tag{15.39}$$

Considering the case $K > \rho^{-m}(L + M)$ it must follow from (15.39) that $P(A) > 0$. On the other hand,

$$P(B|A) = P\big(\rho^m X_0 + Y_m \leq L \,|\, \rho^m X_0 \geq L + M\big).$$

If A occurs then B cannot also occur unless $Y_m \leq -M$. Hence by (15.38), $P(B|A) < \delta$ and so $\phi_m \geq |P(B|A) - P(B)| > 1 - 2\delta$. Since δ is arbitrary this means $\phi_m = 1$ for every m. □

Processes with Gaussian increments fall into the category covered by this example and if ϕ-mixing fails in the first-order AR case it is pretty clear that counterexamples exist for more general MA(∞) cases too. The conditions for uniform mixing in linear processes are evidently extremely restrictive, perhaps too restrictive for this mixing condition to be very useful. In the applications to be studied in later chapters most of the results are found to hold in some form for strong mixing processes. However, the ability to assert uniform mixing usually allows a relaxation of conditions elsewhere in the problem, so it is still desirable to develop the parallel results for the uniform case.

The strong restrictions needed to ensure processes are mixing that these examples point out (to be explored further in the next section) threaten to limit the usefulness of the mixing concept. However, technical infringements like the ones demonstrated are often innocuous in practice. Only certain aspects of mixing, encapsulated in the concept of a *mixingale*, are required for many important limit results to hold. These are shared with the so-called near-epoch dependent functions of a mixing process which include cases like **15.7**. These dependence concepts are to be examined in detail in Chapters 17 and 18.

15.4 Sufficient Conditions for Strong and Uniform Mixing

The problems in the counterexamples above are with the form of the marginal shock distributions—discrete or unbounded, as the case may be. For strong mixing, a degree of smoothness of the distributions appears necessary in addition to summability conditions on the coefficients of linear processes. Several sufficient conditions have been derived, both for general MA(∞) processes and for autoregressive and ARMA processes. The sufficiency result for strong mixing proved below is based on the theorems of Chanda ([31]) and Gorodetskii ([84]). These conditions are not the weakest possible in all circumstances, but they have the virtues of generality and comparative ease of verification.

15.9 Theorem Let $X_t = \sum_{j=0}^{\infty} \theta_j Z_{t-j}$ define a random sequence $\{X_t\}_{-\infty}^{\infty}$, where, for either $0 < r \leq 2$ or r an even positive integer,

(a) Z_t is uniformly L_r-bounded, independent and continuously distributed with p.d.f. f_{Z_t} having the property

$$\sup_t \int_{-\infty}^{+\infty} \left| f_{Z_t}(z+a) - f_{Z_t}(z) \right| \mathrm{d}z \leq M|a|, M < \infty \tag{15.40}$$

whenever $|a| \leq \delta$, for some $\delta > 0$

(b) $\sum_{t=0}^{\infty} G_t(r)^{1/(1+r)} < \infty$ where

$$G_t(r) = \begin{cases} 2\sum_{j=t}^{\infty} |\theta_j|^r, & r \leq 2 \\ 2^{r-1}\left(\sum_{j=t}^{\infty} \theta_j^2\right)^{r/2}, & r \geq 2 \end{cases} \tag{15.41}$$

(c) $\sum_{j=0}^{\infty} |\tau_j| < \infty$ where $\tau(x) = 1/\theta(x)$.

Then $\{X_t\}$ is strong mixing with $\alpha_m = O\left(\sum_{t=m+1}^{\infty} G_t(r)^{1/(1+r)}\right)$. □

Before proceeding to the proof, consider the implications of these three conditions in a bit more detail. Condition **15.9**(a) may be relaxed somewhat as shown below, but begin with this case for simplicity. The following lemma extends the condition to the joint distributions under independence.

15.10 Lemma Inequality (15.40) implies that for $|a_t| \leq \delta$, $t = 1, \ldots, k$, $k < \infty$,

$$\int_{\mathbb{R}^k} \left| \prod_{t=1}^{k} f_{Z_t}(z_t + a_t) - \prod_{t=1}^{k} f_{Z_t}(z_t) \right| \mathrm{d}z_1 \cdots \mathrm{d}z_k \leq M \sum_{t=1}^{k} |a_t|. \tag{15.42}$$

Proof Using Fubini's theorem,

$$\begin{aligned}
&\int_{\mathbb{R}^k}\left|\prod_{t=1}^{k} f_{Z_t}(z_t+a_t)-\prod_{t=1}^{k} f_{Z_t}(z_t)\right|\mathrm{d}z_1\cdots\mathrm{d}z_k \\
&\qquad\le \int_{\mathbb{R}^k}\left|\big(f_{Z_1}(z_1+a_1)-f_{z_1}(z_1)\big)\prod_{t=2}^{k} f_{Z_t}(z_t+a_t)\right|\mathrm{d}z_1\cdots\mathrm{d}z_k \\
&\qquad\quad + \int_{\mathbb{R}^k}\left|f_{Z_1}(z_1)\left(\prod_{t=2}^{k} f_{Z_t}(z_t+a_t)-\prod_{t=2}^{k} f_{Z_t}(z_t)\right)\right|\mathrm{d}z_1\cdots\mathrm{d}z_k \\
&\qquad\le M|a_1| + \int_{\mathbb{R}^{k-1}}\left|\prod_{t=2}^{k} f_{Z_t}(z_t+a_t)-\prod_{t=2}^{k} f_{Z_t}(z_t)\right|\mathrm{d}z_2\cdots\mathrm{d}z_k.
\end{aligned}$$

The lemma follows on applying the same inequality to the second term on the right, iteratively for $t = 2,\ldots,k$. ∎

Condition **15.9** (b) is satisfied when $|\theta_j| \ll j^{-\mu}$ for $\mu > 1 + 2/r$ when $r \le 2$ and $\mu > 3/2 + 1/r$ when $r \ge 2$. The double definition of $G_t(r)$ is motivated by the fact that for cases with $r \le 2$ the von Bahr–Esséen inequality (**11.21**) is used to bound a certain sequence in the proof, whereas the case $r > 2$ relies on Lemma **15.11** below. Since the latter result requires r to be an even integer, the conditions in the theorem are to be applied in practice by taking r as the nearest even integer below the highest existing absolute moment. Gorodetskii in [84] achieves a further weakening of these summability conditions for $r > 2$ by the use of an inequality due to Nagaev and Fuk [134]. This extension is foregone both because proof of the Nagaev–Fuk inequalities represents a complicated detour and because the present version of the theorem permits a generalization (Corollary **15.13**) which would otherwise be awkward to implement.

Define $W_t = \sum_{j=0}^{t-1}\theta_j Z_{t-j}$ and $V_t = \sum_{j=t}^{\infty}\theta_j Z_{t-j}$, so that $X_t = W_t + V_t$ and W_t and V_t are independent. V_t is the $\mathcal{F}_{-\infty}^{0}$-measurable 'tail' of X_t whose contribution to the sum should become negligible as $t \to \infty$.

15.11 Lemma If the sequence $\{Z_s\}$ is independent with zero mean then

$$\mathrm{E}(V_t^{2m}) \le 2^{2m-1}\left(\sum_{j=t}^{\infty}\theta_j^2\right)^m \sup_{s\le 0}\mathrm{E}(Z_s^{2m}) \tag{15.43}$$

for each positive integer m such that $\sup_{s\le 0}\mathrm{E}(Z_s^{2m}) < \infty$.

Proof First consider the case where the r.v.s Z_{t-j} are symmetrically distributed, meaning that $-Z_{t-j}$ and Z_{t-j} have the same distributions. In this case all existing odd-order integer moments about 0 are zero and for any $k > 0$,

$$\begin{aligned}
\mathrm{E}\Bigg(\sum_{j=t}^{t+k}\theta_j Z_{t-j}\Bigg)^{2m} &= \sum_{j_1=t}^{t+k}\cdots\sum_{j_{2m}=t}^{t+k}\theta_{j_1}\cdots\theta_{j_{2m}}\mathrm{E}(Z_{t-j_1}\cdots Z_{t-j_{2m}}) \\
&= \sum_{j_1=t}^{t+k}\cdots\sum_{j_m=t}^{t+k}\theta_{j_1}^2\cdots\theta_{j_m}^2\mathrm{E}(Z_{t-j_1}^2\cdots Z_{t-j_m}^2) \\
&\le \Bigg(\sum_{j=0}^{t+k}\theta_j^2\Bigg)^m \sup_{s\le 0}\mathrm{E}(Z_s^{2m}). \qquad (15.44)
\end{aligned}$$

The second equality holds since $\mathrm{E}(Z_{t-j_1}\cdots Z_{t-j_{2m}})$ vanishes unless the factors form matching pairs and the inequality follows since, for any r.v. Y possessing the requisite moments, $\mathrm{E}(Y^{j+k}) \ge \mathrm{E}(Y^j)\mathrm{E}(Y^k)$ (i.e. $\mathrm{Cov}(Y^j, Y^k) \ge 0$) for $j,k > 0$. The result for symmetrically distributed Z_s follows on letting $k \to \infty$.

For general Z_s, let Z_s' be distributed identically to and independent of Z_s for each $s \le 0$. Then $V_t' = \sum_{j=t}^{\infty}\theta_j Z_{t-j}'$ is independent of V_t and $V_t - V_t'$ has symmetrically distributed independent increments $Z_{t-j} - Z_{t-j}'$. Hence

$$\begin{aligned}
\mathrm{E}(V_t^{2m}) \le \mathrm{E}(V_t - V_t')^{2m} &\le \Bigg(\sum_{j=t}^{\infty}\theta_j^2\Bigg)^m \sup_j \mathrm{E}(Z_{t-j} - Z_{t-j}')^{2m} \\
&\le 2^{2m-1}\Bigg(\sum_{j=0}^{\infty}\theta_j^2\Bigg)^m \sup_j \mathrm{E}(Z_{t-j}^{2m}) \qquad (15.45)
\end{aligned}$$

where the first inequality is by **10.20**, the second by (15.43) for the symmetrically distributed case, and the third is the c_r inequality. ∎

Lastly, consider condition **15.9**(c). See Theorem **13.13** and the associated discussion in §13.3 on the question of inverting lag polynomials. Here a nominally stronger condition than the mere summability of the coefficients is imposed, that of absolute summability, although in the context of time series models the possibility of conditional but not absolute summability would be unusual. For a finite number of terms, the linear transformation is conveniently expressed using matrix notation. Let

$$A_n = \begin{bmatrix} 1 & 0 & \cdots & 0 & 0 \\ \theta_1 & 1 & 0 & \cdots & 0 \\ \vdots & \theta_1 & 1 & \ddots & \vdots \\ \theta_{n-2} & \vdots & \ddots & \ddots & 0 \\ \theta_{n-1} & \theta_{n-2} & \cdots & \theta_1 & 1 \end{bmatrix} (n \times n) \tag{15.46}$$

so that the equations $x_t = \sum_{j=0}^{t-1} \theta_j z_{t-j}$, $t = 1,\dots,n$ can be written $\boldsymbol{x} = \boldsymbol{A}_n \boldsymbol{z}$ where $\boldsymbol{x} = (x_1,\dots,x_n)'$ and $\boldsymbol{z} = (z_1,\dots,z_n)'$. $\boldsymbol{A}_n^{-1}$ is also lower triangular with elements τ_j replacing θ_j for $j = 0,\dots,n-1$. If $\boldsymbol{\nu} = (\nu_1,\dots,\nu_n)'$ the vector $\hat{\boldsymbol{\nu}} = \boldsymbol{A}_n^{-1}\boldsymbol{\nu}$ has elements $\sum_{j=0}^{t-1} \tau_j \nu_{t-j}$ for $t = 1,\dots,n$. These operations can be taken to the limit as $n \to \infty$, subject to **15.9**(c).

Proof of 15.9 Without loss of generality, the object is to show that the σ-fields $\mathcal{F}_{-\infty}^{0} = \sigma(\dots, X_{-1}, X_0)$ and $\mathcal{F}_{m+1}^{\infty} = \sigma(X_{m+1}, X_{m+2}, \dots)$ are becoming independent as $m \to \infty$. The result does not depend on the choice of origin for the indices. This is shown for a sequence $\{X_t\}_{t=1-p}^{m+k}$ for finite p and k and since k and p are arbitrary, it then follows by the consistency theorem (**12.4**) that there exists a sequence $\{X_t\}_{-\infty}^{\infty}$ whose finite-dimensional distributions possess the property for every k and p. This sequence is strong mixing on the definition.

Define a $p+m+k$-vector $\boldsymbol{X} = (\boldsymbol{X}_0', \boldsymbol{X}_1', \boldsymbol{X}_2')'$ where $\boldsymbol{X}_0 = (X_{1-p},\dots,X_0)'$ $(p \times 1)$, $\boldsymbol{X}_1 = (X_1,\dots,X_m)'$ $(m \times 1)$ and $\boldsymbol{X}_2 = (X_{m+1},\dots,X_{m+k})'$ $(k \times 1)$. Also define vectors $\boldsymbol{W} = (\boldsymbol{W}_1', \boldsymbol{W}_2')'$ and $\boldsymbol{V} = (\boldsymbol{V}_1', \boldsymbol{V}_2')'$ whose elements W_t and V_t are defined in the paragraph preceding **15.11**, such that $\boldsymbol{X}_1 = \boldsymbol{W}_1 + \boldsymbol{V}_1$ and $\boldsymbol{X}_2 = \boldsymbol{W}_2 + \boldsymbol{V}_2$. The vectors $\boldsymbol{X}_0$ and $\boldsymbol{V}$ are independent of $\boldsymbol{W}$. Using the notation $\mathcal{F}_s^t = \sigma(X_s,\dots,X_t)$, define the sets

$$\begin{aligned} G &= \{\omega : \boldsymbol{X}_0(\omega) \in C\} \in \mathcal{F}_{1-p}^{0} \text{ for some } C \in \mathcal{B}^p \\ H &= \{\omega : \boldsymbol{X}_2(\omega) \in D\} \in \mathcal{F}_{m+1}^{m+k} \text{ for some } D \in \mathcal{B}^k \\ E &= \{\omega : \boldsymbol{V}_2(\omega) \in B\} \in \mathcal{F}_{-\infty}^{0} \end{aligned}$$

where $B = \{\boldsymbol{v}_2 : |\boldsymbol{v}_2| \le \boldsymbol{\eta}\} \in \mathcal{B}^k$, $|\boldsymbol{v}_2|$ denotes the vector whose elements are the absolute values of $\boldsymbol{v}_2$, and $\boldsymbol{\eta} = (\eta_{m+1},\dots,\eta_{m+k})'$ is a vector of positive constants to be chosen. Also define

$$D - \boldsymbol{v}_2 = \{\boldsymbol{w}_2 : \boldsymbol{w}_2 + \boldsymbol{v}_2 \in D\} \in \mathcal{B}^k.$$

H may be thought of as the random event that has occurred when first $\boldsymbol{V}_2 = \boldsymbol{v}_2$ is realized and then $\boldsymbol{W}_2 \in D - \boldsymbol{v}_2$. By independence, the joint c.d.f. of the variables $(\boldsymbol{W}_2, \boldsymbol{V}_2, \boldsymbol{X}_0)$ factorizes as $F = F_{W_2} F_{V_2,X_0}$ (say) and

$$P(H) = P(\boldsymbol{X}_2 \in D) = \int_{\mathbb{R}^p}\int_{\mathbb{R}^k}\int_{D-v2} \mathrm{d}F(\boldsymbol{w}_2, \boldsymbol{v}_2, \boldsymbol{x}_0)$$
$$= \int_{\mathbb{R}^p}\int_{\mathbb{R}^k} \chi(\boldsymbol{v}_2)\mathrm{d}F_{V_2,X_0}(\boldsymbol{v}_2, \boldsymbol{x}_0) \qquad (15.47)$$

where

$$\chi(\boldsymbol{v}_2) = P(\boldsymbol{W}_2 \in D - \boldsymbol{v}_2) = \int_{D-\boldsymbol{v}_2} \mathrm{d}F_{W_2}(\boldsymbol{w}_2). \qquad (15.48)$$

These definitions set the scene for the main business of the proof which is to show that events G and H are tending to independence as m becomes large. Given $\mathcal{F}/\mathcal{B}^{p+m+k}$-measurability of $\boldsymbol{X}$, this is sufficient for the result since C and D are arbitrary. By the same reasoning that gave (15.47),

$$P(G \cap H \cap E) = \int_C\int_B \chi(\boldsymbol{v}_2)\mathrm{d}F_{V_2,X_0}(\boldsymbol{v}_2, \boldsymbol{x}_0). \qquad (15.49)$$

Define $\chi^* = \sup_{\boldsymbol{v}_2 \in B}\chi(\boldsymbol{v}_2)$ and $\chi_* = \inf_{\boldsymbol{v}_2 \in B}\chi(\boldsymbol{v}_2)$ and (15.49) implies

$$\chi_* P(G \cap E) \le P(G \cap H \cap E) \le \chi^* P(G \cap E). \qquad (15.50)$$

Hence,

$$P(G \cap H) = P(G \cap H \cap E) + P(G \cap H \cap E^c)$$
$$\le \chi^* P(G) + P(E^c) \qquad (15.51)$$

and similarly, since $\chi_* \le 1$,

$$P(G \cap H) \ge \chi_* P(G \cap E) + P(G \cap H \cap E^c)$$
$$= \chi_* P(G) - \chi_* P(G \cap E^c) + P(G \cap H \cap E^c)$$
$$\ge \chi_* P(G) - P(E^c). \qquad (15.52)$$

Choosing $G = \Omega$ (i.e., $C = \mathbb{R}^p$) in (15.51) and (15.52) gives in particular

$$\chi_* - P(E^c) \le P(H) \le \chi^* + P(E^c) \qquad (15.53)$$

and combining all these inequalities yields

$$|P(G \cap H) - P(G)P(H)| \le \chi^* - \chi_* + 2P(E^c). \qquad (15.54)$$

Next write $\boldsymbol{W} = \boldsymbol{A}_{m+k}\boldsymbol{Z}$ where $\boldsymbol{Z} = (Z_1,\dots,Z_{m+k})'$ and $\boldsymbol{A}_{m+k}$ is defined by (15.46). Since $|\boldsymbol{A}_{m+k}| = 1$ and the $\{Z_1,\dots,Z_{m+k}\}$ are independent, the change of variable formula from **8.23** yields the result that $\boldsymbol{W}$ is continuously distributed with

$$f_W(\boldsymbol{w}) = f_Z(\boldsymbol{z}) = \prod_{t=1}^{m+k} f_{Z_t}(z_t). \tag{15.55}$$

Define $B' = \{\boldsymbol{v} : \boldsymbol{v}_1 = \boldsymbol{0}, \boldsymbol{v}_2 \in B\} \in \mathcal{B}^{m+k}$. Then the following relations hold:

$$\begin{aligned}\chi^* - \chi_* &\le 2 \sup_{\boldsymbol{v}_2 \in B} |\chi(\boldsymbol{v}_2) - \chi(\boldsymbol{0})| \\ &\le 2 \sup_{\boldsymbol{v}_2 \in B} \int_D \big|f_{W_2}(\boldsymbol{w}_2 + \boldsymbol{v}_2) - f_{W_2}(\boldsymbol{w}_2)\big| \, d\boldsymbol{w}_2 \\ &\le 2 \sup_{\boldsymbol{v} \in B'} \int_{\mathbb{R}^{m+k}} \big|f_W(\boldsymbol{w} + \boldsymbol{v}) - f_W(\boldsymbol{w})\big| \, d\boldsymbol{w} \\ &= 2 \sup_{\boldsymbol{v} \in B'} \int_{\mathbb{R}^{m+k}} \left| \prod_{t=1}^{m+k} f_{Z_t}(z_t + \hat{v}_t) - \prod_{t=1}^{m+k} f_{Z_t}(z_t) \right| d\boldsymbol{z} \\ &\le 2M \sup_{\boldsymbol{v} \in B'} \sum_{t=m+1}^{m+k} |\hat{v}_t|. \end{aligned} \tag{15.56}$$

Here the equality substitutes $\hat{\boldsymbol{v}} = \boldsymbol{A}_{m+k}^{-1}\boldsymbol{v}$ and uses the fact that $\hat{\boldsymbol{v}}_1 = \boldsymbol{0}$ if $\boldsymbol{v}_1 = \boldsymbol{0}$ by lower triangularity of $\boldsymbol{A}_{m+k}$, while the final inequality uses Lemma **15.10** and assumption (a). According to the definitions of B' and B,

$$\begin{aligned}\sup_{\boldsymbol{v} \in B'} \sum_{t=m+1}^{m+k} |\hat{v}_t| = \sup_{\boldsymbol{v} \in B'} \sum_{t=m+1}^{m+k} \left| \sum_{j=0}^{t-m-1} \tau_j v_{t-j} \right| &\le \sum_{t=m+1}^{m+k} \sum_{j=0}^{t-m-1} |\tau_j| \eta_{t-j} \\ &\le \left(\sum_{j=0}^{\infty} |\tau_j| \right) \sum_{t=m+1}^{m+k} \eta_t. \end{aligned} \tag{15.57}$$

By condition **15.9**(c) it is possible to choose $\boldsymbol{\eta}$ with elements small enough that $|\hat{v}_t| \le \delta/2$ for each t, where δ is defined in condition **15.9**(a).

For the final step, choose r to be the largest order of absolute moment if this does not exceed 2, and the largest even integer moment otherwise. Then

$$
\begin{aligned}
P(E^c) &= P(|\boldsymbol{V}_2| > \boldsymbol{\eta}) \\
&\leq P\Bigg(\bigcup_{t=m+1}^{m+k} \{|V_t| > \eta_t\}\Bigg) \\
&\leq \sum_{t=m+1}^{m+k} P(|V_t| > \eta_t) \leq \sum_{t=m+1}^{m+k} \mathrm{E}|V_t|^r \eta_t^{-r}
\end{aligned}
\tag{15.58}
$$

by the Markov inequality. Also, applying **11.21** for $r \leq 2$ (see (11.85) for the required extension) and Lemma **15.11** for $r > 2$,

$$
\mathrm{E}|V_t|^r \leq \sup_s \mathrm{E}|Z_s|^r G_t(r)
\tag{15.59}
$$

where $G_t(r)$ is given by (15.41). Substituting inequalities (15.56), (15.57), (15.58), and (15.59) into (15.54) yields the order-of-magnitude relation

$$
|P(G \cap H) - P(G)P(H)| \ll \sum_{t=m+1}^{m+k} \left(\eta_t + G_t(r)\eta_t^{-r}\right).
\tag{15.60}
$$

Since $G_t(r) \downarrow 0$ as $t \to \infty$ by **15.9**(b), it is possible to choose m large enough that (15.57) and hence (15.60) hold with $\eta_t = G_t(r)^{1/(1+r)} = G_t(r)\eta_t^{-r}$ for each $t > m$. Therefore,

$$
\begin{aligned}
|P(G \cap H) - P(G)P(H)| &\ll \sum_{t=m+1}^{m+k} G_t(r)^{1/(1+r)} \\
&\leq \sum_{t=m+1}^{\infty} G_t(r)^{1/(1+r)}
\end{aligned}
\tag{15.61}
$$

where the right-hand sum is finite by **15.9**(b) and goes to zero as $m \to \infty$. This completes the proof. ∎

It is worth examining this argument with care to see how violation of the conditions can lead to trouble. According to (15.54) mixing will follow from two conditions: the obvious one is that the tail component $\boldsymbol{V}_2$, the $\mathcal{F}_{-\infty}^{0}$-measurable part of $\boldsymbol{X}_2$, becomes negligible, such that $P(E)$ gets close to 1 when m is large, even when $\boldsymbol{\eta}$ is allowed to approach $\mathbf{0}$. But in addition $\chi^* - \chi_*$ must disappear, ensured by condition **15.9**(a) as shown in (15.56). The implication is that $P(\boldsymbol{W}_2 \in D - \boldsymbol{v}_2)$ must approach a unique limit as $\boldsymbol{v}_2 \to \mathbf{0}$, for any D and whatever the path of convergence, which requires some smoothness of the distribution in addition to continuity. When there are atoms, it is not difficult to devise examples where the requirement fails. For example, in **15.6** the set B_t becomes $W_t - B_t$ on being translated a distance of 2^{-t}. For such a case these probabilities evidently do not converge in the limit as $t \to \infty$.

However, this is a sufficiency result and it remains unclear just how much more than the absence of atoms is strictly necessary. Consider an example where the distribution is continuous, having differentiable p.d.f., but condition (15.40) nonetheless fails.

15.12 Example Let

$$f(z) = C_0 z^{-2} \sin^2(z^4), z \in \mathbb{R}. \tag{15.62}$$

This is non-negative, continuous everywhere, and bounded by $C_0 z^{-2}$ and hence integrable. $\int_{-\infty}^{+\infty} f(z)\mathrm{d}z = 1$ by appropriate choice of C_0 so f is a p.d.f. By the mean value theorem,

$$|f(z+a) - f(z)| = |a||f'\left(z + \alpha(z)a\right)|, \ \alpha(z) \in [0,1] \tag{15.63}$$

where $f'(z) = 8C_0 \sin(z^4)\cos(z^4)z - 2C_0 \sin^2(z^4)z^{-3}$. But note that $\int_{-\infty}^{+\infty} |f'(z)|\mathrm{d}z = \infty$ and hence,

$$\frac{1}{|a|}\int_{-\infty}^{+\infty} |f(z+a) - f(z)|\mathrm{d}z \to \infty \text{ as } |a| \to 0. \tag{15.64}$$

This contradicts (15.40). The problem is that the density is varying too rapidly in the tails of the distribution and $|f(z+a) - f(z)|$ does not diminish rapidly enough in these regions as $a \to 0$.

The rate of divergence in (15.64) can be estimated. For fixed (small) a, $|f(z+a) - f(z)|$ is at a local maximum at points z at which $\sin(z+a)^4 = 1$ (or 0) and $\sin z^4 = 0$ (or 1) or, in other words, at points where $(z+a)^4 - z^4 = 4az^3 + O(a^2) = \pm\pi/2$. The solutions to these approximate relations can be written as $z = \pm C_1|a|^{-1/3}$ for $C_1 > 0$. At these points, (15.62) gives the bound

$$|f(z+a) - f(z)| \le 2f(z) \le 2C_0 C_1^{-2}|a|^{2/3}.$$

Within the interval $[-C_1|a|^{-1/3}, C_1|a|^{-1/3}]$ the integral is bounded by $4C_0C_1^{-1}|a|^{1/3}$ which is the area of the rectangle having height $2C_0C_1^{-2}|a|^{2/3}$. Outside the interval, f is bounded by $C_0 z^{-2}$ and the integral over this region is bounded by

$$2C_0 \int_{C_1|a|^{-1/3}}^{+\infty} z^{-2}\mathrm{d}z = 2C_0 C_1^{-1}|a|^{1/3}.$$

Adding up the approximations yields

$$\int_{-\infty}^{+\infty} |f(z+a) - f(z)| \mathrm{d}z \leq M|a|^{1/3} \tag{15.65}$$

for $M < \infty$. □

Condition (15.65) violates **15.9**(a), but it is possible for it to be sufficient at the cost of an additional restriction on the moving average coefficients.

15.13 Corollary Modify the conditions of **15.9** as follows: for $0 < \beta \leq 1$, assume that

(a) Z_t is uniformly L_r-bounded, independent, and continuously distributed with p.d.f. f_{Z_t} and

$$\sup_t \int_{-\infty}^{+\infty} |f_{Z_t}(z+a) - f_{Z_t}(z)| \mathrm{d}z \leq M|a|^{\beta},\ M < \infty \tag{15.66}$$

whenever $|a| \leq \delta$, for some $\delta > 0$

(b) $\sum_{t=0}^{\infty} G_t(r)^{\beta/(\beta+r)} < \infty$, where $G_t(r)$ is defined in (15.41)

(c) $\sum_{j=1}^{\infty} |\tau_j|^{\beta} < \infty$ where $\tau(x) = 1/\theta(x)$.

Then X_t is strong mixing with $\alpha_m = O\big(\sum_{t=m+1}^{\infty} G_t(r)^{\beta/(\beta+r)}\big)$.

Proof This follows the proof of **15.9** until (15.56), which becomes

$$\chi^* - \chi_* \leq 2M \sup_{v \in B'} \sum_{t=m+1}^{m+k} |\hat{v}_t|^{\beta} \tag{15.67}$$

applying the obvious extension of Lemma **15.10**. Note that

$$\sum_{t=m+1}^{m+k} |\hat{v}_t|^{\beta} = \sum_{t=m+1}^{m+k} \left| \sum_{j=0}^{t-m-1} \tau_j v_{t-j} \right|^{\beta} \leq \sum_{j=0}^{\infty} |\tau_j|^{\beta} \sum_{t=m+1}^{m+k} \eta_t^{\beta} \tag{15.68}$$

using **2.21**, since $0 < \beta \leq 1$. Applying assumption **15.13**(c),

$$|P(G \cap H) - P(G)P(H)| \ll \sum_{t=m+1}^{m+k} (\eta_t^{\beta} + G_r(r)\eta_t^{-r}) \tag{15.69}$$

and the result is obtained as before, but in this case setting $\eta_t = G^{1/(\beta+r)}$. ∎

Condition (b) is satisfied when $|\theta_j| \ll j^{-\mu}$ for $\mu > 1/\beta + 2/r$ when $r \leq 2$ and $\mu > 1/2 + 1/r + 1/\beta$ when $r \geq 2$, which shows how the summability restrictions have to be strengthened when β is close to 0. This is nonetheless a useful extension because there are important cases where **15.13**(b) and **15.13**(c) are easily satisfied. In particular, if the process is finite-order ARMA, both $|\theta_j|$ and $|\tau_j|$ either decline geometrically or vanish beyond some finite j and **15.13**(b) and **15.13**(c) both hold.

Condition **15.13**(a) is nonetheless a strengthening of continuity. Look again at Example **15.12** and note that setting $\beta = \frac{1}{3}$ will satisfy condition **15.13**(a). It is easy to generalize this example.

15.14 Example (**15.12** continued) Put $f(z) = C\sin^2(z^k)z^{-2}$ for $k \geq 4$ in place of (15.62). The argument may be modified to show that the integral converges at the rate $|a|^{1/(k-1)}$ and this choice of β is appropriate. However, for $f(z) = C\sin^2(e^z)z^{-2}$ the integral converges more slowly than $|a|^\beta$ for all $\beta > 0$ and condition **15.13**(a) fails. □

To conclude this chapter consider the case of uniform mixing. Manipulating inequalities (15.50)–(15.53) yields

$$|P(H|G) - P(H)| \leq \chi^* - \chi_* + P(E^c)\left(1 + \frac{1}{P(G)}\right) \tag{15.70}$$

which shows that uniform mixing can fail unless $P(E) = 1$ for all m exceeding a finite value. Otherwise, a sequence of events G can always be constructed whose probability is positive but approaching 0 no slower than $P(E^c)$. When the support of $(X_{-p}, \ldots, X_0)$ is unbounded this kind of thing can occur, as illustrated by **15.8**. The essence of this example does not depend on the AR(1) model and similar cases could be constructed in the general MA(∞) framework. Sufficient conditions must include a.s. boundedness of the distributions, and the summability conditions are also modified. The extended version of the strong mixing condition in **15.13** is used, although it is easy to deduce the relationship between these conditions and **15.9** by setting $\beta = 1$ in the next result.

15.15 Theorem Modify the conditions of **15.13** as follows. Let (a) and (c) hold as before, but replace (b) by

(b) $\sum_{t=0}^{\infty}(\sum_{j=t}^{\infty}|\theta_j|)^\beta < \infty$

and add

(d) $\{Z_t\}$ is uniformly bounded a.s.

Then $\{X_t\}$ is uniform mixing with $\phi_m = O\big(\sum_{t=m+1}^{\infty}(\sum_{j=t}^{\infty}|\theta_j|)^{\beta}\big)$.

Proof Follow the proof of **15.9** up to (15.53), but replace (15.54) with (15.70). By condition **15.15**(d), there exists $K < \infty$ such that $\sup_t |Z_t| < K$ a.s. and hence $|X_t| < K\sum_{j=0}^{\infty}|\theta_j|$ a.s. It further follows, recalling the definition of $\boldsymbol{V}_2$, that $P(E) = 1$ when $\eta_t < K\sum_{j=t}^{\infty}|\theta_j|$ for $t = m+1, \ldots, m+k$. Substituting directly into (15.70) from (15.67) and (15.68) and making this choice of η gives, for any G with $P(G) > 0$,

$$|P(H|G) - P(H)| \ll \sum_{t=m+1}^{m+k}\left(\sum_{j=t}^{\infty}|\theta_j|\right)^{\beta}. \tag{15.71}$$

The result now follows by the same considerations as before. ∎

These summability conditions are tougher than in **15.13**. Letting $r \to \infty$ in the latter case for comparability, **15.13**(b) is satisfied when $|\theta_j| = O(j^{-\mu})$ for $\mu > 1/2 + 1/\beta$, while the corresponding implication of **15.15**(b) is $\mu > 1 + 1/\beta$.

16
Martingales

16.1 Sequential Conditioning

It is trivial to observe that the arrow of time is unidirectional. When a random time series $\{\ldots, X_{t-1}, X_t, X_{t+1}, \ldots\}$ is generated, X_t is determined in an environment in which the preceding members X_{t-k} for $k > 0$ are given, whereas the members following remain contingent. The past is known, but the future is unknown. The operation of conditioning sequentially on past events is therefore of central importance in time series modelling. Partial knowledge is characterized by specifying a sub-σ-field of events from $\mathcal{F}$ for which it is known whether each of the events belonging to it has occurred or not. The accumulation of information by an observer as time passes is represented by an increasing (sometimes called *nested*) sequence of sub-σ-fields, $\boldsymbol{F} = \{\mathcal{F}_t\}_{-\infty}^{\infty}$, such that $\ldots \subseteq \mathcal{F}_{-1} \subseteq \mathcal{F}_0 \subseteq \mathcal{F}_1 \subseteq \mathcal{F}_2 \subseteq \ldots \subseteq \mathcal{F}$. $\boldsymbol{F}$ is commonly called a *filtration*. For economy of notation the symbol $\mathcal{F}_t$ is used here to denote what has been previously written as $\mathcal{F}_{-\infty}^{t}$. A sub-$\sigma$-field bearing a time subscript but no superscript will always be interpreted in this way. The quadruple of objects $(\Omega, \mathcal{F}, \boldsymbol{F}, P)$ in which the full σ-field of events $\mathcal{F}$ has been equipped with a filtration is known as a *filtered* probability space.

If the random variable X_t is $\mathcal{F}_t$-measurable for each t, $\boldsymbol{F}$ is said to be *adapted* to the sequence $\{X_t\}_{-\infty}^{\infty}$. The sequence of pairs $\{X_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ is called an adapted sequence. Setting $\mathcal{F}_t = \sigma(X_s, -\infty < s \leq t)$ defines the minimal adapted sequence, but $\mathcal{F}_t$ typically has the interpretation of an observer's information set and can contain more information than the history of a single variable. When X_t is integrable, the conditional expectation $\mathrm{E}(X_t|\mathcal{F}_{t-1})$ is defined and can be thought of as the optimal predictor of X_t from the point of view of observers looking one period ahead (compare **10.12**).

Consider an adapted sequence $\{S_n, \mathcal{F}_n\}_{-\infty}^{\infty}$ in the space $(\Omega, \mathcal{F}, \boldsymbol{F}, P)$ with the properties

$$\mathrm{E}|S_n| < \infty \tag{16.1}$$

$$\mathrm{E}(S_n|\mathcal{F}_{n-1}) = S_{n-1} \text{ a.s.} \tag{16.2}$$

Such a sequence is called a *martingale*. Originally the name of a type of horse's harness, in old-fashioned gambling parlance 'martingale' was the slang term for

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0016

the policy of attempting to recoup a loss by doubling one's stake on the next bet. The modern usage in probability theory is closer to describing a gambler's worth in the course of a sequence of fair bets. In view of (10.18), an alternative version of condition (16.2) is

$$\int_A S_n \mathrm{d}P = \int_A S_{n-1} \mathrm{d}P, \text{ each } A \in \mathcal{F}_{n-1}. \tag{16.3}$$

A commoner form of this definition specifies the sequence $\{S_n, \mathcal{F}_n\}_1^\infty$ where (16.2) holds for every $n > 1$ with S_1 any integrable and $\mathcal{F}_1$-measurable r.v. It will become apparent that an arbitrarily remote start date can be problematic, but it is not ruled out completely.

It is important not to misunderstand the force of the integrability requirement in (16.1). After all, predicting S_n after observing S_{n-1} might seem to be just a matter of knowing something about the distribution of the increment $S_n - S_{n-1}$. The problem is that $\mathrm{E}(S_n|\mathcal{F}_{n-1})$ cannot be treated *as a random variable* without integrability of S_n. Conditioning on $\mathcal{F}_{n-1} = \sigma(S_{n-1}, S_{n-2}, \ldots)$ is not the same proposition as treating S_{n-1} as a constant, which entails restricting the probability space entirely to the set of repeated random drawings of $S_n - S_{n-1}$. The latter problem has no connection with the theory of random sequences.

Here are some well-known cases.

16.1 Example Let $\{X_t\}_1^\infty$ be an i.i.d. integrable sequence with zero mean. If $S_n = \sum_{t=1}^n X_t$ and $\mathcal{F}_n = \sigma(X_n, X_{n-1}, \ldots, X_1)$ then $\{S_n, \mathcal{F}_n\}_1^\infty$ is a martingale, also known as a random walk sequence (see §13.4). Note that $\mathrm{E}|S_n| \le \sum_{t=1}^n \mathrm{E}|X_t| < \infty$ for all finite n. □

16.2 Example Let $\{X_t\}_1^\infty$ be an i.i.d. non-negative sequence with $\mathrm{E}(X_1) = 1$, let $S_n = \prod_{t=1}^n X_t$ and $\mathcal{F}_n = \sigma(X_n, X_{n-1}, \ldots, X_1)$. Then $\{S_n, \mathcal{F}_n\}_1^\infty$ is a martingale, noting that $\mathrm{E}|S_n| = 1$ and $\mathrm{E}(S_n|\mathcal{F}_{n-1}) = S_{n-1}\mathrm{E}(X_n) = S_{n-1}$ a.s. Curiously, the logarithm of this sequence is another martingale. □

16.3 Example Let X be an integrable, $\mathcal{F}/B$-measurable, zero-mean r.v., $\{\mathcal{F}_n\}_{-\infty}^\infty$ a filtration with $\lim_{n\to\infty} \mathcal{F}_n = \mathcal{F}$, and define $S_n = \mathrm{E}(X|\mathcal{F}_n)$. Then

$$\mathrm{E}(S_n|\mathcal{F}_{n-1}) = \mathrm{E}(\mathrm{E}(X|\mathcal{F}_n)|\mathcal{F}_{n-1}) = \mathrm{E}(X|\mathcal{F}_{n-1}) = S_{n-1} \text{ a.s.} \tag{16.4}$$

where the second equality is by **10.27**(i). $\mathrm{E}|S_n| \le \mathrm{E}|X| < \infty$ by **10.15**, so S_n is a martingale. Notice that in this example an arbitrarily remote start date is not problematic. □

The following is an explicit construction of a martingale with the properties of **16.3**, which is worth relating in detail since it has an important application in the sequel; see §28.5.

16.4 Example Let X denote an integrable r.v. from the space $(\mathbb{R}, \mathcal{B}, \mu)$ having zero mean. Derive a filtration $\{\mathcal{F}_n, n \geq 1\}$ from the distribution of X by successive partitions of the line into intervals, their number doubling at each step. Thus, let $\mathcal{F}_n = \sigma(\mathcal{C}_n)$ where $\mathcal{C}_n = \{A^n_{m(n)}, m(n) = 1, \dots, 2^n\}$ and the $A^n_{m(n)}$ are constructed as follows. If H denotes an interval of the line with $\mu(H) > 0$, let

$$M(H) = \frac{1}{\mu(H)} \int_H x \mathrm{d}\mu. \tag{16.5}$$

$M(H)$ is the optimal predictor of X given the knowledge that it lies in H. At step n, if $\mu(A^n_{m(n)}) > 0$ the partition is extended by dividing interval $A^n_{m(n)}$ at the point $M(A^n_{m(n)})$. Since $M(\mathbb{R}) = 0$ by assumption on X the elements of $\mathcal{C}_1$ are $A^1_1 = (-\infty, 0)$ and $A^1_2 = [0, \infty)$. $\mathcal{C}_2$ has the elements $A^2_1 = (-\infty, M(A^1_1))$, $A^2_2 = [M(A^1_1), 0)$, $A^2_3 = [0, M(A^1_2))$, and $A^2_4 = [M(A^1_2), \infty)$; and so forth. In the case $\mu(A^n_{m(n)}) = 0$ simply partition into $A^n_{m(n)}$ and $\emptyset$.

Now construct a random sequence $\{S_n\}_0^\infty$ as follows. Set $S_0 = 0 = M(A^0_{m(0)})$ where $m(0) = 1$ and $A^0_{m(0)} = \mathbb{R}$. Then at step $n \geq 1$, $S_{n-1} = M(A^{n-1}_{m(n-1)})$ and the interval $A^{n-1}_{m(n-1)}$ is partitioned at the point S_{n-1} into the left and right subintervals $A^n_{2m(n-1)-1}$ and $A^n_{2m(n-1)}$. Choose between these at random, either the left-hand interval with probability $\mu(A^n_{2m(n-1)-1})/\mu(A^{n-1}_{m(n-1)})$ or the right-hand interval with probability $\mu(A^n_{2m(n-1)})/\mu(A^{n-1}_{m(n-1)})$. Denoting the chosen interval by $A^n_{m(n)}$ where either $m(n) = 2m(n-1) - 1$ or $m(n) = 2m(n-1)$, set $S_n = M(A^n_{m(n)})$.

Note that $\mathrm{E}|S_n| \leq \mathrm{E}|X|$ by the LIE and **10.15**. By construction, $S_n|\mathcal{F}_{n-1}$ has a two-point distribution (see **9.1**). Since $A^{n-1}_{m(n-1)} = A^n_{2m(n-1)-1} \cup A^n_{2m(n-1)}$ the conditional mean is

$$\begin{aligned}
\mathrm{E}(S_n|\mathcal{F}_{n-1}) &= \frac{\mu(A^n_{2m(n-1)-1})}{\mu(A^{n-1}_{m(n-1)})} M(A^n_{2m(n-1)-1}) + \frac{\mu(A^n_{2m(n-1)})}{\mu(A^{n-1}_{m(n-1)})} M(A^n_{2m(n-1)}) \\
&= \frac{1}{\mu(A^{n-1}_{m(n-1)})} \left(\int_{A^n_{2m(n-1)-1}} x \mathrm{d}\mu + \int_{A^n_{2m(n-1)}} x \mathrm{d}\mu \right) \\
&= M(A^{n-1}_{m(n-1)}) = S_{n-1} \text{ a.s.}
\end{aligned} \tag{16.6}$$

It follows that $\{S_n\}_1^\infty$ is a martingale.

If $\mathcal{F}_n \to \mathcal{B}$ then $S_n \to X$ a.s. where $X \in \bigcap_n A^n_{m(n)}$. Since $P(X \in A^n_{m(n)}) = \mu(A^n_{m(n)})$ for every n by construction, $X \sim_{\mathrm{d}} \mu$. □

The next, closely related concept is the one that arises most frequently in econometric time series analysis. A *martingale difference* (m.d.) sequence $\{X_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ is an adapted sequence in $(\Omega, \mathcal{F}, \boldsymbol{F}, P)$ satisfying the properties

$$\mathrm{E}|X_t| < \infty \tag{16.7}$$

$$\mathrm{E}(X_t|\mathcal{F}_{t-1}) = 0 \text{ a.s.} \tag{16.8}$$

for every t. Evidently, if $\{S_n\}$ is a martingale and $X_t = S_t - S_{t-1}$, then $\{X_t\}$ is a m.d. Any independent integrable sequence is a m.d. A martingale can be defined as the partial sum of a sequence of m.d.s, a natural way to think about the random walk in **16.1** for example, although if X_t has positive variance uniformly in t, condition (16.1) holds for all finite n but *not* uniformly in n. Therefore, to define a martingale by $S_n = \sum_{t=-\infty}^{n} X_t$ can lead to difficulties. For the product martingale **16.2** the m.d.s take the form $S_{n-1}(X_n - 1)$ which can equally be verified to satisfy the definition. Example **16.3** shows how a martingale can arise without any reference to the summation of a difference sequence. The following case further illustrates the generality of the concept.

16.5 Example Let $\{Y_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ denote any integrable adapted sequence. The centred sequence $X_t = Y_t - \mathrm{E}(Y_t|\mathcal{F}_{t-1})$ is a martingale difference since $\mathrm{E}|X_t| \le 2\mathrm{E}|Y_t| < \infty$ by the triangle inequality and LIE and

$$\mathrm{E}(X_t|\mathcal{F}_{t-1}) = \mathrm{E}(Y_t|\mathcal{F}_{t-1}) - \mathrm{E}(Y_t|\mathcal{F}_{t-1}) = 0 \text{ a.s.} \quad □$$

A fundamental property of a m.d. is that it is uncorrelated with any measurable function of its lagged values.

16.6 Theorem If $\{X_t, \mathcal{F}_t\}$ is a m.d. then

$$\mathrm{Cov}\big(X_t, \phi(X_{t-1}, X_{t-2}, \dots)\big) = 0$$

where ϕ is any Borel-measurable, integrable function of the arguments.

Proof By **10.11** (see also the remarks following) noting that $\phi(X_{t-1}, X_{t-2}, \dots)$ is $\mathcal{F}_{t-1}$-measurable. ∎

16.7 Corollary If $\{X_t, \mathcal{F}_t\}$ is a m.d. then $\mathrm{E}(X_t X_{t-k}) = 0$ for all t and all $k \neq 0$.

Proof Put $\phi = X_{t-k}$ in **16.6**. For $k < 0$ redefine the subscripts, replacing t by $t' = t - k$ and $t - k$ by $t' - |k| = t$ and so making the two cases equivalent. ■

In the hierarchy of constraints on the dependence of an integrable sequence, one might think of the m.d. property as intermediate between uncorrelatedness and independence. That is to say, m.d.s are uncorrelated sequences and independent sequences are m.d.s, while the reverse implications don't hold. However, note the asymmetry with respect to time. Reversing the time ordering of an independent sequence yields another independent sequence and likewise a reversed uncorrelated sequence is uncorrelated, but a reversed m.d. is not a m.d. in general.

The *Doob decomposition* of an integrable adapted sequence $\{S_n, \mathcal{F}_n\}_0^\infty$ is

$$S_n = M_n + A_n \tag{16.9}$$

where $A_0 = 0$, $M_0 = S_0$ and for $n \geq 1$,

$$M_n = M_{n-1} + S_n - \mathrm{E}(S_n|\mathcal{F}_{n-1}) \tag{16.10}$$

$$A_n = A_{n-1} + \mathrm{E}(S_n|\mathcal{F}_{n-1}) - S_{n-1}. \tag{16.11}$$

A_n is an $\mathcal{F}_{n-1}$-measurable sequence called the *predictable component* of S_n. Writing $Y_n = S_n - S_{n-1}$ and $X_n = M_n - M_{n-1}$, $A_n - A_{n-1} = \mathrm{E}(Y_n|\mathcal{F}_{n-1})$ and

$$X_n = Y_n - \mathrm{E}(Y_n|\mathcal{F}_{n-1}). \tag{16.12}$$

According to Example **16.5**, $\{X_n, \mathcal{F}_n\}_0^\infty$ is a m.d. known as the *innovation sequence* of S_n. Applying the the triangle inequality, conditional modulus inequality, and LIE gives

$$\begin{aligned} \mathrm{E}|X_n| &\leq \mathrm{E}|Y_n| + \mathrm{E}\big(\mathrm{E}(|Y_n||\mathcal{F}_{n-1})\big) \\ &= 2\mathrm{E}|Y_n| \leq 4\mathrm{E}|S_n| < \infty. \end{aligned} \tag{16.13}$$

Hence, $\{M_n, \mathcal{F}_n\}_0^\infty$ is a martingale. If $\{S_n, \mathcal{F}_n\}_0^\infty$ is a martingale then $A_n = 0$ by construction.

The following example emphasizes the generality of the Doob representation, which allows a martingale to be created out of an arbitrary integrable process.

16.8 Example Let S_n be a serially independent integrable process. In this case, $\mathrm{E}(S_n|\mathcal{F}_{n-1}) = \mathrm{E}(S_n)$ and (16.10) and (16.11) have the solutions

$$M_n = \sum_{t=1}^{n}(S_t - \mathrm{E}(S_t)), \quad A_n = \sum_{t=1}^{n}(\mathrm{E}(S_t) - S_{t-1})$$

so that M_n is a martingale (random walk) and (16.9) resolves the telescoping sum. □

Martingales play an indispensable role in modern probability theory because m.d.s behave in many important respects like independent sequences. Independence is the simplifying property that permitted the 'classical' limit results, laws of large numbers and central limit theorems, to be proved. However, independence is a constraint on the entire joint distribution of the sequence, aspects of which may be both difficult to verify and innocuous if violated. The m.d. property is a much milder restriction on the memory and yet, as later chapters will show, many limit theorems that hold for independent sequences can also be proved for m.d.s with few if any additional restrictions on the marginal distributions. For time series applications, it makes sense to go directly to the martingale version of any result of interest, unless of course a still-weaker assumption will suffice. A stronger one will be needed only rarely.

To avoid the use of a definition involving σ-fields on an abstract probability space, it is possible to represent a martingale difference as, for example, a sequence with the property

$$\mathrm{E}(X_t|X_{t-1}, X_{t-2}, \dots) = 0 \text{ a.s.} \tag{16.14}$$

When in the sequel a random variable appears in a conditioning set, it is to be understood as representing the corresponding minimal sub-σ-field, in this case $\sigma(X_{t-1}, X_{t-2}, \dots)$. This is appealing at an elementary level since it captures the notion of information available to an observer, in this case the sequence realization to date, but since the conditioning information can extend more widely than the history of the sequence itself, this type of notation is relatively clumsy. Suppose that in a vector sequence $\{(X_t, Z_t)\}$, X_t—though not necessarily Z_t—is a m.d. with respect to $\mathcal{F}_t = \sigma(X_t, Z_t, X_{t-1}, Z_{t-1}, \dots)$ in the sense of (16.8). This case is distinct from (16.14) and shows that definition is inadequate, although (16.8) implies (16.14). More importantly, the representation of conditioning information is not unique and **10.3**(ii) shows that any measurably isomorphic transformation of the conditioning variables contains the same information as the original variables. Indeed, the information need not even be represented by a variable, but is merely knowledge of the occurrence/non-occurrence of certain abstract events.

In §13.3, the concept of a linear process was defined in **13.8** as one having i.i.d. increments. It was also emphasized that simple uncorrelatedness of the increments, as in the Wold moving average representation (**13.14**), did not impart the special properties of linearity. There is now a third possibility that arises very frequently in applications, that of a moving average of martingale differences. Such processes have a number of useful properties that will be exploited in the sequel, and the notion of unpredictability in levels is often more easily justified empirically than the rather strong assumption of serial independence. The term linear process is often used loosely in this context, but the more neutral expression 'moving average' may be found preferable.

16.2 Extensions of the Martingale Concept

An adapted triangular array $\left\{\{X_{nt},\mathcal{F}_{nt}\}_{t=1}^{k_n}\right\}_{n=1}^{\infty}$, where $\{k_n\}_{n=1}^{\infty}$ is some increasing sequence of integers, for which

$$\mathrm{E}|X_{nt}| < \infty \tag{16.15}$$

$$\mathrm{E}(X_{nt}|\mathcal{F}_{n,t-1}) = 0 \text{ a.s.} \tag{16.16}$$

for each $t = 1,\ldots,k_n$ and $n \geq 1$, is called a *martingale difference array.* In many applications $k_n = n$. The double subscripting of the $\mathcal{F}_{nt}$ may be superfluous if the information content of the array does not depend on n, with $\mathcal{F}_{nt} = \mathcal{F}_t$ for each n, but the additional generality given by the definition is harmless and could be useful. The sequence $\{S_n,\mathcal{F}_n\}_1^{\infty}$ where $S_n = \sum_{t=1}^{k_n} X_{nt}$ and $\mathcal{F}_n = \mathcal{F}_{n,k_n}$ is not a martingale, but the properties of martingales can be profitably used to analyse its behaviour. Consider the case $S_n = n^{-1/2}\sum_{t=1}^{n} X_t$ where $\{X_t,\mathcal{F}_t\}$ is a m.d. Such scaling by sample size may ensure that the distribution of S_n is non-degenerate in the limit. S_n is not a martingale since

$$\mathrm{E}(S_n|\mathcal{F}_{n-1}) = \left((n-1)/n\right)^{1/2} S_{n-1} \tag{16.17}$$

but each column of the m.d. array

$$\begin{bmatrix} X_1 & 2^{-1/2}X_1 & 3^{-1/2}X_1 & 4^{-1/2}X_1 & \cdots \\ & 2^{-1/2}X_2 & 3^{-1/2}X_2 & 4^{-1/2}X_2 & \cdots \\ & & 3^{-1/2}X_3 & 4^{-1/2}X_3 & \cdots \\ & & & 4^{-1/2}X_4 & \cdots \\ & & & & \cdots \end{bmatrix} \tag{16.18}$$

is a m.d. sequence and S_n is the sum of column n. It is a term in a martingale sequence even though this is not the sequence $\{S_n\}$.

An adapted sequence $\{S_n, \mathcal{F}_n\}_{-\infty}^{\infty}$ of L_1-bounded variables satisfying

$$\mathrm{E}(S_n|\mathcal{F}_{n-1}) \geq S_{n-1} \text{ a.s.} \tag{16.19}$$

is called a *submartingale*, in which case $X_n = S_n - S_{n-1}$ is a submartingale difference having the property $\mathrm{E}(X_n|\mathcal{F}_{n-1}) \geq 0$ a.s. Reversing the inequality defines a *supermartingale*, although, since $-S_n$ is a supermartingale whenever S_n is a submartingale, this is a minor extension. A supermartingale might represent a gambler's worth when a sequence of bets is unfair because of a house percentage. The generic term *semimartingale* covers all the possibilities.

16.9 Theorem Let $\phi(\cdot) : \mathbb{R} \mapsto \mathbb{R}$ be continuous and convex. If $\{S_n, \mathcal{F}_n\}$ is a martingale and $\mathrm{E}|\phi(S_n)| < \infty$, then $\{\phi(S_n), \mathcal{F}_n\}$ is a submartingale. If ϕ is also non-decreasing, $\{\phi(S_n), \mathcal{F}_n\}$ is a submartingale if $\{S_n, \mathcal{F}_n\}$ is a submartingale.

Proof For the martingale case,

$$\mathrm{E}\big(\phi(S_{n+1})|\mathcal{F}_n\big) \geq \phi\big(\mathrm{E}(S_{n+1}|\mathcal{F}_n)\big) = \phi(S_n) \text{ a.s.} \tag{16.20}$$

by the conditional Jensen inequality (**10.19**). For the submartingale case, '=' becomes '≥' in (16.20) when $x_1 \leq x_2 \Rightarrow \phi(x_1) \leq \phi(x_2)$. ∎

If $\{X_t, \mathcal{F}_t\}_1^{\infty}$ is a martingale difference, $\{Z_t, \mathcal{F}_t\}_0^{\infty}$ any adapted sequence, and

$$S_n = \sum_{t=1}^{n} X_t Z_{t-1} \tag{16.21}$$

then $\{S_n, \mathcal{F}_n\}_1^{\infty}$ is a martingale since $\mathrm{E}(X_{n+1}Z_n|\mathcal{F}_n) = Z_n \mathrm{E}(X_{n+1}|\mathcal{F}_n) = 0$ and hence

$$\mathrm{E}(S_{n+1}|\mathcal{F}_n) = \sum_{t=1}^{n} X_t Z_{t-1} = S_n. \tag{16.22}$$

If $\{X_t\}$ is a submartingale difference and $\{Z_t\}$ a non-negative sequence then S_n is a submartingale, with the extra term $Z_n \mathrm{E}(X_{n+1}|\mathcal{F}_n)$ now appearing in (16.22). Think of X_t as the random return on a stake of 1 unit in a sequence of bets and the sequence $\{Z_t\}$ as representing a betting system, a rule based on information available at time $t-1$ for deciding how many units to bet in the next game. The implication of (16.22) is that, if the basic game (in which the same stake is bet every time) is fair, there is no betting system (based on no more than information about past play) that can turn it into a game favouring the player, or for that matter turn a game favouring the house into a fair game.

In the filtered space $(\Omega, \mathcal{F}, \boldsymbol{F}, P)$ a *stopping time* $\tau(\omega)$, $\omega \in \Omega$, is a random integer having the property $\{\omega : \tau(\omega) = t\} \in \mathcal{F}_t$. The classic example is a gambling policy which entails withdrawing from the game whenever a certain condition depending only on the outcomes to date (such as one's losses exceeding some limit, or a certain number of successive wins) is realized. If τ is the random variable defined as the first time the said condition is met in a sequence of bets, it is a stopping time.

Let τ be a stopping time of $\boldsymbol{F}$ and consider

$$S_{n\wedge\tau} = \begin{cases} S_n, & n \le \tau \\ S_\tau, & n > \tau \end{cases} \tag{16.23}$$

where $n \wedge \tau$ stands for $\min\{n, \tau\}$. $\{S_{n\wedge\tau}, \mathcal{F}_n\}_{n=1}^{\infty}$ is called a *stopped process.*

16.10 Theorem If $\{S_n, \mathcal{F}_n\}_1^\infty$ is a martingale (submartingale), then $\{S_{n\wedge\tau}, \mathcal{F}_n\}_1^\infty$ is a martingale (submartingale).

Proof Since $\{\mathcal{F}_n\}_1^\infty$ is an increasing sequence, $\{\omega : \tau(\omega) = k\} \in \mathcal{F}_n$ for $k < n$ and also $\{\omega : \tau(\omega) \ge n\} \in \mathcal{F}_n$ by complementation. Write $S_{n\wedge\tau} = \sum_{k=1}^{n-1} S_k 1_{\{k=\tau\}} + S_n 1_{\{n\le\tau\}}$, where the indicator functions are all $\mathcal{F}_n$-measurable. It follows by **3.32** and **3.40** that $S_{n\wedge\tau}$ is $\mathcal{F}_n$-measurable and

$$\begin{aligned} \mathrm{E}|S_{n\wedge\tau}| &\le \sum_{k=1}^{n-1} \mathrm{E}|S_k 1_{\{k=\tau\}}| + \mathrm{E}|S_n 1_{\{n\le\tau\}}| \\ &\le \sum_{k=1}^{n-1} \mathrm{E}|S_k| + \mathrm{E}|S_n| < \infty, \; n \ge 1. \end{aligned} \tag{16.24}$$

If $\{S_n, \mathcal{F}_n\}_1^\infty$ is a martingale then for $A \in \mathcal{F}_n$, applying (16.3),

$$\begin{aligned} \int_A S_{(n+1)\wedge\tau} \mathrm{d}P &= \int_{A\cap\{n\le\tau\}} S_{n+1} \mathrm{d}P + \int_{A\cap\{n>\tau\}} S_\tau \mathrm{d}P \\ &= \int_{A\cap\{n\le\tau\}} S_n \mathrm{d}P + \int_{A\cap\{n>\tau\}} S_\tau \mathrm{d}P = \int_A S_{n\wedge\tau} \mathrm{d}P \end{aligned} \tag{16.25}$$

showing that $\{S_{n\wedge\tau}, \mathcal{F}_n\}_1^\infty$ is a martingale. The submartingale case follows easily on replacing the second equality by the required inequality in (16.25). ∎

The general conclusion is that a gambler cannot alter the basic fairness characteristics of a game, *whatever* gambling policy (betting system plus stopping rule) he or she selects.

All these concepts have a natural extension to random vectors. An adapted sequence $\{\boldsymbol{X}_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ where $\boldsymbol{X}_t$ is $m \times 1$ is defined to be a *vector martingale difference* if and only if $\{\boldsymbol{\lambda}'\boldsymbol{X}_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ is a scalar m.d. sequence for all conformable fixed m-vectors $\boldsymbol{\lambda} \neq \mathbf{0}$. As is easily seen by taking the cases with $\boldsymbol{\lambda}$ the columns of $\boldsymbol{I}_m$ in turn and stacking, it has the property

$$\mathrm{E}(\boldsymbol{X}_{t+1}|\mathcal{F}_t) = \mathbf{0} \ \ (m \times 1). \tag{16.26}$$

The one thing to remember is that a vector martingale difference is not the same thing as a vector of martingale differences. A simple counterexample is the two-element vector $\boldsymbol{X}_t = (X_t, X_{t-1})'$, where X_t is a m.d.; $\{\lambda_1 X_t + \lambda_2 X_{t-1}, \mathcal{F}_t\}$ is an adapted sequence, but

$$\mathrm{E}(\lambda_1 X_t + \lambda_2 X_{t-1}|\mathcal{F}_{t-1}) = \lambda_2 X_{t-1} \neq 0$$

so it is not a m.d. On the other hand,

$$\mathrm{E}(\lambda_1 X_{t+1} + \lambda_2 X_t|\mathcal{F}_{t-1}) = 0$$

but $\{\lambda_1 X_{t+1} + \lambda_2 X_t, \mathcal{F}_t\}$ is not adapted, since X_{t+1} is not $\mathcal{F}_t$-measurable.

16.3 Martingale Convergence

Applying **16.9** to the case $\phi(\cdot) = |\cdot|^p$ and taking unconditional expectations shows that every L_p-bounded martingale or submartingale for $p \geq 1$ has the property

$$\mathrm{E}|S_{n+1}|^p \geq \mathrm{E}|S_n|^p. \tag{16.27}$$

By **2.1** the sequence of p^{th} absolute moments converges as $n \to \infty$, either to a finite limit or to $+\infty$. If in particular the sequence of L_1-norms is uniformly bounded, (sub)martingales must converge almost surely to some point which is random in the sense of having a distribution over realizations, but does not change from one time period t to the next.

The intuition is reasonably transparent. $\{\mathcal{F}_n\}$ is an increasing sequence of σ-fields which converges to a limit $\mathcal{F}_\infty$, the σ-field that contains $\mathcal{F}_n$ for every n. Since $\mathrm{E}(S_n|\mathcal{F}_n) = S_n$, the convergence of the sequence $\{\mathcal{F}_n\}$ implies the convergence of a uniformly bounded sequence with the property $\mathrm{E}(S_{n+1}|\mathcal{F}_n) \geq S_n$, so long as these expectations remain well-defined in the limit.

16.11 Theorem If $\{S_n, \mathcal{F}_n\}_1^\infty$ is a submartingale sequence and $\sup_n \mathrm{E}|S_n| \leq M < \infty$, then $S_n \to S_\infty$ a.s. where S_∞ is a $\mathcal{F}$-measurable random variable with $\mathrm{E}|S_\infty| \leq M$. □

To fix ideas, consider the quoted examples. Example **16.1**, the random walk, is a non-convergent martingale. There is no random variable S_∞ to which it tends and $\mathrm{E}|S_n|$ diverges as $n \to \infty$. On the other hand, if $S_n = \mathrm{E}(X|\mathcal{F}_n)$ for a r.v. X, the condition $\mathrm{E}|X| < \infty$ is sufficient for $\{S_n\}$ to be uniformly integrable by **12.14** and hence the uniform L_1-boundedness condition of **16.11** is satisfied. Examples **16.3** and **16.4** are both evidently convergent cases where the limit S_∞ is clearly identified. Example **16.2** is another convergent case, noting that the process is non-negative and $\mathrm{E}|S_n| = \mathrm{E}(S_n) = 1$ for every n. However, see **16.15** for an additional point of interest about product martingales.

The proof of **16.11**, due to Doob, makes use of a result called the *upcrossing inequality,* which is proved as a preliminary lemma. Considering the path of a submartingale through time, an *upcrossing* of an interval $[\alpha, \beta]$ is a succession of steps starting at or below α and terminating at or above β. To complete more than one upcrossing, there must be one and only one intervening downcrossing, so downcrossings do not require separate consideration. Figure 16.1 shows two upcrossings of $[\alpha, \beta]$, spanning the periods marked by dots on the abscissa.

Let the r.v. Y_k be the indicator of an upcrossing. To be precise, set $Y_1 = 0$ and then, for $k = 2, 3, \dots, n$,

$$Y_k = \begin{cases} 0 \text{ if either } Y_{k-1} = 0, S_{k-1} > \alpha, \text{ or } Y_{k-1} = 1, S_{k-1} \geq \beta \\ 1 \text{ if either } Y_{k-1} = 0, S_{k-1} \leq \alpha, \text{ or } Y_{k-1} = 1, S_{k-1} < \beta. \end{cases} \tag{16.28}$$

The values of Y_k appear at the bottom of Figure 16.1. Observe that an upcrossing begins the period after S_k falls to or below α and ends at the first step thereafter where β is reached or exceeded. Y_k is a function of S_{k-1} and an $\mathcal{F}_{k-1}$-measurable random variable.

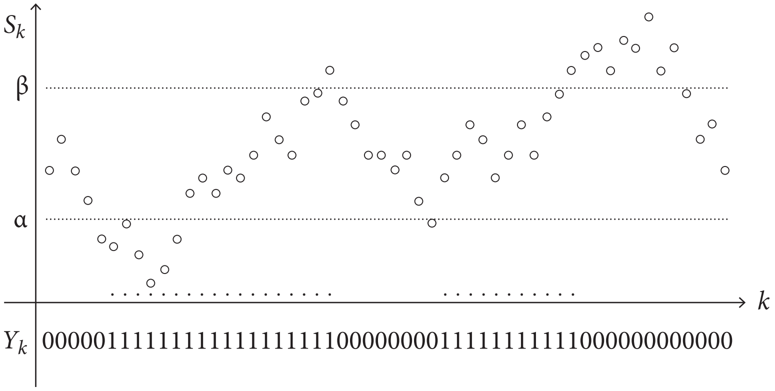

Figure 16.1

The number of upcrossings of $[\alpha,\beta]$ up to time n of the sequence $\{S_n(\omega)\}_1^\infty$, to be denoted $U_n(\omega)$, is an $\mathcal{F}_n$-measurable random variable. The sequence $\{U_n(\omega)\}_1^\infty$ is monotone, but it satisfies the following condition.

16.12 Lemma (upcrossing inequality) The number of upcrossings of $[\alpha,\beta]$ by a submartingale $\{S_n,\mathcal{F}_n\}_1^\infty$ satisfies

$$\mathrm{E}(U_n) \le \frac{\mathrm{E}|S_n| + |\alpha|}{\beta - \alpha}. \tag{16.29}$$

Proof Define $S_n' = \max\{S_n,\alpha\}$, a continuous, convex, non-decreasing function of S_n, such that $\{S_n',\mathcal{F}_n\}$ is an adapted sequence and also a submartingale. U_n is the number of upcrossings up to time n for $\{S_n'\}$ as well as for $\{S_n\}$. Let $X_k' = S_k' - S_{k-1}'$ and so write

$$S_n' - S_1' = \sum_{k=2}^{n} Y_k X_k' + \sum_{k=2}^{n} (1 - Y_k) X_k' \tag{16.30}$$

where Y_k is from (16.28) and X_k' is a submartingale difference. Then

$$\begin{aligned}\mathrm{E}\biggl(\sum_{k=2}^{n}(1 - Y_k)X_k'\biggr) &= \sum_{k=2}^{n}\int_{\{Y_k=0\}} X_k'\mathrm{d}P \\ &= \sum_{k=2}^{n}\int_{\{Y_k=0\}} \mathrm{E}(X_k'|\mathcal{F}_{k-1})\mathrm{d}P \ge 0\end{aligned} \tag{16.31}$$

using the definition of a conditional expectation in the second equality (recalling that Y_k is $\mathcal{F}_{k-1}$ measurable) and the submartingale property to give the inequality. This shows that

$$\mathrm{E}(S_n' - S_1') \ge \mathrm{E}\biggl(\sum_{k=2}^{n} Y_k X_k'\biggr). \tag{16.32}$$

$\sum_{k=2}^{n} Y_k X_k'$ is the sum of the steps made during upcrossings, by definition of Y_k. Since the sum of the X_k' over an upcrossing equals at least $\beta - \alpha$ by definition,

$$\sum_{k=2}^{n} Y_k X_k' \ge (\beta - \alpha) U_n \tag{16.33}$$

where U_n is the number of upcrossings completed by time n. Since $S_1' \ge \alpha$, taking the expectation of (16.33) and substituting from (16.32) gives

$$\begin{aligned}(\beta-\alpha)\mathrm{E}(U_n) \le \mathrm{E}(S_n' - S_1') &\le \mathrm{E}(S_n' - \alpha)\\ &= \int_{\{S_n>\alpha\}} (S_n-\alpha)\mathrm{d}P\\ &\le \mathrm{E}|S_n-\alpha| \le \mathrm{E}|S_n| + |\alpha|. \quad \blacksquare \end{aligned} \tag{16.34}$$

The upcrossing inequality contains the implication that if the sequence is uniformly bounded in L_1, the expected number of upcrossings is finite, even as $n \to \infty$. This is the heart of the convergence proof, for it means that the sequence has to be settling down somewhere beyond a certain point.

Proof of 16.11 Fix α and $\beta > \alpha$. By **16.12**,

$$\mathrm{E}(U_n) \le \frac{\mathrm{E}|S_n| + |\alpha|}{\beta-\alpha} \le \frac{M+|\alpha|}{\beta-\alpha} < \infty. \tag{16.35}$$

For $\omega \in \Omega$, $\{U_n(\omega)\}_1^\infty$ is a positive, non-decreasing sequence and either diverges to $+\infty$ or converges to a finite limit $U(\omega)$ as $n \to \infty$. Divergence for $\omega \in C$ with $P(C) > 0$ would imply $\mathrm{E}(U_n) \to \infty$, which contradicts (16.35), so $U_n \to U$ a.s. where $\mathrm{E}(U) < \infty$.

Define $\bar{S}(\omega) = \limsup_{n\to\infty} S_n$ and $\underline{S}(\omega) = \liminf_{n\to\infty} S_n$. If $\underline{S}(\omega) < \alpha < \beta < \bar{S}(\omega)$, the sequence visits points both above β and below α an infinite number of times as $n \to \infty$ and hence the interval $[\alpha, \beta]$ is crossed an infinite number of times. It therefore must be the case that $P(\underline{S} < \alpha < \beta < \bar{S}) = 0$. This is true for any pair α, β. Hence consider

$$\{\omega : \underline{S}(\omega) < \bar{S}(\omega)\} = \bigcup_{\alpha,\beta} \{\underline{S} \le \alpha < \beta \le \bar{S}\} \tag{16.36}$$

where the union on the right is taken over rational values of α and β. Evidently, $P(\underline{S} < \bar{S}) = 0$ by **3.12**(ii), which is the same as $\underline{S} = \bar{S} = S_\infty$ a.s. where S_∞ is the limit of $\{S_n\}$. Finally, note that

$$\mathrm{E}|S_\infty| \le \liminf_{n\to\infty} \mathrm{E}|S_n| \le \sup_n \mathrm{E}|S_n| \le M \tag{16.37}$$

where the first inequality is from Fatou's lemma and the last is by assumption. This completes the proof. ∎

16.13 Corollary Let $\{S_n, \mathcal{F}_n\}_{-\infty}^{\infty}$ be a doubly infinite martingale. Then $S_n \to S_{-\infty}$ a.s. as $n \to -\infty$, where $S_{-\infty}$ is an L_1-bounded r.v.

Proof Let U_{-n} denote the number of upcrossings of $[\alpha,\beta]$ performed by the sequence $\{S_j : -1 \geq j \geq -n\}$. The argument of **16.12** shows that

$$\mathrm{E}(U_{-n}) \leq \frac{\mathrm{E}|S_1| + |\alpha|}{\beta - \alpha}, \text{ all } n \geq 1. \tag{16.38}$$

Arguments precisely analogous to those of **16.11** show that

$$P\Big(\liminf_{n\to-\infty} S_n < \limsup_{n\to-\infty} S_n\Big) = 0 \tag{16.39}$$

so that the limit $S_{-\infty}$ exists a.s. The sequence $\{\mathrm{E}|S_n|\}_{-\infty}^{-1}$ is non-negative, non-increasing as n decreases by (16.27) and $\mathrm{E}|S_{-1}| < \infty$ by definition of a martingale. Hence $\mathrm{E}|S_{-\infty}| < \infty$. ■

Of the examples quoted earlier, the non-convergent case **16.1** does not satisfy the conditions of **16.11**. A random walk does not converge but wanders forever with a variance that is an increasing function of time. However, be careful to appreciate that 'non-convergence' does not mean convergence in $\bar{\mathbb{R}}$, or in other words heading off to $+\infty$ or $-\infty$ never to return. These are events that occur only with probability 0. The next result shows, subject to the increments having a suitably bounded distribution, that a non-convergent martingale eventually visits all regions of the real line, almost surely.

16.14 Theorem Let $\{X_t, \mathcal{F}_t\}$ be a m.d. sequence with $\mathrm{E}(\sup_t |X_t|) < \infty$ and let $S_n = \sum_{t=1}^n X_t$. If $C = \{\omega : S_n(\omega) \text{ converges}\}$ and

$$E = \big\{\omega : \text{ either } \inf_n S_n(\omega) > -\infty \text{ or } \sup_n S_n(\omega) < \infty\big\} \tag{16.40}$$

then $P(E - C) = 0$.

Proof For a constant $M > 0$, define the stopping time $\tau_M(\omega)$ as the smallest integer n such that $S_n(\omega) > M$ if one exists and $\tau_M(\omega) = \infty$ otherwise. The stopped process $\{S_{n\wedge\tau_M}, \mathcal{F}_n\}_{n=1}^{\infty}$ is a martingale (**16.10**) and $S_{(n-1)\wedge\tau_M} \leq M$ for all n. By **16.11**, $S_{n\wedge\tau_M}$ converges a.s. if $\sup_n \mathrm{E}|S_{n\wedge\tau_M}| < \infty$. To show this let $S_n^+ = \max\{S_n, 0\}$ and $X_n^+ = \max\{X_n, 0\}$ and note that

$$S_{n\wedge\tau_M}^+ \leq S_{n-1\wedge\tau_M}^+ + X_{n\wedge\tau_M}^+ \leq M + \sup_n |X_n|. \tag{16.41}$$

Because $\mathrm{E}(S_{n\wedge\tau_M}) = 0$, $\mathrm{E}|S_{n\wedge\tau_M}| = 2\mathrm{E}(S^+_{n\wedge\tau_M})$ so that $\sup_n \mathrm{E}|S_{n\wedge\tau_M}|$ is bounded according to (16.41) and the assumption. Since $S_{n\wedge\tau_M}(\omega) = S_n(\omega)$ on the set $\{\omega : \sup_n S_n(\omega) \le M\}$, $S_n(\omega)$ converges a.s. on the same set. Letting $M \to \infty$ yields the result that $S_n(\omega)$ converges a.s. on the set $\{\omega : \sup_n S_n(\omega) < \infty\}$.

Now, applying the same argument to $-S_n$ yields the matching conclusion that $S_n(\omega)$ converges a.s. on the set $\{\omega : \inf_n S_n(\omega) > -\infty\}$ and hence, combining the two conclusions, on the set E. The implication is that $P(C \cap E) = P(E)$ which is equivalent to the stated result. ∎

Note that $E^c = \{\omega : \sup_n S_n(\omega) = +\infty \text{ and } \inf_n S_n(\omega) = -\infty\}$. Since $P(E^c) = P((E \cap C)^c) = P(E^c \cup C^c)$, a direct consequence of the theorem is that $C^c \subseteq E^c \cup N$ where $P(N) = 0$. This is precisely the claim made above, that nonconvergent martingales are elements of E^c with probability 1.

A remaining question of interest is the nature of convergence in the product martingale of Example **16.2**. It turns out there are two possibilities depending on the distribution of the increments, either convergence to 1, the expected value of the increments, or convergence to 0. The following is known as Kakutani's theorem (see e.g. [190] 14.12).

16.15 Theorem In the sequence defined in **16.2** let $a_t = \mathrm{E}\left(X_t^{1/2}\right)$ and $A_n = \prod_{t=1}^n a_t$.

(i) If $A_\infty > 0$ then $S_n \to S_\infty$ a.s. and $\mathrm{E}(S_\infty) = 1$.

(ii) If $A_\infty = 0$, then $S_n \to 0$ a.s.

Proof The sequence meets the conditions of Theorem **16.11** and so converges a.s. to a limit S_∞. Define $N_n = S_n^{1/2}/A_n$. This is another product martingale with independent increments $X_t^{1/2}/a_t$ having expectation 1 and hence also converges a.s.

Since $a_t \le \mathrm{E}(X_t)^{1/2} = 1$ by **9.19**, A_n is a non-increasing sequence. If $A_\infty > 0$ then

$$\sup_n \mathrm{E}(N_n^2) = \sup_n \mathrm{E}(S_n)A_n^{-2} = \sup_n A_n^{-2} \le A_\infty^{-2} < \infty \tag{16.42}$$

and therefore

$$\mathrm{E}\left(\sup_n |S_n|\right) = \mathrm{E}\left(\sup_n N_n^2 A_n^2\right) \le \mathrm{E}\left(\sup_n N_n^2\right) \le 4 \sup_n \mathrm{E}(N_n^2) < \infty \tag{16.43}$$

where the second inequality applies Doob's inequality (see **16.21** below). Since $\sup_n |S_n|$ is a dominating function for S_n, this shows by **12.11** that the sequence $\{S_n, n \in \mathbb{N}\}$ is uniformly integrable and hence by **12.8** that $\mathrm{E}(S_n) \to 1$. This proves (i).

On the other hand if $A_\infty = 0$, since N_n converges a.s. the first equality of (16.42) shows that $S_n \to 0$ a.s., proving (ii). ∎

16.4 Convergence and the Conditional Variances

If $\{S_n\}$ is a square-integrable martingale with differences $\{X_n\}$ then

$$\mathrm{E}(S_n^2|\mathcal{F}_{n-1}) = \mathrm{E}(S_{n-1}^2 + X_n^2 + 2X_nS_{n-1}|\mathcal{F}_{n-1}) \geq S_{n-1}^2 \text{ a.s.}$$

showing that S_n^2 is a submartingale. The Doob decomposition of the sequence of squares has the form $S_n^2 = M_n + \langle S\rangle_n$ where $M_n - M_{n-1} = X_n^2 + 2X_nS_{n-1} - \mathrm{E}(X_n^2|\mathcal{F}_{n-1})$ and $\langle S\rangle_n - \langle S\rangle_{n-1} = \mathrm{E}(X_n^2|\mathcal{F}_{n-1})$. The sequence $\{\langle S\rangle_n\}$, being the partial sum of the conditional increment variances, is called the *quadratic variation* of $\{S_n\}$. The following theorem reveals an intimate link between martingale convergence and the summability of these conditional variances; the latter property implies the former almost surely. If the quadratic variation converges so that $\sum_{t=1}^\infty \mathrm{E}(X_t^2|\mathcal{F}_{t-1}) < \infty$ a.s., then $S_n \to S$ a.s.

16.16 Theorem Let $\{X_t, \mathcal{F}_t\}_1^\infty$ be a m.d. sequence and $S_n = \sum_{t=1}^n X_t$. If

$$D = \left\{\omega : \sum_{t=1}^\infty \mathrm{E}(X_t^2|\mathcal{F}_{t-1})(\omega) < \infty\right\} \in \mathcal{F}$$
$$C = \{\omega : S_n(\omega) \text{ converges}\} \in \mathcal{F}$$

then $P(D - C) = 0$.

Proof Fix $M > 0$ and define the stopping time $\tau_M(\omega)$ as the smallest value of n having the property

$$\sum_{t=1}^n \mathrm{E}(X_t^2|\mathcal{F}_{t-1})(\omega) \geq M. \tag{16.44}$$

If there is no finite integer with this property then $\tau_M(\omega) = \infty$. If $D_M = \{\omega : \tau_M(\omega) = \infty\}$, $D = \lim_{M\to\infty} D_M$. The r.v. $1_{\{\tau_M \geq n\}}(\omega)$ is $\mathcal{F}_{n-1}$-measurable since it is known at time $n-1$ whether or not the inequality in (16.44) is true. The stopped process

$$S_{n\wedge\tau_M} = \sum_{t=1}^n X_t 1_{\{\tau_M \geq t\}} \tag{16.45}$$

is a martingale by **16.10**. The increments are orthogonal and

$$\begin{aligned}\sup_n \mathrm{E}(S^2_{n\wedge\tau_M}) &= \sup_n \mathrm{E}\Big(\sum_{t=1}^n X_t^2 1_{\{\tau_M \geq t\}}\Big)\\ &= \sup_n \mathrm{E}\Big(\sum_{t=1}^n \mathrm{E}(X_t^2|\mathcal{F}_{t-1}) 1_{\{\tau_M \geq t\}}\Big) < M \qquad (16.46)\end{aligned}$$

where the final inequality holds for the expectation since it holds for each $\omega \in \Omega$ by definition of $\tau_M(\omega)$. By Liapunov's inequality,

$$\sup_n \mathrm{E}|S_{n\wedge\tau_M}| \leq \sup_n \|S_{n\wedge\tau_M}\|_2 < M^{1/2}$$

and hence $S_{n\wedge\tau_M}$ converges a.s. by **16.11**. If $\omega \in D_M$, $S_{n\wedge\tau_M}(\omega) = S_n(\omega)$ for every $n \in \mathbb{N}$ and hence $S_n(\omega)$ converges, except for ω in a set of zero measure. That is, $P(D_M \cap C) = P(D_M)$. The theorem follows on taking the complement and letting $M \to \infty$. ∎

16.17 Example To get an idea of what convergence entails, consider the case of $\{X_t\}$ an i.i.d. sequence with mean 0 and variance $\sigma^2 < \infty$ (compare **16.1**). Then $\{X_t/a_t\}$ is a m.d. sequence for any sequence $\{a_t\}$ of positive constants. Since $\mathrm{E}(X_t^2|\mathcal{F}_{t-1}) = \mathrm{E}(X_t^2) = \sigma^2$, $S_n = \sum_{t=1}^n X_t/a_t$ is an a.s. convergent martingale whenever $\sum_{t=1}^\infty 1/a_t^2 < \infty$. For example, $a_t = t$ would satisfy the requirement. □

In the almost sure case of Theorem **16.16** (when $P(C) = P(D) = 1$) the summability of the conditional variances transfers to that of the ordinary variances, $\sigma_t^2 = \mathrm{E}(X_t^2)$. Also when $\mathrm{E}(\sup_t X_t^2) < \infty$, the summability of the conditional variances is almost equivalent to the summability of the X_t^2 themselves. These are consequences of the following pair of useful results.

16.18 Theorem Let $\{Z_t\}$ be any non-negative stochastic sequence.

(i) $\sum_{t=1}^\infty \mathrm{E}(Z_t) < \infty$ iff $\sum_{t=1}^\infty \mathrm{E}(Z_t|\mathcal{F}_{t-1}) < \infty$ a.s.

(ii) If $\mathrm{E}(\sup_t Z_t) < \infty$ then $P(D\Delta E) = 0$ where

$$\begin{aligned}D &= \Big\{\omega : \sum\nolimits_{t=1}^\infty \mathrm{E}(Z_t|\mathcal{F}_{t-1})(\omega) < \infty\Big\}\\ E &= \Big\{\omega : \sum\nolimits_{t=1}^\infty Z_t(\omega) < \infty\Big\}.\end{aligned}$$

Proof (i) The first of the sums is the expected value of the second, so the 'only if' part is immediate. Since $\mathrm{E}(Z_t|\mathcal{F}_{t-1})$ is undefined unless $\mathrm{E}(Z_t) < \infty$, assume $\sum_{t=1}^n \mathrm{E}(Z_t) < \infty$ for each finite n. These partial sums form a monotone series which either converges to a finite limit or diverges to $+\infty$. Suppose $\sum_{t=1}^{n} \mathrm{E}(Z_t|\mathcal{F}_{t-1})$ converges a.s. implying, by the Cauchy criterion, that $\sum_{t=n+1}^{n+m} \mathrm{E}(Z_t|\mathcal{F}_{t-1}) \to 0$ a.s. as $m \wedge n \to \infty$. Then $\sum_{t=n+1}^{n+m} \mathrm{E}(Z_t) \to 0$ by the monotone convergence theorem, so that by the same criterion $\sum_{t=1}^{n} \mathrm{E}(Z_t) \to \sum_{t=1}^{\infty} \mathrm{E}(Z_t) < \infty$ as required.

(ii) Define the m.d. sequence $X_t = Z_t - \mathrm{E}(Z_t|\mathcal{F}_{t-1})$ and let $S_n = \sum_{t=1}^{n} X_t$. Clearly, $\sup_n S_n(\omega) \le \sum_{t=1}^{\infty} Z_t(\omega)$ and if the majorant side of this inequality is finite, $S_n(\omega)$ converges in almost every case, by the assumption and **16.14**. Given the definition of X_t, this implies in turn that $\sum_{t=1}^{\infty} \mathrm{E}(Z_t|\mathcal{F}_{t-1})(\omega) < \infty$. In other words, $P(E - D) = 0$. Now apply the same argument to $-X_t = \mathrm{E}(Z_t|\mathcal{F}_{t-1}) - Z_t$ to show that the reverse implication holds almost surely and $P(D - E) = 0$ also. ∎

16.5 Martingale Inequalities

Of the many interesting results that can be proved for martingales, certain inequalities are essential tools of limit theory. Of particular importance are *maximal inequalities,* which place bounds on the extreme behaviour a sequence is capable of over a succession of steps. There are two related results of this type, the first a sophisticated cousin of the Markov inequality.

16.19 Theorem If $\{S_n, \mathcal{F}_n\}_1^{\infty}$ is an L_p-bounded martingale for $p \ge 1$, then

$$P\Big(\max_{1\le k\le n} |S_k| > \varepsilon\Big) \le \frac{\mathrm{E}(|S_n|^p 1_{\{\max_{1\le k\le n}|S_k|>\varepsilon\}})}{\varepsilon^p}. \tag{16.47}$$

Proof Define the disjoint collection of events $A_1, \ldots, A_n$ where $A_1 = \big\{\omega : |S_1(\omega)| > \varepsilon\big\}$ and for $k = 2, \ldots, n$,

$$A_k = \Big\{\omega : \max_{1\le j<k} |S_j(\omega)| \le \varepsilon,\ |S_k(\omega)| > \varepsilon\Big\} \in \mathcal{F}_k$$

with union

$$A = \bigcup_{k=1}^{n} A_k = \Big\{\max_{1\le k\le n} |S_k| > \varepsilon\Big\}. \tag{16.48}$$

Since $A_k \subseteq \{|S_k| > \varepsilon\}$, the generalized Markov inequality (**9.17**) gives

$$P(A_k) \le \varepsilon^{-p}\mathrm{E}(|S_k|^p 1_{A_k}). \tag{16.49}$$

By **16.9**, $|S_n|^p$ for $p \geq 1$ is a submartingale so $|S_k|^p \leq \mathrm{E}(|S_n|^p|\mathcal{F}_k)$ a.s., for $1 \leq k \leq n$. Since $A_k \in \mathcal{F}_k$, it follows that

$$\mathrm{E}(|S_k|^p 1_{A_k}) \leq \mathrm{E}\big(\mathrm{E}(|S_n|^p|\mathcal{F}_k)1_{A_k}\big) = \mathrm{E}(|S_n|^p 1_{A_k}) \tag{16.50}$$

where the equality applies (10.18). Noting from (16.48) that $1_A = \sum_{k=1}^{n} 1_{A_k}$, (16.48)–(16.50) imply

$$P(A) = \sum_{k=1}^{n} P(A_k) \leq \frac{\sum_{k=1}^{n} \mathrm{E}(|S_n|^p 1_{A_k})}{\varepsilon^p} = \frac{\mathrm{E}(|S_n|^p 1_A)}{\varepsilon^p}. \quad \blacksquare \tag{16.51}$$

The following immediate corollary of **16.19** is the version of the result most often quoted in applications.

16.20 Corollary If $\{S_n, \mathcal{F}_n\}_1^\infty$ is an L_p-bounded martingale for $p \geq 1$, then

$$P\Big(\max_{1\leq k\leq n} |S_k| > \varepsilon\Big) \leq \frac{\mathrm{E}(|S_n|^p)}{\varepsilon^p}. \quad \square \tag{16.52}$$

16.20 was originally proved by Kolmogorov for the case where $\{X_t\}$ is independent rather than a m.d. and in this form is known as *Kolmogorov's inequality.*

The second key result is *Doob's inequality*, converting the probability bound of **16.20** into a moment inequality as follows.

16.21 Theorem If $\{S_n, \mathcal{F}_n\}_1^\infty$ is an L_p-bounded martingale for $p > 1$, then

$$\mathrm{E}|S_n|^p \leq \mathrm{E}\Big(\max_{1\leq k\leq n} |S_k|^p\Big) \leq \Big(\frac{p}{p-1}\Big)^p \mathrm{E}|S_n|^p. \quad \square \tag{16.53}$$

The downside inequality of (16.53) is obvious and is included in the statement for completeness. The proof of the upside inequality uses the following ingenious lemma.

16.22 Lemma Let X and Y be non-negative r.v.s. If

$$P(X > \varepsilon) \leq \varepsilon^{-1}\mathrm{E}(Y 1_{\{X>\varepsilon\}})$$

for all $\varepsilon > 0$, then for $p > 1$

$$\mathrm{E}(X^p) \leq \Big(\frac{p}{p-1}\Big)^p \mathrm{E}(Y^p).$$

Proof By Corollary **9.22**,

$$\mathrm{E}(X^p) = \int_0^\infty p\xi^{p-1}P(X > \xi)\mathrm{d}\xi. \tag{16.54}$$

Let F_{XY} denote the joint c.d.f. of X and Y and write $(\mathbb{R}^2)^+$ for $[0,\infty)\times[0,\infty)$, the non-negative orthant of $\mathbb{R}^2$. Substituting the assumption of the lemma into (16.54) gives

$$\begin{aligned}\mathrm{E}(X^p) &\le p\int_0^\infty \xi^{p-2}\mathrm{E}(Y1_{\{X>\xi\}})\mathrm{d}\xi\\ &= p\int_0^\infty \xi^{p-2}\left(\int_{(\mathbb{R}^2)^+} y1_{\{x>\xi\}}\mathrm{d}F_{XY}(x,y)\right)\mathrm{d}\xi\\ &= p\int_{(\mathbb{R}^2)^+} y\left(\int_0^x \xi^{p-2}\mathrm{d}\xi\right)\mathrm{d}F_{XY}(x,y)\\ &= \frac{p}{p-1}\int_{(\mathbb{R}^2)^+} yx^{p-1}\mathrm{d}F_{XY}(x,y)\\ &= \frac{p}{p-1}\mathrm{E}(YX^{p-1}). \end{aligned} \tag{16.55}$$

The second equality is permitted by Tonelli's theorem, noting that the function $F_{XY}\xi$ defines a σ-finite product measure on $(\mathbb{R}^3)^+$. By Hölder's inequality,

$$\mathrm{E}(YX^{p-1}) \le \left(\mathrm{E}(Y^p)\right)^{1/p}\left(\mathrm{E}(X^p)\right)^{1-1/p}.$$

Substituting into the majorant side of (16.55) and simplifying gives the result. ∎

Proof of 16.21 Consider (16.51) for the case $p = 1$, that is,

$$P\left(\max_{1\le k\le n}|S_k| > \varepsilon\right) \le \varepsilon^{-1}\mathrm{E}(|S_n|1_{\{\max_{1\le k\le n}|S_k|>\varepsilon\}}). \tag{16.56}$$

To complete the proof apply **16.22** to (16.56), putting $X = \max_{1\le k\le n}|S_k|$ and $Y = |S_n|$, to yield the upside inequality of (16.53). ∎

The orthogonality of the differences implies the interesting property of an L_2-bounded martingale $\{S_n\}_1^\infty$ that

$$\mathrm{E}(S_n^2) = \mathrm{E}\left(\sum_{t=1}^n X_t^2\right) \tag{16.57}$$

where, with $S_0 = 0$, $X_t = S_t - S_{t-1}$. This lets (16.52) and (16.53) for the case $p = 2$ be extended to link $P(\max_{1\le k\le n}|S_k| > \varepsilon)$ and $\mathrm{E}(\max_{1\le k\le n} S_k^2)$ directly with the variance of the increments. It would be most useful if this type of property extended to other values of p, in particular for $p < 2$.

One approach to this problem is the von Bahr–Esséen inequality of §11.7. Obviously, **11.21** has a direct application to martingales.

16.23 Theorem If $\{X_t, \mathcal{F}_t\}_1^\infty$ is an L_p-bounded m.d. sequence and $S_n = \sum_{t=1}^n X_t$,

$$\mathrm{E}|S_n|^p \le 2\sum_{t=1}^n \mathrm{E}|X_t|^p \tag{16.58}$$

for $0 < p \le 2$.

Proof This is by iterating **11.21** with $Y = X_n$, $\mathcal{G} = \mathcal{F}_{n-1}$ and $X = S_{n-1}$, as in the argument leading to (11.85); note that the latter holds for m.d. sequences just as for independent sequences. ∎

Another route to this type of result goes via the *Burkholder inequalities* ([28], [29]). The following result extends the well-known Marcinkiewicz–Zygmund inequality for independent sequences ([133]) to martingale differences. The scene is set by defining what is sometimes called the martingale square function, that is, the length of the vector of increments. For a martingale S_n with $S_0 = 0$ and increments $X_t = S_t - S_{t-1}$ for $t = 1, \dots, n$, the square function will be written compactly using the notation

$$Q(S_n) = \left(\sum_{t=1}^n X_t^2\right)^{1/2}.$$

The Burkholder result shows that a martingale grows with n in L_p-norm at the same rate as its square function, and this holds for any $p > 1$ such that the L_p-norm exists.

16.24 Theorem (*Burkholder*) If $\{S_n, \mathcal{F}_n\}_1^\infty$ is an L_p-bounded martingale for $p > 1$,

$$c_p\|Q(S_n)\|_p \le \|S_n\|_p \le C_p\|Q(S_n)\|_p \tag{16.59}$$

where $c_p = (18p^{3/2}/(p-1))^{-1}$ and $C_p = 18p^{3/2}/(p-1)^{1/2}$. □

The proof makes use of two lemmas, as follows.

16.25 Lemma Let $\{T_t,\mathcal{F}_t\}_{t=1}^n$ be an L_1-bounded non-negative submartingale. For $\lambda > 0$ define the stopping time $\nu = \min\{t \le n : T_t > \lambda\}$ if such a t exists and $\nu = n+1$ otherwise. Then

$$\mathrm{E}\big(Q(T_{\nu-1})^2\big) \le 2\lambda \mathrm{E}(T_n). \tag{16.60}$$

Proof Let $T_{n+1} = T_n$. For any $m \le n+1$ it can be verified that

$$Q(T_{m-1})^2 = 2T_mT_{m-1} - T_{m-1}^2 - 2\sum_{t=2}^{m} T_{t-1}(T_t - T_{t-1}) \tag{16.61}$$

where

$$\mathrm{E}\Bigg(\sum_{t=2}^{m} T_{t-1}(T_t - T_{t-1})\Bigg) = \mathrm{E}\Bigg(\sum_{t=2}^{m} T_{t-1}(\mathrm{E}(T_t|\mathcal{F}_{t-1}) - T_{t-1})\Bigg) \ge 0.$$

Setting $m = \nu$, since $T_{\nu-1} \le \lambda$, $\mathrm{E}(T_\nu T_{\nu-1}) \le \lambda \mathrm{E}(T_\nu) \le \lambda \mathrm{E}(T_n)$. Hence taking the expectation of (16.61) gives $\mathrm{E}(Q(T_{\nu-1})^2) \le 2\lambda\mathrm{E}(T_n) - \mathrm{E}(T_{\nu-1}^2)$ and (16.60) follows. ∎

16.26 Lemma If $\{R_t,\mathcal{F}_t\}_{t=1}^n$ is an L_p-bounded non-negative submartingale for $1 < p < \infty$,

$$\|Q(R_n)\|_p \le 9\frac{p^{3/2}}{p-1}\|R_n\|_p. \tag{16.62}$$

Proof For brevity, set $R_n^* = \max_{1\le t\le n} R_t$. For some $\theta > 0$ define

$$Y = \max(\theta Q(R_n), R_n^*). \tag{16.63}$$

Then for $\beta = (1+2\theta^2)^{1/2} > 1$ and $\lambda > 0$, consider the event

$$B = \{\theta Q(R_n) > \beta\lambda, R_n^* \le \lambda\}.$$

Note that $\{Y > \beta\lambda\} \subseteq B \cup \{R_n^* > \lambda\}$, hence by subadditivity

$$P(Y > \beta\lambda) \le P(B) + P(R_n^* > \lambda). \tag{16.64}$$

Define the sequence $\{T_t,\mathcal{F}_t\}_{t=1}^n$ by $T_t = R_t 1_{\{\theta Q(R_t)>\lambda\}}$. T_t is a non-negative submartingale since

$$\begin{aligned}\mathrm{E}(T_t|\mathcal{F}_{t-1}) &\geq \mathrm{E}(R_t 1_{\{\theta Q(R_{t-1})>\lambda\}}|\mathcal{F}_{t-1})\\ &= \mathrm{E}(R_t|\mathcal{F}_{t-1})1_{\{\theta Q(R_{t-1})>\lambda\}}\\ &\geq R_{t-1}1_{\{\theta Q(R_{t-1})>\lambda\}} = T_{t-1}.\end{aligned}$$

Also define the stopping time $\tau = \min\{t \leq n : \theta Q(R_t) > \lambda\}$ if $\theta Q(R_n) > \lambda$ and $\tau = n$ otherwise. Let $U_1 = R_1$ and $U_t = R_t - R_{t-1}$ for $t = 2, \dots, n$. When the event B has occurred, $|U_\tau| \leq \max\{R_{\tau-1}, R_\tau\} \leq \lambda$. Hence, B implies that

$$\begin{aligned}\lambda^2 + 2\theta^2\lambda^2 = \beta^2\lambda^2 &< \theta^2 Q(R_n)^2\\ &= \theta^2\sum_{t=1}^{\tau-1} U_t^2 + \theta^2 U_\tau^2 + \theta^2\sum_{t=\tau+1}^{n} U_t^2\\ &\leq \lambda^2 + \theta^2\lambda^2 + \theta^2 Q(T_n)^2,\end{aligned}$$

which, after simplification and cancellation, gives $Q(T_n) > \lambda$. Thus, also noticing that $\{R_n^* \leq \lambda\} \subseteq \{T_n^* \leq \lambda\}$ where $T_n^* = \max_{1\leq t\leq n} T_t$,

$$B \subseteq C = \{Q(T_n) > \lambda, T_n^* \leq \lambda\}.$$

Hence, by **9.17**,

$$\begin{aligned}\lambda P(B) \leq \lambda P(C) &\leq \lambda^{-1}\mathrm{E}\big(Q(T_n)^2 1_{\{T_n^*\leq\lambda\}}\big)\\ &\leq 2\mathrm{E}(T_n)\\ &\leq 2\mathrm{E}\big(R_n 1_{\{Y>\lambda\}}\big) \qquad (16.65)\end{aligned}$$

where the third inequality is by Lemma **16.25** in view of the fact that if $T_n^* \leq \lambda$ then the condition $\nu = n + 1$ applies. The last inequality of (16.65) uses the fact that $\{Y > \lambda\} = \{\theta Q(R_n) > \lambda\} \cup \{R_n^* > \lambda\}$, and hence $T_n \leq R_n 1_{\{Y>\lambda\}}$ by construction. Also, by **16.19**,

$$\lambda P(R_n^* > \lambda) \leq \mathrm{E}(R_n 1_{\{R_n^*>\lambda\}}) \leq \mathrm{E}(R_n 1_{\{Y>\lambda\}}). \qquad (16.66)$$

Putting together (16.64) with (16.65) and (16.66) produces

$$\lambda P(Y > \beta\lambda) \leq 3\mathrm{E}\big(R_n 1_{\{Y>\lambda\}}\big). \qquad (16.67)$$

Applying Corollary **9.22** with a change of variable and substitution from (16.67) now gives

$$\begin{aligned}
\mathrm{E}(Y^p) &= p\beta^p \int_0^\infty \lambda^{p-1} P(Y > \beta\lambda)\mathrm{d}\lambda \\
&\le 3p\beta^p \int_0^\infty \lambda^{p-2}\mathrm{E}(R_n 1_{\{Y>\lambda\}})\mathrm{d}\lambda \\
&= 3p\beta^p \mathrm{E}\Big(R_n \int_0^Y \lambda^{p-2}\mathrm{d}\lambda\Big) \\
&= 3\frac{p}{p-1}\beta^p \mathrm{E}(R_n Y^{p-1}) \\
&\le 3\frac{p}{p-1}\beta^p \|R_n\|_p \|Y\|_p^{p-1}. \qquad (16.68)
\end{aligned}$$

The second equality of (16.68) applies Tonelli's theorem **4.26** and the final inequality is by the Hölder inequality. Hence by (16.63),

$$\theta\|Q(R_n)\|_p \le \|Y\|_p \le 3\frac{p}{p-1}\beta^p\|R_n\|_p,$$

where the second inequality is obtained by rearrangement and cancellation in (16.68). Finally set $\theta = p^{-1/2}$ so that $\beta^p = (1 + 2/p)^{p/2} < 3$ for any value of $p > 1$ (being bounded above by 'e', note) to give (16.62). ∎

Proof of 16.24 Define the non-negative sequences S_n^+ and S_n^- to be the positive and negative parts of S_n, such that $S_n = S_n^+ - S_n^-$. The sequences $\{\mathrm{E}(S_n^+|\mathcal{F}_t)\}_{t=1}^n$ and $\{\mathrm{E}(S_n^-|\mathcal{F}_t)\}_{t=1}^n$ are martingales, and are non-negative. Define $U_t = \mathrm{E}(S_n^+|\mathcal{F}_t) - \mathrm{E}(S_n^+|\mathcal{F}_{t-1})$ and $V_t = \mathrm{E}(S_n^-|\mathcal{F}_t) - \mathrm{E}(S_n^-|\mathcal{F}_{t-1})$ for $t = 2, \dots, n$ with $U_1 = \mathrm{E}(S_n^+|\mathcal{F}_1)$ and $V_1 = \mathrm{E}(S_n^-|\mathcal{F}_1)$. Since $S_n = X_1 + \cdots + X_n$ is a martingale it is immediate that $U_t - V_t = X_t$. By **2.24**, $Q(S_n) \le Q(S_n^+) + Q(S_n^-)$ and hence, by **9.31** and Lemma **16.26** setting successively $R_n = S_n^+$ and $R_n = S_n^-$,

$$\begin{aligned}
\|Q(S_n)\|_p &\le \|Q(S_n^+)\|_p + \|Q(S_n^-)\|_p \\
&\le 9\frac{p^{3/2}}{p-1}\big(\|S_n^+\|_p + \|S_n^-\|_p\big) \\
&\le 18\frac{p^{3/2}}{p-1}\|S_n\|_p. \qquad (16.69)
\end{aligned}$$

This establishes the downside inequality of (16.59). For the upside inequality, define the sequence $\{R_n, \mathcal{F}_n\}_1^\infty$ where

$$R_n = \frac{\operatorname{sgn}(S_n)|S_n|^{p-1}}{||S_n||_p^{p-1}}.$$

The sequence $\{\mathrm{E}(R_n|\mathcal{F}_t)\}_{t=1}^n$ is a martingale. Let $W_1 = \mathrm{E}(R_n|\mathcal{F}_1)$ and $W_t = \mathrm{E}(R_n|\mathcal{F}_t) - \mathrm{E}(R_n|\mathcal{F}_{t-1})$ for $t = 2, \dots, n$. Note that $||S_n||_p = \mathrm{E}(S_nR_n)$ and hence

$$\begin{aligned} ||S_n||_p = \mathrm{E}\Big(\sum_{t=1}^n X_tW_t\Big) &\leq \mathrm{E}\big(Q(S_n)Q(R_n)\big) \\ &\leq ||Q(S_n)||_p||Q(R_n)||_{p/(p-1)}, \end{aligned} \tag{16.70}$$

where the first inequality uses **2.22** with $p = 2$ and the second one is by **9.29**. Applying (16.69) with S_n standing for R_n and $p/(p-1)$ replacing p and also using the fact that $||R_n||_{p/(p-1)} = 1$ gives

$$\begin{aligned} ||Q(R_n)||_{p/p-1} &\leq 18\frac{(p/(p-1))^{3/2}}{p/(p-1)-1}||R_n||_{p/p-1} \\ &= 18p\Big(\frac{p}{p-1}\Big)^{1/2}. \end{aligned} \tag{16.71}$$

Substituting (16.71) into (16.70) gives the upside inequality of (16.59). ∎

16.27 Corollary For $p > 1$,

$$c_p||Q(S_n)||_p \leq ||\max_{1\leq t\leq n}|S_t|||_p \leq C_p^*||Q(S_n)||_p \tag{16.72}$$

where $c_p = (18p^{3/2}/(p-1))^{-1}$ and $C_p^* = 18p^{5/2}/(p-1)^{3/2}$.

Proof This is by application of the Doob inequality (16.53) to the upside of (16.59). ∎

The upside inequality of **16.24** is generally the more important in applications. Provided the martingale is L_2-bounded this inequality can be extended to $p \leq 1$ although with a different constant factor, a helpful fact in particular because it includes the case $p = 1$. The following inequality for $0 < p \leq 2$ is adapted from Novikov's [138] inequality for Brownian motion stochastic integrals, and hence imposes L_2-boundedness.

16.28 Theorem Let $\{S_n,\mathcal{F}_n\}_1^\infty$ be an L_2-bounded martingale with $S_0=0$ and let $Q(S_n)=(\sum_{t=1}^n X_t^2)^{1/2}$ for $X_t=S_t-S_{t-1}$. For $0<p\le 2$,

$$\|S_n\|_p \le C_p\|Q(S_n)\|_p \tag{16.73}$$

for $C_p<\infty$ depending only on p.

Proof For the case $p=2$ the inequality of (16.73) can be set to equality as (16.57) with $C_2=1$, to reproduce the orthogonality property of the martingale. Hence, consider $p<2$. The following argument applies except in one case, that where $X_1=\cdots=X_{n-1}=0$ a.s. In this case $\|S_n\|_p=\|Q(S_n)\|_p=\|X_n\|_p$ and (16.73) holds as an equality with $C_p=1$.

For convenience of notation write $m=p/2<1$ and also let Q_n stand for $Q(S_n)$. Let $B_n=2\sum_{t=2}^n\sum_{s=1}^{t-1}X_sX_t$ so that $S_n^2=Q_n^2+B_n$. Then for $\delta>0$ and $\mu>0$ to be chosen and $n\ge 1$ define

$$Y_n=\delta(\mu+Q_n^2)+S_n^2=\delta\mu+(1+\delta)Q_n^2+B_n. \tag{16.74}$$

With $m<1$ the function x^m is concave so that $2^{m-1}(x^m+y^m)\le(x+y)^m$ for $x,y\ge 0$. Apply this formula to the first equality of (16.74) and take expectations to give the inequality

$$2^{m-1}\big(\delta^m\mathrm{E}(\mu+Q_n^2)^m+\mathrm{E}|S_n|^{2m}\big)\le\mathrm{E}(Y_n^m). \tag{16.75}$$

Next write the telescoping sum

$$Y_n^m=Y_1^m+\sum_{t=2}^n(Y_t^m-Y_{t-1}^m). \tag{16.76}$$

Letting the second equality of (16.74) define Y_t, note that $Q_t^2-Q_{t-1}^2=X_t^2$ and so

$$Y_t-Y_{t-1}=(1+\delta)X_t^2+\Delta B_t$$

where $\Delta B_t=2\sum_{s=1}^{t-1}X_sX_t$. Taylor expansions of the terms in (16.76) to second order for $t=2,\dots,n$ yield

$$\begin{aligned}Y_t^m-Y_{t-1}^m&=mY_{t-1}^{m-1}\big((1+\delta)X_t^2+\Delta B_t\big)\\&\quad+\tfrac{1}{2}m(m-1)\big(Y_{t-1}+\theta_t((1+\delta)X_t^2+\Delta B_t)\big)^{m-2}\\&\qquad\times((1+\delta)X_t^2+\Delta B_t)^2\end{aligned} \tag{16.77}$$

with $\theta_t \in [0,1]$. Since $m < 1$ the second-order term in (16.77) is non-positive. For the case $t = 1$, noting that $Q_1^2 = S_1^2 = X_1^2$ and $Q_0 = B_0 = B_1 = 0$, write

$$Y_1^m = \left(\delta\mu + (1+\delta)X_1^2\right)^m = mY_0^{m-1}(1+\delta)X_1^2 \tag{16.78}$$

where $Y_0 = \delta\mu + (1+\delta)\theta_1$ and θ_1 is defined by the second equality of (16.78). With μ small, $\theta_1 \approx m^{-1/(m-1)}X_1^2$. Combine (16.76) with (16.78) and also substitute (16.77) omitting the nonpositive final terms. Taking expectations, noting $\mathrm{E}(\Delta B_t|\mathcal{F}_{t-1}) = 0$ and then applying the LIE, gives

$$\mathrm{E}(Y_n^m) \le m(1+\delta)\sum_{t=1}^{n}\mathrm{E}(Y_{t-1}^{m-1}X_t^2). \tag{16.79}$$

By choice of μ the functions $Y_{t-1} = \delta\mu + \delta Q_t^2 + S_t^2$ are bounded away from zero and hence the expectations exist, for each $n \ge 1$.

Next define $\tilde{Y}_t = \delta^{-1}Y_t$ for $t = 1,\dots,n-1$ and $\tilde{Y}_0 = \mu + (1+\delta^{-1})\theta_1$. Since $Y_{t-1}^{m-1} = \delta^{m-1}\tilde{Y}_{t-1}^{m-1}$, inequality (16.79) has the equivalent form

$$\mathrm{E}(Y_n^m) \le m(1+1/\delta)\delta^m\sum_{t=1}^{n}\mathrm{E}(\tilde{Y}_{t-1}^{m-1}X_t^2). \tag{16.80}$$

Note that $\tilde{Y}_{t-1} = \mu + Q_t^2 + \delta^{-1}S_{t-1}^2 - X_t^2$. The fact that there exists $\delta > 0$ small enough that

$$\sum_{t=1}^{n}\mathrm{E}(\tilde{Y}_{t-1}^{m-1}X_t^2) \le \sum_{t=1}^{n}\mathrm{E}((\mu + Q_t^2)^{m-1}X_t^2) \tag{16.81}$$

is shown by contradiction. Suppose that for some t, $\mathrm{E}(\tilde{Y}_{t-1}^{m-1}X_t^2) > \mathrm{E}((\mu + Q_t^2)^{m-1}X_t^2)$ held for all $\delta > 0$. Then, either $S_{t-1}^2 = 0$ with probability 1 or letting $\delta \downarrow 0$ would give $\mathrm{E}((\mu + Q_t^2)^{m-1}X_t^2) \le 0$, which is impossible unless $X_1 = \cdots = X_t = 0$. Therefore, $\sum_{t=1}^{n}\mathrm{E}(\tilde{Y}_{t-1}^{m-1}X_t^2) > \sum_{t=1}^{n}\mathrm{E}((\mu + Q_t^2)^{m-1}X_t^2)$ for all $\delta > 0$ implies that $S_{t-1}^2 = 0$ a.s. for $t = 2,\dots,n$, which is the exceptional case identified above.

By the mean value theorem there exist for $t = 2,\dots,n$ r.v.s $\eta_t \in [0,1]$ such that

$$m(\mu + Q_{t-1}^2 + \eta_t X_t^2)^{m-1}X_t^2 = (\mu + Q_t^2)^m - (\mu + Q_{t-1}^2)^m.$$

For the case $t = 1$ the same equality holds with $Q_0 = 0$ where, for small enough μ, $\eta_1 \approx m^{-1/(m-1)} < 1$. Since $Q_{t-1}^2 + \eta_t X_t^2 \le Q_t^2$ and $m < 1$ it follows similarly to (16.81) that

$$\begin{aligned} m\sum_{t=1}^{n}\mathrm{E}((\mu+Q_t^2)^{m-1}X_t^2) &\leq m\sum_{t=1}^{n}\mathrm{E}\big((\mu+Q_{t-1}^2+\eta_t X_t^2)^{m-1}X_t^2\big) \\ &= \mathrm{E}(\mu+Q_1^2)^m+\sum_{t=2}^{n}\big(\mathrm{E}(\mu+Q_t^2)^m-\mathrm{E}(\mu+Q_{t-1}^2)^m\big) \\ &= \mathrm{E}(\mu+Q_n^2)^m. \qquad (16.82) \end{aligned}$$

Combining (16.75), (16.79), (16.81), and (16.82) gives

$$2^{m-1}(\delta^m\mathrm{E}(\mu+Q_n^2)^m+\mathrm{E}|S_n|^{2m}) \leq (1+1/\delta)\delta^m\mathrm{E}(\mu+Q_n^2)^m,$$

which rearranges, after restoring $p = 2m$, as

$$\mathrm{E}|S_n|^p \leq 2^{1-p/2}((1+1/\delta)-1)\delta^{p/2}\mathrm{E}(\mu+Q_n^2)^{p/2}.$$

Since μ is arbitrary it can be set as small as desired. Letting $\mu \to 0$, the proof of (16.73) is completed by setting

$$C_p = (2^{1-p/2}(1+1/\delta_p)-1)^{1/p}\delta_p^{1/2},$$

where δ_p is largest value of δ that satisfies (16.81) for every $n \geq 1$. ∎

The importance of the L_2-boundedness assumption for this result should not be overlooked, even though the inequality relates to moments of order $p < 2$.

As stated, this result is in the nature of a possibility theorem, since δ_p is not given a value and C_p could need to be arbitrarily large. It is of interest to consider how inequality (16.81) is satisfied in specific cases. If $n = 1$ then given the definition of θ_1 it holds for $\delta \leq (m^{1/m-1}-1)^{-1}$. If $n = 2$, which is the worst case, the second term in the sum has the form $\mathrm{E}((\mu+X_1^2+\delta^{-1}X_1^2)^{m-1}X_2^2)$ on the minorant side and $\mathrm{E}((\mu+X_1^2+X_2^2)^{m-1}X_2^2)$ on the majorant side. The largest δ satisfying (16.81) depends on the distribution of the increments, although it is necessarily positive unless $X_1 = 0$ a.s. On the other hand, if n is large the inequality holds if $\delta^{-1}S_{t-1}^2$ dominates X_t^2 'on average' over $1 \leq t \leq n$, which becomes a difference of orders of magnitude. Asymptotically, δ_p may be set simply to minimize C_p as a function of p. With $p = 1$ for example, $\min C_p = 1.53$ is found at $\delta = 3.4$.

One further useful inequality relating to the sum of squares, adapted from [29], shows how the distribution of the squared process is linked to the first absolute moment of a martingale.

16.29 Theorem Let $\{S_n, \mathcal{F}_n\}_1^\infty$ be a martingale with increments $X_t = S_t - S_{t-1}$ and $S_0 = 0$. If $Q_n = \left(\sum_{t=1}^n X_t^2\right)^{1/2}$ then for $\lambda > 0$,

$$P(Q_n > \lambda) \le \frac{3}{\lambda}\mathrm{E}|S_n|. \tag{16.83}$$

Proof It can be verified that for $1 \le m < n$,

$$2S_mS_n - Q_m^2 - S_m^2 = 2\sum_{t=1}^{m} S_{t-1}X_t + 2S_m \sum_{t=m+1}^{n} X_t, \tag{16.84}$$

where the right-hand terms have expectation zero. Define the stopping time $\mu(\omega) = \inf\{n : |S_n(\omega)| > \lambda\}$ so that $\mathrm{E}(S_{\mu-1}S_n) \le \lambda\mathrm{E}|S_n|$. Setting $m = \mu - 1$ in (16.84), taking expectations and rearranging yields

$$\mathrm{E}Q_{\mu-1}^2 = 2\mathrm{E}(S_{\mu-1}S_n) - \mathrm{E}S_{\mu-1}^2 \le 2\lambda\mathrm{E}|S_n|. \tag{16.85}$$

Now, defining $S_n^* = \max_{1\le k\le n}|S_k|$, consider the inequality

$$P(Q_n > \lambda) \le P(S_n^* > \lambda) + P\big(\{Q_n > \lambda\} \cap \{S_n^* \le \lambda\}\big). \tag{16.86}$$

Applying **16.20** with $p = 1$ gives

$$P(S_n^* > \lambda) \le \lambda^{-1}\mathrm{E}|S_n| \tag{16.87}$$

and in view of the fact that $Q_{\mu-1} = Q_n$ on the set $\{S_n^* \le \lambda\}$,

$$P\big(\{Q_n > \lambda\} \cap \{S_n^* \le \lambda\}\big) = P(Q_{\mu-1} > \lambda) \le \lambda^{-2}\mathrm{E}Q_{\mu-1}^2 \le 2\lambda^{-1}\mathrm{E}|S_n| \tag{16.88}$$

where the inequalities are by the Chebyshev inequality and (16.85). Substituting (16.88) and (16.87) into (16.86) gives (16.83) and completes the proof. ∎

The final result of this section is a so-called *exponential inequality.* This gives a probability bound for martingale processes whose increments are a.s. bounded and is accordingly related directly to the bounding constants rather than to absolute moments. This inequality is due in a slightly different form to Azuma ([14]). The corresponding result for independent sequences is Hoeffding's inequality ([102]).

16.30 Theorem If $\{X_t, \mathcal{F}_t\}_1^\infty$ is a m.d. sequence with $|X_t| \le B_t$ a.s., where $\{B_t\}$ is a sequence of positive constants and $S_n = \sum_{t=1}^n X_t$,

$$P\big(|S_n| > \varepsilon\big) \le 2\exp\Big\{-\varepsilon^2\Big(2\sum_{t=1}^{n} B_t^2\Big)^{-1}\Big\}. \quad \square \tag{16.89}$$

The chief interest of the result is the fact that the tail probabilities are declining exponentially as ε increases. To fix ideas, consider the case $B_t = B$ for all t, so that the probability bound in (16.89) becomes $P(|S_n| > \varepsilon) \le 2\exp\{-\varepsilon^2/2nB^2\}$. This is trivial when n is small since of course $P(|S_n| > nB) = 0$ by construction. However, choosing $\varepsilon = O(n^{1/2})$ allows estimation of the tail probabilities associated with the quantity $n^{-1/2}S_n$. The fact that these are becoming exponential suggests an interesting connection with the central limit results to be studied in Chapter 25.

Proof of 16.30 By convexity, every $x \in [-B_t, B_t]$ satisfies

$$\mathrm{e}^{\alpha x} \le \frac{(B_t + x)\mathrm{e}^{\alpha B_t} + (B_t - x)\mathrm{e}^{-\alpha B_t}}{2B_t} \tag{16.90}$$

for any $\alpha > 0$. Hence, by the m.d. property $\mathrm{E}(X_t|\mathcal{F}_{t-1}) = 0$,

$$\mathrm{E}(\mathrm{e}^{\alpha X_t}|\mathcal{F}_{t-1}) \le \tfrac{1}{2}\big(\mathrm{e}^{\alpha B_t} + \mathrm{e}^{-\alpha B_t}\big) \le \mathrm{e}^{\alpha^2 B_t^2/2} \text{ a.s.} \tag{16.91}$$

where the second inequality can be verified using the series expansion (11.7). Now employ a neat recursion of **10.10**:

$$\begin{aligned}\mathrm{E}(\mathrm{e}^{\alpha S_n}|\mathcal{F}_{n-1}) &= \mathrm{E}(\mathrm{e}^{\alpha S_{n-1} + \alpha X_n}|\mathcal{F}_{n-1})\\ &= \mathrm{e}^{\alpha S_{n-1}}\mathrm{E}(\mathrm{e}^{\alpha X_n}|\mathcal{F}_{n-1})\\ &\le \mathrm{e}^{\alpha S_{n-1}}\exp\{\tfrac{1}{2}\alpha^2 B_t^2\} \text{ a.s.}\end{aligned} \tag{16.92}$$

Generalizing this idea yields

$$\begin{aligned}\mathrm{E}(\mathrm{e}^{\alpha S_n}) &= \mathrm{E}(\mathrm{E}(\cdots\mathrm{E}(\mathrm{E}(\mathrm{e}^{\alpha S_n}|\mathcal{F}_{n-1})|\mathcal{F}_{n-2})\cdots)|\mathcal{F}_1)\\ &\le \exp\{\tfrac{1}{2}\alpha^2 B_n^2\}\mathrm{E}(\mathrm{E}(\cdots\mathrm{E}(\mathrm{e}^{\alpha S_n}|\mathcal{F}_{n-2})\cdots)|\mathcal{F}_1)\\ &\le \cdots\\ &\le \exp\{\tfrac{1}{2}\alpha^2 \textstyle\sum_{t=1}^{n} B_t^2\}.\end{aligned} \tag{16.93}$$

Combining (16.93) with the generalized Markov inequality **9.18** gives

$$P(S_n > \varepsilon) \le \exp\{-\alpha\varepsilon + \tfrac{1}{2}\alpha^2 \textstyle\sum_{t=1}^{n} B_t^2\} \tag{16.94}$$

for $\varepsilon > 0$, which for the choice $\alpha = \varepsilon/\left(\sum_{t=1}^{n} B_n^2\right)$ becomes

$$P(S_n > \varepsilon) \leq \exp\left\{-\varepsilon^2/\left(\sum_{t=1}^{n} B_t^2\right)\right\}. \tag{16.95}$$

The result follows on repeating the argument in respect of $-S_n$ and summing the two inequalities. ∎

A practical application of this sort of result is to team it with a truncation or uniform integrability argument under which the probabilities of the bound B being exceeded can also be suitably controlled.

17
Mixingales

17.1 Definition and Examples

Martingale differences are sequences of a rather special kind. One-step-ahead unpredictability is not a feature encountered in observed time series except in special cases. This chapter introduces a concept of *asymptotic* unpredictability.

17.1 Definition The sequence of pairs $\{X_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ in a filtered probability space $(\Omega, \mathcal{F}, \boldsymbol{F}, P)$ where the X_t are integrable r.v.s is called an L_p-mixingale if, for $p \geq 1$, there exist sequences of non-negative constants $\{c_t\}_{-\infty}^{\infty}$ and $\{\zeta_m\}_0^{\infty}$ such that $\zeta_m \to 0$ as $m \to \infty$ and

$$||\mathrm{E}(X_t|\mathcal{F}_{t-m})||_p \leq c_t\zeta_m \tag{17.1}$$

$$||X_t - \mathrm{E}(X_t|\mathcal{F}_{t+m})||_p \leq c_t\zeta_{m+1} \tag{17.2}$$

hold for all t and $m \geq 0$. □

A martingale difference is a mixingale having $\zeta_m = 0$ for all $m > 0$. Indeed, ‘mixingale differences’ might appear the more logical terminology, but for the fact that the counterpart of the martingale (i.e. the cumulation of a mixingale sequence) does not play any direct role in this theory. The present terminology, due to Donald McLeish who invented the concept, is standard. Many of the results of this chapter are basically due to McLeish, although his theorems are for the case $p = 2$.

Unlike martingales, mixingales form a very general class of stochastic processes; many of the processes for which limit theorems are known to hold can be characterized as mixingales, although supplementary conditions are generally needed. Mixingales are not adapted sequences in general. X_t is not assumed to be $\mathcal{F}_t$-measurable, although if it is then $\mathrm{E}(X_t|\mathcal{F}_{t+m}) = X_t$ a.s. for $m \geq 0$ and (17.2) holds trivially. The mixingale property captures the idea that the sequence $\{\mathcal{F}_s\}$ contains progressively more information about X_t as s increases. In the remote past nothing is known according to (17.1), whereas in the remote future everything will eventually be known according to (17.2).

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
DOI: 10.1093/oso/9780192844507.003.0017

The constants c_t are scaling factors to make the choice of ζ_m scale-independent and multiples of $\|X_t\|_p$ will often fulfil this role. The 'size' terminology is applied here analogously to mixing processes and the sequence is said to be of mixingale size $-\varphi_0$ if $\zeta_m = O(m^{-\varphi})$ for $\varphi > \varphi_0$. The discussion on page 291 in the paragraph following equation (15.6) also applies to this case.

17.2 Example Consider the moving average process

$$X_t = \sum_{j=-\infty}^{\infty} \theta_j U_{t-j}, \tag{17.3}$$

where $\{U_s\}_{-\infty}^{\infty}$ is an L_p-bounded martingale difference sequence, with $p \geq 1$. Also let $\mathcal{F}_t = \sigma(U_s, s \leq t)$. Then

$$\mathrm{E}(X_t|\mathcal{F}_{t-m}) = \sum_{j=m}^{\infty} \theta_j U_{t-j}, \text{ a.s.} \tag{17.4}$$

$$X_t - \mathrm{E}(X_t|\mathcal{F}_{t+m}) = \sum_{j=m+1}^{\infty} \theta_{-j} U_{t+j}, \text{ a.s.} \tag{17.5}$$

Assuming $\{U_s\}_{-\infty}^{\infty}$ to be *uniformly* L_p-bounded, the Minkowski inequality shows that (17.1) and (17.2) are satisfied with $c_t = \sup_s \|U_s\|_p$ for every t and $\zeta_m = \sum_{j=m}^{\infty}(|\theta_j| + |\theta_{-j}|)$. $\{X_t, \mathcal{F}_t\}$ is therefore an L_p-mixingale if $\sum_{j=m}^{\infty}(|\theta_j| + |\theta_{-j}|) \to 0$ as $m \to \infty$ and hence if the coefficients $\{\theta_j\}_{-\infty}^{\infty}$ are absolutely summable. The 'one-sided' process in which $\theta_j = 0$ for $j < 0$ arises more commonly in the econometric modelling context and in this case X_t is $\mathcal{F}_t$-measurable. It is then possible to set $c_t = \sup_{s \leq t} \|U_s\|_p$, which may increase with t and does not have to be bounded in the limit to satisfy the definition. To prove X_t integrable, given integrability of the U_s, requires the absolute summability of the coefficients. In this sense, integrability is effectively sufficient for a linear process with independent increments to be an L_1-mixingale. □

Mixingales are to mixing processes as martingale differences are to independent processes; in each case, a restriction on arbitrary dependence is replaced by a restriction on a simple type of dependence, predictability of the level of the process. Just as martingale differences need not be independent, so mixingales need not be mixing. An important distinction is that the mixing property is defined with respect to the filtration generated by the history of the process itself, $\mathcal{F}_{t_1}^{t_2} = \sigma(X_{t_1}, \ldots, X_{t_2})$, whereas a mixingale is defined with respect to some more general filtration containing progressively increasing information about X_t but to which X_t

need not be adapted. A zero-mean α-mixing process of size $-\varphi$ is adapted to $\{\mathcal{F}_{-\infty}^{t}\}$ by construction and if L_r-bounded for $r > 1$ is an L_p-mixingale with respect to the indicated filtration for $1 \leq p < r$, since (17.1) holds for ζ_m of size $-\varphi(1/p - 1/r)$ by **15.2**. Similarly, the mean deviations of an L_r-bounded ϕ-mixing process form an L_p-mixingale for $1 \leq p \leq r$ of size $-\varphi(1 - 1/r)$, by **15.4**. The reader can supply suitable definitions of c_t in each case.

It is interesting that the mixingale size is lower (absolutely) than the mixing size, except only for mixing sequences having finite sup-norm (L_r-bounded for all finite r) and for the case $p = 1$ in the strong mixing case. Although these relative sizes could be an artefact of the inequalities that can be proved rather than the sharpest available, this is not an unreasonable result. If a sequence has so many outliers that it fails to possess higher-order moments, it would not be surprising to find that it can be predicted further into the future than a sequence with the same dependence structure but more restricted variations.

The next examples show the type of cases arising in the sequel.

17.3 Example An L_r-bounded, zero-mean adapted sequence is an L_2-mixingale of size $-\frac{1}{2}$ if either $r > 2$ and the sequence is α-mixing of size $-r/(r-2)$, or $r \geq 2$ and it is ϕ-mixing of size $-r/2(r-1)$. □

17.4 Example Consider for any $j \geq 0$ the adapted zero-mean sequence

$$\{X_t X_{t+j} - \sigma_{t,t+j}, \mathcal{F}_{t+j}\}$$

where $\sigma_{t,t+j} = \mathrm{E}(X_t X_{t+j})$ and $\{X_t\}$ has one of the properties specified in **17.3**. By **15.1** this is α-mixing (ϕ-mixing) of the same size as X_t for finite j and is $L_{r/2}$-bounded, since

$$\|X_t X_{t+j}\|_{r/2} \leq \|X_t\|_r \|X_{t+j}\|_r$$

by the Cauchy–Schwarz inequality. Assuming $r > 2$ and applying **15.2**, this is an L_1-mixingale of size $-(1 - 2/r)r/(r-2) = -1$ in the α-mixing case. To get this result under ϕ-mixing also requires a mixing size of $-r/(r-2)$, by **15.4**, but such a sequence is also α-mixing of size $-r/(r-2)$ so there is no separate result for the ϕ-mixing case. □

17.5 Example If $\{X_t, \mathcal{F}_t\}$ and $\{Y_t, \mathcal{F}_t\}$ are L_p-mixingales with sizes $-\varphi_X$ and $-\varphi_Y$ and constants c_t^X and c_t^Y respectively, then $\{X_t + Y_t, \mathcal{F}_t\}$ is an L_p-mixingale with size of $-\min\{\varphi_X, \varphi_Y\}$. This follows by the linearity of conditional expectations (10.25), the Minkowski inequality, and the fact that

$$c_t^X \zeta_m^X + c_t^Y \zeta_m^Y \le \max\{c_t^X, c_t^Y\}(\zeta_m^X + \zeta_m^Y)$$

where $\zeta_m^X + \zeta_m^Y = O(m^{-\min\{\varphi_X, \varphi_Y\}})$. □

Mixingales generalize naturally from sequences to arrays.

17.6 Definition The integrable array $\{\{X_{nt}, \mathcal{F}_{nt}\}_{t=-\infty}^{\infty}\}_{n=1}^{\infty}$ is an L_p-mixingale if for $p \ge 1$ there exists an array of non-negative constants $\{c_{nt}\}$ and a non-negative sequence $\{\zeta_m\}_0^{\infty}$ such that $\zeta_m \to 0$ as $m \to \infty$ and

$$\|\mathrm{E}(X_{nt}|\mathcal{F}_{n,t-m})\|_p \le c_{nt}\zeta_m \tag{17.6}$$

$$\|X_{nt} - \mathrm{E}(X_{nt}|\mathcal{F}_{n,t+m})\|_p \le c_{nt}\zeta_{m+1} \tag{17.7}$$

hold for all t, n and $m \ge 0$. □

Other details of the definition are as in **17.1**, it being understood that the sequences $\{\mathcal{F}_{nt}\}_{t=-\infty}^{\infty}$ are nondecreasing in t for each $n \ge 1$. In most applications the array formulation of the filtration is purely formal because the array elements X_{nt} depend on n only through a change of scaling constants, but it is harmless and is retained for full generality. Many results for mixingales can be proved for either the sequence or the array case and the proofs may differ by no more than the inclusion or exclusion of the extra subscript. Unless the changes are more fundamental than this, usually the sequence case is treated with the details of the array case left to the reader.

One word of caution. Mixingale is a low-level property adapted to the easy proof of convergence theorems, but it is not a useful construct at the level of time-series modelling. Although examples such as **17.4** can be exhibited, the mixingale property is not generally preserved under transformations, in the manner of **15.1** for example. Mixingales have too little structure to permit results of that sort. The mixingale concept is mainly useful in conjunction with either mixing assumptions or approximation results of the kind to be studied in Chapter 18. The mixingale property holds for processes for which quite general results on transformations are available.

17.2 Telescoping Sum Representations

Mixingale theory is useful mainly because of an ingenious approximation method. A sum of mixingales is 'nearly' a martingale process, involving a remainder which can be neglected asymptotically under various assumptions limiting the dependence. For the sake of brevity, let $\mathrm{E}_s X_t$ stand for $\mathrm{E}(X_t|\mathcal{F}_s)$. Then note the simple identity, for any integrable random variable X_t and any $m \geq 1$,

$$X_t = \sum_{k=-m}^{m} (\mathrm{E}_{t+k}X_t - \mathrm{E}_{t+k-1}X_t) + \mathrm{E}_{t-m-1}X_t + (X_t - \mathrm{E}_{t+m}X_t). \tag{17.8}$$

Verify that each term on the right-hand side of (17.8) appears twice with opposite signs, except for X_t. For any k, the sequence

$$\{\mathrm{E}_{t+k}X_t - \mathrm{E}_{t+k-1}X_t, \mathcal{F}_{t+k}\}_{t=1}^{\infty}$$

is a martingale difference, since $\mathrm{E}_{t+k-1}(\mathrm{E}_{t+k}X_t - \mathrm{E}_{t+k-1}X_t) = 0$ by the LIE. When $\{X_t, \mathcal{F}_t\}$ is a mixingale, the remainder terms can be made negligible by taking m large enough. Observe that $\{\mathrm{E}_{t+m}X_t, \mathcal{F}_{t+m}\}_{m=-\infty}^{\infty}$ is a martingale and since $\sup_m \mathrm{E}|\mathrm{E}_{t+m}X_t| \leq \mathrm{E}|X_t| < \infty$ by **10.15** it converges a.s. both as $m \to \infty$ and as $m \to -\infty$, by **16.11** and **16.13**, respectively. In view of the fact that $\|\mathrm{E}_{t-m}X_t\|_p \to 0$ and $\|X_t - \mathrm{E}_{t+m}X_t\|_p \to 0$, the respective a.s. limits must be 0 and X_t, and hence

$$X_t = \sum_{k=-\infty}^{\infty} (\mathrm{E}_{t+k}X_t - \mathrm{E}_{t+k-1}X_t), \text{ a.s.} \tag{17.9}$$

Now letting $S_n = \sum_{t=1}^{n} X_t$, there is the corresponding decomposition

$$S_n = \sum_{k=-m}^{m} Y_{nk} + \sum_{t=1}^{n} \mathrm{E}_{t-m-1}X_t + \sum_{t=1}^{n}(X_t - \mathrm{E}_{t+m}X_t) \tag{17.10}$$

where

$$Y_{nk} = \sum_{t=1}^{n}(\mathrm{E}_{t+k}X_t - \mathrm{E}_{t+k-1}X_t) \tag{17.11}$$

and the processes $\{Y_{nk}, \mathcal{F}_{n+k}\}$ are martingales for each k. By taking m large enough with fixed n the remainders can again be made as small as desired. The advantage of this approach is that martingale properties can be exploited in studying the convergence characteristics of sequences of the type S_n. Results of this type are elaborated in §17.3 and §17.4.

If the sequence $\{X_t\}$ is stationary, the constants $\{c_t\}$ can be set to 1 with no loss of generality. In this case, a modified form of telescoping sum actually yields a representation of a partial sum of mixingales as a *single* martingale process, plus a remainder whose behaviour can be suitably controlled by limiting the dependence. The following result is adapted from Hall and Heyde ([88] th. 5.4).

17.7 Theorem Let $\{X_t, \mathcal{F}_t\}$ be a stationary L_1-mixingale of size -1. There exists the decomposition

$$X_t = W_t + Z_t - Z_{t+1} \tag{17.12}$$

where $\mathrm{E}|Z_t| < \infty$ and $\{W_t, \mathcal{F}_t\}$ is a stationary m.d. sequence. □

There is the immediate corollary that

$$S_n = Y_n + Z_1 - Z_{n+1} \tag{17.13}$$

where $\{Y_n, \mathcal{F}_n\}$ is a martingale.

Proof of 17.7 Start with the identity

$$X_t = W_{mt} + Z_{mt} - Z_{m,t+1} \tag{17.14}$$

where, for $m \geq 1$,

$$W_{mt} = \sum_{s=-m}^{m} (\mathrm{E}_t X_{t+s} - \mathrm{E}_{t-1} X_{t+s}) + \mathrm{E}_t X_{t+m+1} + X_{t-m-1} - \mathrm{E}_{t-1} X_{t-m-1} \tag{17.15}$$

and

$$Z_{mt} = \sum_{s=0}^{m} (\mathrm{E}_{t-1} X_{t+s} - X_{t-s-1} + \mathrm{E}_{t-1} X_{t-s-1}). \tag{17.16}$$

As in (17.8), every term appears twice with different sign on the right-hand side of (17.14), except for X_t. Consider the limiting cases of these random variables as $m \to \infty$, to be designated W_t and Z_t respectively. By stationarity,

$$\mathrm{E}|\mathrm{E}_{t-1} X_{t+s}| = \mathrm{E}|\mathrm{E}_{t-s-1} X_t|$$

and

$$\mathrm{E}|X_{t-s-1} - \mathrm{E}_{t-1} X_{t-s-1}| = \mathrm{E}|X_t - \mathrm{E}_{t+s} X_t|.$$

Hence, applying the triangle inequality and recalling $c_t = 1$,

$$\mathrm{E}|Z_t| \leq \sum_{s=0}^{\infty} \mathrm{E}|\mathrm{E}_{t-s-1} X_t| + \sum_{s=0}^{\infty} \mathrm{E}|X_t - \mathrm{E}_{t+s} X_t| \leq 2 \sum_{s=0}^{\infty} \zeta_s < \infty \tag{17.17}$$

by the assumption of size -1. Writing $W_t = X_t - Z_t + Z_{t+1}$, note that

$$\mathrm{E}|W_t| \le \mathrm{E}|X_t| + 2\mathrm{E}|Z_t| < \infty, \tag{17.18}$$

and it remains to be shown that W_t is a m.d. sequence. Applying **10.27**(i) to (17.15),

$$\mathrm{E}_{t-1}W_{mt} = \mathrm{E}_{t-1}X_{t+m+1} \text{ a.s.} \tag{17.19}$$

and stationarity and (17.1) imply that

$$\mathrm{E}|\mathrm{E}_{t-1}X_{t+m+1}| = \mathrm{E}|\mathrm{E}_{t-m-2}X_t| \to 0 \tag{17.20}$$

as $m \to \infty$, so that $\mathrm{E}|\mathrm{E}_{t-1}W_{mt}| \to 0$ also. Anticipating a result from the theory of stochastic convergence (**19.6**), this convergence of the L_1-norm means that every subsequence $\{m_k, k \in \mathbb{N}\}$ contains a further subsequence $\{m_{k(j)}, j \in \mathbb{N}\}$ such that $|E_{t-1}W_{m_{k(j)},t}| \to 0$ a.s. as $j \to \infty$. Since $W_{m_{k(j)},t} \to W_t$ for every such subsequence, it is possible to conclude that $\mathrm{E}(W_t|\mathcal{F}_{t-1}) = 0$ a.s. This completes the proof. ∎

The technical argument in the final paragraph of this proof can be better appreciated after studying Chapter 19. It is neither possible nor necessary in this approach to assert that $\mathrm{E}(W_{mt}|\mathcal{F}_{t-1}) \to 0$ a.s. as $m \to \infty$.

In view of **17.7**, taking conditional expectations of (17.12) yields

$$\mathrm{E}(X_t|\mathcal{F}_{t-1}) = Z_t - Z_{t+1} \text{ a.s.} \tag{17.21}$$

It follows that W_t is almost surely equal to the centred r.v. $X_t - \mathrm{E}(X_t|\mathcal{F}_{t-1})$.

17.8 Example Consider the moving average process from **17.2** with $\{U_t\}$ a stationary integrable m.d. sequence. Then X_t is stationary and

$$\mathrm{E}|X_t| \le \mathrm{E}|U_1| \sum_{j=-\infty}^{\infty} |\theta_j| < \infty.$$

If the coefficients satisfy a stronger summability condition, i.e.

$$\sum_{m=1}^{\infty}\sum_{j=m}^{\infty}(|\theta_j| + |\theta_{-j}|) = \sum_{m=1}^{\infty} m|\theta_m| + \sum_{m=1}^{\infty} m|\theta_{-m}| < \infty \tag{17.22}$$

then X_t is an L_1-mixingale of size -1. A rearrangement of terms gives the decomposition of (17.12) with

$$W_t = \Big(\sum_{j=-\infty}^{\infty} \theta_j\Big) U_t \tag{17.23}$$

and

$$Z_t = \sum_{m=1}^{\infty}\bigg(\Big(\sum_{j=m}^{\infty}\theta_j\Big)U_{t-m} - \Big(\sum_{j=m}^{\infty}\theta_{-j}\Big)U_{t+m-1}\bigg), \tag{17.24}$$

where $\mathrm{E}|Z_t| < \infty$ by (17.22). □

17.3 Maximal Inequalities

As with martingales, maximal inequalities are central to applications of the mixingale concept in limit theory. The basic idea of these results is to extend Doob's inequality **16.21** by exploiting the representation as a telescoping sum of martingale differences. McLeish's idea is to let m go to ∞ in (17.10) and accordingly write

$$S_n = \sum_{k=-\infty}^{\infty} Y_{nk} \text{ a.s.} \tag{17.25}$$

17.9 Lemma Suppose $\{S_j\}_1^n$ has the representation in (17.25). Let $\{a_k\}_{-\infty}^{\infty}$ be a summable collection of non-negative real numbers, with $a_k = 0$ if $Y_{nk} = 0$ a.s. and $a_k > 0$ otherwise. For any $p > 1$,

$$\mathrm{E}\Big(\max_{1\le j\le n}|S_j|^p\Big) \le \Big(\frac{p}{p-1}\Big)^p\Big(\sum_{k=-\infty}^{\infty} a_k\Big)^{p-1}\sum_{a_k>0} a_k^{1-p}\mathrm{E}|Y_{nk}|^p. \tag{17.26}$$

Proof For a real sequence $\{x_k\}_{-\infty}^{\infty}$ and summable, positive real sequence $\{a_k\}_{-\infty}^{\infty}$, let $K = \sum_{k=-\infty}^{\infty} a_k$ and note that

$$\Big|\sum_{k=-\infty}^{\infty} x_k\Big|^p = K^p\Big|\sum_{k=-\infty}^{\infty}\frac{x_k}{a_k}\frac{a_k}{K}\Big|^p \le K^{p-1}\sum_{k=-\infty}^{\infty} a_k^{1-p}|x_k|^p, \tag{17.27}$$

where the weights a_k/K sum to unity and the inequality is by (2.23) for the case $\phi(\cdot) = |\cdot|^p$. Clearly, (17.27) remains true if the terms corresponding to zero x_k are omitted from the sums and for these cases set $a_k = 0$ without loss of generality. Set $x_k = Y_{jk}$, so that the sum on the left-hand side of (17.27) is S_j according to (17.25), Taking expectations gives

$$\mathrm{E}\Big(\max_{1\le j\le n}|S_j|^p\Big) \le \Big(\sum_{k=-\infty}^{\infty} a_k\Big)^{p-1}\sum_{a_k>0} a_k^{1-p}\mathrm{E}\Big(\max_{1\le j\le n}|Y_{jk}|^p\Big), \tag{17.28}$$

taking care to note that replacing $|x_k|^p$ on the majorant side by $\max_{1\le j\le n}|Y_{jk}|^p$ in (17.28) cannot contradict the inequality. To get (17.26), apply Doob's inequality **16.21** on the right-hand side. ∎

This lemma yields the key step in the proof of the next theorem, a maximal inequality for L_2-mixingales. This may not appear a very appealing result at first sight, but of course the interesting applications arise by judicious choice of the sequence $\{a_k\}$.

17.10 Theorem ([125] th.1.6) Let $\{X_t,\mathcal{F}_t\}_{-\infty}^{\infty}$ be an L_2-mixingale, let $S_n = \sum_{t=1}^{n} X_t$, and let $\{a_k\}_0^\infty$ be any summable sequence of positive reals. Then

$$\mathrm{E}\Big(\max_{1\le j\le n} S_j^2\Big) \le 8\Big(\sum_{k=0}^{\infty} a_k\Big)\Big((\zeta_0^2+\zeta_1^2)a_0^{-1} + 2\sum_{k=1}^{\infty}\zeta_k^2(a_k^{-1}-a_{k-1}^{-1})\Big)\Big(\sum_{t=1}^{n} c_t^2\Big). \quad (17.29)$$

Proof To get a doubly infinite sequence $\{a_k\}_{-\infty}^{\infty}$, put $a_{-k}=a_k$ for $k>0$. Then, applying **17.9** for the case $p=2$,

$$\mathrm{E}\Big(\max_{1\le j\le n} S_j^2\Big) \le 4\Big(\sum_{k=-\infty}^{\infty} a_k\Big)\Big(\sum_{k=-\infty}^{\infty} a_k^{-1}\mathrm{E}(Y_{nk}^2)\Big). \quad (17.30)$$

Since the terms making up Y_{nk} are martingale differences and pairwise uncorrelated,

$$\mathrm{E}(Y_{nk}^2) = \sum_{t=1}^{n}\mathrm{E}(\mathrm{E}_{t+k}X_t - \mathrm{E}_{t+k-1}X_t)^2. \quad (17.31)$$

Now, $\mathrm{E}(\mathrm{E}_{t+k}X_t\mathrm{E}_{t+k-1}X_t) = \mathrm{E}(\mathrm{E}_{t+k-1}(\mathrm{E}_{t+k}X_t\mathrm{E}_{t+k-1}X_t)) = \mathrm{E}(\mathrm{E}_{t+k-1}^2X_t)$ by the LIE, from which it follows that

$$\mathrm{E}(\mathrm{E}_{t+k}X_t - \mathrm{E}_{t+k-1}X_t)^2 = \mathrm{E}(\mathrm{E}_{t+k}^2X_t - \mathrm{E}_{t+k-1}^2X_t). \quad (17.32)$$

Also let $Z_{tk} = X_t - E_{t+k}X_t$ and it is similarly easy to verify that

$$\mathrm{E}(\mathrm{E}_{t+k}X_t - \mathrm{E}_{t+k-1}X_t)^2 = \mathrm{E}(Z_{t,k-1} - Z_{tk})^2 = \mathrm{E}(Z_{t,k-1}^2 - Z_{tk}^2). \quad (17.33)$$

Now apply Abel's partial summation formula (**2.25**) to get

$$\sum_{k=-\infty}^{\infty} a_k^{-1}\mathrm{E}(Y_{nk}^2) = \sum_{t=1}^{n}\sum_{k=-\infty}^{\infty} a_k^{-1}\mathrm{E}(E_{t+k}X_t - \mathrm{E}_{t+k-1}X_t)^2$$

$$= \sum_{t=1}^{n}\Bigg(\sum_{k=0}^{\infty} a_k^{-1}\mathrm{E}(\mathrm{E}_{t-k}^2 X_t - \mathrm{E}_{t-k-1}^2 X_t)$$

$$+ \sum_{k=1}^{\infty} a_k^{-1}\mathrm{E}(Z_{t,k-1}^2 - Z_{tk}^2)\Bigg)$$

$$= \sum_{t=1}^{n}\Bigg(a_0^{-1}\mathrm{E}(\mathrm{E}_t^2 X_t) + \sum_{k=1}^{\infty}\mathrm{E}(\mathrm{E}_{t-k}^2 X_t)(a_k^{-1} - a_{k-1}^{-1})$$

$$+ a_1^{-1}\mathrm{E}(Z_{t0}^2) + \sum_{k=1}^{\infty}\mathrm{E}(Z_{tk}^2)(a_{k+1}^{-1} - a_k^{-1})\Bigg). \quad (17.34)$$

The second equality in (17.34) follows by substituting (17.32) for the cases $k \le 0$ and (17.33) for the cases $k > 0$. (17.29) now follows, noting from (17.1) with $p = 2$ that $\mathrm{E}(\mathrm{E}_{t-k}^2 X_t) \le c_t^2\zeta_k^2$ and from (17.2) that $\mathrm{E}(Z_{tk}^2) \le c_t^2\zeta_{k+1}^2$ likewise. ∎

Putting

$$K = 8\Bigg(\sum_{k=0}^{\infty} a_k\Bigg)\Bigg(a_0^{-1}(\zeta_0^2 + \zeta_1^2) + 2\sum_{k=1}^{\infty}\zeta_k^2(a_k^{-1} - a_{k-1}^{-1})\Bigg), \quad (17.35)$$

this result poses the question, does there exist a summable sequence $\{a_k\}_0^\infty$ such that $K < \infty$? There is no loss of generality in letting the sequence $\{\zeta_k\}_0^\infty$ be monotone. If $\zeta_m = 0$ for $m < \infty$, then $\zeta_{m+j} = 0$ for all $j > 0$ and in this case one may choose $a_k = 1$ for $k = 0, \dots, m+1$ and 0 otherwise and K reduces to $8(m+1)(\zeta_0^2 + \zeta_1^2)$. Alternatively, consider the case where $\zeta_k > 0$ for every k. One straightforward choice is to set $a_k = 1/(k\log^2 k)$. This sequence is summable according to **2.18** and, as pointed out in the proof of **2.26**,

$$a_k^{-1} - a_{k-1}^{-1} = \log^2 k + O(\log k).$$

Hence, according to **2.29** it will suffice for $K < \infty$ to have a mixingale size of $-\frac{1}{2}$ in the sense that $\zeta_k = O(k^{-1/2-\delta})$ for $\delta > 0$.

17.11 Corollary Let $\{X_t, \mathcal{F}_t\}$ be an L_2-mixingale of size $-\frac{1}{2}$. Then

$$\mathrm{E}\Big(\max_{1\le j\le n} S_j^2\Big) \le K\sum_{t=1}^{n} c_t^2 \quad (17.36)$$

where $K < \infty$. □

This is the conventional approach.

However, there is another way to think about this question. Put $a_0 = \zeta_0$ and define the recursion

$$a_k = \frac{\zeta_k}{2a_{k-1}}\left((\zeta_k^2 + 4a_{k-1}^2)^{1/2} - \zeta_k\right). \tag{17.37}$$

This construction ensures a_k is real and positive if this is true of a_{k-1}, and since the sequence depends directly on ζ_k, its summability depends on ζ_k converging fast enough. It can also be verified (rearrange (17.37) and then square both sides) that the relation

$$a_k^{-1} - a_{k-1}^{-1} = \zeta_k^{-2} a_k \tag{17.38}$$

is satisfied for each k. Therefore, since $a_0^{-1}(\zeta_0^2 + \zeta_1^2) \le 2a_0$,

$$K = 8\left(\sum_{k=0}^{\infty} a_k\right)\left(a_0^{-1}(\zeta_0^2 + \zeta_1^2) + 2\sum_{k=1}^{\infty} a_k\right) \le 16\left(\sum_{k=0}^{\infty} a_k\right)^2. \tag{17.39}$$

Equation (17.38) implies

$$\zeta_k^{-2} = (a_k^{-1} - a_{k-1}^{-1})a_k^{-1} \le a_k^{-2} - a_{k-1}^{-2} \tag{17.40}$$

for $k > 0$ and hence, recalling $\zeta_0 = a_0$,

$$\sum_{k=0}^{m} \zeta_k^{-2} \le a_0^{-2} + \sum_{k=1}^{m}(a_k^{-2} - a_{k-1}^{-2}) = a_m^{-2}. \tag{17.41}$$

It follows that

$$\sum_{m=0}^{\infty} a_m \le \sum_{m=0}^{\infty}\left(\sum_{k=0}^{m} \zeta_k^{-2}\right)^{-1/2}. \tag{17.42}$$

If $\zeta_k = O(k^{-1/2-\delta})$ for $\delta > 0$ then $\sum_{k=1}^{m} \zeta_k^{-2} = O(m^{2+2\delta})$ by **2.17**(i) and so

$$\left(\sum_{k=1}^{m} \zeta_k^{-2}\right)^{-1/2} = O(m^{-1-\delta})$$

and hence is summable over m, as required. However, the condition

$$\sum_{m=0}^{\infty}\left(\sum_{k=0}^{m} \zeta_k^{-2}\right)^{-1/2} < \infty \tag{17.43}$$

is weaker than $\zeta_k = O(k^{-1/2-\delta})$ since it allows a slowly varying component. Consider the case $\zeta_k = (k+2)^{-1/2}(\log k+2)^{-1-\varepsilon}$ for $\varepsilon > 0$, so that $k^{1/2+\delta}\zeta_k \to \infty$ for every $\delta > 0$. Then

$$\sum_{k=0}^{m} \zeta_k^{-2} = \sum_{k=0}^{m}(k+2)(\log k+2)^{2+2\varepsilon} \le (m+2)^2(\log m+2)^{2+2\varepsilon}, \tag{17.44}$$

and (17.43) follows by **2.18**. One may therefore prefer to define the notion of 'size $= -\frac{1}{2}$' in terms of the summability condition (17.43), rather than by orders of magnitude in m. However, in a practical context, assigning an order of magnitude to ζ_m is a convenient way to bound the dependence. These summability arguments are greatly simplified when the order-of-magnitude calculus can be routinely applied. The existence of slightly weaker conditions is an issue that can always be explored in context if the application suggests it.

Theorem **17.10** has no obvious generalization from the L_2-mixingale case to general L_p for $p > 1$, as in **16.21**, because (17.31) hinges on the uncorrelatedness of the terms. However, since second moments may not exist, a comparable result for $1 < p < 2$ would be valuable. This is attainable by a slightly different approach, although at the cost of raising the mixingale size from $-\frac{1}{2}$ to -1; in other words, for the majorant side of the inequality to be finite, the mixingale numbers will need to be summable.

17.12 Theorem Let $\{X_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ be an L_p-mixingale, $1 < p < 2$, of size -1 and let $S_n = \sum_{t=1}^{n} X_t$. Then

$$\mathrm{E}\Big(\max_{1\le j\le n}|S_j|^p\Big) \le 4^p C_p\Big(\frac{p}{p-1}\Big)^p\Big(\sum_{k=0}^{\infty}\zeta_k\Big)^p\sum_{t=1}^{n} c_t^p, \tag{17.45}$$

where C_p is a positive constant.

Proof Let Y_{nk} be defined as in (17.11) and apply the Burkholder inequality **16.24** and then Loève's c_r inequality (**9.32**) with $r = p/2 \in (\frac{1}{2}, 1)$ to obtain

$$\begin{aligned}\mathrm{E}|Y_{nk}|^p &\le C_p\mathrm{E}\left|\sum_{t=1}^{n}(\mathrm{E}_{t+k}X_t - \mathrm{E}_{t+k-1}X_t)^2\right|^{p/2} \\ &\le C_p\sum_{t=1}^{n}\mathrm{E}|\mathrm{E}_{t+k}X_t - \mathrm{E}_{t+k-1}X_t|^p.\end{aligned} \tag{17.46}$$

Now apply the mixingale inequalities

$$\|\mathrm{E}_{t+k}X_t - \mathrm{E}_{t+k-1}X_t\|_p \le \|\mathrm{E}_{t+k}X_t\|_p + \|\mathrm{E}_{t+k-1}X_t\|_p \le 2c_t\zeta_{|k|} \tag{17.47}$$

for $k < 0$ and

$$\|\mathrm{E}_{t+k}X_t - \mathrm{E}_{t+k-1}X_t\|_p = \|Z_{t,k-1} - Z_{tk}\|_p \le \|Z_{t,k-1}\|_p + \|Z_{tk}\|_p \le 2c_t\zeta_k \tag{17.48}$$

for $k > 0$, where Z_{tk} is defined above (17.33). Hence, (17.46) gives

$$\mathrm{E}|Y_{nk}|^p \le 2^p C_p \zeta_k^p \sum_{t=1}^{n} c_t^p \tag{17.49}$$

(put $\zeta_0 = 1$) and substitution in (17.26), with $\{a_k\}_0^\infty$ a positive sequence and $-a_k = a_k$, gives

$$\mathrm{E}\Big(\max_{1\le j\le n}|S_j|^p\Big) \le 2^{p+1}C_p\Big(\frac{p}{p-1}\Big)^p\Big(\sum_{k=0}^{\infty}a_k\Big)^{p-1}\Big(\sum_{k=0}^{\infty}a_k^{1-p}\zeta_k^p\Big)\sum_{t=1}^{n}c_t^p, \tag{17.50}$$

Both a_k and $a_k^{1-p}\zeta_k^p$ can be summable for $p > 1$ only in the case $\zeta_k = O(a_k)$, and the conclusion follows. ∎

The moving average process of **17.2** specializes **17.12** in the following way:

17.13 Corollary Let $\{U_t\}_{-\infty}^{\infty}$ be an L_p-bounded m.d. sequence with $1 < p \le 2$.

(i) If $X_t = \sum_{j=-\infty}^{\infty}\theta_j U_{t-j}$, then

$$\mathrm{E}\Big(\max_{1\le j\le n}|S_j|^p\Big) \le C_p\Big(\frac{p}{p-1}\Big)^p\Big(|\theta_0| + \sum_{k=1}^{\infty}(|\theta_k| + |\theta_{-k}|)\Big)^p n\sup_s \mathrm{E}|U_s|^p.$$

(ii) If $X_t = \sum_{j=0}^{\infty}\theta_j U_{t-j}$, then

$$\mathrm{E}\Big(\max_{1\le j\le n}|S_j|^p\Big) \le C_p\Big(\frac{p}{p-1}\Big)^p\Big(\sum_{k=0}^{\infty}|\theta_k|\Big)^p\sum_{t=1}^{n}\sup_{s\le t}\mathrm{E}|U_s|^p.$$

Proof In this case, $\mathrm{E}_{t-k}X_t - \mathrm{E}_{t-k-1}X_t = \theta_k U_{t-k}$. Letting $\{a_k\}_0^\infty$ be any non-negative constant sequence and $a_{-k} = a_k$,

$$\sum_{a_k\neq 0} a_k^{1-p}\mathrm{E}|Y_{nk}|^p \leq C_p \sum_{a_k\neq 0} a_k^{1-p}|\theta_k|^p \sum_{t=1}^{n} c_t^p, \tag{17.51}$$

where $c_t = \sup_s \|U_s\|_p$ in case (i) and $c_t = \sup_{s\leq t}\|U_s\|_p$ in case (ii). Choosing $a_k = |\theta_k|$ and substituting in (17.26) yields the results. ∎

Recall that the mixingale coefficients in this case are $\zeta_m = \sum_{j=m}^{\infty}(|\theta_j| + |\theta_{-j}|)$, so linearity yields a dramatic relaxation of the conditions, relative to (17.45). Absolute summability of the θ_j is sufficient, which corresponds simply to $\zeta_m \to 0$. A mixingale size of zero suffices. Moreover, there is no separate result for L_2-bounded linear processes. Putting $p = 2$ yields a result that is correspondingly superior to **17.11** in terms of mixingale size restrictions.

17.4 Uniform Square-Integrability

One of the most important of McLeish's mixingale theorems is a further consequence of **17.10**. It is not a maximal inequality, but belongs to the same family of results and has a related application. The question at issue is the uniform integrability of the sequence of squared partial sums.

17.14 Theorem (from [126] lm. 6.5, [127] lm. 3.5) Let $\{X_t, \mathcal{F}_t\}$ be an L_2-mixingale of size $-\frac{1}{2}$, $S_n = \sum_{t=1}^{n} X_t$, and $\nu_n^2 = \sum_{t=1}^{n} c_t^2$ where c_t is defined in (17.1)–(17.2). If the sequence $\{X_t^2/c_t^2\}_{t=1}^{\infty}$ is uniformly integrable, then so is the sequence $\{\max_{1\leq j\leq n} S_j^2/\nu_n^2\}_{n=1}^{\infty}$.

Proof A preliminary step is to decompose X_t into three components. Choose positive numbers B and m (to be specified below), let $1_t^B = 1_{\{|X_t|\leq Bc_t\}}$, and then define

$$U_t = X_t - E_{t+m}X_t + E_{t-m}X_t \tag{17.52}$$
$$Y_t = E_{t+m}X_t 1_t^B - E_{t-m}X_t 1_t^B \tag{17.53}$$
$$Z_t = \mathrm{E}_{t+m}X_t(1 - 1_t^B) - \mathrm{E}_{t-m}X_t(1 - 1_t^B), \tag{17.54}$$

such that $X_t = U_t + Y_t + Z_t$. The following properties of this decomposition with respect to a positive integer k can be verified by application of the LIE. First,

$$\mathrm{E}(\mathrm{E}_{t-k}U_t)^2 = \mathrm{E}(\mathrm{E}_{t-(k\vee m)}X_t)^2 \leq c_t^2\zeta_{k\vee m}^2 \tag{17.55}$$
$$\mathrm{E}(U_t - \mathrm{E}_{t+k}U_t)^2 = \mathrm{E}(X_t - \mathrm{E}_{t+(k\vee m)}X_t)^2 \leq c_t^2\zeta_{(k\vee m)+1}^2, \tag{17.56}$$

where $k \vee m = \max\{k, m\}$. Second,

$$\mathrm{E}\mathrm{E}_{t-k}^2 Y_t = \mathrm{E}(\mathrm{E}_{t-(k\wedge m)}^2 X_t 1_t^B - \mathrm{E}_{t-m}^2 X_t 1_t^B) \tag{17.57}$$

$$\mathrm{E}(Y_t - \mathrm{E}_{t+k} Y_t)^2 = \mathrm{E}(\mathrm{E}_{t+m}^2 X_t 1_t^B - \mathrm{E}_{t+(k\wedge m)}^2 X_t 1_t^B), \tag{17.58}$$

where $k \wedge m = \min\{k, m\}$. The terms are both zero if $k \geq m$ and are otherwise bounded by $\mathrm{E}(X_t^2 1_t^B) \leq Bc_t$. Third,

$$\mathrm{E}\mathrm{E}_{t-k}^2 Z_t = \mathrm{E}\big(\mathrm{E}_{t-(k\wedge m)}^2 X_t(1 - 1_t^B) - \mathrm{E}_{t-m}^2 X_t(1 - 1_t^B)\big) \tag{17.59}$$

$$\mathrm{E}(Z_t - \mathrm{E}_{t+k} Z_t)^2 = \mathrm{E}\big(\mathrm{E}_{t+m}^2 X_t(1 - 1_t^B) - \mathrm{E}_{t+(k\wedge m)}^2 X_t(1 - 1_t^B)\big), \tag{17.60}$$

where the terms are zero for $k \geq m$ and bounded by $\mathrm{E}(X_t^2(1 - 1_t^B))$ otherwise. Note that $\mathrm{E}(X_t^2(1 - 1_t^B))/c_t^2 = \mathrm{E}((X_t/c_t)^2 1_{\{|X_t/c_t| > B\}}) \to 0$ as $B \to \infty$ uniformly in t, by the assumption of uniform square-integrability.

For any $1 \leq j \leq n$, define

$$x_j = \frac{S_j}{\nu_n},\ u_j = \frac{\sum_{t=1}^j U_t}{\nu_n},\ y_j = \frac{\sum_{t=1}^j Y_t}{\nu_n},\ z_j = \frac{\sum_{t=1}^j Z_t}{\nu_n}.$$

Since $x_j = u_j + y_j + z_j$, the inequality

$$x_j^2 \leq 3(u_j^2 + y_j^2 + z_j^2) \tag{17.61}$$

follows from multiplying out the sum and noting $2u_j y_j \leq u_j^2 + y_j^2$, etc. Further, defining $\hat{x}_n^2 = \max_{1\leq j\leq n} x_j^2$ and letting $\hat{u}_n^2$, $\hat{y}_n^2$, and $\hat{z}_n^2$ be defined similarly, it also follows that

$$\hat{x}_n^2 \leq 3(\hat{u}_n^2 + \hat{y}_n^2 + \hat{z}_n^2). \tag{17.62}$$

For any r.v. $X \geq 0$ and constant $M > 0$, introduce the notation $\mathcal{E}_M(X) = \mathrm{E}(1_{\{X>M\}} X)$. The object of the proof is to show that $\sup_n \mathcal{E}_M(\hat{x}_n^2) \to 0$ as $M \to \infty$ and, to this end, begin by noting that

$$\mathcal{E}_M(\hat{x}_n^2) \leq 6\big(\mathrm{E}(\hat{u}_n^2) + \mathcal{E}_{M/6}(\hat{y}_n^2) + \mathrm{E}(\hat{z}_n^2)\big). \tag{17.63}$$

To interpret this relation note that $a \leq b + c$ for positive numbers a, b, and c implies

$$a1_{\{a>M\}} \leq 2(b + c1_{\{c>M/2\}}). \tag{17.64}$$

This is clearly true if $a \leq M$. For $a > M$, consider the two cases $b \geq c$ and $b < c$. In the first case, $a \leq 2b$. In the second case it is likewise true that $a < 2c$ which implies that $2c > M$. Inequality (17.64) is therefore true in all cases. Replacing a by $\hat{x}_n^2$, b by $3(\hat{u}_n^2 + \hat{z}_n^2)$, c by $3\hat{y}_n^2$, and taking expectations gives (17.63).

For any $\varepsilon > 0$, each of the expectations on the right-hand side of (17.63) can be bounded by ε by choosing m large enough. To show this, first consider $\mathrm{E}(\hat{u}_n^2)$. Given (17.55) and (17.56) and assuming $\zeta_m = O(m^{-1/2-\delta})$, apply **17.10** setting $a_k = m^{-1-\delta}$ for $k \le m$ and $a_k = k^{-1-\delta}$ for $k > m$. Applying (17.29) with $\sum_{t=1}^{j} U_t$ substituted for S_j in that expression produces

$$\begin{aligned}\mathrm{E}(\hat{u}_n^2) &\le 8\Big((m+1)m^{-1-\delta} + \sum_{k=m+1}^{\infty} k^{-1-\delta}\Big)\Big(\zeta_m^2 m^{1+\delta} + 2\sum_{k=m+1}^{\infty} \zeta_k^2 k^{\delta}\Big)\\ &= O(m^{-\delta}) \qquad (17.65)\end{aligned}$$

where the order of magnitude in m follows from **2.17**(iii). Evidently m can be chosen large enough that $\mathrm{E}(\hat{u}_n^2) < \varepsilon$. Henceforth, let m be fixed at this value.

A similar argument is applied to $\mathrm{E}(\hat{z}_n^2)$ but, in view of (17.59) and (17.60) and the fact that these terms vanish for $k \ge m$, write formally $\mathrm{EE}_{t-k}^2 Z_t \le c_t^2 \zeta_k^2$ and $\mathrm{E}(Z_t - \mathrm{E}_{t+k} Z_t)^2 \le c_t^2 \zeta_k^2$ where

$$\zeta_k^2 = \begin{cases} \max_{1\le t\le n} \mathrm{E}((X_t/c_t)^2 1_{\{|X_t/c_t| > B\}}), & k < m \\ 0, & k \ge m. \end{cases} \qquad (17.66)$$

Also, choosing $a_k = 1$ for $k = 0, \dots, m$ and $a_k = 0$ otherwise, application of (17.29) leads to

$$\mathrm{E}(\hat{z}_n^2) \le 16(m+1) \max_{1\le t\le n} \mathrm{E}((X_t/c_t)^2 1_{\{|X_t/c_t| > B\}}). \qquad (17.67)$$

This term goes to zero as $B \to \infty$, so let B be fixed at a value large enough that $\mathrm{E}(\hat{z}_n^2) < \varepsilon$.

For the remaining term, notice that $Y_t = \sum_{k=-m+1}^{m} \xi_{tk}$ where

$$\xi_{tk} = \mathrm{E}_{t+k} X_t 1_t^B - \mathrm{E}_{t+k-1} X_t 1_t^B. \qquad (17.68)$$

For each k, $\{\xi_{tk}, \mathcal{F}_{t+k}\}$ is a m.d. sequence. Applying **17.9** for the case $p = 4$ and $a_k = 1$ for $|k| \le m$ and 0 otherwise yields, not forgetting that $(\max_j y_j^2)^2 = \max_j y_j^4$,

$$\mathrm{E}(\hat{y}_n^4) = \frac{1}{\nu_n^4} \mathrm{E}\Big(\max_{1\le j\le n}\Big|\sum_{t=1}^{j} Y_t\Big|^4\Big) \le \frac{1}{\nu_n^4}\Big(\frac{4}{3}\Big)^4 (2m+1)^3 \sum_{k=-m}^{m} \mathrm{E}(Y_{nk}^4), \qquad (17.69)$$

where $Y_{nk} = \sum_{t=1}^{n} \xi_{tk}$. Given $Y_{nk} = Y_{n-1,k} + \xi_{nk}$, consider the recursion

$$\begin{aligned}\mathrm{E}(Y_{nk}^4) = \mathrm{E}(Y_{n-1,k}^4) &+ 4\mathrm{E}(Y_{n-1,k}^3 \xi_{nk}) + 6\mathrm{E}(Y_{n-1,k}^2 \xi_{nk}^2)\\ &+ 4\mathrm{E}(Y_{n-1,k} \xi_{nk}^3) + \mathrm{E}(\xi_{nk}^4). \qquad (17.70)\end{aligned}$$

Noting the ξ_{tk} are bounded absolutely by $2Bc_t$, consider the terms on the right-hand side of (17.70). The second one vanishes, by the m.d. property. For the third one,

$$\mathrm{E}(Y^2_{n-1,k}\xi^2_{nk}) \leq \mathrm{E}(Y^2_{n-1,k})(2Bc_n)^2 \leq (2B)^4\nu^2_{n-1}c^2_n \tag{17.71}$$

and for the fourth one the Cauchy–Schwarz inequality gives

$$\mathrm{E}|Y_{n-1,k}\xi^3_{nk}| \leq (2B)^4\nu_{n-1}c^3_n. \tag{17.72}$$

Making these substitutions into (17.70) and solving the implied inequality recursively yields

$$\begin{aligned}\mathrm{E}(Y^4_{nk}) &\leq (2B)^4\left(6\sum_{t=1}^{n}\nu^2_{t-1}c^2_t + 4\sum_{t=2}^{n}\nu_{t-1}c^3_t + \sum_{t=1}^{n}c^4_t\right)\\ &\leq 11(2B)^4\nu^4_n.\end{aligned} \tag{17.73}$$

Plugging this bound into (17.69) and applying the inequality $a\mathcal{E}_a(X) \leq \mathrm{E}(X^2)$ for $X \geq 0$ and $a > 0$ yields finally

$$\mathcal{E}_{M/6}(\hat{y}^2_n) \leq \frac{6}{M}\mathrm{E}(\hat{y}^4_n) \leq \left(\frac{4}{3}\right)^4\frac{6(2m+1)^4 11(2B)^4}{M}. \tag{17.74}$$

By choice of M, this quantity can be made smaller than ε.

Substituting the various bounds into (17.63) shows than $\mathcal{E}_M(\hat{x}^2_n) < 6\varepsilon$ for large enough M, or, equivalently,

$$\mathcal{E}_M(\hat{x}^2_n) \to 0 \text{ as } M \to \infty. \tag{17.75}$$

By assumption on the increments the foregoing argument applies uniformly in n, so the proof is complete. ∎

The array version of this result, which is effectively identical, is quoted for the record.

17.15 Corollary Let $\{X_{nt}, \mathcal{F}_{nt}\}$ be an L_2-mixingale array of size $-\frac{1}{2}$ and let $S_n = \sum_{t=1}^{n} X_{nt}$ and $\nu^2_n = \sum_{t=1}^{n} c^2_{nt}$, where c_{nt} is given by (17.6)–(17.7). If $\{X^2_{nt}/c^2_{nt}\}$ is uniformly integrable, $\{\max_{1\leq j\leq n} S^2_j/\nu^2_n\}_{n=1}^{\infty}$ is uniformly integrable.

Proof As for **17.14**, after inserting the subscript n as required. ∎

17.5 Autocovariances

Here is an application of the telescoping sum representation that will play an important role in proving central limit results in Chapters 25 and 26. The behaviour of autocovariances as the time separation of the coordinates increases is a leading issue in these arguments and the mixingale assumption is a neat way to characterize the required memory restrictions.

17.16 Theorem Let $\{X_t, \mathcal{F}_t\}$ be an L_2-mixingale of size $-\frac{1}{2}$ with scale constants $c_t = O(1)$ as $t \to \infty$. Then

$$\frac{1}{n}\sum_{t=1}^{n}\sum_{m=l}^{n-t} |\mathrm{E}(X_t X_{t+m})| = O(l^{-\delta}) \tag{17.76}$$

for $\delta > 0$, as $n \to \infty$. □

The convention here is that a sum takes the value zero when the lower limit exceeds the upper. An alternative notation would be $\sum_{t=1}^{n}\sum_{m=0}^{n-t} 1_{\{m\geq l\}}|\mathrm{E}(X_t X_{t+m})|$. The theorem can be applied setting $l = 0$ and the order of magnitude to $O(1)$ to establish that the normalized sum converges; and also to cases in which $l \to \infty$ while $l/n \to 0$ as $n \to \infty$.

A lemma is needed for the proof of **17.16** as follows. For economy of notation let $\mathrm{E}_s(\cdot)$ denote $\mathrm{E}(\cdot|\mathcal{F}_s)$.

17.17 Lemma $|\mathrm{E}(X_t X_{t+m})| \leq \sum_{j=0}^{\infty} p_{tj}p_{t+m,j} + \sum_{j=1}^{\infty} q_{tj}q_{t+m,j}$ where for $m \geq 0$,

$$p_{t+m,j} = \left(\mathrm{E}\mathrm{E}_{t-j}^2 X_{t+m} - \mathrm{E}\mathrm{E}_{t-j-1}^2 X_{t+m}\right)^{1/2} \tag{17.77}$$

$$q_{t+m,j} = \left(\mathrm{E}(X_{t+m} - \mathrm{E}_{t+j-1}X_{t+m})^2 - \mathrm{E}(X_{t+m} - \mathrm{E}_{t+j}X_{t+m})^2\right)^{1/2}. \tag{17.78}$$

Proof In the telescoping sum representation of the process, $X_t = \sum_{j=-\infty}^{\infty} Z_{tj}$ a.s. where $Z_{tj} = \mathrm{E}_{t-j}X_t - \mathrm{E}_{t-j-1}X_t$. If $j \geq 0$, applying the LIE, it can be verified that $\|Z_{t+m,j+m}\|_2 = p_{t+m,j}$ for $m \geq 0$, whereas for the cases with $j < 0$, replace j by $-j > 0$ and note that

$$Z_{t+m,j+m} = (X_{t+m} - \mathrm{E}_{t+j-1}X_{t+m}) - (X_{t+m} - \mathrm{E}_{t+j}X_{t+m})$$

and hence $\|Z_{t+m,j+m}\|_2 = q_{t+m,j}$ for $m \geq 0$.

Next note that

$$\mathrm{E}(X_tX_{t+m}) = \sum_{j=-\infty}^{\infty}\sum_{i=-\infty}^{\infty}\mathrm{E}(Z_{tj}Z_{t+m,i+m}) = \sum_{j=-\infty}^{\infty}\mathrm{E}(Z_{tj}Z_{t+m,j+m}) \tag{17.79}$$

since if $i > j$,

$$\mathrm{E}(Z_{tj}Z_{t+m,i+m}) = \mathrm{E}(Z_{t+m,i+m}\mathrm{E}_{t-i}Z_{tj}) = 0$$

by further application of the LIE, and if $j > i$ then similarly,

$$\mathrm{E}(Z_{tj}Z_{t+m,i+m}) = \mathrm{E}(Z_{tj}\mathrm{E}_{t-j}Z_{t+m,i+m}) = 0.$$

Applying the Cauchy–Schwarz inequality to the terms of (17.79) yields

$$|\mathrm{E}(X_tX_{t+m})| \le \sum_{j=-\infty}^{\infty}\|Z_{tj}\|_2\|Z_{t+m,j+m}\|_2. \quad \blacksquare$$

Proof of 17.16 Substituting the notation of Lemma **17.17** into the sum (17.76), write

$$\begin{aligned}\frac{1}{n}\sum_{t=1}^{n}\sum_{m=l}^{n-t}|\mathrm{E}(X_tX_{t+m})| &= \frac{1}{n}\sum_{j=0}^{\infty}\left(\sum_{t=1}^{n}\sum_{m=l}^{n-t}p_{tj}p_{t+m,j}\right) + \frac{1}{n}\sum_{j=1}^{\infty}\left(\sum_{t=1}^{n}\sum_{m=l}^{n-t}q_{tj}q_{t+m,j}\right)\\ &= T_1 + T_2\end{aligned}$$

to define terms T_1 and T_2. The analysis of these terms is identical and the result is shown for T_1. Define a positive, summable sequence $\{\eta_m\}_{m=0}^{\infty}$ by setting $\eta_0 = \eta_1 = 1$ and $\eta_m = m^{-1}(\log m)^{-2}$ for $m > 1$. Two applications of the Cauchy–Schwarz inequality for sums (set $p = 2$ in **2.22**) give

$$\begin{aligned}T_1 &= \frac{1}{n}\sum_{j=0}^{\infty}\left(\sum_{t=1}^{n}p_{tj}\sum_{m=l}^{n-t}p_{t+m,j}\right)\\ &\le \sum_{j=0}^{\infty}\left(\frac{1}{n}\sum_{t=1}^{n}p_{tj}^2\right)^{1/2}\left(\frac{1}{n}\sum_{t=1}^{n}\left(\sum_{m=l}^{n-t}p_{t+m,j}\eta_m^{-1/2}\eta_m^{1/2}\right)^2\right)^{1/2}\\ &\le \sum_{j=0}^{\infty}\left(\frac{1}{n}\sum_{t=1}^{n}p_{tj}^2\right)^{1/2}\left(\frac{B}{n}\sum_{s=1}^{n}\sum_{m=l}^{s-1}p_{s,j+m}^2\eta_m^{-1}\right)^{1/2}\end{aligned} \tag{17.80}$$

where $B = \sum_{m=l}^{\infty}\eta_m < \infty$. Note the rearrangement of terms in the last member of (17.80), where, putting $s = t + m$, it can be verified that $\sum_{t=1}^{n}\sum_{m=l}^{n-t}p_{t+m,j}^2\eta_m^{-1} = \sum_{s=1}^{n}\sum_{m=l}^{s-1}p_{s,j+m}^2\eta_m^{-1}$.

In the majorant of (17.80), consider the j^{th} term of the sum. Applying Abelian partial summation (**2.25**) to the second factor gives

$$\begin{aligned}\frac{1}{n}\sum_{s=1}^{n}\sum_{m=l}^{s-1}p_{s,j+m}^2\eta_m^{-1} &= \frac{1}{n}\sum_{s=1}^{n}\sum_{m=l}^{s-1}(\mathrm{EE}_{s-j-m}^2X_s - \mathrm{EE}_{s-j-m-1}^2X_s)\eta_m^{-1}\\ &= \frac{1}{n}\sum_{s=1}^{n}\Big(\mathrm{EE}_{s-j-l}^2X_s\eta_l^{-1} - \mathrm{EE}_{-j}^2X_s\eta_{s-1}^{-1}\\ &\qquad + \sum_{m=l}^{s-2}(\eta_{m+1}^{-1} - \eta_m^{-1})\mathrm{EE}_{s-m-j-1}^2X_s\Big)\\ &= O\big((j+l)^{-2\delta}\big). \end{aligned} \tag{17.81}$$

Noting $\eta_{m+1}^{-1} - \eta_m^{-1} = O(\log^2 m)$, the order of magnitude of (17.81) follows since $\mathrm{EE}_{s-m-j}^2X_s \le c_s^2\zeta(m+j)^2 = O((m+j)^{-1-2\delta})$ for $\delta > 0$ and $m \ge l$, where the indices c_s are bounded and the mixingale size is $-\frac{1}{2}$. Turning to the first factor in (17.80), a third application of the Cauchy–Schwarz inequality for sums followed by Abelian summation gives

$$\begin{aligned}&\sum_{j=0}^{\infty}\Big(\frac{1}{n}\sum_{t=1}^{n}p_{tj}^2\eta_j^{-1/2}\eta_j^{1/2}\Big)^{1/2}\\ &\qquad \le B^{1/2}\Big(\sum_{j=0}^{\infty}\frac{1}{n}\sum_{t=1}^{n}p_{tj}^2\eta_j^{-1}\Big)^{1/2}\\ &\qquad \le B^{1/2}\Big(\frac{1}{n}\sum_{t=1}^{n}\Big(\mathrm{EE}_t^2X_t + \sum_{j=0}^{\infty}(\eta_{j+1}^{-1} - \eta_j^{-1})\mathrm{EE}_{t-j-1}^2X_t\Big)\Big)^{1/2}\\ &\qquad = O(1). \end{aligned} \tag{17.82}$$

Substituting (17.81) and (17.82) into (17.80), the conclusion $T_1 = O(l^{-\delta})$ follows. For the case of T_2, replace $j < 0$ by $-j > 0$ and substitute q for p in (17.80) to obtain the result $T_2 = O(l^{-\delta})$. ∎

Taking $l = 0$ in the stationary case in which $\mathrm{E}(X_tX_{t+m}) = \gamma_m$ not depending on t, (17.76) reduces to

$$\sum_{m=0}^{n-1}\Big(1 - \frac{m}{n}\Big)|\gamma_m| = O(1) \text{ as } n \to \infty. \tag{17.83}$$

If $\gamma_m \simeq m^{-1-\delta}$ then $\delta > 0$ is necessary and sufficient for (17.83) by **2.17** and **2.19**. The conditions of the theorem imply absolute summability of the autocovariance sequence.

Theorem **17.16** has a direct extension to the array case as defined in **17.6**.

17.18 Corollary Let $\{X_{nt}, \mathcal{F}_{nt}\}$ be an L_2-mixingale triangular array of size $-\frac{1}{2}$, with scale constants $\{c_{nt}\}$. If $\sum_{t=1}^{n} c_{nt}^2 = O(1)$ as $n \to \infty$ then for $\delta > 0$ and $l \geq 0$,

$$\sum_{t=1}^{n} \sum_{m=l}^{n-t} |\mathrm{E}(X_{nt}X_{n,t+m})| = O(l^{-\delta}). \quad \square \tag{17.84}$$

The modification to the argument is straightforward, replacing $n^{-1/2}c_t$ with c_{nt}. A natural extension is to the standardized sequence $X_{nt} = X_t/s_n$ with $c_{nt} = \sigma_t/s_n$ where $\sigma_t^2 = \mathrm{E}(X_t^2)$ and $s_n^2 = \sum_{t=1}^{n} \sigma_t^2$. This allows for global nonstationarity, for example with $\sigma_t^2 \simeq t^\beta$, the conditions of **17.18** being met for $\beta > -1$.

Theorems **17.16** and **17.18** also have a straightforward generalization to the multivariate context where 'cross-autocovariances' as well as 'own-autocovariances' need to be considered.

17.19 Theorem Let $\{(X_{nt}, Y_{nt}), \mathcal{F}_{nt}\}$ be an L_2-mixingale triangular array of pairs of size $-\frac{1}{2}$, with scale constants $\{c_{nt}^X, c_{nt}^Y\}$. If $\sum_{t=1}^{n}(c_{nt}^X)^2 = O(1)$ and $\sum_{t=1}^{n}(c_{nt}^Y)^2 = O(1)$ as $n \to \infty$, then for $\delta > 0$ and $l \geq 0$,

$$\sum_{t=1}^{n} \sum_{m=l}^{n-t} |\mathrm{E}(X_{nt}Y_{n,t+m})| = O(l^{-\delta}). \quad \square \tag{17.85}$$

The required modifications to the preceding proof are straightforward and so not spelled out in detail. Lemma **17.17** for the sequence case must now be modified to show

$$|\mathrm{E}(X_t Y_{t+m})| \leq \sum_{j=0}^{\infty} p_{tj}^X p_{t+m,j}^Y + \sum_{j=1}^{\infty} q_{tj}^X q_{t+m,j}^Y,$$

where p_{tj}^X and q_{tj}^X are defined as in (17.77) and (17.78) and p_{tj}^Y and q_{tj}^Y in the obvious complementary manner. With these substitutions the proof of **17.16** goes through virtually unchanged and the array generalization likewise.

Now consider a related issue, treated explicitly in the multivariate framework. The expression with alternative forms

$$\frac{1}{n}\sum_{m=0}^{n-1}\sum_{t=m+1}^{n}X_{t-m}Y_t = \frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{t}X_sY_t \tag{17.86}$$

represents (assuming zero means) the sum of the empirical covariances of Y_t with successive lags of X_t but equivalently the empirical covariance of Y_t with partial sums of X_t. While $Y_t = X_t$ is a leading case linking to **17.16** and **17.18**, the multivariate generalization plays an important role in the analysis of integrated processes in Part VI. For present purposes the inclusion or omission of the zero-lag terms (and hence inclusion or exclusion of the variance or contemporaneous covariance in the sum) is optional. Here it is included, but changing the initial m from 0 to 1 on the left-hand side gives the other case and the next result works for either. Our interest is in whether or not the sum in (17.86) diverges as $n \to \infty$.

In the case where X_t and Y_t are mixingales it is shown, essentially, that if the mixingales are of size $-\frac{1}{2}$ and the scale constants are bounded, the sum in (17.86) is of $O_p(1)$. In contrast to **17.16**, the expected absolute value of (17.86) is analysed rather than the sum of the absolute expectations.

17.20 Theorem Let $\{X_t, \mathcal{F}_t\}$ and $\{Y_s, \mathcal{F}_s\}$ be L_2-mixingales with mixingale numbers respectively $\{\zeta_X(j)\}$ and $\{\zeta_Y(j)\}$ and scale constants $\{c_t^X\}$ and $\{c_s^Y\}$. Then

$$\mathrm{E}\left|\frac{1}{n}\sum_{t=1}^{n}\sum_{s=1}^{t}X_sY_t\right| \ll \sqrt{M_n^X M_n^Y} \tag{17.87}$$

where

$$M_n^X = \frac{1}{n}\sum_{t=1}^{n}(c_t^X)^2\sum_{j=1}^{\infty}(\log j)^2\zeta_X(j)^2$$

with matching definition for M_n^Y.

Proof By the usual telescoping sum representation write $X_s = \sum_{l=-\infty}^{\infty} X_{sl}$ where $X_{sl} = \mathrm{E}_{s-l}X_s - \mathrm{E}_{s-l-1}X_s$, and $Y_t = \sum_{l=-\infty}^{\infty} Y_{ti}$ where $Y_{ti} = \mathrm{E}_{t-i}Y_t - \mathrm{E}_{t-i-1}Y_t$. By the Minkowski inequality,

$$\mathrm{E}\left|\sum_{t=1}^{n}\sum_{s=1}^{t}X_sY_t\right| \le \sum_{l=-\infty}^{\infty}\sum_{i=-\infty}^{\infty}\mathrm{E}\left|\sum_{t=1}^{n}\sum_{s=1}^{t}X_{sl}Y_{ti}\right|. \tag{17.88}$$

For simplicity of notation, $1/n$ is omitted throughout in these expressions.

Fixing the pair (l, i), the trick is now to break up the sums on the majorant side of (17.88) into three components according the time ordering of the pairs. Write

$$\mathrm{E}\left|\sum_{t=1}^{n}\sum_{s=1}^{t}X_{sl}Y_{ti}\right| \leq \mathrm{E}\left|\sum_{t=1}^{n}\sum_{s=1}^{t}X_{sl}Y_{ti}1_{\{t-i>s-l\}}\right| + \mathrm{E}\left|\sum_{s=1}^{n}X_{sl}\sum_{t=s}^{n}Y_{ti}1_{\{t-i<s-l\}}\right|$$
$$+ \mathrm{E}\left|\sum_{t=1}^{n}X_{t-i+l,l}Y_{ti}1_{\{1\leq t-i+l\leq n\}}\right|. \qquad (17.89)$$

By construction, the first and second terms of the majorant of (17.89) contain n-fold sums of martingale differences, measurable with respect to $\mathcal{F}_{t-i}$ in the first case and $\mathcal{F}_{s-l}$ in the second case. Consider the first term. Theorem **16.28** with $p = 1$, then the Cauchy–Schwarz inequality, and lastly (noting that $\{X_{tl}\}$ is a m.d. without restriction) the Doob inequality (**16.21**) give

$$\begin{aligned}
&\sum_{l=-\infty}^{\infty}\sum_{i=-\infty}^{\infty}\mathrm{E}\left|\sum_{t=1}^{n}\sum_{s=1}^{t}X_{sl}Y_{ti}1_{\{t-i>s-l\}}\right| \\
&\quad \ll \sum_{l=-\infty}^{\infty}\sum_{i=-\infty}^{\infty}\mathrm{E}\left(\sum_{t=1}^{n}Y_{ti}^{2}\left(\sum_{s=1}^{t}X_{sl}1_{\{t-i>s-l\}}\right)^{2}\right)^{1/2} \\
&\quad \leq \sum_{l=-\infty}^{\infty}\sum_{i=-\infty}^{\infty}\mathrm{E}\left(\left(\sum_{t=1}^{n}Y_{ti}^{2}\right)\max_{1\leq t\leq n}\left(\sum_{s=1}^{t}X_{sl}\right)^{2}\right)^{1/2} \\
&\quad \leq \sum_{i=-\infty}^{\infty}\left(\sum_{t=1}^{n}\mathrm{E}Y_{ti}^{2}\right)^{1/2}\sum_{l=-\infty}^{\infty}\left(\mathrm{E}\max_{1\leq t\leq n}\left(\sum_{s=1}^{t}X_{sl}\right)^{2}\right)^{1/2} \\
&\quad \ll \sum_{i=-\infty}^{\infty}\left(\sum_{t=1}^{n}\mathrm{E}Y_{ti}^{2}\right)^{1/2}\sum_{l=-\infty}^{\infty}\left(\sum_{t=1}^{n}\mathrm{E}X_{tl}^{2}\right)^{1/2}. \qquad (17.90)
\end{aligned}$$

The same argument applies to the second term, with one modification. To apply the Doob inequality note that $\sum_{t=1}^{s-1}Y_{ti}+\sum_{t=s}^{n}Y_{ti}=\sum_{t=1}^{n}Y_{ti}$ for $1 \leq s \leq n$ (empty sums equalling zero) and hence,

$$\begin{aligned}
\max_{1\leq s\leq n}\left|\sum_{t=s}^{n}Y_{ti}\right| &\leq \max_{1\leq s\leq n}\left|\sum_{t=1}^{s-1}Y_{ti}\right| + \left|\sum_{t=1}^{n}Y_{ti}\right| \\
&\leq \max_{1\leq s\leq n}\left|\sum_{t=1}^{s}Y_{ti}\right| + \left|\sum_{t=1}^{n}Y_{ti}\right| \leq 2\max_{1\leq s\leq n}\left|\sum_{t=1}^{s}Y_{ti}\right|.
\end{aligned}$$

For the third term in (17.89) the Cauchy–Schwarz inequality gives

$$\begin{aligned}
&\sum_{l=-\infty}^{\infty}\sum_{i=-\infty}^{\infty}\mathrm{E}\left|\sum_{s=1}^{n}X_{sl}Y_{s-l+i,i}1_{\{1\leq s-l+i\leq n\}}\right| \\
&\quad \ll \sum_{l=-\infty}^{\infty}\left(\sum_{s=1}^{n}\mathrm{E}X_{sl}^{2}\right)^{1/2}\sum_{i=-\infty}^{\infty}\left(\sum_{t=1}^{n}\mathrm{E}Y_{ti}^{2}\right)^{1/2}. \qquad (17.91)
\end{aligned}$$

Now consider the terms such as appear in the majorants of (17.90) and (17.91). First note that $\left(\sum_{t=1}^{n} \mathrm{E}X_{sl}^2\right)^{1/2} \leq \sum_{t=1}^{n}(\mathrm{E}X_{sl}^2)^{1/2}$ (see **2.21**) and that

$$\mathrm{E}X_{sl}^2 = \mathrm{E}\mathrm{E}_{s-l}^2X_t - \mathrm{E}\mathrm{E}_{s-l-1}^2X_t = \mathrm{E}(X_s - \mathrm{E}_{s-l}X_s)^2 - \mathrm{E}(X_s - \mathrm{E}_{s-l-1}^2X_s)^2,$$

with the second form being relevant to the case $l < 0$. It follows by Theorem **2.26** and the mixingale assumption that

$$\begin{aligned}\sum_{l=-\infty}^{\infty}\left(\sum_{s=1}^{n}\mathrm{E}X_{sl}^2\right)^{1/2} &\leq \sum_{l=-\infty}^{\infty}\sum_{s=1}^{n}(\mathrm{E}\mathrm{E}_{s-l}^2X_s - \mathrm{E}\mathrm{E}_{s-l-1}^2X_s)^{1/2}\\ &\ll \left(\sum_{s=1}^{n}\left(\sum_{l=0}^{\infty}(\log j)^2\mathrm{E}\mathrm{E}_{s-l}^2X_s + \sum_{l=-\infty}^{-1}(\log j)^2\mathrm{E}(X_s - \mathrm{E}_{s-l}X_s)^2\right)\right)^{1/2}\\ &\ll \left(\sum_{s=1}^{n}(c_s^X)^2\sum_{l=1}^{\infty}(\log j)^2\zeta_X(j)^2\right)^{1/2}.\end{aligned}$$

Applying the same reasoning to the case of Y_t completes the proof. ∎

It is easily seen that the same argument can be applied to triangular arrays. The proof of the following is identical apart from applying the subscript n to the various objects defined.

17.21 Corollary Let $\{X_{ns}, \mathcal{F}_{ns}\}$ and $\{Y_{nt}, \mathcal{F}_{nt}\}$ be L_2-mixingale arrays with mixingale numbers respectively $\{\zeta_X(j)\}$ and $\{\zeta_Y(j)\}$ and scale constants $\{c_{ns}^X\}$ and $\{c_{nt}^Y\}$. Then

$$\mathrm{E}\left|\sum_{t=1}^{n}\sum_{s=1}^{t}X_{ns}Y_{nt}\right| \ll \sqrt{M_n^X M_n^Y} \tag{17.92}$$

where

$$M_n^X = \sum_{s=1}^{n}(c_{ns}^X)^2\sum_{j=1}^{\infty}(\log j)^2\zeta_X(j)^2$$

with matching definition for M_n^Y. □

18
Near-Epoch Dependence

18.1 Definitions and Examples

As noted in §15.3, the mixing concept has a serious drawback from the viewpoint of applications in time-series modelling, in that a function of a mixing sequence (even an independent sequence) that depends on an infinite number of lags and/or leads of the sequence is not generally mixing. Let

$$X_t = g_t(\dots,\boldsymbol{V}_{t-1},\boldsymbol{V}_t,\boldsymbol{V}_{t+1},\dots), \tag{18.1}$$

where $\boldsymbol{V}_t$ is a vector of mixing processes. The idea to be developed in this chapter is that although X_t may not be mixing, if it depends almost entirely on the 'near epoch' of $\{\boldsymbol{V}_t\}$ it will often have properties permitting the application of limit theorems, of which the mixingale property is the most important.

This idea goes back to Ibragimov ([103]) and had been formalized in different ways by Billingsley ([21]), McLeish ([125]), Bierens ([19]), Gallant and White ([76]), Andrews ([6]), and Pötscher and Prucha ([149]) among others. The following definitions encompass and extend most existing ones. Consider first a definition for sequences.

18.1 Definition For a stochastic sequence $\{\boldsymbol{V}_t\}_{-\infty}^{+\infty}$, possibly vector-valued, in a probability space $(\Omega,\mathcal{F},P)$ let $\mathcal{F}_{t-m}^{t+m} = \sigma(\boldsymbol{V}_{t-m},\dots,\boldsymbol{V}_{t+m})$, such that $\{\mathcal{F}_{t-m}^{t+m}\}_{m=0}^{\infty}$ is a non-decreasing sequence of σ-fields. If for $p > 0$ a sequence of integrable r.v.s $\{X_t\}_{-\infty}^{+\infty}$ satisfies

$$\|X_t - \mathrm{E}(X_t|\mathcal{F}_{t-m}^{t+m})\|_p \le d_t\nu_m \tag{18.2}$$

where $\nu_m \to 0$ and $\{d_t\}_{-\infty}^{+\infty}$ is a sequence of positive constants, X_t will be said to be *near-epoch dependent in L_p-norm* (L_p-NED) on $\{\boldsymbol{V}_t\}_{-\infty}^{+\infty}$. □

Many results in this literature are proved for the case $p = 2$ (see [76] for example). The term near-epoch dependence, without qualification, might be used in this case. As for mixingales, there is an extension to the array case.

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.

DOI: 10.1093/oso/9780192844507.003.0018

18.2 Definition For a stochastic array $\{\{V_{nt}\}_{t=-\infty}^{+\infty}\}_{n=1}^{\infty}$, possibly vector-valued, in a probability space $(\Omega,\mathcal{F},P)$, let $\mathcal{F}_{n,t-m}^{t+m} = \sigma(V_{n,t-m},\dots,V_{n,t+m})$. If an integrable array $\{\{X_{nt}\}_{t=-\infty}^{+\infty}\}_{n=1}^{\infty}$ satisfies

$$\|X_{nt} - \mathrm{E}(X_{nt}|\mathcal{F}_{n,t-m}^{t+m})\|_p \le d_{nt}\nu_m \tag{18.3}$$

where $\nu_m \to 0$ and $\{d_{nt}\}$ is an array of positive constants, it is said to be L_p-NED on $\{V_{nt}\}$. □

The sequence case is treated below with the extensions to the array case being easily supplied when needed. The size terminology defined for mixing processes and mixingales is also applicable here. The sequence or array is said to be L_p-NED of size $-\varphi_0$ when $\nu_m = O(m^{-\varphi})$ for $\varphi > \varphi_0$. It is also said to be geometrically L_p-NED in the case $\nu_m = O(\mathrm{e}^{-\rho m})$ for $\rho > 0$, hence of infinite size.

According to the Minkowski and conditional modulus inequalities,

$$\begin{aligned}\|X_t - \mathrm{E}(X_t|\mathcal{F}_{t-m}^{t+m})\|_p &\le \|X_t - \mu_t\|_p + \|\mathrm{E}(X_t - \mu_t|\mathcal{F}_{t-m}^{t+m})\|_p \\ &\le 2\|X_t - \mu_t\|_p \end{aligned}\tag{18.4}$$

where $\mu_t = \mathrm{E}(X_t)$. The role of the sequence $\{d_t\}$ in (18.2) is usually to account for the possibility of trending moments, and when $\|X_t - \mu_t\|_p$ is uniformly bounded d_t can be set equal to a finite constant for all t. However, a drawback with the definition is that $\{d_t\}$ can always be chosen in such a way that

$$\inf_t\left\{\frac{\|X_t - \mathrm{E}(X_t|\mathcal{F}_{t-m}^{t+m})\|_p}{d_t}\right\} = 0$$

for every m, so that the near-epoch dependence property can break down in the limit without violating (18.2). Indeed, (18.2) might not hold except with such a choice of constants. In many applications this would represent an undesirable weakening of the condition, which can be avoided by imposing the requirement $d_t \le 2\|X_t - \mu_t\|_p$, or for the array case, $d_{nt} \le 2\|X_{nt} - \mu_{nt}\|_p$. Under this restriction $\nu_m \le 1$ can be set with no loss of generality.

Near-epoch dependence is not an alternative to a mixing assumption; it is a property of the *mapping* from $\{V_t\}$ to $\{X_t\}$, not of the random variables themselves. The concept acquires importance when $\{V_t\}$ is a mixing process, because then $\{X_t\}$ inherits certain useful characteristics. Note that $\mathrm{E}(X_t|\mathcal{F}_{t-m}^{t+m})$ is a finite-lag, $\mathcal{F}_{t-m}^{t+m}/\mathcal{B}$-measurable function of a mixing process and hence is also mixing, by **15.1**. Near-epoch dependence implies that $\{X_t\}$ is 'approximately' mixing in the sense of being well approximated by a mixing process. And as shown below, a near-epoch-dependent function of a mixing process can be a mixingale subject to suitable restrictions on the moments, so that the various inequalities of §17.2 can be exploited in this case.

From the point of view of applications, near-epoch dependence captures nicely the characteristics of a stable dynamic econometric model in which a dependent variable X_t depends mainly on the recent histories of a collection of explanatory variables or shock processes $\boldsymbol{V}_t$, which might be assumed to be mixing. The symmetric dependence on past and future embodied in the definition of an L_p-NED function has no obvious relevance to this case, but it is at worst a harmless generalization. In fact, such cases do arise in various practical contexts, such as the application of two-sided seasonal adjustment procedures or similar smoothing filters; since most published seasonally adjusted time series are the output of a two-sided filter, none of these variables is strictly measurable without reference to future events.

18.3 Example Let $\{V_t\}_{-\infty}^{+\infty}$ be a zero-mean, L_p-bounded scalar sequence and define as in (13.10) the moving average process

$$X_t = \sum_{j=-\infty}^{\infty} \theta_j V_{t-j}. \tag{18.5}$$

By the Minkowski inequality,

$$\begin{aligned}\|X_t - \mathrm{E}(X_t|\mathcal{F}_{t-m}^{t+m})\|_p &= \left\| \sum_{j=m+1}^{\infty} (\theta_j(V_{t-j} - \mathrm{E}_{t-m}^{t+m}V_{t-j}) + \theta_{-j}(V_{t+j} - \mathrm{E}_{t-m}^{t+m}V_{t+j})) \right\|_p \\ &\le d_t \nu_m, \end{aligned} \tag{18.6}$$

where $\nu_m = \sum_{j=m+1}^{\infty}(|\theta_j| + |\theta_{-j}|)$ and $d_t = 2\sup_s \|V_s\|_p$, all t. Clearly, $\nu_m \to 0$ if the sequence $\{\theta_j\}$ is absolutely summable and ν_m is of size $-\varphi_0$ if $|\theta_j| + |\theta_{-j}| = O(j^{-1-\varphi})$ for $\varphi > \varphi_0$. In the one-sided case with $\theta_j = 0$ for $j < 0$, put $d_t = \sup_{s\le t} \|V_s\|_p$ which may be an increasing function of t; compare **17.2**. □

Of the many types of nonlinear time series model that might be chosen for illustration, one of the simplest is the *bilinear* variant of the ARMA that has been studied by Tong ([179]) and Priestley ([152]) *inter alia*. Consider for further simplicity the first-order case. The corresponding analysis for two or more lags can be found in [42].

18.4 Example Let $\{u_t\}_{-\infty}^{\infty}$ be an L_4-bounded, symmetrically distributed zero mean i.i.d. sequence. The process $\{x_t\}$ generated by the bilinear equation

$$x_t = \lambda x_{t-1} + \beta x_{t-1}u_{t-1} + u_t \tag{18.7}$$

is geometrically L_2-NED on $\{u_t\}$ under the same conditions that it is covariance stationary. The solution is most easily obtained by writing $w_t = (\lambda + \beta u_t)x_t$ so that after m substitutions,

$$\begin{aligned} x_t &= u_t + (\lambda + \beta u_{t-1})x_{t-1} \\ &= \cdots \\ &= u_t + \sum_{j=1}^{m}\prod_{i=1}^{j}(\lambda + \beta u_{t-i})u_{t-j} + \prod_{i=1}^{m}(\lambda + \beta u_{t-i})w_{t-m-1}. \end{aligned} \tag{18.8}$$

All but the last of the terms of the right-hand-side member of (18.8) are $\mathcal{F}_{t-m}^{t+m}$-measurable where $\mathcal{F}_{t-m}^{t+m} = \sigma(u_{t-m}, \dots, u_{t+m})$, so that the object of interest is the rate of approach to zero in L_2-norm of

$$x_t - \mathrm{E}_{t-m}^{t+m}x_t = \prod_{i=1}^{m}(\lambda + \beta u_{t-i})(w_{t-m-1} - \mathrm{E}_{t-m}^{t+m}w_{t-m-1}). \tag{18.9}$$

The terms of the product in (18.9) are independent, so assuming covariance stationarity conditions apply the squared norm is

$$\prod_{i=1}^{m}\mathrm{E}(\lambda + \beta u_{t-i})^2\mathrm{E}(w_{t-m-1} - \mathrm{E}_{t-m}^{t+m}w_{t-m-1})^2 \le 2(\lambda^2 + \beta^2\sigma^2)^m\mathrm{E}(w_t^2),$$

where $\sigma^2 = \mathrm{E}(u_t^2)$. This yields the condition $\lambda^2 + \beta^2\sigma^2 < 1$ as sufficient for geometric L_2-NED, provided

$$\mathrm{E}(w_t^2) = \lambda^2\mathrm{E}(x_t^2) + 2\lambda\beta\mathrm{E}(x_t^2u_t) + \beta^2\mathrm{E}(x_t^2u_t^2) < \infty.$$

The moment calculations are straightforward but tedious. Write $\mathrm{E}(u_t^4) = \mu_4$, also noting $\mathrm{E}(u_t^3) = 0$ by assumption. Assuming $|\lambda| < 1$ note first that

$$\mathrm{E}(x_t^2u_t) = 2\frac{\beta\sigma^4}{1-\lambda} = \xi$$

where the last equality is to define ξ. Then it can be shown that under the assumptions on u_t,

$$\mathrm{E}(x_t^2) = \frac{\beta^2(\mu_4 - \sigma^4) + \sigma^2 + 2\lambda\beta\xi}{1 - \lambda^2 - \beta^2\sigma^2}$$

and

$$\mathrm{E}(x_t^2u_t^2) = \frac{(1-\lambda^2)(\mu_4-\sigma^4)+\sigma^4+2\sigma^2\lambda\beta\xi}{1-\lambda^2-\beta^2\sigma^2}.$$

It follows that the conditions $|\lambda| < 1$ and $\lambda^2+\beta^2\sigma^2 < 1$ are sufficient. □

An example suggested by Gallant and White ([76]) illustrates how near-epoch dependence might hold for a wide class of lag functions subject to a dynamic stability condition.

18.5 Example Let $\{V_t\}$ be an L_p-bounded stochastic sequence for $p \geq 2$ and let a sequence $\{X_t\}$ be generated by the nonlinear difference equation

$$X_t = f_t(V_t, X_{t-1}) \tag{18.10}$$

where $\{f_t(\cdot,\cdot)\}$ is a sequence of differentiable functions satisfying

$$\sup_{v,x}\left|\frac{\partial f_t(\nu,x)}{\partial x}\right| \leq b < 1. \tag{18.11}$$

As a function of x, f_t is called a *contraction mapping*. Abstracting from the stochastic aspect of the problem, write ν_t as the dummy first argument of f_t. By repeated substitution,

$$f_t = f_t(\nu_t, f_{t-1}(\nu_{t-1}, f_{t-2}(\nu_{t-2},\dots))) = g_t(\nu_t,\nu_{t-1},\nu_{t-2},\dots) \tag{18.12}$$

and by the chain rule of differentiation,

$$\left|\frac{\partial g_t}{\partial \nu_{t-j}}\right| \leq b^{j-1}\left|\frac{\partial f_{t-j}}{\partial \nu_{t-j}}\right|. \tag{18.13}$$

Replace the arguments with lag exceeding m by zeros to define the near-epoch approximation

$$g_t^m(\nu_t,\dots,\nu_{t-m}) = g_t(\nu_t,\dots,\nu_{t-m},0,0,\dots). \tag{18.14}$$

By the mean value theorem the approximation error has the form

$$g_t - g_t^m = \sum_{j=m+1}^{\infty}\left(\frac{\partial g_t}{\partial \nu_{t-j}}\right)^{*}\nu_{t-j}, \tag{18.15}$$

where * denotes evaluation of the derivatives at points in the intervals $[0,\nu_{t-j}]$.

Now define the stochastic sequence $\{X_t\}$ by evaluating g_t at random variables $(V_t, V_{t-1}, \dots)$ and note that by **10.12**,

$$\|X_t - \mathrm{E}(X_t|\mathcal{F}_{t-m}^{t+m})\|_2 \leq \|X_t - g_t^m(V_t, \dots, V_{t-m})\|_2. \tag{18.16}$$

Let G_{t-j} be the random variable defined by evaluating $(\partial g_t/\partial \nu_{t-j})^*$ at the random argument with F_{t-j} bearing the corresponding relationship with $\partial f_{t-j}/\partial \nu_{t-j}$. The Minkowski inequality, (18.15), and then (18.13) imply that

$$\begin{aligned}\|X_t - g_t^m\|_2 &\leq \sum_{j=m+1}^{\infty} \|G_{t-j}V_{t-j}\|_2 \\ &\leq \sum_{j=m+1}^{\infty} b^{j-1}\|F_{t-j}V_{t-j}\|_2 \\ &\leq \Big(\frac{1}{1-b}\Big)b^m \sup_{j>m} \|F_{t-j}V_{t-j}\|_2, \end{aligned} \tag{18.17}$$

so X_t is L_2-NED on $\{V_t\}$ of size $-\infty$ with constants $d_t \ll \sup_{s\leq t}\|F_sV_s\|_2$. □

With a sacrifice of simplicity the rather strong restriction in (18.11) might well be relaxed, by having f_t be a contraction mapping with just sufficiently high probability to bound the L_2-norms in (18.17) as required.

18.2 Near-Epoch Dependence and Mixingales

The usefulness of the near-epoch dependence concept is due largely to the next theorem.

18.6 Theorem Let $\{X_t\}_{-\infty}^{\infty}$ be an L_r-bounded sequence, for $r > 1$ and L_p-NED of size $-b$ on a process $\{\boldsymbol{V}_t\}$ for $1 \leq p \leq r$ with constants $\{d_t\}$.

(i) If $\{\boldsymbol{V}_t\}$ is α-mixing of size $-a$ and $p < r$, $\{X_t - \mathrm{E}(X_t), \mathcal{F}_{-\infty}^t\}$ is an L_p-mixingale of size $-\min\{b, a(1/p - 1/r)\}$ with constants $c_t \leq \max\{d_t, \|X_t\|_r\}$.

(ii) If $\{\boldsymbol{V}_t\}$ is ϕ-mixing of size $-a$, $\{X_t - \mathrm{E}(X_t), \mathcal{F}_{-\infty}^t\}$ is an L_p-mixingale of size $-\min\{b, a(1 - 1/r)\}$, with constants $c_t \leq \max\{d_t, \|X_t\|_r\}$.

Proof For brevity write $\mathrm{E}_s^t(\cdot) = \mathrm{E}(\cdot|\mathcal{F}_s^t)$ where $\mathcal{F}_s^t = \sigma(\boldsymbol{V}_s, \dots, \boldsymbol{V}_t)$. Also for $m \geq 1$ let $k = [m/2]$, the largest integer not exceeding $m/2$. By the Minkowski inequality,

$$\|\mathrm{E}_{-\infty}^{t-m}(X_t - \mathrm{E}(X_t))\|_p \leq \|\mathrm{E}_{-\infty}^{t-m}(X_t - \mathrm{E}_{t-k}^{t+k}X_t)\|_p + \|\mathrm{E}_{-\infty}^{t-m}(\mathrm{E}_{t-k}^{t+k}X_t - \mathrm{E}(X_t))\|_p. \tag{18.18}$$

Consider each term of the majorant. First,

$$\|\mathrm{E}_{-\infty}^{t-m}(X_t - \mathrm{E}_{t-k}^{t+k}X_t)\|_p \leq \left(\mathrm{E}(\mathrm{E}_{-\infty}^{t-m}|X_t - \mathrm{E}_{t-k}^{t+k}X_t|^p)\right)^{1/p} = \|X_t - \mathrm{E}_{t-k}^{t+k}X_t\|_p \leq d_t\nu_k \tag{18.19}$$

using the conditional Jensen inequality and LIE. Second, $\mathrm{E}_{t-k}^{t+k}X_t - \mathrm{E}(X_t)$ is a finite-lag measurable function of $\boldsymbol{V}_{t-k}, \ldots, \boldsymbol{V}_{t+k}$ and hence mixing of the same size as $\{\boldsymbol{V}_t\}$ for finite k. Hence from **15.2**,

$$\|\mathrm{E}_{-\infty}^{t-m}(\mathrm{E}_{t-k}^{t+k}X_t - \mathrm{E}(X_t))\|_p \leq 6\alpha_k^{1/p-1/r}\|\mathrm{E}_{t-k}^{t+k}X_t\|_r \leq 6\alpha_k^{1/p-1/r}\|X_t\|_r. \tag{18.20}$$

Combining (18.19) and (18.20) into (18.18),

$$\|\mathrm{E}_{-\infty}^{t-m}(X_t - \mathrm{E}(X_t))\|_p \leq \max\{d_t, \|X_t\|_r\}\zeta_m \tag{18.21}$$

holds for the case $\zeta_m = 6\alpha_k^{1/p-1/r} + 2\nu_k$. Also, applying **10.29** gives

$$\|X_t - \mathrm{E}_{-\infty}^{t+m}X_t\|_p \leq 2\|X_t - \mathrm{E}_{t-k}^{t+k}X_t\|_p \leq 2d_t\nu_k \leq d_t\zeta_m. \tag{18.22}$$

Since ζ_m is of size $-\min\{b, a(1/p - 1/r)\}$, part (i) of the theorem follows. The proof of part (ii) is identical except that in place of (18.20) substitute, by **15.4**,

$$\|\mathrm{E}_{-\infty}^{t-m}(\mathrm{E}_{t-k}^{t+k}X_t - \mathrm{E}(X_t))\|_p \leq 2\phi_k^{1-1/r}\|\mathrm{E}_{t-k}^{t+k}X_t\|_r \leq 2\phi_k^{1-1/r}\|X_t\|_r \tag{18.23}$$

and define $\zeta_m = 2\phi_m^{1-1/r} + 2\nu_m$ accordingly. ∎

The incorporation of constant factors into the mixingale index is an arbitrary choice but this is in any case defined only as an order of magnitude. This convention is relatively tidier than including them in the definition of the scale constants. The case when $\boldsymbol{Z}_t$ is a serially independent process often arises in applications. In that case the size is $-b$ and r may be arbitrarily close to p in the α-mixing case, with $a = \infty$.

The following corollary of **18.6** is given for future reference with the mean subtracted implicitly for the sake of simplicity. The proof is in essence just a matter of inserting n before the t subscript wherever required in the last proof.

18.7 Corollary Let $\{\{X_{nt}\}_{t=-\infty}^{+\infty}\}_{n=1}^{\infty}$ be an L_r-bounded zero-mean array, $r > 1$ and L_p-NED of size $-b$ for $1 \le p \le r$ with constants $\{d_{nt}\}$ on an array $\{\boldsymbol{V}_{nt}\}$.

(i) If $\{\boldsymbol{V}_{nt}\}$ is α-mixing of size $-a$ and $p < r$, $\{X_{nt}, \mathcal{F}_{n,-\infty}^{t}\}$ is an L_p-mixingale of size $-\min\{b, a(1/p - 1/r)\}$ with respect to constants $c_{nt} \le \max\{\|X_{nt}\|_r, d_{nt}\}$.

(ii) If $\{\boldsymbol{V}_{nt}\}$ is ϕ-mixing of size $-a$, $\{X_{nt}, \mathcal{F}_{n,-\infty}^{t}\}$ is an L_p-mixingale of size $-\min\{b, a(1 - 1/r)\}$ with respect to constants $c_{nt} \le \max\{\|X_{nt}\|_r, d_{nt}\}$. □

The replacement of $\boldsymbol{V}_t$ by $\boldsymbol{V}_{nt}$ and $\mathcal{F}_s^t$ by $\mathcal{F}_{ns}^t$ is basically a formality since no applications will make use of it. The role of the array notation is generally to indicate a transformation by a function of sample size, typically the normalization of the partial sums to zero mean and unit variance and in these cases $\mathcal{F}_{ns}^t = \mathcal{F}_s^t$ for all n.

Reconsider the AR process of **15.7**. As a special case of **18.3** it is clear that in that example X_t is L_p-NED of size $-\infty$ on Z_t which is an independent process, where since Z_t has moments of all orders p is arbitrary. X_t is therefore an L_p-mixingale of size $-\infty$ for every $p > 0$. There is no need to impose smoothness assumptions on the marginal distributions to obtain these properties, which are usually all that are needed to apply limit theory to the process.

These results allow fine-tuning of the assumptions on the rates of mixing and near-epoch dependence to ensure specific low-level properties needed to prove convergence theorems. Among the most important of these is summability of the sequences of autocovariances. Conditions for absolute summability have been derived for the mixingale case in Theorem **17.16** and Corollary **17.18**, and sufficient conditions for processes that are L_2-NED-on-mixing are easily derived via **18.6** and **18.7**.

18.3 Transformations

Suppose that $(X_{1t}, \ldots, X_{kt})' = \boldsymbol{X}_t = \boldsymbol{g}(\ldots, \boldsymbol{V}_{t-1}, \boldsymbol{V}_t, \boldsymbol{V}_{t+1}, \ldots)$ is a k-vector of L_p-NED functions and interest focuses on the scalar sequence $\{\phi_t(\boldsymbol{X}_t)\}$, where

$$\phi_t : \mathbb{T} \mapsto \mathbb{R}, \mathbb{T} \subseteq \mathbb{R}^k$$

is a $\mathcal{B}^k/\mathcal{B}$-measurable function. Under certain conditions on the function, $\phi_t(\boldsymbol{X}_t)$ will be near-epoch dependent on $\{\boldsymbol{V}_t\}$ if the elements of $\boldsymbol{X}_t$ are. This setup subsumes the important case $k = 1$ where the question at issue is the direct effect of nonlinear transformations on the NED property. The dependence of the functional form $\phi_t(\cdot)$ on t is only occasionally needed, but is worth making explicit.

Consider first the sums and products of pairs of sequences, for which specialized results exist.

18.8 Theorem Let X_t and Y_t be L_p-NED on $\{V_t\}$ of respective sizes $-\varphi_X$ and $-\varphi_Y$. Then $X_t + Y_t$ is L_p-NED of size $-\min\{\varphi_X, \varphi_Y\}$.

Proof Minkowski's inequality gives

$$\begin{aligned} \|(X_t + Y_t) - \mathrm{E}_{t-m}^{t+m}(X_t + Y_t)\|_p &\le \|X_t - \mathrm{E}_{t-m}^{t+m}X_t\|_p + \|Y_t - \mathrm{E}_{t-m}^{t+m}Y_t\|_p \\ &\le d_t^X \nu_m^X + d_t^Y \nu_m^Y \\ &\le d_t \nu_m \end{aligned} \tag{18.24}$$

where $d_t = \max\{d_t^X, d_t^Y\}$ and $\nu_m = \nu_m^X + \nu_m^Y = O(m^{-\min\{\varphi_X, \varphi_Y\}})$. ∎

A variable that is L_q-NED is L_p-NED for $1 \le p \le q$, by the Liapunov inequality, so there is no loss of generality in equating the orders of norm in this result. The same consideration applies to the next theorem.

18.9 Theorem Let X_t and Y_t be L_p-NED on $\{V_t\}$ with $p \ge 2$ of respective sizes $-\varphi_X$ and $-\varphi_Y$. Then, X_tY_t is $L_{p/2}$-NED of size $-\min\{\varphi_Y, \varphi_X\}$.

Proof

$$\begin{aligned} &\|X_tY_t - \mathrm{E}_{t-m}^{t+m}(X_tY_t)\|_{p/2} \\ &\quad = \|(X_tY_t - X_t\mathrm{E}_{t-m}^{t+m}Y_t) + (X_t - \mathrm{E}_{t-m}^{t+m}X_t)\mathrm{E}_{t-m}^{t+m}Y_t \\ &\qquad\qquad - \mathrm{E}_{t-m}^{t+m}((X_t - \mathrm{E}_{t-m}^{t+m}X_t)(Y_t - \mathrm{E}_{t-m}^{t+m}Y_t))\|_{p/2} \\ &\quad \le \|X_t(Y_t - \mathrm{E}_{t-m}^{t+m}Y_t)\|_{p/2} + \|(X_t - \mathrm{E}_{t-m}^{t+m}X_t)\mathrm{E}_{t-m}^{t+m}Y_t\|_{p/2} \\ &\qquad\qquad + \|(X_t - \mathrm{E}_{t-m}^{t+m}X_t)(Y_t - \mathrm{E}_{t-m}^{t+m}Y_t)\|_{p/2} \\ &\quad \le \|X_t\|_p\|Y_t - \mathrm{E}_{t-m}^{t+m}Y_t\|_p + \|\mathrm{E}_{t-m}^{t+m}Y_t\|_p\|X_t - \mathrm{E}_{t-m}^{t+m}X_t\|_p \\ &\qquad\qquad + \|X_t - \mathrm{E}_{t-m}^{t+m}X_t\|_p\|Y_t - \mathrm{E}_{t-m}^{t+m}Y_t\|_p \\ &\quad \le \|X_t\|_p d_t^Y \nu_m^Y + \|Y_t\|_p d_t^X \nu_m^X + d_t^Y \nu_m^Y d_t^X \nu_m^X \\ &\quad \le d_t \nu_m \end{aligned} \tag{18.25}$$

where $d_t = \max\{\|X_t\|_p d_t^Y, \|Y_t\|_p d_t^X, d_t^Y d_t^X\}$ and

$$\nu_m = \nu_m^Y + \nu_m^X + \nu_m^Y \nu_m^X = O(m^{-\min\{\varphi_X, \varphi_Y\}}).$$

The first inequality of (18.25) uses the Minkowski and conditional Jensen inequalities and the second applies (9.57). ■

Both of the last results hold for the case $Y_t = X_{t+j}$ for some finite j, although with a slight modification of the argument.

18.10 Theorem If X_t is L_p-NED on $\{V_t\}$, so is X_{t+j} for $0 < j < \infty$.

Proof If X_t is L_p-NED then

$$\begin{aligned}\|X_{t+j} - \mathrm{E}(X_{t+j}|\mathcal{F}_{t-j-m}^{t+j+m})\|_p &\leq 2\|X_{t+j} - \mathrm{E}(X_{t+j}|\mathcal{F}_{t+j-m}^{t+j+m})\|_p \\ &\leq d_t'^X \nu_m \end{aligned} \tag{18.26}$$

using **10.29**, where $d_t'^X = 2d_{t+j}^X$. Accordingly, write

$$\|X_{t+j} - \mathrm{E}(X_{t+j}|\mathcal{F}_{t-m}^{t+m})\|_p \leq d_t'^X \nu_m', \tag{18.27}$$

where

$$\nu_m' = \begin{cases} \nu_0, & m \leq j \\ \nu_{m-j}, & m > j \end{cases}$$

and ν_m' is of size $-\varphi$ if ν_m is of size $-\varphi$. ■

Putting the last two results together gives the following corollary.

18.11 Corollary If X_t and Y_t are L_p-NED for $p \geq 2$ of size $-\varphi_X$ and $-\varphi_Y$, $X_t Y_{t+k}$ is $L_{p/2}$-NED of size $-\min\{\varphi_Y, \varphi_X\}$. □

By considering $Z_t = X_{t-[k/2]} Y_{t+k-[k/2]}$, the $L_{p/2}$-NED numbers are

$$\nu_m' = \begin{cases} \nu_0, & m \leq [k/2] + 1 \\ \nu_{m-[k/2]-1}, & m > [k/2] + 1 \end{cases}$$

where $\nu_m = \nu_m^Y + \nu_m^X + \nu_m^Y \nu_m^X$ and the constants are $4d_{t-[k/2]}^X d_{t+k-[k/2]}^Y$, assuming that d_t^X and d_t^Y are not smaller than the corresponding L_p norms.

All these results extend to the array case as before, by simply including the extra subscript throughout. Corollary **18.11** should be compared with **17.16**. In the former case k is fixed and finite, whereas the latter result deals with the case as $m \to \infty$. The two theorems naturally complement each other in applying truncation arguments to infinite sums of products.

More general classes of function can be treated under an assumption of continuity, provided $\phi_t(\boldsymbol{X}_t)$ is L_2-NED. Let

$\phi(\boldsymbol{x}) : \mathbb{T} \mapsto \mathbb{R},\ \mathbb{T} \subseteq \mathbb{R}^k$

be a function of k real variables and use the taxicab metric on $\mathbb{R}^k$,

$$\rho(\boldsymbol{x}^1, \boldsymbol{x}^2) = \sum_{i=1}^{k} |x_i^1 - x_i^2| \tag{18.28}$$

to measure the distance between points $\boldsymbol{x}^1$ and $\boldsymbol{x}^2$. The following collection of results imposes restrictions of differing severity on the types of function allowed, but offers a trade-off with the severity of the moment restrictions. To begin with, impose the uniform Lipschitz condition,

$$|\phi_t(\boldsymbol{X}^1) - \phi_t(\boldsymbol{X}^2)| \le B_t \rho(\boldsymbol{X}^1, \boldsymbol{X}^2) \text{ a.s.} \tag{18.29}$$

where B_t is a finite constant.

18.12 Theorem Let X_{it} be L_2-NED of size $-a$ on $\{V_t\}$ for $i = 1, \ldots, k$ with constants d_{it}. If (18.29) holds, $\{\phi_t(\boldsymbol{X}_t)\}$ is also L_2-NED on $\{V_t\}$ of size $-a$ with constants $d_t = kB_t \max_i\{d_{it}\}$.

Proof $\phi_t(\mathrm{E}_{t-m}^{t+m}\boldsymbol{X}_t)$ is an $\mathcal{F}_{t-m}^{t+m}$-measurable random variable. Hence,

$$\begin{aligned} \|\phi_t(\boldsymbol{X}_t) - \mathrm{E}_{t-m}^{t+m}\phi_t(\boldsymbol{X}_t)\|_2 &\le \|\phi_t(\boldsymbol{X}_t) - \phi_t(\mathrm{E}_{t-m}^{t+m}\boldsymbol{X}_t)\|_2 \\ &\le B_t\|\rho(\boldsymbol{X}_t, \mathrm{E}_{t-m}^{t+m}\boldsymbol{X}_t)\|_2 \\ &\le B_t \sum_{i=1}^{k} \|X_{it} - \mathrm{E}_{t-m}^{t+m}X_{it}\|_2 \\ &\le B_t \sum_{i=1}^{k} d_{it}\nu_{im} \le d_t\nu_m, \end{aligned} \tag{18.30}$$

where $\nu_m = k^{-1}\sum_{i=1}^{k}\nu_{im}$ is of size $-a$ by assumption. The first inequality of (18.30) holds by **10.12**, the second is by (18.29), and the third is by Minkowski's inequality and (18.28). ∎

If X_{it} are L_p-NED on $\boldsymbol{V}_t$ for some $\rho \in [1,2)$, this argument fails, but there is a way to get the result if the functions ϕ_t are bounded almost surely.

18.13 Theorem Let X_{it} be L_p-NED of size $-a$ on $\{\boldsymbol{V}_t\}$ for $1 \le p \le 2$, with constants d_{it}, $i = 1,\dots,k$. Suppose that $|\phi_t(\boldsymbol{X}_t)| \le M < \infty$ a.s. for each t and also that

$$|\phi_t(\boldsymbol{X}^1) - \phi_t(\boldsymbol{X}^2)| \le \min\{B_t\rho(\boldsymbol{X}^1,\boldsymbol{X}^2), 2M\} \text{ a.s.} \tag{18.31}$$

where $B_t < \infty$. Then $\{\phi_t(\boldsymbol{X}_t)\}$ is L_2-NED on $\{\boldsymbol{V}_t\}$ of size $-ap/2$, with constants $d_t = B_t^{p/2}(2M)^{1-p/2}k^{p/2}\max_i\{d_{it}^{p/2}\}$.

Proof For brevity, write $\phi_t^i = \phi_t(\boldsymbol{X}^i)$ and let $Z = B_t\rho(\boldsymbol{X}^1,\boldsymbol{X}^2)/2M$, so that $|\phi_t^1 - \phi_t^2| \le 2M\min\{Z,1\}$. Then

$$\begin{aligned}
\mathrm{E}(\phi_t^1 - \phi_t^2)^2 &= \int_{\{Z\le 1\}}(\phi_t^1 - \phi_t^2)^2\mathrm{d}P + \int_{\{Z>1\}}(\phi_t^1 - \phi_t^2)^2\mathrm{d}P \\
&\le (2M)^2\left(\int_{\{Z\le 1\}}Z^2\mathrm{d}P + \int_{\{Z>1\}}\mathrm{d}P\right) \\
&\le (2M)^2\mathrm{E}(Z^p) \\
&= B_{1t}^2\mathrm{E}(\rho(\boldsymbol{X}^1,\boldsymbol{X}^2)^p)
\end{aligned} \tag{18.32}$$

where $B_{1t} = B_t^{p/2}(2M)^{1-p/2}$. The first inequality of (18.30) in combination with (18.32) gives

$$\begin{aligned}
\|\phi_t(\boldsymbol{X}_t) - \mathrm{E}_{t-m}^{t+m}\phi_t(\boldsymbol{X}_t)\|_2 &\le \|\phi_t(\boldsymbol{X}_t) - \phi_t(\mathrm{E}_{t-m}^{t+m}\boldsymbol{X}_t)\|_2 \\
&\le B_{1t}\|\rho(\boldsymbol{X}_t,\mathrm{E}_{t-m}^{t+m}\boldsymbol{X}_t)\|_p^{p/2} \\
&\le B_{1t}\left(\sum_{i=1}^{k}\|X_{it} - \mathrm{E}_{t-m}^{t+m}X_{it}\|_p\right)^{p/2} \\
&\le B_{1t}\left(\sum_{i=1}^{k}d_{it}\nu_{im}\right)^{p/2} \le d_t\nu_m,
\end{aligned} \tag{18.33}$$

where $d_t = B_{1t}k^{p/2}\max_i\{d_{it}^{p/2}\}$ and $\nu_m = (k^{-1}\sum_{i=1}^{k}\nu_{im})^{p/2}$ which is of size $-ap/2$ by assumption. ∎

An important example of this case (with $k = 1$) is the truncation of X_t, although this must be defined as a continuous transformation.

18.14 Example For $M > 0$ let

$$\phi_M(x) = \begin{cases} x, & |x| \le M \\ M, & x > M \\ -M, & x < -M \end{cases} \tag{18.34}$$

or, equivalently, $\phi_M(x) = x1_{\{|x|\le M\}} + M(x/|x|)1_{\{|x|>M\}}$. In this case

$$|\phi_M(X^1) - \phi_M(X^2)| \le |X^1 - X^2|, \tag{18.35}$$

so set $B_t = 1$ and **18.13** can be used to show that $\{\phi_M(X_t)\}$ is L_2-NED if $\{X_t\}$ is L_p-NED. The more conventional truncation,

$$X_t1_{\{|x|\le M\}} = \begin{cases} X_t, & |X_t| \le M \\ 0, & \text{otherwise} \end{cases} \tag{18.36}$$

cannot be shown to be near-epoch dependent by this approach, because of the lack of continuity. □

A further variation on **18.12** is to relax the Lipschitz condition (18.29) by letting the scale factor B_t be a possibly unbounded function of the random variables. Assume

$$|\phi_t(\boldsymbol{X}^1) - \phi_t(\boldsymbol{X}^2)| \le B_t(\boldsymbol{X}^1,\boldsymbol{X}^2)\rho(\boldsymbol{X}^1,\boldsymbol{X}^2) \text{ a.s.} \tag{18.37}$$

where, for each t,

$$B_t(\boldsymbol{X}^1,\boldsymbol{X}^2) : \mathbb{T}\times\mathbb{T} \mapsto \mathbb{R}^+ \tag{18.38}$$

is a non-negative, $\mathcal{B}^{2k}/\mathcal{B}$-measurable function. To deal with this case requires a lemma due to Gallant and White ([76]). The object of this result is to allow the Hölder inequality to be applied to a squared product while having $q \le 2$ in the L_q-norm of at least one of the factors of the majorant.

18.15 Lemma Let B and ρ be non-negative r.v.s and assume $\|\rho\|_q < \infty$, $\|B\|_{q/(q-1)} < \infty$ and $\|B\rho\|_r < \infty$, for $q \ge 1$ and $r > 2$. Then

$$\|B\rho\|_2 \leq 2(\|\rho\|_q^{r-2}\|B\|_{q/(q-1)}^{r-2}\|B\rho\|_r^r)^{1/2(r-1)}. \tag{18.39}$$

Proof Define

$$C = (\|\rho\|_q\|B\|_{q/(q-1)}\|B\rho\|_r^{-r})^{1/(1-r)} \tag{18.40}$$

and let $B_1 = 1_{\{B\rho \leq C\}}B$. Then by the Minkowski inequality,

$$\|B\rho\|_2 \leq \|B_1\rho\|_2 + \|(B - B_1)\rho\|_2. \tag{18.41}$$

The right-hand-side terms are bounded as follows. First, applying the Hölder inequality,

$$\begin{aligned}\|B_1\rho\|_2 &= \left(\int_{B\rho\leq C}(B\rho)^2\mathrm{d}P\right)^{1/2}\\ &\leq C^{1/2}\left(\int B\rho\mathrm{d}P\right)^{1/2}\\ &\leq C^{1/2}(\|\rho\|_q\|B\|_{q/(q-1)})^{1/2}.\end{aligned} \tag{18.42}$$

Second,

$$\begin{aligned}\|(B-B_1)\rho\|_2 &= \left(\int_{B\rho>C}(B\rho)^2\mathrm{d}P\right)^{1/2}\\ &\leq C^{1-r/2}\left(\int_{B\rho>C}(B\rho)^r\mathrm{d}P\right)^{1/2}\\ &\leq C^{1-r/2}\|B\rho\|_r^{r/2}\end{aligned} \tag{18.43}$$

where the first inequality follows from $r > 2$ and $B\rho/C > 1$. Substituting for C in (18.42) and (18.43) yields the same expression, so applying (18.41) gives the result. ∎

The general result is then as follows.

18.16 Theorem Let $\{\boldsymbol{X}_t\}$ be a k-dimensioned random sequence, of which each element is L_2-NED of size $-a$ on $\{\boldsymbol{V}_t\}$ and suppose that $\phi_t(\boldsymbol{X}_t)$ is L_2-bounded. Suppose further that for $1 \leq q \leq 2$,

$$\|\rho(\boldsymbol{X}_t, \mathrm{E}_{t-m}^{t+m}\boldsymbol{X}_t)\|_q < \infty$$

$$\|B_t(\boldsymbol{X}_t, \mathrm{E}_{t-m}^{t+m}\boldsymbol{X}_t)\|_{q/(q-1)} < \infty$$

and for $r > 2$,

$$\|B_t(\boldsymbol{X}_t, \mathrm{E}_{t-m}^{t+m}\boldsymbol{X}_t)\rho(\boldsymbol{X}_t, \mathrm{E}_{t-m}^{t+m}\boldsymbol{X}_t)\|_r < \infty.$$

Then $\{\phi_t(\boldsymbol{X}_t)\}$ is L_2-NED on $\{\boldsymbol{V}_t\}$ of size $-a(r-2)/2(r-1)$.

Proof For ease of notation, write ρ for $\rho(\boldsymbol{X}_t, \mathrm{E}_{t-m}^{t+m}\boldsymbol{X}_t)$ and B for $B_t(\boldsymbol{X}_t, \mathrm{E}_{t-m}^{t+m}\boldsymbol{X}_t)$. Then, similarly to (18.30) and (18.33) but now applying **18.15** with $q = 2$,

$$\begin{aligned}\|\phi(\boldsymbol{X}_t) - \mathrm{E}_{t-m}^{t+m}\phi(\boldsymbol{X}_t)\|_2 &\le \|\phi(\boldsymbol{X}_t) - \phi(\mathrm{E}_{t-m}^{t+m}\boldsymbol{X}_t)\|_2 \\ &\le \|B\rho\|_2 \\ &\le 2(\|\rho\|_2^{r-2}\|B\|_2^{r-2}\|B\rho\|_r^r)^{1/2(r-1)}. \end{aligned} \tag{18.44}$$

Since

$$\|\rho\|_2 \le \sum_{i=1}^{k} \|X_{it} - \mathrm{E}_{t-m}^{t+m}X_{it}\|_2 \le \sum_{i=1}^{k} d_{it}\nu_{im} = d_t\nu_m \tag{18.45}$$

with $d_t = k\max_i\{d_{it}\}$, and $\nu_m = k^{-1}\sum_{i=1}^{k}\nu_{im}$ which is of size $-a$ by assumption,

$$\|\phi_t(\boldsymbol{X}_t) - \mathrm{E}_{t-m}^{t+m}\phi_t(\boldsymbol{X}_t)\|_2 \le d_t'\nu_m^{(r-2)/2(r-1)}, \tag{18.46}$$

where $d_t' = \|B\|_2^{(r-2)/2(r-1)}\|B\rho\|_r^{r/2(r-1)}d_t$. ∎

The exponent in (18.46) is a small number unless r is large and at best bounded by $\frac{1}{2}$, so by comparison with **18.12** the cost of relaxing (18.29) in this result proves to be substantial. However, without **18.15** the best that could be done in (18.44) would be to apply the Hölder inequality directly to obtain

$$\|B\rho\|_2 \le \|\rho\|_{2q}\|B\|_{2q/(q-1)}, q \ge 1. \tag{18.47}$$

The minimum requirement for this inequality to be useful is that B be bounded almost surely permitting the choice $q = 1$. This is merely the case covered by **18.12** with the constant scale factors set to ess sup $B_t(\boldsymbol{X}^1, \boldsymbol{X}^2)$.

The following application of **18.16** may be contrasted with **18.9**. In addition to the noted penalty on the L_2-NED size the moment conditions have to be strengthened by a factor of at least 2, to ensure that the product of L_2-NED functions is also L_2-NED rather than just L_1-NED.

18.17 Example Let $\boldsymbol{X}_t = (X_t, Y_t)$ and $\phi(\boldsymbol{X}_t) = X_t Y_t$. Assume that $\|X_t\|_{2r} < \infty$ and $\|Y_t\|_{2r} < \infty$ for $r > 2$ and that X_t and Y_t are L_2-NED on $\{V_t\}$ of size $-a$. Then

$$\begin{aligned}|X_t^1 Y_t^1 - X_t^2 Y_t^2| &\le |X_t^1|\,|Y_t^1 - Y_t^2| + |X_t^1 - X_t^2|\,|Y_t^2| \\ &\le (|X_t^1| + |Y_t^2|)(|Y_t^1 - Y_t^2| + |X_t^1 - X_t^2|) \\ &= B(\boldsymbol{X}_t^1, \boldsymbol{X}_t^2)\rho(\boldsymbol{X}_t^1, \boldsymbol{X}_t^2),\end{aligned} \tag{18.48}$$

defining B and ρ. For q in the range $[4/3, 4]$ the assumptions imply

$$\|B(\boldsymbol{X}_t^1, \boldsymbol{X}_t^2)\|_{q/(q-1)} \le \|X_t^1\|_{2r} + \|Y_t^2\|_{2r} < \infty \tag{18.49}$$

$$\|\rho(\boldsymbol{X}_t^1, \boldsymbol{X}_t^2)\|_q \le \|Y_t^1\|_{2r} + \|Y_t^2\|_{2r} + \|X_t^1\|_{2r} + \|X_t^2\|_{2r} < \infty \tag{18.50}$$

and

$$\begin{aligned}\|B(\boldsymbol{X}_t^1, \boldsymbol{X}_t^2)\rho(\boldsymbol{X}_t^1, \boldsymbol{X}_t^2)\|_r &\le \|X_t^1\|_{2r}\|Y_t^1\|_{2r} + \|X_t^1\|_{2r}\|Y_t^2\|_{2r} + \|X_t^1\|_{2r}^2 \\ &\quad + \|X_t^1\|_{2r}\|X_t^2\|_{2r} + \|Y_t^2\|_{2r}\|Y_t^1\|_{2r} + \|Y_t^2\|_{2r}^2 \\ &\quad + \|Y_t^2\|_{2r}\|X_t^1\|_{2r} + \|Y_t^2\|_{2r}\|X_t^2\|_{2r} \\ &< \infty.\end{aligned} \tag{18.51}$$

Putting $\boldsymbol{X}_t^1 = \boldsymbol{X}_t$ and $\boldsymbol{X}_t^2 = \mathrm{E}_{t-m}^{t+m}\boldsymbol{X}_t$, the conditions of **18.16** are satisfied for q in the range $[4/3, 2]$ and $X_t Y_t$ is L_2-NED of size $-a(r-2)/2(r-1)$. □

18.4 Adaptation

A natural case to consider is when the sequence $\{X_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ is adapted, so that $\mathrm{E}(X_t|\mathcal{F}_t) = X_t$ a.s. This has the simple implication that the process is causal; if the progression of the sequence is identified with the passage of time, then under the adaptation assumption it is driven by present and past shocks alone. This is a very natural assumption in the context of econometric time series modelling, implying that knowledge of future shocks cannot improve the prediction of X_t once its history is known. In the mixingale context, it means that the norm in (17.2) is equal to zero for every $m \ge 0$, whereas the NED-defining inequality becomes

$$\|X_t - \mathrm{E}(X_t|\mathcal{F}_{t-m}^t)\|_p \le d_t \nu_m \tag{18.52}$$

in place of (18.2). It is conventional in econometric applications not to assume adaptation for the sake of full generality, but some arguments are considerably simplified by doing so.

The chief implication of adaptation is that $\mathrm{E}_{t+s}X_t = X_t$ when $s \geq 0$. The mixingale property corresponds to just (17.1). Consider specifically the properties of the products X_tX_{t+s}, when X_t is an adapted sequence that is NED on a mixing process and hence a mixingale according to Theorem **18.6**.

18.18 Theorem Let $\{X_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ be an L_r-bounded, adapted, zero-mean sequence for $r \geq q > p \geq 2$. Assume that for $s \geq 0$, $\{\mathrm{E}_tX_{t+s}, \mathcal{F}_t\}$ is L_q-NED of size $-a$ on a process $\{\boldsymbol{V}_t\}$ with constants $d_t \leq \|X_{t+s}\|_q$. Either

(i) $\{\boldsymbol{V}_t\}$ is α-mixing of size $-3apr/2(r-p)$ and $q = 3pr/(p+2r) < r$, or

(ii) $\{\boldsymbol{V}_t\}$ is ϕ-mixing of size $-ar/(r-2)$ and $q = r$.

Then, $\{X_tX_{t+s} - \sigma_{t,t+s}, \mathcal{F}_{t+s}\}_{-\infty}^{\infty}$ is an $L_{p/2}$-mixingale of size $-a$ with constants $c_{t+s} = \|X_t\|_r\|X_{t+s}\|_r$. □

It can be verified that in part (i), q is defined as the solution to the equation $2/p - 2/q = 1/q - 1/r$, a condition required by the proof.

Unlike Corollary **18.11** which is limited to a finite fixed lag separation, in **18.18** the rate of memory decay is independent of s, even as $s \to \infty$. This property is the consequence of assuming the sequences of conditional means are NED, subsuming the usual NED property for X_t which is the case $s = 0$. Under adaptation this assumption appears quite natural. By the law of iterated expectations the L_p-NED condition bounds $\|\mathrm{E}_{-\infty}^{t}X_{t+s} - \mathrm{E}_{t-m}^{t}X_{t+s}\|_p$. This clearly converges to zero as $m \to \infty$ and the only question arising is whether the rate of convergence might depend adversely on s. For illustration, consider the linear adapted process $\{X_t, \mathcal{F}_t\}$ where $X_t = \sum_{j=0}^{\infty}\theta_j\varepsilon_{t-j}$, with $\mathcal{F}_t = \sigma(\varepsilon_k, k \leq t)$ and $\{\varepsilon_t\}$ an L_2-bounded independent sequence with zero mean. Here, $\mathrm{E}_tX_{t+s} = \sum_{j=0}^{\infty}\theta_{j+s}\varepsilon_{t-j}$ and if $|\theta_j| = O(j^{-1-\varphi})$ for $\varphi > \varphi_0 \geq 0$, then for $p \leq 2$,

$$\begin{aligned}\|\mathrm{E}_tX_{t+s} - \mathrm{E}_{t-m}^{t}X_{t+s}\|_p &= \|\textstyle\sum_{j=m+1}^{\infty}\theta_{j+s}\varepsilon_{t-j}\|_p \\ &\leq \sup_{k \leq t-m}\|\varepsilon_k\|_p\textstyle\sum_{j=s+m+1}^{\infty}|\theta_j| \\ &= O((s+m)^{-\varphi}). \qquad (18.53)\end{aligned}$$

The sequences $\{\mathrm{E}_tX_{t+s}\}$ are therefore L_p-NED on $\{\varepsilon_t\}$ of size φ_0 for all $s \geq 0$.

Proof of 18.18 Case (i) will be proved first, followed by the modifications of the argument for case (ii). With $s = 0$ the assumptions imply $\{X_t\}$ is a mixingale and

$$||\mathrm{E}_{t-m}X_t||_q \le ||X_t||_r\zeta_m \tag{18.54}$$

where, for $k = [m/2]$, $\zeta_m = 2\nu_k + 6\alpha_k^{1/q-1/r}$ is the mixingale index derived in **18.6** with NED index $\nu_k = O(k^{-a})$. Under the stated conditions $\alpha_k^{1/q-1/r} = O(k^{-a})$ and hence $\zeta_m = O(m^{-a})$.

By (18.54), the s^{th} autocovariance is bounded as

$$\begin{aligned}|\sigma_{t,t+s}| = |\mathrm{E}(X_t\mathrm{E}_tX_{t+s})| &\le ||X_t\mathrm{E}_tX_{t+s}||_{q/2}\\ &\le ||X_t||_q||\mathrm{E}_tX_{t+s}||_q \le ||X_t||_q||X_{t+s}||_r\zeta_s,\end{aligned} \tag{18.55}$$

where these inequalities apply successively the LIE, **10.10**, the modulus and Liapunov inequalities, then Cauchy–Schwarz. If $m \le s$, similar arguments and (18.55) give

$$\begin{aligned}||\mathrm{E}_{t+s-m}(X_tX_{t+s}) - \sigma_{t,t+s}||_{p/2} &\le ||X_t\mathrm{E}_{t+s-m}X_{t+s}||_{p/2} + |\sigma_{t,t+s}|\\ &\le ||X_t||_p||\mathrm{E}_{t+s-m}X_{t+s}||_q + |\sigma_{t,t+s}|\\ &\le ||X_t||_q||X_{t+s}||_r(\zeta_m + \zeta_s)\\ &= O(m^{-a}).\end{aligned} \tag{18.56}$$

Next, suppose $m > s$. To show that the sequence $\{X_t\mathrm{E}_tX_{t+s}\}$ is $L_{p/2}$-NED let $P_t = X_t - \mathrm{E}_{t+s-m}^tX_t$ and $Q_t = \mathrm{E}_tX_{t+s} - \mathrm{E}_{t+s-m}^t\mathrm{E}_tX_{t+s}$. By assumption,

$$||P_t||_q \le ||X_t||_q\nu_{m-s} \tag{18.57}$$

and

$$||Q_t||_q \le ||\mathrm{E}_tX_{t+s}||_q\nu_{m-s} \le ||X_{t+s}||_r\nu_{m-s}\zeta_s. \tag{18.58}$$

Therefore by a variant of Theorem **18.9**,

$$\begin{aligned}&||X_t\mathrm{E}_tX_{t+s} - \mathrm{E}_{t+s-m}^t(X_t\mathrm{E}_tX_{t+s})||_{p/2}\\ &\quad = ||X_tQ_t + P_t\mathrm{E}_{t+s-m}^t\mathrm{E}_tX_{t+s} - \mathrm{E}_{t+s-m}^t(P_tQ_t)||_{p/2}\\ &\quad \le ||X_tQ_t||_{p/2} + ||P_t\mathrm{E}_{t+s-m}^t\mathrm{E}_tX_{t+s}||_{p/2} + ||P_tQ_t||_{p/2}\\ &\quad \le ||X_t||_q||Q_t||_q + ||P_t||_q||\mathrm{E}_tX_{t+s}||_q + ||P_t||_q||Q_t||_q\\ &\quad \le ||X_t||_q||X_{t+s}||_r(2 + \nu_{m-s})\nu_{m-s}\zeta_s.\end{aligned} \tag{18.59}$$

Here, the first inequality of (18.59) applies the Minkowski and conditional Jensen inequalities; the second applies Cauchy–Schwarz, the conditional Jensen a second time, and Liapunov; and the last substitutes from (18.54), (18.57), and (18.58).

To show that $\{X_t\mathrm{E}_tX_{t+s} - \sigma_{t,t+s}\}$ is an $L_{p/2}$-mixingale, a modification of the argument in **18.6** is applied. In the counterpart of inequality (18.18), write $2(m-s)$ in place of 'm' and so compute the bound for $\|\mathrm{E}_{t+2(s-m)}(X_t\mathrm{E}_tX_{t+s} - \sigma_{t,t+s})\|_{p/2}$. In view of (18.19), the role of the first majorant term of (18.18) is taken by (18.59). For the second term, noting $\mathrm{E}^t_{t+s-m}X_t\mathrm{E}_tX_{t+s}$ is mixing of the same size as $\{\boldsymbol{V}_t\}$, the counterpart of (18.20) is

$$\begin{aligned}\|\mathrm{E}_{t+2(s-m)}(\mathrm{E}^t_{t+s-m}X_t\mathrm{E}_tX_{t+s}) - \sigma_{t,t+s}\|_{p/2}&\\ \le 6\alpha_{m-s}^{2/p-2/q}\|\mathrm{E}^t_{t+s-m}(X_t\mathrm{E}_tX_{t+s})\|_{q/2},&\end{aligned}\tag{18.60}$$

where

$$\begin{aligned}\|\mathrm{E}^t_{t+s-m}(X_t\mathrm{E}_tX_{t+s})\|_{q/2} &\le \|X_t\mathrm{E}_tX_{t+s}\|_{q/2}\\ &\le \|X_t\|_q\|\mathrm{E}_tX_{t+s}\|_q\\ &\le \|X_t\|_q\|X_{t+s}\|_r\zeta_s.\end{aligned}\tag{18.61}$$

Writing $\xi_m = 6\alpha_{m-s}^{2/p-2/q}\zeta_s + 2\nu_{m-s}\zeta_s$, since $2/p - 2/q = 1/q - 1/r$ it can be verified that $\xi_m = O(m^{-a})$. Thus, combining (18.60) and (18.61) with (18.59) in the manner of (18.18), also noting (18.56) in the case $m \le s$, shows that $\{X_t\mathrm{E}_tX_{t+s} - \sigma_{t,t+s}\}$ is an $L_{p/2}$ mixingale. The conclusion

$$\begin{aligned}\|\mathrm{E}_{t+s-m}(X_tX_{t+s}) - \sigma_{t,t+s}\|_{p/2} &= \|\mathrm{E}_{t+s-m}(X_t\mathrm{E}_tX_{t+s} - \sigma_{t,t+s})\|_{p/2}\\ &\le \|X_t\|_r\|X_{t+s}\|_r\xi_m\end{aligned}\tag{18.62}$$

now follows by the LIE.

For part (ii), in (18.54) $\zeta_m = 2\phi_k^{1-1/r} + \nu_k$ and since $-a(1-1/r)r/(r-2) < -a$, $\zeta_m = O(\nu_m) = O(m^{-a})$. The argument now proceeds in the same way as before, with $q = r$, until (18.60) which in view of (18.23) becomes

$$\begin{aligned}\|\mathrm{E}_{t+2(s-m)}(\mathrm{E}^t_{t+s-m}X_t\mathrm{E}_tX_{t+s}) - \sigma_{t,t+s}\|_{p/2}&\\ \le 2\phi_{m-s}^{1-2/r}\|\mathrm{E}^t_{t+s-m}(X_t\mathrm{E}_tX_{t+s})\|_{r/2}.&\end{aligned}\tag{18.63}$$

Under the stated conditions, $\phi_{m-s}^{1-2/r} = O(m^{-a})$ and the same argument as before now leads to (18.62) with $\xi_m = 2\phi_{m-s}^{1-2/r}\zeta_s + 2\nu_{m-s}\zeta_s = O(m^{-a})$. ∎

18.5 Approximability

Most of the applications to be studied in later chapters exploit the fact that L_p-NED functions of mixing processes are L_p-mixingales. Another way to look at the L_p-NED property is in terms of the existence of a finite lag approximation to the process. The conditional mean $\mathrm{E}(X_t|\mathcal{F}_{t-m}^{t+m})$ can be thought of as a function of the variables $\boldsymbol{V}_{t-m}, \ldots, \boldsymbol{V}_{t+m}$, and if $\{\boldsymbol{V}_t\}$ is a mixing sequence so is $\{\mathrm{E}(X_t|\mathcal{F}_{t-m}^{t+m})\}$ by **15.1**. This is the trick used in **18.6**, in particular. One way to overcome the problem that lag functions of mixing processes need not be mixing is to team a limit theorem for mixing processes with a proof that the difference between the actual sequence and its approximating sequence can be neglected.

However, the conditional mean might not be the only function to possess the desired approximability property. More generally, consider a definition of the following sort. Letting the vector $\boldsymbol{V}_t$ be $l \times 1$, think of $h_t^m : \mathbb{R}^{l(2m+1)} \mapsto \mathbb{R}$ as a $\mathcal{F}_{t-m}^{t+m}/\mathcal{B}$-measurable function, where $\mathcal{F}_{t-m}^{t+m} = \sigma(\boldsymbol{V}_{t-m}, \ldots, \boldsymbol{V}_{t+m})$.

18.19 Definition The sequence $\{X_t\}$ will be called L_p-approximable ($p > 0$) on the sequence $\{\boldsymbol{V}_t\}$ if for each $m \in \mathbb{N}$ there exists a sequence $\{h_t^m\}$ of $\mathcal{F}_{t-m}^{t+m}$-measurable random variables and

$$\|X_t - h_t^m\|_p \leq d_t \nu_m, \tag{18.64}$$

where $\{d_t\}$ is a non-negative constant sequence and $\nu_m \to 0$ as $m \to \infty$. $\{X_t\}$ will also be said to be *approximable in probability* (or L_0-approximable) on $\{\boldsymbol{V}_t\}$ if there exist $\{h_t^m\}$, $\{d_t\}$, and $\{\nu_m\}$ as above such that for every $\delta > 0$,

$$P(|X_t - h_t^m| > d_t\delta) \leq \nu_m. \quad \square \tag{18.65}$$

The usual size terminology can be applied here. There is also the usual extension to arrays, by inclusion of the additional subscript wherever appropriate.

If a sequence is L_p-approximable for $p > 0$, then by the Markov inequality

$$P(|X_t - h_t^m| > d_t\delta) \leq (d_t\delta)^{-p}\|X_t - h_t^m\|_p^p \leq \nu_m', \tag{18.66}$$

where $\nu_m' = \delta^{-p}\nu_m^p$; hence an L_p-approximable process is also L_0-approximable. An L_p-NED sequence is L_p-approximable, although only in the case $p = 2$ can it be claimed (from **10.12**) that $\mathrm{E}(X_t|\mathcal{F}_{t-m}^{t+m})$ is the best L_p-approximator in the sense that the p-norms in (18.64) are smaller than for any alternative choice of h_t^m.

18.20 Example Consider the moving average process of **18.3**. The function

$$h_t^m = \sum_{j=-m}^{m} \theta_j V_{t-j} \tag{18.67}$$

is different from $\mathrm{E}(X_t|\mathcal{F}_{t-m}^{t+m})$ unless $\{V_t\}$ is an independent process, but is also an L_p-approximator for X_t since

$$\|X_t - h_t^m\|_p = \left\| \sum_{j=m+1}^{\infty} (\theta_j V_{t-j} + \theta_{-j} V_{t+j}) \right\|_p \le d_t \nu_m, \tag{18.68}$$

where $\nu_m = \sum_{j=m+1}^{\infty} (|\theta_j| + |\theta_{-j}|)$ and $d_t = \sup_t \|V_t\|_p$. □

18.21 Example In **18.5**, the functions g_t^m are L_p-approximators for X_t of infinite size whenever $\sup_{s\le t}\|F_s V_s\|_p < \infty$. □

One reason why approximability might have advantages over the L_p-NED property is the ease of handling transformations. The results of §18.3 show that transferring the L_p-NED property to transformations of the original functions can present difficulties and impose undesirable moment restrictions. With approximability, these difficulties can be largely overcome. The first step is to show that subject to L_r-boundedness, an L_0-approximable sequence is also L_p-approximable for $p < r$ and the approximator functions can be bounded for each finite m. The following is adapted from [149].

18.22 Theorem If $\{X_t\}$ is L_r-bounded for $r > 1$ and L_0-approximable by h_t^m, it is L_p-approximable for $0 < p < r/2$ by

$$\tilde{h}_t^m = h_t^m 1_{\{|h_t^m| \le d_t M_m\}}$$

where $M_m < \infty$ for each $m \in \mathbb{N}$.

Proof Since h_t^m is an L_0-approximator of X_t there exists a positive sequence $\{\delta_m\}$ such that $\delta_m \to 0$ and yet

$$P(|X_t - h_t^m| > d_t \delta_m) \le \nu_m^{pr/(r-p)} \to 0 \tag{18.69}$$

as $m \to \infty$. Also, a sequence of numbers $\{M_m\}$ can be chosen having the properties $M_m \to \infty$ but $M_m \nu_m \to 0$. For example, $M_m = \nu_m^{-1/2}$ would serve. There is no loss of generality in assuming $\sup_m M_m^{-1} \le 1$. By Minkowski's inequality,

$$\|X_t - \tilde{h}_t^m\|_p \le A_{tm}^1 + A_{tm}^2 + A_{tm}^3 \tag{18.70}$$

where

$$\begin{aligned}
A^1_{tm} &= \|(X_t - \tilde{h}^m_t)1_{\{|X_t - h^m_t| > d_t\delta_m\}}\|_p \\
A^2_{tm} &= \|(X_t - \tilde{h}^m_t)1_{\{|X_t - h^m_t| \le d_t\delta_m, |h^m_t| > d_tM_m\}}\|_p \\
A^3_{tm} &= \|(X_t - \tilde{h}^m_t)1_{\{|X_t - h^m_t| \le d_t\delta_m, |h^m_t| \le d_tM_m\}}\|_p.
\end{aligned}$$

These norms may be decomposed by applying the Hölder inequality in the form

$$\|XY\|_p \le \|X\|_{pq}\|Y\|_{pq/(q-1)}, q > 1 \tag{18.71}$$

to A^i_{mt} for $i = 1, 2, 3$, setting $q = r/p$. Also by Minkowski's inequality and the definition of $\tilde{h}^m_t$,

$$\|X_t - \tilde{h}^m_t\|_r \le \|X_t\|_r + d_tM_m. \tag{18.72}$$

First, noting that $\|1_E\|_{pq/(q-1)} = P(E)^{1/p-1/r}$, (18.71) and then (18.72) together with (18.69) imply

$$\begin{aligned}
A^1_{tm} &\le \|X_t - \tilde{h}^m_t\|_r P(|X_t - h^m_t| > d_t\delta_m)^{1/p-1/r} \\
&\le d_t(\|X_t/d_t\|_r M_m^{-1} + 1)M_m\nu_m.
\end{aligned} \tag{18.73}$$

Second, for any x, y note that if $|x - y| \le A$ and $|y| > B$ then $|x| > B - A$. This is immediate if $|x| \ge |y|$, otherwise use $A \ge |x - y| \ge |y| - |x| > B - |x|$. It follows that

$$\{|X_t - h^m_t| \le d_t\delta_m, |h^m_t| > d_tM_m\} \subseteq \{|X_t| > d_t(M_m - \delta_m)\}$$

so that when $M_m > \delta_m$ the Markov inequality gives

$$\begin{aligned}
P(|X_t - h^m_t| \le d_t\delta_m, |h^m_t| > d_tM_m) &\le P(|X_t| > d_t(M_m - \delta_m)) \\
&\le \|X_t\|^r_r d_t^{-r}(M_m - \delta_m)^{-r}.
\end{aligned} \tag{18.74}$$

Thus, (18.71) and then (18.72) together with (18.74) give

$$\begin{aligned}
A^2_{tm} &\le \|X_t - \tilde{h}^m_t\|_r P(|X_t - h^m_t| \le d_t\delta_m, |h^m_t| > d_tM_m)^{1/p-1/r} \\
&\le (\|X_t\|_r + d_tM_m)\|X_t\|_r^{r/p-1} d_t^{1-r/p}(M_m - \delta_m)^{1-r/p} \\
&\le d_t(\|X_t/d_t\|_r^{r/p} + \|X_t/d_t\|_r^{r/p-1})M_m(M_m - \delta_m)^{1-r/p},
\end{aligned} \tag{18.75}$$

where the final inequality is from replacing M_m^{-1} with 1. Lastly,

$$A_{tm}^3 \leq d_t\delta_m \tag{18.76}$$

in view of the fact that $\tilde{h}_t^m = h_t^m$ on the set $\{|h_t^m| \leq d_t M_m\}$. Therefore,

$$\|X_t - \tilde{h}_t^m\|_p \leq d_t'\nu_m' \tag{18.77}$$

where

$$d_t' = d_t \max\{\|X_t/d_t\|_r, \|X_t/d_t\|_r^{r/p} + \|X_t/d_t\|_r^{r/p-1}, 1\} \tag{18.78}$$

and

$$\nu_m' = M_m\nu_m + M_m(M_m - \delta_m)^{1-r/p} + \delta_m, \tag{18.79}$$

so that $\nu_m' \to 0$ by assumption, noting $1 - r/p < -1$. ∎

If L_0-approximability is satisfied with $d_t \ll \|X_t\|_r$, then $d_t' = d_t$.

The value of this result is that to preserve L_p-approximability under the transformation, it only has to be shown that the transformation of an L_0-approximable variable is also L_0-approximable, and then to establish the existence of the requisite moments. Consider the Lipschitz condition specified in (18.37). The conditions that need to be imposed on $B(\cdot,\cdot)$ are notably weaker than those in **18.16** for the L_p-NED case. This is shown as follows.

18.23 Theorem Let $\boldsymbol{h}_t^m = (h_{1t}^m, \ldots, h_{kt}^m)'$ be the L_0-approximator of $\boldsymbol{X}_t = (X_{1t}, \ldots, X_{kt})'$ with approximability indices $\boldsymbol{\nu}_t$ $(k \times 1)$ of size $-\varphi$ and scale constants $\boldsymbol{d}_t$ $(k \times 1)$. If $\phi_t : \mathbb{R}^k \mapsto \mathbb{R}$ satisfies (18.37) and $\mathrm{E}(B_t(\boldsymbol{X}_t, \boldsymbol{h}_t^m)^\varepsilon) < \infty$ for $\varepsilon > 0$, then $\phi_t(\boldsymbol{X}_t)$ is L_0-approximable of size $-\varphi$.

Proof Fix $\delta > 0$ and $M > 0$ and define $d_t = \sum_{i=1}^k d_{it}$. The Markov inequality gives

$$\begin{aligned} &P(|\phi_t(\boldsymbol{X}_t) - \phi_t(\boldsymbol{h}_t^m)| > d_t\delta) \\ &\qquad \leq P(B_t(\boldsymbol{X}_t, \boldsymbol{h}_t^m)\rho(\boldsymbol{X}_t, \boldsymbol{h}_t^m) > d_t\delta, B_t(\boldsymbol{X}_t, \boldsymbol{h}_t^m) > M) \\ &\qquad\qquad + P(B_t(\boldsymbol{X}_t, \boldsymbol{h}_t^m)\rho(\boldsymbol{X}_t, \boldsymbol{h}_t^m) > d_t\delta, B_t(\boldsymbol{X}_t, \boldsymbol{h}_t^m) \leq M) \\ &\qquad \leq \mathrm{E}(B_t(\boldsymbol{X}_t, \boldsymbol{h}_t^m)^\varepsilon)/M^\varepsilon + P(\rho(\boldsymbol{X}_t, \boldsymbol{h}_t^m) > d_t\delta/M). \end{aligned} \tag{18.80}$$

Since M is arbitrary the first term on the majorant side can be made as small as desired. The proof is completed by noting that

$$\begin{aligned}P(\rho(\boldsymbol{X}_t,\boldsymbol{h}_t^m) > d_t\delta/M) &= P\Big(\sum_{i=1}^{k}|X_{it}-h_{it}^m| > d_t\delta/M\Big)\\ &\le P\Big(\bigcup_{i=1}^{k}\{|X_{it}-h_{it}^m| > d_{it}\delta/M\}\Big)\\ &\le \sum_{i=1}^{k}P(|X_{it}-h_{it}^m| > d_{it}\delta/M)\\ &\le \sum_{i=1}^{k}\nu_{im} \to 0 \text{ as } m\to\infty. \quad \blacksquare \qquad (18.81)\end{aligned}$$

Given an L_r-bounded, L_2-NED sequence with $r > 4$ which is accordingly L_2-approximable and hence L_0-approximable, it might appear that any transformation satisfying the conditions of **18.23** is L_2-approximable by **18.22** and therefore also L_2-NED, by **10.12**. This argument would circumvent the need to check the moment restrictions of **18.16**. The catch is that it is not possible to specify the L_2-NED size of the transformed sequence. In (18.79) it is not possible to put a bound on the rate at which the sequence $\{\delta_m\}$ may converge without specifying the distributions of the X_{it} in greater detail. However, if it *is* possible to do this in a given application, here is an alternative route to dealing with transformations.

Pötscher and Prucha ([149]), to whom the concepts of this section are due, define approximability in a slightly different way, in terms of the convergence of the Cesàro sums of the p-norms or probabilities. These authors say that X_t is L_p-approximable ($p > 0$) if

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\|X_t-h_t^m\|_p \to 0 \text{ as } m\to\infty \qquad (18.82)$$

and is L_0-approximable if, for every $\delta > 0$,

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}P(|X_t-h_t^m| > \delta) \to 0 \text{ as } m\to\infty. \qquad (18.83)$$

It is clear that near-epoch dependence and the mixingale property might be defined in an analogous manner, leading to a whole class of alternative convergence results. Comparing these alternatives, it turns out that neither definition dominates, each permitting a form of behaviour by the sequences which is ruled out by the other. If (18.64) holds,

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\|X_t - h_t^m\|_p \leq \limsup_{n\to\infty}\left(\frac{1}{n}\sum_{t=1}^{n} d_t\right)\nu_m \to 0 \tag{18.84}$$

so long as the limsup on the majorant side is bounded. On the other hand, if (18.82) holds, define

$$\nu_m = \limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\|X_t - h_t^m\|_p \tag{18.85}$$

and then $d_t = \sup_m\{\|X_t - h_t^m\|_p/\nu_m\}$ will satisfy **18.19** *so long as* it is finite for finite t. Evidently, (18.82) permits the existence of a set of sequence coordinates for which the p-norms fail to converge to 0 with m, so long as these are ultimately negligible, accumulating at a rate strictly less than n as n increases. On the other hand (18.64) permits trending moments, with for example $d_t = O(t^\lambda)$, $\lambda > 0$, which would contradict (18.82).

Similarly, for $\delta_m > 0$ and $\nu_m > 0$, define d_{tm} by the relation

$$P(|X_t - h_t^m| > d_{tm}\delta_m) = \nu_m, \tag{18.86}$$

and then, allowing $\nu_m \to 0$ and $\delta_m \to 0$, define $d_t = \sup_m d_{tm}$. (18.65) is satisfied if $d_t < \infty$ for each finite t; this latter condition need not hold under (18.83). On the other hand, (18.83) could fail in cases where, for *fixed* δ and every m, $P(|X_t - h_t^m| > \delta)$ is tending to unity as $t \to \infty$.

18.6 NED in Volatility

Models of time-varying conditional volatility are typically constructed to explain financial series such as stock returns that are unpredictable in levels. These applications are instructive by showing how near-epoch dependence measures general dependence, in particular dependence that is not measured by autocorrelations.

Consider an adapted process $\{X_t, \mathcal{F}_t\}$ where $X_t = \sigma_t U_t$ and U_t is L_r-bounded for $r \geq 2$ and identically and independently distributed with mean 0 and variance 1. σ_t^2 is an $\mathcal{F}_{t-1}$-measurable non-negative process that here may be called the conditional variance, although in a general context caution is needed with this interpretation. Assume the existence of fourth moments and of a positive constant γ such that $\sigma_t^2 \geq \gamma$ almost surely. Then, since $\mathrm{E}(X_t|\mathcal{F}_{t-m}^{t+m}) = U_t\mathrm{E}(\sigma_t|\mathcal{F}_{t-m}^{t+m})$ a.s. for any $m \geq 1$ and $\mathrm{E}U_t^2 = 1$, in the L_2-NED case

$$\begin{aligned}\|X_t - \mathrm{E}(X_t|\mathcal{F}_{t-m}^{t+m})\|_2 &= \|\sigma_t - \mathrm{E}(\sigma_t|\mathcal{F}_{t-m}^{t+m})\|_2 \\ &\le \|\sigma_t - \mathrm{E}(\sigma_t^2|\mathcal{F}_{t-m}^{t+m})^{1/2}\|_2 \\ &\le \gamma^{-1/2}\|\sigma_t^2 - \mathrm{E}(\sigma_t^2|\mathcal{F}_{t-m}^{t+m})\|_2 \end{aligned} \tag{18.87}$$

where the first inequality is because the conditional expectation is the MSE predictor. Hence the L_2-NED characteristics of the 'levels' process depend on those of the conditional variance and it suffices to prove the NED property for this process. By contrast the L_1-NED case does not call for fourth moments. An approximator for X_t itself is not so easy to construct in this case, but since X_t is serially uncorrelated there is in fact no need to show that it has limited memory in levels. To show limited memory in volatility means proving the result for X_t^2 and given the properties of U_t,

$$\|X_t^2 - \mathrm{E}(X_t^2|\mathcal{F}_{t-m}^{t+m})\|_1 = \|\sigma_t^2 - \mathrm{E}(\sigma_t^2|\mathcal{F}_{t-m}^{t+m})\|_1. \tag{18.88}$$

The so-called GARCH(1,1) is probably the most commonly applied example of this class of model, the conditional variance equation taking the form

$$\sigma_t^2 = \gamma + \alpha X_{t-1}^2 + \beta\sigma_{t-1}^2 \tag{18.89}$$

where γ and α are positive parameters and $0 \le \beta < 1$. The process is stationary and has a finite unconditional variance if $\alpha + \beta < 1$, since in that case

$$\mathrm{E}(X_t^2) = \mathrm{E}(\sigma_t^2) = \gamma + \alpha\mathrm{E}(X_{t-1}^2) + \beta\mathrm{E}(\sigma_{t-1}^2) = \frac{\gamma}{1-\alpha-\beta} < \infty. \tag{18.90}$$

GARCH stands for 'generalized autoregressive conditional heteroscedasticity'. Writing $V_t = (U_t^2 - 1)\sigma_t^2$, rearrangement of (18.89) produces

$$X_t^2 = \gamma + (\alpha+\beta)X_{t-1}^2 + V_t - \beta V_{t-1},$$

so GARCH(1,1) may be viewed as an ARMA(1,1) process in the squares, although while V_t is an uncorrelated process it is *not* independent. The GARCH(p, q) class where p and q represent the orders of lag are in turn members of the so-called ARCH(∞) class of processes. These have the general form

$$\sigma_t^2 = \gamma + \sum_{j=1}^{\infty}\theta_j X_{t-j}^2 \tag{18.91}$$

so that GARCH(1,1) in solved form has $\theta_j = \alpha\beta^{j-1}$. There is a large literature on these models and for simplicity of exposition, not least because the formulae involved are not excessively complex, the GARCH(1,1) is used to illustrate the derivation of NED properties.

18.24 Theorem The GARCH(1,1) model is

(i) Geometrically L_1-NED if $\alpha + \beta < 1$.

(ii) Geometrically L_2-NED if $\mathrm{E}(U_t^4) = \mu_4 < \infty$ and $\mu_4\alpha^2 + 2\alpha\beta + \beta^2 < 1$.

Proof In view of (18.88) and (18.87) respectively, it suffices to show the results for the process σ_t^2. To solve the difference equation (18.89), set $Z_t = \alpha U_t^2 + \beta$ and apply recursive substitution to obtain

$$\begin{aligned}\sigma_t^2 &= \gamma + \sigma_{t-1}^2 Z_{t-1}\\ &= \gamma + \gamma Z_{t-1} + \sigma_{t-2}^2 Z_{t-1} Z_{t-2}\\ &= \cdots\\ &= \gamma\left(1 + \sum_{j=1}^{m-1}\prod_{i=1}^{j} Z_{t-i}\right) + \sigma_{t-m}^2 \prod_{i=1}^{m} Z_{t-i}.\end{aligned} \tag{18.92}$$

The Z_{t-i} for $i = 1, \dots, m$ are non-negative, $\mathcal{F}_{t-m}^{t+m}$-measurable, independently and identically distributed, and in particular independent of σ_{t-m}^2. Hence,

$$\sigma_t^2 - \mathrm{E}(\sigma_t^2 | \mathcal{F}_{t-m}^{t+m}) = \left(\sigma_{t-m}^2 - \mathrm{E}(\sigma_{t-m}^2 | \mathcal{F}_{t-m}^{t+m})\right)\prod_{i=1}^{m} Z_{t-i}.$$

It follows, applying (18.90) and noting $\mathrm{E}(Z_{t-i}) = \alpha + \beta$ for each i, that

$$\begin{aligned}||\sigma_t^2 - \mathrm{E}(\sigma_t^2 | \mathcal{F}_{t-m}^{t+m})||_1 &= \mathrm{E}|\sigma_{t-m}^2 - \mathrm{E}(\sigma_{t-m}^2 | \mathcal{F}_{t-m}^{t+m})| \prod_{i=1}^{m} \mathrm{E}Z_{t-i}\\ &\le 2\mathrm{E}(\sigma_{t-m}^2)(\alpha + \beta)^m\\ &= \frac{2\gamma}{1 - \alpha - \beta}(\alpha + \beta)^m.\end{aligned} \tag{18.93}$$

Since $\alpha + \beta < 1$, this formula shows geometric convergence, proving part (i).

Under the assumptions of part (ii), note that $\alpha + \beta < 1$ follows from $\mathrm{E}(Z_t^2) = \mu_4\alpha^2 + 2\alpha\beta + \beta^2 < 1$, since $\mu_4 \ge 1$. $\mathrm{E}(X_t^4) = \mu_4 \mathrm{E}(\sigma_t^4)$ and by stationarity,

$$\begin{aligned}\mathrm{E}(\sigma_t^4) &= \gamma^2 + 2\gamma \mathrm{E}(Z_{t-1})\mathrm{E}(\sigma_{t-1}^2) + \mathrm{E}(Z_{t-1}^2)\mathrm{E}(\sigma_{t-1}^4) \\ &= \frac{\gamma^2 + 2\gamma \mathrm{E}(Z_t)\mathrm{E}(\sigma_t^2)}{1 - \mathrm{E}(Z_t^2)} \\ &= \frac{\gamma^2(1+\alpha+\beta)}{(1-\alpha-\beta)(1-\mu_4\alpha^2 - 2\alpha\beta - \beta^2)}.\end{aligned}$$

Hence, $||\sigma_t^2||_2 < \infty$ not depending on t and arguing similarly to (18.93),

$$\begin{aligned}||\sigma_t^2 - \mathrm{E}(\sigma_t^2|\mathcal{F}_{t-m}^{t+m})||_2 &= ||\sigma_{t-m}^2 - \mathrm{E}(\sigma_{t-m}^2|\mathcal{F}_{t-m}^{t+m})||_2 \prod_{i=1}^{m} ||Z_{t-i}||_2 \\ &\le 2||\sigma_t^2||_2(\mu_4\alpha^2 + 2\alpha\beta + \beta^2)^{m/2}.\end{aligned}$$

This completes the proof of (ii). ∎

These conditions for NED to hold have required that the second and fourth moments are finite, respectively. These are the covariance stationarity and fourth-order stationarity conditions and either is a sufficient condition for the GARCH process to be stationary in the strict sense. They are not necessary conditions for stationarity, although they are necessary for NED because the NED property requires integrability. So-called IGARCH (integrated GARCH), where $\alpha + \beta = 1$, appears to be characteristic of many financial time series and with suitable coefficient values it is stationary and has short memory, just like covariance-stationary GARCH. The stationarity and ergodicity of the GARCH(1,1) process was investigated by Daniel Nelson ([135]) who showed that a necessary and sufficient condition is

$$\mathrm{E}(\log(\alpha U_t^2 + \beta)) < 0. \tag{18.94}$$

This is a condition on the distribution of the shocks as well as on the parameters, but as Nelson showed for the case where U_t is Gaussian, $\alpha + \beta \ge 1$ is permissible provided $\beta < 1$. The same condition must determine the memory of the sequence but this cannot be measured in terms of conditional variances, remembering that these exist only if X_t^2 is integrable. The process designated σ_t^2 as in (18.89), while always a valid measure of time-varying volatility, may only be equated with $\mathrm{E}(X_t^2|\mathcal{F}_{t-1})$ when $\mathrm{E}(X_t^2) < \infty$.

However, while in these cases the NED property is not available, what can be done is to show L_0-approximability. Showing limited memory in volatility means proving the result for X_t^2 as before, although in this case there is no counterpart for (18.88).

18.25 Theorem If (18.94) holds, the GARCH(1,1) model is geometrically L_0-approximable.

Proof Initially, consider the L_0-approximability of σ_t^2. Let the $\mathcal{F}_{t-m}^{t+m}$-measurable approximation to σ_t^2 be h_t^m where

$$h_t^m = \gamma\Big(1 + \sum_{j=1}^{m-1}\prod_{i=1}^{j} Z_{t-i}\Big)$$

so that

$$\sigma_t^2 - h_t^m = \sigma_{t-m}^2 \prod_{i=1}^{m} Z_{t-i}.$$

According to (18.65), under stationarity L_0-approximability requires $P(|\sigma_t^2 - h_t^m| > \delta) \to 0$ as $m \to \infty$ for any $\delta > 0$, where in this case geometric convergence is expected. Equivalently, noting that $\sigma_t^2 - h_t^m > 0$, consider the logarithm

$$\log(\sigma_t^2 - h_t^m) = \sum_{i=1}^{m} \log Z_{t-i} + \log \sigma_{t-m}^2. \tag{18.95}$$

By the assumption of (18.94) define $\zeta = \mathrm{E}(\log Z_t) < 0$ and also let $\tau^2 = \mathrm{Var}(\log Z_t)$. When m is large, the collection $(\log Z_{t-i} - \zeta)/\tau$ for $i = 1, \ldots, m$ are i.i.d. random variables with zero mean and unit variance. The condition $\log \sigma_{t-m}^2 = O_p(1)$ follows from the noted fact that X_t is stationary and ergodic under (18.94). Therefore, (18.95) shows that $(\log(\sigma_t^2 - h_t^m) - m\zeta)/\tau m^{1/2}$ is approximately standard normal when m is large. Also, since $\zeta < 0$, for any choice of δ there exists m large enough that $(\log\delta - m\zeta)/\tau m^{1/2} \geq 1/\sqrt{2\pi}$. It follows by (8.22) that for such an m,

$$\begin{aligned} P(|\sigma_t^2 - h_t^m| > \delta) &= P\Big(\frac{\log(\sigma_t^2 - h_t^m) - m\zeta}{\tau m^{1/2}} > \frac{\log\delta - m\zeta}{\tau m^{1/2}}\Big) \\ &\leq \exp\Big\{-\frac{(\log\delta - m\zeta)^2}{2\tau^2 m}\Big\} \\ &= O(\mathrm{e}^{-m\zeta^2/2\tau^2}), \end{aligned} \tag{18.96}$$

noting that the terms in the exponent containing $\log\delta$ are $O(1)$ as $m \to \infty$.

Now consider $X_t^2 = U_t^2\sigma_t^2$. The $\mathcal{F}_{t-m}^{t+m}$-measurable approximation to $U_t^2\sigma_t^2$ is of the form $U_t^2 h_t^m$, so it is required to bound $P(U_t^2|\sigma_t^2 - h_t^m| > \delta)$. Set $Y = \log\delta - \log U_t^2$ and consider the convolution of the independent r.v.s Y and $\log(\sigma_t^2 - h_t^m)$ (compare **11.1**). Similarly to (18.96) with Y replacing $\log\delta$ and F_Y denoting its distribution function,

$$
\begin{aligned}
P(U_t^2(\sigma_t^2 - h_t^m) > \delta) &= P(\log(\sigma_t^2 - h_t^m) > Y) \\
&= \int P(\log(\sigma_t^2 - h_t^m) > y)\mathrm{d}F_Y(y) \\
&\leq \mathrm{e}^{-m\zeta^2/2\tau^2} \int \exp\left\{\frac{\zeta y}{\tau^2} - \frac{y^2}{2m\tau^2}\right\}\mathrm{d}F_Y(y) \\
&= O(\mathrm{e}^{-m\zeta^2/2\tau^2}). \quad \blacksquare
\end{aligned}
$$

For derivation of NED and approximability conditions for ARCH(∞) models, see for example [44]. The analysis is substantially more complicated than that given here, entailing Volterra-type expansions of the infinite-order nonlinear moving averages, but the basic arguments are similar. Also see [48] for the extension of the Nelson-type analysis of stationarity/ergodicity conditions to the non-integrable ARCH(∞) case.

PART IV

THE LAW OF LARGE NUMBERS

19
Stochastic Convergence

19.1 Almost Sure Convergence

Almost sure convergence was defined formally in §12.2. Sometimes the condition is stated in the form

$$P\Big(\limsup_{n\to\infty}|X_n - X| > \varepsilon\Big) = 0, \text{ for all } \varepsilon > 0. \tag{19.1}$$

Yet another way to express the same idea is to say that $P(C) = 1$ where for each $\omega \in C$ and any $\varepsilon > 0$, $|X_n(\omega) - X(\omega)| > \varepsilon$ for at most a finite number of the sequence coordinates. This condition is also written as

$$P(|X_n - X| > \varepsilon, \text{ i.o.}) = 0, \text{ all } \varepsilon > 0 \tag{19.2}$$

where i.o. stands for 'infinitely often'.

Note that the probability in (19.2) is assigned to an attribute of the whole sequence, not to a particular n. One way to grasp the 'infinitely often' idea is to consider the event $\bigcup_{n=m}^{\infty}\{|X_n - X| > \varepsilon\}$; in words, 'the event that has occurred whenever $\{|X_n - X| > \varepsilon\}$ occurs for at least one n beyond a given point m in the sequence'. If this event occurs for every m, no matter how large, $\{|X_n - X| > \varepsilon\}$ occurs infinitely often. In other words,

$$\begin{aligned}\{|X_n - X| > \varepsilon, \text{ i.o.}\} &= \bigcap_{m=1}^{\infty}\bigcup_{n=m}^{\infty}\{|X_n - X| > \varepsilon\}\\ &= \limsup_{n\to\infty}\{|X_n - X| > \varepsilon\}.\end{aligned} \tag{19.3}$$

Useful facts about this set and its complement are contained in the following lemma.

19.1 Lemma Let $\{E_n \in \mathcal{F}\}_1^\infty$ be an arbitrary sequence. Then

(i) $P\Big(\limsup_{n\to\infty} E_n\Big) = \lim_{n\to\infty} P\Big(\bigcup_{m=n}^{\infty} E_m\Big)$

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0019

(ii) $P\Big(\liminf_{n\to\infty} E_n\Big) = \lim_{n\to\infty} P\Big(\bigcap_{m=n}^{\infty} E_m\Big)$.

Proof The sequence $\{\bigcup_{n=m}^{\infty} E_n\}_{m=1}^{\infty}$is decreasing monotonically to $\limsup_n E_n$. Part (i) therefore follows by **3.7**. Part (ii) follows in exactly the same way, since the sequence $\{\bigcap_{n=m}^{\infty} E_n\}_{m=1}^{\infty}$ increases monotonically to $\liminf_n E_n$. ∎

A fundamental tool in proofs of a.s. convergence is the *Borel–Cantelli lemma.* This has two parts, the 'convergence' part and the 'divergence' part. The former is the most useful since it yields a very general sufficient condition for convergence. The second part, which generates a necessary condition for convergence, requires independence of the sequence.

19.2 Lemma (Borel–Cantelli)

(i) For an arbitrary sequence of events $\{E_n \in \mathcal{F}\}_1^{\infty}$,

$$\sum_{n=1}^{\infty} P(E_n) < \infty \Rightarrow P(E_n \text{ i.o.}) = 0. \tag{19.4}$$

(ii) For a sequence $\{E_n \in \mathcal{F}\}_1^{\infty}$ of totally independent events,

$$\sum_{n=1}^{\infty} P(E_n) = \infty \Rightarrow P(E_n \text{ i.o.}) = 1. \tag{19.5}$$

Proof By countable subadditivity,

$$P\Big(\bigcup_{n=m}^{\infty} E_n\Big) \leq \sum_{n=m}^{\infty} P(E_n). \tag{19.6}$$

The premise in (19.4) is that the majorant side of (19.6) is finite for $m = 1$. This implies $\sum_{n=m}^{\infty} P(E_n) \to 0$ as $m \to \infty$ (by **2.15**), which further implies

$$\lim_{m\to\infty} P\Big(\bigcup_{n=m}^{\infty} E_n\Big) = 0. \tag{19.7}$$

Part (i) now follows by part (i) of **19.1**.

To prove (ii), note by **7.5** that the collection $\{E_n^c \in \mathcal{F}\}_1^{\infty}$ is independent; hence for any $m > 0$ and $m' > m$,

$$P\Big(\bigcap_{n=m}^{m'} E_n^c\Big) = \prod_{n=m}^{m'} P(E_n^c) = \prod_{n=m}^{m'}(1 - P(E_n))$$
$$\leq \exp\Big\{-\sum_{n=m}^{m'} P(E_n)\Big\} \to 0 \text{ as } m' \to \infty \tag{19.8}$$

by hypothesis, since $e^{-x} \geq 1 - x$. (19.8) holds for all m, so

$$P\big(\liminf_{n\to\infty} E_n^c\big) = \lim_{m\to\infty} P\Big(\bigcap_{n=m}^{\infty} E_n^c\Big) = 0 \tag{19.9}$$

by **19.1**(ii). Hence by **1.1**(iii),

$$P(E_n \text{ i.o.}) = P\big(\limsup_{n\to\infty} E_n\big) = 1 - P\big(\liminf_{n\to\infty} E_n^c\big) = 1. \quad \blacksquare \tag{19.10}$$

To appreciate the role of this result (the convergence part) in showing a.s. convergence, consider the particular case

$$E_n = \{\omega : |X_n(\omega) - X(\omega)| > \varepsilon\}.$$

If $\sum_{n=1}^{\infty} P(E_n) < \infty$, the condition $P(E_n) > 0$ can hold for at most a finite number of n. The lemma shows that $P(E_n$ i.o.) has to be zero to avoid a contradiction.

Yet another way to characterize a.s. convergence is suggested by the following theorem.

19.3 Theorem $\{X_n\}$ converges a.s. to X iff for all $\varepsilon > 0$,

$$\lim_{m\to\infty} P\Big(\sup_{n\geq m} |X_n - X| \leq \varepsilon\Big) = 1. \tag{19.11}$$

Proof Let

$$A_m(\varepsilon) = \bigcap_{n=m}^{\infty} \{\omega : |X_n(\omega) - X(\omega)| \leq \varepsilon\} \in \mathcal{F} \tag{19.12}$$

and then (19.11) can be written in the form $\lim_{m\to\infty} P(A_m(\varepsilon)) = 1$. The sequence $\{A_m(\varepsilon)\}_1^{\infty}$ is non-decreasing, so $A_m(\varepsilon) = \bigcup_{j=1}^{m} A_j(\varepsilon)$; letting $A(\varepsilon) = \bigcup_{m=1}^{\infty} A_m(\varepsilon)$, (19.11) can be stated as $P(A(\varepsilon)) = 1$.

Define the set C by the property that, for each $\omega \in C$, $\{X_n(\omega)\}_1^\infty$ converges. That is, for $\omega \in C$,

$$\exists\ m(\omega) \text{ such that } \sup_{n \geq m(\omega)} |X_n(\omega) - X(\omega)| \leq \varepsilon, \text{ for all } \varepsilon > 0. \tag{19.13}$$

Given $\varepsilon > 0$, $\omega \in C$ implies $\omega \in A_m(\varepsilon)$ for some m, so that $C \subseteq A(\varepsilon)$. Hence $P(C) = 1$ implies $P(A(\varepsilon)) = 1$. Since ε is arbitrary, this proves 'only if'.

To show 'if', assume $P(A(\varepsilon)) = 1$ for all $\varepsilon > 0$. Set $\varepsilon = 1/k$ for positive integer k and define

$$A^* = \bigcap_{k=1}^{\infty} A(1/k) = \left(\bigcup_{k=1}^{\infty} A(1/k)^c\right)^c \tag{19.14}$$

where the second equality is by **1.1** (iv). By **3.12**(ii), $P(A^*) = 1 - P\big(\bigcup_{k=1}^{\infty} A(1/k)^c\big) = 1$. But every element of A^* is a convergent outcome in the sense of (19.13), hence $A^* \subseteq C$ and the conclusion follows. ∎

The last theorem characterizes a.s. convergence in terms of the *uniform* proximity of the tail sequences $\{|X_n(\omega) - X(\omega)|\}_{n=m}^{\infty}$ to zero, on a set A_m whose measure approaches 1 as $m \to \infty$. A related but distinct result is Egorov's theorem. Recall from §2.4 the distinction between pointwise and uniform convergence. Uniform convergence on a set $C \subseteq \Omega$ means that for $\omega \in C$ and each $m > 0$, $|X_n(\omega) - X(\omega)| < 1/m$ for $n \geq k(m)$ where $k(m)$ is a function of m that does not depend on ω. The next result establishes a direct link between a.s. convergence of a random sequence and uniform convergence on subsets of Ω.

19.4 Theorem (Egorov) If $X_n \to_{\text{a.s.}} X$ there exists for every $\delta > 0$ a set $C(\delta)$ with $P(C(\delta)) \geq 1 - \delta$ such that $X_n(\omega) \to X(\omega)$ uniformly on $C(\delta)$.

Proof Define

$$A_m(\delta) = \bigcap_{n=k(m)}^{\infty} \{\omega : |X_n(\omega) - X(\omega)| < 1/m\} \tag{19.15}$$

where $k(m)$ is chosen to satisfy the condition $P(A_m(\delta)) \geq 1 - 2^{-m}\delta$. In view of a.s. convergence and **19.3**, the existence of finite $k(m)$ is assured for each m. By construction, convergence is uniform in the set

$$C(\delta) = \bigcap_{m=1}^{\infty} A_m(\delta). \tag{19.16}$$

Applying **1.1**(iii) and subadditivity gives

$$\begin{aligned} P\big(C(\delta)\big) &= 1 - P\Big(\bigcup_{m=1}^{\infty} A_m(\delta)^c\Big) \\ &\geq 1 - \sum_{m=1}^{\infty} \big(1 - P(A_m(\delta))\big) \\ &\geq 1 - \delta. \quad ∎ \end{aligned} \tag{19.17}$$

19.2 Convergence in Probability

In spite of its conceptual simplicity, the theory of almost sure convergence cannot easily be appreciated without a grasp of probability fundamentals, and traditionally an alternative convergence concept has been preferred in econometric theory. If for any $\varepsilon > 0$

$$\lim_{n\to\infty} P(|X_n - X| > \varepsilon) = 0 \tag{19.18}$$

then X_n is said to *converge in probability* (in pr.) to X. Here the convergent sequences are specified to be not random elements $\{X_n(\omega)\}_1^\infty$, but the nonstochastic sequences $\{P(|X_n - X| > \varepsilon)\}_1^\infty$. The probability of the convergent subset of Ω is left unspecified. However, the following relation is immediate from **19.3**, since (19.11) implies (19.18).

19.5 Theorem If $X_n \to_{\text{a.s.}} X$ then $X_n \to_{\text{pr}} X$. □

The converse does not hold. Convergence in probability imposes a limiting condition on the marginal distribution of the n^{th} member of the sequences as $n \to \infty$. The probability that the deviation of X_n from X is negligible approaches 1 as n increases. Almost sure convergence, on the other hand, requires that beyond a certain point in the sequence the probability that deviations are negligible *from there on* approaches 1. While it may not be intuitively obvious that a sequences can converge in pr. but not a.s., in **19.17** below it is shown that convergence in pr. is compatible with a.s. *non*convergence.

However, convergence in probability is equivalent to a.s. convergence *on a subsequence*; given a sequence that converges in pr. it is always possible, by throwing away some of the members of the sequence, to be left with an a.s. convergent sequence.

19.6 Theorem $X_n \to_{\text{pr}} X$ iff every subsequence $\{X_{nk}, k \in \mathbb{N}\}$ contains a further subsequence $\{X_{n_{k(j)}}, j \in \mathbb{N}\}$ that converges a.s. to X.

Proof To prove 'only if', suppose $P(|X_n - X| > \varepsilon) \to 0$ for any $\varepsilon > 0$. This means that, for any sequence of integers $\{n_k, k \in \mathbb{N}\}$, $P(|X_{n_k} - X| > \varepsilon) \to 0$. Hence for each $j \in \mathbb{N}$ there exists an integer $k(j)$ such that

$$P(|X_{n_k} - X| > 1/j) < 2^{-j}, \text{ all } k \geq k(j). \tag{19.19}$$

Since this sequence of probabilities is summable over j, it follows from the first Borel–Cantelli lemma that

$$P(|X_{n_{k(j)}} - X| > 1/j \text{ i.o.}) = 0. \tag{19.20}$$

It follows by consideration of the infinite subsequences $\{n_{k(j)}, j \geq 1/\varepsilon\}$ for every $\varepsilon > 0$ that $P(|X_{n_{k(j)}} - X| > \varepsilon \text{ i.o.}) = 0$ and hence the subsequence $\{X_{n_{k(j)}}\}$ converges a.s. as required.

To prove 'if': if $\{X_n\}$ does not converge in probability, there must exist a subsequence $\{n_k\}$ such that $\inf_k\{P(|X_{n_k} - X| > \varepsilon)\} \geq \varepsilon$, for some $\varepsilon > 0$. This rules out convergence in pr. on any subsequence of $\{n_k\}$, which rules out convergence a.s. on the same subsequence, by **19.5**. ∎

19.3 Transformations and Convergence

The following set of results on convergence, a.s. and in pr., are fundamental tools of asymptotic theory. For completeness they are given for the vector case, even though most applications in subsequent chapters are to scalar sequences. A random k-vector $\boldsymbol{X}_n$ is said to converge a.s. (in pr.) to a vector $\boldsymbol{X}$ if each element of $\boldsymbol{X}_n$ converges a.s. (in pr.) to the corresponding element of $\boldsymbol{X}$. The notation $||\boldsymbol{X}||$ here denotes the Euclidean norm $\sqrt{\boldsymbol{X}'\boldsymbol{X}}$, otherwise the length, of a k-vector $\boldsymbol{X}$.[1]

19.7 Lemma $\boldsymbol{X}_n \to \boldsymbol{X}$ a.s. (in pr.) iff $||\boldsymbol{X}_n - \boldsymbol{X}|| \to 0$ a.s. (in pr.).

Proof Take first the case of a.s. convergence. The relation $||\boldsymbol{X}_n - \boldsymbol{X}|| \to_{\text{a.s.}} 0$ may be expressed as

$$P\left(\lim_{n\to\infty} \sum_{i=1}^{k} (X_{ni} - X_i)^2 < \varepsilon^2\right) = 1 \tag{19.21}$$

[1] To avoid confusion between the Euclidean norm and the L_2 norm of a random variable the latter is always written with subscript 2, the former without.

for any $\varepsilon > 0$. But (19.21) implies that

$$P\Big(\lim_{n\to\infty} |X_{ni} - X_i| < \varepsilon,\ i = 1,\dots,k\Big) = 1, \tag{19.22}$$

proving 'if'. To prove 'only if' observe that if (19.22) holds then

$$P\Big(\lim_{n\to\infty} \|\boldsymbol{X}_n - \boldsymbol{X}\| < k^{1/2}\varepsilon\Big) = 1$$

for any $\varepsilon > 0$. To get the proof for convergence in pr., replace $P(\lim_{n\to\infty}\cdots)$ everywhere by $\lim_{n\to\infty} P(\cdots)$ and the arguments are identical. ∎

There are two approaches to the problem of preserving convergence (a.s. or in pr.) under transformations. In the first of these results the limit of the sequence can be a random variable provided the function is continuous almost everywhere.

19.8 Theorem Let $g : \mathbb{R}^k \to \mathbb{R}$ be a Borel function, let $C_g \subseteq \mathbb{R}^k$ be the set of continuity points of g, and assume $P(\boldsymbol{X} \in C_g) = 1$.[2]

(i) If $\boldsymbol{X}_n \to_{\text{a.s.}} \boldsymbol{X}$ then $g(\boldsymbol{X}_n) \to_{\text{a.s.}} g(\boldsymbol{X})$.

(ii) If $\boldsymbol{X}_n \to_{\text{pr}} \boldsymbol{X}$ then $g(\boldsymbol{X}_n) \to_{\text{pr}} g(\boldsymbol{X})$.

Proof For case (i), there is by hypothesis a set $D \in \mathcal{F}$, with $P(D) = 1$, such that $\boldsymbol{X}_n(\omega) \to \boldsymbol{X}(\omega)$, each $\omega \in D$. Continuity and **19.7** together imply that $g(\boldsymbol{X}_n(\omega)) \to g(\boldsymbol{X}(\omega))$ for each $\omega \in \boldsymbol{X}^{-1}(C_g) \cap D$. This set has probability 1 by **3.12**(iii).

To prove (ii), analogous reasoning shows that for each $\varepsilon > 0\ \exists\ \delta > 0$ such that

$$\{\omega : \|\boldsymbol{X}_n(\omega) - \boldsymbol{X}(\omega)\| < \delta\} \cap \boldsymbol{X}^{-1}(C_g) \subseteq \{\omega : |g(\boldsymbol{X}_n(\omega)) - g(\boldsymbol{X}(\omega))| < \varepsilon\}. \tag{19.23}$$

If $P(B) = 1$ for a set $B \in \mathcal{F}$ then for any $A \in \mathcal{F}$,

$$P(A \cap B) = 1 - P(A^c \cup B^c) \geq P(A) - P(B^c) = P(A) \tag{19.24}$$

by de Morgan's law and subadditivity of P. In particular, when $P(\boldsymbol{X} \in C_g) = 1$, (19.23) and monotonicity imply

$$P(\|\boldsymbol{X}_n - \boldsymbol{X}\| < \delta) \leq P(|g(\boldsymbol{X}_n) - g(\boldsymbol{X})| < \varepsilon). \tag{19.25}$$

[2] Implicitly, it is assumed that $C_g \in \mathcal{B}^k$ or equivalently that the probability space is complete (see page 56).

Taking the limit as $n \to \infty$ of each side of the inequality, the minorant side tends to 1 by hypothesis. ∎

The second result specifies convergence to a constant limit, but relaxes the continuity requirements.

19.9 Theorem Let $g : \mathbb{R}^k \to \mathbb{R}$ be a Borel function, continuous at $\boldsymbol{a}$.

(i) If $\boldsymbol{X}_n \to_{\text{a.s.}} \boldsymbol{a}$ then $g(\boldsymbol{X}_n) \to_{\text{a.s.}} g(\boldsymbol{a})$.

(ii) If $\boldsymbol{X}_n \to_{\text{pr}} \boldsymbol{a}$ then $g(\boldsymbol{X}_n) \to_{\text{pr}} g(\boldsymbol{a})$.

Proof By the hypothesis of (i) there is a set $D \in \mathcal{F}$ with $P(D) = 1$ such that $\boldsymbol{X}_n(\omega) \to \boldsymbol{a}$ for each $\omega \in D$. Continuity implies $g(\boldsymbol{X}_n(\omega)) \to g(\boldsymbol{a})$ for $\omega \in D$, proving (i). Likewise, continuity implies that for $\varepsilon > 0\ \exists\ \delta > 0$ such that

$$\{\omega : \|\boldsymbol{X}_n(\omega) - \boldsymbol{a}\| < \delta\} \subseteq \{\omega : |g(\boldsymbol{X}_n(\omega)) - g(\boldsymbol{a})| < \varepsilon\} \tag{19.26}$$

and (ii) follows much as in **19.8**. ∎

Theorem **19.9**(ii) is commonly known as Slutsky's theorem (Slutsky [173]). These results have a vast range of applications and represent one of the chief reasons why limit theory is useful. Having established the convergence of one set of statistics, such as the first few empirical moments of a distribution, one can then deduce the convergence of any continuous function of these. Many commonly used estimators fall into this category.

19.10 Example Let $\boldsymbol{A}_n$ be a random matrix, nonsingular with probability 1, whose elements converge a.s. (in pr.) to a nonsingular limit $\boldsymbol{A}$. Since the matrix inversion mapping is continuous at all points where $\det\boldsymbol{A} \neq 0$, the results a.s.lim $\boldsymbol{A}_n^{-1} = \boldsymbol{A}^{-1}$ (plim $\boldsymbol{A}_n^{-1} = \boldsymbol{A}^{-1}$) follow on applying **19.8** element by element. □

There are cases where only the difference of two sequences is convergent. Provided the sequences themselves don't diverge the convergence is preserved under a continuous transformation.

19.11 Theorem Let $\{\boldsymbol{X}_n\}$ and $\{\boldsymbol{Z}_n\}$ be sequences of random k-vectors and let $C_g \subseteq \mathbb{R}^k$ be the set of continuity points of a function $g : \mathbb{R}^k \mapsto \mathbb{R}$ where $P(\boldsymbol{X}_n \in C_g) = 1$ for each n.

(i) If $\{\boldsymbol{X}_n\}$ and $\{\boldsymbol{Z}_n\}$ are uniformly bounded a.s. and $\|\boldsymbol{Z}_n - \boldsymbol{X}_n\| \to_{\text{a.s.}} 0$, then $|g(\boldsymbol{Z}_n) - g(\boldsymbol{X}_n)| \to_{\text{a.s.}} 0$.

(ii) If $\{\boldsymbol{X}_n\}$ and $\{\boldsymbol{Z}_n\}$ are uniformly bounded in probability and $\|\boldsymbol{Z}_n - \boldsymbol{X}_n\| \to_{\text{pr}} 0$, then $|g(\boldsymbol{Z}_n) - g(\boldsymbol{X}_n)| \to_{\text{pr}} 0$.

Proof Write $\boldsymbol{Y}_n = \boldsymbol{Z}_n - \boldsymbol{X}_n$ so that $\boldsymbol{Z}_n = \boldsymbol{X}_n + \boldsymbol{Y}_n$ where $\|\boldsymbol{Y}_n\|$ converges to 0, whereas $\{\boldsymbol{X}_n\}$ may be a bounded sequence having two or more cluster points. Define $E_n = \boldsymbol{X}_n^{-1}(C_g)$ so that $P(E_n) = 1$ and $P(E) = 1$ where $E = \bigcap_{n=1}^{\infty} E_n$, by Theorem **3.12**(iii).

Uniform boundedness a.s. implies that $\sup_n \|\boldsymbol{X}_n(\omega)\| \leq B$ with $B < \infty$ for $\omega \in D$ with $P(D) = 1$. Hence for $\omega \in E \cap D$, $g(\boldsymbol{X}_n(\omega))$ is a continuous function on a compact set and is bounded and uniformly continuous by **2.9**. If $\|\boldsymbol{Y}_n(\omega)\| \to 0$ for $\omega \in C$ and $P(C) = 1$ then by Lemma **19.7**, $|g(\boldsymbol{X}_n(\omega) + \boldsymbol{Y}_n(\omega)) - g(\boldsymbol{X}_n(\omega))| \to 0$ as $n \to \infty$ for $\omega \in E \cap D \cap C$ where $P(E \cap D \cap C) = 1$. This shows (i).

Next, to show (ii) define

$$A_n(\eta) = \{\omega : \|\boldsymbol{X}_n(\omega)\| \leq B_\eta\} \cap E$$

where B_η may depend on η but is finite for all $\eta > 0$. $P(A_n(\eta)) = P(\|\boldsymbol{X}_n\| \leq B_\eta)$ by (19.24). Uniform boundedness in probability means that $\sup_n P(A_n(\eta)^c) < \eta$ for any $\eta > 0$. By **2.9**, g is bounded and uniformly continuous on $A_n(\eta)$. Next define for $\delta > 0$,

$$C_n(\delta) = \{\omega : \|\boldsymbol{Y}_n(\omega)\| < \delta\}$$

and also for $\varepsilon > 0$,

$$D_n(\varepsilon) = \{\omega : |g(\boldsymbol{X}_n(\omega) + \boldsymbol{Y}_n(\omega)) - g(\boldsymbol{X}_n(\omega))| < \varepsilon\}.$$

Uniform continuity and Lemma **19.7** imply that for each $\varepsilon > 0$ $\exists$ $\delta > 0$ such that

$$C_n(\delta) \cap A_n(\eta) \subseteq D_n(\varepsilon). \tag{19.27}$$

Hence,

$$P(D_n(\varepsilon)) \geq P\big(C_n(\delta) \cap A_n(\eta)\big) \geq 1 - P(C_n(\delta)^c) - P(A_n(\eta)^c) \tag{19.28}$$

where the second inequality uses the de Morgan law and subadditivity. Since $P(C_n(\delta)^c) \to 0$ as $n \to \infty$ by assumption, so $P(D_n(\varepsilon)) \geq 1 - \eta$ when n is large enough. This completes the proof since ε and η are arbitrary. ∎

Notice the importance of the uniform boundedness in this result. The result does not necessarily follow in cases where $\boldsymbol{X}_t$ diverges. As a counterexample consider $X_n = n$ and $Z_n = n + 1/n$. Then, $|Z_n - X_n| \to 0$ but $|Z_n^2 - X_n^2| \to 2$.

In the following extension, g is replaced by a sequence of non-stochastic functions, g_n. The case of a function which is itself random in addition to depending on random arguments is different. This case is treated in Chapter 22—see in particular Theorem **22.6**.

19.12 Corollary Let $\{g_n : \mathbb{R}^k \to \mathbb{R};\ n \in \mathbb{N}\}$ denote a collection of bounded non-stochastic Borel functions, where g_n is continuous at $\boldsymbol{a}$ for each n and $g_n(\boldsymbol{a}) \to g(\boldsymbol{a})$.

(i) If $\boldsymbol{X}_n \to_{\text{a.s.}} \boldsymbol{a}$ then $g_n(\boldsymbol{X}_n) \to_{\text{a.s.}} g(\boldsymbol{a})$.
(ii) If $\boldsymbol{X}_n \to_{\text{pr}} \boldsymbol{a}$ then $g_n(\boldsymbol{X}_n) \to_{\text{pr}} g(\boldsymbol{a})$.

Proof Define the array $\{\{A_{mn}\}_{n=1}^\infty\}_{m=1}^\infty$ by $A_{mn} = |g_m(\boldsymbol{X}_n) - g(\boldsymbol{a})|$ and apply the triangle inequality to write

$$A_{mn} \le A_{mn}^1 + A_m^2 \tag{19.29}$$

where $A_{mn}^1 = |g_m(\boldsymbol{X}_n) - g_m(\boldsymbol{a})|$ and $A_m^2 = |g_m(\boldsymbol{a}) - g(\boldsymbol{a})|$. In case (i), for each fixed $m \in \mathbb{N}$, $A_{mn}^1(\omega)$ converges to zero as $n \to \infty$ for $\omega \in C$ with $P(C) = 1$, by **19.9**(i). By (19.29) the sequence $\{A_{mn}(\omega)\}_{n=1}^\infty$ is bounded in the limit by A_m^2 on the same set, where $A_m^2 \to 0$ as $m \to \infty$ by assumption. Therefore, the sequence $\{A_{nn}(\omega)\}_{n=1}^\infty$ formed by taking for each n the array element $A_{mn}(\omega)$ with $m = n$ converges to zero for $\omega \in C$, by an application of the diagonal theorem **2.36**.

In case (ii) the argument is similar, with **19.9**(ii) giving the result $P(A_{mn}^1 > \varepsilon) \to 0$ for $\varepsilon > 0$, $m \in \mathbb{N}$ and hence $P(A_{nn} > \varepsilon) \to 0$ as $n \to \infty$, similarly. ∎

Another useful supplementary result is for a case not covered by the Slutsky theorem because Y_n is not required to converge in any sense.

19.13 Theorem Let a sequence $\{Y_n\}_1^\infty$ be bounded in probability (i.e. $Y_n = O_p(1)$ as $n \to \infty$); if $X_n \to_{\text{pr}} 0$, then $X_n Y_n \to_{\text{pr}} 0$.

Proof For a constant $B > 0$, define $Y_n^B = Y_n 1_{\{|Y_n| \le B\}}$. The event $\{|X_n Y_n| \ge \varepsilon\}$ for $\varepsilon > 0$ is expressible as a disjoint union:

$$\{|X_n Y_n| \ge \varepsilon\} = \{|X_n||Y_n^B| \ge \varepsilon\} \cup \{|X_n||Y_n - Y_n^B| \ge \varepsilon\}. \tag{19.30}$$

For any $\varepsilon > 0$, $\{|X_n||Y_n^B| \ge \varepsilon\} \subseteq \{|X_n| \ge \varepsilon/B\}$ and

$$P(|X_n||Y_n^B| \ge \varepsilon) \le P(|X_n| \ge \varepsilon/B) \to 0. \tag{19.31}$$

By the $O_p(1)$ assumption there exists, for each $\delta > 0$, $B_\delta < \infty$ such that $P(|Y_n - Y_n^{B_\delta}| > 0) < \delta$ for $n \in \mathbb{N}$. Since $\{|X_n \| Y_n - Y_n^B| \geq \varepsilon\} \subseteq \{|Y_n - Y_n^B| > 0\}$, (19.30) and additivity imply, putting $B = B_\delta$ in (19.31), that

$$\lim_{n\to\infty} P(|X_n Y_n| \geq \varepsilon) < \delta. \tag{19.32}$$

The theorem follows since both ε and δ are arbitrary. ∎

19.4 Convergence in L_p Norm

Recall that when $\|X_n\|_p < \infty$, X_n is said to be L_p-bounded. Consider for $p > 0$ the sequence $\{\|X_n - X\|_p\}_1^\infty$. If $\|X_n\|_p < \infty$ for all n, $\|X\|_p < \infty$ and $\lim_{n\to\infty} \|X_n - X\|_p = 0$, X_n is said to *converge in L_p norm* to X (write $X_n \to_{L_p} X$). The case $p = 2$ is called *convergence in mean square* (m.s.).

Convergence in probability is sometimes called L_0-convergence, terminology that can be explained by the fact that L_p-convergence implies L_q-convergence for $0 < q < p$ by Liapunov's inequality, together with the following relationship, which is immediate from the Markov inequality.

19.14 Theorem If $X_n \to_{L_p} X$ for any $p > 0$, then $X_n \to_{\text{pr}} X$. □

The converse does not follow in general, but see the following theorem.

19.15 Theorem If $X_n \to_{\text{pr}} X$ and $\{|X_n|^p\}_1^\infty$ is uniformly integrable, then $X_n \to_{L_p} X$.

Proof For $\varepsilon > 0$,

$$\begin{aligned} \mathrm{E}|X_n - X|^p &= \mathrm{E}(1_{\{|X_n - X| > \varepsilon\}}|X_n - X|^p) + \mathrm{E}(1_{\{|X_n - X| \leq \varepsilon\}}|X_n - X|^p) \\ &\leq \mathrm{E}(1_{\{|X_n - X| > \varepsilon\}}|X_n - X|^p) + \varepsilon^p. \end{aligned} \tag{19.33}$$

Convergence in pr. means that $P(|X_n - X| > \varepsilon) \to 0$ as $n \to \infty$. Uniform integrability therefore implies, by **12.10**, that the expectation on the majorant side of (19.33) converges to zero. The theorem follows since ε is arbitrary. ∎

The a.s. counterpart of this result was proved in effect as **12.8**, whose conclusion can be written as: $|X_n - X| \to_{\text{a.s.}} 0$ implies $\mathrm{E}|X_n - X| \to 0$. The extension from the L_1 case to the L_p case is easily obtained by applying that result to the sequence $\{|X_n - X|^p\}$.

One of the useful features of L_p convergence is that the L_p norms of $X_n - X$ define a sequence of constants whose order of magnitude in n may be determined, providing a measure of the rate of approach to the limit. For example, X_n converges to X in mean square at the rate n^k if $\|X_t - X\|_2 = O(n^{-k})$, but not $o(n^{-k})$. This is useful because the scaled random variable $n^k(X_t - X)$ may be non-degenerate in the limit, in the sense of having positive but finite limiting variance. Determining this rate of convergence is often the first step in the analysis of limiting distributions, as is discussed in Part V below.

19.5 Examples

Convergence in pr. is a weak mode of convergence in the sense that without side conditions it does not imply, yet is implied by, a.s. convergence and L_p convergence. However, there is no implication from a.s. convergence to L_p convergence, or *vice versa.* A good way to appreciate the distinctions is to consider 'pathological' cases where one or other mode of convergence fails to hold.

19.16 Example Look again at **12.7**, in which $X_n = 0$ with probability $1 - 1/n$ and $X_n = n$ with probability $1/n$, for $n = 1,2,3,\ldots$. A convenient model for this sequence is to let ω be a drawing from the space $([0,1], \mathcal{B}_{[0,1]}, m)$ where m is Lebesgue measure and define the random variable

$$X_n(\omega) = \begin{cases} n, & \omega \in [0, 1/n) \\ 0, & \text{otherwise.} \end{cases} \tag{19.34}$$

The set $\{\omega : \lim_n X_n(\omega) \neq 0\}$ consists of the point $\{0\}$ and has p.m. zero, so that $X_n \to_{\text{a.s.}} 0$ according to (19.1). But $\mathrm{E}|X_n|^p = 0.(1 - 1/n) + n^p/n = n^{p-1}$. It will be recalled that this sequence is not uniformly integrable. It fails to converge in L_p for any $p > 1$, but if $p = 1$ then $\mathrm{E}(X_n) = 1$ for every n. The limiting expectation of X_n is therefore different from its almost sure limit. □

The same device can be used to define a.s. convergent sequences that do not converge in L_p for any $p > 0$. It is left to the reader to construct examples.

19.17 Example Let a sequence be generated as follows: $X_1 = 1$ with probability 1; (X_2, X_3) are either $(0,1)$ or $(1,0)$ with equal probability; (X_4, X_5, X_6) are chosen from $(1,0,0)$, $(0,1,0)$, $(0,0,1)$ with equal probability; and so forth. For $k = 1,2,3,\ldots$ the next k members of the sequence are randomly selected such that one of them is unity, the others zero. Hence, for n in the range $[\frac{1}{2}k(k-1) + 1, \frac{1}{2}k(k+1)]$, $P(X_n = 1) = 1/k$, as well as $\mathrm{E}|X_n|^p = 1/k$ for $p > 0$. Since $k \to \infty$ as

$n \to \infty$, it is clear that X_n converges to zero both in pr. and in L_p-norm. But since, for any n, $X_{n+j} = 1$ a.s. for infinitely many j,

$$P(|X_n| < \varepsilon, \text{ i.o.}) = 0 \tag{19.35}$$

for $0 \le \varepsilon \le 1$.The sequence not only fails to converge a.s. but converges with probability 0.

Consider for the same X_n the sequence $\{k^{1/r}X_n\}$, whose members are either 0 or $k^{1/r}$ in the range $[\frac{1}{2}k(k-1)+1, \frac{1}{2}k(k+1)]$. Then $\text{E}(|k^{1/r}X_n|^p) = k^{p/r-1}$ and for $p > r$ the sequence does not converge in L_p. With $r = p = 1$, $\text{E}(kX_n) = 1$ for all n but the sequence is not uniformly integrable, as in **19.16**. The limiting expectation of the sequence exists but is different from the probability limit, which is 0 as before. □

In these non-uniformly integrable cases in which the sequence converges in L_1 but not in $L_{1+\theta}$ for any $\theta > 0$, the expectation remains formally well defined in the limit but loses its intuitive interpretation as the limit of a sample average. Example **19.16** is related to the well-known St Petersburg paradox. Consider a game of chance in which the player announces a number $n \in \mathbb{N}$ and bets that a succession of coin tosses will produce n heads before tails comes up, the pay-off for a correct prediction being £2^{n+1}. The probability of winning is 2^{-n-1}, so the expected winnings are £1; that is to say, it is a 'fair game' if the stake is fixed at £1. The sequence of random winnings X_n generated by choosing $n = 1, 2, 3, \ldots$ is exactly the process specified in **19.16**.[3] If n is chosen to be a very large number, a moment's reflection shows that the probability limit is a much better guide to one's prospective winnings in a finite number of plays than the expectation. The paradox that with large n no one would be willing to bet on this apparently fair game has been explained by appeal to psychological notions such as risk aversion, but it would appear to be an adequate explanation that for large enough n the expectation is simply not a useful predictor of the outcome of the next play.

19.6 Laws of Large Numbers

Let $\{X_t\}_1^\infty$ be a stochastic sequence and define $\bar{X}_n = n^{-1}\sum_{t=1}^n X_t$. Suppose that $\text{E}(X_t) = \mu_t$ and $n^{-1}\sum_{t=1}^n \mu_t \to \mu$ with $|\mu| < \infty$; this is trivial in the mean-stationary case in which $\mu_t = \mu$ for all t. In this simple setting, the sequence is said to obey the weak law of large numbers (WLLN) when $\bar{X}_n \to_{\text{pr}} \mu$ and the strong law of large numbers (SLLN) when $\bar{X}_n \to_{\text{a.s.}} \mu$.

[3] The original St Petersburg paradox due to Daniel Bernoulli (1738) considered a game in which the player wins £2^{n-1} if the first head appears on the n^{th} toss for *any n*. Here the expected winnings are infinite, but the principle involved is the same in either case.

These statements of the LLNs are standard and familiar, but as characterizations of a class of convergence results they are rather restrictive. The case $\mu = 0$ corresponds to considering the centred sequence $\{X_t - \mu_t\}_1^\infty$. With centring it is no longer necessary for the time average of the means to converge in the manner specified and $n^{-1}\sum_{t=1}^n \mu_t \to \infty$ is compatible with $n^{-1}\sum_{t=1}^n (X_t(\omega) - \mu_t) \to 0$. In such cases the law of large numbers requires a modified interpretation since it does not make sense to speak of convergence of the sequence of sample means.

Such convergence is possible even when the series coordinates do not possess first moments. Consider the following case.

19.18 Theorem Let the sequence $X_1, \ldots, X_n$ be i.i.d. with the property

$$nP(|X_1| > n) \to 0 \tag{19.36}$$

as $n \to \infty$. There exists a nonstochastic sequence $\{C_n\}$ such that $\bar{X}_n - C_n \to_{\text{pr}} 0$.

Proof Write $S_n = \sum_{t=1}^n X_t$ and also let $S_n^* = \sum_{t=1}^n Y_{nt}$ where $Y_{nt} = X_t 1_{\{|X_t| \le n\}}$ for $t = 1, \ldots, n$. For any $\varepsilon > 0$,

$$P(|\bar{X}_n - n^{-1}S_n^*| > \varepsilon) \le P(S_n \ne S_n^*) \le P\big(\textstyle\bigcup_t \{X_t \ne Y_{nt}\}\big) \le nP(|X_1| > n)$$

where the last inequality is by subadditivity. Hence, $\bar{X}_n - n^{-1}S_n^* \to_{\text{pr}} 0$ by (19.36) and the proof is completed by showing that $n^{-1}S_n^* - C_n \to_{L_2} 0$ where $C_n = n^{-1}\mathrm{E}(S_n^*) = \mathrm{E}(X_1 1_{\{|X_1| \le n\}})$. By independence of the coordinates,

$$\mathrm{E}\big(n^{-1}S_n^* - C_n\big)^2 \le \frac{1}{n^2}\sum_{t=1}^n \mathrm{E}(Y_{nt}^2) = \frac{1}{n}\mathrm{E}\big(X_1^2 1_{\{|X_1| \le n\}}\big).$$

To show that the majorant vanishes, define $h(n) = 2nP(|X_1| > n)$ and note that $h(n) \le 2n$ and $h(n) \to 0$ as $n \to \infty$ by (19.36). For any positive constant M note that by **9.21** with $r = 2$,

$$\begin{aligned}\frac{1}{n}\mathrm{E}\big(X_1^2 1_{\{|X_1| \le n\}}\big) &\le \frac{1}{n}\int_0^n h(x)\mathrm{d}x \\ &\le \frac{M\sup_{n \le M} h(n)}{n} + \frac{n-M}{n}\sup_{n>M} h(n) \\ &\to \sup_{n>M} h(n) \text{ as } n \to \infty.\end{aligned}$$

The limit on the right-hand side can be made as small as desired by taking M large enough. ∎

Condition (19.36) does not imply integrability. If $P(|X_1| > n) \simeq (n\log n)^{-1}$, for example, then (19.36) is satisfied but $\mathrm{E}|X_1| = \infty$ according to **9.22** with $r = 1$ and (2.20). In such a case, C_n may not converge as $n \to \infty$ and then it is not clear what the result tells us, although note that if X_1 is symmetrically distributed about zero then $C_n = 0$ for all finite n. The Cauchy distribution (**8.15**) is a well-known case of failure to obey the law of large numbers and in this case (19.36) fails, noting that

$$nP(|X| > n) = \frac{2n}{\pi}\int_n^\infty \frac{1}{(1+x^2)}\mathrm{d}x \sim \frac{2n}{\pi}\int_n^\infty x^{-2}\mathrm{d}x = \frac{2}{\pi}.$$

More general modes of convergence also exist. It is possible that $\bar{X}_n$ does not converge in the manner specified, even after centring, but that there exists a sequence of positive constants $\{a_n\}_1^\infty$ such that $a_n \uparrow \infty$ and $a_n^{-1}\sum_{t=1}^n X_t \to 0$, where a_n may diverge either faster or slower than n. These possibilities and others too may be subsumed in a fully general array formulation of the problem. If $\{\{X_{nt}\}_{t=1}^{k_n}\}_{n=1}^\infty$ is a triangular stochastic array with $\{k_n\}_{n=1}^\infty$ an increasing integer sequence, results to be presented in §20.3–§20.6 will prove conditions for

$$S_n = \sum_{t=1}^{k_n} X_{nt} \overset{\mathrm{pr}}{\to} 0. \tag{19.37}$$

A result in this form can be specialized to the familiar case with $X_{nt} = a_n^{-1}(X_t - \mu_t)$ and $a_n = k_n = n$, but there are important applications where the greater generality is essential.

According to **14.6**, $\bar{X}_n \to_{\mathrm{a.s.}} \mu = \mathrm{E}(X_1)$ when $\{X_t\}$ is a stationary ergodic sequence and $\mathrm{E}|X_1| < \infty$. Here is an example where the sequence is i.i.d., which is sufficient for ergodicity.

19.19 Example Consider a sequence of independent Bernoulli variables X_t with $P(X_t = 1) = P(X_t = 0) = \frac{1}{2}$ for $t = 1, 2, 3, \ldots$. These may be coin tosses expressed in binary form, as in **12.1**. The conditions of the ergodic theorem are clearly satisfied and $n^{-1}\sum_{t=1}^n X_t \to_{\mathrm{a.s.}} \mathrm{E}(X_1) = \frac{1}{2}$. This is *Borel's normal number theorem.* In the construction of **12.1**, a normal number is defined as one in which 0s and 1s occur in its binary expansion with equal frequency in the limit. The normal number theorem therefore states that almost every point of the unit interval is a normal number; that is, the set of normal numbers has Lebesgue measure 1.

Any number with a terminating expansion is clearly non-normal and all such numbers are rationals. However, rationals can be normal, such as for example $\frac{1}{3}$, which has the binary expansion 0.01010101010101.... This is a different result from the well-known zero measure of the rationals and is much stronger, because the non-normal numbers include irrationals and form an uncountable set. An example would be any number with a binary expansion of the form $0.11b_1 11b_2 11b_3 11\ldots$where the b_i are arbitrary digits. These numbers are non-normal, yet they can be put into 1–1 correspondence with the expansions $0.b_1 b_2 b_3 \ldots$, in other words, with the points of the whole interval. The set of non-normal numbers is equipotent with the reals, but it nonetheless has Lebesgue measure 0. □

The stationary ergodic property is preserved under measurable transformations. If $\{X_t\}$ is stationary and ergodic, so is the sequence $\{g(X_t)\}$ whenever $g : \mathbb{R} \mapsto \mathbb{R}$ is a measurable function, a special case of **14.9**. For example, it is enough to know that $\mathrm{E}(X_1^2) < \infty$ to be able to assert that $n^{-1}\sum_{t=1}^{n} X_t^2 \to_{\text{a.s.}} \mathrm{E}(X_1^2)$. The ergodic theorem serves to establish the strong law for most stationary sequences likely to be encountered, recalling from §14.4 that ergodicity is a weaker property than regularity or mixing. The interesting problems in stochastic convergence arise when the distributions of sequence coordinates are heterogeneous, so that it is not trivial to assume that averaging of coordinates is a stable procedure in the limit.

Another result known to yield a strong law is the martingale convergence theorem (**16.11**), which has the interpretation that $a_n^{-1}\sum_{t=1}^{n} X_t \to_{\text{a.s.}} 0$ whenever $\{\sum_{t=1}^{n} X_t\}$ is a submartingale with $\mathrm{E}|\sum_{t=1}^{n} X_t| < \infty$ uniformly in n and $a_n \to \infty$. This particular strong law needs to be combined with additional results to give it a broad application, but this is readily done as shown in §21.3.

However, lest the law of large numbers appear an altogether trivial problem, it is a good idea to exhibit some cases where convergence fails to occur.

19.20 Example Let $\{X_t\}$ denote a sequence of independent Cauchy random variables with characteristic function $\phi_{X_t}(\lambda) = \mathrm{e}^{-|\lambda|}$ for each t (**11.13**). It is easy to verify using formulae (11.24) and (11.27) that $\phi_{\bar{X}_n}(\lambda) = \mathrm{e}^{-n|\lambda|/n} = \mathrm{e}^{-|\lambda|}$. According to the inversion theorem, the average of n independent Cauchy variables is also a Cauchy variable. This result holds for any n, contradicting the possibility that $\bar{X}_n$ could converge to a constant. The failure of condition (19.36) in this case has been noted above. □

19.21 Example Consider a process

$$X_t = \sum_{s=1}^{t} \Psi_s Z_s = X_{t-1} + \Psi_t Z_t,\ t = 1, 2, 3, \ldots \tag{19.38}$$

with $X_0 = 0$, where $\{Z_t\}_1^\infty$ is an i.i.d. sequence with mean 0 and variance σ^2 and $\{\Psi_s\}_1^\infty$ is a sequence of constant coefficients. Notice that these are indexed with the absolute date rather than the lag relative to time t, as in the moving average processes considered in §13.3. For $m > 0$,

$$\mathrm{Cov}(X_t, X_{t+m}) = \mathrm{Var}(X_t) = \sigma^2 \sum_{s=1}^{t} \Psi_s^2. \tag{19.39}$$

For $\{X_t\}_1^\infty$ to be uniformly L_2-bounded requires $\sum_{s=1}^{\infty} \Psi_s^2 < \infty$. In this case the effect of the innovations declines to zero with t and X_t approaches a limiting random variable. Without the square-summability assumption, $\mathrm{Var}(X_t) \to \infty$; an example is the random walk process in which $\Psi_s = 1$, all s (see §13.4). In either case the sequence of means $\{\bar{X}_n\}$ fails to converge to a fixed limit, being either stochastic asymptotically, or divergent. □

These counterexamples illustrate the fact that to obey the law of large numbers, a sequence must satisfy regularity conditions relating to two distinct factors: the probability of outliers (limited by bounding absolute moments) and the degree of dependence between coordinates. **19.20** is a case where the mean fails to exist and **19.21** an example of long-range dependence. In neither case can $\bar{X}_n$ be thought of as a sample statistic which is estimating a parameter of the underlying distribution in any meaningful fashion. Chapters 20 and 21 derive sets of regularity conditions sufficient for weak and strong laws to operate, constraining both characteristics in different configurations. The *necessity* of a set of regularity conditions is usually hard to prove (the exception being when the sequences are serially independent) but various configurations of mixing and L_p-boundedness conditions are shown to be sufficient. These results usually exhibit a trade-off between the two dimensions of regularity; the stronger the moment restrictions are, the weaker dependence restrictions can be, and vice versa.

One word of caution before proceeding to the theorems. In §9.1 there is an attempt to motivate the idea of an expected value by viewing it as the limit of the empirical average. There is then a temptation to attempt to *define* an expectation as such a limit; but to do this would inevitably involve circular reasoning since the arguments establishing convergence are couched in the language of probability. The aim of the theory is to establish convergence in particular sampling schemes. It cannot be used to validate the frequentist interpretation of probability. However, it *does* show that axiomatic probability yields predictions that accord with the frequentist model and in this sense the laws of large numbers are among the most fundamental results in probability theory.

20
Convergence in L_p Norm

20.1 Weak Laws by Mean Square Convergence

This chapter surveys a range of techniques for proving (mainly) weak laws of large numbers. The common theme in these results is that they depend on showing convergence in L_p-norm, where in general p lies in the interval [1,2]. Consider initially the case $p = 2$. The regularity conditions for these results relate directly to the variances and covariances of the process. While for subsequent results these moments will not need to exist, the L_2 case is of interest both because the conditions are familiar and intuitive and because in certain respects the results available are more powerful.

Consider a stochastic sequence $\{X_t\}_1^\infty$, with sequence of means $\{\mu_t\}_1^\infty$ and variances $\{\sigma_t^2\}_1^\infty$. There is no loss of generality in setting $\mu_t = 0$ by considering the centred case of $\{X_t - \mu_t\}_1^\infty$, but to focus the discussion on a familiar case, the assumption maintained in this section and the next is that $\bar{\mu}_n = n^{-1}\sum_{t=1}^n \mu_t \to \mu < \infty$. In this case the question to be posed is, what are sufficient conditions for $\mathrm{E}(\bar{X}_n - \mu)^2 \to 0$? This condition yields a weak law of large numbers by the application of **19.14**, often referred to as Chebyshev's theorem in this context. An elementary relation is

$$\mathrm{E}(\bar{X}_n - \mu)^2 = \mathrm{Var}(\bar{X}_n) + \big(\mathrm{E}(\bar{X}_n) - \mu\big)^2 \tag{20.1}$$

where the second term on the right-hand side converges to zero by definition of μ. Thus the question becomes: when does $\mathrm{Var}(\bar{X}_n) \to 0$?

There are various scenarios under which this property might be examined, but a good place to start is with the wide-sense stationary case in which $\mathrm{Cov}(X_t, X_{t-m}) = \gamma_m$, depending on m but not on t. An implication of stationarity is that $\gamma_m = \gamma_{-m}$ and

$$\mathrm{Var}(\bar{X}_n) = \frac{1}{n^2}\sum_{t=1}^n\sum_{s=1}^n \gamma_{t-s} = \frac{1}{n}\Big(\gamma_0 + 2\sum_{m=1}^{n-1}\Big(1 - \frac{m}{n}\Big)\gamma_m\Big). \tag{20.2}$$

In this context, the condition for mean square convergence that is both sufficient and necessary is that the autocovariance sequence $\{\gamma_m\}_1^\infty$ be 'zero on average'; that is to say, it has a Cesàro sum of zero.

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0020

20.1 Theorem In a wide-sense stationary process, $\lim_{n\to\infty}\mathrm{Var}(\bar{X}_n) = 0$ iff $\lim_{n\to\infty} n^{-1}\sum_{m=1}^{n}\gamma_m = 0$.

Proof Write $S_n = n^{-1}\sum_{m=1}^{n}\gamma_m$. Since $|\gamma_m| \le \gamma_0$ for $m > 0$ by the Cauchy–Schwarz inequality, note that

$$|S_n| \le \frac{1}{n}\sum_{m=1}^{n}|\gamma_m| \le \gamma_0 < \infty$$

and also that

$$|S_n - S_{n-1}| = \frac{|\gamma_n - S_{n-1}|}{n} = O(n^{-1}).$$

It follows that the sequence $\{\gamma_m\}_1^\infty$ is Cesàro-summable with Cesàro sum S_∞. Similarly, define

$$S_n^* = \frac{1}{n}\sum_{m=1}^{n-1}\Big(1 - \frac{m}{n}\Big)\gamma_m = \frac{1}{2}\Big(\mathrm{Var}(\bar{X}_n) - \frac{\gamma_0}{n}\Big) \tag{20.3}$$

corresponding to the term appearing on the right-hand side of (20.2).

To show 'if', suppose $S_\infty = 0$. There are two ways in which this condition can arise. The first is that at most a finite number of γ_m are nonzero, in which case it is immediate that $S_\infty^* = 0$ also. The second possibility is that nonzero autocovariances of both signs arise at all orders, but $S_\infty^+ = S_\infty^-$ where $S_n^+ = n^{-1}\sum_{m=1}^{n} 1_{\{\gamma_m \ge 0\}}\gamma_m$ and $S_n^- = S_n^+ - S_n$. In this case a proportion of the autocovariances are negative even though (20.3) shows that $S_n^* \ge 0$ when n is large. Note that the condition $S_n^* > S_n$ necessarily implies a pattern in the autocovariance sequence. The 'remote' γ_m receiving the smallest weights in S_n^* must make the greater negative contributions to the sum, while those with smaller m, whose weights are larger, must make the greater positive contributions. However, given $S_\infty = 0$ such imbalances cannot exist for every n. Therefore, $S_n^* \to 0$ and $\mathrm{Var}(\bar{X}_n) \to 0$ according to (20.3).

To show 'only if', invert this argument. Having $\mathrm{Var}(\bar{X}_n) > 0$ imposes a sign pattern on the autocovariances. The condition $S_n^* > 0$ when $S_n = 0$ would imply γ_m is more often positive when m is close to 1 than when m is close to n. Such a difference cannot exist for every n, hence $S_\infty^* > 0$ implies $S_\infty > 0$. ∎

It is not impossible for a stationary sequence to have nonzero autocovariances of all orders. Here are two cases.

20.2 Example Consider $X_t = z + u_t$ where u_t is an i.i.d. sequence with mean zero and variance σ_x^2 and z is an independently drawn r.v. with mean zero and variance σ_z^2, so that $\gamma_0 = \sigma_u^2 + \sigma_z^2$ and $\gamma_m = \sigma_z^2$ for all $m > 0$. This sequence is strictly stationary, but is not ergodic. The sample mean does not converge in mean square and converges in probability to z, not to $\mathrm{E}(X_t) = 0$. □

20.3 Example Consider the sequence formed as $X_t = z$ when t is even and $X_t = -z$ when t is odd, where $z \sim_{\mathrm{d}} \mathrm{N}(0,\sigma_z^2)$. This sequence is strictly stationary (since z is symmetrically distributed about zero) and the autocorrelation sequence has the form $\gamma_m = \sigma_z^2$ for $m = 0,2,4,6,\ldots$ and $\gamma_m = -\sigma_z^2$ for $m = 1,3,5,7,\ldots$. The sequence is not ergodic, but the sample mean is either 0 (n even) or $-z/n$ (n odd), with respective variances 0 (n even) and σ_z^2/n^2 (n odd), and hence it converges in mean square. This is a case when the positive and negative autocorrelations do not tend to 0 but do have equal Cesàro sums. □

An alternative approach to conditions for L_2 convergence is to relax the stationarity requirement. Defining $\sigma_t^2 = \mathrm{Var}(X_t)$ and $\sigma_{ts} = \mathrm{Cov}(X_t, X_s)$, the counterpart of (20.2) is

$$\mathrm{Var}(\bar{X}_n) = \frac{1}{n^2}\left(\sum_{t=1}^{n}\sigma_t^2 + 2\sum_{t=2}^{n}\sum_{s=1}^{t-1}\sigma_{ts}\right). \tag{20.4}$$

Assume initially that the sequence is uncorrelated, with $\sigma_{ts} = 0$ for $t \neq s$ in (20.4). Then there is the following well-known result.

20.4 Theorem If $\{X_t\}_1^\infty$ is an uncorrelated sequence and

$$\sum_{t=1}^{\infty}\sigma_t^2/t^2 < \infty \tag{20.5}$$

then $\bar{X}_n \to_{L_2} \mu$.

Proof This is an application of Kronecker's lemma (**2.35**), by which (20.5) implies $\mathrm{Var}(\bar{X}_n) = n^{-2}\sum_t \sigma_t^2 \to 0$. ∎

An (amply) sufficient condition for (20.5) is that the variances are uniformly bounded with, say, $\sup_t \sigma_t^2 \leq B < \infty$. Wide-sense stationary sequences as in **20.1** fall into this class and for these cases $\mathrm{Var}(\bar{X}_n) = O(n^{-1})$. But since all that is needed for L_2 convergence is $\mathrm{Var}(\bar{X}_n) = o(1)$, allowing $\sigma_t^2 \to \infty$ is permissible. If $\sigma_t^2 \simeq t^{1-\delta}$ for $\delta > 0$, then $\sum_{t=1}^{n} t^{-2}\sigma_t^2$ has terms of $O(t^{-1-\delta})$ and therefore converges by **2.17**.

An alternative approach to (20.4) is to let the magnitude of the covariances be suitably controlled. Imposing uniform L_2-boundedness to allow the maximum relaxation of constraints on dependence, the Cauchy–Schwarz inequality gives $|\sigma_{ts}| \le B$ for all t and s. Rearranging and bounding (20.4), defining $B_m = \sup_t |\sigma_{t,t-m}|$ so that $B_m \le B$, all $m \ge 1$, yields

$$\begin{aligned}\mathrm{Var}(\bar{X}_n) &= \frac{1}{n^2}\Bigg(\sum_{t=1}^{n}\sigma_t^2 + 2\sum_{t=2}^{n}\sigma_{t,t-1} + 2\sum_{t=3}^{n}\sigma_{t,t-2} + \ldots + 2\sigma_{n1}\Bigg)\\ &\le \frac{1}{n^2}\sum_{t=1}^{n}\sigma_t^2 + \frac{2}{n^2}\sum_{m=1}^{n-1}\sum_{t=m+1}^{n}|\sigma_{t,t-m}|\\ &\le \frac{B}{n} + \frac{2}{n^2}\sum_{m=1}^{n-1}(n-m)B_m. \qquad (20.6)\end{aligned}$$

This suggests the following variant on **20.4**.

20.5 Theorem If $\{X_t\}_1^\infty$ is a uniformly L_2-bounded sequence and $\sum_{m=1}^{\infty} m^{-1} B_m < \infty$, then $\bar{X}_n \to_{L_2} \mu$.

Proof Since $(n-m)/n < 1$ it is sufficient by (20.6) to show the convergence of $(2/n)\sum_{m=1}^{n-1} B_m$ to zero. This follows immediately from the stated condition and Kronecker's lemma. ∎

A sufficient condition in view of **2.18** is $B_m = O((\log m)^{-1-\delta})$, $\delta > 0$, a very mild restriction on the autocovariances.

There are two observations to be made about these results. The first is to point to the trade-off between the dimensions of dependence and the growth of the variances. Theorems **20.4** and **20.5** are easily combined and it is found that by tightening the rate at which the covariances diminish the variances can grow faster and vice versa. The reader can explore these possibilities using the rather simple techniques of the above proofs, remembering that the $|\sigma_{t,t-m}|$ need to be treated as growing with t as well as diminishing with m. Analogous trade-offs are derived in a different context below.

The second is to note that the order of magnitude in n of $\mathrm{Var}(\bar{X}_n)$ measures the *rate* of convergence. With no autocorrelation and bounded variances, convergence is at the rate $n^{-1/2}$ in the sense that $\mathrm{Var}(\bar{X}_n) = O(n^{-1})$. However, $B_m = O(m^{-\delta})$ in (20.6) implies that $\mathrm{Var}(\bar{X}_n) = O(n^{-\delta})$. If convergence rates are thought of as indicating the number of sample observations required to get $\bar{X}_n$ close to μ with high confidence, the weakest sufficient conditions evidently yield convergence only in a notional sense. It is less easy in some of the more general results below to link explicitly the rate of convergence with the degree of dependence and/or nonstationarity; this is always an issue to keep in mind.

Mixing sequences have the property that the covariances tend to zero and the mixing inequalities of §15.2 give the following corollary to **20.5**.

20.6 Corollary If $\{X_t\}_1^\infty$ is either (i) uniformly $L_{2+\delta}$-bounded for $\delta > 0$ and strong mixing with

$$\sum_{m=1}^{\infty} m^{-1}\alpha_m^{\delta/(2+\delta)} < \infty \tag{20.7}$$

or (ii) uniformly L_2-bounded and uniform mixing with

$$\sum_{m=1}^{\infty} m^{-1}\phi_m^{1/2} < \infty \tag{20.8}$$

then $\bar{X}_n \to_{L_2} \mu$.

Proof For part (i), **15.3** for the case $p = r = 2 + \delta$ yields $B_m \leq 6\|X_t\|_{2+\delta}^2\alpha_m^{\delta/(2+\delta)}$. For part (ii), **15.5** for the case $r = 2$ yields the inequality $B_m \leq 2B\phi_m^{1/2}$. Noting that $B \leq \|X_t\|_{2+\delta}^2$, the conditions of **20.5** are satisfied in both cases. ∎

A sufficient condition for **20.6**(i) is $\alpha_m = O((\log m)^{-(1+2/\delta)(1+\varepsilon)})$ for any $\varepsilon > 0$. For **20.6**(ii), $\phi_m = O((\log m)^{-2-\varepsilon})$ for $\varepsilon > 0$ is sufficient. In the size terminology of §15.1, mixing of any size will ensure these conditions. The most significant cost of using the strong mixing condition is that simple existence of the variances is not sufficient. This is not to say that no weak law exists for L_2-bounded strong mixing processes, but more subtle arguments such as those of §20.5 are needed for the proof.

20.2 Almost Sure Convergence by the Method of Subsequences

Almost sure convergence does not follow from convergence in mean square (a counterexample is **19.17**), but a clever adaptation of the above techniques yields a result. The proof of the following theorems makes use of the method of subsequences, exploiting the relation between convergence in pr. and convergence a.s. demonstrated in **19.6**.

Mainly for the sake of clarity, the first result proved is for the uncorrelated case. Notice how the conditions have to be strengthened, relative to **20.4**.

20.7 Theorem If $\{X_t\}_1^\infty$ is uniformly L_2-bounded and uncorrelated, $\bar{X}_n \to_{\text{a.s.}} \mu$.

□

A natural place to start in a sufficiency proof of the strong law is with the convergence part of the Borel–Cantelli lemma. Under the stated conditions the Chebyshev inequality yields

$$P(|\bar{X}_n - \bar{\mu}_n| > \varepsilon) \le \frac{\mathrm{Var}(\bar{X}_n)}{\varepsilon^2} \le \frac{B}{n\varepsilon^2} \tag{20.9}$$

for $B < \infty$, with the probability on the left-hand side going to zero with the right-hand side as $n \to \infty$. One approach to the problem of bounding the quantity $P(|\bar{X}_n - \bar{\mu}_n| > \varepsilon, \text{i.o.})$ would be to add up the inequalities in (20.9) over n. Since the partial sums of $1/n$ form a divergent sequence, a direct attack on these lines does not succeed. However, $\sum_{n=1}^{\infty} n^{-2} = \pi^2/6 < 1.65$ and the *sub*sequence of the probabilities in (20.9), for $n = 1, 4, 9, 16, \ldots$, can be summed as follows.

Proof of 20.7 By (20.9),

$$\sum_{n^2} P(|\bar{X}_{n^2} - \bar{\mu}_{n^2}| > \varepsilon) \le \frac{1.65B}{\varepsilon^2} < \infty. \tag{20.10}$$

Now **19.2**(i) yields the result that the subsequence $\{\bar{X}_{n^2}, n \in \mathbb{N}\}$ converges a.s. The proof is completed by showing that the maximum deviation of the omitted terms from the nearest member of $\{\bar{X}_{n^2}\}$ also converges in mean square. For each n define

$$D_{n^2} = \max_{n^2 \le k < (n+1)^2} |\bar{X}_k - \bar{X}_{n^2}| \tag{20.11}$$

and consider the variance of D_{n^2}. Given the assumptions, the sequence of the $\mathrm{Var}(\bar{X}_n) = (1/n^2)\sum_{t=1}^{n} \sigma_t^2$ tends monotonically to zero. For $n^2 < k < (n+1)^2$, rearrangement of the terms produces

$$\bar{X}_k - \bar{X}_{n^2} = \Big(\frac{n^2}{k} - 1\Big)\bar{X}_{n^2} + \frac{1}{k}\sum_{t=n^2+1}^{k} X_t \tag{20.12}$$

and when the sequence is uncorrelated the two terms on the right are also uncorrelated. Hence

$$\begin{aligned}
\mathrm{Var}(\bar{X}_k - \bar{X}_{n^2}) &= \Big(1 - \frac{n^2}{k}\Big)^2 \mathrm{Var}(\bar{X}_{n^2}) + \Big(\frac{1}{k}\Big)^2 \sum_{t=n^2+1}^{k} \sigma_t^2 \\
&\le \Big(1 - \frac{n^2}{k}\Big)^2 \frac{B}{n^2} + (k - n^2)\frac{B}{k^2} \\
&= B\Big(\frac{1}{n^2} - \frac{1}{k}\Big) \le B\Big(\frac{1}{n^2} - \frac{1}{(n+1)^2}\Big).
\end{aligned} \tag{20.13}$$

$\operatorname{Var}(D_{n^2})$ cannot exceed the last term in (20.13) and

$$\sum_{n^2}\Big(\frac{1}{n^2}-\frac{1}{(n+1)^2}\Big)<\sum_{n}\Big(\frac{1}{n^2}-\frac{1}{(n+1)^2}\Big)=1 \tag{20.14}$$

so the Chebyshev inequality gives

$$\sum_{n^2}P(D_{n^2}>\varepsilon)\leq\frac{B}{\varepsilon^2}<\infty \tag{20.15}$$

and by **19.2**(i) the subsequence $\{D_{n^2}, n\in\mathbb{N}\}$ also converges a.s. to zero. Since $D_{n^2}\geq|\bar{X}_k-\bar{X}_{n^2}|$ for any k between n^2 and $(n+1)^2$, the triangle inequality gives

$$\begin{aligned}|\bar{X}_k-\bar{\mu}_k|&\leq|\bar{X}_{n^2}-\bar{\mu}_{n^2}|+|\bar{X}_k-\bar{X}_{n^2}|+|\bar{\mu}_k-\bar{\mu}_{n^2}|\\&\leq|\bar{X}_{n^2}-\bar{\mu}_{n^2}|+D_{n^2}+|\bar{\mu}_k-\bar{\mu}_{n^2}|.\end{aligned} \tag{20.16}$$

The sequences on the majorant side are each positive and converge a.s. to zero, hence so does their sum. But (20.16) holds for $n^2\leq k<(n+1)^2$ for $n\in\mathbb{N}$ so that k ranges over every integer value, hence $\bar{X}_n\to_{\text{a.s.}}\mu$. ∎

The same technique can be generalized to allow autocorrelation.

20.8 Corollary If $\{X_t\}_1^\infty$ is uniformly L_2-bounded and

$$B^*=\sum_{m=1}^{\infty}B_m<\infty \tag{20.17}$$

where $B_m=\sup_t|\sigma_{t,t-m}|$, then $\bar{X}_n\to_{\text{a.s.}}\mu$. □

Note how much tougher these conditions are than those of **20.5**. Instead of having the autocovariances merely decline to zero, now their summability is required.

Proof of 20.8 By (20.6), $\operatorname{Var}(\bar{X}_n)\leq(B+2B^*)/n$ and hence equation (20.10) holds in the modified form,

$$\sum_{n^2}P(|\bar{X}_{n^2}-\bar{\mu}_{n^2}|>\varepsilon)\leq\frac{1.65(B+2B^*)}{\varepsilon^2}<\infty. \tag{20.18}$$

In contrast to (20.13), multiplying out and taking expectations gives

$$\operatorname{Var}(\bar{X}_k-\bar{X}_{n^2})=\operatorname{Var}\Big(\frac{1}{k}\sum_{t=n^2+1}^{k}X_t-\Big(1-\frac{n^2}{k}\Big)\bar{X}_{n^2}\Big)$$

$$
\begin{aligned}
&= \Big(1 - \frac{n^2}{k}\Big)^2 \mathrm{Var}(\bar{X}_{n^2}) \\
&\quad + \frac{1}{k^2}\Bigg(\sum_{t=n^2+1}^{k} \sigma_t^2 + 2 \sum_{t=n^2+2}^{k} \sum_{m=1}^{t-n^2-1} \sigma_{t,t-m} \Bigg) \\
&\qquad - \frac{2}{k}\Big(1 - \frac{n^2}{k}\Big)\Bigg(\sum_{t=n^2+1}^{k} \sum_{m=t-n^2}^{t-1} \sigma_{t,t-m} \Bigg). \qquad (20.19)
\end{aligned}
$$

The first term on the right-hand side is bounded by $(1 - n^2/k)^2(B + 2B^*)/n^2$, the second by $(k - n^2)(B + 2B^*)/k^2$, and the third (absolutely) by $2(1 - n^2/k)^2B^*$. Adding together these latter terms and simplifying yields

$$
\begin{aligned}
\mathrm{Var}\big(\bar{X}_k - \bar{X}_{n^2}\big) &\le \Big(\frac{1}{n^2} - \frac{1}{k}\Big)B + 2\Bigg(\frac{1}{n^2} - \frac{1}{k} + \Big(1 - \frac{n^2}{k}\Big)^2\Bigg)B^* \\
&\le \Big(\frac{1}{n^2} - \frac{1}{(n+1)^2}\Big)B \\
&\quad + 2\Bigg(\frac{1}{n^2} - \frac{1}{(n+1)^2} + \Big(1 - \frac{n^2}{(n+1)^2}\Big)^2\Bigg)B^*. \qquad (20.20)
\end{aligned}
$$

Note that $(1 - n^2/(n+1)^2)^2 = O(n^{-2})$, so the term in B^* is summable. In place of (20.15) write

$$\sum_{n^2} P(D_{n^2} > \varepsilon) \le \frac{B + K_1B^*}{\varepsilon^2} < \infty \qquad (20.21)$$

where K_1 is a finite constant and from here follow the proof of **20.7**. ∎

Again there is a straightforward extension to mixing sequences by direct analogy with **20.6**.

20.9 Corollary If $\{X_t\}_1^\infty$ is either (i) uniformly $L_{2+\delta}$-bounded for $\delta > 0$ and strong mixing with

$$\sum_{m=1}^{\infty} \alpha_m^{\delta/(2+\delta)} < \infty \qquad (20.22)$$

or (ii) uniformly L_2-bounded and uniform mixing with

$$\sum_{m=1}^{\infty} \phi_m^{1/2} < \infty \qquad (20.23)$$

then $\bar{X}_n \to_{\text{a.s.}} \mu$. □

Let it be emphasized that these results have no pretensions to being sharp! They are given here as an illustration of technique and also to define the limits of this approach to strong convergence. Chapter 21 shows how they can be improved upon.

20.3 Truncation Arguments

There are a number of ways of dealing with processes not possessing a variance, the basic common idea being truncation. Given a sequence $\{X_t\}_1^\infty$, define $Y_t = 1_{\{|X_t|\le B\}}X_t$ which equals X_t when $|X_t| \le B < \infty$ and 0 otherwise. The argument proceeds by showing L_2-convergence of (say) $\bar{Y}_n$ while also showing the remainders $X_t - Y_t$ to be collectively negligible. The following argument for independent r.v.s is one of the simplest and most elegant of the type and is closely affiliated with Theorem **19.18**. To generalize to non-identical distributions it is convenient to adopt an array formulation. For an increasing sequence $\{k_n, n \in \mathbb{N}\}$ let $\{X_{nt}\}_{t=1}^{k_n}$ for $n = 1, 2, 3, \ldots$ be a triangular stochastic array.

20.10 Theorem For each n let $\{X_{nt}\}_{t=1}^{k_n}$ be an independent sequence and define $Y_{nt} = X_{nt}1_{\{|X_{nt}|\le 1\}}$. If

$$\sum_{t=1}^{k_n} P(|X_{nt}| > 1) \to 0 \tag{20.24}$$

and

$$\sum_{t=1}^{k_n} \mathrm{E}(Y_{nt}^2) \to 0 \tag{20.25}$$

then $\sum_{t=1}^{k_n}(X_{nt} - \mathrm{E}(Y_{nt})) \to_{\mathrm{pr}} 0$.

Proof Put $X_n = \sum_{t=1}^{k_n} X_{nt}$, $Y_n = \sum_{t=1}^{k_n} Y_{nt}$, and $A_n = \mathrm{E}(Y_n)$. Also define $B_n = \{Y_n = X_n\}$, the event that occurs when $|X_{nt}| \le 1$ for every t. Now, for $\varepsilon > 0$,

$$\begin{aligned} P(|X_n - A_n| \ge \varepsilon) &= P\big(\{|X_n - A_n| \ge \varepsilon\} \cap B_n\big) + P\big(\{|X_n - A_n| \ge \varepsilon\} \cap B_n^c\big) \\ &\le \frac{1}{\varepsilon^2}\mathrm{E}(Y_n - A_n)^2 + P\big(\textstyle\bigcup_{t=1}^{k_n}\{|X_{nt}| > 1\}\big) \\ &\le \frac{1}{\varepsilon^2}\sum_{t=1}^{k_n}\mathrm{E}(Y_{nt}^2) + \sum_{t=1}^{k_n}P(|X_{nt}| > 1). \end{aligned} \tag{20.26}$$

The first inequality applies Chebyshev (**9.16**) and the second one the assumption of independence, also applying subadditivity in the second term. To complete the proof, apply (20.24) and (20.25). ∎

The probability limit indicated raises the question of what happens as n increases to the array $\{\mathrm{E}(Y_{nt})\}$, which appears to depend on the arbitrarily chosen truncation point. However, since $\sup_t P(|X_{nt}| > 1) \to 0$ as $n \to \infty$ according to (20.24) it follows according to Lemma **12.9** that for any $\varepsilon > 0$,

$$|\mathrm{E}(X_{nt}) - \mathrm{E}(Y_{nt})| \le \mathrm{E}|X_{nt} - Y_{nt}| = \mathrm{E}(|X_{nt}| 1_{\{|X_{nt}|>1\}}) < \varepsilon$$

when n is large enough. The conditions of the theorem therefore imply that $\mathrm{E}(X_{nt})$ exists in the limit, although this may merely reflect the fact that X_{nt} is becoming infinitesimal and hence arbitrarily close to a constant. Theorem **19.18** shows, in effect, that in the i.i.d. case condition (20.25) follows from (20.24). On setting $k_n = n$ and $X_{nt} = X_t/n$ the latter condition is seen to be identical with (19.36) and shows how the result might be interpreted in the absence of first moments.

Gnedenko and Kolmogorov ([82]), from whom **20.10** is adapted, show using a characteristic function argument that the conditions of the theorem are necessary as well as sufficient for convergence in probability in the independent case. This extension is not attempted here, although note that the 'three-series theorem' **21.8** sets necessary and sufficient conditions of a similar sort for the strong law, which by **21.12** is equivalent to the weak law under independence.

20.4 A Martingale Weak Law

The object here is to prove L_p-convergence for $p < 2$ in cases where the variance may not exist. Again, the basic tool to implement these results is a truncation argument in which the r.v. is divided into additive components, the bounded component and a remainder. Given a sequence $\{X_t\}_1^\infty$ assumed to have mean 0, define $Y_t = 1_{\{|X_t| \le B\}} X_t$. Letting $Z_t = X_t - Y_t = 1_{\{|X_t|>B\}} X_t$ denote the 'tail component' of X_t, notice that $\mathrm{E}(Z_t) = -\mathrm{E}(Y_t)$ by construction and $\bar{X}_n = \bar{Y}_n + \bar{Z}_n$. Since Y_t is a.s. bounded and possesses all its moments, arguments of the type used in §20.1 might be brought to bear to show that $\bar{Y}_n \to_{L_2} \mu_y$ (say). Some other approach must then be used to show that $\bar{Z}_n \to_{\mathrm{pr}} \mu_z = -\mu_y$. An obvious technique would be to assume uniform integrability of $\{|X_t|^p\}$. In this case, $\sup_t \mathrm{E}|Z_t|^p$ can be made as small as desired by choosing B large enough, leading (via the Minkowski inequality, for example) to an L_p-convergence result for $\bar{Z}_n$.

In §20.1 dependence was modelled by the autocorrelation structure, but a different approach to limiting dependence is called for here. The property of uncorrelatedness, that $\mathrm{E}(X_t X_s) = \mathrm{E}(X_t)\mathrm{E}(X_s)$ for $t \neq s$, does not require the existence of second moments but it does not follow that Y_t is serially uncorrelated just because X_t is. Serial independence of X_t would imply serial independence of Y_t but this is a strong restriction. However, if X_t has the martingale difference property, a mild strengthening of uncorrelatedness, *this* property is passed on to Y_t after a centring adjustment. This is the clever idea behind the next result, based on a theorem of Y. S. Chow ([32]). In §20.5 the m.d. assumption will be relaxed to a mixingale assumption.

Again, an array formulation is adopted here. Similarly to **20.10** the theorems are easily specialized to the case of sample averages, as in §19.6, but in subsequent chapters array results will be indispensable.

20.11 Theorem Let $\{X_{nt}, \mathcal{F}_{nt}\}$ be a m.d., $\{c_{nt}\}$ a positive array, and $\{k_n\}$ an increasing integer sequence with $k_n \uparrow \infty$. If, for $1 \le p \le 2$,

(a) $\{|X_{nt}/c_{nt}|^p\}$ is uniformly integrable

(b) $\limsup_{n\to\infty} \sum_{t=1}^{k_n} c_{nt} < \infty$

(c) $\lim_{n\to\infty} \sum_{t=1}^{k_n} c_{nt}^2 = 0,$

then $\sum_{t=1}^{k_n} X_{nt} \to_{L_p} 0.$ □

The leading specialization of this result is where $X_{nt} = X_t/a_n$, where $\{X_t, \mathcal{F}_t\}$ is a m.d. sequence with $\mathcal{F}_{nt} = \mathcal{F}_t$ and $\{a_n\}$ is a positive constant sequence. This deserves stating as a corollary, since the formulation can be made slightly more transparent.

20.12 Corollary Suppose $\{X_t, \mathcal{F}_t\}_0^\infty$ is a m.d. sequence and $\{b_t\}$, $\{a_n\}$, and $\{k_n\}$ are constant positive sequences with $a_n \uparrow \infty$ and $k_n \uparrow \infty$ and satisfying

(a) $\{|X_t/b_t|^p\}$ is uniformly integrable, $1 \le p \le 2$

(b) $\sum_{t=1}^{k_n} b_t = O(a_n)$

(c) $\sum_{t=1}^{k_n} b_t^2 = o(a_n^2),$

then $a_n^{-1} \sum_{t=1}^{k_n} X_t \to_{L_p} 0.$

Proof Immediate from **20.11**, defining $X_{nt} = X_t/a_n$ and $c_{nt} = b_t/a_n$. ∎

Be careful to distinguish the constants a_n and k_n. Although both are equal to n in the sample-average case, more generally their roles are quite different. The case with k_n different from n typically arises in 'blocking' arguments, where the array coordinates are generated from successive blocks of underlying sequence coordinates. One possibility is $k_n = [n^a]$ for $\alpha \in (0,1)$ ($[x]$ denoting the largest integer below x) where the length of a block does not exceed $[n^{1-\alpha}]$. For an application of this sort see §25.4.

Conditions **20.12**(b) and (c) together imply $a_n \uparrow \infty$, so this does not need to be separately asserted. To form a clear idea of the role of the assumptions, it is helpful to suppose that b_t and a_n are regularly varying functions of their arguments. It is easily verified by **2.17** that the conditions are observed if $b_t \sim t^\beta$ for any $\beta \geq -1$, by choosing $a_n \sim n^{1+\beta}$ for $\beta > -1$ and $a_n \sim \log n$ for $\beta = -1$. In particular, setting $b_t = 1$ for all t and $a_n = k_n = n$ yields

$$||\bar{X}_n||_p \to 0. \tag{20.27}$$

Choosing $a_n = \sum_{t=1}^{k_n} b_t$ will automatically satisfy condition (b), and condition (c) will also hold when $b_t = O(t^\beta)$. On the other hand, a case where the conditions fail is where $b_1 = 2$ and for $t > 1$, $b_t = 2 + \sum_{s=1}^{t-1} b_s = 2^t$. In this case condition (b) imposes the requirement $b_n = O(a_n)$ so that $b_n^2 = O(a_n^2)$, contradicting condition (c). The growth rate of b_t exceeds that of t^β for every $\beta > 0$.

Proof of 20.11 Uniform integrability implies that

$$\sup_{n,t} \mathrm{E}(|X_{nt}/c_{nt}|^p 1_{\{|X_{nt}/c_{nt}|>M\}}) \to 0 \text{ as } M \to \infty.$$

Therefore, for $\varepsilon > 0$ there is a constant $B_\varepsilon < \infty$ such that

$$\sup_{n,t}\{||X_{nt}1_{\{|X_{nt}|>B_\varepsilon c_{nt}\}}||_p/c_{nt}\} \leq \varepsilon. \tag{20.28}$$

Define $Y_{nt} = X_{nt}1_{\{|X_{nt}|\leq B_\varepsilon c_{nt}\}}$ and $Z_{nt} = X_{nt} - Y_{nt}$. Then, since $E(X_{nt}|\mathcal{F}_{n,t-1}) = 0$, $X_{nt} = Y_{nt} - E(Y_{nt}|\mathcal{F}_{n,t-1}) + Z_{nt} - E(Z_{nt}|\mathcal{F}_{n,t-1})$. By the Minkowski inequality

$$\begin{aligned}\left\|\sum_{t=1}^{k_n} X_{nt}\right\|_p \leq \left\|\sum_{t=1}^{k_n}(Y_{nt} - \mathrm{E}(Y_{nt}|\mathcal{F}_{n,t-1}))\right\|_p \\ + \left\|\sum_{t=1}^{k_n}(Z_{nt} - \mathrm{E}(Z_{nt}|\mathcal{F}_{n,t-1}))\right\|_p.\end{aligned} \tag{20.29}$$

Consider each of these right-hand-side terms. First,

$$\begin{aligned}\left\|\sum_{t=1}^{k_n}(Y_{nt}-\mathrm{E}(Y_{nt}|\mathcal{F}_{n,t-1}))\right\|_p &\leq \left\|\sum_{t=1}^{k_n}(Y_{nt}-\mathrm{E}(Y_{nt}|\mathcal{F}_{n,t-1}))\right\|_2\\ &= \left(\sum_{t=1}^{k_n}\mathrm{E}(Y_{nt}-\mathrm{E}(Y_{nt}|\mathcal{F}_{n,t-1}))^2\right)^{1/2}\\ &\leq \left(\sum_{t=1}^{k_n}\mathrm{E}Y_{nt}^2\right)^{1/2} \leq B_\varepsilon\left(\sum_{t=1}^{k_n}c_{nt}^2\right)^{1/2}. \end{aligned} \tag{20.30}$$

The first inequality in (20.30) is Liapunov's inequality and the equality follows because $\{Y_{nt}-E(Y_{nt}|\mathcal{F}_{n,t-1})\}$ is a m.d. and hence orthogonal. Second,

$$\begin{aligned}\left\|\sum_{t=1}^{k_n}(Z_{nt}-\mathrm{E}(Z_{nt}|\mathcal{F}_{n,t-1}))\right\|_p &\leq \sum_{t=1}^{k_n}\|Z_{nt}\|_p+\sum_{t=1}^{k_n}\|\mathrm{E}(Z_{nt}|\mathcal{F}_{n,t-1})\|_p\\ &\leq 2\sum_{t=1}^{k_n}\|Z_{nt}\|_p \leq 2\varepsilon\sum_{t=1}^{k_n}c_{nt}. \end{aligned} \tag{20.31}$$

The second inequality here follows because

$$\mathrm{E}|\mathrm{E}(Z_{nt}|\mathcal{F}_{n,t-1})|^p \leq \mathrm{E}(\mathrm{E}(|Z_{nt}|^p|\mathcal{F}_{n,t-1})) = \mathrm{E}|Z_{nt}|^p$$

from, respectively, the conditional Jensen inequality (**10.19**) and the LIE. The last is by (20.28). It follows by (c) that for $\varepsilon > 0$ there exists $N_\varepsilon \geq 1$ such that, for $n \geq N_\varepsilon$,

$$\sum_{t=1}^{k_n}c_{nt}^2 \leq B_\varepsilon^{-2}\varepsilon^2. \tag{20.32}$$

Putting together (20.29) with (20.30) and (20.31) shows that

$$\left\|\sum_{t=1}^{k_n}X_{nt}\right\|_p \leq B\varepsilon \tag{20.33}$$

for $n \geq N_\varepsilon$, where $B = 1+2\sum_{t=1}^{k_n}c_{nt} < \infty$ by condition (b). Since ε is arbitrary, this completes the proof. ∎

The weak law for martingale differences follows directly, on applying **19.14**.

20.13 Corollary Under the conditions of **20.11** or **20.12**, $\sum_{t=1}^{k_n} X_{nt} \to_{\text{pr}} 0$. □

Taking the case $p = 1$, $c_{nt} = 1/n$, and $k_n = n$ gives the result that uniform integrability of $\{X_t\}$ is sufficient for convergence in probability of the sample mean $\bar{X}_n$. This cannot be significantly weakened even if the martingale difference assumption is replaced by independence. Assuming identically distributed coordinates, the explicit requirement of uniform integrability can also be dropped and L_1-boundedness is enough, though of course this is only because the uniform property is subsumed under the stationarity.

Condition (b) in **20.11** can be replaced by

$$\text{(b}'\text{)} \quad \limsup_{n\to\infty} k_n^{p-1} \sum_{t=1}^{k_n} c_{nt}^p < \infty.$$

It suffices for the two terms on the majorant side of (20.29) to converge in L_p and the c_r inequality (**9.32**) can be used instead of the Minkowski inequality in (20.31) to obtain

$$\mathrm{E}\left|\sum_{t=1}^{k_n}(Z_{nt} - \mathrm{E}(Z_{nt}|\mathcal{F}_{n,t-1}))\right|^p \le \varepsilon(2k_n)^{p-1}\sum_{t=1}^{k_n} c_{nt}^p. \tag{20.34}$$

However, the gain in generality is notional. Condition (b′) imposes the condition $\limsup_{t,n\to\infty} k_n^p c_{nt}^p < \infty$ and if this is true the same property obviously extends to $\{k_n c_{nt}\}$. For concreteness put $c_{nt} = b_t/a_n$ as in **20.12** with $b_t \sim t^\beta$ and $a_n \sim n^\gamma$, where β and γ can be any real constants. With $k_n \sim n^\alpha$ for $\alpha > 0$, the majorant side of (20.34) is bounded if $\alpha(1+\beta) - \gamma \le 0$, *independent* of the value of p. This condition is automatically satisfied as an equality by setting $a_n = \sum_{t=1}^{k_n} b_t$, but note how the choice of a_n can accommodate different choices of k_n.

Nonetheless, in some situations condition (b) is stronger than what is known to be sufficient. For the case $p = 2$ it can be omitted, in addition to weakening the martingale difference assumption to uncorrelatedness and uniform integrability to simple L_2-boundedness. Here is the array version of **20.4**, with the conditions cast in the framework of **20.11** for comparability, although all they do is to ensure that the variance of the partial sums goes to zero.

20.14 Corollary If $\{X_{nt}\}$ is a zero-mean stochastic array with $\mathrm{E}(X_{nt}X_{ns}) = 0$ for $t \neq s$ and

(a) $\{X_{nt}/c_{nt}\}$ is uniformly L_2-bounded

(b) $\lim_{n\to\infty}\sum_{t=1}^{k_n} c_{nt}^2 = 0$,

then $\sum_{t=1}^{k_n} X_{nt} \to_{L_2} 0$. □

20.5 Mixingale Weak Laws

To generalize the last results from martingale differences to mixingales is not too difficult. The basic tool is the 'telescoping series' argument developed in §17.2. The array element X_{nt} can be decomposed into a finite sum of martingale differences, to which **20.11** can be applied, and two residual components that can be treated as negligible. The following result, from [39], is an extension to the heterogeneous case of a theorem due to Andrews ([6]).

20.15 Theorem Let the array $\{X_{nt}, \mathcal{F}_{nt}\}_{-\infty}^{\infty}$ be an L_1-mixingale with respect to a constant array $\{c_{nt}\}$ and k_n an increasing integer-valued function of n with $k_n \uparrow \infty$. If

(a) $\{X_{nt}/c_{nt}\}$ is uniformly integrable

(b) $\limsup_{n\to\infty}\sum_{t=1}^{k_n} c_{nt} < \infty$

(c) $\lim_{n\to\infty}\sum_{t=1}^{k_n} c_{nt}^2 = 0$,

then $\sum_{t=1}^{k_n} X_{nt} \to_{L_1} 0$.

There is no restriction on the mixingale size here. It suffices simply for the mixingale coefficients to tend to zero. The remarks following **20.11** apply here in just the same way. In particular, if X_t is an L_1-mixingale sequence and $\{X_t/b_t\}$ is uniformly integrable for positive constants $\{b_t\}$, the theorem holds for $X_{nt} = X_t/a_n$ and $c_{nt} = b_t/a_n$ where $a_n = \sum_{t=1}^{n} b_t$. The corresponding results for mixing sequences can be found from Theorems **15.2** and **15.4** and those for processes that are NED on a mixing sequence from **18.6** and **18.7**. It is sufficient for, say, X_{nt} to be L_r-bounded for $r > 1$ and L_p-NED on an α-mixing process for $p \geq 1$. Again, no size restrictions need to be specified. Uniform integrability of $\{X_{nt}/c_{nt}\}$ will obtain in those cases where $\|X_{nt}\|_r$ is finite for $r > 1$ and each t and the NED constants likewise satisfy $d_{nt} \gg \|X_{nt}\|_r$.

A simple lemma is required for the proof. For an arbitrary $\mathcal{F}$-measurable r.v. X_{nt} let $\mathrm{E}_s X_{nt}$ stand for $\mathrm{E}(X_{nt}|\mathcal{F}_{ns})$. There is no requirement for X_{nt} to be $\mathcal{F}_{nt}$-measurable and $t - s$ can be of either sign.

20.16 Lemma If the array $\{X_{nt}/c_{nt}\}$ is uniformly integrable, so is the array $\{\mathrm{E}_s X_{nt}/c_{nt}\}$.

Proof By the necessity part of **12.10**, for any $\varepsilon > 0 \ \exists\ \delta > 0$ such that

$$\sup_{n,t}\left\{\sup \int_E |X_{nt}/c_{nt}|\mathrm{d}P\right\} < \varepsilon \tag{20.35}$$

where the inner supremum is taken over all $E \in \mathcal{F}$ satisfying $P(E) < \delta$. Since $\mathcal{F}_{ns} \subseteq \mathcal{F}$, (20.35) also holds when the supremum is taken over $E \in \mathcal{F}_{ns}$ satisfying $P(E) < \delta$. For any such E,

$$\int_E |X_{nt}/c_{nt}|\mathrm{d}P = \int_E \mathrm{E}_s|X_{nt}/c_{nt}|\mathrm{d}P \geq \int_E |\mathrm{E}_s X_{nt}/c_{nt}|\mathrm{d}P \tag{20.36}$$

by definition of $\mathrm{E}_s(\cdot)$ and the conditional Jensen inequality (**10.19**). Therefore, for $\varepsilon > 0 \ \exists\ \delta > 0$ such that

$$\sup_{n,t}\left\{\sup \int_E |\mathrm{E}_s X_{nt}/c_{nt}|\mathrm{d}P\right\} < \varepsilon \tag{20.37}$$

where the inner supremum is over $E \in \mathcal{F}_{ns}$ satisfying $P(E) < \delta$. Since $\mathrm{E}_s X_{nt}$ is $\mathcal{F}_{ns}$-measurable, uniform integrability holds by the sufficiency part of **12.10**. ∎

Proof of 20.15 Fix an integer j and let

$$Y_{nj} = \sum_{t=1}^{k_n}(\mathrm{E}_{t+j}X_{nt} - \mathrm{E}_{t+j-1}X_{nt}).$$

The sequence $\{Y_{nj}, \mathcal{F}_{n,n+j}\}_{n=1}^{\infty}$ is a martingale, for each j. Since the array

$$\{(\mathrm{E}_{t+j}X_{nt} - \mathrm{E}_{t+j-1}X_{nt})/c_{nt}\}$$

is uniformly integrable by (a) and **20.16**, it follows by (b) and (c) and **20.11** that

$$Y_{nj} \xrightarrow{L_1} 0. \tag{20.38}$$

Now express $\sum_{t=1}^{k_n} X_{nt}$ as a telescoping sum. For any $M \geq 1$,

$$\sum_{j=1-M}^{M-1} Y_{nj} = \sum_{t=1}^{k_n} \mathrm{E}_{t+M-1}X_{nt} - \sum_{t=1}^{k_n} \mathrm{E}_{t-M}X_{nt} \tag{20.39}$$

and hence

$$\sum_{t=1}^{k_n} X_{nt} = \sum_{j=1-M}^{M-1} Y_{nj} + \sum_{t=1}^{k_n}(X_{nt} - \mathrm{E}_{t+M-1}X_{nt}) + \sum_{t=1}^{k_n} \mathrm{E}_{t-M}X_{nt}. \tag{20.40}$$

The triangle inequality and the L_1-mixingale property now give

$$\begin{aligned} \mathrm{E}\left|\sum_{t=1}^{k_n} X_{nt}\right| &\leq \sum_{j=1-M}^{M-1} \mathrm{E}|Y_{nj}| + \sum_{t=1}^{k_n} \mathrm{E}|X_{nt} - \mathrm{E}_{t+M-1}X_{nt}| + \sum_{t=1}^{k_n} \mathrm{E}|\mathrm{E}_{t-M}X_{nt}| \\ &\leq \sum_{j=1-M}^{M-1} \mathrm{E}|Y_{nj}| + 2\zeta_M \sum_{t=1}^{k_n} c_{nt} \end{aligned} \tag{20.41}$$

where ζ_M is the sequence defined in **17.1**. According to the assumptions, the second term in the third member of (20.41) is $O(M^{-\delta})$ for some $\delta > 0$ and given $\varepsilon > 0$ there exists M_ε such that $\zeta_M \sum_{t=1}^{k_n} c_{nt} < \frac{1}{2}\varepsilon$ for $M \geq M_\varepsilon$. There is an n large enough that the sum of $2M - 1$ terms on the right-hand side of (20.41) is smaller than $\frac{1}{2}\varepsilon$ for any finite M, by (20.38). So by choosing $M \geq M_\varepsilon$, $E|\sum_{t=1}^{k_n} X_{nt}| < \varepsilon$ when n is large enough. The theorem is now proved since ε is arbitrary. ∎

A comparison with the results of §20.1 is instructive. In an L_2-bounded process, the L_2-mixingale property would be a stronger form of dependence restriction than the limiting uncorrelatedness specified in **20.5**, just as the martingale property is stronger than simple uncorrelatedness. The value of the present result is the substantial weakening of the moment conditions.

Theorem **20.15** does not establish mean squared convergence, but this can be shown straightforwardly in the following result for sequences, also from Andrews ([6]).

20.17 Theorem Let $\{X_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ be an L_2-mixingale with mixingale indices ζ_m and constant array $c_t \geq \|X_t\|_2$ and $k_n \uparrow \infty$ an increasing integer-valued function of n. If $\sup_t c_t < \infty$ and $n^{-1}\sum_{m=1}^{n} \zeta_m \to 0$, then $\bar{X}_n \to_{L_2} 0$.

Proof For $m = 1,2,3,4,\ldots$ let $s = [m/2]$. Note that

$$\begin{aligned}|\mathrm{E}(X_tX_{t+m})| &= |\mathrm{E}((X_t - \mathrm{E}_{t+s}X_t)X_{t+m}) + \mathrm{E}(X_{t+m}\mathrm{E}_{t+s}X_t)| \\ &\le |\mathrm{E}((X_t - \mathrm{E}_{t+s}X_t)X_{t+m})| + |\mathrm{E}(\mathrm{E}_{t+s}X_{t+m}\mathrm{E}_{t+s}X_t)| \\ &\le \|X_{t+m}\|_2\|X_t - \mathrm{E}_{t+s}X_t\|_2 + \|\mathrm{E}_{t+s}X_{t+m}\|_2\|\mathrm{E}_{t+s}X_t\|_2 \\ &\le \|X_{t+m}\|_2 c_t\zeta(s+1) + \|X_t\|_2 c_{t+m}\zeta(m-s) \\ &\le 2\sup_t c_t^2\zeta([m/2]),\end{aligned}$$

invoking first the LIE and the triangle inequality, next the Cauchy–Schwarz inequality, and finally the mixingale assumption. Hence,

$$\begin{aligned}\mathrm{E}\Bigl(\frac{1}{n}\sum_{t=1}^{n}X_t\Bigr)^2 &\le \frac{1}{n^2}\Bigl(\sum_{t=1}^{n}\mathrm{E}X_t^2 + 2\sum_{m=1}^{n-t}\mathrm{E}(X_tX_{t+m})\Bigr) \\ &\le \frac{1}{n}\sup_t c_t^2\Bigl(1 + 8\sum_{m=1}^{(n-t)/2}\zeta(m)\Bigr) \to 0\end{aligned}$$

as $n \to \infty$. ∎

The conditions of this result are essentially comparable to those of Theorem **20.1**, interpreted here within the mixingale framework. It is sufficient that the mixingale indices tend to zero eventually, no matter how slowly. It is also clear that summability of the mixingale indices is sufficient for $\mathrm{E}(\bar{X}_n)^2 = O(n^{-1})$, to match the rate of convergence achieved in an independent sequence.

20.6 Approximable Processes

There remains the possibility of cases in which the mixingale property is not easily established—perhaps because of a nonlinear transformation of an L_p-NED process which cannot be shown to preserve the requisite moments for application of the results in §18.3. In such cases the theory of §18.5 may yield a result. On the assumption that the approximator sequence is mixing, so that its mean deviations converge in probability by **20.15**, it will be sufficient to show that this implies the convergence of the approximable sequence. This is the object of the following theorem.

20.18 Theorem Suppose that, for each $m \in \mathbb{N}$, $\{h_{nt}^m\}$ is a stochastic array and the centred array $\{h_{nt}^m - \mathrm{E}(h_{nt}^m)\}$ satisfies the conditions of **20.15**. If the array

$\{X_{nt}\}$ is L_1-approximable by $\{h_{nt}^m\}$ with respect to a constant array $\{d_{nt}\}$ and $\limsup_{n\to\infty} \sum_{t=1}^{k_n} d_{nt} \le B < \infty$, then $\sum_{t=1}^{k_n} X_{nt} \to_{\text{pr}} 0$. □

Establishing the conditions of the theorem will typically be achieved using **18.22**, by showing that X_{nt} is L_r-bounded for $r > 1$ and approximable in probability on h_{nt}^m for each m, the latter being m-order lag functions of a mixing array of any size.

Proof of 20.18 Since

$$\left|\sum_{t=1}^{k_n} X_{nt}\right| \le \left|\sum_{t=1}^{k_n}(X_{nt} - h_{nt}^m)\right| + \left|\sum_{t=1}^{k_n}(h_{nt}^m - \mathrm{E}(h_{nt}^m))\right| + \left|\sum_{t=1}^{k_n}\mathrm{E}(h_{nt}^m)\right| \tag{20.42}$$

by the triangle inequality,

$$\begin{aligned} P\left(\left|\sum_{t=1}^{k_n} X_{nt}\right| > \delta\right) \le P\left(\left|\sum_{t=1}^{k_n}(X_{nt} - h_{nt}^m)\right| > \frac{\delta}{3}\right) + P\left(\left|\sum_{t=1}^{k_n}(h_{nt}^m - \mathrm{E}(h_{nt}^m))\right| > \frac{\delta}{3}\right) \\ + P\left(\left|\sum_{t=1}^{k_n}\mathrm{E}(h_{nt}^m)\right| > \frac{\delta}{3}\right) \end{aligned} \tag{20.43}$$

for $\delta > 0$ by subadditivity, since the event whose probability is on the minorant side implies at least one of those on the majorant. By the Markov inequality,

$$\begin{aligned} P\left(\left|\sum_{t=1}^{k_n}(X_{nt} - h_{nt}^m)\right| > \frac{\delta}{3}\right) &\le \frac{3}{\delta}\mathrm{E}\left|\sum_{t=1}^{k_n}(X_{nt} - h_{nt}^m)\right| \\ &\le \frac{3}{\delta}\sum_{t=1}^{k_n}\mathrm{E}|X_{nt} - h_{nt}^m| \\ &\le \frac{3}{\delta}\left(\sum_{t=1}^{k_n} d_{nt}\right)\nu_m \end{aligned} \tag{20.44}$$

where ν_m is defined in **18.19**. $P(|\sum_{t=1}^{k_n}\mathrm{E}(h_{nt}^m)| > \delta/3)$ is equal to either 0 or 1, depending on whether the non-stochastic inequality holds or does not hold. By the fact that $\mathrm{E}(X_{nt}) = 0$ and L_1-approximability,

$$|\mathrm{E}(h_{nt}^m)| = |\mathrm{E}(X_{nt}) - \mathrm{E}(h_{nt}^m)| \le \mathrm{E}|X_{nt} - h_{nt}^m| \le d_{nt}\nu_m \tag{20.45}$$

and hence

$$\left|\sum_{t-1}^{k_n}\mathrm{E}(h_{nt}^m)\right| \le \sum_{t=1}^{k_n}|\mathrm{E}(h_{nt}^m)| \le \sum_{t=1}^{k_n} d_{nt}\nu_m \le B\nu_m. \tag{20.46}$$

Therefore for each $m \in \mathbb{N}$,

$$
\begin{aligned}
\limsup_{n\to\infty} P\left(\left|\sum_{t=1}^{k_n} X_{nt}\right| > \delta\right) & \\
&\leq \frac{3B}{\delta}\nu_m + \limsup_{n\to\infty} P\left(\left|\sum_{t=1}^{k_n}(h_{nt}^m - \mathrm{E}(h_{nt}^m))\right| > \frac{\delta}{3}\right) + 1_{\{B\nu_m > \delta/3\}} \\
&= \frac{3B}{\delta}\nu_m + 1_{\{B\nu_m > \delta/3\}} \qquad (20.47)
\end{aligned}
$$

by the assumption that h_{nt}^m satisfies the WLLN for each $m \in \mathbb{N}$. The proof is completed by letting $m \to \infty$. ∎

21

The Strong Law of Large Numbers

21.1 Technical Tricks for Proving LLNs

This chapter explores the strong law under a range of different assumptions, from independent sequences to near-epoch dependent functions of mixing processes. Many of the proofs are based on one or more of the following collection of ingenious technical lemmas. The reader has the option of skipping ahead to §21.2 and referring back to this section as necessary, but there is something to be said for forming an impression of the method of attack at the outset.

To begin, here is a basic result on convergence that shows why maximal inequalities (e.g. **16.20**, **16.21**, **17.10**, and **17.12**) are important.

21.1 Lemma Let $\{X_t\}_1^\infty$ be a stochastic sequence in a probability space $(\Omega, \mathcal{F}, P)$ and let $S_n = \sum_{t=1}^n X_t$ for $n \geq 1$ and $S_0 = 0$. For $\omega \in \Omega$, let

$$M(\omega) = \inf_m \Big(\sup_{j>m} |S_j(\omega) - S_m(\omega)| \Big). \tag{21.1}$$

If $P(M > \varepsilon) = 0$ for all $\varepsilon > 0$, then $S_n \to_{\text{a.s.}} S$.

Proof By the Cauchy criterion for convergence, the realization $\{S_n(\omega)\}$ converges if there is an m such that $|S_j - S_m| \leq \varepsilon$ for all $j > m$, for all $\varepsilon > 0$; in other words, it converges if $M(\omega) \leq \varepsilon$ for all $\varepsilon > 0$. ■

This result may be applied in the following way.

21.2 Corollary Let $\{c_t\}_1^\infty$ be a sequence of constants and suppose there exists $p > 0$ such that for every $m \geq 0$ and $n > m$ and every $\varepsilon > 0$,

$$P\Big(\max_{m<j\leq n} |S_j - S_m| > \varepsilon \Big) \leq \frac{K}{\varepsilon^p} \sum_{t=m+1}^{n} c_t^p \tag{21.2}$$

where K is a finite constant. If $\sum_{t=1}^\infty c_t^p < \infty$, then $S_n \to_{\text{a.s.}} S$. □

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0021

The specification of the majorant side of (21.2) is of course to allow this result to be meshed with known properties of the sequence via, for example, the Kolmogorov inequality (**16.20**).

Proof of 21.2 Since $\{c_t^p\}$ is summable it follows by **2.15** that $\lim_{m\to\infty}\sum_{t=m+1}^{\infty} c_t^p = 0$. Let M be the r.v. in (21.1). By definition, $M \le \sup_{j>m}|S_j - S_m|$ for any $m > 0$ and hence

$$\begin{aligned} P(M>\varepsilon) &\le \lim_{m\to\infty} P\Big(\sup_{j>m}|S_j - S_m| > \varepsilon\Big) \\ &\le \frac{K}{\varepsilon^p}\lim_{m\to\infty}\sum_{t=m+1}^{\infty} c_t^p = 0 \end{aligned} \tag{21.3}$$

where the final inequality is the limiting case of (21.2). **21.1** completes the proof. ∎

Notice how this proof does not make a direct appeal to the Borel–Cantelli lemma to get a.s. convergence. The method is closer to that of **19.3**. The essential trick with a maximal inequality is to put a bound on the probability of *all* occurrences of a certain type of event by specifying a probability for the most extreme of them.

Since S is finite almost surely, $\bar{X}_n \to_{\text{a.s.}} 0$ is an instant corollary of **21.2**. However, the result can be used in a more subtle way in conjunction with Kronecker's lemma (**2.35**). If $\sum_{t=1}^{n} Y_t$ converges a.s. where $Y_t = X_t/a_t$ and $\{a_t\}$ is a monotone sequence of positive constants with $a_n \uparrow \infty$, it follows that $a_n^{-1}\sum_{t=1}^{n} X_t \to_{\text{a.s.}} 0$. This is a much weaker condition than the convergence of $\sum_{t=1}^{n} X_t$ itself. Most applications feature $a_t = t$, but the more general formulation also has uses.

There is a standard device for extending a.s. convergence to a wider class of sequences, once it has been proved for a given class. This is the method of *equivalent sequences*. Sequences $\{X_t\}_1^\infty$ and $\{Y_t\}_1^\infty$ are said to be equivalent if

$$\sum_{t=1}^{\infty} P(X_t \ne Y_t) < \infty. \tag{21.4}$$

By the first Borel–Cantelli lemma (**19.2**(i)), (21.4) implies $P(X_t \ne Y_t,\ \text{i.o.}) = 0$. In other words, only on a set of probability measure zero are there more than a finite number of t for which $X_t(\omega) \ne Y_t(\omega)$.

21.3 Theorem If X_t and Y_t are equivalent, $\sum_{t=1}^{n}(X_t - Y_t)$ converges a.s.

Proof By definition of equivalence and **19.2**(i) there exists a subset C of Ω, with $P(\Omega - C) = 0$ and with the following property: for all $\omega \in C$, there is a finite $n_0(\omega)$ such that $X_t(\omega) = Y_t(\omega)$ for $t > n_0(\omega)$. Hence

$$\sum_{t=1}^{n}\big(X_t(\omega)-Y_t(\omega)\big)=\sum_{t=1}^{n_0(\omega)}\big(X_t(\omega)-Y_t(\omega)\big),\ \forall\, n\geq n_0(\omega)$$

and the sum converges, for all $\omega \in C$. ∎

The equivalent sequences concept is often put to use by means of the following theorem.

21.4 Theorem Let $\{X_t\}_1^\infty$ be a zero-mean random sequence satisfying

$$\sum_{t=1}^{\infty}\mathrm{E}|X_t|^p/a_t^p<\infty \tag{21.5}$$

for some $p \geq 1$ and a sequence of positive constants $\{a_t\}$. Then, putting 1_t^a for the indicator function $1_{\{|X_t|\leq a_t\}}(\omega)$,

$$\sum_{t=1}^{\infty}P(|X_t|>a_t)<\infty, \tag{21.6}$$

$$\sum_{t=1}^{\infty}|\mathrm{E}(X_t1_t^a)|/a_t<\infty, \tag{21.7}$$

and for any $r \geq p$,

$$\sum_{t=1}^{\infty}\mathrm{E}(|X_t|^r1_t^a)/a_t^r<\infty. \quad \square \tag{21.8}$$

The idea behind this result may be apparent. The indicator function is used to truncate a sequence, replacing a member by 0 if it exceeds a given absolute bound. The ratio of the truncated sequence to the bound cannot exceed 1 and possesses all its absolute moments, while according to inequality (21.6), the truncated sequence is equivalent to the original under condition (21.5). Proving a strong law under (21.5) can therefore be accomplished by proving a strong law for a truncated sequence, subject to (21.7) and (21.8).

Proof of Theorem 21.4 The t^{th} term of each of the sums (21.6), (21.7), and (21.8) is shown to be dominated by the corresponding term in (21.5). First, since $|X_t(\omega)|^p/a_t^p>1$ for $\omega\in\{|X_t|>a_t\}$ and $\mathrm{E}(|X_t|^p1_t^a)\geq 0$,

$$\begin{aligned}P(|X_t|>a_t)&=\mathrm{E}(1-1_t^a)\\&\leq \mathrm{E}\big(|X_t|^p(1-1_t^a)\big)/a_t^p\\&\leq \mathrm{E}(|X_t|^p)/a_t^p.\end{aligned} \tag{21.9}$$

Note, the random variable appearing in the second member is either 0 or 1, while that appearing in the third member is either 0 or exceeds 1. Next, since $\mathrm{E}(X_t) = 0$, $\mathrm{E}(X_t 1_t^a) = -\mathrm{E}\big(X_t(1 - 1_t^a)\big)$ and hence

$$\begin{aligned}
|\mathrm{E}(X_t 1_t^a)|/a_t &= |\mathrm{E}\big(X_t(1 - 1_t^a)\big)|/a_t \\
&\le \mathrm{E}\big(|X_t|(1 - 1_t^a)\big)/a_t \\
&\le \mathrm{E}\big(|X_t|^p(1 - 1_t^a)\big)/a_t^p \\
&\le \mathrm{E}(|X_t|^p)/a_t^p. \qquad (21.10)
\end{aligned}$$

The first inequality is the modulus inequality and the second is because on the event $\{|X_t| > a_t\}$, $(|X_t|/a_t)^p \ge |X_t|/a_t$ for $p \ge 1$. Finally, by similar arguments,

$$\begin{aligned}
\mathrm{E}(|X_t|^r 1_t^a)/a_t^r &\le \mathrm{E}(|X_t|^p 1_t^a)/a_t^p \text{ for } p \le r \\
&\le \mathrm{E}(|X_t|^p)/a_t^p. \qquad (21.11)
\end{aligned}$$

The theorem follows on summing over t. ∎

There are a number of variations on this basic result. The first is a version for martingale differences in terms of the one-step-ahead conditional moments, where the weight sequence is also allowed to be stochastic. The style of this result is appropriate to the class of martingale limit theorems to be examined in §21.4, in which almost sure equivalence is established between sets on which certain conditions obtain and on which sequences converge.

21.5 Corollary Let $\{X_t, \mathcal{F}_t\}$ be a m.d. sequence, let $\{W_t\}$ be a sequence of positive $\mathcal{F}_{t-1}$-measurable r.v.s, and for some $p \ge 1$ let

$$D = \left\{\omega : \sum_{t=1}^{\infty} \mathrm{E}(|X_t|^p|\mathcal{F}_{t-1})(\omega)/W_t^p(\omega) < \infty\right\} \in \mathcal{F}. \qquad (21.12)$$

Also define, for any $r \ge p$,

$$D_1 = \left\{\omega : \sum_{t=1}^{\infty} P(|X_t| > W_t|\mathcal{F}_{t-1})(\omega) < \infty\right\} \in \mathcal{F} \qquad (21.13)$$

$$D_2 = \left\{\omega : \sum_{t=1}^{\infty} |\mathrm{E}(X_t 1_t^W|\mathcal{F}_{t-1})(\omega)|/W_t(\omega) < \infty\right\} \in \mathcal{F} \qquad (21.14)$$

$$D_3 = \left\{\omega : \sum_{t=1}^{\infty} \mathrm{E}(|X_t|^r 1_t^W|\mathcal{F}_{t-1})(\omega)/W_t^r(\omega) < \infty\right\} \in \mathcal{F} \qquad (21.15)$$

where $1_t^W = 1_{\{|X_t| \le W_t\}}$ and let $D' = D_1 \cap D_2 \cap D_3$. Then $P(D - D') = 0$. In particular, if $P(D) = 1$ then $P(D') = 1$.

Proof It suffices to prove the three inequalities (21.9), (21.10), and (21.11) for the case of conditional expectations. Noting that $\mathrm{E}(X_t|\mathcal{F}_{t-1}) = 0$ a.s. and using the fact that W_t is $\mathcal{F}_{t-1}$-measurable, all of these go through unchanged, except that the conditional modulus inequality **10.15** is used to get (21.14). It follows that almost every $\omega \in D$ is in D'. ∎

Another version of this theorem uses a different truncation, with the truncated variable chosen to be a continuous function of X_t; see **18.13** and **18.14** to appreciate why this variation might be useful.

21.6 Corollary Let $\{X_t\}_1^\infty$ be a zero-mean random sequence satisfying (21.5) for $p \ge 1$. Define

$$Y_t = X_t 1_t^a / a_t + (X_t/|X_t|)(1 - 1_t^a) = \begin{cases} X_t/a_t, & |X_t| \le a_t \\ 1, & X_t > a_t \\ -1, & X_t < -a_t. \end{cases} \tag{21.16}$$

Then

$$\sum_{t=1}^{\infty} |\mathrm{E}(Y_t)| < \infty \tag{21.17}$$

and

$$\sum_{t=1}^{\infty} \mathrm{E}|Y_t|^r < \infty,\ r \ge p. \tag{21.18}$$

Proof Write $\pm a_t$ to denote $a_t X_t/|X_t|$. Inequalities (21.10) and (21.11) of **21.4** are adapted as follows.

$$\begin{aligned} |\mathrm{E}(Y_t)| &= |\mathrm{E}\big(X_t 1_t^a + (1 - 1_t^a)(\pm a_t)\big)|/a_t \\ &= |\mathrm{E}\big(X_t - (\pm a_t)\big)(1 - 1_t^a)|/a_t \\ &\le \mathrm{E}|X_t|(1 - 1_t^a)/a_t + \mathrm{E}|1 - 1_t^a| \\ &\le \mathrm{E}\big(|X_t|^p(1 - 1_t^a)\big)/a_t^p + P(|X_t| > a_t) \\ &\le 2\mathrm{E}(|X_t|^p)/a_t^p. \end{aligned} \tag{21.19}$$

The second equality in (21.19) is again because $\mathrm{E}(X_t) = 0$. The first inequality is an application of the modulus inequality and triangle inequalities in succession and the last one uses (21.9). By similar arguments, except that here the c_r inequality (**9.32**) is used in the second line,

$$\begin{aligned}\mathrm{E}(|Y_t|^r) &\leq \mathrm{E}|X_t 1_t^a + (1 - 1_t^a)(\pm a_t)|^r/a_t \\ &\leq 2^{r-1}\big(\mathrm{E}|X_t 1_t^a|^r/a_t^r + \mathrm{E}|(1 - 1_t^a)|^r\big) \\ &\leq 2^{r-1}\big(\mathrm{E}(|X_t|^p 1_t^a)/a_t^p + P(|X_t| > a_t)\big) \text{ for } p \leq r \\ &\leq 2^r \mathrm{E}(|X_t|^p)/a_t^p. \end{aligned} \tag{21.20}$$

The theorem follows on summing over t as before. ∎

Clearly, **21.5** could be adapted to this case if desired, but that extension will not be needed subsequently.

The last extension is relatively modest but permits summability conditions for norms to be applied.

21.7 Corollary (21.6), (21.7), (21.8), (21.17), and (21.18) all continue to hold if (21.5) is replaced by

$$\sum_{t=1}^{\infty} \mathrm{E}(|X_t|^p)^{1/q}/a_t^{p/q} < \infty \tag{21.21}$$

for any $q \geq 1$.

Proof The modified forms of (21.9) and of (21.19) and (21.20) (say) are

$$P(|X_t| > a_t) \leq P(|X_t| > a_t)^{1/q} \leq \big(\mathrm{E}(|X_t|^p)/a_t^p\big)^{1/q} \tag{21.22}$$

$$|\mathrm{E}(Y_t)| \leq |\mathrm{E}(Y_t)|^{1/q} \leq 2^{1/q}\big(\mathrm{E}(|X_t|^p)/a_t^p\big)^{1/q} \tag{21.23}$$

$$\mathrm{E}|Y_t|^r \leq (\mathrm{E}|Y_t|^r)^{1/q} \leq 2^{r/q}\big(\mathrm{E}(|X_t|^p)/a_t^p\big)^{1/q} \tag{21.24}$$

where in each case the first inequality is because the left-hand-side member does not exceed 1.

For example, by choosing $p = q$ the condition that the sequence $\{\|X_t/a_t\|_p\}$ is summable is seen to be sufficient for **21.4** and **21.6**.

21.2 The Case of Independence

The classic results on strong convergence are for the case of independent sequences. The following is the 'three-series theorem' of Kolmogorov:

21.8 Theorem Let $\{X_t\}$ be an independent sequence and $S_n = \sum_{t=1}^n X_t$. Iff

$$\sum_{t=1}^{\infty} P(|X_t| > a) < \infty \tag{21.25}$$

$$\sum_{t=1}^{\infty} \mathrm{E}(1_{\{|X_t|\le a\}}X_t) < \infty \tag{21.26}$$

$$\sum_{t=1}^{\infty} \mathrm{Var}(1_{\{|X_t|\le a\}}X_t) < \infty \tag{21.27}$$

for some fixed $a > 0$, then $S_n \to_{\text{a.s.}} S$. □

Since the event $\{S_n \to S\}$ is the same as the event $\{S_{n+1} \to S\}$, convergence is invariant to shift transformations. It is a remote event by **14.14** and hence in independent sequences occurs with probability either 0 or 1, according to **14.12**. **21.8** gives the conditions under which the probability is 1, rather than 0. As noted previously, the theorem has the immediate corollary that if $X_t = Y_t/a_t$ where $\{a_t\}$ is a monotone sequence of constants and $a_n \uparrow \infty$, then $a_n^{-1}\sum_{t=1}^n Y_t \to_{\text{a.s.}} 0$ by the Kronecker lemma.

The basic idea is to prove the convergence result for the truncated variables $1_{\{|X_t|\le a\}}X_t$ and then use the equivalent sequences theorem to extend it to X_t itself. A point to notice about the proof is that the necessity part does not assign a value to a. Convergence implies that (21.25)–(21.27) hold for *every* $a > 0$.

Proof of 21.8 Write $Y_t = 1_{\{|X_t|\le a\}}X_t$, so that the summands in (21.26) and (21.27) are respectively the means and variances of Y_t. The sequence $\{Y_t - \mathrm{E}(Y_t)\}$ is independent and hence a martingale difference. Putting S_n' to denote the partial sum of these terms, $S_n' - S_m' = \sum_{t=m+1}^n (Y_t - \mathrm{E}(Y_t))$ is a martingale for fixed $m \ge 0$ and $\sum_{t=m+1}^n \mathrm{Var}(Y_t) = \mathrm{Var}(S_n' - S_m')$. **16.20** combined with **21.2**, setting $p = 2$ in each case and putting $c_t^2 = \mathrm{Var}(Y_t)$ and $K = 1$, together yield the result that $S_n' \to_{\text{a.s.}} S'$ when (21.27) holds. If (21.26) holds, this further implies that $\sum_{t=1}^n Y_t$ itself converges. And then if (21.25) holds the sequences $\{X_t\}$ and $\{Y_t\}$ are equivalent and so $S_n \to_{\text{a.s.}} S$, by **21.3**. This proves sufficiency of the three conditions.

Conversely, suppose $S_n \to_{\text{a.s.}} S$. By **2.15** applied to $S_n(\omega)$ for each $\omega \in \Omega$, it follows that $\lim_{m\to\infty}\sum_{t=m}^{\infty} X_t = 0$ a.s. This means that $P(|X_t| > a, \text{i.o.}) = 0$, for any $a > 0$ and so (21.25) must follow by the divergence part of the Borel–Cantelli lemma (**19.2**(ii)). **21.3** then assures us that $\sum_{t=1}^n Y_t$ also converges a.s.

Write $s_n^2 = \sum_{t=1}^n \mathrm{Var}(Y_t)$. If $s_n^2 \to \infty$ as $n \to \infty$, $\sum_{t=1}^n (Y_t - \mathrm{E}(Y_t))/s_n$ fails to converge, but is asymptotically distributed as a standard Gaussian r.v. (This is the central limit theorem—see **24.6**.) This fact contradicts the possibility of $\sum_{t=1}^n Y_t$ converging, so s_n^2 is bounded in the limit which is equivalent to (21.27).

Finally, consider the sequence $\{Y_t - \mathrm{E}(Y_t)\}$. This has mean zero, the same variance as Y_t, and $P\big(|Y_t - \mathrm{E}(Y_t)| > 2a\big) = 0$ for all t. Hence, it satisfies the conditions (21.25)–(21.27) (in respect of the constant $2a$) and the sufficiency part of the theorem implies that $\sum_{t=1}^{n}\big(Y_t - \mathrm{E}(Y_t)\big)$ converges. And since $\sum_{t=1}^{n} Y_t$ converges, (21.26) must hold. This completes the proof of necessity. ∎

The sufficiency part of this result is subsumed under the weaker conditions of **21.13** below and is now mainly of historical interest; it is the necessity proof that is interesting, since it has no counterpart in the LLNs for dependent sequences. In these cases the divergence part of the Borel–Cantelli lemma is not available and it appears difficult to rule out special cases in which convergence is achieved with arbitrary moment conditions. Incidentally, Kolmogorov originally proved the maximal inequality of **16.20** cited in the proof for the independent case; but again, his result can now be subsumed under the case of martingale differences and does not need to be quoted separately.

An interesting contrast with the three-series theorem is provided by a result due to Etemadi ([72]) which uses a completely different approach. Initially, Etemadi considers the case where the series in question is non-negative.

21.9 Theorem If $\{X_t\}_{t=1}^{\infty}$ is an identically distributed and pairwise-independent non-negative sequence with $\mathrm{E}(X_1) < \infty$, then $n^{-1}\sum_{t=1}^{n} X_t \to \mathrm{E}(X_1)$ a.s. as $n \to \infty$.

Proof Let $Y_t = X_t 1_{\{X_t \le t\}}$ and observe that

$$\begin{aligned}\sum_{t=1}^{\infty} P(X_t \ne Y_t) &= \sum_{t=1}^{\infty} P(X_1 > t)\\ &\le \sum_{t=1}^{\infty}\int_{t-1}^{t} P(X_1 > x)\mathrm{d}x\\ &= \int_{0}^{\infty} P(X_1 > x)\mathrm{d}x = \mathrm{E}(X_1) < \infty\end{aligned}$$

where the final equality is by **9.22** with $r = 1$. Hence $\{X_t\}$ and $\{Y_t\}$ are equivalent sequences and by the Borel–Cantelli lemma **19.2**(i) $P(X_t \ne Y_t \text{ i.o.}) = 0$. Putting $\bar{Y}_n = n^{-1}\sum_{t=1}^{n} Y_t$, it will therefore suffice to show that $\bar{Y}_n \to \mathrm{E}(X_1)$. Noting that $\mathrm{E}(Y_t) = \mathrm{E}(X_1 1_{\{X_1 \le t\}}) \to \mathrm{E}(X_1)$ as $t \to \infty$ by the monotone convergence theorem (**4.7**), $\mathrm{E}(\bar{Y}_n) \to \mathrm{E}(X_1)$ by **2.16** so the condition to be shown is equivalent to

$$\bar{Y}_n - \mathrm{E}(\bar{Y}_n) \to 0 \text{ a.s. as } n \to \infty. \tag{21.28}$$

To show (21.28) it is convenient to consider the subsequence $\{n_k, k = 1, 2, \ldots\}$ where $n_k = [\rho^k]$ for some $\rho > 1$. Given n, set k so that $n_k \le n < n_{k+1}$. There exists n large enough that $n_k/n > 1/\rho$ and $n_{k+1}/n < \rho$, and hence since $Y_t \ge 0$ for all t, that

$$\frac{1}{\rho}\bar{Y}_{n_k} < \frac{n_k}{n}\bar{Y}_{n_k} \le \bar{Y}_n \le \frac{n_{k+1}}{n}\bar{Y}_{n_k+1} < \rho\bar{Y}_{n_k+1}.$$

Since $\rho > 1$ is arbitrary, to prove the theorem it is therefore sufficient to show $\bar{Y}_{n_k} \to \mathrm{E}(X_1)$ a.s. as $k \to \infty$.

To do this, apply the Chebyshev inequality (**9.16** for $p = 2$), the pairwise independence assumption, and the Jensen inequality **9.19** to get

$$P(|\bar{Y}_{n_k} - \mathrm{E}(\bar{Y}_{n_k})| > \varepsilon) \le \frac{\mathrm{Var}(\bar{Y}_{n_k})}{\varepsilon^2} \le \frac{1}{\varepsilon^2 n_k^2}\sum_{t=1}^{n_k}\mathrm{E}(Y_t^2)$$

and hence, summing over k,

$$\begin{aligned}\sum_{k=1}^{\infty} P(|\bar{Y}_{n_k} - \mathrm{E}(\bar{Y}_{n_k})| > \varepsilon) &\le \frac{1}{\varepsilon^2}\sum_{k=1}^{\infty}\frac{1}{n_k^2}\sum_{t=1}^{n_k}\mathrm{E}(Y_t^2)\\ &= \frac{1}{\varepsilon^2}\sum_{t=1}^{\infty}\mathrm{E}(Y_t^2)\sum_{k:n_k\ge t}\frac{1}{n_k^2}\\ &\le \frac{C_1}{\varepsilon^2}\sum_{t=1}^{\infty}\frac{\mathrm{E}(Y_t^2)}{t^2}. \end{aligned} \tag{21.29}$$

The equality here is found by reordering the terms of the double sum, and the final inequality follows because

$$\sum_{k:n_k\ge t}\frac{1}{n_k^2} \le \sum_{k:\rho^{k-1}\ge t}\frac{1}{\rho^{2(k-1)}} \le C_1\frac{1}{t^2}$$

for some $C_1 < \infty$, making use of the fact that the tail of a convergent geometric series has the same order of magnitude as its leading term. Moreover,

$$\begin{aligned}\sum_{t=1}^{\infty}\frac{\mathrm{E}(Y_t^2)}{t^2} &= \sum_{t=1}^{\infty}\frac{\mathrm{E}(X_1^2 1_{\{X_1\le t\}})}{t^2}\\ &= \sum_{t=1}^{\infty}\sum_{j=1}^{t}\frac{\mathrm{E}(X_1^2 1_{\{j-1\le X_1\le j\}})}{t^2}\\ &= \sum_{j=1}^{\infty}\mathrm{E}\big(X_1^2 1_{\{j-1\le X_1\le j\}}\big)\sum_{t=j}^{\infty}t^{-2}\\ &\le C_2\sum_{j=1}^{\infty} j\mathrm{E}\big(X_1 1_{\{j-1\le X_1\le j\}}\big)j^{-1} = C_2\mathrm{E}(X_1) \end{aligned} \tag{21.30}$$

for $C_2 < \infty$. Putting (21.29) and (21.30) together shows that

$$\sum_{k=1}^{\infty} P(|\bar{Y}_{n_k} - \mathrm{E}(\bar{Y}_{n_k})| > \varepsilon) < \infty$$

and the Borel–Cantelli lemma (**19.2**(i)) then gives $\bar{Y}_{n_k} - \mathrm{E}(\bar{Y}_{n_k}) \to 0$ a.s. to complete the proof. ∎

It might be thought that this result was of limited application, but Etemadi's singular insight was to notice that if the assumptions of the theorem hold for an arbitrary sequence, they also hold for its positive and negative parts.

21.10 Theorem If $\{X_t\}$ is any identically distributed and pairwise-independent sequence with $\mathrm{E}|X_1| < \infty$, then $n^{-1}\sum_{t=1}^{n} X_t \to \mathrm{E}(X_1)$ a.s.

Proof Let $X_t^+ = \max\{X_t, 0\} \geq 0$ and $X_t^- = X_t - X_t^+ \geq 0$. These random variables have identical distributions for all t, are pairwise independent, and on the assumptions of the theorem have finite means. $\mathrm{E}(X_1^+) - \mathrm{E}(X_1^-) = \mathrm{E}(X_1)$ whereas $\mathrm{E}(X_1^+) + \mathrm{E}(X_1^-) = \mathrm{E}|X_1|$. Applying Theorem **21.9** gives $\bar{X}_n^+ = n^{-1}\sum_{t=1}^{n} X_t^+ \to \mathrm{E}(X_1^+)$ a.s. and $\bar{X}_n^- = n^{-1}\sum_{t=1}^{n} X_t^- \to \mathrm{E}(X_1^-)$ a.s. and hence $\bar{X}_n = \bar{X}_n^+ - \bar{X}_n^- \to \mathrm{E}(X_1)$ a.s. ∎

Circumstances in which pairwise independence might arise in the absence of total independence do not spring readily to mind and this feature of the conditions is more incidental than critical. What chiefly makes the Etemadi theorem of interest is that it delivers the SLLN without reliance on the standard tricks of a maximal inequality teamed with Kronecker's lemma. While it is a sufficiency result, given the i.i.d. framework there is really no way that the conditions might be relaxed further while yielding a finite limit. However, one evident extension is the following.

21.11 Corollary For a sequence $\{X_t, t \geq 1\}$ let X_t^- and X_t^- be defined as in the proof of **21.10**. If $\mathrm{E}(X_1^+) = \infty$ and $\mathrm{E}(X_1^-) < \infty$ then $\bar{X}_n \to \infty$ a.s., while if $\mathrm{E}(X_1^+) < \infty$ and $\mathrm{E}(X_1^-) = \infty$ then $\bar{X}_n \to -\infty$ a.s. □

In the event that both positive and negative parts of the sequence have infinite means, no convergence can be demonstrated and the limit is undefined.

A further reason why the independence case is of interest is the following very elegant result due to Lévy. This shows that when dealing with partial sums of independent sequences, the concepts of weak and strong convergence coincide.

21.12 Theorem When $\{X_t\}$ is an independent sequence and $S_n = \sum_{t=1}^{n} X_t$, $S_n \to_{\text{pr}} S$ iff $S_n \to_{\text{a.s.}} S$.

Proof Sufficiency is by **19.5**. It is the necessity that is unique to the particular case cited. Let $S_{mn} = \sum_{t=m+1}^{n} X_t$ and for some $\varepsilon > 0$ consider the various ways in which the event $\{|S_{mn}| > \varepsilon\}$ can occur. In particular, consider the disjoint collection

$$\Big\{\max_{m\le j\le k-1} |S_{mj}| \le 2\varepsilon, |S_{mk}| > 2\varepsilon\Big\},\ k = m+1, \dots, n.$$

For each $k > m$ this is the event that the sum from m onwards exceeds 2ε absolutely for the *first* time at time k and thus

$$\bigcup_{k=m+1}^{n} \Big\{\max_{m\le j\le k-1} |S_{mj} \le 2\varepsilon, |S_{mk}| > 2\varepsilon\Big\} = \Big\{\max_{m\le j\le n} |S_{mj}| > 2\varepsilon\Big\} \tag{21.31}$$

where the sets of the union are disjoint. It is also the case that

$$\bigcup_{k=m+1}^{n} \Big\{\max_{m\le j\le k-1} |S_{mj}| \le 2\varepsilon, |S_{mk}| > 2\varepsilon\Big\} \cap \{|S_{kn}| \le \varepsilon\} \subseteq \{|S_{mn}| > \varepsilon\} \tag{21.32}$$

where the inclusion is ensured by imposing the extra condition for each k. The events in this union are still disjoint and by the assumption of an independent sequence the intersections are of independent pairs of events. On applying (21.31), it follows from (21.32) that

$$\begin{aligned}
&P\Big(\max_{m\le j\le n} |S_{mj}| > 2\varepsilon\Big) \min_{m<k\le n} P(|S_{kn}| \le \varepsilon)\\
&\qquad \le \Bigg(\sum_{k=m+1}^{n} P\Big(\max_{m\le j\le k-1} |S_{mj}| \le 2\varepsilon, |S_{mk}| > 2\varepsilon\Big)\Bigg) \min_{m<k\le n} P(|S_{kn}| \le \varepsilon)\\
&\qquad \le \sum_{k=m+1}^{n} P\Big(\max_{m\le j\le k-1} |S_{mj}| \le 2\varepsilon, |S_{mk}| > 2\varepsilon\Big) P(|S_{kn}| \le \varepsilon)\\
&\qquad \le P(|S_{mn}| > \varepsilon).
\end{aligned} \tag{21.33}$$

If $S_n \to_{\text{pr}} S$, there exists by definition $m \ge 1$ such that

$$P(|S_{mn}| > \varepsilon) < \varepsilon \tag{21.34}$$

for all $n > m$. According to (21.34), for this m and any n the second factor on the minorant side of (21.33) is at least as great as $1 - \varepsilon$, so for $0 < \varepsilon < 1$,

$$P\Big(\max_{m\le j\le n}|S_{mj}| > 2\varepsilon\Big) < \frac{\varepsilon}{1-\varepsilon}. \tag{21.35}$$

Letting $n \to \infty$ and then $m \to \infty$, since ε is arbitrary the theorem now follows by **19.3**. ∎

This equivalence of weak and strong results is one of the chief benefits stemming from the independence assumption. Since the three-series theorem is equivalent to a weak law according to **21.12**, the result gives necessary conditions for convergence in probability. As far as sufficiency results go, however, practically nothing is lost by passing from the independent to the martingale case, and since showing convergence is usually of disproportionately greater importance than showing nonconvergence, the absence of necessary conditions may be regarded as a small price to pay.

However, a feature of the three-series theorem that is common to all the strong law results of this chapter is that it is not an array result. Being based on the convergence lemma, all these proofs depend on teaming a convergent stochastic sequence with an increasing constant sequence, such that their ratio goes to zero. Although the results can be written down in array form, there is no counterpart of the weak law of **20.11** more general than its specialization in **20.12**.

21.3 Martingale Strong Laws

Martingale limit results are remarkably powerful. So long as a sequence is a martingale difference, no further restrictions on its dependence are required and the moment assumptions called for are scarcely tougher than those imposed in the independent case. Moreover, while the m.d. property is stronger than the uncorrelatedness assumed in §20.2, the distinction is very largely technical. Given the nature of econometric time-series models, one can usually assert that a sequence is uncorrelated *because* it is a m.d.; basically a sequence that is not forecastable in mean one step ahead. The case when it is uncorrelated with its own past values but not with some other function of lagged information could arise, but would be in the nature of a special case.

The results in this section and the next one are drawn or adapted chiefly from Stout ([176]) and Hall and Heyde ([88]), although many of the ideas go back to Doob ([60]). The first one is a standard SLLN for L_2-bounded sequences.

21.13 Theorem Let $\{X_t, \mathcal{F}_t\}_0^\infty$ be a m.d. sequence with variance sequence $\{\sigma_t^2\}$ and $\{a_t\}$ a positive constant sequence with $a_t \uparrow \infty$. If

$$\sum_{t=1}^{\infty} \sigma_t^2/a_t^2 < \infty \tag{21.36}$$

then $S_n/a_n \to_{\text{a.s.}} 0$. □

There are (at least) two ways to prove this result. The first is to use the martingale convergence theorem (**16.11**) directly and the second is to combine the maximal inequality of **16.20** with the convergence lemma **21.2**. In effect, the second line of argument provides an alternative proof of martingale convergence for the square-integrable case, providing an interesting comparison of techniques.

First proof of 21.13 Define $T_n = \sum_{t=1}^{n} X_t/a_t$, so that $\{T_n, \mathcal{F}_n\}$ is a square-integrable martingale. The norm inequality and orthogonality of $\{X_t\}$ give

$$\sup_n \mathrm{E}|T_n| \le \sup_n \mathrm{E}(T_n^2)^{1/2} = \left(\sum_{t=1}^{\infty} \sigma_t^2/a_t^2\right)^{1/2} < \infty \tag{21.37}$$

leading directly to the conclusion $T_n \to T$ a.s., by **16.11**. Then apply the Kronecker lemma to the sequences $\{T_n(\omega)\}$ for $\omega \in \Omega$ to show that $S_n/a_n \to_{\text{a.s.}} 0$. ∎

Second proof of 21.13 For $m \ge 0$, $\{T_n - T_m, \mathcal{F}_n\}$ is a martingale with

$$\mathrm{E}(T_n - T_m)^2 = \sum_{t=m+1}^{n} \sigma_t^2/a_t^2. \tag{21.38}$$

Apply **16.20** for $p = 2$ and then **21.2** with $c_t^2 = \sigma_t^2/a_t^2$. Finally apply the Kronecker lemma as before. ∎

Compare this result with **20.7**. If $\mathrm{Var}(X_t) = \sigma_t^2 \le B < \infty$ (say), then setting $a_t = t$,

$$\sum_{t=1}^{\infty} \sigma_t^2/t^2 \le B \sum_{t=1}^{\infty} 1/t^2 \le 1.65B < \infty$$

and the condition of the theorem is satisfied, hence $\bar{X}_n = S_n/n \to_{\text{a.s.}} 0$, the same conclusion as before. But the conditions on the variances are now a lot weaker and in effect the weak law of **20.4** has been converted into a strong law at the small cost of substituting the m.d. assumption for orthogonality. As an example of the general formulation, suppose the sequence satisfies

$$\sum_{t=1}^{\infty} \mathrm{E}(X_t^2)/t^4 < \infty. \tag{21.39}$$

While $\bar{X}_n$ does not necessarily converge to zero, putting $a_t = t^2$ shows that $n^{-2}\sum_{t=1}^n X_t = \bar{X}_n/n$ does so converge.

The limitation of **21.13** is that it calls for square integrability. The next step is to use **21.4** to extend it to the class of cases that satisfy

$$\sum_{t=1}^{\infty} \mathrm{E}|X_t|^p/a_t^p < \infty \tag{21.40}$$

for $1 \le p \le 2$ and some $\{a_t\} \uparrow \infty$. It is important to appreciate that (21.40) for $p < 2$ is not a *weaker* condition than for $p = 2$ and the latter does not imply the former. For contrast, consider $p = 1$. The Kronecker lemma applied to (21.40) then implies that

$$a_n^{-1}\sum_{t=1}^{n} \mathrm{E}|X_t| \to 0. \tag{21.41}$$

For $a_n \sim n$, such a sequence has got to be zero or very close to it most of the time. In fact, there is a trivially direct proof of convergence. Note that (21.41) implies

$$\begin{aligned} \mathrm{E}\Big(\lim_{n\to\infty} a_n^{-1}|S_n|\Big) &\le \mathrm{E}\Big(\lim_{n\to\infty} a_n^{-1}\sum_{t=1}^{n}|X_t|\Big) \\ &= \lim_{n\to\infty} a_n^{-1}\sum_{t=1}^{n}\mathrm{E}|X_t| = 0 \end{aligned} \tag{21.42}$$

where the first equality follows by the dominated convergence theorem **4.16** with bounding function $g = \sum_{t=1}^{\infty}|X_t|/a_t$, where $\mathrm{E}(g)$ is finite by assumption. For any random variable X, $\mathrm{E}|X| = 0$ if and only if $X = 0$ a.s. Nothing more is needed to show that S_n/a_n converges, regardless of other conditions.

Thus, having latitude in the value of p is really a matter of being able to trade off the existence of absolute moments against the rate of damping necessary to make them summable. Interesting cases in which (21.40) holds for $p < 2$ may arise only rarely, but since this extension is available at small extra cost in complexity it makes sense to take advantage of it.

21.14 Theorem If $\{X_t, \mathcal{F}_t\}_1^{\infty}$ is a m.d. sequence satisfying (21.40) for $1 \le p \le 2$, $S_n/a_n \to_{\text{a.s.}} 0$.

Proof Let $Y_t = 1_{\{|X_t| \le a_t\}}X_t$ and note that $\{X_t\}$ and $\{Y_t\}$ are equivalent under (21.40) by **21.4**. Y_t is also $\mathcal{F}_t$-measurable and hence the centred sequence $\{Z_t, \mathcal{F}_t\}$, where $Z_t = Y_t - \mathrm{E}(Y_t|\mathcal{F}_{t-1})$, is a m.d. Now,

$$\begin{aligned}\mathrm{E}(Z_t^2) &= \mathrm{E}\big(\mathrm{E}(Z_t^2|\mathcal{F}_{t-1})\big)\\ &= \mathrm{E}\big(\mathrm{E}(Y_t^2|\mathcal{F}_{t-1}) - \mathrm{E}(Y_t|\mathcal{F}_{t-1})^2\big)\\ &= \mathrm{E}(Y_t^2) - \mathrm{E}\big(\mathrm{E}(Y_t|\mathcal{F}_{t-1})^2\big). \qquad (21.43)\end{aligned}$$

According to **21.4** with $r = 2$, (21.40) implies that $\sum_{t=1}^{\infty}\mathrm{E}(Y_t^2)/a_t^2 < \infty$ and so since $\mathrm{E}(Z_t^2) \le \mathrm{E}(Y_t^2)$ by (21.43),

$$\sum_{t=1}^{\infty}\mathrm{E}(Z_t^2)/a_t^2 < \infty. \qquad (21.44)$$

By **21.13** this is sufficient for $\sum_{t=1}^{n} Z_t/a_t \to_{\text{a.s.}} S_1$, where S_1 is some random variable. But

$$\sum_{t=1}^{n} Z_t/a_t = \sum_{t=1}^{n} Y_t/a_t - \sum_{t=1}^{n}\mathrm{E}(Y_t|\mathcal{F}_{t-1})/a_t. \qquad (21.45)$$

By **16.18**(i), (21.40) is equivalent to

$$\sum_{t=1}^{\infty}\mathrm{E}(|X_t|^p|\mathcal{F}_{t-1})/a_t^p < \infty, \text{ a.s.} \qquad (21.46)$$

According to **21.5**, (21.46) implies that $\sum_{t=1}^{\infty}|\mathrm{E}(Y_t|\mathcal{F}_{t-1})|/a_t < \infty$, a.s. Absolute convergence of a series implies convergence by **2.14**, so $\sum_{t=1}^{n}\mathrm{E}(Y_t|\mathcal{F}_{t-1})/a_t \to_{\text{a.s.}} S_2$. Hence, $\sum_{t=1}^{n} Y_t/a_t \to_{\text{a.s.}} S_1 + S_2$ and so $a_n^{-1}\sum_{t=1}^{n} Y_t \to_{\text{a.s.}} 0$ by the Kronecker lemma. It follows by **21.3** and the equivalence of X_t and Y_t implied by (21.40) that $S_n/a_n \to_{\text{a.s.}} 0$. ∎

Notice that in this proof there are no short cuts through the martingale convergence theorem. While $\sum_{t=1}^{n} X_t/a_t$ is a martingale, the problem is to establish that it is uniformly L_1-bounded from the information in (21.40). Going by way of a result for $p = 2$ to exploit orthogonality is where the truncation arguments come in handy.

21.4 Conditional Variances and Random Weighting

A feature of martingale theory exploited in the last theorem is the possibility of relating convergence to the behaviour of the sequences of one-step-ahead conditional moments; the next step is to extend this principle to the conditional variances $\mathrm{E}(X_t^2|\mathcal{F}_{t-1})$. The elegant results of this section contain those such as **21.13** and **21.14**.

The conditional variance of a centred coordinate is the variance of the innovation, that is, of $X_t - \mathrm{E}(X_t|\mathcal{F}_{t-1})$, and in some circumstances it may be more natural to place restrictions on the behaviour of the innovations than on the original sequence. In regression models, for example, the innovations may correspond to the regression disturbances. Moreover, the fact that the conditional moments are $\mathcal{F}_{t-1}$-measurable random variables, so that any constraint upon them is probabilistic, permits a generalization of the concept of convergence, following the results of §16.4; confidence in the summability of the weighted conditional variances translates into a probability that the sequence converges, in the manner of the following theorem. A nice refinement is that the constant weight sequence $\{a_t\}$ can be replaced by a sequence of $\mathcal{F}_{t-1}$-measurable random weights.

21.15 Theorem Let $\{X_t, \mathcal{F}_t\}_1^\infty$ be a m.d. sequence, let $\{W_t\}$ be a non-decreasing sequence of positive, $\mathcal{F}_{t-1}$-measurable r.v.s, and let $S_n = \sum_{t=1}^n X_t$. Then

$$P\left(\left\{\sum_{t=1}^{\infty} \mathrm{E}(X_t^2|\mathcal{F}_{t-1})/W_t^2 < \infty\right\} \cap \{W_t \uparrow \infty\} - \{S_n/W_n \to 0\}\right) = 0. \quad \square \qquad (21.47)$$

The last statement is perhaps a little opaque, but roughly translated it says that the probability of convergence, that is to say of the event $\{S_n/W_n \to 0\}$, is not less than that of the intersection of the two other events in (21.47). In particular, when one probability is 1, so is the other.

Proof of 21.15 If $\{X_t\}$ is a m.d. sequence so is $\{X_t/W_t\}$, since W_t is $\mathcal{F}_{t-1}$ measurable and $T_n = \sum_{t=1}^n X_t/W_t$ is a martingale. For $\omega \in \Omega$, if $T_n(\omega) \to T(\omega)$ and $W_n(\omega) \uparrow \infty$ then $S_n(\omega)/W_n(\omega) \to 0$ by Kronecker's lemma. Applying **16.16** completes the proof. ∎

See how this result contains **21.13**, corresponding to the case of a fixed, divergent weight sequence and a.s. summability.

As before, the next step is to weaken the summability conditions from conditional variances to p^{th} absolute moments for $1 \le p \le 2$. However, to exploit **21.5** outside the almost sure case requires a modification to the equivalent sequences argument **21.3**, which is as follows.

21.16 Theorem If $\{X_t\}$ and $\{Y_t\}$ are sequences of $\mathcal{F}_t$-measurable r.v.s,

$$P\left(\left\{\sum_{t=1}^{n} P(X_t \neq Y_t|\mathcal{F}_{t-1}) < \infty\right\} \triangle \left\{\sum_{t=1}^{n}(X_t - Y_t) \text{ converges}\right\}\right) = 0. \qquad (21.48)$$

Proof Let $E_t = \{X_t \neq Y_t\} \in \mathcal{F}_t$, so that $P(X_t \neq Y_t|\mathcal{F}_{t-1}) = \mathrm{E}(1_{E_t}|\mathcal{F}_{t-1})$. According to **16.18**(ii),

$$P\left(\left\{\omega : \sum_{t=1}^{\infty} \mathrm{E}(1_{E_t}|\mathcal{F}_{t-1})(\omega) < \infty\right\} \triangle \left\{\omega : \sum_{t=1}^{\infty} 1_{E_t}(\omega) < \infty\right\}\right) = 0. \tag{21.49}$$

But $\sum_{t=1}^{\infty} 1_{E_t}(\omega) < \infty$ means that the number of coordinates for which $X_t(\omega) \neq Y_t(\omega)$ is finite and hence $\sum_{t=1}^{\infty}\big(X_t(\omega) - Y_t(\omega)\big) < \infty$. (21.49) therefore implies (21.48). ∎

With this result it is possible to prove the following extension of **21.14**.

21.17 Theorem For $1 \leq p \leq 2$, let $E_1 = \{\sum_{t=1}^{\infty} \mathrm{E}(|X_t|^p|\mathcal{F}_{t-1})/W_t^p < \infty\}$ and $E_2 = \{W_t \uparrow \infty\}$. Under the conditions of **21.15**,

$$P\big((E_1 \cap E_2) - \{S_n/W_n \to 0\}\big) = 0. \tag{21.50}$$

Proof The basic line of argument follows closely that of **21.14**. As before, let $Y_t = 1_{\{|X_t| \leq a_t\}}X_t$, so that $Z_t = Y_t - \mathrm{E}(Y_t|\mathcal{F}_{t-1})$ is a m.d. and

$$\begin{aligned} \mathrm{E}(Z_t^2|\mathcal{F}_{t-1}) &= \mathrm{E}(Y_t^2|\mathcal{F}_{t-1}) - \big(\mathrm{E}(Y_t|\mathcal{F}_{t-1})\big)^2 \\ &\leq \mathrm{E}(Y_t^2|\mathcal{F}_{t-1}) \text{ a.s.} \end{aligned} \tag{21.51}$$

Let $D = \{\sum_{t=1}^{\infty} \mathrm{E}(Z_t^2|\mathcal{F}_{t-1})/W_t^2 < \infty\}$. Applying **21.5** (the case of D_3) and the last inequality gives

$$P(E_1 - D) = 0. \tag{21.52}$$

Also let $C = \{\sum_{t=1}^{n} Z_t/W_t \to S_1\}$ where S_1 is some a.s. finite random variable. It follows by **16.16** and the fact that $E_1 - C \subseteq (E_1 - D) \cup (D - C)$ that

$$P(E_1 - C) = 0. \tag{21.53}$$

Next, a second application of **21.5** (the case of D_2) gives

$$P\left(E_1 - \left\{\sum_{t=1}^{\infty} |\mathrm{E}(Y_t|\mathcal{F}_{t-1})|/W_t < \infty\right\}\right) = 0 \tag{21.54}$$

which is equivalent (by **2.14**) to

$$P\left(E_1 - \left\{\sum_{t=1}^{n} \mathrm{E}(Y_t|\mathcal{F}_{t-1})/W_t \to S_2\right\}\right) = 0 \tag{21.55}$$

where S_2 is another a.s. finite r.v. Now a third application of **21.5** (the case of D_1) together with **21.16** gives

$$P\left(E_1 - \left\{\sum_{t=1}^{n} X_t - \sum_{t=1}^{n} Y_t \to S_3\right\}\right) = 0 \tag{21.56}$$

for some a.s. finite r.v. S_3. Finally, (21.53), (21.55), (21.56), the definition of Z_t, the Kronecker lemma, and some more set algebra yield, as required,

$$\begin{aligned} 0 &= P\left(\left(E_1 - \left\{\sum_{t=1}^{n} Y_t/W_t \to S_1 + S_2\right\}\right) \cap E_2\right) \\ &= P\left((E_1 \cap E_2) - \left\{W_n^{-1} \sum_{t=1}^{n} Y_t \to 0\right\}\right) \\ &= P\big((E_1 \cap E_2) - \{S_n/W_n \to 0\}\big). \quad \blacksquare \end{aligned} \tag{21.57}$$

21.5 Strong Laws for Mixingales

The martingale difference assumption is specialized and the last results are not sufficient to support a general treatment of dependent processes, although they are the central prop. The key to extending them, as in the weak law case, is the mixingale concept. This section contrasts alternative approaches to proving mixingale strong convergence.

The first of these applies a straightforward generalization of the methods introduced by McLeish ([125]); see also Hansen ([93], [94]) for related results. There are two versions of the theorem to choose from, with a milder constraint on the dependence being available in return for the existence of second moments.

21.18 Theorem Let the sequence $\{X_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ be an L_p-mixingale with respect to constants $\{c_t\}$, for either

(i) $p = 2$, with mixingale size $-\frac{1}{2}$, or

(ii) $1 < p < 2$, with mixingale size -1.

Let $\{a_t\}_1^{\infty}$ be a positive sequence with $a_t \uparrow \infty$. If

$$\sum_{t=1}^{\infty} c_t^p / a_t^p < \infty \tag{21.58}$$

then $a_n^{-1} \sum_{t=1}^{n} X_t \to_{\text{a.s.}} 0$.

Proof Let $S_n = \sum_{t=1}^{n} X_t/a_t$. By **17.11** in the case of (i) and **17.12** in case (ii),

$$\mathrm{E}\Big(\max_{1\le j\le n}|S_j|^p\Big) \le K\sum_{t=1}^{n} c_t^p/a_t^p \tag{21.59}$$

where K is a finite constant. By relabelling coordinates this inequality can be expressed in the form

$$\mathrm{E}\Big(\max_{m\le j\le n}|S_j - S_m|^p\Big) \le K\sum_{t=m+1}^{n} c_t^p/a_t^p \tag{21.60}$$

for any choice of m and n. Moreover,

$$\begin{aligned} P\Big(\max_{m<j\le n}|S_j - S_m| > \varepsilon\Big) &= P\Big(\max_{m<j\le n}|S_j - S_m|^p > \varepsilon^p\Big) \\ &\le \frac{1}{\varepsilon^p}\mathrm{E}\Big(\max_{m<j\le n}|S_j - S_m|^p\Big) \end{aligned} \tag{21.61}$$

by the Markov inequality. The result follows from inequalities (21.60) and (21.61), using the convergence lemma (**21.2**) and then the Kronecker lemma **2.35**. ∎

This is a fairly straightforward argument, but the resulting conditions are not as sharp as they could be. More elaborate arguments based around the key idea of the telescoping sum of martingale differences can refine it in different directions. As in **21.18** these results specify properties for the sequences $\{c_t\}_{-\infty}^{\infty}$ and $\{\zeta_m\}_0^{\infty}$ of size $-\varphi_0$, the norm index p defining the mixingale $\{X_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ and also the normalizing sequence $\{a_n\}_1^{\infty}$ with $a_n \uparrow \infty$. All exploit in different ways a trade-off between the mixingale size and the heterogeneity of the marginal distributions of coordinates, represented by scale factors c_t. This approach is in contrast with the McLeish method of specifying separate summability conditions for the moments and mixingale numbers, as detailed in §17.3.

To clarify the arguments it is often useful initially to keep in mind the case $c_t = 1$ and $a_t = t$, corresponding to the mean $\bar{X}_n$ of a stationary sequence. If $c_t/a_t = O(t^{\xi})$ then $p\xi < -1$ is the required condition to satisfy (21.58). Suppose $\xi < -\frac{1}{2}$ and $p = 2$. In this case, replacing a_t with t implies that $\bar{X}_n = O(n^{-1/2})$ a.s. according to Kronecker's lemma, which is the 'square root rule' of convergence. With $p < 2$ and also in the presence of a drift in the p^{th} moment such that (say) $c_t = O(t^{\beta})$ for $\beta > 0$, $\bar{X}_n \to_{\text{a.s.}} 0$ may still happen but at a slower rate than $n^{-1/2}$. However, if (21.58) fails with $a_t = t$ then $\bar{X}_n$ fails to converge.

The following three results, dealing respectively with the cases $p = 2$, $1 < p \le 2$, and $p = 1$, adopt more elaborate variants of (21.58) allowing the summability condition to depend on the mixingale index, hence trading off dependence for heterogeneity. Theorems **21.19** and **21.21** are adapted from [45] and **21.23** from [49]. Throughout, the notation $\mathrm{E}_s X_t$ is used as the abbreviated form of $\mathrm{E}(X_t|\mathcal{F}_s)$.

21.19 Theorem Let $\{X_t,\mathcal{F}_t\}_{-\infty}^{\infty}$ be an L_2-mixingale with respect to constants $\{c_t\}_{-\infty}^{\infty}$ and mixingale numbers $\{\zeta_j\}_0^{\infty}$ and let $\{a_n\}_1^{\infty}$ be an increasing constant sequence such that $a_n/\sqrt{n} \uparrow \infty$. If

$$\sum_{t=1}^{\infty}\Bigl(1+\sum_{j=0}^{t} L_j\zeta_j^2\Bigr)\frac{c_t^2}{a_t^2} < \infty \tag{21.62}$$

where $L_j = \log(j+1)(\log\log(j+2))^{1+\delta}$ for $\delta > 0$, then $a_n^{-1}\sum_{t=1}^{n} X_t \to_{\text{a.s.}} 0$.

Proof Define S_{n1}, S_{n2}, and S_{n3} by the decomposition

$$\begin{aligned}\sum_{t=1}^{n} X_t &= S_{n1} + S_{n2} + S_{n3}\\ &= \sum_{t=1}^{n}\mathrm{E}_{-2}X_t + \sum_{t=1}^{n}\sum_{j=-t-1}^{t+1}(\mathrm{E}_{t+j}X_t - \mathrm{E}_{t+j-1}X_t) + \sum_{t=1}^{n}(X_t - \mathrm{E}_{2t+1}X_t).\end{aligned}$$

First considering S_{n2}, also define

$$T_{n2} = \sum_{t=1}^{n}\Bigl(\sum_{j=-t-1}^{t+1}\frac{\mathrm{E}_{t+j}X_t - \mathrm{E}_{t+j-1}X_t}{a_t}\Bigr).$$

It can be verified that for $0 \le m < n$,

$$T_{n2} - T_{m2} = \sum_{j=-n-1}^{n+1} Y_{nmj}$$

where

$$Y_{nmj} = \sum_{t=\max\{m+1,|j|-1\}}^{n}\frac{\mathrm{E}_{t+j}X_t - \mathrm{E}_{t+j-1}X_t}{a_t}$$

is a sum of martingale differences. Given a summable sequence $\{m_k\}_0^{\infty}$ to be chosen, Lemma **17.9** and then by reasoning closely paralleling **17.10**, but with sums of $2t+3$ terms replacing the infinite sums in (17.35),

$$\mathrm{E}\Bigl(\max_{m<k\le n}(T_{k2} - T_{m2})^2\Bigr) \le \sum_{t=m+1}^{n} C_t^2\frac{c_t^2}{a_t^2} \tag{21.63}$$

where

$$C_t^2 = K\left(\frac{\zeta_0^2 + \zeta_1^2}{m_0} + 2\sum_{j=1}^{t+1}\left(\frac{1}{m_j} - \frac{1}{m_{j-1}}\right)\zeta_j^2 - 2\frac{\zeta_{t+2}^2}{m_{t+2}}\right) \tag{21.64}$$

for $K < \infty$. Setting $m_0 = 1$ and $m_j = \left(jL_j\right)^{-1}$ for $j \geq 1$, note that this sequence is summable according to **2.18** and

$$1/m_j - 1/m_{j-1} = L_j + O\big((\log\log(j+2))^\delta\big)$$

where $\delta > 0$ is arbitrary. If (21.62) holds, it follows from **21.1**, **21.2**, (21.61), and (21.63) that $T_{n2} < \infty$ a.s. and hence by Kronecker's lemma that $S_{n2}/a_n \to 0$ a.s.

Next note that, as a case of the elementary inequality $\left(\sum x_t\right)^2 \leq n\sum x_t^2$,

$$\frac{S_{n1}^2}{a_n^2} \leq \frac{n}{a_n^2}\sum_{t=1}^{n}(\mathrm{E}_{-2}X_t)^2.$$

Since $a_n^2/n \to \infty$ by assumption, it follows by Kronecker's lemma that $S_{n1}/a_n \to 0$ a.s. if

$$\sum_{t=1}^{\infty}\frac{t(\mathrm{E}_{-2}X_t)^2}{a_t^2} < \infty \text{ a.s.} \tag{21.65}$$

and this is verified by showing the expectation to be finite. Applying Theorem **10.28** gives the inequality

$$\begin{aligned}\mathrm{E}\left(\sum_{t=1}^{\infty}\frac{t(\mathrm{E}_{-2}X_t)^2}{a_t^2}\right) &\leq \sum_{t=1}^{\infty}\frac{1}{a_t^2}\sum_{j=1}^{t}\mathrm{E}(\mathrm{E}_{t-j-2}X_t)^2 \\ &\leq \sum_{t=1}^{\infty}\frac{c_t^2}{a_t^2}\sum_{j=1}^{t}\zeta_{j+2}^2 < \infty \end{aligned} \tag{21.66}$$

where the finiteness is by (21.62) with L_j replaced by 1. The demonstration of $S_{3n}/a_n \to 0$ a.s. is effectively identical. ∎

It is of interest to compare the following corollary with **21.18**(i):

21.20 Corollary With mixingale size $-\varphi_0$, a sufficient condition for (21.62) is $c_t/a_t = O(t^\xi)$ where $\xi < \min\{-\frac{1}{2}, \varphi_0 - 1\}$.

Proof Note that (21.62) is equivalent to

$$\sum_{t=1}^{\infty}\frac{c_t^2}{a_t^2} + \sum_{j=1}^{\infty}L_j\zeta_j^2\sum_{t=j}^{\infty}\frac{c_t^2}{a_t^2} < \infty.$$

This condition holds if both $2\xi < -1$ and $2\xi + 1 - 2\varphi_0 < -1$. ∎

This is a substantial weakening of the memory condition over **21.18**. Any $\varphi_0 > 0$ allows convergence with $\xi > -1$, implying a positive convergence rate for the sample mean of a stationary process, while $\varphi_0 \geq \frac{1}{2}$ implies square-root convergence.

The next result applies the same type of argument while weakening the moment condition.

21.21 Theorem Let $\{X_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ be an L_p-mixingale, $1 < p \leq 2$, with respect to constants $\{c_t\}_{-\infty}^{\infty}$ and mixingale numbers $\{\zeta_j\}_0^{\infty}$ and let $\{a_n\}_1^{\infty}$ be an increasing positive constant sequence such that $a_n/n^{1-1/p} \uparrow \infty$ and $\sum_{j=0}^{\infty} m_j < \infty$ where

$$m_j = \zeta_j \Bigg(\sum_{t=\max\{1,[j^{p/(p-1)}]\}}^{\infty} \frac{c_t^p}{a_t^p} \Bigg)^{1/p}. \tag{21.67}$$

Then, $a_n^{-1} \sum_{t=1}^{n} X_t \to_{\text{a.s.}} 0$.

Proof For an integer sequence $\{b_t\}_1^{\infty}$ to be specified, define

$$Y_{tj} = \begin{cases} (E_{t+j}X_t - E_{t+j-1}X_t)/a_t, & -b_t \leq j \leq b_t \\ 0, & \text{otherwise} \end{cases}$$

so that the sequences $\{Y_{tj}, \mathcal{F}_{t+j}\}_{j=-\infty}^{\infty}$ are martingale differences for $t = 1, \ldots, n$, and also,

$$\|Y_{tj}\|_p \leq 2\zeta_{|j|} 1_{\{|j| \leq b_t\}} c_t / a_t. \tag{21.68}$$

Hence define S_{n1}, S_{n2}, and S_{n3} by the decomposition

$$\begin{aligned} \sum_{t=1}^{n} X_t &= S_{n1} + S_{n2} + S_{n3} \\ &= \sum_{t=1}^{n} \mathrm{E}_{-b_t} X_t + \sum_{t=1}^{n} \sum_{j=1-b_t}^{b_t} (\mathrm{E}_{t+j} X_t - \mathrm{E}_{t+j-1} X_t) + \sum_{t=1}^{n} (X_t - \mathrm{E}_{b_t} X_t). \end{aligned} \tag{21.69}$$

Define $T_{n2} = \sum_{t=1}^{n} \sum_{j=1-b_t}^{b_t} Y_{tj}$. For $\{m_j\}_{-\infty}^{\infty}$ as specified in (21.67) with $m_{-j} = m_j$, **17.9** and (21.68) give for $K < \infty$,

$$\mathrm{E}\Big(\max_{m<k\le n}|T_{k2}-T_{m2}|^p\Big)\le K\sum_{j=-\infty}^{\infty}m_j^{1-p}\mathrm{E}\Big|\sum_{t=m+1}^{n}Y_{tj}\Big|^p$$
$$\le K\sum_{t=m+1}^{n}\frac{c_t^p}{a_t^p}\sum_{j=0}^{\infty}m_j^{1-p}\zeta_j^p1_{\{j\le b_t\}}.$$

Choose $b_t=\min\{j:[j^{p/(p-1)}]\ge t\}$ and note that

$$\{t:|j|\le b_t\}=\{t:t\ge[j^{p/(p-1)}]\}\tag{21.70}$$

and hence by (21.67),

$$\sum_{t=1}^{n}\frac{c_t^p}{a_t^p}\sum_{j=0}^{\infty}m_j^{1-p}\zeta_j^p1_{\{j\le b_t\}}=\sum_{j=0}^{\infty}m_j^{1-p}\zeta_j^p\sum_{t=\max\{1,[j^{p/(p-1)}]\}}^{n}\frac{c_t^p}{a_t^p}$$
$$=\sum_{j=0}^{\infty}m_j<\infty.$$

It follows by **21.2** that T_{n2} converges and $S_{n2}/a_n\to 0$ a.s.

Considering the terms S_{n1} and S_{n3} in (21.69), these are treated identically so consider S_{n1}. Define $S'_{n1}(k)=\sum_{t=1}^{\min\{n,k\}}\mathrm{E}_{t-b_t}X_t$ and $S''_{n1}(k)=S_{n1}-S'_{n1}(k)$, so that $S''_{n1}(k)=0$ for $n\le k$. If k is fixed then since $a_n\to\infty$,

$$\lim_{m\to\infty}P\Big(\sup_{n\ge m}|S'_{n1}(k)|/a_n>\frac{\varepsilon}{2}\Big)=0\tag{21.71}$$

for any $\varepsilon>0$. Next, considering $S''_{n1}(k)$, apply the Markov inequality to get

$$\lim_{m\to\infty}P\Big(\sup_{n\ge m}|S''_{n1}(k)|/a_n>\frac{\varepsilon}{2}\Big)\le\lim_{m\to\infty}2\frac{\mathrm{E}\sup_{n\ge m}|S''_{n1}(k)|}{a_n\varepsilon}.\tag{21.72}$$

Next, for $p>1$ and any $n>k$,

$$\Big|\frac{S''_{n1}(k)}{a_n}\Big|=\frac{1}{a_n}\Big|\sum_{t=k+1}^{n}\mathrm{E}_{t-b_t}X_t\Big|$$
$$\le\sum_{j=1}^{b_n}\frac{1}{a_n}\Big|\sum_{t=k+1}^{n}\frac{1_{\{j\le b_t\}}}{b_t}\mathrm{E}_{t-b_t}X_t\Big|$$

$$\leq \sum_{j=1}^{b_n} \biggl(\frac{(n-k)^{p-1}}{a_n^p} \sum_{t=k+1}^{n} \frac{1_{\{j \leq b_t\}}}{b_t^p} |\mathrm{E}_{t-b_t} X_t|^p \biggr)^{1/p}$$

$$\leq \sum_{j=1}^{b_n} \biggl(\sum_{t=k+1}^{n} \frac{t^{p-1} 1_{\{j \leq b_t\}}}{a_t^p b_t^p} |\mathrm{E}_{t-b_t} X_t|^p \biggr)^{1/p}.$$

Here, the first inequality is by the triangle inequality after inserting the terms $\sum_{j=1}^{b_n} b_t^{-1} 1_{\{j \leq b_t\}} = 1$ for each t, the second is by convexity of the power transform (see **2.21**), while the last one uses the fact that $n^{p-1} a_n^{-p}$ is decreasing in n by assumption. Therefore,

$$\frac{\mathrm{E}|S_{n1}''(k)|}{a_n} \leq \sum_{j=1}^{b_n} \biggl(\sum_{t=k+1}^{n} \frac{t^{p-1} 1_{\{j \leq b_t\}}}{a_t^p b_t^p} \mathrm{E}|\mathrm{E}_{t-j} X_t|^p \biggr)^{1/p}$$

$$\leq \sum_{j=1}^{\infty} \zeta_j \biggl(\sum_{t=\max\{k+1, [j^{p/(p-1)}]\}}^{\infty} \frac{c_t^p}{a_t^p} \biggr)^{1/p} = \delta_k$$

defining δ_k, where the first inequality is by the reverse Jensen (**9.19**) under the concavity of $(\cdot)^{1/p}$ and also by the fact that $\mathrm{E}|\mathrm{E}_{t-b_t} X_t|^p \leq E|\mathrm{E}_{t-j} X_t|^p$ for $j \leq b_t$ which is also the condition $t \geq [j^{p/(p-1)}]$ by (21.70). The second one, likewise, used the fact that $b_t^p \geq t^{p-1}$ by definition of b_t.

The points to notice here are that $\delta_k < \infty$ by (21.67) and that $\delta_k \to 0$ as $k \to \infty$. Moreover, δ_k does not depend on n and hence bounds the majorant side of (21.72), which can be made as close to 0 as desired by taking k large enough. Summing (21.72) and (21.71) shows that

$$\lim_{m \to \infty} P\Bigl(\sup_{n \geq m} |S_{n1}|/a_n > \varepsilon \Bigr) = 0$$

which is sufficient for $S_{n1}/a_n \to_{\text{a.s.}} 0$. The same argument can be applied to S_{n3}, completing the proof. ∎

21.22 Corollary If the mixingale size is $-\varphi_0$, a sufficient condition for the sequence in (21.67) to be summable is $c_t/a_t = O(t^{\xi})$ where

$$\xi < \min\{-1/p, (1-1/p)\varphi_0 - 1\}.$$

Proof It is necessary that the sequence c_t^p/a_t^p is summable, which sets $\xi p < -1$. Subject to this, Theorem **2.17**(iii) sets the summability condition as

$$\frac{1+\xi p}{p-1} - \varphi_0 < -1. \quad \blacksquare$$

This shows the utility of Theorem **21.19**, since for $p = 2$, **21.22** gives the condition $\xi < \frac{1}{2}\varphi_0 - 1$ to compare with $\xi < \varphi_0 - 1$ from **21.20**. The existence of second moments is always helpful to trade off with memory restrictions. With $p < 2$, suppose $\varphi_0 \geq 1$ and $c_t = 1$. If for example $p = 3/2$ in this situation then $a_t = t^{2/3}$ suffices for convergence and so $\bar{X}_n = O(n^{-1/3})$ a.s. The general rule here is $\bar{X}_n = O(n^{1/p-1})$. With either p arbitrarily close to 1 or φ_0 arbitrarily small, the convergence rate for $\bar{X}_n$ is arbitrarily slow. Other such calculations can be performed as required.

The third strong law is for L_1-mixingales, with the novelty that the main summability condition exploits Azuma's inequality and so involves exponentials instead of powers.

21.23 Theorem Let $\{X_t, \mathcal{F}_t\}$ be an L_r-bounded, L_1-mixingale for $r \geq 1$ with respect to constants $\{c_t\}_{-\infty}^{\infty}$ and mixingale numbers $\{\zeta_j\}_0^{\infty}$. If $\{a_n\}_1^{\infty}$ is a positive increasing constant sequence, $\{B_t\}$ a positive constant sequence, and $\{M_t\}$ a positive increasing integer sequence such that

$$\sum_{n=1}^{\infty} M_n \exp\left\{-\varepsilon^2 a_n^2 \left(32 M_n^2 \sum_{t=1}^{n} B_t^2\right)^{-1}\right\} < \infty \tag{21.73}$$

for any $\varepsilon > 0$ and also

$$\sum_{t=1}^{\infty} B_t^{1-r} \mathrm{E}|X_t|^r / a_t < \infty \tag{21.74}$$

$$\sum_{t=1}^{\infty} \zeta_{M_t} c_t / a_t < \infty \tag{21.75}$$

then $a_n^{-1} \sum_{t=1}^{n} X_t \to_{\text{a.s.}} 0$. □

Here, $\{B_t\}$ and $\{M_t\}$ are chosen freely to satisfy the conditions given $\{a_t\}$ and $\{c_t\}$, which suggests a considerable amount of flexibility in application. The sequence $\{B_t\}$ is used to define a truncation of $\{X_t\}$, the role that was played by $\{a_t\}$ in **21.14**. The most interesting of the conditions is (21.75), which explicitly trades off the rate of decrease of the mixingale coefficients with that of the sequence $\{c_t/a_t\}$.

Proof of 21.23 Writing 1_t^B for $1_{\{|X_t|\le B_t\}}$, start by noting that

$$\mathrm{E}_{t+j}X_t = \mathrm{E}_{t+j}1_t^B X_t + \mathrm{E}_{t+j}(1-1_t^B)X_t.$$

and so write the identity

$$\begin{aligned} X_t =& (\mathrm{E}_{t+M_t-1}1_t^B X_t - \mathrm{E}_{t-M_t}1_t^B X_t) \\ &+ \big(\mathrm{E}_{t+M_t-1}(1-1_t^B)X_t - \mathrm{E}_{t-M_t}(1-1_t^B)X_t\big) \\ &+ (X_t - \mathrm{E}_{t+M_t-1}X_t) + \mathrm{E}_{t-M_t}X_t. \end{aligned} \tag{21.76}$$

By the usual telescoping sum argument,

$$\mathrm{E}_{t+M_t-1}1_t^B X_t - \mathrm{E}_{t-M_t}1_t^B X_t = \sum_{j=1-M_t}^{M_t-1} Z_{jt} \tag{21.77}$$

where $Z_{jt} = \mathrm{E}_{t+j}1_t^B X_t - \mathrm{E}_{t+j-1}1_t^B X_t$ and $\{Z_{jt}, \mathcal{F}_{t+j}\}$ is a m.d. sequence. Note that $|Z_{jt}| \le 2B_t$ a.s. by a double application of **10.14**(ii). Summing yields

$$\begin{aligned} \sum_{t=1}^n X_t =& \sum_{t=1}^n \sum_{j=1-M_t}^{M_t-1} Z_{jt} + \sum_{t=1}^n \mathrm{E}_{t+M_t-1}(1-1_t^B)X_t \\ &- \sum_{t=1}^n \mathrm{E}_{t-M_t}(1-1_t^B)X_t + \sum_{t=1}^n (X_t - \mathrm{E}_{t+M_t-1}X_t) \\ &+ \sum_{t=1}^n \mathrm{E}_{t-M_t}X_t \\ =& S_{1n} + S_{2n} - S_{3n} + S_{4n} + S_{5n}. \end{aligned} \tag{21.78}$$

The object is to show that $S_{kn}/a_n \to_{\text{a.s.}} 0$ for $k = 1,\ldots,5$. Starting with S_{1n}, the main task is to reorganize the double sum. Let $q_j = 1$ for $-M_1 < j < M_1$ and for $s = 2,\ldots,n$, let $q_j = s$ for $-M_s < j \le -M_{s-1}$ and $M_{s-1} \le j < M_s$. Then it can be verified by inspection that

$$\begin{aligned} \sum_{t=1}^n \sum_{j=1-M_t}^{M_t-1} Z_{jt} &= \sum_{j=1-M_n}^{M_n-1} \sum_{t=q_j}^n Z_{jt} \\ &= \sum_{j=1-M_n}^{M_n-1} \sum_{t=1}^n Z_{jt} - \Bigg(\sum_{j=1-M_n}^{-M_2} + \sum_{j=M_2}^{M_n-1} \Bigg) \sum_{t=1}^{q_j-1} Z_{jt}. \end{aligned} \tag{21.79}$$

Note that for arbitrary numbers $x_1,\dots,x_k$, $\{|\sum_{i=1}^k x_i| > \varepsilon\} \subseteq \bigcup_{i=1}^k \{|x_i| > \varepsilon/k\}$. Hence by subadditivity and Azuma's inequality **16.30**,

$$\begin{aligned}P(|S_{1n}| > a_n\varepsilon) &\le \sum_{j=1-M_n}^{M_n-1} P\biggl(\Bigl|\sum_{t=1}^{n} Z_{jt}\Bigr| > \frac{a_n\varepsilon}{4M_n}\biggr)\\ &\qquad + \biggl(\sum_{j=1-M_n}^{-M_2} + \sum_{j=M_2}^{M_n-1}\biggr) P\biggl(\Bigl|\sum_{t=1}^{q_j-1} Z_{jt}\Bigr| > \frac{a_n\varepsilon}{4M_n}\biggr)\\ &\le 2M_n \exp\Bigl\{-\varepsilon^2 a_n^2\bigl(32M_n^2\sum_{t=1}^{n} B_t^2\bigr)^{-1}\Bigr\}\\ &\qquad + \biggl(\sum_{j=1-M_n}^{-M_2} + \sum_{j=M_2}^{M_n-1}\biggr)\exp\biggl\{-\varepsilon^2 a_n^2\biggl(32M_n^2\sum_{t=1}^{q_j-1} B_t^2\biggr)^{-1}\biggr\}\\ &\le 4M_n \exp\Bigl\{-\varepsilon^2 a_n^2\bigl(32M_n^2\sum_{t=1}^{n} B_t^2\bigr)^{-1}\Bigr\}. \end{aligned} \tag{21.80}$$

Under (21.73), these probabilities are summable over n and so $S_{1n}/a_n \to_{\text{a.s.}} 0$ by the first Borel–Cantelli lemma.

To deal with S_{2n}, S_{3n}, S_{4n}, and S_{5n} let $\{Y_t\}_1^n$ be any integrable sequence. Define S_n as the unweighted sum of the sequence and

$$S_n^a = \sum_{t=1}^{n} Y_t/a_t. \tag{21.81}$$

By the Markov inequality,

$$\begin{aligned}P\Bigl(\max_{m<j\le n}|S_j^a - S_m^a| > \varepsilon\Bigr) &\le \frac{1}{\varepsilon}\mathrm{E}\Bigl(\max_{m<j\le n}|S_j^a - S_m^a|\Bigr)\\ &\le \frac{1}{\varepsilon}\sum_{t=m+1}^{n} \mathrm{E}|Y_t|/a_t. \end{aligned} \tag{21.82}$$

If

$$\sum_{t=1}^{\infty} \mathrm{E}|Y_t|/a_t < \infty \tag{21.83}$$

then $S_n^a \to_{\text{a.s.}} S^a$ by an application of **21.2** and hence $S_n/a_n \to_{\text{a.s.}} 0$ by Kronecker's lemma. For the case of S_{2n} put $Y_t = \mathrm{E}_{t+M_t-1}X_1(1 - 1_t^B)$ and note that

$$\sum_{t=1}^{n} \mathrm{E}|Y_t|/a_t \leq \sum_{t=1}^{n} \mathrm{E}|(1-1_t^B)X_t|/a_t \leq \sum_{t=1}^{n} B_t^{1-r}\mathrm{E}|X_t|^r/a_t \tag{21.84}$$

using the fact that $|X_t(1-1_t^B)/B_t| \leq |X_t(1-1_t^B)/B_t|^r$. S_{3n} is dealt with in exactly the same way. For S_{4n} and S_{5n}, put successively $Y_t = X_t - \mathrm{E}_{t+M_t-1}X_t$ and $Y_t = \mathrm{E}_{t-M_t}X_t$, and by the mixingale assumption,

$$\sum_{t=1}^{n} \mathrm{E}|Y_t|/a_t \leq \sum_{t=1}^{n} (c_t/a_t)\zeta_{M_r}. \tag{21.85}$$

The proof is completed by noting that the majorant terms of (21.84) and (21.85) are bounded in the limit by assumption. ∎

The conditions of **21.23** are difficult to apply and interpret, but restricting them very slightly allows derivation of a parametric condition to compare with those of **21.18**, **21.20**, and **21.22**.

21.24 Corollary Let the conditions of **21.23** hold where the L_1-mixingale size is $-\varphi_0$, $c_t \leq \|X_t\|$, and $\{B_t\}$, $\{a_t\}$, and $\{M_t\}$ are regularly varying sequences. If

$$\sum_{t=1}^{n} (c_t/a_t)^{\psi} < \infty$$

where

$$\psi = \frac{2r\varphi + 2(r-1)}{(1+r)\varphi + 2(r-1)} \tag{21.86}$$

for $\varphi > \varphi_0$, then $a_n^{-1}\sum_{t=1}^{n} X_t \to_{\text{a.s.}} 0$.

Proof In what follows, let $o_n = (\log n)^{1+\delta}$ where $\delta > 0$ denotes a generic constant to be set as small as required. Consider conditions sufficient for (21.73) to hold. In view of **2.18** a sufficient condition for summability of a series of the form $\sum_n U_1(n)\exp\{-\eta U_2(n)\}$ with $\eta > 0$, where U_1 and U_2 are regularly varying functions, is

$$n o_n U_1(n)\exp\{-\eta U_2(n)\} \to 0. \tag{21.87}$$

Suppose $U_i(n) = n^{\rho_i}L_i(n)$ for $i = 1$ and 2 where $\rho_i \geq 0$ and the $L_i(n)$ are slowly varying. Taking the logarithm of (21.87) and noting that the terms $\log o_n$ and $\log L_1(n)$ can be neglected, an equivalent condition to (21.87) is

$$(1+\rho_1)\log n - \eta n^{\rho_2}L_2(n) \to -\infty. \tag{21.88}$$

For any ρ_1 and any $\eta > 0$, sufficient conditions are that either $\rho_2 > 0$ or $\rho_2 = 0$ and $L_2(n) \geq o_n$. Recalling that $\{B_t\}$ is monotone, a sufficient condition for (21.73) is therefore

$$nM_n^2 B_n^2 / a_n^2 \ll 1/o_n. \tag{21.89}$$

Similarly, conditions (21.74) and (21.75) are satisfied if, respectively,

$$B_t^{1-r} \mathrm{E}|X_t|^r / a_t \leq B_t^{1-r} c_t^r / a_t \ll (to_t)^{-1} \tag{21.90}$$

and

$$M_t^{-\varphi} c_t / a_t \ll (to_t)^{-1},\ \varphi > \varphi_0. \tag{21.91}$$

To identify bounding cases of B_t and M_t that satisfy these conditions, replace the order-of-magnitude inequality signs in (21.90) and (21.91) by equalities, leaving the required scaling constants implicit. Solving for M_t and B_t in this way, substituting into (21.89) with $t = n$ and simplifying yields the condition

$$\left(\frac{c_n}{a_n}\right)^{\psi} \ll \frac{1}{no_n} \tag{21.92}$$

where ψ is defined by (21.86). This condition is sufficient for (21.73), (21.74), and (21.75) to hold. ∎

It will be noted that $1 \leq \psi \leq 2$, the upper bound being attained as both $\varphi \to \infty$ and $r \to \infty$ while the lower bound applies when $\varphi = 0$ and also when $r = 1$, in which case the expression does not depend on φ. In the case of the mean of a stationary series where $c_t = 1$ and $a_t = t$, the sufficient condition for convergence is $\psi > 1$ for which $r > 1$ and $\varphi > 0$ are jointly sufficient. If $r = 1$ then the case with $c_t = 1$ and $a_t = to_t$ meets the condition for any mixingale size; in other words, the sample mean does not converge but it is slowly varying in the limit almost surely.

For comparability with Corollaries **21.20** and **21.22**, write the sufficient condition from **21.24** in the form $c_t/a_t = O(t^{\xi})$ where $\xi < -1/\psi$. While r is the order of existing moment here, not the mixingale order, an L_p-mixingale for $p > 1$ is also an L_1-mixingale so the conditions may be compared on that basis. Since ξ is negative, the 'best' result with respect to a given pair (r, φ) is the one that delivers the largest value of ξ, meaning that convergence is shown compatible with the slowest rate of decline of c_t/a_t. In the case $r = 2$, **21.20** gives the best result for all values of φ. However, for $r < 2$, whether **21.22** or **21.24** dominates depends on the magnitude of φ. Figure 21.1 plots the formulae from the indicated corollaries for the case $r = 1.1$, showing values of ξ against φ. In this particular case, **21.24**

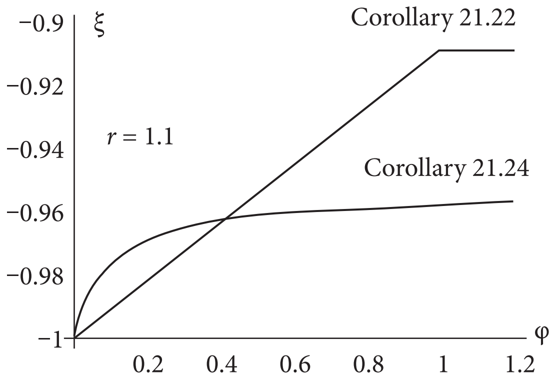

Figure 21.1

delivers the best result for $\varphi < 0.4$ and **21.22** is best otherwise. Both results have a contribution to make in cases where the variance may not exist.

Here is a final, more specialized result. The moving average of martingale differences with summable coefficients is a case of particular interest since it unifies the two approaches to the strong law.

21.25 Theorem Let $X_t = \sum_{j=-\infty}^{\infty} \theta_j U_{t-j}$ where $\{U_t\}$ is a uniformly L_p-bounded m.d. sequence with $1 < p \leq 2$ and $\sum_{j=-\infty}^{\infty} |\theta_j| < \infty$. Then $\bar{X}_n \to_{\text{a.s.}} 0$.

Proof Letting $Y_t = X_t/t$ and $\mathcal{F}_t = \sigma(U_s, s \leq t)$, the sequence $\{Y_t, \mathcal{F}_t\}_1^{\infty}$ is an L_p-mixingale with $c_t \ll 1/t$ and arbitrary size. For this case the maximal inequality from Theorem **17.13**(i) has the form

$$\mathrm{E}\Big(\max_{1\leq j\leq n} |S_j|^p\Big) \leq K \sum_{t=1}^{n} t^{-p}$$

where

$$K = C_p\Big(\frac{p}{p-1}\Big)^p \Big(|\theta_0| + \sum_{k=1}^{\infty}(|\theta_k| + |\theta_{-k}|)\Big)^p \sup_s \mathrm{E}|U_s|^p.$$

Applying the argument used to prove **21.18** leads directly to the result. Alternatively, Theorem **21.21** and Corollary **21.22** can be applied, setting $a_t = t$ and $\varphi_0 = 0$. ∎

This result establishes the convergence of the sample mean, but at an arbitrarily slow rate. The convergence rate can be increased by having the moving average weights go to zero faster. The remarks following **21.22** apply here. When $\varphi_0 \geq 1$, meaning that $\sum_{k=m}^{\infty}(|\theta_k| + |\theta_{-k}|) = O(m^{-1-\delta})$ for $\delta > 0$, then $\xi < -1/p$ so that $\bar{X}_n = O(n^{1/p-1})$ a.s.

21.6 NED and Mixing Processes

Results for mixingales are useful in most cases because results for NED and mixing processes can be inferred from them. Let $\{X_t\}_{-\infty}^{\infty}$ be a zero-mean, L_r-bounded stochastic sequence where X_t is L_p-NED for $1 \leq p \leq 2$ of size $-b$ with respect to scale constants $\{d_t\}_{-\infty}^{\infty}$ on a possibly vector-valued sequence $\{\boldsymbol{V}_t\}_{-\infty}^{\infty}$ which is either α-mixing of size $-a_\alpha$ and $r > p$, or ϕ-mixing of size $-a_\phi$ and $r \geq p$. There is no loss of generality in assuming zero means since X_t can always be regarded as a mean deviation. Letting $\mathcal{F}_s^t = \sigma(\boldsymbol{V}_s, \dots, \boldsymbol{V}_t)$ it follows by **18.6** that $\{X_t, \mathcal{F}_{-\infty}^t\}$ is an L_p-mixingale with respect to constants $c_t \ll \max\{d_t, \|X_t\|_r\}$ of size $-\varphi_0$ where in the α-mixing case

$$\varphi_0 = \min\{b, a_\alpha(1/p - 1/r)\} \tag{21.93}$$

and in the ϕ-mixing case

$$\varphi_0 = \min\{b, a_\phi(1 - 1/r)\}. \tag{21.94}$$

Theorems **21.19**, **21.21**, and **21.23**, or more conveniently the size-based corollaries **21.20**, **21.22**, and **21.24**, can be applied directly, specifying the constraints on the mixingale size directly from equations (21.93) and (21.94). These results do not require separate statements.

However, where a mixing component exists these results have the drawback of requiring L_r-boundedness for $r > p$ and in the case of α-mixing the mixingale size depends critically on the magnitude of the difference according to (21.93). Therefore, to have a result where no moments of order greater than p need exist is useful and the natural approach to follow is a truncation argument.

21.26 Theorem $\{X_t\}_{-\infty}^{\infty}$ is a zero-mean stochastic sequence where X_t is L_p-NED for $1 \leq p \leq 2$ of size $-1/p$ with respect to scale constants $d_t \leq \|X_t\|_p$ on $\{\boldsymbol{V}_t\}_{-\infty}^{\infty}$ which is either

(i) α-mixing with $a_\alpha = -r/(r-2)$ for $r > 2$, or
(ii) ϕ-mixing with $a_\phi = -r/2(r-1)$ for $r \geq 2$.

If

$$\sum_{t=1}^{\infty} \|X_t/a_t\|_p^{2p/r} < \infty \tag{21.95}$$

for a positive increasing sequence $\{a_n\}$, then $a_n^{-1}\sum_{t=1}^n X_t \to_{\text{a.s.}} 0$. □

The L_p-NED size has to be linked to p under these assumptions, embodying a trade-off between memory and moment summability. Take care to note that the role of r is not to set a bound on existing moments of X_t in this case, but is simply to specify the α-mixing or ϕ-mixing size.

Proof of 21.26 The strategy is to show that there is a sequence that is equivalent to $\{X_t/a_t\}$ and satisfying the conditions of **21.20**. As in **21.6** let

$$Y_t = X_t 1_t^a / a_t \pm (1 - 1_t^a) \tag{21.96}$$

where $1_t^a = 1_{\{|X_t| \le a_t\}}$ and '±' denotes '+' if $X_t > 0$, '−' otherwise. Note that $\{X_t/a_t\}$ is L_p-NED with constants d_t/a_t and Y_t is a continuous function of X_t/a_t with $|Y_t| \le 1$ a.s. Applying **18.13** shows that Y_t is L_2-NED on $\{\boldsymbol{V}_t\}$ of size $-\frac{1}{2}$ with respect to constants $2^{1-p/2}(d_t/a_t)^{p/2}$. Since $\|Y_t\|_r < \infty$ for any finite r, it further follows by **18.6** and the assumptions that if $\mathcal{F}_t = \sigma(\boldsymbol{V}_{t-s}, s \ge 0)$, $\{Y_t - \mathrm{E}(Y_t), \mathcal{F}_t\}_0^\infty$ is an L_2-mixingale of size $-\frac{1}{2}$ with constants

$$c_t \ll \max\{(d_t/a_t)^{p/2}, \|Y_t\|_r\}. \tag{21.97}$$

Setting $q = r$ in the second inequality of (21.24) yields $\|Y_t\|_r \le 2\|X_t/a_t\|_p^{p/r}$ for any $r \ge p$. If the sequence $\{c_t^2\}$ is summable then necessarily $\|X_t/a_t\|_p < 1$ except for at most a finite number of t and for these cases, since $r \ge 2$, $\|X_t/a_t\|_p^{p/2} \le \|X_t/a_t\|_p^{p/r}$. Also, $(d_t/a_t)^{p/2} \le \|X_t/a_t\|_p^{p/2}$ by assumption. Hence it follows by (21.97) that $c_t^2 \ll \|X_t/a_t\|_p^{2p/r}$. At this point the arguments in the proof of Theorem **21.18**(i) can be adapted with Y_t replacing X_t/a_t. Given condition (21.95), the arguments leading to (21.60) and (21.61) show that $\sum_{t=1}^n \big(Y_t - \mathrm{E}(Y_t)\big) \to_{\text{a.s.}} S_1$ where S_1 is some random variable.

According to **21.7** condition (21.95) is also sufficient for $\sum_{t=1}^\infty |\mathrm{E}(Y_t)| < \infty$. The series $\sum_{t=1}^n \mathrm{E}(Y_t)$ therefore converges to a finite limit by **2.14**, say $\sum_{t=1}^\infty \mathrm{E}(Y_t) = C_1$ and

$$\sum_{t=1}^{n} Y_t \overset{\text{a.s.}}{\to} S_1 + C_1. \tag{21.98}$$

Inequality (21.22) further implies that Y_t and X_t/a_t are equivalent sequences and hence

$$\sum_{t=1}^{n}\left(X_t/a_t - Y_t\right) \overset{\text{a.s.}}{\to} S_2 \tag{21.99}$$

where S_2 is another random variable, by **21.3**. Putting (21.98) and (21.99) together,

$$\sum_{t=1}^{n} X_t/a_t \overset{\text{a.s.}}{\to} S_1 + S_2 + C_1 = S_3. \tag{21.100}$$

It follows by Kronecker's lemma that $a_n^{-1}\sum_{t=1}^{n} X_t \to_{\text{a.s.}} 0$, the required conclusion. ∎

To compare this result with the size-based corollaries of §21.5, note the implication of (21.95) that $\|X_t/a_t\|_p = O(t^{\xi})$ for $\xi < -r/2p$. As already noted, for the sample mean of a stationary sequence to converge a.s. requires that this ξ be larger (i.e. smaller absolutely) than -1. Consider two contrasting examples. First, suppose the process $\boldsymbol{V}_t$ is α-mixing fast enough that $\varphi_0 = b = 1/p$ according to (21.93), with r arbitrarily close to 2. Taking the example $p = 3/2$, $\xi < -2/3$ is sufficient for convergence with $\bar{X}_n = O(n^{-1/3})$ a.s. Compare this to the α-mixing case with $a_\alpha = -2$ so that $r = 4$. If the fourth moment exists and assuming $b > 5/6$, the formula in (21.93) gives the mixingale size of X_t as $-\varphi_0 = -5/6$. Since $(1 - 1/p)\varphi_0 - 1 = -13/18 < -2/3$ the formula in **21.22** gives $\xi = -13/18$ and the condition for convergence is satisfied. However, $-r/2p = -4/3$ and condition (21.95) fails. This shows the cost of foregoing the assumption of a finite fourth moment.

In these results, four features summarize the relevant characteristics of the stochastic process: the order of existing moments, the summability characteristics of the moments, and the sizes of the mixing and near-epoch dependence numbers. These characteristics can, to an extent, be traded off one against another to explore the envelope of sufficient conditions. It may be clear from the variety of results presented that the exploration of this space is by no means complete.

22
Uniform Stochastic Convergence

22.1 Stochastic Functions on a Parameter Space

The setting for this chapter is the class of functions

$f : \Omega \times \Theta \mapsto \bar{\mathbb{R}}$

where $(\Omega, \mathcal{F}, \mu)$ is a measure space and (Θ, ρ) is a metric space. The notation $f(\omega, \theta)$ represents the real value assumed by f at the point (ω, θ). At fixed θ this is a random variable, whereas $f(\omega, \cdot)$, alternatively written just $f(\omega)$, is *not* a random variable but a random element of a space of functions.

Econometric analysis is frequently concerned with this type of object. Log-likelihoods, sums of squares, and other criterion functions for the estimation of econometric models and also the first and second derivatives of these criterion functions are all the subject of important convergence theorems on which proofs of consistency and the derivation of limiting distributions are based. Except in a restricted class of linear models, all of these are typically functions both of the model parameters and of random data.

To deal with convergence on a function space, it is necessary to have a criterion by which to judge when two functions are close to one another. This chapter examines the questions posed by stochastic convergence (almost sure or in probability) when the relevant space of functions is endowed with the uniform metric. A class of set functions that are therefore going to be central to the discussion have the form $f^* : \Omega \mapsto \bar{\mathbb{R}}$ where

$$f^*(\omega) = \sup_{\theta \in \Theta} f(\omega, \theta). \tag{22.1}$$

For example, if g and h are two stochastic functions whose uniform proximity is at issue, the object of interest is the supremum with respect to Θ of

$$f(\omega, \theta) = |g(\omega, \theta) - h(\omega, \theta)|.$$

An important technical problem arises here that ought to be confronted at the outset. No results have so far been given that would justify treating f^* as a random variable when (Θ, ρ) may be an arbitrary metric space. It is possible to write

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0022

$$\{\omega : f^*(\omega) > x\} = \bigcup_{\theta\in\Theta}\{\omega : f(\theta,\omega) > x\} \tag{22.2}$$

and the results of **3.33** show that $\{\omega : f^*(\omega) > x\} \in \mathcal{F}$ when $\{\omega : f(\theta,\omega) > x\} \in \mathcal{F}$ for each θ where Θ is a *countable* set. But typically, Θ is a subset of $(\mathbb{R}^k, d_E)$ or something of the kind and is uncountable.

This is one of a class of measurability problems having ramifications far beyond the uniform convergence issue and to handle this question properly requires a mathematical apparatus going beyond what is covered in Chapter 3. To deal with it in depth goes beyond the scope of the chapter and no proofs will be offered in this instance, merely an outline of the main features of the theory required for its solution. The essential step is to recognize that the set on the left-hand side of (22.2) can be expressed as a projection.

Let $\mathcal{B}_\Theta$ denote the Borel field of subsets of Θ, that is, the smallest σ-field containing the sets of Θ that are open with respect to ρ. Then let $(\Omega\times\Theta, \mathcal{F}\otimes\mathcal{B}_\Theta)$ denote the product space endowed with the product σ-field (the σ-field generated from the measurable rectangles of $\mathcal{F}$ and $\mathcal{B}_\Theta$) and suppose that $f(\cdot,\cdot)$ is $\mathcal{F}\otimes\mathcal{B}_\Theta/\bar{\mathcal{B}}$-measurable. Observe that if

$$A_x = \{(\omega,\theta) : f(\omega,\theta) > x\} \in \mathcal{F}\otimes\mathcal{B}_\Theta, \tag{22.3}$$

the projection of A_x into Ω is

$$\begin{aligned} E_x &= \{\omega : f(\omega,\theta) > x, \theta \in \Theta\} \\ &= \{\omega : f^*(\omega) > x\}. \end{aligned} \tag{22.4}$$

In view of **3.31**, measurability of f^* is equivalent to the condition that $E_x \in \mathcal{F}$ for rational x. However, projections are *not* as a rule measurable transformations. As the remarks following **3.25** point out, the projection of a Borel set in a product space need not be a Borel set of the factor space. It is true that in topological spaces projections are continuous and hence measurable under the product topology (see §6.5), but the abstract space $(\Omega,\mathcal{F})$ lacks topological structure and this reasoning does not apply.

However, under certain conditions it can be shown that $E_x \in \mathcal{F}^P$ where $(\Omega, \mathcal{F}^P, \bar{P})$ is the completion of the probability space. The key notion is that of an *analytic* set. A standard reference on this topic is Dellacherie and Meyer ([55]); see also Dudley ([63] ch. 13) and Stinchcombe and White ([174]). The latter authors provide the following definition. Letting $(\Omega,\mathcal{F})$ be a measurable space, a set $E \subset \Omega$ is called $\mathcal{F}$-analytic if there exists a compact metric space (Θ,ρ) such that E is the projection onto Ω of a set $A \in \mathcal{F}\otimes\mathcal{B}_\Theta$. The collection of $\mathcal{F}$-analytic sets is written

$\mathcal{A}(\mathcal{F})$. Also, a function $f : \Omega \mapsto \bar{\mathbb{R}}$ is called $\mathcal{F}$-analytic if $\{\omega : f(\omega) \leq x\} \in \mathcal{A}(\mathcal{F})$ for each $x \in \bar{\mathbb{R}}$.

Note that $\mathcal{F} \subseteq \mathcal{A}(\mathcal{F})$, in view of the fact that every $E \in \mathcal{F}$ is the projection of $E \times \Theta \in \mathcal{F} \otimes \mathcal{B}_\Theta$. A measurable set (or function) is therefore also analytic. $\mathcal{A}(\mathcal{F})$ is not in general a σ-field. It can be shown to be closed under countable unions and countable intersections, but the complement of an analytic set is not necessarily analytic. The conditions under which an image under projection is known to be analytic are somewhat weaker than the definition might suggest and it will actually suffice to let $(\Theta, \mathcal{B}_\Theta)$ be a *Souslin space*, that is, a space that is measurably isomorphic to an analytic subset of a compact metric space. A sufficient condition, whose proof can be extracted from the results in Stinchcombe and White [174], is the following:

22.1 Theorem Let $(\Omega, \mathcal{F})$ be a measurable space and $(\Theta, \mathcal{B}_\Theta)$ a Souslin space. If $B \in \mathcal{A}(\mathcal{F} \otimes \mathcal{B}_\Theta)$, the projection of B onto Ω is in $\mathcal{A}(\mathcal{F})$. □

In other words, if B is itself an analytic set resulting from a projection onto $\Omega \times \Theta$, the further projection of B onto Ω preserves that property.

Recall from **3.14** and associated discussion that the completion $(\Omega, \mathcal{F}^\mu, \bar{\mu})$ of a probability space $(\Omega, \mathcal{F}, \mu)$ augments $\mathcal{F}$ with all subsets of sets of measure zero under μ and defines $\bar{\mu}(E) = \mu(E)$ for each $E \in \mathcal{F}$ and also $\bar{\mu}(F) = \mu(E)$ for each set $F \in \mathcal{F}^\mu$ such that $E \in \mathcal{F}$ and $\mu(E \triangle F) = 0$. Given the measurable space $(\Omega, \mathcal{F})$ the *universally measurable* sets are the collection $\mathcal{F}^U = \bigcap_\mu \mathcal{F}^\mu$, where the intersection is taken over the completions with respect to all p.m.s μ defined on the space. The key conclusion, from Dellacherie and Meyer [55] (III.33(a)), is the following.

22.2 Theorem For a measurable space $(\Omega, \mathcal{F})$,

$$\mathcal{A}(\mathcal{F}) \subseteq \mathcal{F}^U. \quad \square \tag{22.5}$$

It follows that sets of $\mathcal{A}(\mathcal{F})$ are measurable under the completion of $(\Omega, \mathcal{F}, \mu)$ for any choice of μ. If E is analytic there exist $A, B \in \mathcal{F}$ such that $A \subseteq E \subseteq B$ and $\mu(A) = \mu(B)$ and in this sense analytic sets are 'nearly' measurable. All the standard probabilistic arguments, and in particular the values of integrals, are unaffected by this technical non-measurability and it can be ignored with impunity. It is legitimate to treat $f^*(\omega)$ as a random variable, *provided* the conditions on Θ are observed and $f(\cdot, \cdot)$ is (near-) $\mathcal{F} \otimes \mathcal{B}_\Theta / \bar{\mathcal{B}}$-measurable.

An analytic subset of a compact space need not be compact but must be totally bounded. It is convenient not to have to insist on compactness of the parameter space, since the latter is often required to be open thanks to strict inequality constraints (think of variances, stable roots of polynomials, and the like). In the convergence results below, Θ will in any case have to be totally bounded for completely different reasons such as to ensure equicontinuity and to ensure that the stochastic functions have bounded moments. When a stochastic criterion function is being optimized with respect to Θ, the optimum is usually required to lie almost surely in the interior of a compact set. Hence, total boundedness is not an extra restriction in practice.

The measurability condition on $f(\omega,\theta)$ might be verifiable using an argument from simple functions. It is certainly necessary by **4.24** that the cross-section functions $f(\cdot,\theta) : \Omega \mapsto \bar{\mathbb{R}}$ and $f(\omega,\cdot) : \Theta \mapsto \bar{\mathbb{R}}$ be, respectively, $\mathcal{F}/\bar{\mathcal{B}}$-measurable for each $\theta \in \Theta$ and $\mathcal{B}_\Theta/\bar{\mathcal{B}}$-measurable for each $\omega \in \Omega$. For a finite partition $\{\Theta_1, \dots, \Theta_m\}$ of Θ by $\mathcal{B}_\Theta$-sets, consider the function

$$f_{(m)}(\omega,\theta) = f(\omega,\theta_j), \theta \in \Theta_j, j = 1, \dots, m \tag{22.6}$$

where θ_j is a point of Θ_j. If $E_j^x = \{\omega : f(\omega,\theta_j) \le x\} \in \mathcal{F}$ for each j, then

$$A_x = \{(\omega,\theta) : f_{(m)}(\omega,\theta) \le x\} = \bigcup_j E_j^x \times \Theta_j \in \mathcal{F} \otimes \mathcal{B}_\Theta \tag{22.7}$$

being a finite union of measurable rectangles. Since this is true for any x, $f_{(m)}$ is $\mathcal{F} \otimes \mathcal{B}_\Theta/\bar{\mathcal{B}}$-measurable. The question to be addressed in any particular case is whether a sequence of such partitions can be constructed such that $f_{(m)} \to f$ as $m \to \infty$.

Henceforth it is assumed without further comment that suprema of stochastic functions are random variables. The following inequality should be carefully noted. The *envelope* of a class of random functions $f(\omega,\theta)$ is the random function that assumes the value $\sup_{\theta\in\Theta} f(\omega,\theta)$ at each point of Ω.

22.3 Theorem $\sup\limits_{\theta\in\Theta} \mathrm{E}(f(\theta)) \le \mathrm{E}\Big(\sup\limits_{\theta\in\Theta} f(\theta)\Big)$.

Proof Appealing to **3.35**, it suffices to prove the inequality for simple functions. A simple function depending on θ has the form

$$\varphi(\omega,\theta) = \sum_{i=1}^m \alpha_i(\theta) 1_{E_i}(\omega) = \alpha_i(\theta), \omega \in E_i. \tag{22.8}$$

Defining $\alpha_i^* = \sup_{\theta\in\Theta} \alpha_i(\theta)$,

$$\sup_{\theta\in\Theta} \varphi(\omega,\theta) = \alpha_i^*, \omega \in E_i. \tag{22.9}$$

Hence

$$\sup_{\theta\in\Theta} \mathrm{E}\big(\varphi(\theta)\big) - \mathrm{E}\Big(\sup_{\theta\in\Theta}\varphi(\theta)\Big) = \sup_{\theta\in\Theta}\sum_{i=1}^{m}\big(\alpha_i(\theta)-\alpha_i^*\big)P(E_i) \le 0 \tag{22.10}$$

where the final inequality is by definition of α_i^*. ∎

22.2 Pointwise and Uniform Convergence

Consider the convergence (a.s., in pr., in L_p, etc.) of the stochastic sequence $\{Q_n(\theta)\}$ to a limit function $Q(\theta)$. Typically such questions relate to the law of large numbers, with

$$Q_n(\theta) = \sum_{t=1}^{n} q_{nt}(\theta) \tag{22.11}$$

(using array notation for generality but the case $q_{nt} = q_t/n$ may generally be assumed) where the limit function by hypothesis is $Q(\theta) = \lim_{n\to\infty}\mathrm{E}(Q_n(\theta))$. Alternatively, consider the case $G_n(\theta) \to 0$ where

$$G_n(\theta) = \sum_{t=1}^{n}\big(q_{nt}(\theta) - \mathrm{E}(q_{nt}(\theta))\big). \tag{22.12}$$

Consideration of (22.12) divides the problem into two parts: the stochastic convergence of the sum of the mean deviations to zero, and the nonstochastic convergence assumed in the definition of $Q(\theta)$. This raises the separate question of whether the latter convergence is uniform, but this is a matter for the problem at hand and is not treated here.

Obedience to a law of large numbers calls for both the boundedness and the dependence of the sequence to be controlled. In the case of a function on Θ, the dependence question presents no extra difficulty; for example, if $q_{nt}(\theta_1)$ is a mixing or near-epoch dependent array of a given class, the property will generally be shared by $q_{nt}(\theta_2)$, for any $\theta_1, \theta_2 \in \Theta$. But the existence of particular moments is clearly not independent of θ. If there exists a positive array $\{D_{nt}\}$ such that $|q_{nt}(\theta)| \le D_{nt}$ for all $\theta \in \Theta$ and $\|D_{nt}\|_r < \infty$ uniformly in t and n, $q_{nt}(\theta)$ is said to be *L_r-dominated.* To ensure convergence pointwise on Θ the $q_{nt}(\theta)$ need to be bounded functions of θ. More generally it is necessary to bound Θ itself, but since Θ will often have to be bounded for a different set of reasons this does not necessarily present an additional restriction.

Given restrictions on the dependence plus suitable domination conditions, pointwise stochastic convergence follows by considering $\{G_n(\theta)\}$ as an ordinary stochastic sequence, for each $\theta \in \Theta$. However, this line of argument does not guarantee that there is a minimum rate of convergence that applies for all θ, which is the condition of uniform convergence. If pointwise convergence of $\{G_n(\theta)\}$ to the limit $G(\theta) = 0$ is defined by

$$G_n(\theta) \to 0 \text{ (a.s., in } L_p\text{, or in pr.), each } \theta \in \Theta, \tag{22.13}$$

a sequence of stochastic functions $\{G_n(\theta)\}$ is said to *converge uniformly* (a.s., in L_p, or in pr.) in Θ if

$$\sup_{\theta\in\Theta} |G_n(\theta)| \to 0 \text{ (a.s., in } L_p\text{, or in pr.).} \tag{22.14}$$

To appreciate the difference, consider the following example.

22.4 Example Let $\Theta = [0, \infty)$ and define a zero-mean array $\{g_{nt}(\theta)\}$ where

$$g_{nt}(\theta) = \frac{h_t}{n} + \begin{cases} Z\theta & 0 \le \theta \le 1/2n \\ Z(1/n - \theta), & 1/2n < \theta \le 1/n \\ 0, & 1/n < \theta < \infty \end{cases} \tag{22.15}$$

where $\{h_t\}$ is a zero-mean stochastic sequence and Z is a two-point r.v. with $P(Z = 1) = P(Z = -1) = \frac{1}{2}$. Then $G_n(\theta) = \sum_{t=1}^n g_{nt}(\theta) = H_n + K_n(\theta)$, where $H_n = n^{-1} \sum_{t=1}^n h_t$, and

$$K_n(\theta) = \begin{cases} Zn\theta, & 0 \le \theta \le 1/2n \\ Z(1 - n\theta), & 1/2n < \theta \le 1/n \\ 0, & 1/n < \theta < \infty. \end{cases} \tag{22.16}$$

Assume $H_n \to_{\text{a.s.}} 0$. Since $G_n(\theta) = H_n$ for $\theta > 1/n$ as well as for $\theta = 0$, $G_n(\theta) \to_{\text{a.s.}} 0$ for each fixed $\theta \in \Theta$. In other words, $G_n(\theta)$ converges pointwise to zero, a.s. However, $\sup_{\theta\in\Theta}|K_n(\theta)| = |\frac{1}{2}Z| = \frac{1}{2}$ for every $n \ge 1$. Because H_n converges a.s. there will exist N such that $|H_n| < \frac{1}{4}$ for all $n \ge N$ with probability 1. When $|H_n| < \frac{1}{4}$ the supremum on Θ of $|H_n + K_n(\theta)|$ is always attained at the point $\theta = 1/2n$. Hence, with probability 1,

$$\begin{aligned} \sup_{\theta\in\Theta} |G_n(\theta)| &= |H_n + \tfrac{1}{2}Z| \text{ for } n \ge N \\ &\to \tfrac{1}{2} \text{ as } n \to \infty. \end{aligned} \tag{22.17}$$

It follows that the uniform a.s. limit of $G_n(\theta)$ is not zero. Similarly, for $n \ge N$ and $\varepsilon > 0$,

$$P\Big(\sup_{\theta\in\Theta}|G_n(\theta)| \geq \varepsilon\Big) = P\big(|H_n + \tfrac{1}{2}Z| \geq \varepsilon\big) \to P\big(|\tfrac{1}{2}Z| \geq \varepsilon\big) = 1 \tag{22.18}$$

so that the uniform probability limit is not zero either, although the pointwise probability limit must equal the pointwise a.s. limit. □

The *Glivenko–Cantelli theorem* is a classic of the probability literature and also of interest as being a case outside the class of functions to be considered subsequently. For a collection of identically distributed r.v.s $\{X_1(\omega), \ldots, X_n(\omega)\}$ on the probability space $(\Omega, \mathcal{F}, P)$, the *empirical distribution function* is defined as

$$F_n(x,\omega) = \frac{1}{n}\sum_{t=1}^{n} 1_{(-\infty,x]}\big(X_t(\omega)\big). \tag{22.19}$$

In other words, the random variable $F_n(x,\omega)$ is the relative frequency of the variables in the set not exceeding x. A natural question to pose is whether (and in what sense) F_n converges to F, the true marginal c.d.f. for the distribution.

For fixed x, $\{F_n(x,\omega)\}_1^\infty$ is a stochastic sequence being the average of n Bernoulli-distributed random variables taking the value 1 with probability $F(x)$ and 0 otherwise. If these form a stationary ergodic sequence, for example, $F_n(x,\omega) \to F(x)$ a.s. for each $x \in \mathbb{R}$ and the strong law of large numbers is said to hold pointwise on $\mathbb{R}$. Convergence is achieved at x for all $\omega \in C_x$, where $P(C_x) = 1$. The problem is that to say that the *functions* F_n converge a.s. requires that a.s. convergence is achieved at each of an uncountable set of points. It is not legitimate to appeal to **3.12**(iii) to claim that $P(\bigcap_{x\in\mathbb{R}} C_x) = 1$ and hence the assertion that $F_n(x,\omega) \to F(x)$ with probability 1 *at a point x not specified beforehand* cannot be proved in this manner. This is a problem for a.s. convergence additional to the possibility of convergence breaking down at certain points of the parameter space, illustrated by **22.4**. However, uniform convergence is the condition that suffices to rule out either difficulty. Since the c.d.f. is bounded, monotone, and right-continuous, uniform continuity can be proved by establishing a.s. convergence just at a countable collection of points of $\mathbb{R}$.

22.5 Theorem (Glivenko–Cantelli) If $F_n(x,\omega) \to F(x)$ a.s. pointwise for $x \in \mathbb{R}$, then

$$\sup_x |F_n(x,\omega) - F(x)| \to 0 \text{ a.s.} \tag{22.20}$$

Proof First define in parallel with F_n,

$$F'_n(x,\omega) = \frac{1}{n}\sum_{t=1}^{n} 1_{(-\infty,x)}\big(X_t(\omega)\big) \tag{22.21}$$

(with the indicated sets open above) and note that $F_n'(x,\omega) \to F(x-)$ for all ω in a set C_x', where $P(C_x') = 1$. For an integer $m > 1$ let

$$x_{jm} = \inf\{x \in \mathbb{R} : F(x) \geq j/m\},\ j = 1, \ldots, m-1 \tag{22.22}$$

and also let $x_{0m} = -\infty$ and $x_{mm} = +\infty$ so that by construction,

$$F(x_{jm}-) - F(x_{j-1,m}) \leq 1/m,\ j = 1, \ldots, m. \tag{22.23}$$

Lastly, let

$$M_{mn}(\omega) = \max_{1\leq j\leq m}\{\max\{|F_n(x_{jm},\omega) - F(x_{jm})|, |F_n'(x_{jm},\omega) - F(x_{jm}-)|\}\}. \tag{22.24}$$

Then, for $j = 1, \ldots, m$ and $x \in [x_{j-1,m}, x_{jm})$,

$$\begin{aligned} F(x) - \frac{1}{m} - M_{mn}(\omega) &\leq F(x_{j-1,m}) - M_{mn}(\omega) \\ &\leq F_n(x_{j-1,m},\omega) \leq F_n(x,\omega) \leq F_n'(x_{jm},\omega) \\ &\leq F(x_{jm}-) + M_{mn}(\omega) \leq F(x) + \frac{1}{m} + M_{mn}(\omega). \end{aligned} \tag{22.25}$$

That is to say, $|F_n(x,\omega) - F(x)| \leq 1/m + M_{mn}(\omega)$ for every $x \in \mathbb{R}$. By pointwise strong convergence, $\lim_{n\to\infty} M_{mn}(\omega) = 0$ for finite m and hence

$$\lim_{n\to\infty} \sup_x |F_n(x,\omega) - F(x)| \leq 1/m$$

for all $\omega \in C_m^*$ where

$$C_m^* = \bigcap_{j=1}^{m} (C_{x_{jm}} \cap C'_{x_{jm}}). \tag{22.26}$$

But $P(\lim_{m\to\infty} C_m^*) = 1$ by **3.12**(iii) and this completes the proof. ∎

Another quite separate problem calling for uniform convergence is when a sample statistic is not merely a stochastic function of parameters but is to be evaluated at a random point in the parameter space. Estimates of covariance matrices of estimators generally have this character, for example. One way such estimates are obtained is as the inverted negative Hessian matrix of the associated sample log-likelihood function, evaluated at estimated parameter values. The problem of proving consistency involves two distinct stochastic convergence phenomena and it does not suffice to appeal to an ordinary law of large numbers to establish convergence to the true function evaluated at the true point. The following theorem gives sufficient conditions for the double convergence to hold.

22.6 Theorem Let $(\Omega, \mathcal{F}, P)$ be a probability space and (Θ, ρ) a metric space and let $Q_n : \Theta \times \Omega \mapsto \mathbb{R}$ be $\mathcal{F}/\mathcal{B}$-measurable for each $\theta \in \Theta$. If

(a) $\theta_n^* \to_{\text{pr}} \theta_0$ and

(b) $Q_n(\theta) \to_{\text{pr}} Q(\theta)$ uniformly in an open set $B_0 \subseteq \Theta$ containing θ_0 where $Q(\theta)$ is a nonstochastic function continuous at θ_0

then $Q_n(\theta_n^*) \to_{\text{pr}} Q(\theta_0)$.

Proof Uniform convergence in probability of Q_n in B_0 implies that for any $\varepsilon > 0$ and $\delta > 0$ $\exists$ $N_1 \geq 1$ large enough that for $n \geq N_1$,

$$P\Big(\sup_{\theta \in B_0} |Q_n(\theta) - Q(\theta)| < \tfrac{1}{2}\varepsilon\Big) \geq 1 - \tfrac{1}{4}\delta. \tag{22.27}$$

Also, since $\theta_n^* \to_{\text{pr}} \theta_0$ $\exists$ N_2 such that for $n \geq N_2$,

$$P(\theta_n^* \in B_0) \geq 1 - \tfrac{1}{4}\delta. \tag{22.28}$$

To consider the joint occurrence of these two events use the elementary relation

$$P(A \cap B) \geq P(A) + P(B) - 1. \tag{22.29}$$

Since

$$\{\theta_n^* \in B_0\} \cap \Big\{\sup_{\theta \in B_0} |Q_n(\theta) - Q(\theta)| < \tfrac{1}{2}\varepsilon\Big\} \subseteq \{|Q_n(\theta_n^*) - Q(\theta_n^*)| < \tfrac{1}{2}\varepsilon\} \tag{22.30}$$

it follows that for $n \geq \max(N_1, N_2)$,

$$P\big(|Q_n(\theta_n^*) - Q(\theta_n^*)| < \tfrac{1}{2}\varepsilon\big) \geq 2(1 - \tfrac{1}{4}\delta) - 1 = 1 - \tfrac{1}{2}\delta. \tag{22.31}$$

Using continuity at θ_0 and **19.9**(ii), there exists N_3 large enough that, for $n \geq N_3$,

$$P\big(|Q(\theta_n^*) - Q(\theta_0)| < \tfrac{1}{2}\varepsilon\big) \geq 1 - \tfrac{1}{2}\delta. \tag{22.32}$$

By the triangle inequality,

$$|Q_n(\theta_n^*) - Q(\theta_n^*)| + |Q(\theta_n^*) - Q(\theta_0)| \geq |Q_n(\theta_n^*) - Q(\theta_0)| \tag{22.33}$$

and hence

$$\{|Q_n(\theta_n^*) - Q(\theta_n^*)| < \tfrac{1}{2}\varepsilon\} \cap \{|Q(\theta_n^*) - Q(\theta_0)| < \tfrac{1}{2}\varepsilon\} \\ \subseteq \{|Q_n(\theta_n^*) - Q(\theta_0)| < \varepsilon\}. \tag{22.34}$$

Applying (22.29) again gives for $n \geq \max(N_1, N_2, N_3)$,

$$P\big(|Q_n(\theta_n^*) - Q(\theta_0)| < \varepsilon\big) \geq 1 - \delta. \tag{22.35}$$

The theorem follows since δ and ε are arbitrary. ∎

Notice why uniform convergence is needed here. Pointwise convergence would not allow the assertion of (22.27) for a *single* N_1 that works for all $\theta \in B_0$. There would be the risk of a sequence of points existing in B_0 on which N_1 is diverging. Suppose that in Example **22.4**, $G_n(\theta) = Q_n(\theta) - Q(\theta)$ and $\theta_0 = 0$. For arbitrary $\varepsilon > 0$ and $\delta > 0$, a sequence $\{1/m, m \in \mathbb{N}\}$ approaching θ_0 has the property

$$P\big(|Q_n(1/m) - Q(1/m)| < \tfrac{1}{2}\varepsilon\big) \geq 1 - \tfrac{1}{4}\delta \tag{22.36}$$

only for $n > m$. There is no finite n for which (22.31) holds and the proof collapses.

22.3 Stochastic Equicontinuity

In Example **22.4** the sequence of functions $\{G_n(\theta)\}$ is continuous for each n, but the continuity breaks down in the limit. This points to a link between uniform convergence and continuity. Continuity was not needed to prove the Glivenko–Cantelli theorem but the c.d.f. is rather a special type of function, with behaviour at discontinuities (and elsewhere) subject to tight limitations. In the wider class of functions, not necessarily bounded and monotone, continuity is the condition that has generally been exploited to get uniform convergence results. Imposing continuity uniformly over the sequence could suffice to eliminate failures of uniform convergence.

Care with the terminology is needed here because 'uniform continuity' is a well-established term for something completely different. The required concept is *equicontinuity* or, to be more precise, *asymptotic uniform equicontinuity* as defined by (5.47). The results to follow are based on the following version of the Arzelà–Ascoli theorem (**5.28**).

22.7 Theorem Let $\{f_n(\theta), n \in \mathbb{N}\}$ be sequence of functions on a totally bounded parameter space (Θ, ρ). Then, $\sup_{\theta\in\Theta} |f_n(\theta)| \to 0$ iff $f_n(\theta) \to 0$ for all $\theta \in \Theta_0$, where Θ_0 is a dense subset of Θ and $\{f_n\}$ is asymptotically uniformly equicontinuous. □

In the notation introduced in §5.6, the set $\mathbb{F} = \{f_n, n \in \mathbb{N}\} \cup \{0\}$ endowed with the uniform metric is a subspace of (C_Θ, d_U). By definition, convergence of f_n to 0 in the uniform metric is the same thing as uniform convergence on Θ. According to **5.12** compactness of $\mathbb{F}$ is equivalent to the property that every sequence in $\mathbb{F}$ has a cluster point. In view of the pointwise convergence the cluster point must be unique and equal to 0, so that the conclusion of this theorem is really identical is that of the Arzelà–Ascoli theorem, although the method of proof is adapted to the present case. A version of the notation defined in (5.48) for functions of a real variable is adopted, focusing in this case on the second argument. The modulus of continuity of f_n is the mapping $w(f_n, \cdot) : \mathbb{R}^+ \mapsto \mathbb{R}^+$ where

$$w(f_n, \delta) = \sup_{\theta\in\Theta} \sup_{\theta'\in S(\theta,\delta)} |f_n(\theta') - f_n(\theta)|. \tag{22.37}$$

Asymptotic uniform equicontinuity of the sequence $\{f_n\}$ is the property that $\limsup_n w(f_n, \delta) \downarrow 0$ as $\delta \downarrow 0$.

Proof of 22.7 To prove 'if': given $\varepsilon > 0$, there exists by assumption $\delta > 0$ to satisfy

$$\limsup_{n\to\infty} w(f_n, \delta) < \varepsilon. \tag{22.38}$$

Since Θ is totally bounded it has a finite cover $\{S(\theta_i, \delta/2), i = 1, \ldots, m\}$. For each i, choose $\tilde{\theta}_i \in \Theta_0$ such that $\rho(\theta_i, \tilde{\theta}_i) < \delta/2$ (possible because Θ_0 is dense in Θ) and note that $\{S(\tilde{\theta}_i, \delta), i = 1, \ldots, m\}$ is also a cover for Θ. Every $\theta \in \Theta$ is contained in $S(\tilde{\theta}_i, \delta)$ for some i and for this i,

$$|f_n(\theta)| \le \sup_{\theta'\in S(\tilde{\theta}_i,\delta)} |f_n(\theta')| \le \sup_{\theta'\in S(\tilde{\theta}_i,\delta)} |f_n(\theta') - f_n(\tilde{\theta}_i)| + |f_n(\tilde{\theta}_i)|. \tag{22.39}$$

Therefore,

$$\begin{aligned} \sup_{\theta\in\Theta} |f_n(\theta)| &\le \max_{1\le i\le m} \sup_{\theta'\in S(\tilde{\theta}_i,\delta)} |f_n(\theta') - f_n(\tilde{\theta}_i)| + \max_{1\le i\le m} |f_n(\tilde{\theta}_i)| \\ &\le w(f_n, \delta) + \max_{1\le i\le m} |f_n(\tilde{\theta}_i)|. \end{aligned} \tag{22.40}$$

Sufficiency follows on taking the limsup of both sides of this inequality.

'Only if' follows from the facts that uniform convergence entails pointwise convergence and that

$$w(f_n, \delta) \leq 2 \sup_{\theta \in \Theta} |f_n(\theta)|. \quad \blacksquare \tag{22.41}$$

To apply this result to the stochastic convergence problem calls for concepts of *stochastic* equicontinuity. Of the several such definitions that can be devised the important variants are for weak convergence (in pr.) and strong convergence (a.s.). Let (Θ, ρ) be a metric space and $(\Omega, \mathcal{F}, P)$ a probability space and let $\{G_n(\theta, \omega), n \in \mathbb{N}\}$ be a sequence of stochastic functions $G_n : \Theta \times \Omega \mapsto \mathbb{R}$, $\mathcal{F}/\mathcal{B}$-measurable for each $\theta \in \Theta$. The sequence is said to be *asymptotically uniformly stochastically equicontinuous* (in pr.) if for all $\varepsilon > 0 \ \exists \ \delta > 0$ such that

$$\limsup_{n\to\infty} P\big(w(G_n, \delta) \geq \varepsilon\big) < \varepsilon. \tag{22.42}$$

And it is said to be *strongly asymptotically uniformly stochastically equicontinuous* if for all $\varepsilon > 0 \ \exists \ \delta > 0$ such that

$$P\Big(\limsup_{n\to\infty} w(G_n, \delta) \geq \varepsilon\Big) = 0. \tag{22.43}$$

Clearly, there is a bit of a terminology problem here! The qualifiers 'asymptotic' and 'uniform' will be adopted in all the applications in this chapter, so let these be understood and speak simply of stochastic equicontinuity and strong stochastic equicontinuity. The abbreviations s.e. and s.s.e. will sometimes be used.

22.4 Generic Uniform Convergence

Uniform convergence results and their application in econometrics have been researched by several authors including Hoadley ([101]), Bierens ([20]), Andrews ([5], [8]), Newey ([136]), and Pötscher and Prucha ([148], [151]). The material in the remainder of this chapter is drawn mainly from the work of Andrews and of Pötscher and Prucha. These authors have developed alternative 'generic' uniform convergence results, applicable in a variety of modelling situations.

These methods rely on establishing a stochastic equicontinuity condition. Given **22.7**, the proof of uniform almost sure convergence is direct.

22.8 Theorem Let $\{G_n(\theta), n \in \mathbb{N}\}$ be a sequence of stochastic real-valued functions on a totally bounded metric space (Θ, ρ). Then,

$$\sup_{\theta \in \Theta} |G_n(\theta)| \xrightarrow{\text{a.s.}} 0 \tag{22.44}$$

iff

(a) $G_n(\theta) \to_{\text{a.s.}} 0$ for each $\theta \in \Theta_0$, where Θ_0 is a dense subset of Θ; and
(b) $\{G_n\}$ is strongly stochastically equicontinuous.

Proof Because (Θ, ρ) is totally bounded it is separable (**5.7**) and Θ_0 can be chosen to be a countable set, say $\Theta_0 = \{\theta_k, k \in \mathbb{N}\}$. Condition (a) means that for $k = 1, 2, \ldots$ there is a set C_k with $P(C_k) = 1$ such that $G_n(\theta_k, \omega) \to 0$ for $\omega \in C_k$. Condition (b) means that the sequences $\{G_n(\omega)\}$ are asymptotically equicontinuous for all $\omega \in C'$, with $P(C') = 1$. By the sufficiency part of **22.7**, $\sup_{\theta\in\Theta}|G_n(\theta,\omega)| \to 0$ for $\omega \in C^* = \bigcap_{k=1}^{\infty} C_k \cap C'$. $P(C^*) = 1$ by **3.12**(iii), proving 'if'.

'Only if' follows from the necessity part of **22.7** applied to $\{G_n(\omega)\}$ for each $\omega \in C^*$. ∎

The corresponding 'in probability' result follows very similar lines. The proof cannot exploit **22.7** quite so directly, but the family resemblance in the arguments will be noted.

22.9 Theorem Let $\{G_n(\theta), n \in \mathbb{N}\}$ be a sequence of stochastic real-valued functions on a totally bounded metric space (Θ, ρ). Then,

$$\sup_{\theta\in\Theta}|G_n(\theta)| \xrightarrow{\text{pr}} 0 \tag{22.45}$$

iff

(a) $G_n(\theta) \to_{\text{pr}} 0$ for each $\theta \in \Theta_0$, where Θ_0 is a dense subset of Θ; and
(b) $\{G_n\}$ is stochastically equicontinuous.

Proof To show 'if', let $\{S(\tilde{\theta}_i, \delta), i = 1, \ldots, m\}$ with $\tilde{\theta}_i \in \Theta_0$ be a finite cover for Θ. This exists by the assumption of total boundedness and the argument used in the proof of **22.7**. Then,

$$\begin{aligned} P\Big(\sup_{\theta\in\Theta}|G_n(\theta)| \geq 2\varepsilon\Big) &\leq P\Big(\max_{1\leq i\leq m}\sup_{\theta'\in S(\tilde{\theta}_i,\delta)}\big(|G_n(\theta') - G_n(\theta_i)| + |G_n(\theta_i)|\big) \geq 2\varepsilon\Big) \\ &\leq P\big(w(G_n,\delta) \geq \varepsilon\big) + P\Big(\max_{1\leq i\leq m}|G_n(\theta_i)| \geq \varepsilon\Big) \\ &\leq P\big(w(G_n,\delta) \geq \varepsilon\big) + P\Big(\bigcup_{i=1}^{m}\{|G_n(\tilde{\theta}_i)| \geq \varepsilon\}\Big) \\ &\leq P\big(w(G_n,\delta) \geq \varepsilon\big) + \sum_{i=1}^{m} P\big(|G_n(\tilde{\theta}_i)| \geq \varepsilon\big) \end{aligned} \tag{22.46}$$

using

$$\{x+y \geq 2\varepsilon\} \subseteq \{x \geq \varepsilon\} \cup \{y \geq \varepsilon\} \tag{22.47}$$

for real numbers x and y to get the second inequality. The m probabilities in the sum in the majorant of (22.46) each vanish in the limit by (a) and hence so does their sum. Therefore, taking the limsup, (a) and (b) and (22.42) together imply that

$$\limsup_{n\to\infty} P\Big(\sup_{\theta\in\Theta}|G_n(\theta)| \geq 2\varepsilon\Big) < \varepsilon. \tag{22.48}$$

To prove 'only if', pointwise convergence follows immediately from uniform convergence, so it remains to be shown that s.e. holds; but this follows easily in view of the fact (see (22.41)) that

$$P\big(w(G_n,\delta) \geq \varepsilon\big) \leq P\Big(\sup_{\theta\in\Theta}|G_n(\theta)| \geq \varepsilon/2\Big). \quad \blacksquare \tag{22.49}$$

A leading case of these theorems is $G_n(\theta) = Q_n(\theta) - \bar{Q}_n(\theta)$, where $\bar{Q}_n$ is a nonstochastic function which may really depend on n or just be a limit function so that $\bar{Q}_n = \bar{Q}$. In the former case there is no need for Q_n to converge, as long as $Q_n - \bar{Q}_n$ does. Applying the triangle inequality and taking complements in (22.47),

$$\big\{w(Q_n,\delta) < \varepsilon\big\} \cap \big\{w(\bar{Q}_n,\delta) < \varepsilon\big\} \subseteq \big\{w(Q_n - \bar{Q}_n,\delta) < 2\varepsilon\big\}. \tag{22.50}$$

This means that $\{Q_n - \bar{Q}_n\}$ is s.e. or s.s.e., as the case may be, provided that $\{Q_n\}$ is s.e. or s.s.e.; and $\{\bar{Q}_n\}$ is asymptotically equicontinuous in the ordinary sense of §5.6. The extension of **22.8** is obvious and in **22.9** insert the step

$$P\big(w(Q_n - \bar{Q}_n,\delta) \geq 2\varepsilon\big) \leq P\big(w(Q_n,\delta) \geq \varepsilon\big) + 1_{\{w(\bar{Q}_n,\delta)\geq\varepsilon\}} \tag{22.51}$$

into (22.46), where the second term on the right is 0 or 1 depending on whether the indicated nonstochastic condition holds and this term will vanish when $n \geq N$ for some $N \geq 1$, by assumption.

The s.e. and s.s.e. conditions may not be particularly easy to verify directly and the existence of Lipschitz-type sufficient conditions could then be very convenient. Andrews ([8]) suggests conditions of the following sort.

22.10 Theorem Suppose there exists $N \geq 1$ such that

$$|Q_n(\theta') - Q_n(\theta)| \leq B_n h\big(\rho(\theta,\theta')\big) \text{ a.s.} \tag{22.52}$$

holds for all $\theta, \theta' \in \Theta$ and $n \geq N$, where h is nonstochastic and $h(x) \downarrow 0$ as $x \downarrow 0$ and $\{B_n\}$ is an a.s. positive stochastic sequence not depending on θ. Then,

(i) $\{Q_n\}$ is s.e. if $B_n = O_p(1)$.

(ii) $\{Q_n\}$ is s.s.e. if $\limsup_n B_n < \infty$, a.s.

Proof The definitions imply that $w(Q_n, \delta) \leq B_n h(\delta)$ a.s. for $n \geq N$. To prove (i) note that, for any $\varepsilon > 0$ and $\delta > 0$,

$$\limsup_{n\to\infty} P\big(w(Q_n,\delta) \geq \varepsilon\big) \leq \limsup_{n\to\infty} P\big(B_n \geq \varepsilon/h(\delta)\big). \tag{22.53}$$

By definition of $O_p(1)$ the right-hand side can be made arbitrarily small by choosing $\varepsilon/h(\delta)$ large enough. In particular fix $\varepsilon > 0$ and then, by definition of h, δ can be small enough that $\limsup_{n\to\infty} P\big(B_n \geq \varepsilon/h(\delta)\big) < \varepsilon$. For (ii), in the same way,

$$P\Big(\limsup_{n\to\infty} w(Q_n,\delta) \geq \varepsilon\Big) \leq P\Big(\limsup_{n\to\infty} B_n \geq \varepsilon/h(\delta)\Big) < \varepsilon \tag{22.54}$$

when δ is small enough. ∎

A sufficient condition for $B_n = O_p(1)$ is to have B_n uniformly bounded in L_1 norm, that is, $\sup_n \mathrm{E}(B_n) < \infty$ (see **12.13**), and it is sufficient for $\limsup_n B_n$ to be a.s. bounded if, in addition to this, $B_n - \mathrm{E}(B_n) \to_{\text{a.s.}} 0$.

The conditions of **22.10** offer a striking contrast in restrictiveness. Think of (22.52) as a continuity condition, which says that $Q_n(\theta')$ must be close to $Q_n(\theta)$ when θ' is close to θ. When Q_n is stochastic these conditions are very hard to satisfy for *fixed* B_n, because random changes of scale may lead the condition to be violated from time to time even if $Q_n(\theta, \omega)$ is a continuous function for all ω and n. The purpose of the factor B_n is to allow for such random scale variations.

Under s.e., the probability of large variations must decline as their magnitude increases; this is what $O_p(1)$ means. But in the s.s.e. case, the requirement that $\{B_n\}$ be bounded a.s. except for at most a finite number of terms implies that $\{Q_n\}$ must satisfy the same condition. This is very restrictive. It means for example that $Q_n(\theta)$ cannot be Gaussian nor have any other distribution with infinite support. In such a case, no matter what $\{B_n\}$ and h were chosen, the condition in (22.52) would be violated eventually. It does not matter that the probability of large deviations might be extremely small, because over an *infinite* number of sequence coordinates they will still arise with probability 1.

Thus, strong uniform convergence is a phenomenon confined to a.s. bounded sequences. Although (22.52) is only a sufficient condition, it can be verified that this feature of s.s.e. is implicit in the definition. This fact puts the relative merits of working with strong and weak laws of large numbers in a new light. The former are simply not available in many important cases. Fortunately, 'in probability' results are often sufficient for the purpose at hand, for example determining the limits in distribution of estimators and sample statistics. See §26.1 for more details.

Supposing $(\Theta, \rho) \subset (\mathbb{R}^k, d_E)$, suppose further that $Q_n(\theta)$ is differentiable a.s. at each point of Θ. To be precise, specify differentiability a.s. at each point of an open convex set Θ^* containing Θ. (A set $B \subset \mathbb{R}^k$ is said to be convex if $\boldsymbol{x} \in B$ and $\boldsymbol{y} \in B$ imply $\lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{y} \in B$ for $\lambda \in [0,1]$.) The mean value theorem yields the result that at a pair of points $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta^*$,[1]

$$Q_n(\boldsymbol{\theta}) - Q_n(\boldsymbol{\theta}') = \sum_{i=1}^{k} \left.\frac{\partial Q_n}{\partial \theta_i}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} (\theta_i - \theta_i') \text{ a.s.} \tag{22.55}$$

where $\boldsymbol{\theta}^* \in \Theta^*$ is a point on the line segment joining $\boldsymbol{\theta}$ and $\boldsymbol{\theta}'$, which exists by convexity of Θ^*. Applying the Cauchy–Schwarz inequality,

$$|Q_n(\boldsymbol{\theta}) - Q_n(\boldsymbol{\theta}')| \leq \sum_{i=1}^{k} \left|\left.\frac{\partial Q_n}{\partial \theta_i}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*}\right| |\theta_i - \theta_i'| \leq B_n \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \text{ a.s.} \tag{22.56}$$

where, letting $\partial Q_n / \partial \boldsymbol{\theta}$ denote the gradient vector whose elements are the partials of Q_n with respect to the θ_i, and $\|\cdot\|$ denote the Euclidean length,

$$B_n = \sup_{\boldsymbol{\theta}^* \in \Theta^*} \left\| \left.\frac{\partial Q_n}{\partial \boldsymbol{\theta}}\right|_{\boldsymbol{\theta}=\boldsymbol{\theta}^*} \right\|. \tag{22.57}$$

Clearly, (22.52) is satisfied by taking h as the identity function and B_n defined in (22.57) is a random variable for all n. Subject to this condition and B_n satisfying the conditions specified in **22.10**, a.s. differentiability emerges as a sufficient condition for s.e.

22.5 Uniform Laws of Large Numbers

In the last section it was shown that stochastic equicontinuity (strong or in pr.) is a necessary and sufficient condition to go from pointwise to uniform convergence (strong or in pr.). The next task is to find sufficient conditions for stochastic

[1] Since $\boldsymbol{\theta}$ is here a real k-vector it is written in bold face by convention, notwithstanding that θ is used to denote the generic element of (Θ, ρ) in the abstract.

equicontinuity when $\{Q_n(\theta)\}$ is a sequence of partial sums and hence to derive uniform laws of large numbers. There are several possible approaches to this problem, of which perhaps the simplest is to establish the Lipschitz condition of **22.10**.

22.11 Theorem Let $\{\{q_{nt}(\omega,\theta)\}_{t=1}^{n}\}_{n=1}^{\infty}$ denote a triangular array of real stochastic functions with domain (Θ,ρ) satisfying, for $N \geq 1$,

$$|q_{nt}(\theta') - q_{nt}(\theta)| \leq B_{nt}h\big(\rho(\theta,\theta')\big), \text{ a.s.} \tag{22.58}$$

for all $\theta,\theta' \in \Theta$ and $n \geq N$, where h is nonstochastic, $h(x) \downarrow 0$ as $x \downarrow 0$, and $\{B_{nt}\}$ is an a.s. positive stochastic array not depending on θ with $\sum_{t=1}^{n} \mathrm{E}(B_{nt}) = O(1)$. If $Q_n(\theta) = \sum_{t=1}^{n} q_{nt}(\theta)$ then

(i) Q_n is s.e.

(ii) Q_n is s.s.e. if $\sum_{t=1}^{n}(B_{nt} - \mathrm{E}(B_{nt})) \to_{\text{a.s.}} 0$.

Proof For (i) it is only necessary by **22.10**(i) and the triangle inequality to establish that $\sum_{t=1}^{n} B_{nt} = O_p(1)$. This follows from the stated condition by the Markov inequality. Likewise, (ii) follows directly from **22.10**(ii). ∎

A second class of conditions is obtained by applying a form of s.e. to the summands. Specify G_n to be an unweighted average of n functions, since the conditions to be imposed take the form of Cesàro summability of certain related sequences. It is convenient to confine attention to the case

$$G_n(\omega,\theta) = \frac{1}{n}\sum_{t=1}^{n}\big(q_t(X_t(\omega),\theta) - \mathrm{E}(q_t(X_t,\theta))\big) \tag{22.59}$$

where $X_t \in \mathbb{X}$ is a random element drawn from the probability space $(\mathbb{X},\mathcal{X},\mu_t)$. Typically, though *not* necessarily, X_t is a vector of real r.v.s with $\mathbb{X}$ a subset of $\mathbb{R}^m$, $m \geq 1$, $\mathcal{X}$ being the restriction of $\mathcal{B}^m$ to $\mathbb{X}$. The point here is not to restrict the form of the functional relation between q_t and ω but to specify the existence of marginal derived measures $\mu_t(A) = P(X_t \in A)$ for $A \in \mathcal{X}$. The usual context will have G_n the sample average of functions that are stochastic through their dependence on some kind of data set, indexed on t. The functions themselves, not just their arguments, can be different for different t.

To find conditions on both the functions $q_t(\cdot,\cdot)$ and the p.m.s μ_t that yield the s.e. condition on G_n, first establish conditions on the stochastic functions $q_t(\theta)$ that have to be satisfied for s.e. to hold. Andrews (1992) gives the following result.

22.12 Theorem If G_n is defined by (22.59) where

(a) $\exists$ a positive stochastic sequence $\{d_t\}$ satisfying

$$\sup_{\theta\in\Theta}|q_t(\theta)| \le d_t, \text{ all } t \tag{22.60}$$

and

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathrm{E}\big(d_t 1_{\{d_t>M\}}\big) \to 0 \text{ as } M\to\infty \tag{22.61}$$

(b) $\forall\, \varepsilon > 0\ \exists\, \delta > 0$ such that

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}P\big(w(q_t,\delta)>\varepsilon\big) < \varepsilon \tag{22.62}$$

then G_n is s.e. □

Condition (22.61) is an interesting Cesàro-sum variation on uniform integrability. Actual uniform integrability of $\{d_t\}$ is sufficient, although not necessary. Condition (a) is a domination condition, while condition (b) is called by Andrews *termwise stochastic equicontinuity*.

Proof Given $\varepsilon > 0$, choose M such that $\limsup_{n\to\infty}n^{-1}\sum_{t=1}^{n}\mathrm{E}(2d_t 1_{\{2d_t>M\}}) < \frac{1}{6}\varepsilon^2$ and then δ such that

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}P\big(w(q_t,\delta)>\tfrac{1}{6}\varepsilon^2\big) < \tfrac{1}{6}M^{-1}\varepsilon^2. \tag{22.63}$$

Next note that

$$w\big((q_t-\mathrm{E}(q_t)),\delta\big) \le w(q_t,\delta) + w\big(\mathrm{E}(q_t),\delta\big) \le w(q_t,\delta) + \mathrm{E}\big(w(q_t,\delta)\big) \tag{22.64}$$

where the last inequality is an application of **22.3**. Applying (22.64) and using Markov's inequality,

$$\begin{aligned} P\big(w(G_n,\delta)>\varepsilon\big) &\le P\Big(\frac{1}{n}\sum_{t=1}^{n}w\big(q_t-\mathrm{E}(q_t),\delta\big)>\varepsilon\Big)\\ &\le P\Big(\frac{1}{n}\sum_{t=1}^{n}\big(w(q_t,\delta)+\mathrm{E}(w(q_t,\delta))\big)>\varepsilon\Big) \end{aligned}$$

$$\begin{aligned}
&\leq \frac{2}{n\varepsilon}\sum_{t=1}^{n}\mathrm{E}(w(q_t,\delta)) \\
&= \frac{2}{n\varepsilon}\sum_{t=1}^{n}\mathrm{E}\big(w(q_t,\delta)(1_{\{w(q_t,\delta)\leq\varepsilon^2/6\}} \\
&\qquad + 1_{\{\varepsilon^2/6<w(q_t,\delta)\leq M\}} + 1_{\{w(q_t,\delta)>M\}})\big) \qquad (22.65)
\end{aligned}$$

where the indicator functions in the last member add up to 1. Concerning the middle term of this last expression, note that

$$\mathrm{E}\big(w(q_t,\delta)1_{\{\varepsilon^2/6<w(q_t,\delta)\leq M\}}\big) \leq MP\big(w(q_t,\delta) > \tfrac{1}{6}\varepsilon^2\big).$$

Also, using the fact that $w(q_t,\delta) \leq 2d_t$ and hence $\{w(q_t,\delta) > M\} \subseteq \{2d_t > M\}$ and taking the limsup, in view of the values chosen for M and δ and the assumptions,

$$\begin{aligned}
\limsup_{n\to\infty} P\big(w(G_n,\delta) > \varepsilon\big) &\leq \frac{2}{\varepsilon}\bigg(\tfrac{1}{6}\varepsilon^2 + M\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}P\big(w(q_t,\delta) > \tfrac{1}{6}\varepsilon^2\big) \\
&\qquad + \limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathrm{E}(2d_t1_{\{2d_t>M\}})\bigg) \\
&< \varepsilon. \quad \blacksquare \qquad (22.66)
\end{aligned}$$

Whether condition **22.12**(a) is satisfied depends on both the distribution of X_t and functional form of $q_t(\cdot)$, but something relatively general can be said about termwise s.e. (condition **22.12**(b)). Assume following Pötscher and Prucha ([148]) that

$$q_t(x,\theta) = \sum_{k=1}^{p} r_{kt}(x)s_{kt}(x,\theta) \qquad (22.67)$$

where $r_{kt} : \mathbb{X} \mapsto \mathbb{R}$ and $s_{kt}(\cdot,\theta) : \mathbb{X} \mapsto \mathbb{R}$ for fixed θ are $\mathcal{X}/\mathcal{B}$-measurable functions. The idea here is to allow more freedom to the factors r_{kt} as functions of X_t than to the factors s_{kt}, with discontinuities permitted, for example. Suppose that

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mathrm{E}\left|r_{kt}(X_t)\right| \leq B < \infty,\ k = 1,\ldots,p \qquad (22.68)$$

while the factors $s_{kt}(x,\theta)$ are assumed to be asymptotically equicontinuous for a sufficiently large set of x values. Specifically, there is a sequence of sets $\{K_m \in \mathcal{X}, m = 1,2,\ldots\}$ such that

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mu_t(K_m^c)\to 0 \text{ as } m\to\infty \tag{22.69}$$

and for each $m \geq 1$ and $\varepsilon > 0$, $\exists\ \delta > 0$ such that

$$\limsup_{t\to\infty}\ \sup_{x\in K_m} w\big(s_{kt}(x,\cdot),\delta\big) < \varepsilon,\ k = 1,\dots,p. \tag{22.70}$$

Notice that (22.70) is a *non*stochastic equicontinuity condition, but under condition (22.69) it holds 'almost surely, on average' when the r.v. X_t is substituted into the formula.

These conditions suffice to give termwise s.e. and hence can be used to prove s.e. of G_n by application of **22.12**.

22.13 Theorem If $q_t(X_t,\theta)$ is defined by (22.67), and (22.68), (22.69), and (22.70) hold, then for every $\varepsilon > 0$ there exists $\delta > 0$ such that

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}P\big(w(q_t,\delta)>\varepsilon\big)<\varepsilon. \tag{22.71}$$

Proof Fix $\varepsilon > 0$ and first note that

$$\begin{aligned}
P\big(w(q_t,\delta)>\varepsilon\big) &\leq P\Big(\sum_{k=1}^{p}|r_{kt}|w(s_{kt},\delta)>\varepsilon\Big)\\
&\leq P\Big(\bigcup_{k=1}^{p}\{|r_{kt}|w(s_{kt},\delta)>\varepsilon/p\}\Big)\\
&\leq \sum_{k=1}^{p}P\big(|r_{kt}|w(s_{kt},\delta)>\varepsilon/p\big)\\
&\leq \sum_{k=1}^{p}\Big(P\Big(|r_{kt}|w(s_{kt},\delta)1_{K_m}>\frac{\varepsilon}{2p}\Big)\\
&\qquad\qquad + P\Big(|r_{kt}|w(s_{kt},\delta)1_{K_m^c}>\frac{\varepsilon}{2p}\Big)\Big).
\end{aligned} \tag{22.72}$$

Consider any one of these p terms. Choose m large enough that

$$\limsup_{n\to\infty}\frac{1}{n}\sum_{t=1}^{n}\mu_t(K_m^c)<\frac{\varepsilon}{2p}. \tag{22.73}$$

For this m there exist, by (22.70), $\delta > 0$ and $t_0 \geq 1$ such that

$$\sup_{x \in K_m} w\big(s_{kt}(x, \cdot), \delta\big) < \frac{\varepsilon^2}{4Bp^2} \tag{22.74}$$

for $t > t_0$, where B is from (22.68). Applying (22.74), the Markov inequality, and then (22.68),

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} P\Big(|r_{kt}| w(s_{kt}, \delta) 1_{K_m} > \frac{\varepsilon}{2p}\Big) \\
&\leq \limsup_{n\to\infty} \frac{1}{n} \Bigg(t_0 + \sum_{t=t_0+1}^{n} P\Big(|r_{kt}| \frac{\varepsilon^2}{4Bp^2} > \frac{\varepsilon}{2p}\Big)\Bigg) \\
&\leq \limsup_{n\to\infty} \frac{1}{n} \sum_{t=t_0+1}^{n} \mathrm{E}|r_{kt}| \frac{\varepsilon}{2Bp} \\
&\leq \frac{\varepsilon}{2p}
\end{aligned} \tag{22.75}$$

whereas by (22.73),

$$\begin{aligned}
\limsup_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} P\Big(|r_{kt}| w(s_{kt}, \delta) 1_{K_m^c} > \frac{\varepsilon}{2p}\Big) &\leq \limsup_{n\to\infty} \frac{1}{n} \sum_{t=1}^{n} P(X_t \notin K_m) \\
&\leq \frac{\varepsilon}{2p}.
\end{aligned} \tag{22.76}$$

Substituting these bounds into (22.72) yields the result. ∎

PART V

THE CENTRAL LIMIT THEOREM

23
Weak Convergence of Distributions

23.1 Basic Concepts

The objects examined in this part of the book are not sequences of random variables, but sequences of marginal distribution functions. There will of course be associated sequences of r.v.s generated from these distributions, but the concept of convergence arising here is quite distinct. Formally, if $\{F_n\}_1^\infty$ is a sequence of c.d.f.s, the sequence is said to *converge weakly* to a limit F if $F_n(x) \to F(x)$ pointwise for each $x \in C$, where $C \subseteq \mathbb{R}$ is the set of points at which F is continuous. A common notation is $F_n \Rightarrow F$.

If X_n has c.d.f. F_n and X has c.d.f. F it is commonly said with a mild abuse of terminology that the random sequence X_n *converges in distribution* to X, written as $X_n \to_{\mathrm{d}} X$. Although this latter form of notation is customary in econometrics it is also really irregular. To say a sequence of r.v.s converges in distribution means only that the distributions of the coordinates of the sequence of r.v.s approximate the given distribution when n is large. It does not assert that the sequence has a limit in the usual sense. If both X and Y have the distribution specified by F, then $X_n \to_{\mathrm{d}} X$ and $X_n \to_{\mathrm{d}} Y$ are equivalent statements. It is also customary to write things like '$X_n \to_{\mathrm{d}} \mathrm{N}(0,1)$' to indicate that the limiting distribution is standard Gaussian, where '$\mathrm{N}(0,1)$' is shorthand for 'a r.v. having the standard Gaussian distribution with mean 0 and variance 1'. It does not denote a particular r.v.

Convergence of the distribution functions at continuity points is all that is needed, remembering that F is non-decreasing, right-continuous, bounded by 0 and 1, and that every point is either a continuity point or a jump point. It is possible that F could possess a jump at a point x_0 which is a continuity point of F_n for all finite n and in these cases $F_n(x_0)$ does not have a unique limit since any point between $F(x_0-)$ and $F(x_0)$ is a candidate. However, the jump points of F are at most countable in number. According to **8.5** the true F can be constructed by assigning the value $F(x_0)$ at every jump point x_0 and hence the above definition is adequate.

If μ represents the corresponding probability measure such that $F(x) = \mu\big((-\infty, x]\big)$ for each $x \in \mathbb{R}$, μ and F are equivalent representations of the same measure (see §8.2) and similarly for μ_n and F_n. Hence the statement $\mu_n \Rightarrow \mu$ is equivalent to $F_n \Rightarrow F$. The corresponding notion of weak convergence for the sequence of measures $\{\mu_n\}$ is given by the following theorem.

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0023

23.1 Theorem $\mu_n \Rightarrow \mu$ iff $\mu_n(A) \to \mu(A)$ for every $A \in \mathcal{B}$ having $\mu(\partial A) = 0$. □

The proof of this theorem is postponed to a later point in the development (see page 504). Note meanwhile that the exclusion of events whose boundary points have positive probability corresponds to the exclusion of jump points of F, where the events in question have the form $\{(-\infty, x]\}$.

Just as the theory of the expectation is an application of the general theory of integrals, so the theory of weak convergence is a general theory for sequences of finite measures. The results below do not generally depend on the condition $\mu_n(\mathbb{R}) = 1$ for their validity, provided definitions are adjusted appropriately. However, a serious concern of the theory is whether a sequence of distribution functions has a distribution function as its limit; more specifically, should it follow because $\mu_n(\mathbb{R}) = 1$ for every n that $\mu(\mathbb{R}) = 1$? This is a question that is taken up in §23.5. Meanwhile, the reader should not be distracted by the use of the convenient notations $\mathrm{E}(\cdot)$ and $P(\cdot)$ from appreciating the generality of the theory.

23.2 Example Recall Example **8.9**. In the sequence of binomial$(n, \lambda/n)$ distributions it was shown that

$$P(X_n = x) \to \frac{\lambda^x}{x!}\mathrm{e}^{-\lambda},\ x = 0,1,2,\ldots \tag{23.1}$$

as $n \to \infty$ and accordingly

$$F_n(a) = \sum_{0 \le x \le a} P(X_n = x) \to \mathrm{e}^{-\lambda} \sum_{0 \le x \le a} \frac{\lambda^x}{x!} \tag{23.2}$$

at all points $0 \le a < \infty$ and in particular at all continuity points. The Poisson distribution with parameter λ is the weak limit of the sequence. □

23.3 Example A sequence of discrete distributions on $[0,1]$ is defined by

$$P(X_n = x) = \begin{cases} 1/n, & x = i/n \\ 0 & \text{otherwise} \end{cases},\ i = 1,\ldots,n. \tag{23.3}$$

This sequence converges weakly to Lebesgue measure m on $[0,1]$. For any $x \in [0,1]$, $\mu_n\big([0,x]\big) = [nx]/n \to x = m\big([0,x]\big)$, where $[nx]$ denotes the largest integer not exceeding nx. There are sets for which convergence fails, notably the set $\mathbb{Q}_{[0,1]}$ of all rationals in $[0,1]$, in view of the fact that $\mu_n(\mathbb{Q}_{[0,1]}) = 1$ for every n and $m(\mathbb{Q}_{[0,1]}) = 0$. But $\bar{\mathbb{Q}}_{[0,1]} = [0,1]$ and so $m\big(\partial\mathbb{Q}_{[0,1]}\big) = 1$, thus the definition of weak convergence in **23.1** is not violated. □

Although convergence in distribution is fundamentally different from convergence a.s. and in pr., the latter imply the former. In the next result, '$\to_{\text{a.s.}}$' can be substituted for '$\to_{\text{pr}}$', by **19.5**.

23.4 Theorem If $X_n \to_{\text{pr}} X$, then $X_n \to_{\text{d}} X$.

Proof For $\varepsilon > 0$,

$$\begin{aligned} P(X_n \le x) &= P(\{X_n \le x\} \cap \{|X_n - X| \le \varepsilon\}) \\ &\qquad + P(\{X_n \le x\} \cap \{|X_n - X| > \varepsilon\}) \\ &\le P(X \le x + \varepsilon) + P(|X_n - X| > \varepsilon) \end{aligned} \tag{23.4}$$

where the events whose probabilities appear on the right-hand side of the inequality contain (and hence are at least as probable as) the corresponding events on the left. $P(|X_n - X| > \varepsilon) \to 0$ by hypothesis and hence

$$\limsup_{n\to\infty} P(X_n \le x) \le P(X \le x + \varepsilon). \tag{23.5}$$

Similarly,

$$\begin{aligned} P(X \le x - \varepsilon) &= P(\{X \le x - \varepsilon\} \cap \{|X_n - X| \le \varepsilon\}) \\ &\qquad + P(\{X \le x - \varepsilon\} \cap \{|X_n - X| > \varepsilon\}) \\ &\le P(X_n \le x) + P(|X_n - X| > \varepsilon) \end{aligned} \tag{23.6}$$

and so

$$P(X \le x - \varepsilon) \le \liminf_{n\to\infty} P(X_n \le x). \tag{23.7}$$

Since ε is arbitrary, it follows that $\lim_{n\to\infty} P(X_n \le x) = P(X \le x)$ at every point x for which $\lim_{\varepsilon\downarrow 0} P(X \le x - \varepsilon) = P(X \le x)$. These are the points at which $P(X = x) = 0$, hence the condition is equivalent to weak convergence. ∎

The converse of **23.4** is not true in general, but the two conditions are equivalent when the probability limit in question is a constant. A *degenerate distribution* defined for a real constant a has the form

$$F(x) = \begin{cases} 0, & x < a \\ 1, & x \ge a. \end{cases} \tag{23.8}$$

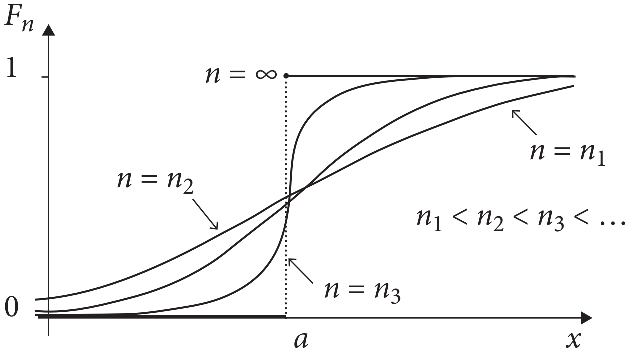

Figure 23.1

If a random variable is converging to a constant, its c.d.f. converges to the step function (23.8) through a sequence of the sort illustrated in Figure 23.1.

23.5 Theorem X_n converges in probability to a constant a iff its c.d.f. converges to a step function with jump at a.

Proof For any $\varepsilon > 0$,

$$\begin{aligned} P\big(|X_n - a| < \varepsilon\big) &= P(a - \varepsilon \le X_n \le a + \varepsilon) \\ &= F_n(a+\varepsilon) - F_n\big((a-\varepsilon)-\big). \end{aligned} \tag{23.9}$$

Convergence to a step function with jump at a implies $\lim_{n\to\infty} F_n(a+\varepsilon) = F(a+\varepsilon) = 1$ and similarly $\lim_{n\to\infty} F_n\big((a-\varepsilon)-\big) = F\big((a-\varepsilon)-\big) = 0$ for all $\varepsilon > 0$. The sufficiency part follows from (23.9) and the definition of convergence in probability. For the necessity, let the left-hand side of (23.9) have a limit of 1 as $n \to \infty$, for all $\varepsilon > 0$. This implies

$$\lim_{n\to\infty} \big(F_n(a+\varepsilon) - F_n\big((a-\varepsilon)-\big)\big) = 1. \tag{23.10}$$

Since $0 \le F \le 1$, (23.10) will be satisfied for all $\varepsilon > 0$ only if $F(a) = 1$ and $F(a-) = 0$, which defines the function in (23.8). ∎

23.2 The Skorokhod Representation Theorem

Notwithstanding the fact that $X_n \to_{\mathrm{d}} X$ does not imply $X_n \to_{\mathrm{a.s.}} X$, whenever a sequence of distributions $\{F_n\}$ converges weakly to F it is possible to construct a sequence of r.v.s with distributions F_n that converges almost surely to a limit having distribution F. Shown by Skorokhod ([170]) in a more general context (see §29.2), this is an immensely useful fact for proving results about weak convergence.

Consider the sequence $\{F_n\}$ converging to F. Each of these functions is a monotone mapping from $\bar{\mathbb{R}}$ to the interval $[0,1]$. The idea is to invert this mapping. Let a random variable ω be defined on the probability space $\big([0,1], \mathcal{B}_{[0,1]}, m\big)$ where $\mathcal{B}_{[0,1]}$ is the Borel field on the unit interval and m is the Lebesgue measure. Define for $\omega \in [0,1]$

$$Y_n(\omega) = \inf\big\{x : \omega \le F_n(x)\big\}. \tag{23.11}$$

In words, Y_n is the random variable obtained by using the inverse distribution function to map from the uniform distribution on $[0,1]$ onto $\bar{\mathbb{R}}$, taking care of any discontinuities in $F_n^{-1}(\omega)$ (corresponding to intervals with zero probability mass under F_n) by taking the infimum of the eligible values. Y_n is therefore a non-decreasing, left-continuous function. Figure 23.2 illustrates the construction, essentially the same as used in the proof of **8.6** (compare Figure 8.2 on page 159). When F_n has discontinuities it is only possible to assert (by right-continuity) that $F_n(Y_n(\omega)) \ge \omega$, whereas $Y_n(F_n(x)) \le x$, by left-continuity of Y_n.

The first important feature of the Skorokhod construction is that, for any constant $a \in \mathbb{R}$,

$$P(Y_n(\omega) \le a) = P(\omega \le F_n(a)) = F_n(a) \tag{23.12}$$

where the last equality follows from the fact that ω is uniformly distributed on $[0,1]$. Thus, F_n is the c.d.f. of Y_n. Letting F be a c.d.f. and Y the r.v. corresponding to F according to (23.11), the second important feature of the construction is contained in the following result.

23.6 Theorem If $F_n \Rightarrow F$ then $Y_n \to Y$ a.s. $[m]$ as $n \to \infty$. □

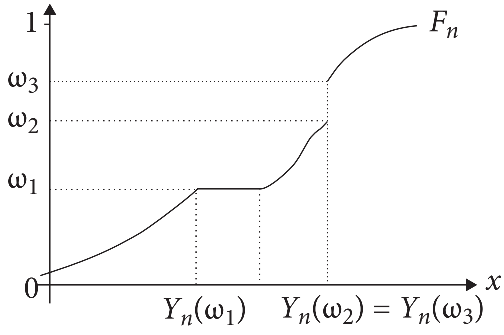

Figure 23.2

In working through the proof, it may be helpful to check each assertion about the functions F and Y against the example in Figure 23.3. This represents the extreme case where F and hence also Y is a step function; of course, if F is everywhere continuous and increasing, the mappings are 1–1 and the problem becomes trivial.

Proof of 23.6 Let ω be any continuity point of Y, excluding the end points 0 and 1. For any $\varepsilon > 0$, choose x as a continuity point of F satisfying $Y(\omega) - \varepsilon < x < Y(\omega)$. Given the countability of the discontinuities of F, such a point will always exist and according to the definition of Y it must have the property $F(x) < \omega$. If $F_n(x) \to F(x)$, there will be n large enough that $F_n(x) < \omega$ and hence $x < Y_n(\omega)$, by definition. Therefore,

$$Y(\omega) - \varepsilon < x < Y_n(\omega). \tag{23.13}$$

Without presuming that $\lim_{n\to\infty} Y_n(\omega)$ exists, since ε is arbitrary (23.13) allows the conclusion that $\liminf_{n\to\infty} Y_n(\omega) \geq Y(\omega)$.

Next, choose y as a continuity point of F satisfying $Y(\omega) < y < Y(\omega) + \varepsilon$. The properties of F give $\omega \leq F(Y(\omega)) \leq F(y)$. For large enough n, $\omega \leq F_n(y)$ and hence, again by definition of Y_n,

$$Y_n(\omega) \leq y < Y(\omega) + \varepsilon. \tag{23.14}$$

It follows in the same way as before that $\limsup_{n\to\infty} Y_n(\omega) \leq Y(\omega)$. The superior and inferior limits are therefore equal and $\lim_{n\to\infty} Y_n(\omega) = Y(\omega)$.

This result only holds for continuity points of Y. However, there is a 1–1 correspondence between the discontinuity points of Y and intervals having zero probability under μ in $\mathbb{R}$. A collection of disjoint intervals on the line is at most countable (**1.11**) and hence the discontinuities of Y (plus the points 0 and 1) are countable and have Lebesgue measure zero. Hence, $Y_n \to Y$ w.p.1 $[m]$, as asserted. ∎

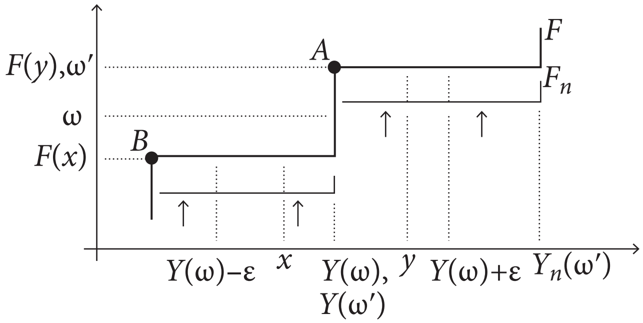

Figure 23.3

In Figure 23.3, notice how both functions take their values at the discontinuities at the points marked A and B. Thus, $F(Y(\omega)) = \omega' > \omega$. Inequality (23.14) holds for ω, but need not hold for ω', a discontinuity point. A counterexample is the sequence of functions F_n obtained by vertical translations of the fixed graph, converging to F from below, as illustrated. In this case $Y_n(\omega') > Y(\omega') + \varepsilon$ for every n.

23.7 Corollary Define random variables Y_n' so that $Y_n'(\omega) = Y_n(\omega)$ at each ω where the function is continuous and $Y_n'(\omega) = 0$ at discontinuity points and at $\omega = 0$ and 1. Define Y' similarly. If $F_n \Rightarrow F$ then $Y_n'(\omega) \to Y'(\omega)$ for every $\omega \in [0,1]$ and F_n and F are the distribution functions of Y_n' and Y'.

Proof The convergence for every ω is immediate. The equivalence of the distributions follows from **8.5**, since the discontinuity points are countable and their complement is dense in $[0,1]$, by **1.39**. ∎

In the form given, **23.6** does not generalize very easily to distributions in $\mathbb{R}^k$ for $k > 1$, although a generalization does exist. This can be deduced as a special case of **29.6**, which derives the Skorokhod representation for distributions on general metric spaces of suitable type.

A final point to observe about Skorokhod's representation is its generalization to any finite measure. If F_n is a non-decreasing right-continuous function with codomain $[a,b]$, (23.11) defines a function $Y_n(\omega)$ on a measure space $\big([a,b], \mathcal{B}_{[a,b]}, m\big)$ where m is Lebesgue measure as before. With appropriate modifications, all the foregoing remarks continue to apply in this case.

The following application of the Skorokhod representation yields a different, but equivalent, characterization of weak convergence. The necessity half of the result is known as the Helly–Bray theorem.

23.8 Theorem $X_n \to_{\mathrm{d}} X$ iff

$$\lim_{n\to\infty} \mathrm{E}\big(f(X_n)\big) = \mathrm{E}\big(f(X)\big) \tag{23.15}$$

for every bounded, continuous real function f.

Proof To prove sufficiency, construct an example. For $a \in \mathbb{R}$ and $\delta > 0$, let

$$f(x) = \begin{cases} 1, & x \le a - \delta \\ (a-x)/\delta, & a - \delta < x \le a \\ 0, & x > a. \end{cases} \tag{23.16}$$

This is the 'smoothed indicator' of the set $(-\infty, a]$ (see Figure 23.4). It is a continuous function with the properties

$$F_n(a-\delta) \le \int f\mathrm{d}F_n \le F_n(a) \tag{23.17}$$

for all n and

$$F(a-\delta) \le \int f\mathrm{d}F \le F(a). \tag{23.18}$$

By hypothesis, $\int f\mathrm{d}F_n \to \int f\mathrm{d}F$ and hence

$$\limsup_{n\to\infty} F_n(a-\delta) \le \int f\mathrm{d}F \le \liminf_{n\to\infty} F_n(a). \tag{23.19}$$

Letting $\delta \to 0$, combining (23.18) and (23.19) yields

$$\limsup_{n\to\infty} F_n(a-) \le F(a) \tag{23.20}$$

$$F(a-) \le \liminf_{n\to\infty} F_n(a). \tag{23.21}$$

These inequalities show than $\lim_n F_n(a)$ exists and is equal to $F(a)$ whenever $F(a-) = F(a)$, that is, $F_n \Rightarrow F$.

To prove necessity, let f be a bounded function whose points of discontinuity are contained in a set D_f where $\mu(D_f) = 0$, μ being the p.m. such that $F(x) = \mu((-\infty, x])$. When $F_n \Rightarrow F$ (F_n being the c.d.f. of X_n and F that of X) $Y_n'(\omega) \to Y'(\omega)$ for every $\omega \in [0,1]$, where $Y_n'(\omega)$ and $Y'(\omega)$ are the Skorokhod variables defined in **23.7**. Since $m(\omega : Y'(\omega) \in D_f) = \mu(D_f) = 0, f(Y_n') \to f(Y')$ a.s.$[\mu]$ by **19.8**(i). The bounded convergence theorem then implies $\mathrm{E}(f(Y_n')) \to \mathrm{E}(f(Y'))$, or

$$\int f(y)\mathrm{d}\mu_n(y) \to \int f(y)\mathrm{d}\mu(y) \tag{23.22}$$

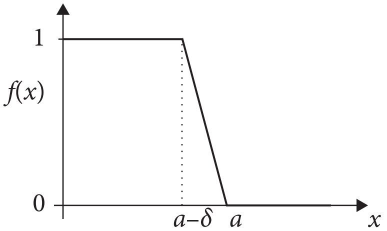

Figure 23.4

where μ_n is the p.m. corresponding to F_n. But **9.8** implies that

$$\int f(y)\mathrm{d}\mu_n(y) = \int y\mathrm{d}\mu_n f^{-1}(y) = \int x\mathrm{d}\mu_n f^{-1}(x) = \mathrm{E}\big(f(X_n)\big) \qquad (23.23)$$

with a similar equality for $\mathrm{E}(f(X))$. (The trivial change of dummy argument from y to x is just to emphasize the equivalence of the two formulations.) Hence, $\mathrm{E}(f(X_n)) \to \mathrm{E}(f(X))$. The result certainly holds for the case $D_f = \emptyset$, so 'only if' is proved. ∎

Notice how the proof cleverly substitutes $([0,1],\mathcal{B},m)$ for the fundamental probability space $(\Omega,\mathcal{F},P)$ generating $\{X_n\}$, exploiting the fact that the derived distributions are the same. This result does not say that the expectations converge *only* for bounded continuous functions; it is simply that convergence is implied at least for all members of this large class of functions. The theorem also holds for any subclass of the class of bounded continuous functions that contains at least the smoothed indicator functions of half-lines, for example the bounded uniformly continuous functions.

23.9 Example Consider Example **23.3**. The expectation of a function f defined on the discrete points of the interval is

$$\int f\mathrm{d}\mu_n = \frac{1}{n}\sum_{i=1}^{n} f(i/n). \qquad (23.24)$$

The limit of the expression on the right of (23.24) as $n \to \infty$ is by definition the Riemann integral of f on the unit interval. Since this agrees with the Lebesgue integral, weak convergence is proved in this case. □

The extension of Theorem **23.8** to general finite measures is given as a corollary, the modifications to the proof being left to the reader to supply. This is mainly a matter of modifying the notation to suit.

23.10 Corollary Let $\{F_n\}$ be a sequence of bounded, non-decreasing, right-continuous functions. $F_n \Rightarrow F$ iff

$$\int f\mathrm{d}F_n \to \int f\mathrm{d}F \qquad (23.25)$$

for every bounded, continuous real function f. □

A proof that was deferred earlier can now be given.

Proof of 23.1 To show sufficiency, consider $A = (-\infty, x]$, for which $\partial A = \{x\}$. Weak convergence is defined by the condition $\mu_n((-\infty, x]) \to \mu((-\infty, x])$ whenever $\mu(\{x\}) = 0$. To show necessity, consider in the necessity part of **23.8** the case $f(x) = 1_A(x)$ for any $A \in \mathcal{B}$. The discontinuity points of this function are contained in the set ∂A and if $\mu(\partial A) = 0$, $\mu_n(A) \to \mu(A)$ as a case of (23.15), when $F_n \Rightarrow F$. ∎

23.3 Weak Convergence and Transformations

The next result is the well-known *continuous mapping theorem*. This might be thought of as the weak convergence counterpart of **19.8**.

23.11 Theorem Let $h : \mathbb{R} \mapsto \mathbb{R}$ be Borel-measurable with discontinuity points confined to a set D_h, where $\mu(D_h) = 0$. If $\mu_n \Rightarrow \mu$, then $\mu_n h^{-1} \Rightarrow \mu h^{-1}$.

Proof Let Y_n' and Y' be defined as in **23.7**. By the argument used to prove the Helly–Bray theorem (necessity part of **23.8**), $h(Y_n') \to h(Y')$ a.s.$[\mu]$ under the stated conditions. It follows from **23.4** that $h(Y_n') \to_{\mathrm{d}} h(Y')$. Since $m(\omega : Y_n'(\omega) \in A) = \mu_n(A)$,

$$m\big(\omega : h(Y_n'(\omega)) \in A\big) = m\big(\omega : Y_n'(\omega) \in h^{-1}(A)\big) = \mu_n h^{-1}(A) \qquad (23.26)$$

for each $A \in \mathcal{B}$, using **3.28**. Similarly, $m(h(Y') \in A) = \mu h^{-1}(A)$. According to the definition of weak convergence, $h(Y_n') \to_{\mathrm{d}} h(Y')$ is equivalent to $\mu_n h^{-1} \Rightarrow \mu h^{-1}$. ∎

23.12 Corollary If h is the function of **23.11** and $X_n \to_{\mathrm{d}} X$, then $h(X_n) \to_{\mathrm{d}} h(X)$.

Proof Immediate from the theorem, given that $X_n \sim_{\mathrm{d}} \mu_n$ and $X \sim_{\mathrm{d}} \mu$. ∎

23.13 Example If $X_n \to_{\mathrm{d}} \mathrm{N}(0,1)$, then $X_n^2 \to_{\mathrm{d}} \chi^2(1)$. □

A second result on transformations is from [36] and is sometimes called Cramér's theorem:

23.14 Theorem (Cramér) If $X_n \to_{\mathrm{d}} X$ and $Y_n \to_{\mathrm{pr}} a$ (constant), then

(i) $X_n + Y_n \to_{\mathrm{d}} X + a$
(ii) $Y_n X_n \to_{\mathrm{d}} aX$
(iii) If $a \neq 0$, $X_n / Y_n \to_{\mathrm{d}} X/a$.

Proof This is by an extension of the type of argument used in **23.4**.

$$\begin{aligned} P(X_n + Y_n \leq x) &= P\big(X_n + Y_n \leq x, |Y_n - a| < \varepsilon\big) \\ &\quad + P\big(X_n + Y_n \leq x, |Y_n - a| \geq \varepsilon\big) \\ &\leq P(X_n \leq x - a + \varepsilon) + P\big(|Y_n - a| \geq \varepsilon\big). \end{aligned} \tag{23.27}$$

Similarly,

$$P(X_n \leq x - a - \varepsilon) \leq P(X_n + Y_n \leq x) + P\big(|Y_n - a| \geq \varepsilon\big) \tag{23.28}$$

and putting these inequalities together,

$$\begin{aligned} P(X_n \leq x - a - \varepsilon) - P\big(|Y_n - a| \geq \varepsilon\big) &\leq P(X_n + Y_n \leq x) \\ &\leq P(X_n \leq x - a + \varepsilon) + P\big(|Y_n - a| \geq \varepsilon\big). \end{aligned} \tag{23.29}$$

Let F_{X_n} and $F_{X_n+Y_n}$ denote the c.d.f.s of X_n and $X_n + Y_n$ respectively and let F_X be the c.d.f. of X, such that $F_X(x) = \lim_{n\to\infty} F_{X_n}(x)$ at all continuity points of F_X. Since $\lim_{n\to\infty} P(|Y_n - a| \geq \varepsilon) = 0$ for all $\varepsilon > 0$ by assumption, (23.29) implies

$$\begin{aligned} F_X(x - a - \varepsilon) &\leq \liminf_{n\to\infty} F_{X_n+Y_n}(x) \\ &\leq \limsup_{n\to\infty} F_{X_n+Y_n}(x) \leq F_X(x - a + \varepsilon). \end{aligned} \tag{23.30}$$

Taking ε arbitrarily close to zero shows that

$$\lim_{n\to\infty} F_{X_n+Y_n}(x) = F_X(x - a) = F_{X+a}(x) \tag{23.31}$$

whenever $x - a$ is a continuity point of F_x. This proves (i).

To prove (ii), suppose first that $a > 0$. Taking $\varepsilon > 0$ small enough ensures $a - \varepsilon > 0$ and applying the type of argument used in (i) with obvious variations,

$$\begin{aligned} P\big(X_n(a + \varepsilon) \leq x\big) - P\big(|Y_n - a| \geq \varepsilon\big) &\leq P(X_n Y_n \leq x) \\ &\leq P\big(X_n(a - \varepsilon) \leq x\big) + P\big(|Y_n - a| \geq \varepsilon\big). \end{aligned} \tag{23.32}$$

Taking limits gives

$$\begin{aligned} F_X\big(x/(a + \varepsilon)\big) &\leq \liminf_{n\to\infty} F_{X_n Y_n}(x) \\ &\leq \limsup_{n\to\infty} F_{X_n Y_n}(x) \leq F_X\big(x/(a - \varepsilon)\big) \end{aligned} \tag{23.33}$$

and thus

$$\lim_{n\to\infty} F_{X_nY_n}(x) = F_X(x/a) = F_{aX}(x). \tag{23.34}$$

If $a < 0$, replace Y_n by $-Y_n$ and a by $-a$, repeat the preceding argument and then apply **23.12.** And if $a = 0$, (23.32) becomes

$$\begin{aligned} P(X_n\varepsilon \le x) - P\big(|Y_n| \ge \varepsilon\big) &\le P(X_nY_n \le x) \\ &\le P(-X_n\varepsilon \le x) + P\big(|Y_n| \ge \varepsilon\big). \end{aligned} \tag{23.35}$$

For $x > 0$, this yields in the limit $F_{X_nY_n}(x) = 1$ and for $x < 0$, $F_{X_nY_n}(x) = 0$, which defines the degenerate distribution with the mass concentrated at 0. In this case $X_nY_n \to_{\text{pr}} 0$ in view of **23.5**. (Alternatively, see **19.13**.)

To prove (iii) it suffices to note by **19.9**(ii) that $\text{plim}1/Y_n = 1/a$ if $a \neq 0$. Replacing Y_n by $1/Y_n$ in (ii) yields the result directly. ∎

23.4 Convergence of Moments and Characteristic Functions

Paralleling the sequence of distribution functions, there may be sequences of moments. If $X_n \to_{\text{d}} X$ where the c.d.f. of X is F, then $\text{E}(X) = \int x\text{d}F(x)$, where it exists, is sometimes called the *asymptotic expectation* of X_n. There is a temptation to write $\text{E}(X) = \lim_{n\to\infty} \text{E}(X_n)$, but there are cases where $\text{E}(X_n)$ does not exist for any finite n while $\text{E}(X)$ exists and also cases where $\text{E}(X_n)$ exists for every n but $\text{E}(X)$ does not. This usage is therefore best avoided except in specific circumstances when the convergence is known to obtain.

Theorem **23.8** shows that expectations of bounded random variables converge under weak convergence of the corresponding measures. The following theorems indicate how far this result can be extended to more general cases. Recall that $\text{E}|X|$ is defined for every X, although it may take the value $+\infty$.

23.15 Theorem If $X_n \to_{\text{d}} X$ then $\text{E}|X| \le \liminf_{n\to\infty} \text{E}|X_n|$.

Proof The function $h_\alpha(x) = |x|1_{\{|x|\le\alpha\}}$ for $\alpha > 0$ is real and bounded. If $P\big(|X| = \alpha\big) = 0$ so that $\pm\alpha$ are continuity points of the limit distribution, it follows by **23.11** that $h_\alpha(X_n) \to_{\text{d}} h_\alpha(X)$ and from **23.8** (letting f be the identity function which is bounded in this case) that

$$\text{E}\big(h_\alpha(X)\big) = \lim_{n\to\infty} \text{E}\big(h_\alpha(X_n)\big) \le \liminf_{n\to\infty} \text{E}|X_n|. \tag{23.36}$$

The result follows on letting α approach $+\infty$ through continuity points of the distribution. ∎

The following theorem gives a sufficient condition for $\mathrm{E}(X)$ to exist, given that $\mathrm{E}(X_n)$ exists for each n.

23.16 Theorem If $X_n \to_{\mathrm{d}} X$ and $\{X_n\}$ is uniformly integrable, then $\mathrm{E}|X| < \infty$ and $\mathrm{E}(X_n) \to \mathrm{E}(X)$.

Proof Let Y_n and Y be the Skorokhod variables of (23.11), so that $Y_n \to_{\mathrm{a.s.}} Y$. Since X_n and Y_n have the same distribution, uniform integrability of $\{X_n\}$ implies that of $\{Y_n\}$. Hence invoke **12.8** to show that $\mathrm{E}(Y_n) \to \mathrm{E}(Y)$, Y being integrable. Reversing the argument then gives $\mathrm{E}|X| < \infty$ and $\mathrm{E}(X_n) \to \mathrm{E}(X)$ as required. ∎

Uniform integrability is a sufficient condition and although where it fails the existence of $\mathrm{E}(X)$ may not be ruled out, **12.7** showed that its interpretation is questionable in these circumstances.

A sequence of complex r.v.s that is always uniformly integrable is $\{\mathrm{e}^{\mathrm{i}tX_n}\}$, for any sequence $\{X_n\}$, since $|\mathrm{e}^{\mathrm{i}tX_n}| = 1$. Given the sequence of characteristic functions $\{\phi_{X_n}(t)\}$, if $F_n \Rightarrow F$ then

$$\phi_{X_n}(t) \to \phi_X(t) \tag{23.37}$$

(pointwise on $\mathbb{R}$), where the indicated limit should be the characteristic function associated with F. In view of the inversion theorem **11.18**, $X_n \to_{\mathrm{d}} X$ only if (23.37) holds where $\phi_X(t)$ is the ch.f. of X. However, it is the 'if' rather than the 'only if' that is the point of interest here. If a sequence of characteristic functions converges pointwise to a limit, under what circumstances is the limit a ch.f. in the sense that inverting it yields a c.d.f.? A sufficient condition for this is provided by the *Lévy continuity theorem*:

23.17 Theorem Suppose that $\{F_n\}$ is a sequence of c.d.f.s and $F_n \Rightarrow F$, where F is a non-negative, bounded, non-decreasing, right-continuous function. If

$$\phi_n(t) = \int_{-\infty}^{+\infty} \mathrm{e}^{\mathrm{i}tx}\mathrm{d}F_n \to \phi(t) \tag{23.38}$$

and $\phi(t)$ is continuous at the point $t = 0$, then F is a c.d.f. (i.e. $\int \mathrm{d}F = 1$) and ϕ is its ch.f. □

The fact that the conditions imposed on the limit F in this theorem are not unreasonable will be established by the Helly selection theorem, to be discussed in the next section.

Proof of 23.17 For any n, by (23.38) and the fact that F_n is a c.d.f.,

$$\phi_n(0) = \int_{-\infty}^{+\infty} \mathrm{d}F_n = 1. \tag{23.39}$$

For $\nu > 0$,

$$\frac{1}{\nu}\int_0^{\nu} \phi_n(t)\mathrm{d}t = \int_{-\infty}^{+\infty}\left(\frac{1}{\nu}\int_0^{\nu} \mathrm{e}^{\mathrm{i}tx}\mathrm{d}t\right)\mathrm{d}F_n = \int_{-\infty}^{+\infty}\left(\frac{\mathrm{e}^{\mathrm{i}\nu x}-1}{\mathrm{i}\nu x}\right)\mathrm{d}F_n \tag{23.40}$$

where the change in the order of integration is permitted by **9.36**. By assumption on $\{F_n\}$ and **23.10**, which extends to complex-valued functions by linearity of the integral,

$$\int_{-\infty}^{+\infty}\left(\frac{\mathrm{e}^{\mathrm{i}\nu x}-1}{\mathrm{i}\nu x}\right)\mathrm{d}F_n \to \int_{-\infty}^{+\infty}\left(\frac{\mathrm{e}^{\mathrm{i}\nu x}-1}{\mathrm{i}\nu x}\right)\mathrm{d}F = \frac{1}{\nu}\int_0^{\nu}\phi(t)\mathrm{d}t \tag{23.41}$$

as $n \to \infty$ where the equality is by (23.40) and the definition of ϕ in (23.38). Since ϕ is continuous at $t = 0$, by assumption, for any $\varepsilon > 0$ there is $\nu > 0$ small enough that

$$\left|\frac{1}{\nu}\int_0^{\nu}\phi(t)\mathrm{d}t - \phi(0)\right| \le \frac{1}{\nu}\int_0^{\nu}|\phi(t)-\phi(0)|\mathrm{d}t < \varepsilon \tag{23.42}$$

where the first inequality is by **11.3**. It follows by (23.42) and then by (23.41) that

$$\phi(0) = \lim_{\nu\to 0}\frac{1}{\nu}\int_0^{\nu}\phi(t)\mathrm{d}t = \int_{-\infty}^{+\infty}\lim_{\nu\to 0}\left(\frac{\mathrm{e}^{\mathrm{i}\nu x}-1}{\mathrm{i}\nu x}\right)\mathrm{d}F = \int_{-\infty}^{+\infty}\mathrm{d}F. \tag{23.43}$$

Since $\phi_n \to \phi$, it also follows from (23.39) that $\phi(0) = 1$ which in view of the other conditions imposed means F is a c.d.f. If X is a random variable having c.d.f. F then by (23.38) and **23.8** applied to the real and imaginary parts individually, $\phi(t) = \mathrm{E}(\mathrm{e}^{\mathrm{i}tX})$. ∎

The continuity theorem provides the basic justification for investigating limiting distributions by evaluating the limits of sequences of ch.f.s and then using the inversion theorem of §11.6. The next two chapters are devoted to developing these methods. Here is one useful application, a result similar to **23.4** that may also be proved as a corollary.

23.18 Theorem If $|X_n - Z_n| \to_{\mathrm{pr}} 0$ and $\{X_n\}$ converges in distribution, then $\{Z_n\}$ converges in distribution to the same limit.

Proof X_n converges and hence is $O_p(1)$, so $|\mathrm{e}^{\mathrm{i}tX_n} - \mathrm{e}^{\mathrm{i}tZ_n}| \to_{\mathrm{pr}} 0$ by **19.11**(ii). Since $|\mathrm{e}^{\mathrm{i}tX}| = 1$ these functions are L_∞-bounded and the sequence $\{|\mathrm{e}^{\mathrm{i}tX_n} - \mathrm{e}^{\mathrm{i}tZ_n}|\}$ is uniformly integrable. So by **19.15**, $|\mathrm{e}^{\mathrm{i}tX_n} - \mathrm{e}^{\mathrm{i}tZ_n}| \to_{L_1} 0$. However, the complex modulus inequality (11.15) gives

$$\mathrm{E}|\mathrm{e}^{\mathrm{i}tX_n} - \mathrm{e}^{\mathrm{i}tZ_n}| \geq |\mathrm{E}(\mathrm{e}^{\mathrm{i}tX_n}) - \mathrm{E}(\mathrm{e}^{\mathrm{i}tZ_n})| \tag{23.44}$$

so that a further consequence is $|\phi_{X_n}(t) - \phi_{Z_n}(t)| \to 0$ as $n \to \infty$, pointwise on $\mathbb{R}$. Given the assumption of weak convergence, the conclusion now follows from the inversion theorem. ∎

To get the alternative proof of **23.4**, set $Z_n = X$ for each n.

23.5 Criteria for Weak Convergence

Not every sequence of c.d.f.s has a c.d.f. as its limit. Counterexamples are easy to construct.

23.19 Example Consider the uniform distribution on the interval $[-n, n]$, such that $F_n(a) = \frac{1}{2}(1 + a/n)$, $-n \leq a \leq n$. Then $F_n(a) \to \frac{1}{2}$ for all $a \in \mathbb{R}$. □

23.20 Example Consider the degenerate r.v. $X_n = n$ w.p.1. The c.d.f. is a step function with jump at n. $F_n(a) \to 0$, all $a \in \mathbb{R}$. □

Although F_n is a c.d.f. for all n, in neither of these cases is the limit F a c.d.f., in the sense that $F(a) \to 1$ (0) as $a \to \infty$ $(-\infty)$. Nor does intuition suggest that the limiting distributions are well defined. The difficulty in the first example is that the probability mass is getting smeared out evenly over an infinite support, so that the density is tending everywhere to zero. It does not make sense to define a random variable that can take any value in $\mathbb{R}$ with equal probability, any more than it does to make a random variable infinite almost surely, which is the limiting case of the second example.

In view of these pathological cases, it is important to establish the conditions under which a sequence of measures can be expected to converge weakly. A measure μ is said to be *tight* if for every $\varepsilon > 0$ there exists a compact set K_ε of the domain such that $\mu(K_\varepsilon^c) \leq \varepsilon$. Every p.m. on $(\mathbb{R}, \mathcal{B})$ is tight, although this is not necessarily the case in more general probability spaces. (See §29.2 for details.) The condition that ensures the limit of a sequence of tight measures is well-defined is accordingly called *uniform tightness*. The sequence $\{\mu_n\}$ of p.m.s on $\mathbb{R}$ is uniformly tight if for any $\varepsilon > 0$ there exists a finite interval $(a, b]$ such that $\sup_n \mu_n((a, b]) > 1 - \varepsilon$. Equivalently, if $\{F_n\}$ is the sequence of c.d.f.s corresponding to $\{\mu_n\}$, uniform tightness is the condition that for any $\varepsilon > 0$ there exist a and b with $0 < b - a < \infty$ and

$$\sup_n \{F_n(b) - F_n(a)\} > 1 - \varepsilon. \tag{23.45}$$

It is easy to see that examples **23.19** and **23.20** both fail to satisfy the uniform tightness condition. However, provided a sequence of p.m.s $\{\mu_n\}$ is uniformly tight it does converge to a limit μ which is a p.m.

An essential ingredient in this argument is a classic result in analysis, Helly's selection theorem.

23.21 Theorem If $\{F_n\}$ is any sequence of c.d.f.s, there exists a subsequence $\{n_k, k = 1, 2, \dots\}$ such that $F_{n_k} \Rightarrow F$, where F is bounded, non-decreasing, and right-continuous and $0 \leq F \leq 1$.

Proof Consider the bounded array $\{\{F_n(x_i), n \in \mathbb{N}\}, i \in \mathbb{N}\}$ where $\{x_i, i \in \mathbb{N}\}$ is an enumeration of the rationals. By **2.36**, this array converges on a subsequence so that $\lim_{k\to\infty} F_{n_k}(x_i) = F^*(x_i)$ for every i. Note that $F^*(x_{i_1}) \leq F^*(x_{i_2})$ whenever $x_{i_1} < x_{i_2}$, since this property is satisfied by F_n for every n. Hence consider the non-decreasing function on $\mathbb{R}$,

$$F(x) = \inf_{x_i > x} F^*(x_i). \tag{23.46}$$

Clearly $0 \leq F^*(x_i) \leq 1$ for all i, since the $F_{n_k}(x_i)$ have this property for every k. By definition of F, for $x \in \mathbb{R}$ $\exists$ $x_i > x$ such that $F(x) \leq F^*(x_i) < F(x) + \varepsilon$ for any $\varepsilon > 0$, showing that F is right-continuous since $F^*(x_i) = F(x_i)$. Further, for continuity points x of F there exist $x_{i_1} < x$ and $x_{i_2} > x$ such that

$$F(x) - \varepsilon < F^*(x_{i_1}) \leq F^*(x_{i_2}) < F(x) + \varepsilon. \tag{23.47}$$

The following inequalities hold in respect of these points:

$$\begin{aligned} F^*(x_{i_1}) = \lim_{k\to\infty} F_{n_k}(x_{i_1}) &\leq \liminf_{k\to\infty} F_{n_k}(x) \\ &\leq \limsup_{k\to\infty} F_{n_k}(x) \leq \lim_{k\to\infty} F_{n_k}(x_{i_2}) = F^*(x_{i_2}). \end{aligned} \tag{23.48}$$

Combining (23.47) with (23.48),

$$F(x) - \varepsilon < \liminf_{k\to\infty} F_{n_k}(x) \leq \limsup_{k\to\infty} F_{n_k}(x) < F(x) + \varepsilon. \tag{23.49}$$

Since ε is arbitrary, $\lim_{k\to\infty} F_{n_k}(x) = F(x)$ at all continuity points of F. ∎

The only problem here is that F need not be a c.d.f., as in **23.19** and **23.20**. Tightness is the required property to ensure that $F(x) \to 1$ (0) as $x \to \infty$ $(-\infty)$.

23.22 Theorem Let $\{F_n\}$ be a sequence of c.d.f.s. If

(a) $F_{n_k} \Rightarrow F$ for every convergent subsequence $\{n_k\}$ and

(b) the sequence is uniformly tight

then $F_n \Rightarrow F$ where F is a c.d.f. Condition (b) is also necessary. □

Helly's theorem tells us that $\{F_n\}$ has a cluster point F. Condition (a) requires that this F be the *unique* cluster point, regardless of the subsequence chosen and the argument of **2.3** applied pointwise to $\{F_n\}$ implies that F is the actual limit of the sequence. Uniform tightness is necessary and sufficient for this limit F to be a c.d.f.

Proof of 23.22 Let x be a continuity point of F and suppose $F_n(x) \nrightarrow F(x)$. Then $|F_n(x) - F(x)| \geq \varepsilon > 0$ for an infinite subsequence of integers, say $\{n_k, k \in \mathbb{N}\}$. Define a sequence of c.d.f.s by $F'_k = F_{n_k}$, $k = 1,2,\ldots$. According to Helly's theorem, this sequence contains a convergent subsequence $\{k_i, i \in \mathbb{N}\}$ (say) such that $F'_{k_i} \Rightarrow F'$. But $F' \neq F$ contradicts assumption (a). Hence, $F_n \Rightarrow F$.

Since F_n is a c.d.f. for every n, $F_n(b) - F_n(a) > 1 - \varepsilon$ for some $b - a < \infty$, for any $\varepsilon > 0$. Since $F_n \to F$ at continuity points, increase b and reduce a as necessary to make them continuity points of F. Assuming uniform tightness, $F(b) - F(a) > 1 - \varepsilon$ by (23.45) as required. It follows that $\lim_{x\to\infty} F(x) = 1$ and $\lim_{x\to-\infty} F(x) = 0$. Given the monotonicity and right continuity of F established by Helly's theorem, this means that F is a c.d.f.

On the other hand, if the sequence is not uniformly tight, $F(b) - F(a) \leq 1 - \varepsilon$ for some $\varepsilon > 0$ and every $b > a$. Letting $b \to +\infty$ and $a \to -\infty$, $F(+\infty) - F(-\infty) \leq 1 - \varepsilon < 1$. Hence, either $F(+\infty) < 1$ or $F(-\infty) > 0$ or both and F is not a c.d.f. ∎

The role of the continuity theorem (**23.17**) should now be apparent. Helly's theorem ensures that the limit F of a sequence of c.d.f.s has all the properties of a c.d.f. except possibly that of $\int \mathrm{d}F = 1$. Uniform tightness ensures this property and the continuity of the limiting ch.f. at the origin can now be interpreted as a sufficient condition for tightness of the sequence. It is of interest to note what happens in the cases of counterexamples **23.19** and **23.20**. The ch.f. corresponding to **23.19** is

$$\phi_n(\nu) = \frac{1}{2n}\int_{-n}^{n} (\cos\nu x + \mathrm{i}\sin\nu x)\mathrm{d}x = \frac{\sin\nu n}{\nu n}. \tag{23.50}$$

L'Hôpital's rule shows that $\phi_n(0) = 1$ for every n whereas $\phi_n(\nu) \to 0$ as $n \to \infty$ for all $\nu \neq 0$. In the case of **23.20**,

$$\phi_n(\nu) = \cos \nu n + \mathrm{i} \sin \nu n \tag{23.51}$$

which fails to converge except at the point $\nu = 0$.

23.6 Convergence of Random Sums

Most of the important weak convergence results concern sequences of partial sums of a random array $\{X_{nt}, t = 1, \ldots, n, n \in \mathbb{N}\}$. Let

$$S_n = \sum_{t=1}^{n} X_{nt} \tag{23.52}$$

and consider the distributions of the sequence $\{S_n\}$ as $n \to \infty$. The array notation (double indexing) permits a normalization depending on n to be introduced. Central limit theorems, where typically $X_{nt} = n^{-1/2} X_t$ and S_n converges to the Gaussian limit, are examined in detail in the following chapters. However, these are not the only possibility.

23.23 Example From **8.8** the binomial$(n, \lambda/n)$ is the distribution of the sum of n independent Bernoulli random variables X_{nt}, where $P(X_{nt} = 1) = \lambda/n$ and $P(X_{nt} = 0) = 1 - \lambda/n$. From **23.2**,

$$S_n = \sum_{t=1}^{n} X_{nt} \xrightarrow{\mathrm{d}} \text{Poisson with mean } \lambda. \quad \square \tag{23.53}$$

From **11.1**, the distribution of a sum of independent r.v.s is given by the convolution of the distributions of the summands. The weak limits of independent sum distributions therefore have to be expressible as infinite convolutions, or in other words to be infinitely divisible. The properties of these distributions, specifically of their characteristic functions, were discussed in §11.5.

A certain subclass of the infinitely divisible distributions has the further property that the distribution functions F and F_n actually match, apart from possible changes of scale and location. The question is the following. If X is a drawing from distribution F and $X_1, X_2, \ldots, X_n$ a set of independent drawings from F, do there exist nonstochastic functions of n, $a_n > 0$ and b_n, such that

$$X_1 + X_2 + \cdots + X_n \overset{\mathrm{d}}{\sim} a_n X + b_n? \tag{23.54}$$

This would imply the identity

$$\phi(t)^n = \mathrm{e}^{\mathrm{i}b_n t}\phi(a_n t) \tag{23.55}$$

which imposes further restrictions on the functional form of ϕ, beyond those already detailed in §11.5. A distribution with this property is called *stable*. In the case where $b_n = 0$ the designation *strictly stable* is generally adopted. A more familiar representation of the stability relation is

$$\frac{X_1 + X_2 + \cdots + X_n - b_n}{a_n} \overset{\mathrm{d}}{\sim} X. \tag{23.56}$$

If it is the case that a normalized sum of independent and identically distributed random variables has a weak limit in distribution as the number of terms becomes large, that limit can only be a member of the stable class.

The significance of the stability property is that a stable distribution can act as an 'attractor' for sums. The following simplified story may help to explain this phenomenon. In a large number of independent drawings from some distribution, suppose that by chance a proportion p of these drawings 'have the attractor distribution'; that is to say, they would have had a relatively high probability of arising under the attractor. Perhaps p is small. However, now let pairs of independent drawings from this distribution be summed and normalized, as in (23.56) with $n = 2$. Two independent drawings that are both in the body of the attractor distribution arise with probability p^2. However, the attractor property ensures that their normalized sum falls in the same region. Of the other cases where either or both of the pair fall outside the attractor region, suppose that nonetheless a proportion p of their sums fall in the attractor region. Then, the total proportion of such sums of pairs falling in the attractor region would be $p_{(2)} = p^2 + (1 - p^2)p > p$. Now repeat the exercise: make drawings of summed pairs and form sums of pairs of these drawings, to give normalized sums of four terms. The proportion of *these* sums falling in the attractor region is $p_{(3)} = p_{(2)}^2 + (1 - p_{(2)}^2)p_{(2)} > p_{(2)}$. Iterating this doubling of sample sizes defines a sequence $\{p_{(j)}, j = 2, 3, \ldots\}$ that is of course monotone and converging to 1.

There are two key questions. The first, to be addressed in the next section, is what do these stable distributions look like? The second is what locates a distribution in the domain of attraction of a particular stable distribution? The Gaussian case is familiar, and the Gaussian domain of attraction contains, broadly, any distribution having a finite variance. Its boundaries are explored in detail in §24.5. The case of distributions lying outside the Gaussian domain is taken up in §24.6.

23.7 Stable Distributions

The original research on stable random variables was due mainly to Lévy and Khinchine in the interwar years, although the key reference to which readers are often directed for the details is the comprehensive monograph of Gnedenko and Kolmogorov ([82]). In addition to the standard texts such as Breiman ([26]), Chow and Teicher ([33]), and Feller ([74]), the specialized monographs of Ibragimov and Linnik ([105]), Zolotarev ([195]), and Uchaikin and Zolotarev ([180]) also offer excellent treatments.

It is clear that stable distributions must be infinitely divisible. Of the examples of infinitely divisible distributions listed in Section §11.5 the following are also stable.

23.24 Example The Gaussian family: by (11.41),

$$\phi(t;\mu,\sigma^2)^n = \exp\{\mathrm{i}\mu nt - \tfrac{1}{2}\sigma^2 nt^2\} = \mathrm{e}^{\mathrm{i}(n-n^{1/2})\mu t}\phi(n^{1/2}t;\mu,\sigma^2). \tag{23.57}$$

The distribution is stable with index $a_n = n^{1/2}$ and $b_n = (n - n^{1/2})\mu$. If $X_t - \mu \sim_{\mathrm{d}} \mathrm{N}(0,\sigma^2)$ for $t = 1,\dots,n$ then $n^{-1/2}\sum_{t=1}^{n}(X_t - \mu) \sim_{\mathrm{d}} \mathrm{N}(0,\sigma^2)$. □

23.25 Example The Cauchy family: by (11.42),

$$\phi(t;\nu,\delta)^n = \exp\{\mathrm{i}tn\nu - n\delta|t|\} = \phi(nt;\nu,\delta). \tag{23.58}$$

These cases are stable with $a_n = n$ and $b_n = 0$. If $X_t \sim_{\mathrm{d}} C(\nu,\delta)$ for $t = 1,\dots,n$, then $\bar{X}_n \sim_{\mathrm{d}} C(\nu,\delta)$. This result reflects the fact already noted (see **19.20**) that Cauchy variates fail to observe the law of large numbers. The sample mean does not converge to a nonstochastic limit. □

The defining property of a stable distribution is given in (23.55), and this can conveniently be restated in logarithmic form. There exist sequences a_n and b_n such that the ch.f. has the property

$$n\log\phi(t) = \log\phi(a_n t) + \mathrm{i}b_n t \tag{23.59}$$

for all t.

23.26 Theorem If $\phi(t)$ in equation (11.57) is the ch.f. of a stable distribution satisfying (23.59), then for $0 < \alpha \leq 2$ and constants $c_1 \geq 0$ and $c_2 \geq 0$,

(i) $M(x) = c_1|x|^{-\alpha}$, $x < 0$

(ii) $N(x) = -c_2 x^{-\alpha}$, $x > 0$

(iii) $a_n = n^{1/\alpha}$

(iv) $\sigma^2 = 0$ if $\alpha < 2$ and $c_1 = c_2 = 0$ if $\alpha = 2$.

Proof Making the change of variable $y = a_n x$, formula (11.57) becomes according to (23.59),

$$n\log\phi(t) = \mathrm{i}(a_n\gamma + b_n)t - \frac{\sigma^2 a_n^2 t^2}{2} + \int_{-\infty}^{0}\Big(\mathrm{e}^{\mathrm{i}ty} - 1 - \frac{\mathrm{i}ty}{1+y^2}\Big)\mathrm{d}M(y/a_n) \\ + \int_{0}^{\infty}\Big(\mathrm{e}^{\mathrm{i}ty} - 1 - \frac{\mathrm{i}ty}{1+y^2}\Big)\mathrm{d}N(y/a_n) \tag{23.60}$$

where

$$b_n = \int_{-\infty}^{0} y\Big(\frac{1}{1+y^2} - \frac{1}{1+y^2/a_n^2}\Big)\mathrm{d}M(y/a_n) \\ + \int_{0}^{\infty} y\Big(\frac{1}{1+y^2} - \frac{1}{1+y^2/a_n^2}\Big)\mathrm{d}N(y/a_n).$$

To show (i), note that comparing (11.57) with equation (23.60) implies

$$nM(y) = M(y/a_n) \tag{23.61}$$

for $y < 0$. If $M(y)$ is not identically zero (in which case there is nothing to show), (23.61) implies $M(y)$ is positive everywhere on $(-\infty, 0)$. It also implies $M(y)/n = M(a_n y)$ and hence $(m/n)M(y) = M((a_n/a_m)y)$. Defining a function on the positive rationals $A(m/n) = a_n/a_m$, this becomes

$$rM(y) = M(A(r)y) \text{ for } r \in \mathbb{Q}^+. \tag{23.62}$$

Next, define $c_1 = M(-1)$, so that $rc_1 = M(-A(r))$. Setting $u_j = -A(r_j)$ for arbitrary r_1 and r_2, (23.62) yields $c_1^{-2}M(u_1)M(u_2) = c_1^{-1}r_1 M(u_2) = c_1^{-1}M(-u_1u_2)$ or equivalently

$$c_1^{-2}M(-|u_1|)M(-|u_2|) = c_1^{-1}M(-|u_1u_2|). \tag{23.63}$$

It follows according to Lemma **2.31** that $M(u) = M(-|u|) = c_1|u|^{-\alpha}$ for $\alpha \in \mathbb{R}$ whenever $u = -A(r)$ for rational $r > 0$. Arguments from continuity allow this formulation to be extended to $u \in \mathbb{R}^-$ and since $M(-\infty) = 0$ it must be the case that $\alpha > 0$.

The argument for (ii) closely parallels that for (i) with $(0, \infty)$ replacing $(-\infty, 0)$, N replacing M, and defining $c_2 = N(1)$.

To show (iii), write (23.61) in the form

$$nc_1|y|^{-\alpha} = c_1|y/a_n|^{-\alpha} \tag{23.64}$$

with solution $a_n = n^{1/\alpha}$. The corresponding relation for N is identical except for c_2 replacing c_1. To show (iv), comparison of (11.57) with (23.60) gives $\sigma^2(n - a_n^2) = 0$. If and only if $a_n = n^{1/2}$, this has solution $\sigma^2 > 0$. The leading term in the expansion of $e^{ity} - 1 - ity/(1 + y^2)$ is of $O(y^2)$ as $y \to 0$ and hence the solution of (23.60) must accommodate the condition $\int_{-1}^{0} u^2 dM(u) < \infty$. Since

$$\mathrm{d}M(u) = -\alpha c_1|u|^{-\alpha-1}\mathrm{d}u \tag{23.65}$$

it must be the case either that $\alpha < 2$ or that $c_1 = 0$. Likewise, by the same reasoning either $\alpha < 2$ or $c_2 = 0$. ∎

In the Gaussian case it is easily seen that the formula of (23.57) matches the conditions of the theorem with $\sigma^2 > 0$. For the case $\alpha < 2$, substitute (23.65) and its counterpart for N to write the canonical form of the stable ch.f. as

$$\begin{aligned}\log\phi(t) = \mathrm{i}t\gamma + \alpha c_1 \int_{-\infty}^{0} \left(e^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{1+x^2}\right)\frac{\mathrm{d}x}{|x|^{1+\alpha}} \\ + \alpha c_2 \int_{0}^{\infty} \left(e^{\mathrm{i}tx} - 1 - \frac{\mathrm{i}tx}{1+x^2}\right)\frac{\mathrm{d}x}{x^{1+\alpha}}.\end{aligned} \tag{23.66}$$

It remains to solve the integrals in (23.66), a computation whose details are omitted but can be found spelled out for the curious in [82] and [105] *inter alia*. The final formula is as follows.

23.27 Corollary The ch.f. of an α-stable distribution with $\alpha < 2$ is

$$\phi(t; \alpha, \delta, \beta, \nu) = \exp\{\mathrm{i}\nu t - \delta|t|^{\alpha}(1 + \mathrm{i}\beta\omega(t))\} \tag{23.67}$$

where $\beta = (c_1 - c_2)/(c_1 + c_2)$,

$$\omega(t) = \begin{cases} -\tan\left(\frac{\pi\alpha}{2}\right)\operatorname{sgn}(t), & \alpha \neq 1 \\ \frac{2}{\pi}\log|t|\operatorname{sgn}(t), & \alpha = 1 \end{cases} \tag{23.68}$$

and

$$\delta = (c_1 + c_2)\begin{cases} \frac{\Gamma(2-\alpha)}{1-\alpha}\cos\left(\frac{\pi\alpha}{2}\right), & 1 < \alpha < 2 \\ \frac{\pi}{2}, & \alpha = 1 \\ \Gamma(1-\alpha)\cos\left(\frac{\pi\alpha}{2}\right), & 0 < \alpha < 1. \quad \square \end{cases} \tag{23.69}$$

The Cauchy ch.f. in (23.58) is the case of (23.67) with $\alpha = 1$ and $\beta = 0$ and bears a passing resemblance to (23.57), but this is deceptive. The term $-n\delta|t|$ is the sum of the two integrals in (23.66) with $\delta = \pi(c_1 + c_2)/2$, while the term containing σ^2 in (23.57) has vanished.

Samorodnitsky and Taqqu ([163]) denote membership of the α-stable family having the ch.f. of (23.67) by the abbreviation $S_\alpha(\delta^{1/\alpha},\beta,\nu)$.[1] The parameters ν,δ,β, and α can be viewed as having roles analogous to the mean, variance, skewness, and kurtosis of distributions possessing the corresponding moments. However, it should be noted that δ as a function of α is diverging as α approaches 2 and, as was pointed out above, $\frac{1}{2}\sigma^2$ is not as it may appear the limiting case of δ. Condition (iv) of Theorem **23.26** shows that σ vanishes when $\alpha < 2$ and δ vanishes at $\alpha = 2$, with $c_1 + c_2 = 0$. The model space has a point of discontinuity at $\alpha = 2$ and the parameters need to be interpreted in this light.

In view of (11.24) the effect of linear transformations of the random variable can be deduced from the form of (23.67). The ch.f. of $aX + b$ has scale parameter $|a|^\alpha\delta$, skewness parameter $\operatorname{sgn}(a)\beta$, and location parameter $a\nu + b$ unless $\alpha = 1$ and $\beta \neq 0$, in which case the form of $\omega(t)$ shows that the location has the additional term $-(2/\pi)\delta\beta a\log|a|$. Except in this case, a standardized form of the random variable is $(X - \nu)/\delta^{1/\alpha}$ which can be denoted in the notation of ([163]) as $S_\alpha(1,\beta,0)$, or for brevity simply as $S_\alpha(\beta)$. The case of particular interest is (23.55) which holds with $a_n = n^{1/\alpha}$ and $b_n = \nu(n - n^{1/\alpha})$, except if $\alpha = 1$ in which case $b_n = (2/\pi)\beta\delta n\log n$. Another consideration is that δ depends on α as well as $c_1 + c_2$ according to (23.69), so setting $\delta = 1$ as a scale standardization is not comparable to setting $\sigma^2 = 1$ in the Gaussian case. Setting $c_1 + c_2 = 1$ would be the most closely comparable normalization, but of course all these choices are essentially arbitrary and subordinate to units of measurement.

[1] These authors employ a different parameterization, denoting scale by $\sigma = \delta^{1/\alpha}$ by analogy with the Gaussian case.

In the case of symmetry about zero, where $\nu = \beta = 0$, the ch.f. is real-valued and reduces to the simple formula

$$\phi(t;\alpha,\delta,0,0) = \exp\{-\delta|t|^{\alpha}\}$$

of which the Cauchy (**11.13**) is the most familiar case. These distributions are designated in ([163]) as $S\alpha S$, standing for 'symmetric alpha-stable'. A notable fact the analysis has revealed is that the Gaussian family is the unique class of stable distributions that are necessarily symmetric. By contrast, while the Cauchy family (**23.25**) also has $\beta = 0$ there do exist α-stable distributions having $\alpha = 1$ and $\beta \neq 0$ with the noted quirk in the location shift rule.

Closed forms for the density functions exist only in three special cases of the α-stable family. In addition to the Gaussian and the Cauchy, the third case is the so-called *Lévy distribution* which is defined as follows.

23.28 Example The Lévy family of distributions with location and scale parameters ν and δ are the cases of (23.67) with $\alpha = \frac{1}{2}$ and $\beta = 1$. The family is supported on $[\nu,\infty)$ with p.d.f.

$$f(x;\nu,\delta) = \frac{\delta}{\sqrt{2\pi}} e^{-\delta^2/2(x-\nu)}(x-\nu)^{-3/2}. \quad \square \tag{23.70}$$

The designation notwithstanding, the α-stable class are also sometimes referred to collectively as 'Lévy distributions', so care needs to be taken with the terminology here. In the standard case of **23.28** with $\nu = 0$ and $\delta = 1$, the Lévy is the distribution of Z^{-2} where $Z \sim_{\rm d} N(0,1)$. Thus, $P(X < x) = P(Z^{-2} < x) = 2P(Z > x^{-1/2})$ and the derivative of

$$F(x) = \sqrt{\frac{2}{\pi}} \int_{x^{-1/2}}^{\infty} e^{-z^2/2} dz$$

can be shown to be the particular case of (23.70). Like the Cauchy, this distribution has no moments. The further interesting feature is that the mean of n independent drawings not merely fails to converge but diverges like $O(n)$.

However, with only a single exception, all stable distributions are known to be continuous and unimodal, possessing derivatives of all orders. The exception is the degenerate (sometimes called improper) distribution that assigns probability 1 to a single point, say $P(X = \nu) = 1$. This distribution has ch.f. $\phi_X(t) = e^{i\nu t}$ which can be seen to belong to the class (23.67) with $\delta = 0$.

Concerning the existence of moments, the Gaussian is the only member of the α-stable class having a finite variance and it possesses all the higher moments too. An α-stable r.v. X with $\alpha < 2$ has $\mathrm{E}|X|^{\gamma} < \infty$ if and only if $\gamma < \alpha$. Members of the class with $1 < \alpha \leq 2$ possess a mean, equal to ν, but the Cauchy and all other cases having $\alpha \leq 1$ are not integrable. When $\mathrm{E}|X| = \infty$ the variance of X may also be said to be infinite but the expected value of X is undefined, given that $\infty - \infty$ is not a number.

24
The Classical Central Limit Theorem

24.1 The I.I.D. Case

The ‘normal law of error’ is justly the most famous result in statistics and to the susceptible mind has an almost mystical fascination. If a sequence of random variables $\{X_t\}_1^\infty$ have means of zero and the partial sums $\sum_{t=1}^n X_t$, $n = 1, 2, 3, \ldots$ have variances s_n^2 tending to infinity with n although finite for each finite n, then, subject to rather mild additional conditions on the distributions and the sampling process,

$$S_n = \frac{1}{s_n}\sum_{t=1}^{n} X_t \xrightarrow{\mathrm{d}} \mathrm{N}(0,1). \tag{24.1}$$

This is the central limit theorem (CLT). Establishing sets of sufficient conditions is the main business of this chapter and the next, but before getting into the formal results it might be of interest to illustrate the operation of the CLT as an approximation theorem. Particularly if the distribution of the X_t is symmetric, the approach to the limit can be very rapid.

24.1 Example The distribution of the sum of two independent $U[0,1]$ drawings was derived in **11.2**. Similarly, the sum of three such drawings has density

$$f_{X+Y+Z}(\omega) = \int_0^1 \int_0^1 1_{\{w-z-1,w-z\}}\mathrm{d}y\mathrm{d}z \tag{24.2}$$

which is plotted in Figure 24.1. This function is actually piecewise quadratic (the three segments are on [0,1], [1,2], and [2.3] respectively), but it lies remarkably close to the density of the Gaussian r.v. having the same mean and variance as $X + Y + Z$ (also plotted). The sum of 10 or 12 independent uniform r.v.s is almost indistinguishable from a Gaussian variate; indeed, the formula $S = \sum_{i=1}^{12} X_i - 6$, which has mean 0 and variance 1 when $X_i \sim_{\mathrm{d}} U[0,1]$ and independent, provides a simple and perfectly adequate device for simulating a standard Gaussian variate in computer modelling exercises. □

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.

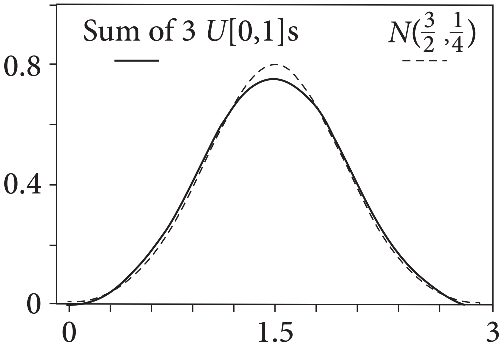

Figure 24.1

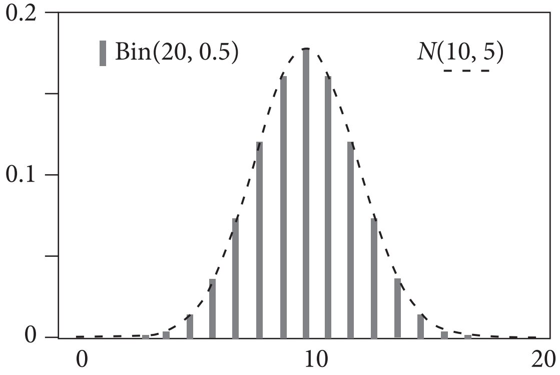

Figure 24.2

24.2 Example For a contrast in the *manner* of convergence consider the sum of n Bernoulli(p) variates for fixed $p \in (0,1)$ defining the binomial(n,p) distribution. The probabilities for $p = \frac{1}{2}$ and $n = 20$ are plotted in Figure 24.2, with the Gaussian density with matching mean and variance. The distributions are of course discrete for every finite n and continuous only in the limit. The correspondence of the ordinates is remarkably close, although remember that for $p \neq \frac{1}{2}$ the binomial distribution is not symmetric and the convergence is correspondingly slower. This example should be compared with **23.2**, the non-Gaussian limit in the latter case being obtained by having p decline as a function of n. □

Proofs of the CLT, like proofs of stochastic convergence, depend on establishing properties for certain *non*stochastic sequences. Previously, sample points $|X_n(\omega) - X(\omega)|$ for $\omega \in C$ with $P(C) = 1$, probabilities $P(|X_n - X| > \varepsilon)$, and moments $\mathrm{E}|X_n - X|^p$ were considered as sequences to be shown to converge to 0, to establish the convergence of X_n to X, respectively a.s., in pr., or in L_p. In the present case the expectations of certain functions of the S_n are considered, the key result being Theorem **23.8**. The practical trick is to find functions that will fingerprint the limiting distribution conclusively. The characteristic function is by common consent the convenient choice, exploiting the multiplicative property for independent sums, but this is not the only possible method. The reader can find an alternative approach in Pollard ([147]: III.4), for example.

The simplest case is where the sequence $\{X_t\}$ is both stationary and independently drawn.

24.3 Theorem (Lindeberg–Lévy) If $\{X_t\}_1^\infty$ is an i.i.d. sequence having zero mean and variance σ^2,

$$S_n = n^{-1/2}\sum_{t=1}^{n} X_t/\sigma \xrightarrow{\mathrm{d}} \mathrm{N}(0,1). \tag{24.3}$$

Proof The ch.f.s $\phi_X(\lambda)$ of X_t are identical for all t,[1] so from (11.24) and (11.27),

$$\phi_{S_n}(\lambda) = \left(\phi_X(\lambda\sigma^{-1}n^{-1/2})\right)^n. \tag{24.4}$$

Applying **11.5** with $k = 2$ yields the expansion

$$|\phi_X(\lambda\sigma^{-1}n^{-1/2}) - 1 - \lambda^2/2n| \le \mathrm{E}\min\left\{\frac{(\lambda X_t)^2}{\sigma^2 n}, \frac{|\lambda X_t|^3}{6\sigma^3 n^{3/2}}\right\} \tag{24.5}$$

which makes it possible to write, for fixed λ,

$$\phi_X(\lambda\sigma^{-1}n^{-1/2}) = 1 - \lambda^2/2n + O(n^{-3/2}). \tag{24.6}$$

Since $(1 + a/n)^n \to \mathrm{e}^a$ as $n \to \infty$, raising (24.6) to the n^{th} power with $a = -\frac{1}{2}\lambda^2 + O(n^{-1/2})$ yields

$$\lim_{n\to\infty} \phi_{S_n}(\lambda) = \mathrm{e}^{-\lambda^2/2}. \tag{24.7}$$

Comparing this formula with (11.40) and appealing to the inversion theorem (**11.18**) establishes the limiting distribution to be standard Gaussian. ∎

Be careful to note how the existence of $\mathrm{E}|X|^3$ is not necessary for the expansion in (24.6) to hold. The 'min' function whose expectation appears on the majorant

[1] λ is used here as the argument of the ch.f. instead of the t used in Chapters 11 and 23 to avoid confusion with the time subscript.

side of (24.5) is unquestionably of $O(n^{-3/2})$, but also integrable for each n on the assumption of finite variance.

The Lindeberg–Lévy theorem imposes strong assumptions, but offers the benefit of a simple and transparent proof. All the key features of the central limit property are discernible. In (24.6) the expansion of the ch.f. of $n^{-1/2}X_t$ consists either of terms common to every centred distribution with finite variance, or of terms that can be neglected asymptotically, a fact that ensures that the limiting sum distribution is invariant to the component distributions. The imaginary part of $\phi_X(\lambda\sigma^{-1}n^{-1/2})$ is of smaller order than the real part, which requires a symmetric limit according to the remarks following (11.25). The coincidence of these facts with the fact that the centred Gaussian is the only stable symmetric distribution having a second moment appears to rule out any alternative to the central limit property under the specified conditions, that is, zero mean and finite variance. The earlier remark that symmetry of the distribution of $n^{-1/2}X_t$ improves the rate of convergence to the limit can be also appreciated here. Assuming $\mathrm{E}(X_t^3) = 0$, the expansion in (24.5) can be taken to third order and the remainder in (24.6) is of $O(n^{-2})$.

On the other hand, if the variance does not exist the expansion of (24.6) fails. Indeed, in the specific case in which the X_t are centred Cauchy, $n^{1/2}\bar{X}_n = O(n^{1/2})$, the sequence of distributions of $\{n^{1/2}\bar{X}_n\}$ is not tight and there is no weak convergence. The limit law for the sum is itself Cauchy under the appropriate normalization with n^{-1}. These issues are revisited in §24.6.

The distinction between convergence in distribution and convergence in probability, and in particular the fact that the former does not imply the latter, can be demonstrated here by means of a counterexample. Consider the sequence $\{X_t\}_1^\infty$ defined in the statement of the Lindeberg–Lévy theorem and the corresponding S_n in (24.3).

24.4 Theorem S_n does not converge in probability.

Proof If it was true that $\operatorname{plim}_{n\to\infty} S_n = Z$, it would also be the case that $\operatorname{plim}_{n\to\infty} S_{2n} = Z$, implying

$$\operatorname*{plim}_{n\to\infty}(S_{2n} - S_n) = 0. \tag{24.8}$$

To show that (24.8) is false, write

$$S_{2n} = (2n)^{-1/2}\left(\sum_{t=1}^{n} X_t/\sigma + \sum_{t=n+1}^{2n} X_t/\sigma\right) = (S_n + S_n')/\sqrt{2} \tag{24.9}$$

where $S_n' = n^{-1/2}\sum_{t=n+1}^{2n} X_t/\sigma$ and hence

$$S_{2n} - S_n = S_n\big(1/\sqrt{2} - 1\big) + S_n'/\sqrt{2}. \tag{24.10}$$

According to the Lindeberg–Lévy theorem, $\phi_{S_n}(\lambda) \to \exp\{-\frac{1}{2}\lambda^2\}$ and $\phi_{S_n'}(\lambda) \to \exp\{-\frac{1}{2}\lambda^2\}$. Since no X_t is contained in both sums, S_n and S_n' are independent, each with mean zero and variance 1. Noting $(1/\sqrt{2}-1)^2 + (1/\sqrt{2})^2 = 2 - \sqrt{2}$ and applying the properties of ch.f.s,

$$\phi_{S_{2n}-S_n}(\lambda) \to \exp\big\{-\tfrac{1}{2}\lambda^2(2-\sqrt{2})\big\}. \tag{24.11}$$

In other words

$$S_{2n} - S_n \xrightarrow{\mathrm{d}} N\big(0, 2-\sqrt{2}\big), \tag{24.12}$$

which is the required contradiction of (24.8). ∎

Compare the sequence $\{S_n(\omega)\}_1^\infty$ with, say, $\{X(\omega) + Y(\omega)/n\}_1^\infty$, $\omega \in \Omega$, where X and Y are random variables. For given ω, the latter sequence converges to a fixed limit $X(\omega)$. On the other hand, each new contribution to S_n has *equal* weight with the others, ensured by rescaling the sum as the sample size increases. For given ω, $S_n(\omega) = n^{-1/2}\sum_{t=1}^n X_t(\omega)$ is *not* a convergent sequence, for as **24.4** shows $S_{2n}(\omega)$ is not necessarily close to $S_n(\omega)$ no matter how large n becomes. Weak convergence of the distribution functions does not imply convergence of the random sequence. The calculation in (24.10) has the implication that if the sample size n is doubled, the change in the normalized sum in (24.3) is a random variable with variance $2 - \sqrt{2} \approx 0.58$. Note, this is true whether n equals 1 or 1 million!

Characteristic function-based arguments can also be used to show convergence in distribution to a degenerate limit. What is often called 'Khinchine's theorem' is a well-known derivation of the weak law of large numbers for i.i.d. sequences that circumvents the need to show L_p convergence, even for $p = 1$.

24.5 Theorem If $\{X_t\}_1^\infty$ is an identically and independently distributed sequence with finite mean μ, then $\bar{X}_n = n^{-1}\sum_{t=1}^n X_t \to_{\text{pr}} \mu$.

Proof The characteristic function of $\bar{X}_n$ has the form

$$\phi_{\bar{X}n}(\lambda) = \big(\phi_X(\lambda/n)\big)^n \tag{24.13}$$

where by application of the argument used for **24.3**, $\phi_X(\lambda/n) = 1 + \mathrm{i}\lambda\mu/n + O(\lambda^2/n^2)$. By analogy with (24.7),

$$\lim_{n\to\infty} \phi_{\bar{X}_n}(\lambda) = \mathrm{e}^{\mathrm{i}\lambda\mu} \tag{24.14}$$

whereas $\mathrm{E}(\mathrm{e}^{\mathrm{i}\lambda X}) = \mathrm{e}^{\mathrm{i}\lambda\mu}$ only in the case where $X = \mu$ with probability 1. The distribution is degenerate and convergence in probability follows from **23.5**. ∎

24.2 Independent Heterogeneous Sequences

The Lindeberg–Lévy theorem imposes conditions that are too strong for the result to have wide practical applications in econometrics. In the remainder of this chapter the assumption of independence is retained (to be relaxed in Chapter 25) but the summands are allowed to have different distributions. In this theory it is convenient to work with normalized variables such that the partial sums always have unit variance, which entails a double indexing scheme. Define the triangular array

$$\{X_{nt}, t = 1, \dots, n, n \in \mathbb{N}\}$$

with elements having zero means and variances σ_{nt}^2, such that if

$$S_n = \sum_{t=1}^{n} X_{nt} \tag{24.15}$$

then (under independence)

$$\mathrm{E}(S_n^2) = \sum_{t=1}^{n} \sigma_{nt}^2 = 1. \tag{24.16}$$

Typically, $X_{nt} = (Y_t - \mu_t)/s_n$, where $\{Y_t\}$ is the raw sequence under study with means $\{\mu_t\}$ and $s_n^2 = \sum_{t=1}^{n} \mathrm{E}(Y_t - \mu_t)^2$. In this case $\sigma_{nt}^2 = \mathrm{E}(Y_t - \mu_t)^2/s_n^2$, so that these variances sum to unity by construction. It is also possible to have $X_{nt} = (Y_{nt} - \mu_{nt})/s_n$, the double indexing of the mean arising in situations where the sequence depends on a parameter whose value in turn depends on n. This case arises, for example, in the study of the limiting distributions of test statistics under a sequence of 'local' deviations from the null hypothesis, the device known as Pitman drift.

The existence of each variance σ_{nt}^2 is going to be a necessary baseline condition in all the theorems, just as the existence of the common variance σ^2 was required in the Lindeberg–Lévy theorem. However, with heterogeneity not even uniformly bounded variances are sufficient to get a central limit result. If the Y_t are identically distributed there is no possibility of a small (i.e. finite) number of members of the sequence exhibiting such extreme behaviour as to influence the distribution of the sum as a whole, even in the limit, but in a heterogeneous sequence this is possible. This could interfere with convergence to the normal which in most cases depends on the contribution of each individual member of the sequence being negligible.

The standard result for independent, non-identically distributed sequences is the Lindeberg–Feller theorem, which establishes that a certain condition on the distributions of the summands is sufficient and in some circumstances also necessary. Lindeberg is credited with the sufficiency part and Feller the necessity, to be dealt with in the next section.

24.6 Theorem Let the array $\{X_{nt}\}$ be independent with zero mean and variance sequence $\{\sigma_{nt}^2\}$ satisfying (24.16). Then, $S_n \to_{\mathrm{d}} \mathrm{N}(0,1)$ if for all $\varepsilon > 0$

$$\lim_{n\to\infty}\sum_{t=1}^{n}\int_{\{|X_{nt}|>\varepsilon\}} X_{nt}^2 \mathrm{d}P = 0. \quad \square \tag{24.17}$$

Equation (24.17) is known as the Lindeberg condition.

The proof of the Lindeberg theorem requires a couple of purely mechanical lemmas.

24.7 Lemma If $x_1, \dots, x_n$ and $y_1, \dots, y_n$ are collections of complex numbers with $|x_t| \le 1$ and $|y_t| \le 1$ for $t = 1, \dots, n$, then

$$\left|\prod_{t=1}^{n} x_t - \prod_{t=1}^{n} y_t\right| \le \sum_{t=1}^{n} |x_t - y_t|. \tag{24.18}$$

Proof For $n = 2$,

$$\begin{aligned} |x_1x_2 - y_1y_2| &= |(x_1 - y_1)x_2 + (x_2 - y_2)y_1| \\ &\le |x_1 - y_1||x_2| + |x_2 - y_2||y_1| \\ &\le |x_1 - y_1| + |x_2 - y_2|. \end{aligned} \tag{24.19}$$

The general case follows easily by induction. ∎

24.8 Lemma If z is a complex number and $|z| \le \frac{1}{2}$, then $|\mathrm{e}^z - 1 - z| \le |z|^2$.

Proof Using the power series expansion of e^z and the triangle inequality,

$$|e^z - 1 - z| = \left|\sum_{j=2}^{\infty} \frac{z^j}{j!}\right| = \left|z^2 \sum_{j=0}^{\infty} \frac{z^j}{(j+2)!}\right| \le |z|^2 \sum_{j=0}^{\infty} \frac{|z|^j}{(j+2)!}. \tag{24.20}$$

Since $\sum_{j=0}^{\infty} 2^{-j} = 2$, the infinite series on the right-hand side cannot exceed 1. ∎

Proof of 24.6 To show that $\phi_{S_n}(\lambda) \to e^{-\lambda^2/2}$ as $n \to \infty$, note that the difference is bounded by

$$\begin{aligned} |\phi_{S_n}(\lambda) - e^{-\lambda^2/2}| &= \left|\prod_{t=1}^{n} \phi_{X_{nt}}(\lambda) - \prod_{t=1}^{n} e^{-\lambda^2\sigma_{nt}^2/2}\right| \\ &\le \left|\prod_{t=1}^{n} \phi_{X_{nt}}(\lambda) - \prod_{t=1}^{n}(1 - \tfrac{1}{2}\lambda^2\sigma_{nt}^2)\right| \\ &\quad + \left|\prod_{t=1}^{n} e^{-\lambda^2\sigma_{nt}^2/2} - \prod_{t=1}^{n}(1 - \tfrac{1}{2}\lambda^2\sigma_{nt}^2)\right| \end{aligned} \tag{24.21}$$

where the equality is by definition using the fact that $\sum_{t=1}^{n} \sigma_{nt}^2 = 1$ and the inequality is the triangle inequality. The proof is completed by showing that both right-hand-side terms converge to zero.

Writing the integral in (24.17) in the form $E(1_{\{|X_{nt}|>\varepsilon\}}X_{nt}^2)$, notice that for $1 \le t \le n$,

$$\begin{aligned} \sigma_{nt}^2 &= E(1_{\{|X_{nt}|\le\varepsilon\}}X_{nt}^2) + E(1_{\{|X_{nt}|>\varepsilon\}}X_{nt}^2) \\ &\le \varepsilon^2 + E(1_{\{|X_{nt}|>\varepsilon\}}X_{nt}^2) \\ &\to \varepsilon^2 \text{ as } n \to \infty \end{aligned} \tag{24.22}$$

since the Lindeberg condition implies that the second term on the right-hand side of the inequality in (24.22) must converge to zero. Since ε can be chosen arbitrarily small, this shows that

$$\max_{1\le t\le n} \sigma_{nt}^2 \to 0. \tag{24.23}$$

In the first of the two terms on the majorant side of (24.21), the ch.f.s do not exceed 1 in modulus and $|1 - \frac{1}{2}\lambda^2\sigma_{nt}^2| \le 1$ for any fixed value of λ when n is large enough. Hence by **24.7**, for large enough n,

$$\left|\prod_{t=1}^{n} \phi_{X_{nt}}(\lambda) - \prod_{t=1}^{n}(1 - \tfrac{1}{2}\lambda^2\sigma_{nt}^2)\right| \le \sum_{t=1}^{n} |\phi_{X_{nt}}(\lambda) - (1 - \tfrac{1}{2}\lambda^2\sigma_{nt}^2)|. \tag{24.24}$$

To break down the terms on the majorant side of (24.24), note that **11.5** for the case $k = 2$, combined with (11.34), yields

$$\begin{aligned}|\phi_{X_{nt}}(\lambda) - (1 - \tfrac{1}{2}\lambda^2\sigma_{nt}^2)| &\le \mathrm{E}\big(\min\{(\lambda X_{nt})^2, \tfrac{1}{6}|\lambda X_{nt}|^3\}\big)\\ &\le \lambda^2\mathrm{E}(1_{\{|X_{nt}|>\varepsilon\}}X_{nt}^2) + \tfrac{1}{6}|\lambda|^3\varepsilon\sigma_{nt}^2 \qquad (24.25)\end{aligned}$$

for any $\varepsilon > 0$. Hence, recalling $\sum_{t=1}^{n}\sigma_{nt}^2 = 1$,

$$\begin{aligned}\sum_{t=1}^{n}|\phi_{X_{nt}}(\lambda) - (1 - \tfrac{1}{2}\lambda^2\sigma_{nt}^2)| &\le \lambda^2\sum_{t=1}^{n}\mathrm{E}(1_{\{|X_{nt}|>\varepsilon\}}X_{nt}^2) + \tfrac{1}{6}|\lambda|^3\varepsilon\\ &\to \tfrac{1}{6}|\lambda|^3\varepsilon \text{ as } n \to \infty \qquad (24.26)\end{aligned}$$

since the first majorant-side term vanishes by the Lindeberg condition. Since ε is arbitrary, this limit can be made as small as desired.

Similarly, for the second term of (24.21), take n large enough so that

$$\begin{aligned}\left|\prod_{t=1}^{n}\mathrm{e}^{-\lambda^2\sigma_{nt}^2/2} - \prod_{t=1}^{n}(1 - \tfrac{1}{2}\lambda^2\sigma_{nt}^2)\right| &\le \sum_{t=1}^{n}|\mathrm{e}^{-\lambda^2\sigma_{nt}^2/2} - 1 - \tfrac{1}{2}\lambda^2\sigma_{nt}^2|\\ &\le \tfrac{1}{4}\lambda^4\sum_{t=1}^{n}\sigma_{nt}^4. \qquad (24.27)\end{aligned}$$

where the second inequality is got by setting $z = -\frac{1}{2}\lambda^2\sigma_{nt}^2$ (a real number, actually) in **24.8**. However,

$$\sum_{t=1}^{n}\sigma_{nt}^4 \le \Big(\max_{1\le t\le n}\sigma_{nt}^2\Big)\sum_{t=1}^{n}\sigma_{nt}^2 = \max_{1\le t\le n}\sigma_{nt}^2 \to 0 \text{ as } n \to \infty \qquad (24.28)$$

by (24.23). ∎

The Lindeberg condition is subtle and its implications for the behaviour of random sequences call for careful interpretation. In the case $X_{nt} = X_t/s_n$ where X_t has mean 0 and variance σ_t^2 and $s_n^2 = \sum_{t=1}^{n}\sigma_t^2$, the Lindeberg condition becomes

$$\lim_{n\to\infty}\frac{1}{s_n^2}\sum_{t=1}^{n}\int_{\{|X_t|>s_n\varepsilon\}}X_t^2\mathrm{d}P = 0 \text{ for all } \varepsilon > 0. \qquad (24.29)$$

One point easily verified is that, when the summands are identically distributed, $s_n = \sqrt{n}\sigma$ and (24.29) reduces to $\lim_{n\to\infty}\sigma^{-2}\mathrm{E}(X_1^2 1_{\{|X_1|>\sigma\sqrt{n}\varepsilon\}}) = 0$. The Lindeberg condition then holds if and only if X_1 has finite variance, so that the Lindeberg theorem contains the Lindeberg–Lévy as a special case.

The problematic cases that the Lindeberg condition is designed to exclude are those where the behaviour of a finite subset of sequence elements dominates all the others, even in the limit. This can occur either by the sequence becoming excessively disorderly in the limit, or (the other side of the same coin, really) by its being not disorderly enough, beyond a certain point.

Thus, the condition clearly fails if the variance sequence $\{\sigma_t^2\}$ is tending to zero in such a way that $s_n^2 = \sum_{t=1}^n \sigma_t^2$ is bounded in n. On the other hand, if $s_n^2 \to \infty$ then $s_n\varepsilon \to \infty$ for any fixed positive ε and the Lindeberg condition resembles a condition of 'average' uniform integrability of $\{X_t^2\}$. The sum of the terms $\mathrm{E}(1_{\{|X_t|>s_n\varepsilon\}}X_t^2)$ must grow less fast than s_n^2, no matter how close ε is to zero. The following is a counterexample (compare **12.7**).

24.9 Example Let $X_t = Y_t - \mathrm{E}(Y_t)$ where $Y_t = 0$ with probability $1 - t^{-2}$ and $Y_t = t$ with probability t^{-2}. Thus $\mathrm{E}(Y_t) = t^{-1}$ and X_t is converging to a degenerate r.v., equal to 0 with probability 1, although $\mathrm{Var}(Y_t) = 1 - t^{-2}$ for every t. The Lindeberg condition fails here. $s_n^2 = n - \sum_{t=1}^n t^{-2}$ and $s_n\varepsilon < t$ for $0 < \varepsilon \le 1$ when $t > n^{1/2}$. In this range the Lindeberg integral takes the value $(t - t^{-1})^2 t^{-2}$ and

$$\frac{1}{s_n^2}\sum_{t=1}^n \int_{\{|X_t|>s_n\varepsilon\}} X_t^2 \mathrm{d}P \ge \frac{1}{n - \sum_{t=1}^n t^{-2}} \sum_{t=[\sqrt{n}]+1}^{n} \frac{(t-t^{-1})^2}{t^2} \to 1 \qquad (24.30)$$

as $n \to \infty$. It is easy to see that the CLT fails. For any $n_0 \ge 1$ put $B_0 = \sum_{t=1}^{n_0} t$ and

$$\sup_n P\Big(\sum_{t=1}^n Y_t > B_0\Big) \le P\Big(\bigcup_{t=n_0+1}^{\infty} \{Y_t > 0\}\Big) \le \sum_{t=n_0+1}^{\infty} t^{-2} \qquad (24.31)$$

where the majorant side can be made as small as desired by choosing n_0 large enough. Hence $\sum_{t=1}^n Y_t = O_p(1)$ and, noting that $\sum_{t=1}^n \mathrm{E}(Y_t) = O(\log n)$, it follows that $\sum_{t=1}^n X_t/s_n \to_{\mathrm{pr}} 0$. □

Uniform square-integrability is neither sufficient nor necessary for the Lindeberg condition, so parallels must be drawn with caution. However, the following theorem gives a simple sufficient condition.

24.10 Theorem The Lindeberg condition (24.29) holds if

(a) $\{X_t^2\}$ is uniformly integrable and

(b) $s_n^2/n \ge B > 0$ for all n.

Proof For any n and $\varepsilon > 0$, (b) implies

$$\frac{1}{s_n^2}\sum_{t=1}^{n} \mathrm{E}(1_{\{|X_t|>s_n\varepsilon\}}X_t^2) \leq \frac{1}{B}\max_{1\leq t\leq n}\{\mathrm{E}(1_{\{|X_t|>s_n\varepsilon\}}X_t^2)\}. \tag{24.32}$$

Hence

$$\begin{aligned}\lim_{n\to\infty}\frac{1}{s_n^2}\sum_{t=1}^{n} \mathrm{E}(1_{\{|X_t|>s_n\varepsilon\}}X_t^2) &\leq \frac{1}{B}\limsup_{n\to\infty}\max_{1\leq t\leq n}\{\mathrm{E}(1_{\{|X_t|>s_n\varepsilon\}}X_t^2)\} \\ &\leq \frac{1}{B}\sup_t \lim_{n\to\infty} \mathrm{E}(1_{\{|X_t|>s_n\varepsilon\}}X_t^2) \\ &= 0 \end{aligned} \tag{24.33}$$

where the last equality follows by (a) since $s_n \to \infty$. ∎

There is no assumption here that the sequence is independent. The conditions involve only the sequence of marginal distributions of the variables. Nonetheless they are stronger than necessary and a counterexample as well as further discussion of the conditions appears in §24.4.

The following, known as Liapunov's theorem, provides a popular sufficient condition for the CLT in independent processes.

24.11 Theorem A sufficient condition for (24.17) is

$$\lim_{n\to\infty}\sum_{t=1}^{n}\mathrm{E}|X_{nt}|^{2+\delta} = 0 \text{ for some } \delta > 0. \tag{24.34}$$

Proof For $\delta > 0$ and $\varepsilon > 0$,

$$\begin{aligned}\mathrm{E}|X_{nt}|^{2+\delta} &\geq \mathrm{E}(1_{\{|X_{nt}|>\varepsilon\}}|X_{nt}|^{2+\delta}) \\ &\geq \varepsilon^{\delta}\mathrm{E}(1_{\{|X_{nt}|>\varepsilon\}}X_{nt}^2). \end{aligned} \tag{24.35}$$

The theorem follows since, if $\varepsilon^{\delta}\lim_{n\to\infty}\sum_{t=1}^{n}\mathrm{E}(1_{\{|X_{nt}|>\varepsilon\}}X_{nt}^2) = 0$ for fixed $\varepsilon > 0$, then the same holds with ε^{δ} replaced by 1. ∎

Condition (24.34) is called the *Liapunov condition,* although this term is also used to refer to Liapunov's original result in which the condition was cast in terms of integer moments, that is,

$$\lim_{n\to\infty}\sum_{t=1}^{n}\mathrm{E}|X_{nt}|^3 = 0. \tag{24.36}$$

Although stronger than necessary, the Liapunov condition has the advantage of being more easily checkable at least in principle than the Lindeberg condition, as the following example illustrates.

24.12 Theorem Liapunov's condition holds for the array $\{X_t/s_n\}$ if $s_n^2/n > 0$ uniformly in n and $\mathrm{E}|X_t|^{2+\delta} < \infty$ uniformly in t, for $\delta > 0$.

Proof Under the stated conditions there exist constants $B_1 < \infty$ and $B_2 > 0$ such that $\sup_t E|X_t|^{2+\delta} \le B_1$ and $B_2 \le \inf_n s_n^2/n$. Then for every n

$$\frac{1}{s_n^2}\sum_{t=1}^{n}\mathrm{E}|X_t|^{2+\delta} \le \frac{nB_1}{s_n^2} \le \frac{B_1}{B_2} < \infty \tag{24.37}$$

and

$$\lim_{n\to\infty}\frac{1}{s_n^{2+\delta}}\sum_{t=1}^{n}\mathrm{E}|X_t|^{2+\delta} = 0 \tag{24.38}$$

follows immediately. ∎

Note that these conditions imply those of **24.10**, by **12.12**. It is sufficient to avoid the 'knife-edge' condition in which variances exist but no moments even fractionally higher, provided the sum of those variances is also $O(n)$.

24.3 Feller's Theorem and Asymptotic Negligibility

The Lindeberg condition is sufficient and sometimes also necessary. The following result, known as Feller's theorem, specifies the side condition that implies necessity.

24.13 Theorem Let $\{X_{nt}\}$ be an independent array with zero mean and variance array $\{\sigma_{nt}^2\}$. If $S_n = \sum_{t=1}^{n} X_{nt} \to_{\mathrm{d}} \mathrm{N}(0,1)$ and

$$\max_{1\le t\le n} P\big(|X_{nt}| > \varepsilon\big) \to 0 \text{ as } n \to \infty, \text{ any } \varepsilon > 0, \tag{24.39}$$

the Lindeberg condition must hold. □

The proof of Feller's theorem is rather fiddly and mechanical, but necessary conditions are rare and difficult to obtain in this theory and it is worth a little study for that reason alone. Several of the arguments are of the type used already in the sufficiency part.

Proof of 24.13 Since $\sigma_{nt}^2 \to 0$ for every t, the series expansion of the ch.f. suggests that $|\phi_{X_{nt}}(\lambda) - 1|$ converges to zero for each t. In fact, the sum of the squares of these terms converges and this is the first step in the proof. Applying **11.5** with $k = 0$ gives

$$|\phi_{X_{nt}}(\lambda) - 1| \leq \mathrm{E}\big(\min\{2, |\lambda X_{nt}|\}\big) \leq 2P\big(|X_{nt}| > \varepsilon\big) + \varepsilon|\lambda| \tag{24.40}$$

for $\varepsilon > 0$ where the second inequality in (24.40) is by the first case of (11.34). Also, since $\mathrm{E}(X_{nt}) = 0$, applying **11.5** with $k = 1$ gives

$$|\phi_{X_{nt}}(\lambda) - 1| \leq \mathrm{E}\big(\min\{2|\lambda X_{nt}|, \tfrac{1}{2}\lambda^2 X_{nt}^2\}\big) \leq \lambda^2\sigma_{nt}^2 \tag{24.41}$$

where the second inequality of (24.41) applies the first inequality of (11.34) with $\varepsilon = \infty$. Squaring $|\phi_{X_{nt}}(\lambda) - 1|$, adding up over t, and substituting from the inequalities, remembering $\sum_{t=1}^n \sigma_{nt}^2 = 1$ and using (24.39), gives

$$\begin{aligned}\sum_{t=1}^n |\phi_{X_{nt}}(\lambda) - 1|^2 &\leq \max_{1\leq t\leq n} |\phi_{X_{nt}}(\lambda) - 1| \sum_{t=1}^n |\phi_{X_{nt}}(\lambda) - 1| \\ &\leq \Big(2\max_{1\leq t\leq n} P\big(|X_{nt}| > \varepsilon\big) + \varepsilon|\lambda|\Big)\lambda^2 \\ &\to |\lambda|^3\varepsilon \text{ as } n \to \infty.\end{aligned} \tag{24.42}$$

This result is used to construct an approximation to $\phi_{S_n}(\lambda)$. Since $\phi_{S_n}(\lambda) = \prod_{t=1}^n \phi_{X_{nt}}(\lambda)$ Lemma **24.7** can be applied with n large enough to give

$$\begin{aligned}\left|\exp\left\{\sum_{t=1}^n (\phi_{X_{nt}}(\lambda) - 1)\right\} - \phi_{S_n}(\lambda)\right| &\leq \sum_{t=1}^n |\exp\{\phi_{X_{nt}}(\lambda) - 1\} - \phi_{X_{nt}}(\lambda)| \\ &\leq \sum_{t=1}^n |\phi_{X_{nt}}(\lambda) - 1|^2\end{aligned} \tag{24.43}$$

where the second inequality is an application of **24.8** with $z = \phi_{X_{nt}}(\lambda) - 1$, noting that the condition of the lemma is satisfied for large enough n according to (24.41) and (24.23). By hypothesis $\phi_{S_n}(\lambda) \to \mathrm{e}^{-\lambda^2/2}$ so by choosing ε arbitrarily small, (24.43) and (24.42) together imply that

$$\exp\{\sum_{t=1}^n (\phi_{X_{nt}}(\lambda) - 1)\} \to \mathrm{e}^{-\lambda^2/2}. \tag{24.44}$$

Take the modulus of (24.44), apply identity (11.13) and then take the logarithm, to give

$$\log\left|\exp\left\{\sum_{t=1}^{n}(\phi_{X_{nt}}(\lambda)-1)\right\}\right| = \sum_{t=1}^{n}\mathrm{E}(\cos\lambda X_{nt}-1) \to -\tfrac{1}{2}\lambda^2. \tag{24.45}$$

This convergence holds for any choice of $\lambda > 0$. Fix $\varepsilon > 0$ and choose $\lambda > 2/\varepsilon$. Then, observe that

$$\begin{aligned}\left(\frac{\lambda^2}{2}-\frac{2}{\varepsilon^2}\right)\sum_{t=1}^{n}\int_{\{|X_{nt}|>\varepsilon\}}X_{nt}^2\mathrm{d}P &\le \sum_{t=1}^{n}\int_{\{|X_{nt}|>\varepsilon\}}\left(\tfrac{1}{2}\lambda^2X_{nt}^2-2\right)\mathrm{d}P\\ &\le \sum_{t=1}^{n}\int_{\{|X_{nt}|>\varepsilon\}}\left(\tfrac{1}{2}\lambda^2X_{nt}^2-1+\cos\lambda X_{nt}\right)\mathrm{d}P\\ &\le \sum_{t=1}^{n}\mathrm{E}\left(\tfrac{1}{2}\lambda^2X_{nt}^2-1+\cos\lambda X_{nt}\right)\\ &\le \tfrac{1}{2}\lambda^2+\sum_{t=1}^{n}\mathrm{E}(\cos\lambda X_{nt}-1) \to 0.\end{aligned} \tag{24.46}$$

The term in parentheses on the minorant side of (24.46) is positive by choice of λ. The third inequality holds because the integrand is positive for every X_{nt}, noting that by the power series expansion, $\cos x > 1-\frac{1}{2}x^2$ for $x \neq 0$. The last inequality substitutes $\sum_t \sigma_{nt}^2 = 1$ and the convergence is from (24.45). Since ε is arbitrary, the Lindeberg condition holds according to (24.46). ∎

Condition (24.39) is a condition of 'asymptotic negligibility', under which no single summand may be so influential as to dominate the sum as a whole. The chief reason why $\phi_{S_n}(\lambda) \to \mathrm{e}^{-\lambda^2/2}$ could happen without the Lindeberg condition, unless (24.39) holds, is that a finite number of summands dominating all the others could happen to be individually Gaussian. The following example illustrates.

24.14 Example Let $X_t \sim_{\mathrm{d}} \mathrm{N}(0,\sigma_t^2)$ where $\sigma_t^2 = 2^t$. Note that $s_n^2 = \sum_{t=1}^{n} 2^t = 2^n\sum_{t=0}^{n-1}2^{-t} = 2^{n+1}-2$, and $X_{nn} = X_n/s_n \sim_{\mathrm{d}} \mathrm{N}(0, 2^n/(2^{n+1}-2))$. Clearly $S_n \sim_{\mathrm{d}} \mathrm{N}(0,1)$ for every n by the linearity property of the Gaussian, but condition (24.39) fails. The Lindeberg condition also fails, since

$$\sum_{t=1}^{n}\int_{\{|X_{nt}|>\varepsilon\}}X_{nt}^2\mathrm{d}P \ge \int_{\{|X_{nn}|>\varepsilon\}}X_{nn}^2\mathrm{d}P \to \tfrac{1}{2}\mathrm{E}(1_{\{|Z|>\sqrt{2}\varepsilon\}}Z^2) > 0 \tag{24.47}$$

where Z is a standard Gaussian variate. □

A condition related to (24.39) is

$$P\Big(\max_{1\le t\le n}|X_{nt}| > \varepsilon\Big) \to 0 \text{ as } n \to \infty, \text{ any } \varepsilon > 0 \tag{24.48}$$

which says that the largest X_{nt} converges in probability to zero.

24.15 Theorem (24.48) implies (24.39).

Proof (24.48) is the same as $P\big(\max_{1\le t\le n}|X_{nt}| \le \varepsilon\big) \to 1$. But

$$\begin{aligned} P\Big(\max_{1\le t\le n}|X_{nt}| \le \varepsilon\Big) &= P\Big(\bigcap_{t=1}^{n}\{|X_{nt}| \le \varepsilon\}\Big) \\ &\le \min_{1\le t\le n} P\big(|X_{nt}| \le \varepsilon\big) \\ &= 1 - \max_{1\le t\le n} P\big(|X_{nt}| > \varepsilon\big) \end{aligned} \tag{24.49}$$

where the inequality is by monotonicity of P. If the first member of (24.49) converges to 1, so does the last. ∎

Also, interestingly enough:

24.16 Theorem The Lindeberg condition implies (24.48).

Proof Another way to write (24.17) (interchanging the order of summation and integration) is

$$\sum_{t=1}^{n} 1_{\{|X_{nt}|>\varepsilon\}} X_{nt}^2 \xrightarrow{L_1} 0, \text{ all } \varepsilon > 0. \tag{24.50}$$

According to **19.14** this implies $\sum_{t=1}^{n} 1_{\{|X_{nt}|>\varepsilon\}} X_{nt}^2 \to_{\text{pr}} 0$, or, equivalently,

$$P\Big(\sum_{t=1}^{n} 1_{\{|X_{nt}|>\varepsilon\}} X_{nt}^2 > \varepsilon^2\Big) \to 0 \text{ as } n \to \infty \tag{24.51}$$

for any $\varepsilon > 0$. But notice that

$$\Big\{\omega : \sum_{t=1}^{n} 1_{\{|X_{nt}|>\varepsilon\}}(\omega) X_{nt}^2(\omega) > \varepsilon^2\Big\} = \Big\{\omega : \max_{1\le t\le n}|X_{nt}(\omega)| > \varepsilon\Big\} \tag{24.52}$$

so (24.51) is equivalent to (24.48). ∎

Note that the last two results hold generally and do not impose independence on the sequence.

The foregoing theorems establish a network of implications that it may be helpful to summarize symbolically. Let

L = the Lindeberg condition;
I = independence of the sequence;
AG = asymptotic Gaussianity ($\phi_{S_n}(\lambda) \to \mathrm{e}^{-\lambda^2/2}$);
AN = asymptotic negligibility (condition (24.39)); and
PM = $\max|X_{nt}| \to_{\mathrm{pr}} 0$ (condition (24.48)).

Then,

$$L + I \Rightarrow AG + PM + I \Rightarrow AG + AN + I \Rightarrow L + I \tag{24.53}$$

where the first implication is the Lindeberg theorem and **24.16**, the second is by **24.15**, and the third is by **24.13**. Under independence, conditions L, $AG + PM$, and $AG + AN$ are therefore equivalent to one another.

24.4 The Case of Trending Variances

The Lindeberg and Feller theorems do not impose uniform integrability, nor L_r-boundedness conditions for any r. A trending variance sequence, with no uniform bound, is compatible with the Lindeberg condition. It would be sufficient if, for example, $\sigma_t^2 < \infty$ for each finite t and the unit-variance sequence $\{X_t/\sigma_t\}$ is uniformly square-integrable, provided that the variances do not grow so fast that the largest of them dominates the Cesàro sum of the sequence. The following is an extension of the sufficient condition of **24.10**.

24.17 Theorem $\{X_t\}$ satisfies the Lindeberg condition (24.29) if

(a) there exists a sequence $\{c_t\}$ of positive constants such that $\{X_t^2/c_t^2\}$ is uniformly integrable and

(b)

$$\sup_n nM_n^2/s_n^2 = C < \infty \tag{24.54}$$

where $M_n = \max_{1\le t\le n} c_t$ and $s_n^2 = \sum_{t=1}^n \sigma_t^2$. □

One way to construct the c_t might be as $\max\{1, \sigma_t\}$. The variances of the transformed sequence are then bounded by 1, but $\sigma_t^2 = 0$ is not ruled out for some t.

Proof of 24.17 The inequality of (24.32) extends by (b) to

$$\begin{aligned}\frac{1}{s_n^2}\sum_{t=1}^{n}\mathrm{E}(1_{\{|X_t|>s_n\varepsilon\}}X_t^2) &\le \frac{n}{s_n^2}\max_{1\le t\le n}\bigl\{c_t^2\mathrm{E}(1_{\{|X_t/c_t|>s_n\varepsilon/c_t\}}(X_t/c_t)^2)\bigr\}\\ &\le C\max_{1\le t\le n}\bigl\{\mathrm{E}(1_{\{|X_t/c_t|>s_n\varepsilon/c_t\}}(X_t/c_t)^2)\bigr\}. \qquad (24.55)\end{aligned}$$

By the analogous modification of (24.33), (a) then gives

$$\sup_t \lim_{n\to\infty}\mathrm{E}(1_{\{|X_t/c_t|>s_n\varepsilon/c_t\}}(X_t/c_t)^2) = 0. \quad ∎ \qquad (24.56)$$

Notice how (24.54) restricts the growth of the variances, whether this be positive or negative. Regardless of the choice of $\{c_t\}$, it requires that $s_n^2/n > 0$ uniformly in n. It permits the c_t to grow without limit so long as they are finite for all t, so the variances can do the same; but the rate of increase must not be so rapid as to have a single coordinate dominate the whole sequence. Letting $c_t = \max\{1, \sigma_t\}$ as above, (24.54) is satisfied (according to **2.17**) when $\sigma_t^2 \sim t^a$ for any $\alpha \ge 0$, but not when $\sigma_t^2 \sim 2^t$.

In fact, the conditions of **24.17** are stronger than necessary in the case of *decreasing* variances. The variance sequence may actually decline to zero without violating the Lindeberg condition, but in this case it is not possible to state a general sufficient condition on the sequence. If $\sigma_t^2 \sim t^a$ with $-1 < \alpha < 0$ then (24.32) is replaced by the condition

$$\frac{1}{s_n^2}\sum_{t=1}^{n}\mathrm{E}(1_{\{|X_t|>s_n\varepsilon\}}X_t^2) \le \frac{1}{Bn^{\alpha}}\max_{1\le t\le n}\bigl\{\mathrm{E}(1_{\{|X_t|>s_n\varepsilon\}}X_t^2)\bigr\} \qquad (24.57)$$

where $B = \inf_n(s_n^2/n^{1+\alpha}) > 0$ by assumption (note, $s_n^2 \sim n^{1+\alpha}$ under independence). Convergence of the majorant side of (24.57) to zero as $n \to \infty$ is not ruled out, but depends on the distribution of the X_t.

The following example illustrates both possibilities.

24.18 Example Let $\{X_t\}$ be a zero-mean independent sequence with $X_t \sim_{\mathrm{d}} \mathrm{U}[-t^{\alpha}, t^{\alpha}]$ for some real α so that $\sigma_t^2 = \frac{1}{3}t^{2\alpha}$ (see **9.12**), either growing with t ($\alpha > 0$) or declining with t ($\alpha < 0$). However, X_t is L_∞-bounded for finite t. The integrals in (24.29) each take the form

$$\frac{1}{2t^{\alpha}}\int_{-t^{\alpha}}^{t^{\alpha}} 1_{\{|\xi|>s_n\varepsilon\}}\xi^2 \mathrm{d}\xi \qquad (24.58)$$

where $s_n^2 = \frac{1}{3}\sum_{\tau=1}^n \tau^{2\alpha}$. Now, $\sum_{\tau=1}^n \tau^{2\alpha} = O(n^{2\alpha+1})$ for $\alpha > -\frac{1}{2}$ and $O(\log n)$ for $\alpha = -\frac{1}{2}$ (**2.17**). Condition (24.54) is satisfied when $\alpha \geq 0$. Note that (24.58) is zero if $\left(\frac{1}{3}\sum_{\tau=1}^n \tau^{2\alpha}\right)^{1/2}\varepsilon > t^\alpha$; $\left(\sum_{\tau=1}^n \tau^{2\alpha}\right)^{1/2}$ grows faster than n^α for all $\alpha \geq 0$ and hence (24.58) vanishes in the limit for every t in these cases and the Lindeberg condition is satisfied. But if $X_t \sim_{\mathrm{d}} \mathrm{U}[-2^t, 2^t]$, 2^t grows at the same rate as $\left(\sum_{\tau=1}^n 2^{2\tau}\right)^{1/2}$, so the above argument does not apply and the Lindeberg condition fails. Note how condition (24.54) is violated in this case.

However, the fact that condition (24.54) is not necessary is evident from the fact that the variance sum diverges at a positive rate when $X_t \sim_{\mathrm{d}} \mathrm{U}[-t^\alpha, t^\alpha]$ for any $\alpha \geq -\frac{1}{2}$ even though the variance sequence itself goes to zero. It can be verified that (24.58) vanishes in the limit and accordingly the Lindeberg condition holds for these cases too. On the other hand, if $\alpha < -\frac{1}{2}$, s_n^2 is bounded in the limit and (24.17) becomes

$$\frac{3}{B}\lim_{n\to\infty}\sum_{t=1}^n \frac{1}{2t^\alpha}\int_{-t^\alpha}^{t^\alpha} 1_{\{|\xi|>\varepsilon\sqrt{B/3}\}}\xi^2 \mathrm{d}\xi \tag{24.59}$$

where $B = \lim_{n\to\infty}\sum_{t=1}^n t^{2\alpha} < \infty$ and by choice of small enough ε, (24.59) can be made arbitrarily close to 1. This is the other extreme at which the Lindeberg condition fails. □

24.5 Gaussianity by Other Means

This section examines some cases where a form of central limit theorem operates in spite of the failure of conditions often assumed to be necessary. Consider first an example of a true CLT operating under asymptotic negligibility but *without* the Lindeberg condition.

24.19 Example This case has similar characteristics to **24.9**. Let $X_t = \frac{1}{2}$ and $-\frac{1}{2}$ with probabilities $\frac{1}{2}(1 - t^{-2})$ each and t and $-t$ with probabilities $\frac{1}{2}t^{-2}$ each, so that $\mathrm{E}(X_t) = 0$ and $\mathrm{Var}(X_t) = \frac{5}{4} - \frac{1}{4}t^{-2}$ and so $s_n^2 = \frac{5}{4}n + O(1)$. For any $\varepsilon > 0$, let n be large enough that $s_n\varepsilon > \frac{1}{2}$ and let n_ε be the smallest integer exceeding $s_n\varepsilon$. Then

$$P(|X_t| > s_n\varepsilon) \leq n_\varepsilon^{-2}. \tag{24.60}$$

That is, if $s_n\varepsilon \leq t$ the probability equals t^{-2} and is otherwise zero. Therefore, since $n_\varepsilon = O(n^{1/2})$, (24.39) holds in this case. However, the Lindeberg condition is not satisfied. For $n_\varepsilon \leq t \leq n$, $\mathrm{E}(1_{\{|X_t|>\varepsilon s_n\}}X_t^2) = t^2/t^2 = 1$ and hence

$$\frac{1}{s_n^2}\sum_{t=1}^{n}\int_{\{|X_t|>s_n\varepsilon\}} X_t^2 \mathrm{d}P \geq \frac{n-n_\varepsilon}{s_n^2} \to \tfrac{4}{5} > 0. \tag{24.61}$$

However, consider the random sequence $\{W_t\}$, where $W_t = X_t$ when $|X_t| = \frac{1}{2}$ and $W_t = 0$ otherwise. As t increases, W_t tends to a centred Bernoulli variate with $p = \frac{1}{2}$ and hence $\mathrm{E}(W_t^2) = \frac{1}{4}$. Therefore

$$\frac{1}{s_n}\sum_{t=1}^{n} W_t \xrightarrow{\mathrm{d}} \mathrm{N}(0, \tfrac{1}{5}). \tag{24.62}$$

$|X_t - W_t| = t$ with probability $1/t^2$ and 0 otherwise and hence, similarly to (24.31),

$$\left|\frac{1}{s_n}\sum_{t=1}^{n} X_t - \frac{1}{s_n}\sum_{t=1}^{n} W_t\right| \leq \frac{1}{s_n}\sum_{t=1}^{n}|X_t - W_t| = O_p(n^{-1/2}).$$

It follows according to **23.18** that $s_n^{-1}\sum_{t=1}^{n} X_t \to_{\mathrm{d}} \mathrm{N}(0, \frac{1}{5})$. □

A CLT does operate in this case and Feller's theorem is not contradicted because the limit is not the *standard* Gaussian. The clue to this apparent paradox is that the sequence is not uniformly square-integrable, having a component that contributes to the variance asymptotically despite vanishing in probability. In these circumstances $s^{-1}\sum_{t=1}^{n} X_t$ can have a 'variance' of 1 for every n in spite of the fact that its limiting distribution has a variance of $\frac{1}{5}$.

The next result shows that, notwithstanding what is sometimes assumed, a finite variance is not a necessary condition for the Gaussian central limit theorem to hold. It is not implied by the Lindeberg condition nor by the other conditions summarized in (24.53).

24.20 Theorem Let $X_1, \ldots, X_n$ be a sequence of i.i.d. random variables having mean zero. There exists a sequence of constants $a_n > 0$ such that $a_n^{-1}\sum_{t=1}^{n} X_t \to_{\mathrm{d}} \mathrm{N}(0, \sigma^2)$ for $\sigma^2 < \infty$ iff there exists a diverging sequence of positive constants $\{C_n\}$ such that as $n \to \infty$,

$$nP(|X_1| > C_n) \to 0 \tag{24.63}$$

and

$$\frac{n}{C_n^2}\mathrm{E}\big(X_1^2 1_{\{|X_1| \leq C_n\}}\big) \to \infty. \tag{24.64}$$

Proof Suppose that (24.64) holds for a given sequence $\{C_n\}$. Set

$$\varepsilon_n = \frac{\sigma C_n}{\sqrt{n\mathrm{E}(X_1^2 1_{\{X_1 \le C_n\}})}} \tag{24.65}$$

so that $\varepsilon_n \to 0$ by construction. Let F denote the common c.d.f. of the variables and set

$$a_n = \frac{C_n}{\varepsilon_n} = \frac{1}{\sigma}\left(n\int_{-C_n}^{C_n} x^2 \mathrm{d}F(x)\right)^{1/2} \to \infty. \tag{24.66}$$

Hence, defining $F_n(y) = F(a_n y)$,

$$\begin{aligned}\sigma^2 &= \frac{n}{a_n^2}\int_{-C_n}^{C_n} x^2 \mathrm{d}F(x)\\ &= n\int_{-\varepsilon_n}^{\varepsilon_n} y^2 \mathrm{d}F_n(y)\\ &\to G(0+) - G(0-) \text{ as } n \to \infty \end{aligned} \tag{24.67}$$

where G is defined in (11.55). Next, noting that $M(-\infty) = N(\infty) = 0$ where M and N are defined respectively in (11.58) and (11.59),

$$\begin{aligned} nP(|X_1| > C_n) &= n\int_{-\infty}^{-C_n} \mathrm{d}F(x) + n\int_{C_n}^{\infty} \mathrm{d}F(x)\\ &= n\int_{-\infty}^{-\varepsilon_n} \mathrm{d}F_n(y) + n\int_{\varepsilon_n}^{\infty} \mathrm{d}F_n(y)\\ &\to \int_{-\infty}^{0} \mathrm{d}M(y) + \int_{0}^{\infty} \mathrm{d}N(y) \text{ as } n \to \infty\\ &= M(0) - N(0). \end{aligned} \tag{24.68}$$

Since both functions are nondecreasing, this shows that (24.63) is necessary and sufficient for $M(x) = 0$ for all $x < 0$ and $N(x) = 0$ for all $x > 0$. Finally if $\mathrm{E}(X_1) = 0$ then $\gamma = 0$ in (11.53). The limiting form of (11.57) is therefore $\log\phi(t) = -\frac{1}{2}\sigma^2 t^2$, proving sufficiency. Necessity follows first from the fact that the condition $M(0) - N(0) \neq 0$ contradicts (24.63) according to (24.68). The condition $G(0+) - G(0-) = \sigma^2 > 0$ is also necessary for normality and $\varepsilon_n \to 0$ in (24.67) implies (24.64) for some $C_n = o(a_n)$. ∎

The key role of (24.63) is to impose asymptotic negligibility as in (24.48). By subadditivity, it implies under the i.i.d. assumption that for all $\varepsilon > 0$,

$$\begin{aligned} P\Big(\max_{1\le t\le n}|X_t/a_n| > \varepsilon\Big) &= P\Big(\bigcup\nolimits_{t=1}^{n}\{|X_t/a_n| > \varepsilon\}\Big) \\ &\le nP(|X_1| > \varepsilon a_n) \to 0. \end{aligned} \tag{24.69}$$

To see how these conditions might fail, consider a case where $P(|X_1| < x) = x^{-\alpha}$ for $\alpha < 2$ and $\mathrm{E}(X_1^2 1_{\{|X_1|<x\}}) = x^{2-\alpha}$. Then, with $\sigma = 1$, (24.65) solves as $C_n = n^{1/\alpha}\varepsilon_n^{2/\alpha}$ but (24.63) yields, if $\varepsilon_n \to 0$,

$$nP(|X_1| > C_n) = O(nC_n^{-\alpha}) = O(\varepsilon_n^{-2}) \to \infty.$$

Distributions having these attributes are in the domain of attraction of a non-Gaussian α-stable distribution; these cases are to be examined in §24.6.

However, although a constant σ^2 is defined in **24.20** as the limiting variance of the normalized sum, take care to note there has been no assumption that this is equal to $\mathrm{E}(X_1^2)$, nor that this latter quantity is finite. The clue is in the unspecified normalizing sequence a_n. (24.66) shows that $a_n = \sqrt{n}$ in the case where $\mathrm{E}(X_1^2) < \infty$, but evidently other possibilities exist. Consider the following example of a distribution that is well-known to have infinite variance, yet satisfies the conditions of **24.20**.

24.21 Example Let $X_1 \sim_{\mathrm{d}} t(2)$, the Student's t distribution with two degrees of freedom. Consider formula (8.33) for the case $\nu = 2$.

$$\begin{aligned} P(|X_1| > C_n) &= \frac{2}{\sqrt{8}}\int_{C_n}^{\infty}(1+u^2/2)^{-3/2}\mathrm{d}u \\ &= \frac{\sqrt{C_n^2+2}-C_n}{\sqrt{C_n^2+2}} \\ &= O(1/C_n^2) \end{aligned}$$

and also

$$\begin{aligned} \mathrm{E}\big(X_1^2 1_{\{|X_1|\le C_n\}}\big) &= \frac{2}{\sqrt{8}}\int_0^{C_n} u^2(1+u^2/2)^{-3/2}\mathrm{d}u \\ &= 2\log(\sqrt{C_n^2+2}+C_n) - \frac{2\sqrt{2}C_n}{\sqrt{1+2C_n^2}} - \log 2 \\ &= 2\log C_n + O(1) \text{ as } C_n \to \infty. \end{aligned}$$

Therefore (24.65) gives $C_n \sim \sqrt{2n \log C_n} \varepsilon_n$. Set $a_n = \sqrt{n \log n}$, $\varepsilon_n = (\log n)^{-1/4}$, and then $C_n = \sqrt{n} (\log n)^{1/4}$. It can be verified that in this case both of conditions (24.63) and (24.64) are satisfied. It follows that $(n \log n)^{-1/2} \sum_{t=1}^{n} X_t \to_{\mathrm{d}} \mathrm{N}(0,1)$ when $X_t \sim_{\mathrm{d}} t(2)$. □

This example differs from the usual CLT by the presence of the slowly varying component in the normalizing sequence. Another way to appreciate the result is to note that under the usual normalization factor of $\sqrt{n}$, the normalized sum appears to approach a Gaussian limit but at the same time the variance is diverging like $\log n$ instead of approaching 1. This offers some intuition into the type of behaviour to be expected at the boundary of the regularity conditions for operation of the CLT.

The next result gives a condition that is both necessary and sufficient for those of **24.20** and may be more convenient to cite.

24.22 Theorem The distribution F belongs to the domain of attraction of the normal law iff

$$\frac{m^2 \int_{|x|>m} \mathrm{d}F(x)}{\int_{|x|\le m} x^2 \mathrm{d}F(x)} \to 0 \text{ as } m \to \infty. \tag{24.70}$$

Proof To show necessity, assume that F belongs to the domain of attraction of the normal law. Let F denote the distribution of X_1 in **24.20** and then by hypothesis and **24.20**, (24.63) and (24.64) hold. For any m large enough, n can be chosen such that such that $C_n \le m \le C_{n+1}$ and the ratio in (24.70) can be made as small as desired by choice of m.

To show sufficiency, define for $\delta > 0$ the sequences

$$C_n(\delta) = \inf \left\{ m : n \int_{|x|>m} \mathrm{d}F(x) \le \delta \right\}.$$

It is evident that $C_n(\delta) \to \infty$ as $n \to \infty$, so that (24.70) implies that as $n \to \infty$,

$$\frac{n}{C_n(\delta)^2} \int_{|x|\le C_n(\delta)} x^2 \mathrm{d}F(x) \to \infty. \tag{24.71}$$

This is true for any fixed $\delta > 0$. Now choose a sequence $\{\delta_n\}$ with $\delta_n \to 0$ as $n \to \infty$ and consider the sequence $C_n(\delta_n)$. With this choice, both (24.63) and (24.64) hold as $n \to \infty$ so that the sufficient condition of **24.20** is verified. ∎

It is easy to check that condition (24.70) holds in the case of Example **24.21**, noting that the numerator of the ratio is $O(1)$ whereas the denominator is $O(\log m)$.

24.6 α-Stable Convergence

In this section it is assumed that the necessary condition of **24.20** fails, so that in particular the variance does not exist. The form of the so-called α-stable attractor distributions was derived in §23.6. A random variable X (assumed standardized for location and scale) lies in the domain of attraction of $S_\alpha(\beta)$ if there exist sequences $\{a_n\}$ and $\{b_n\}$ such that

$$\frac{S_n - b_n}{a_n} \xrightarrow{\mathrm{d}} S_\alpha(\beta)$$

where $S_n = X_1 + \cdots + X_n$ is the partial sum of independent drawings. The familiar case of the normal distribution is exceptional in that every random variable having a finite variance lies in its domain of attraction. This is what the results of §24.1–§24.4 have in various ways established. The case of non-Gaussian attractors is more complicated because of the profusion of possible attractors.

The so-called *generalized central limit theorem* is the result that limits in distribution of normalized partial sums of independent random variables belong to the family of α-stable distributions, either Gaussian or having a ch.f. with the functional form of (23.67). Possession of a variance is sufficient for a distribution to lie in the normal domain of attraction. For the α-stable case with $\alpha < 2$ the following set of conditions applies, notably relating to the behaviour of the distribution in the tails.

24.23 Theorem A random variable X lies in the domain of attraction of $S_\alpha(\beta)$ for $0 < \alpha < 2$ iff

$$P(|X| > x) = x^{-\alpha}L(x) \tag{24.72}$$

where $L(x)$ is a function that is slowly varying as $x \to \infty$, and

$$\frac{P(X < -x)}{P(|X| > x)} \to p, \quad \frac{P(X > x)}{P(|X| > x)} \to 1 - p \tag{24.73}$$

as $x \to \infty$, for a constant $p \in [0, 1]$. □

Condition (24.72) may also be rendered in the form

$$\frac{P(|X| > bx)}{P(|X| > x)} \to b^{-\alpha} \text{ as } x \to \infty.$$

Notice that these are conditions that bind the tails of the distribution. Since slowly varying functions are allowed to have arbitrary local properties, the range of different distributions with infinite variance that satisfy the conditions of Theorem **24.23** for some α is clearly wide, but the factor $L(x)$ must not affect the rate of decay of the tail probabilities according to (24.72). To interpret the conditions, recall from (11.58) and (11.59) the definitions of functions M and N in the Lévy representation of an infinitely divisible distribution and refer to the attributes of a stable distribution itemized in **23.26**. The special cases corresponding to α-stability are that $L(x) = c_1 + c_2$ not depending on x, that $p = c_1/(c_1 + c_2)$, and that there is no discontinuity at 0 in the function $G(x)$ defined in (11.55) so that $\sigma^2 = 0$. What the proof shows is that when a distribution lies in the domain of attraction these conditions are obtained for extreme values, after replacing x by $a_n x$ where the sequence $\{a_n\}$ increases with n at a suitably chosen rate.

The context is that X stands for one of the terms of a normalized sum, say $S_n = a_n^{-1} \sum_{t=1}^{n} X_t$ where $X_1, \dots, X_n$ are i.i.d. random variables whose distribution satisfies the conditions of **24.23**. The key observation is that $P(|X| > a_n x) = P(|X|/a_n > x)$. What is shown is that when $\{a_n\}$ is appropriately defined, specifically requiring $a_n \sim n^{1/\alpha}$, the normalized probabilities $nP(|X_t/a_n| > x)$ tend to $(c_1 + c_2)x^{-\alpha}$ as n increases for any $x > 0$. Of course, the terms X_t/a_n are becoming of small order so that only their extreme deviations can have a non-negligible impact on the sum. It is the fact that the distribution of these extreme deviations has α-stable characteristics that defines membership of the relevant domain of attraction. As shown in **23.26**, the distribution of S_n is then tending with increasing n to obey the relation $\phi(t) = \phi(t/n^{1/\alpha})^n$. The stylized sampling scenario described on page 513 offers a clue to the random mechanism that results in the convergence to the attractor distribution. What distinguishes these cases from the Gaussian limit is that the distribution of the increments must contain the appropriate 'seed' for the limit distribution, specifically the parameters α and p.

Proof of 24.23 According to (24.72), for $x > 0$ and an increasing sequence $\{a_n\}$,

$$nP(|X| > a_n x) = n(a_n x)^{-\alpha} L(a_n x). \tag{24.74}$$

Set

$$a_n = \inf\{z \in \mathbb{R}^+ : nz^{-\alpha} L(z) \le c_1 + c_2\}. \tag{24.75}$$

Then, $a_n \to \infty$ as $n \to \infty$ and according to (24.74),

$$\frac{nP(|X| > a_n x)}{nP(|X| > a_n)} = \frac{n(a_n x)^{-\alpha} L(a_n x)}{n a_n^{-\alpha} L(a_n)} \to x^{-a} \text{ as } n \to \infty. \tag{24.76}$$

However, (24.75) and (24.72) together imply that $nP(|X| > a_n) = c_1 + c_2$ and hence by (24.76) that as $n \to \infty$,

$$nP(|X| > a_n x) \to (c_1 + c_2)x^{-\alpha}. \tag{24.77}$$

Since $c_1 = p(c_1 + c_2)$ (24.77) and (24.73) give

$$nP(X < -a_n x) \to c_1 x^{-\alpha}$$

and similarly

$$nP(X > a_n x) \to c_2 x^{-\alpha}.$$

This shows that (24.72) and (24.73) are sufficient for conditions (i) and (ii) of Theorem **23.26**, referring to (11.58) and (11.59) for the definitions of M and N. With L replaced by $c_1 + c_2$ in (24.75), condition (iii) of **23.26** is also found.

To show condition (iv) of **23.26** for the case $\alpha < 2$, write F for the c.d.f. of X. In view of the definition $\sigma^2 = G(0+) - G(0-)$ and (11.55), the required condition has the form

$$\lim_{\varepsilon \to 0} \limsup_{n \to \infty} n \int_{x<|\varepsilon|} x^2 \mathrm{d}F(a_n x) = 0. \tag{24.78}$$

After a change of variable, Theorem **9.21** implies

$$\begin{aligned} n \int_{x<|\varepsilon|} x^2 \mathrm{d}F(a_n x) &= \frac{n}{a_n^2} \int_{|\xi|<a_n\varepsilon} \xi^2 \mathrm{d}F(\xi) \\ &= \frac{2n}{a_n^2} \int_0^{a_n\varepsilon} \xi P(|X| > \xi)\mathrm{d}\xi - n\varepsilon^2 P(|X| > a_n\varepsilon). \end{aligned} \tag{24.79}$$

Consider the first right-hand-side term of (24.79). Applying Theorem **2.33**(i) for regularly varying functions gives the asymptotic relation

$$\frac{2}{a_n^2} \int_0^{a_n\varepsilon} \xi^{1-\alpha} L(\xi)\mathrm{d}\xi = \frac{2\varepsilon^2}{2-\alpha}(a_n\varepsilon)^{-\alpha} L(a_n\varepsilon)(1 + o(1)). \tag{24.80}$$

Substituting from (24.74), (24.80) gives the corresponding relation

$$\begin{aligned} \frac{2n}{a_n^2} \int_0^{a_n\varepsilon} \xi P(|X| > \xi)\mathrm{d}\xi &= \frac{2\varepsilon^2}{2-\alpha} nP(|X| > a_n\varepsilon)(1 + o(1)) \\ &\to \frac{2}{2-\alpha}(c_1 + c_2)\varepsilon^{2-\alpha} \text{ as } n \to \infty \end{aligned} \tag{24.81}$$

where the limit applies (24.77). Also in view of (24.77), the second right-hand-side term of (24.79) converges as

$$n\varepsilon^2 P(|X| > a_n\varepsilon) \to (c_1 + c_2)\varepsilon^{2-\alpha}. \tag{24.82}$$

Since $\alpha < 2$, both of (24.81) and (24.82) vanish as $\varepsilon \to 0$, confirming (24.78).

This completes the proof of sufficiency. Showing necessity is simply a matter of observing that the parameters α and p (recalling $\beta = 2p - 1$) are common to the hypothesis (24.72)+(24.73) and the limit distribution $S_a(\beta)$. ∎

It might be remarked here that in the standardized distribution $S_a(\beta)$, $c_1 + c_2$ is the function of α implied by imposing $\delta = 1$ in the formulae in (23.69).

A useful device for understanding the properties of these distributions is the series representation introduced by Le Page, Woodroofe, and Zinn ([120]). A feature of the α-stable distributions is that absolute values are distributed independently of signs. The absolute values depend on $c_1 + c_2$ but not on c_1 and c_2 individually, and p is the fixed probability that $X > 0$. Accordingly, write $Y_t = |X_t|$ and so write $X_t = \delta_t Y_t$ where δ_t is a two-point r.v. taking the value 1 if $X_t \geq 0$ and -1 otherwise. When the distribution is α-stable, Y_t and δ_t are distributed independently with $P(\delta_t = -1) = p$ and $P(\delta_t = 1) = 1 - p$.

This fact makes it possible to focus on the distribution of $Y_1, \dots, Y_n$ in isolation. Define

$$G(x) = F(-x) + 1 - F(x) = P(|X| > x)$$

which is a monotone decreasing function of x. Defining the quantile function of G as $G^{-1}(u) = \inf\{x : G(x) \geq u\}$, by an approach closely allied to the Skorokhod construction of §23.2 it is possible to write $Y_t = G^{-1}(U_t)$ where U_t is a drawing from the uniform distribution on $[0, 1]$. In the α-stable case it is possible by appropriate normalization (choice of units of measurement) to write $G(x) = x^{-\alpha}$ and hence, $Y_t = U_t^{-1/\alpha}$.

The next critical step is to sort the variables into order from largest to smallest. Define $Y_{(1)}, \dots, Y_{(n)}$ to be the order statistics of the set $Y_1, \dots, Y_n$, where $Y_{(1)} \geq Y_{(2)} \geq \cdots \geq Y_{(n)}$. Then let $U_{(t)} = Y_{(t)}^{-\alpha}$, such that $U_{(1)} \leq U_{(2)} \leq \cdots \leq U_{(n)}$. It was shown in Theorem **13.7** that the joint distribution of the ordered set $U_{(1)}, \dots, U_{(n)}$ of independently drawn U$[0, 1]$ r.v.s matches that of the arrival times $\Gamma_k/\Gamma_{n+1}, k = 1, \dots, n$ of a normalized Poisson process where the Γ_k are Gamma$(k, 1)$-distributed and k represents the rank of $U_{(t)}$ in the set. Therefore, when the X_t are α-stably distributed,

$$\begin{aligned}
n^{-1/\alpha}Y_{(t)} &= n^{-1/\alpha}G^{-1}(U_{(t)}) \\
&= \inf\{n^{-1/\alpha}y : y^{-\alpha} \geq U_{(t)}\} \\
&= \inf\{z : z \geq (nU_{(t)})^{-1/\alpha}\} \\
&= (nU_{(t)})^{-1/\alpha} \\
&\overset{\mathrm{d}}{\sim} (n\Gamma_k/\Gamma_{n+1})^{-1/\alpha} \\
&\xrightarrow{\mathrm{pr}} \Gamma_k^{-1/\alpha} \text{ as } n \to \infty. \qquad (24.83)
\end{aligned}$$

The equivalence in distribution in the penultimate member of (24.83) is by Theorem **13.7** and the final convergence in probability is because $\mathrm{E}(\Gamma_k) = k$ and $\mathrm{Var}(\Gamma_k) = k$ (see (9.10) and (9.26)) so that $\Gamma_{n+1}/n \to_{L_2} 1$ as $n \to \infty$.

Let $X_1, \dots, X_n$ be an independent collection of α-stable random variables, assumed for simplicity to be centred so that $b_n = 0$. Also letting $\delta_{(t)}$ denote the random sign paired with $Y_{(t)}$, the conclusion from (24.83) is that

$$\frac{1}{n^{1/\alpha}}\sum_{t=1}^{n} X_t = \frac{1}{n^{1/\alpha}}\sum_{t=1}^{n}\delta_{(t)}Y_{(t)} \xrightarrow{\mathrm{d}} \sum_{k=1}^{\infty}\varepsilon_k\Gamma_k^{-1/\alpha} \qquad (24.84)$$

where $\{\varepsilon_k\}_{k=1}^{\infty}$ is an independent sequence with $P(\varepsilon_k = -1) = p$, $P(\varepsilon_k = 1) = 1 - p$. By infinite divisibility, the limit in distribution in (24.84) is a representation of an $S_\alpha(2p-1)$ random variable.

The result in (24.84) is striking in several ways, not least in the fact that it is valid only for $\alpha < 2$. There is a discontinuity in the model space at the point $\alpha = 2$ because the stable distribution at that point is the Gaussian for which $G(y)$ decreases exponentially, not like y^{-2} as continuity would require.

Since Γ_k is the sum of independent exponential variates it has the characteristics of a random walk with positive increments and hence with drift. The terms $\Gamma_k^{-1/\alpha}$ in (24.84) are therefore diminishing as k increases, the faster as α is closer to zero. A striking revelation is that

$$P(\Gamma_k^{-1/\alpha} > x) = P(\Gamma_k < x^{-\alpha}) = \frac{1}{\Gamma(k)}\int_0^{x^{-\alpha}} \xi^{k-1}\mathrm{e}^{-\xi}\mathrm{d}\xi = O(x^{-k\alpha}).$$

In other words, the property $P(|X| > x) \sim x^{-\alpha}$ is due solely to the leading term of the sum in (24.84). If the absolutely largest coordinate of the sequence were to be systematically deleted in repeated drawings, the reduction in tail probabilities of the normalized sum would persist even in the limit. When an α-stable random variable is generated in this way a finite set of observations influences the distribution even as $n \to \infty$. This is in striking contrast to the asymptotic negligibility condition that enables Gaussian convergence. It is only because of the influence of extreme values that the α-stable limit distributions can exhibit skewness. With asymptotically negligible summands such characteristics are averaged away to reveal the symmetric Gaussian limit.

25

CLTs for Dependent Processes

25.1 A General Convergence Theorem

The main results of this chapter are derived from the following fundamental theorem, due to McLeish ([124]).

25.1 Theorem Let $\{Z_{ni}, i = 1, \ldots, r_n, n \in \mathbb{N}\}$ denote a zero-mean stochastic array, where r_n is a positive, increasing integer-valued function of n, and let

$$T_{r_n} = \prod_{i=1}^{r_n}(1 + \mathrm{i}\lambda Z_{ni}),\ \lambda > 0. \tag{25.1}$$

Then, $S_{r_n} = \sum_{i=1}^{r_n} Z_{ni} \to_{\mathrm{d}} \mathrm{N}(0,1)$ if the following conditions hold:

(a) T_{r_n} is uniformly integrable,

(b) $\mathrm{E}(T_{r_n}) \to 1$ as $n \to \infty$,

(c) $\sum_{i=1}^{r_n} Z_{ni}^2 \to_{\mathrm{pr}} 1$ as $n \to \infty$,

(d) $\max_{1 \le i \le r_n} |Z_{ni}| \to_{\mathrm{pr}} 0$ as $n \to \infty$. □

There are a number of features requiring explanation here, regarding both the theorem and the way it has been expressed. This is a generic result in which the elements of the array need not be data points in the conventional way, so that their number r_n does not always correspond with the number of sample observations, n. $r_n = n$ is a leading case, but see **25.6** for another possibility.

It is interesting to note that the Lindeberg condition is not imposed explicitly in **25.1**, nor is anything specific assumed about the dependence of the sequence. Condition **25.1**(d) is condition PM defined in (24.48) and by **24.16** it follows from the Lindeberg condition. According to (24.53), condition (d) is equivalent to the Lindeberg condition in independent sequences where (as is to be proved here) the central limit theorem holds, but without independence no such conclusion can be drawn.

Proof of 25.1 Consider the series expansion of the logarithmic function, defined for $|x| < 1$,

$$\log(1 + x) = x - \tfrac{1}{2}x^2 + \tfrac{1}{3}x^3 - \ldots.$$

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0025

Although a complex number does not possess a unique logarithm (see §11.2) the arithmetic identity obtained by taking the exponential of both sides of this equation is well-defined when x is complex. For $x = \mathrm{i}\lambda Z_{ni}$ the formula yields

$$1 + \mathrm{i}\lambda Z_{ni} = \exp\{\mathrm{i}\lambda Z_{ni}\}\exp\{\tfrac{1}{2}\lambda^2 Z_{ni}^2 + r(\mathrm{i}\lambda Z_{ni})\} \tag{25.2}$$

where the remainder satisfies $|r(x)| \le |x|^3$ for $|x| < 1$. Multiplying the terms for $i = 1, \ldots, r_n$ and rearranging yields

$$\exp\{\mathrm{i}\lambda S_{r_n}\} = T_{r_n} U_{r_n}$$

where T_{r_n} is defined in (25.1) and

$$U_{r_n} = \exp\left\{-\tfrac{1}{2}\lambda^2 \sum_{i=1}^{r_n} Z_{ni}^2 - \sum_{i=1}^{r_n} r(\mathrm{i}\lambda Z_{ni})\right\}. \tag{25.3}$$

Taking expectations, the characteristic function of the sum can be written as

$$\phi_{S_{r_n}}(\lambda) = \mathrm{E}(T_{r_n} U_{r_n}) = \mathrm{e}^{-\lambda^2/2}\mathrm{E}(T_{r_n}) + \mathrm{E}\big(T_{r_n}(U_{r_n} - \mathrm{e}^{-\lambda^2/2})\big) \tag{25.4}$$

so given condition (b) of the theorem, $\phi s_{r_n}(\lambda) \to \mathrm{e}^{-\lambda^2/2}$ if

$$\lim_{n\to\infty} \mathrm{E}\big|T_{r_n}(U_{r_n} - \mathrm{e}^{-\lambda^2/2})\big| = 0. \tag{25.5}$$

The sequence

$$T_{r_n}(U_{r_n} - \mathrm{e}^{-\lambda^2/2}) = \exp\{\mathrm{i}\lambda S_{r_n}\} - T_{r_n}\mathrm{e}^{-\lambda^2/2} \tag{25.6}$$

is uniformly integrable in view of condition (a), the first term on the right-hand side having unit modulus. So in view of **19.15**, it suffices to show that

$$\operatorname*{plim}_{n\to\infty} T_{r_n}(U_{r_n} - \mathrm{e}^{-\lambda^2/2}) = 0. \tag{25.7}$$

Since T_{r_n} is clearly $O_p(1)$, by **19.13** the problem reduces to showing $\mathrm{plim}_{n\to\infty}$ $U_{r_n} = \mathrm{e}^{-\lambda^2/2}$ and for this in turn it suffices, by condition (c), if

$$\operatorname*{plim}_{n\to\infty}\sum_{i=1}^{r_n} r(\mathrm{i}\lambda Z_{ni}) = 0. \tag{25.8}$$

Given the inequality

$$\left|\sum_{i=1}^{r_n} r(\mathrm{i}\lambda Z_{ni})\right| \le |\lambda|^3\sum_{i=1}^{r_n}|Z_{ni}|^3 \le |\lambda|^3\Big(\max_{1\le i\le r_n}|Z_{ni}|\Big)\sum_{i=1}^{r_n}Z_{ni}^2 \tag{25.9}$$

(25.8) follows from conditions (c) and (d) by **19.13**. ∎

It is instructive to compare this proof with that of the Lindeberg theorem. A different series approximation of the ch.f. is used and the assumption from independence that

$$\phi_{S_{r_n}} = \prod\nolimits_i \phi_{Z_{ni}}$$

is avoided. Of course, it has yet to be shown that conditions **25.1**(a) and **25.1**(b) hold under convenient and plausible assumptions about the sequence. With the exception of §25.3 which describes a contrasting approach, the rest of this chapter is devoted to this question. **25.1**(b) will turn out to result from suitable restrictions on the dependence. **25.1**(a) can be shown to follow from a more primitive moment condition by an argument based on the 'equivalent sequences' idea, as follows.

25.2 Theorem For an array $\{Z_{ni}\}$ let

$$J_n = \begin{cases} \min\{j : \sum_{i=1}^{j} Z_{ni}^2 > 2\} & \text{if } \sum_{i=1}^{r_n} Z_{ni}^2 > 2 \\ r_n, & \text{otherwise.}\end{cases}$$

(i) The sequence $T_{J_n} = \prod_{i=1}^{J_n}(1+\mathrm{i}\lambda Z_{ni})$ is uniformly integrable if

$$\sup_n \mathrm{E}\Big(\max_{1\le i\le r_n} Z_{ni}^2\Big) < \infty. \tag{25.10}$$

If $\sum_{i=1}^{r_n} Z_{ni}^2 \to_{\mathrm{pr}} 1$, then

(ii) $\sum_{i=1}^{J_n} Z_{ni}^2 \to_{\mathrm{pr}} 1$

(iii) S_{J_n} has the same limiting distribution as S_{r_n}.

Proof The inequality $1 + x \leq \mathrm{e}^x$ for $x \geq 0$ implies that $\prod_i (1 + x_i) \leq \prod_i \mathrm{e}^{x_i}$ for $x_i > 0$. Hence, since the terms $\lambda^2 Z_{ni}^2$ are real and positive,

$$\begin{aligned}|T_{J_n}|^2 &= \prod_{i=1}^{J_n-1}(1 + \lambda^2 Z_{ni}^2)(1 + \lambda^2 Z_{nJ_n}^2)\\ &\leq \exp\{\lambda^2 \textstyle\sum_{i=1}^{J_n-1} Z_{ni}^2\}(1 + \lambda^2 Z_{nJ_n}^2)\\ &\leq \mathrm{e}^{2\lambda^2}\left(1 + \lambda^2 Z_{nJ_n}^2\right). \qquad (25.11)\end{aligned}$$

Then by (25.10),

$$\sup_n \mathrm{E}|T_{J_n}|^2 \leq \mathrm{e}^{2\lambda^2}\left(1 + \lambda^2 \sup_n \mathrm{E}(Z_{nJ_n}^2)\right) < \infty. \qquad (25.12)$$

Uniform boundedness of $\mathrm{E}|T_{J_n}|^2$ is sufficient for uniform integrability of T_{J_n} by **12.12**, proving (i).

Since by construction $\sum_{i=1}^{J_n-1} Z_{ni}^2 \leq 2$,

$$\begin{aligned}P\left(\sum_{i=1}^{J_n-1} Z_{ni}^2 \neq \sum_{i=1}^{r_n} Z_{ni}^2\right) &= P\left(\sum_{i=1}^{r_n} Z_{ni}^2 > 2\right)\\ &\leq P\left(\left|\sum_{i=1}^{r_n} Z_{ni}^2 - 1\right| > \varepsilon\right) \text{ for } 0 < \varepsilon < 1\\ &\to 0 \text{ as } n \to \infty \qquad (25.13)\end{aligned}$$

by assumption, which proves (ii). In addition,

$$P(S_{J_n} \neq S_{r_n}) \leq P\left(\sum_{i=1}^{r_n} Z_{ni}^2 > 2\right) \to 0 \text{ as } n \to \infty$$

so $|S_{J_n} - S_{r_n}| \to_{\mathrm{pr}} 0$ and by **23.18**, S_{J_n} and S_{r_n} have the same limiting distribution, proving (iii). ∎

25.2 The Martingale Case

Although it permits a law of large numbers, uncorrelatedness is not a strong enough assumption to yield a central limit result. The martingale difference assumption is similar to uncorrelatedness for practical purposes but it is strong enough and is attractive in other ways too. The next theorem shows how **25.1** applies to this case.

25.3 Theorem Let $\{X_{nt}, \mathcal{F}_{nt}\}$ be a martingale difference array with finite unconditional variances $\{\sigma_{nt}^2\}$ and $\sum_{t=1}^n \sigma_{nt}^2 = 1$. If

(a) $\sum_{t=1}^n X_{nt}^2 \to_{\text{pr}} 1$

(b) $\max_{1\le t\le n} |X_{nt}| \to_{\text{pr}} 0$

then $S_n = \sum_{t=1}^n X_{nt} \to_{\text{d}} \text{N}(0,1)$.

Proof Use **25.1** and **25.2** setting $r_n = n$, $i = t$, and $Z_{ni} = X_{nt}$. Conditions (a) and (b) are the same as (c) and (d) of **25.1**, so it remains to be shown that the other conditions of **25.1** are satisfied; not actually by X_{nt}, but by an equivalent sequence in the sense of **25.2**(iii).

If $T_n = \prod_{t=1}^n (1+\text{i}\lambda X_{nt})$, $\lim_{n\to\infty} \text{E}(T_n) = 1$ when $\{X_{nt}\}$ is a m.d. array. To show this, multiply out (25.1) repeatedly to get

$$\begin{aligned} T_n = \prod_{t=1}^n (1+\text{i}\lambda X_{nt}) &= T_{n-1} + \text{i}\lambda T_{n-1}X_{nn} \\ &= \dots \\ &= 1 + \text{i}\lambda \sum_{t=1}^n T_{t-1}X_{nt}. \end{aligned} \tag{25.14}$$

$T_{t-1} = \prod_{s=1}^{t-1}(1+\text{i}\lambda X_{ns})$ is an $\mathcal{F}_{n,t-1}$-measurable r.v., so by the LIE,

$$\begin{aligned} \text{E}(T_n) &= 1 + \text{i}\lambda \sum_{t=1}^n \text{E}(T_{t-1}X_{nt}) \\ &= 1 + \text{i}\lambda \sum_{t=1}^n \text{E}\big(T_{t-1}\text{E}(X_{nt}|\mathcal{F}_{n,t-1})\big) = 1. \end{aligned} \tag{25.15}$$

This is an exact result for any n. If X_{nt} is a m.d. so is the stopped process $\tilde{X}_{nt} = X_{nt}1\big(\sum_{k=1}^{t-1} X_{nk}^2 \le 2\big)$. The sequence $\tilde{T}_n = \prod_{t=1}^n (1+\text{i}\lambda\tilde{X}_{nt})$ satisfies **25.1**(b) as above and $\{\tilde{X}_{nt}\}$ also satisfies **25.1**(d) by condition (b) of the theorem. Since $\sum_{t=1}^n \text{E}(X_{nt}^2) = 1$, condition (25.10) certainly holds for X_{nt}. Hence, $\tilde{X}_{nt}$ satisfies **25.1**(a) and **25.1**(c) according to **25.2**(i) and (ii) and so obeys the CLT. The theorem now follows by **25.2**(iii). ∎

This theorem holds for independent sequences as a special case of m.d. sequences, but the conditions are slightly stronger than those of the Lindeberg theorem. Under independence, **25.3**(b) is equivalent to the Lindeberg condition by (24.53) when the CLT holds. However, **25.3**(a) is not a consequence of the Lindeberg condition. For the purpose of discussion, assume that $X_{nt} = X_t/s_n$, where $s_n^2 = \sum_{t=1}^n \sigma_t^2$ and $\sigma_t^2 = \mathrm{E}(X_t^2)$. Under independence, a sufficient extra condition for **25.3**(a) is that the sequence $\{X_t^2/\sigma_t^2\}$ be uniformly integrable. In this case, independence of $\{X_t\}$ implies independence of $\{X_t^2\}$, $\{X_t^2 - \sigma_t^2, \mathcal{F}_t\}$ is a m.d., and **20.12** (put $a_n = s_n^2$ and $b_t = \sigma_t^2$) gives sufficient conditions for $s_n^{-2}\sum_{t=1}^n (X_t^2 - \sigma_t^2) \to_{\mathrm{pr}} 0$. This is equivalent to **25.3**(a), but of course it is not the case that $\{X_t^2 - \sigma_t^2, \mathcal{F}_t\}$ is a m.d. merely because $\{X_t, \mathcal{F}_t\}$ is a m.d. If **25.3**(a) cannot be imposed in any other manner then $\mathrm{E}(X_t^2|\mathcal{F}_{t-1}\} = \sigma_t^2$ is required, a significant strengthening of the assumptions.

On the other hand, the theorem does not rule out trending variances. Following the approach of §24.4 such a case is as follows.

25.4 Theorem Let $\{X_t, \mathcal{F}_t\}$ be a square-integrable m.d. sequence with the property $\mathrm{E}(X_t^2|\mathcal{F}_{t-1}) = \sigma_t^2$ a.s. If there exists a sequence of positive constants $\{c_t\}$ such that $\{X_t^2/c_t^2\}$ is uniformly integrable and

$$\sup_n nM_n^2/s_n^2 < \infty \tag{25.16}$$

where $M_n^2 = \max_{1\le t\le n} c_t^2$, conditions **25.3**(a) and **25.3**(b) hold for $X_{nt} = X_t/s_n$.

Proof By **24.17** the sequence $\{X_{nt}\}$ satisfies the Lindeberg condition, and hence **25.3**(b) holds by **24.16**. Neither of these results imposes restrictions on the dependence of the sequence. To get **25.3**(a), apply **20.12** to the m.d. sequence $(X_t^2 - \sigma_t^2)$, putting $p = 1$, $b_t = c_t^2$, and $a_n = s_n^2$. The sequence $\{(X_t^2 - \sigma_t^2)/c_t^2\}$ is uniformly integrable on the assumptions and note that $\sum_{t=1}^n b_t \le nM_n^2 = O(s_n^2)$ and that $\sum_{t=1}^n b_t^2 \le nM_n^4 = o(s_n^4)$, both as consequences of (25.16). The conditions of **20.12** are therefore satisfied and the required result follows by **20.13**. ∎

25.3 Stationary Ergodic Sequences

It is easy to see that any stationary ergodic martingale difference having finite variance satisfies the condition of **25.3**. Under stationarity, finite variance is sufficient for the Lindeberg condition, which ensures **25.3**(b) by **24.16** and **25.3**(a) follows from the ergodicity by **14.6**. The next question is how the result might be extended to more general cases of dependence. Before proceeding to consider heterogeneous distributions of the increments the following theorem for the stationary ergodic case, due originally to M. I. Gordin, is of considerable interest.

This result has an unusual provenance, being published originally in Russian ([83]) and then misreported in several sources including [105] and [88].

The record has since been set straight in [24], [98], and [71], the latter source giving a corrected proof. The interest of the result is that a finite variance is not explicitly included in the conditions, although the case is different in kind from those such as **24.21**. The proof shows that the conditions are sufficient for the increments of the approximating martingale to have finite variance. The present statement of the result makes use of the familiar mixingale representation of dependence. The original formulation was in terms of a more primitive set of conditions, that the mixingale assumption conveniently delivers.

25.5 Theorem Let $\{X_t, \mathcal{F}_t\}$ be a stationary ergodic L_1-mixingale of size -1 and with $S_n = \sum_{t=1}^n X_t$ assume that

$$\limsup_{n\to\infty} n^{-1/2}\mathrm{E}|S_n| < \infty. \tag{25.17}$$

Then, $n^{-1/2}\mathrm{E}|S_n| \to \lambda$ for $0 \le \lambda < \infty$. If $\lambda > 0$ then $n^{-1/2}S_n \to_{\mathrm{d}} \mathrm{N}(0,\sigma^2)$ where $\sigma^2 = \frac{1}{2}\pi\lambda^2$.

Proof Apply the decomposition of **17.7** to write $X_t = Y_t + Z_t - Z_{t+1}$ where Y_t is a stationary ergodic m.d. and Z_t is stationary with $\mathrm{E}|Z_1| < \infty$. Hence $S_n = \sum_{t=1}^n Y_t + Z_1 - Z_{n+1}$ and by (25.17) there exists $A < \infty$ such that

$$\begin{aligned}\limsup_{n\to\infty} n^{-1/2}\mathrm{E}\left|\sum_{t=1}^n Y_t\right| &= \limsup_{n\to\infty} n^{-1/2}\mathrm{E}|S_n - Z_1 + Z_{n+1}| \\ &\le \limsup_{n\to\infty} n^{-1/2}\big(\mathrm{E}|S_n| + 2\mathrm{E}|Z_1|\big) \le A. \end{aligned}\tag{25.18}$$

According to the ergodic theorem **14.6**, $n^{-1}\sum_{t=1}^n Y_t^2 \to \sigma^2$ with probability 1. Either $\sigma^2 = \infty$ or $\sigma^2 < \infty$, but in view of Theorem **16.29** and (25.18), a constant $C > 0$ can be chosen large enough that

$$P\left(\left(\frac{1}{n}\sum_{t=1}^n Y_t^2\right)^{1/2} > C\right) \le \frac{3}{C\sqrt{n}}\mathrm{E}\left|\sum_{t=1}^n Y_t\right| \to \frac{3A}{C} < 1. \tag{25.19}$$

(The arrow could indicate convergence on a subsequence.) Since $\sigma^2 = \infty$ would contradict (25.19) it must be the case that $\sigma^2 < \infty$. If $\sigma^2 > 0$, **25.3** and **23.18** imply that $n^{-1/2}S_n \to_{\mathrm{d}} \mathrm{N}(0,\sigma^2)$. In this case, the sequence $n^{-1/2}|\sum_{t=1}^n Y_t|$ is uniformly integrable in view of (25.19) and **12.13** and by the continuous mapping theorem it converges in distribution to the half-Gaussian limit (see **8.20**). Hence, by **9.15**,

$$n^{-1/2}\mathrm{E}\left|\sum_{t=1}^{n} Y_t\right| \to \sigma\sqrt{\frac{2}{\pi}}. \tag{25.20}$$

It follows that $\lambda = \sigma(2/\pi)^{1/2}$ and $\sigma^2 = 0$ implies $\lambda = 0$. ∎

This result might be thought of as a counterpart for dependent sequences of the Lindeberg–Levy theorem, but unlike that case there is no need to assume explicitly that $\mathrm{E}(X_1^2) < \infty$. Independence of $\{X_t\}$ enforces the condition $Z_t = 0$ for all t and then the conditions of the theorem imply $\mathrm{E}(X_1^2) = \sigma^2$. It might appear that the existence of dependence weakens the moment restrictions required for the CLT, but this gain is more technical than real for it is not obvious how to construct a stationary sequence such that S_1 is not square-integrable but Y_1 is. The most useful implication is that the independence assumption can be replaced by arbitrary local dependence (controlled by the mixingale assumption) without weakening any of the conclusions of the Lindeberg–Levy theorem.

25.4 The CLT for Mixingales

The traditional approach to the problem of general dependence is the so-called method of 'Bernstein sums' ([16])—that is, break up S_n into blocks (partial sums) and consider the sequence of blocks. Each block must be so large, relative to the rate at which the memory of the sequence decays, that the degree to which the next block can be predicted from observing the 'current' block is negligible. To render the blocks effectively independent of each other in large samples, small blocks are omitted from the sum to separate the large ones, their omission being asymptotically negligible. Let b_n denote the total length of a block and l_n the length of the little initial block to be omitted. Then $r_n = [n/b_n]$ is the number of complete blocks with a possible final fragment of length $n - r_n b_n$. It must be the case that $b_n/n \to 0$ and also that $l_n/b_n \to 0$. The following theorem of de Jong ([51]) specifies sufficient conditions for a CLT to hold.

25.6 Theorem Let the following conditions hold for a triangular array of random variables $\{X_{nt}\}_{t=1}^{n}$, $n \geq 1$.

(a) $\{X_{nt}\}$ is an L_2-mixingale of size $-\frac{1}{2}$ with respect to filtrations $\{\mathcal{F}_{nt}\}$ and constants $\{c_{nt}\}$.

(b) The array $\{X_{nt}^2/c_{nt}^2\}$ is uniformly integrable.

(c) Setting $b_n = [n^{1-\alpha}]$ for some $\alpha \in (0,1]$ and defining

$$M_{nj} = \max_{(j-1)b_n+1 \leq t \leq jb_n} c_{nt}$$

for $j = 1, \dots, r_n$ and $M_{n,r_n+1} = \max_{r_nb_n+1\le t\le n} c_{nt}$,

$$\max_{1\le j\le r_n+1} M_{nj} = o(b_n^{-1/2}) \tag{25.21}$$

and

$$\sum_{j=1}^{r_n} M_{nj}^2 = O(b_n^{-1}). \tag{25.22}$$

(d) Setting $l_n = [n^{1-\alpha-\varepsilon}]$ for $\varepsilon > 0$, $\sum_{j=1}^{r_n} Z_{nj}^2 \to_{\text{pr}} 1$ where

$$Z_{nj} = \sum_{t=(j-1)b_n+l_n+1}^{jb_n} X_{nt}. \tag{25.23}$$

Then, $\sum_{t=1}^n X_{nt} \to_{\text{d}} \text{N}(0,1)$. □

The roles of these conditions are not particularly transparent and in reviewing them it will be helpful to keep in mind the leading case with $X_{nt} = (Y_t - \mu_t)/s_n$ where $\mu_t = \text{E}(Y_t)$ and $s_n^2 = \text{E}\big(\sum_{t=1}^n (Y_t - \mu_t)\big)^2$, although more general interpretations are possible as noted in §24.2. In this case it would often be legitimate to choose

$$c_{nt} = \max\{\sigma_t, 1\}/s_n \tag{25.24}$$

where σ_t^2 is the variance of Y_t. The c_{nt} have to be thought of as tending to zero with n, although possibly growing or shrinking with t also, subject to **25.6**(c). Because the sequence is in general autocorrelated, s_n^2 is no longer just the partial sum of the variances, but is defined as

$$s_n^2 = \sum_{t=1}^n \sigma_t^2 + 2\sum_{t=2}^n \sum_{k=1}^{t-1} \sigma_{t,t-k} \tag{25.25}$$

where $\sigma_{t,t-k} = \text{Cov}(Y_t, Y_{t-k})$.

Proof of 25.6 This proceeds by decomposing the sum $\sum_{t=1}^n X_{nt}$ into components which either vanish or satisfy conditions for the CLT. The decomposition occurs in two stages. The first is

$$\sum_{t=1}^{n} X_{nt} = \sum_{j=1}^{r_n} Z_{nj} + \sum_{j=1}^{r_n} \sum_{t=(j-1)b_n+1}^{(j-1)b_n+l_n} X_{nt} + \sum_{t=r_nb_n+1}^{n} X_{nt}$$

where Z_{nj} is defined by (25.23). Because $\{X_{nt}\}$ is an L_2-mixingale of size $-\frac{1}{2}$, Corollary **17.11** adapted to the array case implies in view of assumption (c) that

$$\mathrm{E}\Bigg(\sum_{j=1}^{r_n} \sum_{t=(j-1)b_n+1}^{(j-1)b_n+l_n} X_{nt}\Bigg)^2 \ll \sum_{j=1}^{r_n} l_n M_{nj}^2 = O(l_n/b_n) = O(n^{-\varepsilon}). \tag{25.26}$$

Similarly,

$$\mathrm{E}\Bigg(\sum_{t=r_nb_n+1}^{n} X_{nt}\Bigg)^2 \ll b_n M_{n,r_n+1}^2 = o(1). \tag{25.27}$$

The proof is completed by showing the limiting distribution of $\sum_{j=1}^{r_n} Z_{nj}$. Let $\mathrm{E}_{nj}(\cdot)$ stand for $\mathrm{E}(\cdot|\mathcal{F}_{n,jb_n+l_n/2})$, so that the conditioning information for $\mathrm{E}_{nj}(\cdot)$ leads the j^{th} block by $l_n/2$ periods and that for $\mathrm{E}_{n,j-1}(\cdot)$ lags the j^{th} block by $l_n/2$ periods. A martingale difference with respect to this filtration is $\{W_{nj}, \mathcal{F}_{n,jb_n+l_n/2}\}$ where

$$W_{nj} = \mathrm{E}_{nj}Z_{nj} - \mathrm{E}_{n,j-1}Z_{nj} = \sum_{t=(j-1)b_n+l_n+1}^{jb_n} \mathrm{E}_{nj}X_{nt} - \mathrm{E}_{n,j-1}X_{nt}. \tag{25.28}$$

The second decomposition is

$$\sum_{j=1}^{r_n} Z_{nj} = \sum_{j=1}^{r_n} (Z_{nj} - \mathrm{E}_{nj}Z_{nj}) + \sum_{j=1}^{r_n} W_{nj} + \sum_{j=1}^{r_n} \mathrm{E}_{n,j-1}Z_{nj}. \tag{25.29}$$

The first and last right-hand-side terms of (25.29) are shown to vanish in mean square, as follows. Considering the last term, note that for $t > b_n(j-1) + l_n$, if $0 \le m < t - b_n(j-1) - l_n/2$ then

$$||\mathrm{E}(\mathrm{E}_{n,j-1}X_{nt}|\mathcal{F}_{n,t-m})||_2 = ||\mathrm{E}_{n,j-1}X_{nt}||_2 \le c_{nt}\zeta(l_n/2)$$

whereas for $m \ge t - b_n(j-1) - l_n/2$,

$$||\mathrm{E}(\mathrm{E}_{n,j-1}X_{nt}|\mathcal{F}_{n,t-m})||_2 = ||\mathrm{E}_{n,t-m}X_{nt}||_2 \le c_{nt}\zeta(m).$$

Also, because $\mathrm{E}_{n,j-1}X_{nt}$ is $\mathcal{F}_{n,(j-1)b_n+l_n/2}$-measurable,

$$||\mathrm{E}_{n,j-1}X_{nt} - \mathrm{E}(\mathrm{E}_{n,j-1}X_{nt}|\mathcal{F}_{n,t+m})||_2 = 0.$$

It follows that $\{\mathrm{E}_{n,j-1}X_{nt}, \mathcal{F}_{nt}\}$ is an L_2-mixingale with respect to constants c_{nt} and mixingale indices that can be set at either $\zeta(m)$ or at $\zeta(l_n/2)$ and hence at $\zeta(l_n/2)^{\eta}\zeta(m)^{1-\eta}$ for any $\eta \in [0,1]$. By setting η small enough, the constant array can be chosen as $c_{nt}\zeta(l_n/2)^{\eta}$, while the mixingale size can be set as close to $-\frac{1}{2}$ as desired. Therefore, applying **17.11** and (25.22) as above,

$$\begin{aligned}\mathrm{E}\Bigl(\sum_{j=1}^{r_n}\mathrm{E}_{n,j-1}Z_{nj}\Bigr)^2 &= \mathrm{E}\Bigl(\sum_{j=1}^{r_n}\sum_{t=(j-1)b_n+l_n+1}^{jb_n}\mathrm{E}_{n,j-1}X_{nt}\Bigr)^2\\ &\ll \zeta(l_n/2)^{2\eta}b_n\sum_{j=1}^{r_n}M_{nj}^2 = O(\zeta(l_n/2)^{2\eta}) = o(1).\end{aligned} \tag{25.30}$$

Closely analogous arguments applied to the first right-hand-side term of (25.29) show that $\{X_{nt} - \mathrm{E}_{nj}X_{nt}, \mathcal{F}_{nt}\}$ is also an L_2-mixingale. On one hand,

$$||\mathrm{E}(X_{nt} - \mathrm{E}_{nj}X_{nt}|\mathcal{F}_{n,t-m})||_2 = 0.$$

On the other, if $m \le jb_n + l_n/2 - t$ then

$$\begin{aligned}||X_{nt} - \mathrm{E}_{nj}X_{nt} - \mathrm{E}(X_{nt} - \mathrm{E}_{nj}X_{nt}|\mathcal{F}_{n,t+m})||_2 &= ||X_{nt} - \mathrm{E}_{nj}X_{nt}||_2\\ &\le c_{nt}\zeta(l_n/2)\end{aligned}$$

and otherwise,

$$\begin{aligned}||X_{nt} - \mathrm{E}_{nj}X_{nt} - \mathrm{E}(X_{nt} - \mathrm{E}_{nj}X_{nt}|\mathcal{F}_{n,t+m})||_2 &= ||X_{nt} - \mathrm{E}(X_{nt}|\mathcal{F}_{n,t+m})||_2\\ &\le c_{nt}\zeta(m).\end{aligned}$$

It follows exactly as for (25.30) that

$$\mathrm{E}\Bigl(\sum_{j=1}^{r_n}Z_{nj} - \mathrm{E}_{nj}Z_{nj}\Bigr)^2 = O\bigl(\zeta(l_n/2)^{2\eta}\bigr) = o(1). \tag{25.31}$$

The remaining step is to verify the conditions of Theorem **25.3** for the second sum of terms in (25.29). Condition (a) of **25.3** follows from condition (d) of this theorem by the following argument. For random variables $A = Z^2_{nj}$ and $B = W^2_{nj}$ the Cauchy–Schwarz inequality gives

$$\mathrm{E}|A^2 - B^2| = \mathrm{E}|(A-B)(A+B)| \leq ||A-B||_2||A+B||_2.$$

Therefore, substituting from (25.28),

$$\begin{aligned}
&\mathrm{E}\left|\sum_{j=1}^{r_n} Z^2_{nj} - \sum_{j=1}^{r_n} W^2_{nj}\right| \\
&\qquad\leq \sum_{j=1}^{r_n} ||(Z_{nj} - \mathrm{E}_{nj}Z_{nj}) + \mathrm{E}_{n,j-1}Z_{nj}||_2||Z_{nj} + \mathrm{E}_{nj}Z_{nj} - \mathrm{E}_{n,j-1}Z_{nj}||_2 \\
&\qquad\leq 3\sum_{j=1}^{r_n}\big(||\mathrm{E}_{n,j-1}Z_{nj}||_2 + ||Z_{nj} - \mathrm{E}_{nj}Z_{nj}||_2\big)||Z_{nj}||_2 = o(1) \qquad (25.32)
\end{aligned}$$

where the order of magnitude is obtained from (25.30) and (25.31). These two sums of squares converge in L_1 to the same limit and since this limit is 1 in the first case by assumption (d), it is 1 in the second case.

To show that the $\{W_{nj}\}$ satisfy condition (b) of Theorem **25.3**, it is sufficient by **24.16** to show that they satisfy the Lindeberg condition—in other words that $\sum_{j=1}^{r_n} \mathrm{E}(W^2_{nj}1_{\{|W_{nj}|>\varepsilon\}}) \to 0$ as $n \to \infty$ for $\varepsilon > 0$. This is shown as follows. For each $j = 1, \ldots, r_n$ the array $\{\mathrm{E}_{nj}X_{nt} - \mathrm{E}_{n,j-1}X_{nt}, \mathcal{F}_{nt}\}_{t=(j-1)b_n+l_n+1}^{jb_n}$ is an L_2-mixingale of size $-\frac{1}{2}$ with respect to constants $2c_{nt}$. Specifically it is the sum of two L_2-mixingales, each with constants c_{nt}, since

$$||\mathrm{E}(\mathrm{E}_{nj}X_{nt}|\mathcal{F}_{n,t-m})||_2 = ||\mathrm{E}(X_{nt}|\mathcal{F}_{n,t-m})||_2 \leq c_{nt}\zeta(m)$$

and

$$||\mathrm{E}(\mathrm{E}_{n,j-1}X_{nt}|\mathcal{F}_{n,t-m})||_2 \leq c_{nt}\max\{\zeta(m), \zeta(t - b_n(j-1) - l_n/2)\}$$

whereas

$$||\mathrm{E}_{nj}X_{nt} - \mathrm{E}(\mathrm{E}_{nj}X_{nt}|\mathcal{F}_{n,t+m})||_2 \leq c_{nt}\zeta(m)$$

noting that the minorant vanishes for $m \geq jb_n + l_n/2 - t$, while

$$||\mathrm{E}_{n,j-1}X_{nt} - \mathrm{E}(\mathrm{E}_{n,j-1}X_{nt}|\mathcal{F}_{n,t+m})||_2 = 0.$$

Further, the collections $\{(\mathrm{E}_{nj}X_{nt} - \mathrm{E}_{n,j-1}X_{nt})^2/c_{nt}^2, j = 1, \dots, r_n\}$ are uniformly integrable for each t in view of assumption (b) and **12.14**. Hence, defining

$$v_{nj}^2 = \sum_{t=(j-1)b_n+l_n+1}^{jb_n} c_{nt}^2 \tag{25.33}$$

the array $\{W_{nj}^2/v_{nj}^2, j = 1, \dots, r_n, n \in \mathbb{N}\}$ is uniformly integrable according to Corollary **17.15**. This means that

$$\begin{aligned}\sum_{j=1}^{r_n} \mathrm{E}(W_{nj}^2 1_{\{|W_{nj}|>\varepsilon\}}) &\le \max_{1\le j\le r_n} \mathrm{E}(W_{nj}^2/v_{nj}^2 1_{\{|W_{nj}|/v_{nj}>\varepsilon/v_{nj}\}}) \sum_{j=1}^{r_n} b_n M_{nj}^2 \\ &= O\Big(\max_{1\le j\le r_n} \mathrm{E}(W_{nj}^2/v_{nj}^2 1_{\{|W_{nj}|/v_{nj}>\varepsilon/v_{nj}\}})\Big) \\ &= o(1) \end{aligned} \tag{25.34}$$

where the first order of magnitude follows from (25.22) and the second is by the uniform integrability and the fact that

$$v_{nj}^2 \le \max_{1\le j\le r_n} b_n M_{nj}^2 = o(1)$$

by (25.21). This completes the proof. ∎

There is a simpler condition that is sufficient for condition (c) of **25.6**, giving rise to the following corollary.

25.7 Corollary Letting $M_n = \max_{1\le t\le n}\{c_{nt}\}$,

$$\sup_n nM_n^2 < \infty \tag{25.35}$$

is sufficient for (25.21) and (25.22) to hold.

Proof If (25.35) holds then $M_{nj} = O(n^{-1/2})$ for each j. Since $r_n b_n \simeq n$, (25.21) and (25.22) hold for any α in $(0, 1]$. ∎

Assumption (c) of **25.6** permits forms of global nonstationarity, as discussed in §13.1 and §24.4. The following cases illustrate what is allowed.

25.8 Example Let $\{X_t\}$ be an independent zero-mean sequence, so that $s_n^2 = \sum_{t=1}^n \sigma_t^2$, with variances $\sigma_t \simeq t^\beta$ for any $\beta \ge 0$ (compare **13.5**). It is straightforward to verify that (25.35) is satisfied for $c_{nt} = \sigma_t/s_n$ and the theorem then holds for the case $X_{nt} = X_t/s_n$ subject to condition (b). It is however violated when $\sigma_t^2 \simeq 2^t$, a case incompatible with the asymptotic negligibility of individual summands (compare **24.18**). It is also violated when $\beta < 0$. (See the discussion of this case following Lemma **25.10**.) □

25.9 Example Let $\{X_t\}$ be an independent zero-mean sequence with variance sequence generated by the scheme described in **13.6**. Putting $X_{nt} = X_t/s_n$, (25.35) is satisfied with $c_{nt} = 1/s_n$ for all t. □

Among the cases that (25.35) rules out are asymptotically degenerate sequences, having $\sigma_t^2 \to 0$ as $t \to \infty$. In these cases, $\max_{1\le t\le n}\sigma_t^2 = O(1)$ as $n \to \infty$, but if $s_n^2/n \to 0$ then, with $c_{nt} = \max\{\sigma_t, 1\}/s_n$, (25.35) is violated. It is certain of these cases that assumption **25.6**(c) is designed to allow. To see what is going on here, it is easiest to think in terms of the array $\{c_{nt}\}$ as varying regularly with n and t, within certain limits to be determined. The conditions of the following lemma are somewhat more general than required, but are easily specialized.

25.10 Lemma Suppose $c_{nt}^2 \simeq t^\beta n^{-\gamma-1}$ for $\beta, \gamma \in \mathbb{R}$. Then **25.6**(c) holds if

$$\gamma \ge \beta \tag{25.36}$$

and if $\gamma < 0$,

$$\alpha > -\gamma. \tag{25.37}$$

Proof Either $M_{nj}^2 \simeq (jb_n)^\beta n^{-\gamma-1}$ for $\beta \ge 0$, or $M_{nj}^2 \simeq ((j-1)b_n)^\beta n^{-\gamma-1}$ for $\beta < 0$. In both cases,

$$\sum_{j=1}^{r_n} M_{nj}^2 \simeq r_n^{1+\beta} b_n^\beta n^{-\gamma-1} \simeq n^{\alpha(1+\beta)+(1-\alpha)\beta-\gamma-1}. \tag{25.38}$$

Simplifying shows that condition (25.36) is necessary and sufficient for (25.22), independently of the value of α, note. Next, (25.21) is equivalent to

$$\max_{1\le t\le n} c_{nt}^2 = o(n^{\alpha-1}), \tag{25.39}$$

which, since the maximum is at $t = n$ for $\beta \ge 0$ and $t = 1$ otherwise, imposes the requirement

$$\alpha > \max\{\beta, 0\} - \gamma. \tag{25.40}$$

In view of (25.36) the choice of α is constrained by (25.37) only if $\gamma < 0$ and hence $\beta < 0$. ∎

While β and γ can be of either sign subject to the indicated constraints, don't overlook that while the conditions of **25.10** are sufficient for **25.6**(c) the values must also be such as to validate conditions (b) and (d) of the theorem.

Consider in particular how the conditions of **25.6** apply in the case $X_{nt} = (Y_t - \mu_t)/s_n$ with s_n^2 given by (25.25). If $\sigma_t^2 \simeq t^\beta$ with $\beta \geq 0$, c_{nt} can be made monotone in t (not essential, but analytically convenient) by setting

$$c_{nt} = \max_{1\leq s\leq t}\{\sigma_s\}/s_n. \tag{25.41}$$

Further, write the k^{th} autocovariance as $\sigma_{t,t-k} = \sigma_t\sigma_{t-k}\rho_k$ where $\rho_k = O(k^{-b})$. From **17.18** it can be verified that $b > 1$ under the specified conditions. Substituting into (25.25), Theorem **2.19** yields

$$s_n^2 \ll \sum_{t=1}^{n} t^\beta + 2\sum_{t=1}^{n} t^{\beta/2}\sum_{k=1}^{t-1}(t-k)^{\beta/2}\rho_k = O(n^{\beta+1}) \tag{25.42}$$

and hence, $\gamma = \beta$.

Now consider the case $\sigma_t^2 \simeq t^\beta$ with $\beta < 0$. It might be possible to set

$$c_{nt} = \sup_{s\geq t}\{\sigma_s\}/s_n \tag{25.43}$$

and so $c_{nt}^2 \simeq t^\beta n^{-\beta-1}$ under (25.36), but here $M_n^2 = \sigma_{t^*}^2/s_n^2$ for some $t^* < \infty$. Hence, $M_n^2 \simeq n^{-\beta-1}$ and **25.7** ceases to apply. However, (25.21) and (25.22) can still hold, although with $\beta = \gamma$, condition (25.37) now imposes a restriction on the blocking scheme, with $\alpha > -\beta$, showing that $\beta \geq -1$ is the lower bound. The case $\alpha = 1$ applies in the case of martingale differences, where $s_n^2 = \sum_{t=1}^n \sigma_t^2$. Then, (25.22) is satisfied with $b_n = 1$ and $l_n = 0$ by construction and (25.21) requires only that $s_n^2 \uparrow \infty$ and hence $\sigma_t^2 \simeq t^\beta$ is permitted for any $\beta \geq -1$. This result may be compared with **25.4**, whose conditions it extends rather as **25.6** extends **25.7**. As an example, let $Y_t - \mu_t$ be a m.d. with $\sigma_t^2 = \sigma^2/t$ where σ^2 is constant, so that $s_n^2 \simeq \log n$. This case of **25.6** gives the result $(\log n)^{-1/2}\sum_{t=1}^n (Y_t - \mu_t) \to_d \mathrm{N}(0,\sigma^2)$. However, if the variances are summable the central limit theorem surely fails. Here is a well-known case where the non-summability condition is violated.

25.11 Example Consider the sequence $\{Y_t\}$ of first differences with $Y_1 = Z_1$ and

$$Y_t = Z_t - Z_{t-1}, t = 2, 3, \ldots,$$

where $\{Z_t\}$ is an identically and independently distributed sequence. Here $\{Y_t\}$ satisfies **25.6**(a) and **25.6**(b), but $\sum_{t=1}^n Y_t = Z_n$ and $s_n^2 = \mathrm{Var}(Z_n)$. □

25.5 NED Functions of Mixing Processes

Theorem **25.6** is a fundamental result, but it is not operational until a way is found of establishing the high-level conditions (a) and (d). A near-epoch dependence-on-mixing assumption is the natural way to achieve (a), although the pure mixing case is also covered. There are two versions of the result, for the strong mixing and uniform mixing cases respectively.

25.12 Theorem Let the following conditions hold for a triangular array of random variables $\{X_{nt}\}_{t=1}^{n}$, $n \geq 1$.

(a) $\|\sum_{t=1}^{n} X_{nt}\|_2 = 1$.

(b) $\{X_{nt}\}$ is L_2-near epoch dependent of size $-\frac{1}{2}$ with respect to a constant array $\{d_{nt}\}$ on either an α-mixing array $\{V_{nt}\}$ of size $-r/(r-2)$ for $r > 2$ or a ϕ-mixing array $\{V_{nt}\}$ of size $-r/(2r-2)$ for $r \geq 2$.

(c) There exists a positive constant array $\{c_{nt}\}$ such that $\{X_{nt}/c_{nt}\}$ is L_r-bounded uniformly in t and n and uniformly square-integrable if $r = 2$, $\{d_{nt}/c_{nt}\}$ is uniformly bounded in t and n, and condition (c) of Theorem **25.6** is satisfied.

Then, $\sum_{t=1}^{n} X_{nt} \to_{\mathrm{d}} \mathrm{N}(0,1)$.

Proof The method of proof is to show that the conditions of **25.6** are all satisfied. By **18.7**(i) in the α-mixing case and **18.7**(ii) in the ϕ-mixing case, $\{X_{nt}\}$ is an L_2-mixingale of size $-\frac{1}{2}$ with respect to constants $c_{nt} \ll \max\{\|X_{nt}\|, d_{nt}\}$ and **25.6**(b) holds by **12.12** if $r > 2$. It therefore remains to be shown that **25.6**(d) holds.

Invoking assumption (a), write

$$\sum_{j=1}^{r_n} Z_{nj}^2 - 1 = A_n - B_n \tag{25.44}$$

where

$$A_n = \sum_{j=1}^{r_n} (Z_{nj}^2 - \mathrm{E}Z_{nj}^2) \tag{25.45}$$

and

$$B_n = \mathrm{E}\Bigl(\sum_{t=1}^{n} X_{nt}\Bigr)^2 - \sum_{j=1}^{r_n} \mathrm{E}Z_{nj}^2. \tag{25.46}$$

B_n is a nonstochastic sequence that must be shown to converge to zero, while A_n must be shown to converge to zero in probability. This latter result is shown first. Since $\{X_{nt}^2/c_{nt}^2\}$ is uniformly integrable, if the sequence v_{nj}^2 is defined in (25.33) then $\{Z_{nj}^2/v_{nj}^2\}$ is also uniformly integrable by Corollary **17.15**. Let $\tilde{Z}_{nj} = \phi_{K_\delta v_{nj}}(Z_{nj})$

represent the truncation of Example **18.14** where K_δ is a constant to be chosen. Since $\sum_{j=1}^{r_n} v_{nj}^2 = O(1)$ by (25.22), for any $\delta > 0$ there exists K_δ large enough that

$$\begin{aligned}
\mathrm{E}\left|\sum_{j=1}^{r_n}(Z_{nj}^2 - \tilde{Z}_{nj}^2)\right| &\le 2\sum_{j=1}^{r_n}\mathrm{E}\big(Z_{nj}^2 1_{\{Z_{nj}^2 > K_\delta^2 v_{nj}^2\}}\big) \\
&\le 2\sup_n \max_{1\le j\le r_n} \mathrm{E}\big((Z_{nj}^2/v_{nj}^2)1_{\{Z_{nj}^2/v_{nj}^2 > K_\delta^2\}}\big)\sum_{j=1}^{r_n} v_{nj}^2 \\
&\le \delta.
\end{aligned}$$

Therefore it remains to be shown that for any $\delta > 0$,

$$\lim_{n\to\infty} \mathrm{E}\left|\sum_{j=1}^{r_n}(\tilde{Z}_{nj}^2 - \mathrm{E}\tilde{Z}_{nj}^2)\right| = 0. \tag{25.47}$$

The object will be to show that $\tilde{Z}_{nj}^2 - \mathrm{E}\tilde{Z}_{nj}^2$ is an L_2-mixingale of size $-\frac{1}{2}$, allowing the application of **17.11**. The truncation is a useful transformation for this purpose. Recall the notation $\mathrm{E}_{nj}(\cdot) = \mathrm{E}(\cdot|\mathcal{F}_{n,jb_n+l_n/2})$ defined above (25.28). In a similar way, for $k > j$ let $\mathrm{E}_{nj}^k(\cdot) = \mathrm{E}\big(\cdot|\mathcal{F}_{n,jb_n+l_n/2}^{kb_n+l_n/2}\big)$ where $\mathcal{F}_{n,jb_n+l_n/2}^{kb_n+l_n/2} = \sigma(\boldsymbol{V}_{n,(j-1)b_n+l_n/2}, \dots, \boldsymbol{V}_{n,kb_n+l_n/2})$. For $m \ge 1$ note that

$$\begin{aligned}
\left\|\mathrm{E}_{n,j-2m}(\tilde{Z}_{nj}^2 - \mathrm{E}\tilde{Z}_{nj}^2)\right\|_2 &\le \left\|\mathrm{E}_{n,j-2m}(\tilde{Z}_{nj}^2 - \mathrm{E}_{n,j-m}^{j+m}\tilde{Z}_{nj}^2)\right\|_2 \\
&\quad + \left\|\mathrm{E}_{n,j-2m}(\mathrm{E}_{n,j-m}^{j+m}\tilde{Z}_{nj}^2 - \mathrm{E}\tilde{Z}_{nj}^2)\right\|_2 \\
&\le \left\|\tilde{Z}_{nj}^2 - \mathrm{E}_{n,j-m}^{j+m}\tilde{Z}_{nj}^2\right\|_2 \\
&\quad + \left\|\mathrm{E}_{n,j-2m}(\mathrm{E}_{n,j-m}^{j+m}\tilde{Z}_{nj}^2) - \mathrm{E}\tilde{Z}_{nj}^2\right\|_2
\end{aligned} \tag{25.48}$$

where the second inequality uses **10.15**. Also note that

$$\left\|\tilde{Z}_{nj}^2 - \mathrm{E}_{n,j+2m}\tilde{Z}_{nj}^2\right\|_2 \le \left\|\tilde{Z}_{nj}^2 - \mathrm{E}_{n,j-m}^{j+m}\tilde{Z}_{nj}^2\right\|_2 \tag{25.49}$$

since the conditioning information set in the minorant is larger than that in the majorant. These L_2 norms (involving fourth moments) must exist thanks to the truncation. To bound the term appearing on the majorant sides of both (25.48) and (25.49), note that

$$
\begin{aligned}
\left\|\tilde{Z}^2_{nj} - \mathrm{E}^{j+m}_{n,j-m}\tilde{Z}^2_{nj}\right\|_2 &\le \left\|\tilde{Z}^2_{nj} - \phi_{K_\delta \nu_{nj}}(\mathrm{E}^{j+m}_{n,j-m}Z^2_{nj})\right\|_2 \\
&\le K_\delta \nu_{nj}\left\|\phi_{K_\delta \nu_{nj}}(Z_{nj}) - \phi_{K_\delta \nu_{nj}}(\mathrm{E}^{j+m}_{n,j-m}Z_{nj})\right\|_2 \\
&\le K_\delta \nu_{nj} \sum_{t=(j-1)b_n+l_n+1}^{jb_n} \left\|X_{nt} - \mathrm{E}(X_{nt}|\mathcal{F}^{(j+m)b_n+l_n/2}_{n,(j-m-1)b_n+l_n/2})\right\|_2 \\
&\le K_\delta \nu_{nj} \sum_{t=(j-1)b_n+l_n+1}^{jb_n} d_{nt}\nu(mb_n + l_n/2) \\
&= O(b_n M^2_{nj} m^{-1/2-\varepsilon})
\end{aligned}
\qquad (25.50)
$$

for $\varepsilon > 0$. Here, the first inequality uses the fact that the conditional expectation minimizes the mean squared prediction error, the second uses the truncation, the third uses (18.35) and the Minkowski inequality, and the fourth applies assumption (b) of the theorem. The final order of magnitude follows because $K_\delta < \infty$, $\nu_{nj} = O(b_n^{1/2}M_{nj})$ by (25.21) and the d_{nt} is are likewise bounded by M_{nj} according to assumption (c) and $\nu(mb_n + l_n/2) = O(b_n^{-1/2}m^{-1/2-\varepsilon})$.

Consider the second term in the majorant of (25.48). $\mathrm{E}^{j+m}_{n,j-m}\tilde{Z}^2_{nj}$ is a finite-lag function of mixing random variables and hence is itself mixing. By **15.2** and **15.4**,

$$
\begin{aligned}
\left\|\mathrm{E}_{n,j-2m}(\mathrm{E}^{j+m}_{n,j-m}\tilde{Z}^2_{nj}) - \mathrm{E}\tilde{Z}^2_{nj}\right\|_2 &\le \begin{cases} 5\left\|\tilde{Z}^2_{nj}\right\|_r \alpha(mb_n)^{1/2-1/r}, & \alpha\text{-mixing case} \\ 2\left\|\tilde{Z}^2_{nj}\right\|_r \phi(mb_n)^{1-1/r}, & \phi\text{-mixing case} \end{cases} \\
&= O(b_n M^2_{nj} m^{-1/2-\varepsilon}).
\end{aligned}
\qquad (25.51)
$$

Bounding (25.48) and (25.49) by (25.50) and (25.51) leads to the conclusion that $\{\tilde{Z}^2_{nj}\}$ is an L_2-mixingale array of size $-\frac{1}{2}$ with respect to constants $\{\nu^2_{nj}\}$. It follows by **17.11** that

$$
\mathrm{E}\left(\sum_{j=1}^{r_n}(\tilde{Z}^2_{nj} - \mathrm{E}\tilde{Z}^2_{nj})\right)^2 \ll \sum_{j=1}^{r_n}\nu^4_{nj} \le b_n^2 \max_{1\le j\le r_n} M^2_{nj}\sum_{j=1}^{r_n} M^2_{nj} = o(1)
$$

by assumption (c) of the theorem.

The remaining step of the proof is to show $B_n \to 0$ in (25.46). This is done by application of Corollary **17.18**, which supplies a bound on the autocovariances of a mixingale array. The sum may be decomposed as $B_n = B_{1n} + 2B_{2n} + 2B_{3n}$ where, with any empty sums equalling zero,

$$B_{1n} = \sum_{j=1}^{r_n}\Bigg(\sum_{t=(j-1)b_n+1}^{(j-1)b_n+l_n} \mathrm{E}X_{nt}\Bigg)^2 + \Bigg(\sum_{t=r_nb_n+1}^{n} \mathrm{E}X_{nt}\Bigg)^2$$

$$B_{2n} = \sum_{j=1}^{r_n}\Bigg(\sum_{t=(j-1)b_n+1}^{(j-1)b_n+l_n}\;\sum_{s=(j-1)b_n+l_n+1}^{\min\{jb_n+l_n,n\}} + \sum_{t=(j-1)b_n+l_n+1}^{jb_n}\;\sum_{s=jb_n+1}^{\min\{jb_n+l_n,n\}}\Bigg)\mathrm{E}X_{nt}X_{ns}$$

$$B_{3n} = \sum_{j=1}^{r_n}\;\sum_{t=(j-1)b_n+1}^{jb_n}\;\sum_{s=jb_n+l_n+1}^{n} \mathrm{E}X_{nt}X_{ns}.$$

$B_{1n} = o(1)$ by (25.26) and (25.27). Since $\sum_{t=1}^{n} c_{nt} = O(1)$, which follows from (25.22), **17.18** implies that B_{3n} is $O(l_n^{-\delta})$ for $\delta > 0$. To deal with B_{2n} define

$$B_{2n}^{*} = \sum_{j=1}^{r_n}\Bigg(\sum_{t=(j-1)b_n+1}^{(j-1)b_n+l_n}\;\sum_{s=(j-1)b_n+2l_n+1}^{\min\{jb_n+l_n,n\}} + \sum_{t=(j-1)b_n+l_n+1}^{jb_n-l_n}\;\sum_{s=jb_n+1}^{\min\{jb_n+l_n,n\}}\Bigg)\mathrm{E}X_{nt}X_{ns}$$

which omits the terms of B_{2n} with $|t-s| \le l_n$ so that $B_{2n}^{*} = O(l_n^{-\delta})$ by a further application of **17.18**. Also define

$$B_{1n}^{*} = \sum_{j=1}^{r_n}\Bigg(\Bigg(\sum_{t=\max\{1,(j-1)b_n+1-l_n\}}^{(j-1)b_n+l_n} \mathrm{E}X_{nt}\Bigg)^2 + \Bigg(\sum_{t=(j-1)b_n+1}^{(j-1)b_n+2l_n} \mathrm{E}X_{nt}\Bigg)^2\Bigg)$$
$$+ \Bigg(\sum_{t=r_nb_n+1-l_n}^{n} \mathrm{E}X_{nt}\Bigg)^2 - B_{1n}$$

which includes the terms from B_{2n} omitted from B_{2n}^{*}, and B_{1n} is subtracted to avoid double counting. $B_{1n}^{*} = o(1)$ and the inequality

$$B_{1n} + 2B_{2n} \le B_{1n}^{*} + 2B_{2n}^{*} \tag{25.52}$$

holds since the difference between the two sums is the summed elements of $2r_n + 1$ positive definite matrices. ∎

To appreciate the roles of the various components of $\mathrm{E}\big(\sum_{t=1}^{n} X_{nt}\big)^2 = 1$, it may be helpful to consider Figure 25.1. The squares represent $b_n \times b_n$ blocks of autocovariance terms. The shaded areas represent the terms $\mathrm{E}(Z_{n1}^2), \dots, \mathrm{E}(Z_{nr_n}^2)$ belonging to A_n. All the remainder represent the terms of B_n whose contribution is negligible. The diagonal blocks are bordered at top and left sides by bands of width l_n (exaggerated for clarity). The sum B_{1n} contains the elements of the $l_n \times l_n$ diagonal blocks appearing at the top-left corners of the diagonal blocks. B_{2n} is

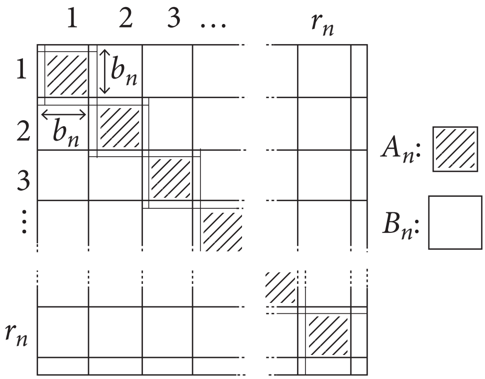

Figure 25.1

defined for the block upper triangle, mirrored in the lower triangle, with row indices t and column indices s. For $j = 1, \ldots, r_n$ these sums contain the border elements, the first l_n rows of diagonal block j and the first l_n columns of the adjacent block to the right, having column index $j + 1$. B_{3n} contains all the other terms in the block upper triangle, indexed similarly. The dependence assumptions ensure that l_n is both small enough to render these components negligible and large enough that autocovariances of higher order can be neglected. The object of defining B_{1n}^* and B_{2n}^* is to show those terms of B_{2n} close to the diagonal to be negligible by a different argument. The difference between the two sides of inequality (25.52) contains the elements of the $l_n \times l_n$ diagonal blocks (not shown) occupying the bottom-right corners of the $b_n \times b_n$ diagonal blocks.

26

Extensions and Complements

26.1 The CLT with Estimated Normalization

The results of the last two chapters, applied to the case $X_{nt} = X_t/s_n$ where $\mathrm{E}(X_t) = 0$ and $s_n^2 = \mathrm{E}\big(\sum_{t=1}^n X_t\big)^2$, would not be particularly useful if it were necessary to *know* the sequences $\{\sigma_t^2\}$ and $\{\sigma_{t,t-k}\}$ for $k \geq 1$ in order to apply them. Obviously, the relevant normalizing constants must be estimated in practice. Consider the independent case initially and let

$$S_n = \frac{1}{s_n}\sum_{t=1}^n X_t \tag{26.1}$$

where $s_n^2 = \sum_{t=1}^n \sigma_t^2$. Letting

$$\hat{s}_n^2 = \sum_{t=1}^n X_t^2 \tag{26.2}$$

write

$$\hat{S}_n = \frac{1}{\hat{s}_n}\sum_{t=1}^n X_t = d_n S_n \tag{26.3}$$

where $d_n = s_n/\hat{s}_n$. $\hat{S}_n$ is often called a *Studentized* statistic, noting that if $X_t \sim_{\mathrm{d}} N(0,\sigma^2)$ and the sample is i.i.d. then $\hat{S}_n \sim_{\mathrm{d}} t(n)$. More generally, if $d_n \to_{\mathrm{pr}} 1$ an appeal to Cramér's theorem **23.14**(ii) shows that $\hat{S}_n \to_{\mathrm{d}} \mathrm{N}(0,1)$ whenever $S_n \to_{\mathrm{d}} \mathrm{N}(0,1)$. When it holds this is an important result, because it means that it is possible to construct a statistic entirely from an observed sample whose limiting distribution is known to be standard normal. Suppose that $\hat{\mu}_n = n^{-1}\sum_{t=1}^n Y_t$ for a sequence Y_t with mean μ, where $\hat{\mu}_n \to_{\mathrm{pr}} \mu$ under conditions sufficient for the CLT by (for example) **20.4**. Then, $X_{nt} = (Y_t - \hat{\mu}_n)/\hat{s}_n$ defines an array about which nothing need be known apart from independence and a fixed mean, yet whose sums are standard normal in the limit. The interesting question is whether the minimal conditions sufficient for the CLT are also sufficient for the convergence in probability of $\hat{s}_n^2$.

If the sequence is stationary as well as independent, existence of the variance σ^2 is sufficient for both the CLT (**24.3**) and for $n^{-1}\sum_{t=1}^n X_t^2 \to_{\mathrm{pr}} \sigma^2$ (applying

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0026

24.5 to X_t^2). In the heterogeneous case, there is no weak law of large numbers for $\{X_t^2\}$ based solely on the Lindeberg condition. However, the various sufficient conditions for the Lindeberg condition given in Chapter 24, based on uniform integrability, *are* sufficient for a WLLN. Without loss of generality, take the case of possibly trending variances.

26.1 Theorem If $\{X_t\}$ is a zero-mean independent sequence satisfying the conditions of **24.17**, then

$$\frac{\hat{s}_n^2}{s_n^2} \xrightarrow{\text{pr}} 1. \tag{26.4}$$

Proof Consider the sequence $(X_t^2 - \sigma_t^2)/c_t^2$. By assumption this has zero mean, is independent (and hence an m.d.), and is uniformly integrable. The conditions of **20.12**, with $p = 1$, $b_t = c_t^2$ and $a_n = s_n^2$, are satisfied since, by (24.54),

$$\sum_{t=1}^{n} c_t^2 \le nM_n^2 = O(s_n^2) \tag{26.5}$$

where $M_n = \max_{1\le t\le n}\{c_t\}$. Hence

$$\mathrm{E}\left|\frac{\sum_{t=1}^n (X_t^2 - \sigma_t^2)}{s_n^2}\right| = \mathrm{E}\left|\frac{\sum_{t=1}^n X_t^2}{s_n^2} - 1\right| \to 0 \tag{26.6}$$

which is sufficient for convergence in probability. ∎

When the sequence $\{X_t\}$ is a martingale difference, supplementary conditions are needed for $\{X_t^2\}$ to obey a WLLN, but these turn out to be the same as are needed for the martingale CLTs of **25.3** and **25.4**. In fact, condition **25.3**(a) corresponds precisely to the required result.

26.2 Theorem Let $\{X_t, \mathcal{F}_t\}$ be a m.d. and let the conditions of **25.3** for $X_{nt} = X_t/s_n$, or **25.4**, be satisfied; then (26.4) holds. □

While in an informal sense the object is to estimate the variance, there is in fact no necessity for $\{s_n^2/n\}$, or any other sequence apart from $\{d_n\}$, to converge. Example **25.9** is a case in point. In those globally covariance stationary cases (see §13.1) where s_n^2/n converges to a finite positive constant, say $\bar{\sigma}^2$, the 'average variance' of the sequence, it is conventional to refer to $\hat{s}_n^2/n$ as a consistent estimator of $\bar{\sigma}^2$. More generally, the same terminology can always be applied to $\hat{s}_n^2$ with respect to s_n^2, in the sense of (26.4).

Alternative variance estimators can sometimes be defined which exploit the particular structure of the random sequence. In regression analysis the CLT is typically applied to sequences of the form $X_t = W_t U_t$ where $\{U_t, \mathcal{F}_t\}$ is assumed to be a m.d. with fixed variance σ^2 (the disturbance) and where W_t (a regressor) is $\mathcal{F}_{t-1}$-measurable. In this case, $\{X_t, \mathcal{F}_t\}$ is a m.d. with variances $\sigma_t^2 = \sigma^2 \mathrm{E}(W_t^2)$, which suggests the estimator $\tilde{s}_n^2 = (n^{-1}\sum_{t=1}^n U_t^2)\sum_{t=1}^n W_t^2$, for s_n^2. This is the usual approach in regression analysis, but of course the method is not robust to the failure of the fixed-variance assumption. By contrast, $\hat{s}_n^2 = \sum_{t=1}^n W_t^2 U_t^2$ possesses the property cited in (26.4) under the stated conditions, regardless of the distributions of the sequence. The latter type of estimator is termed heteroscedasticity-consistent.

Now consider the case of general dependence and heterogeneity. Letting $\sigma_{t,t+k} = \mathrm{E}(X_t X_{t+k})$, the object of concern is the variance of the sum

$$s_n^2 = \mathrm{E}\left(\sum_{t=1}^n X_t\right)^2 = \sum_{t=1}^n \sum_{t=1}^n \mathrm{E}(X_t X_s) = \sum_{k=1-n}^{n-1} \sum_{t \in n(k)} \sigma_{t,t+k} \tag{26.7}$$

where

$$n(k) = \{t : \max(1, 1-k) \le t \le \min(n, n-k)\}. \tag{26.8}$$

It may help to visualize the last representation of the sum in (26.7) as a grouping of the elements of an $n \times n$ covariance matrix by diagonals. The available unbiased estimator of $\mathrm{E}(X_t X_{t+k})$ where this may depend on both t and k is $X_t X_{t+k}$ itself . However, following the pattern of (26.2) plainly does not work. In the case where the process is stationary and weakly dependent with $s_n^2 = O(n)$,

$$\frac{1}{s_n^2}\left(\sum_{t=1}^n X_t\right)^2 \xrightarrow{\mathrm{d}} \chi^2(1)$$

as $n \to \infty$, not in probability to unity. The problem is that most of the n^2 autocovariances are either 0 or arbitrarily close to 0, while the estimators $X_t X_{t+k}$ have distributions with positive variance for any value of k.

Solutions to the problem generally considered introduce a system of weights into the sum, so that elements $X_t X_{t+k}$ having k so large that they are estimating $\sigma_{t,t+k} \approx 0$ are either downsized or deleted. The estimator takes the form

$$\hat{s}_n^2 = \sum_{k=1-n}^{n-1} w_{nk} \sum_{t \in n(k)} X_t X_{t+k} \tag{26.9}$$

where $w_{nk} = w(k/\gamma_n)$, for some suitable function $w(\cdot)$ called the *kernel* and γ_n some increasing function of n, called the *bandwidth*. The requirement is that w_{nk} is close to 1 for small k, but approaching 0 for k approaching n. $w(0) = 1$, $w(x) = w(-x)$, and $\gamma_n/n \to 0$ are all natural requirements and leading choices of kernel actually truncate the sum at a rate determined by γ_n by setting $w(x) = 0$ for $x > 1$. The formula in (26.9) defines a *kernel estimator* of the variance.

One obviously important property of a variance estimator is that it should be positive with probability 1. Kernel schemes do not guarantee this property by default. It is easy to see the requirement by writing (26.9) as a quadratic form. If $\boldsymbol{W}$ is the $n \times n$ symmetric Toeplitz matrix whose k^{th} diagonals contain the w_{nk}, then

$$\hat{s}_n^2 = \boldsymbol{x}'\boldsymbol{W}\boldsymbol{x} \tag{26.10}$$

where $\boldsymbol{x} = (X_1, \dots, X_n)'$ and the requirement is for $\boldsymbol{W}$ to be positive definite. One way to guarantee this is to construct $\boldsymbol{W}$ as the product $\boldsymbol{B}'\boldsymbol{B}$ where the Toeplitz structure is imposed by having $\boldsymbol{B}$ a $2n \times n$ matrix whose columns are the cyclic permutations of a vector $\boldsymbol{b}$, so that column 1 is $(b_1, \dots, b_{2n})'$, column 2 is $(b_{2n}, b_1, \dots, b_{2n-1})'$, column 3 is $(b_{2n-1}, b_{2n}, b_1, \dots, b_{2n-2})'$, and so forth. The inner product of a pair of these columns is

$$w_{nk} = \sum_{j=1}^{2n-k} b_j b_{j+k} + \sum_{j=2n-k+1}^{2n} b_j b_{j-2n+k}$$

depending only on their separation k. A well-known example is found by setting $b_j = \gamma_n^{-1/2}$ for $j = 1, \dots, \gamma_n < n$ and $b_j = 0$ thereafter, which gives rise to $w_{nk} = 1 - k/\gamma_n$ for $k < \gamma_n$ and 0 thereafter. This is the popular *Bartlett* kernel. With this exception kernels are rarely constructed explicitly in this fashion, but an approximate positive definiteness criterion is $\lambda_k > 0$ for $k = 1, \dots, n$ where $\lambda_k = \boldsymbol{x}_k^* \boldsymbol{W} \boldsymbol{x}_k$ and the $\boldsymbol{x}_k$ are the so-called *Fourier vectors*, having complex-valued elements of the form $x_{jk} = \exp\{2\pi \mathrm{i} jk/n\}$ for $j = 1, \dots, n$ and with $\boldsymbol{x}^*$ denoting the transposed complex conjugate. This criterion works because when n is large $\boldsymbol{W}$ approximates to a circulant matrix whose eigenvectors are the $\boldsymbol{x}_k$ with eigenvalues λ_k (see e.g. [86]). The test can be performed on the kernel function directly by verifying

$$\psi(\lambda) = \frac{1}{2\pi}\int_{-\infty}^{\infty} w(x)\mathrm{e}^{\mathrm{i}\lambda x}\mathrm{d}x \geq 0 \text{ for } -\infty < \lambda < \infty. \tag{26.11}$$

In addition to the Bartlett kernel which can be written as

$$w(x) = \begin{cases} 1 - |x| & |x| < 1 \\ 0 & \text{otherwise,} \end{cases}$$

the popular kernel functions examined by Andrews ([7]) that satisfy criterion (26.11) include the *Parzen* kernel defined by

$$w(x) = \begin{cases} 1 - 6x^2 + 6|x|^3 & |x| < \frac{1}{2} \\ 2(1 - |x|)^3 & \frac{1}{2} \le |x| < 1 \\ 0 & \text{otherwise} \end{cases}$$

and the *quadratic spectral* (QS) kernel,

$$w(x) = \frac{3}{u^2}\Big(\frac{\sin u}{u} - \cos u\Big),\ u = 6\pi x/5.$$

Truncation at 1 (i.e. at $k = \gamma_n$) is a feature shared by the Bartlett and Parzen cases but not by the quadratic spectral.

Consistency of a kernel estimator depends on the property

$$\frac{1}{\gamma_n}\sum_{k=1-n}^{n-1} |w_{nk}| \to \int_{-\infty}^{\infty} |w(x)|\mathrm{d}x < \infty \text{ as } n \to \infty \tag{26.12}$$

and given the finiteness of the limit in (26.12), authors have tended to specify continuity of w at 0 and at all but a finite number of other points as a sufficient condition. However, Jansson ([108]) exhibits counterexamples and gives the sufficient consistency condition

$$\int_{-\infty}^{\infty} \bar{w}(x)\mathrm{d}x < \infty \tag{26.13}$$

where $\bar{w}(x) = \sup_{y \ge x} |w(y)|$.

Since a consistent estimator of s_n^2 based on blocking arguments is implicitly constructed in the proof of the central limit theorem **25.12**, it should be feasible to devise an explicit estimator under precisely the same conditions, yet in practice this has proved difficult. A theorem that effectively matches the conditions of **25.12** is de Jong and Davidson ([53]), while Davidson and de Jong ([46]) works with comparable conditions on the process while specifying a restricted class of kernel functions. The first of these results in particular requires a lengthy and technical proof. By contrast, the following result, which is adapted with refinements from Hansen ([95]) and de Jong ([52]), is quoted not as the best available but simply because the proof is brief and relatively transparent, giving some good insight into the issues arising.

26.3 Theorem Let the following assumptions hold.

(a) $\{X_t, \mathcal{F}_t\}$ is a L_r-bounded adapted zero-mean sequence for $r \geq q \geq p > 2$, and for any $s \geq 0$, $\{E_t X_{t+s}, \mathcal{F}_t\}$ is L_q-NED of size $-a$, with scale constants $d_t \leq \|X_{t+s}\|_q$ on either an α-mixing process of size $-3apr/2(r-p)$ with $q = 3pr/(p+2r) < r$ or a ϕ-mixing process of size $-ar/(r-2)$ with $q = r$, and either

  (i) $p \geq 4$ and $a = -\frac{1}{2}$, or

  (ii) $4 > p > 2$ and $a = -1$.

(b) The kernel function w of (26.9) satisfies $w(0) = 1$, $w(x) = w(-x)$, and $|w(x)| < 1$ for all $x \in \mathbb{R}$ and condition (26.13).

(c) $\gamma_n = o(n^{1-2/\min\{p,4\}})$.

Then, $\hat{s}_n^2/s_n^2 - 1 \to_{\text{pr}} 0$. □

The chief distinguishing feature of **26.3** is the adaptation condition requiring X_t to be a causal process. This is unusual in the literature but is not a strong assumption in most econometric contexts. The martingale difference is only the most prominent case of an adapted process. There is further discussion of the extended NED assumption on page 384 in the context of Theorem **18.18**, which is the relevant result to be cited in the proof following.

The conditions of **26.3** accommodate globally nonstationary processes where the variance sequence may either diverge or degenerate. If the series is stationary, it follows by Theorem **17.16** that $s_n^2 = O(n)$ under the conditions of the theorem. If on the other hand $\sigma_t^2 \simeq t^\beta$ for any $\beta > -1$, Corollary **17.18** implies that $s_n^2 \simeq n^{1+\beta}$ under the same conditions and **26.3** holds for these cases. While finite moments strictly greater than of order 2 must exist there is no requirement for finite fourth moments, as are specified in a number of the well-known results in the literature such as Newey and West ([137]) and Gallant and White ([76]). These latter theorems also use the mixing characterization of dependence, which can make their application to linear processes problematic as detailed in §15.3. The alternative versions of **26.3** represent a trade-off between restrictions on dependence and existing moments. Either fourth moments exist as in case (i) of assumption (a), or the NED size needs to be -1 in place of the $-\frac{1}{2}$ specified in Theorem **25.12**, as in case (ii). De Jong and Davidson ([53]) avoid the need to invoke case (ii) by introducing a truncation similar to that invoked in the proof of Theorem **25.12**, so that all moments of the truncated variables exist yet under uniform integrability assumptions the remainders can be neglected. However, implementation of this trick requires a different representation of the estimator and results in a complicated argument. Readers are referred to the cited paper for details.

The constraints on the rate of growth of the bandwidth γ_n are a straightforward trade-off with the order of existing moments. If $p \geq 4$ then $\gamma_n = o(n^{1/2})$ is allowed, while with $p \geq 3$ which is possible in case (a)(ii) the restriction is $\gamma_n = o(n^{1/3})$. Andrews ([7]) showed that given the existence of fourth moments, $\gamma_n = O(n^{1/3})$ is the optimal bandwidth choice for the Bartlett kernel on the minimum mean-squared error (MSE) criterion, while other cases are optimized by $\gamma_n = O(n^{1/5})$. Both choices are compatible with consistency in case (a)(i).

Proof of 26.3

$$\left|\hat{s}_n^2/s_n^2 - 1\right| \leq A_{1n} + A_{2n}$$

where

$$A_{1n} = \frac{1}{s_n^2}\left|\sum_{k=1-n}^{n-1} w_{nk} \sum_{t\in n(k)} (X_tX_{t+k} - \sigma_{t,t+k})\right|,$$

$$A_{2n} = \frac{1}{s_n^2}\left|\sum_{k=1-n}^{n-1} (w_{nk} - 1) \sum_{t\in n(k)} \sigma_{t,t+k}\right|.$$

The task is to show that $A_{1n} \to_{\mathrm{pr}} 0$ and that $A_{2n} = o(1)$.

Theorem **18.18** and the assumptions imply that for $k \geq 0$, $X_tX_{t+k} - \sigma_{t,t+k}$ is an $L_{p/2}$-mixingale of size either $-\frac{1}{2}$ when $p \geq 4$ or -1 when $2 \leq p < 4$. In case (i) the Liapunov inequality and Corollary **17.11** give for each $k \geq 0$,

$$\begin{aligned}\frac{1}{s_n^2}\mathrm{E}\left|\sum_{t\in n(k)} (X_tX_{t+k} - \sigma_{t,t+k})\right| &\ll \frac{1}{s_n^2}\left|\sum_{t\in n(k)} \|X_t\|_r^2\|X_{t+k}\|_r^2\right|^{1/2} \\ &= O(n^{-1/2}).\end{aligned} \tag{26.14}$$

The order of magnitude of (26.14) is transparent in the stationary case with $\|X_t\|_r = O(1)$ and $s_n^2 \simeq n$. In the nonstationary case where $\sigma_{tt} \simeq t^\beta$ and $s_n^2 \simeq n^{1+\beta}$, considering the same relation for the normalized case with X_t replaced by $t^{-\beta/2}X_t$ shows that the terms of the sum in the majorant are of order $t^{2\beta}$.

In case (ii) the mixingale is of order $p/2 < 2$ and in this case the Liapunov inequality and Theorem **17.12** give

$$\frac{1}{s_n^2}\mathrm{E}\left|\sum_{t\in n(k)}(X_tX_{t+k}-\sigma_{t,t+k})\right| \ll \frac{1}{s_n^2}\left(\sum_{t\in n(k)}\|X_t\|_r^{p/2}\|X_{t+k}\|_r^{p/2}\right)^{2/p}$$
$$= O\big(n^{2/p-1}\big). \tag{26.15}$$

The order of magnitude in (26.15) is shown by similar reasoning to that of (26.14). Applying the limit in (26.12) which holds under assumption (b), in case (i),

$$\mathrm{E}(A_{1n}) = O(\gamma_n n^{-1/2})$$

and in case (ii),

$$\mathrm{E}(A_{1n}) = O(\gamma_n n^{2/p-1}).$$

The assumption on γ_n implies $\mathrm{E}(A_{1n}) = o(1)$ in each case, which is sufficient for convergence in probability.

Turning next to A_{2n}, consider Theorem **17.18** for the case $l = 0$ and $c_{nt} = \sigma_{tt}^{1/2}/s_n$. By implication there exists a summable sequence of non-negative weights $\{\mu(k), k \geq 1\}$, not depending on n, where $s_n^{-2}|\sum_{t\in n(k)}\sigma_{t,t+k}| \leq \mu(k)$ for each k and n so that

$$A_{2n} \leq 2\sum_{k=1-n}^{n-1}|w_{nk}-1|\left|\frac{1}{s_n^2}\sum_{t\in n(k)}\sigma_{t,t+k}\right|$$
$$\leq \sum_{k=1-n}^{n-1}|w_{nk}-1|\mu(k). \tag{26.16}$$

The $\mu(k)$ define a finite measure that is absolutely continuous with respect to counting measure on the positive integers and the majorant of (26.16) is the integral of the function $|w(k/\gamma_n) - 1|$ with respect to this measure. According to Theorem **17.18**, μ can be defined so that all but a finite number of the weights are arbitrarily close to zero. Since $w(k/\gamma_n) \to 1$ as $n \to \infty$ for all finite k and in view of assumption (b) it follows by the dominated convergence theorem that $A_{2n} = o(1)$ as $n \to \infty$. This completes the proof. ∎

A further step in the analysis for econometric applications is to let X_t depend on unknown parameters of which $n^{1/2}$-consistent estimators are available. This extension falls outside the scope of the present chapter and the reader is referred to the cited literature for further details of this aspect of the problem.

26.2 The CLT for Linear Processes

Suppose that the sequence $\{U_t, 1 \leq t \leq n\}$ satisfies one of the theorems of Chapters 24 and 25. What, then, can be said about the linear causal process

$$X_t = \theta(L)U_t \tag{26.17}$$

where $\theta(L) = \sum_{j=0}^{\infty} \theta_j L^j$, the lag polynomial defined in (13.13)? Assuming $\theta(1) = \sum_{j=0}^{\infty} \theta_j < \infty$, a fundamental identity for lag polynomials is

$$\theta(L) = \theta(1) + \theta^*(L)(1 - L) \tag{26.18}$$

where $\theta^*(L)$ is the polynomial with lag weights $\theta_j^* = -\sum_{k=j+1}^{\infty} \theta_k$ and $1 - L$ is the difference operator, also denoted by Δ. (26.18) is easily verified by noting that $\theta_j = \theta_j^* - \theta_{j-1}^*$ for $j \geq 1$ and also that $\theta_0 = \theta(1) + \theta_0^*$. The corresponding representation of (26.17) is

$$X_t = \theta(1)U_t + \theta^*(L)\Delta U_t. \tag{26.19}$$

For historical reasons, (26.19) is rather inaccurately known in the econometrics literature as the *Beveridge–Nelson decomposition*, after Beveridge and Nelson ([18]). It is easily verified that $\theta^*(1) = -\sum_{j=1}^{\infty} j\theta_j$. The condition that this latter sum is finite is known as '1-summability', where 'p-summability' is the case of weights j^p for $p \geq 1$.

The implications of the B–N decomposition were pointed out by Phillips and Solo ([146]). Consider the partial-sum process with increments defined by (26.19),

$$\begin{aligned}\sum_{t=1}^{n} X_t &= \theta(1)\sum_{t=1}^{n} U_t + \theta^*(L)\sum_{t=1}^{n} \Delta U_t \\ &= \theta(1)\sum_{t=1}^{n} U_t + \theta^*(L)(U_n - U_0).\end{aligned}$$

Thanks to the telescoping property of the second summation, the sum is decomposed into terms of different orders of magnitude under the 1-summability condition. That is to say, $\theta^*(L)(U_n - U_0) = O_p(1)$ and hence

$$\frac{1}{\sqrt{n}}\sum_{t=1}^{n} X_t = \theta(1)\frac{1}{\sqrt{n}}\sum_{t=1}^{n} U_t + O_p(n^{-1/2}).$$

It follows immediately that if $n^{-1/2}\sum_{t=1}^{n} U_t \to_{\mathrm{d}} \mathrm{N}(0,\sigma^2)$ then $n^{-1/2}\sum_{t=1}^{n} X_t \to_{\mathrm{d}} \mathrm{N}(0,\theta(1)^2\sigma^2)$. To take the best-known case, ARMA processes with autoregressive roots in the stable region are always 1-summable. By comparison with the nonparametric dependence conditions reviewed in Chapter 25, the Phillips–Solo result offers a remarkably straightforward route to establishing central limit theorems for dependent processes.

26.3 The CLT with Random Norming

As remarked in §23.1, convergence in distribution is not an attribute of a sequence of random variables but of a sequence of probability measures, notwithstanding that a sequence of random variables whose distributions approach a specified weak limit can always be defined. However, a related convergence concept does apply explicitly to random sequences. Let $\{Y_n, n \geq 1\}$ be a sequence of random variables defined on a probability space $(\Omega, \mathcal{F}, \mathcal{P})$. If for every $A \in \mathcal{F}$ the sequence of probabilities $P(\{Y_n \leq x\} \cap A)$ converges to a well-defined limit $Q_x(A) = P(\{Y \leq x\} \cap A)$ for all continuity points x of the distribution of Y and $\lim_{x\to\infty} Q_x(A) = P(A)$, the sequence is said to converge *stably* to Y in the sense of Rényi (see [156]), taking care to note that this is a different concept from the stability of a distribution defined in §23.6. Thus, stable convergence implies something about the limit distribution of a random sequence in its relationship with other random variables. For example: the random sequence

$$Y_n = \begin{cases} X, & n \text{ even} \\ X', & n \text{ odd}, \end{cases}$$

where X and X' are an independent pair of random variables sharing the same distribution, converges in distribution trivially. However, the convergence is clearly not stable. The event A might be chosen as $\{X \leq x\}$ and then

$$P(Y_n \leq x, X \leq x) = \begin{cases} P(X \leq x), & n \text{ even} \\ P(X \leq x)^2, & n \text{ odd} \end{cases}$$

for every $n \in \mathbb{N}$.

In the case that $Q_x(A) = P(A)P(Y \leq x)$ for all $A \in \mathcal{F}$, the convergence in distribution is said to be *mixing* in the sense of Rényi (see [157]). Take further care not to confuse mixing convergence with the mixing sequences concept of page 283, although there is clearly a connection between the two. Mixing convergence implies that the limiting random variable is independent of every event of the probability space in which it resides—it has somehow 'escaped from' the elements of which it is composed. However, this is the feature implicitly understood to characterize the central limit theorem; that no single component of the aggregate may be influential enough to affect the limit distribution. Hence, the mixing variant of stable convergence is just the accustomed case. More interesting in the present context are cases where the convergence in distribution is stable but is not mixing.

Consider a sequence $\{X_t\}$ which instead of (26.4) has the property

$$\frac{\sum_{t=1}^{n} X_t^2}{s_n^2} \xrightarrow{\mathrm{d}} \eta^2 \tag{26.20}$$

where η^2 is a random variable. In other words, there is a failure to obey the law of large numbers. In the context of a regression model, suppose for simplicity that $X_t = W_t U_t$ where $\{U_t, \mathcal{F}_t\}_{-\infty}^{\infty}$ is i.i.d. with mean 0 and variance $\sigma^2$and $\{W_t\}_{-\infty}^{\infty}$ is a sequence of r.v.s that are totally independent of U_t at all leads and lags. The implication is that the W_t are 'strongly exogenous' (see e.g. [68]) with respect to the generation mechanism of U_t. In effect, they can be treated as measurable with respect to the remote σ-field $\mathcal{R} = \bigcap_{s=-\infty}^{\infty} \mathcal{F}_s$. This means in particular that

$$\mathrm{E}(X_t|\mathcal{F}_{t-1}) = W_t \mathrm{E}(U_t|\mathcal{F}_{t-1}) = 0 \text{ a.s.} \tag{26.21}$$

since $\mathcal{R} \subseteq \mathcal{F}_{t-1}$ for every t, hence X_t is a m.d.

To see how (26.20) might happen in this context, consider the following case.

26.4 Example Suppose $W_t = \sum_{s=1}^{t} V_s$ where $\{V_t\}$ is i.i.d. with mean 0 and variance τ^2. Then $\mathrm{E}(W_t^2) = t\tau^2$ and

$$s_n^2 = \sum_{t=1}^{n} \mathrm{E}(U_t^2)\mathrm{E}(W_t^2) = \tfrac{1}{2}n(n+1)\tau^2\sigma^2.$$

Anticipating distributional results from Part VI it can be shown (see **32.2**) that

$$\frac{1}{\tau^2 n^2} \sum_{t=1}^{n} W_t^2 \xrightarrow{\mathrm{d}} \int_0^1 B(r)^2 \mathrm{d}r \tag{26.22}$$

where B is a Brownian motion process. The limit in (26.22) is a known distribution, a drawing from which corresponds to $2\eta^2$ where η^2 is the limit in (26.20). □

The key distinction to be drawn is between the distribution of the statistic conditional on $\mathcal{R}$ and the unconditional distribution. In the former case with $\{W_t\}$ treated as a sequence of constants,

$$\mathrm{E}\left(\frac{1}{s_n^2}\sum_{t=1}^{n}X_t^2\middle|\mathcal{R}\right)=\frac{\sigma^2}{s_n^2}\sum_{t=1}^{n}W_t^2\to\eta^2\text{ a.s.}\tag{26.23}$$

The conditional limit distribution (see §10.6 for the relevant theory) has the form

$$\frac{1}{s_n}\sum_{t=1}^{n}X_t\middle|\mathcal{R}\xrightarrow{\mathrm{d}}\mathrm{N}(0,\eta^2)\text{ a.s.}\tag{26.24}$$

This result can also be expressed in the form of the limiting ch.f.

$$\lim_{n\to\infty}\mathrm{E}\left(\exp\left\{\frac{\mathrm{i}\lambda}{s_n}\sum_{t=1}^{n}X_t\right\}\middle|\mathcal{R}\right)=\mathrm{e}^{-\lambda^2\eta^2/2}\text{ a.s.}\tag{26.25}$$

This is a case of stable convergence in Renyi's sense, but not mixing convergence because the limit distribution depends on events from $\mathcal{R}\subseteq\mathcal{F}$. Under the unconditional distribution, averaging the outcomes with respect to the marginal distribution of $\{W_t\}$, this result becomes

$$\lim_{n\to\infty}\mathrm{E}\left(\exp\left\{\frac{\mathrm{i}\lambda}{s_n}\sum_{t=1}^{n}X_t\right\}\right)=\mathrm{E}\left(\mathrm{e}^{-\lambda^2\eta^2/2}\right).\tag{26.26}$$

This is a novel central limit result because $\sum_{t=1}^{n}X_t/s_n$ is not asymptotically Gaussian. The right-hand side of (26.26) is the ch.f. of a *mixed Gaussian* distribution. A way to visualize this distribution is to note that drawings can be generated by a two-step procedure: (i) draw η^2 from an appropriate distribution supported on $(0,\infty)$; (ii) draw a standard Gaussian variate and multiply it by η. If X is mixed Gaussian with respect to a marginal c.d.f. $G(\eta)$ (say) and f_η is the Gaussian density with mean 0 and variance η^2, the moments of the distribution are easily computed. As well as $\mathrm{E}(X)=\mathrm{E}(X^3)=0$,

$$\mathrm{E}(X^2)=\int_0^\infty\int_{-\infty}^{\infty}x^2f_\eta(x)\mathrm{d}x\mathrm{d}G(\eta)=\int_0^\infty\eta^2\mathrm{d}G(\eta)=\mathrm{E}(\eta^2).\tag{26.27}$$

However, the kurtosis is non-Gaussian, for

$$\mathrm{E}(X^4) = \int_0^\infty \int_{-\infty}^\infty x^4 f_\eta(x)\mathrm{d}x\mathrm{d}G(\eta) = 3\int_0^\infty \eta^4 \mathrm{d}G(\eta) = 3\mathrm{E}(\eta^4) \qquad (26.28)$$

where the right-hand side is in general different from $3\mathrm{E}(\eta^2)^2$ (see **9.14**).

What might be the effect of this phenomenon on asymptotic inference, should it occur? As the reader may have already observed, the crucial fact is that when s_n^2 is unknown the feasible test statistic is $\hat{S}_n$ as defined in (26.3). If S_n in (26.1) has the mixed Gaussian limit with conditional variance η^2 then under the conditional distribution $d_n \to 1/\eta^2$ as in (26.23) and so $\hat{S}_n \to_{\mathrm{d}} \mathrm{N}(0,1)$. This is true under the distribution conditional on $\mathcal{R}$ and this distribution is invariant with respect to the realization $\{W_t\}$, hence it is identical with the unconditional distribution.

In this case, it turns out that Studentization is the natural remedy. The paradoxical conclusion is that while normalization by the true standard deviation s_n gives the wrong distribution, the estimated standard error gives the correct one in the limit, even though (or because) $\hat{s}_n$ does not converge in probability but remains a random variable in the limit.

However, it may be remarked that strong exogeneity is often a restrictive assumption in econometric applications. In situations where there is dynamic dependence between $\{V_t\}$ and $\{U_t\}$ in Example **26.4** (say), the distributions are more complicated and the remedies for valid inference correspondingly less tractable.

26.4 The Multivariate CLT

An array $\{\boldsymbol{X}_{nt}\}$ of p-vectors of random variables is said to satisfy the CLT if the joint distribution of $\boldsymbol{S}_n = \sum_{t=1}^n \boldsymbol{X}_{nt}$ converges weakly to the multivariate Gaussian. In its multivariate version, the central limit theorem contributes a new and powerful approximation result. Given a vector of stochastic processes exhibiting arbitrary contemporaneous dependence among themselves, there exist *linear* combinations of the processes whose partial sums are asymptotically *independent* in the limit, in view of the fact that uncorrelated Gaussian variables are independent. This fact is fundamental to the theory of asymptotic inference in econometrics.

The main step in the solution to the multivariate problem sometimes goes by the name of the 'Cramér–Wold device'.

26.5 Theorem (Cramér–Wold) A vector random sequence $\{\boldsymbol{S}_n\}_1^\infty$, $\boldsymbol{S}_n \in \mathbb{R}^k$, converges in distribution to a random vector $\boldsymbol{S}$ iff $\boldsymbol{\alpha}'\boldsymbol{S}_n \to_{\mathrm{d}} \boldsymbol{\alpha}'\boldsymbol{S}$ for every fixed k-vector $\boldsymbol{\alpha} \neq \boldsymbol{0}$.

Proof For given $\boldsymbol{\alpha}$ the characteristic function of the scalar $\boldsymbol{\alpha}'\boldsymbol{S}_n$ is

$$\mathrm{E}\big(\exp\{\mathrm{i}\lambda\boldsymbol{\alpha}'\boldsymbol{S}_n\}\big) = \phi_{n\alpha}(\lambda).$$

By the Lévy continuity theorem (**23.17**), $\boldsymbol{\alpha}'\boldsymbol{S}_n \to_{\mathrm{d}} \boldsymbol{\alpha}'\boldsymbol{S}$ if and only if $\phi_{n\alpha}(\lambda) \to \phi_{\alpha}(\lambda)$ and ϕ_{α} is continuous at $\lambda = 0$. Since $\boldsymbol{\alpha}$ is arbitrary, put $\boldsymbol{t} = \lambda\boldsymbol{\alpha}$ and obtain

$$\mathrm{E}\big(\exp\{\mathrm{i}\boldsymbol{t}'\boldsymbol{S}_n\}\big) \to \mathrm{E}\big(\exp\{\mathrm{i}\boldsymbol{t}'\boldsymbol{S}\}\big) = \psi(\boldsymbol{t}) \tag{26.29}$$

(say) where by assumption the convergence is pointwise on $\mathbb{R}^k$. By (11.44), the left-hand side of (26.29) is the ch.f. of $\boldsymbol{S}_n$ and the right-hand side is the ch.f. of $\boldsymbol{S}$. The continuity of ψ at the origin is ensured by the continuity of ϕ_{α} at 0 for all $\boldsymbol{\alpha}$ and it follows that $\boldsymbol{S}_n \to_{\mathrm{d}} \boldsymbol{S}$. ∎

Now let $\{\boldsymbol{X}_t\}$ be a sequence of random vectors and let $\boldsymbol{\Omega}_n$ be the variance matrix of $\sum_{t=1}^{n} \boldsymbol{X}_t$. Being symmetric and positive semidefinite by construction, this matrix possesses the factorization

$$\boldsymbol{\Omega}_n = \boldsymbol{C}_n\boldsymbol{\Lambda}_n\boldsymbol{C}_n' = \boldsymbol{L}_n\boldsymbol{L}_n' \tag{26.30}$$

where $\boldsymbol{L}_n = \boldsymbol{C}_n\boldsymbol{\Lambda}_n^{1/2}$, $\boldsymbol{C}_n$ being the eigenvector matrix (satisfying $\boldsymbol{C}_n\boldsymbol{C}_n' = \boldsymbol{C}_n'\boldsymbol{C}_n = \boldsymbol{I}_p$) and $\boldsymbol{\Lambda}_n$ the diagonal non-negative matrix of eigenvalues. Let $\boldsymbol{X}_{nt} = \boldsymbol{L}_n^{-}\boldsymbol{X}_t$ where if $\boldsymbol{\Omega}_n$ has full rank then $\boldsymbol{L}_n^{-} = \boldsymbol{L}_n^{-1} = \boldsymbol{\Lambda}_n^{-1/2}\boldsymbol{C}_n'$ so that $\boldsymbol{L}_n^{-\prime}\boldsymbol{L}_n^{-} = \boldsymbol{\Omega}_n^{-1}$ and $\boldsymbol{L}_n^{-}\boldsymbol{\Omega}_n\boldsymbol{L}_n^{-\prime} = \boldsymbol{I}_p$. However, $\boldsymbol{\Omega}_n$ need not have full rank for every n. If it is singular with

$$\boldsymbol{\Lambda}_n = \begin{bmatrix} \boldsymbol{\Lambda}_{1n} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} \end{bmatrix}$$

let $\boldsymbol{L}_n^{-\prime} = [\boldsymbol{C}_{1n}\boldsymbol{\Lambda}_{1n}^{-1/2} \;\; \boldsymbol{0}]$ where $\boldsymbol{C}_{1n}$ is the appropriate submatrix of $\boldsymbol{C}_n$. In this case, $\boldsymbol{L}_n^{-}\boldsymbol{\Omega}_n\boldsymbol{L}_n^{-\prime}$ has either 1s or 0s on the diagonal and $\boldsymbol{\Omega}_n^{-} = \boldsymbol{L}_n^{-\prime}\boldsymbol{L}_n^{-}$ is the *generalized* or Moore–Penrose inverse of $\boldsymbol{\Omega}_n$.

However, $\boldsymbol{\Omega}_n$ can be assumed to be asymptotically of full rank in the sense that $\boldsymbol{L}_n^{-}\boldsymbol{\Omega}_n\boldsymbol{L}_n^{-\prime} \to \boldsymbol{I}_p$. If $\boldsymbol{S}_n = \sum_{t=1}^{n} \boldsymbol{X}_{nt}$ then for any p-vector $\boldsymbol{\alpha}$ with $\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$, $\mathrm{E}(\boldsymbol{\alpha}'\boldsymbol{S}_n)^2 \to 1$. If this condition fails and there exist $\boldsymbol{\alpha} \neq \boldsymbol{0}$ such that $\mathrm{E}(\boldsymbol{\alpha}'\boldsymbol{S}_n)^2 \to 0$, the asymptotic distribution of $\boldsymbol{S}_n$ is said to be singular. In this case, some elements of the limiting vector are linear combinations of the remainder. Their distribution is thereby determined and nothing is lost by dropping these variables from the analysis.

To obtain the multivariate CLT it is necessary to show that the scalar sequences $\{\boldsymbol{\alpha}'\boldsymbol{X}_{nt}\}$ satisfy the ordinary scalar CLT, for any $\boldsymbol{\alpha}$, although the assumption $\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$ sacrifices no generality. If sufficient conditions hold for $\boldsymbol{\alpha}'\boldsymbol{S}_n \to_{\mathrm{d}} \mathrm{N}(0,1)$, the Cramér–Wold theorem implies that $\boldsymbol{S}_n \to_{\mathrm{d}} \boldsymbol{S}$. For any unit-length $\boldsymbol{\alpha}$, the ch.f. of $\boldsymbol{\alpha}'\boldsymbol{S}$ is

$$\phi(\lambda) = \mathrm{E}\big(\exp\{i\lambda\boldsymbol{\alpha}'\boldsymbol{S}\}\big) = \mathrm{e}^{-\lambda^2/2}. \tag{26.31}$$

Letting $\boldsymbol{t} = \lambda\boldsymbol{\alpha}$ be a vector of length λ, it follows from (11.46) that (26.31) is the ch.f. of a standard multivariate Gaussian vector. The inversion theorem yields the required result that $\boldsymbol{S} \sim_{\mathrm{d}} \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_p)$. The following theorem is therefore proved.

26.6 Theorem Let $\{\boldsymbol{X}_t\}$ be a stochastic sequence of p-vectors and let $\boldsymbol{\Omega}_n = \mathrm{E}\big((\sum_{t=1}^n \boldsymbol{X}_t)(\sum_{t=1}^n \boldsymbol{X}_t)'\big)$. If $\boldsymbol{L}_n^- \boldsymbol{\Omega}_n \boldsymbol{L}_n^{-\prime} \to \boldsymbol{I}_p$ and $\sum_{t=1}^n \boldsymbol{\alpha}'\boldsymbol{L}_n^-\boldsymbol{X}_t \to_{\mathrm{d}} \mathrm{N}(0,1)$ for every $\boldsymbol{\alpha}$ satisfying $\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$, then

$$\sum_{t=1}^n \boldsymbol{L}_n^- \boldsymbol{X}_t \xrightarrow{\mathrm{d}} \mathrm{N}(\boldsymbol{0}, \boldsymbol{I}_p). \qquad \square \tag{26.32}$$

In this result the elements of $\boldsymbol{\Omega}_n$ need not have the same orders of magnitude in n. The variances can be tending to infinity for some elements of $\boldsymbol{X}_t$ and to zero for others, within the bounds set by the Lindeberg condition. However, in the case when all of the elements of $\boldsymbol{\Omega}_n$ have the same order of magnitude, say n^δ for some $\delta > 0$ such that $n^{-\delta}\boldsymbol{\Omega}_n \to \boldsymbol{\Omega}$ where $\boldsymbol{\Omega}$ is a finite constant matrix, it is easy to manipulate (26.32) into the form

$$n^{-\delta/2}\sum_{t=1}^n \boldsymbol{X}_t \xrightarrow{\mathrm{d}} \mathrm{N}(\boldsymbol{0}, \boldsymbol{\Omega}). \tag{26.33}$$

Techniques for estimating the normalization factors generalize naturally from the scalar case discussed in §26.1, just like the CLT itself. Consider the martingale difference case in which $\boldsymbol{\Omega}_n = \sum_{t=1}^n \mathrm{E}(\boldsymbol{X}_t\boldsymbol{X}_t')$ and assume this matrix has rank p asymptotically in the sense defined above. Under the assumptions of **26.2**,

$$\frac{\sum_{t=1}^n (\boldsymbol{\alpha}'\boldsymbol{X}_t)^2}{\boldsymbol{\alpha}'\boldsymbol{\Omega}_n\boldsymbol{\alpha}} = \frac{\boldsymbol{\alpha}'(\sum_{t=1}^n \boldsymbol{X}_t\boldsymbol{X}_t')\boldsymbol{\alpha}}{\boldsymbol{\alpha}'\boldsymbol{\Omega}_n\boldsymbol{\alpha}} \xrightarrow{\mathrm{pr}} 1 \tag{26.34}$$

where the ratio is always well defined on taking n large enough, by assumption. This suggests that the positive semidefinite matrix $\hat{\boldsymbol{\Omega}}_n = \sum_{t=1}^n \boldsymbol{X}_t\boldsymbol{X}_t'$ is the natural estimator for $\boldsymbol{\Omega}_n$.

To be more precise: (26.34) says that $P(|\boldsymbol{\alpha}'\hat{\boldsymbol{\Omega}}_n\boldsymbol{\alpha}/\boldsymbol{\alpha}'\boldsymbol{\Omega}_n\boldsymbol{\alpha} - 1| > \varepsilon)$ can be made as small as desired for $\varepsilon > 0$ and *arbitrary* $\boldsymbol{\alpha} \neq \boldsymbol{0}$ by taking n large enough, since the normalization to unit length cancels in the ratio. This is true in the particular case

$\boldsymbol{\alpha}_n^* = \boldsymbol{L}_n^{-\prime}\boldsymbol{\alpha}$ and since $\boldsymbol{\alpha}_n^{*\prime}\boldsymbol{\Omega}_n\boldsymbol{\alpha}_n^* = 1, \boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$ implies $\boldsymbol{\alpha}'\boldsymbol{L}_n^-\hat{\boldsymbol{\Omega}}_n\boldsymbol{L}_n^{-\prime}\boldsymbol{\alpha} \to_{\mathrm{pr}} 1$. If a matrix $\boldsymbol{B}$ $(p \times p)$ is nonsingular and $g(\boldsymbol{\alpha}) = \boldsymbol{\alpha}'\mathbf{B}\boldsymbol{\alpha}/\boldsymbol{\alpha}'\boldsymbol{\alpha} = 1$ for every $\boldsymbol{\alpha} \neq \mathbf{0}$, the gradient is $\boldsymbol{g}'(\boldsymbol{\alpha}) = 2\mathbf{B}\boldsymbol{\alpha}/\boldsymbol{\alpha}'\boldsymbol{\alpha} - 2\boldsymbol{\alpha}/\boldsymbol{\alpha}'\boldsymbol{\alpha}$ and $\boldsymbol{g}'(\boldsymbol{\alpha}) = \mathbf{0}$ has unique solution $\boldsymbol{B} = \boldsymbol{I}_p$. Therefore, $\boldsymbol{L}_n^-\hat{\boldsymbol{\Omega}}_n\boldsymbol{L}_n^{-\prime} \to_{\mathrm{pr}} \boldsymbol{I}_p$. If $\hat{\boldsymbol{L}}_n^-$ is the factorization of $\hat{\boldsymbol{\Omega}}_n$, since $\boldsymbol{L}_n^-$ is asymptotically of rank p it follows by **19.9**(ii) that $\hat{\boldsymbol{L}}_n^-\boldsymbol{L}_n^- \to_{\mathrm{pr}} \boldsymbol{I}_p$ giving the desired conclusion for comparison with (26.32):

$$\hat{\boldsymbol{L}}_n^- \sum_{t=1}^{n} \boldsymbol{X}_t \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{0}, \boldsymbol{I}_p). \tag{26.35}$$

The extension to general dependence is a matter of estimating $\boldsymbol{\Omega}_n$ by a generalization of the consistent methods discussed in §26.1. Thus, let

$$\hat{\boldsymbol{\Omega}}_n = \sum_{k=1-n}^{n-1} w_{nk} \sum_{t \in n(k)} \boldsymbol{X}_t\boldsymbol{X}_{t+k}' \tag{26.36}$$

where w_{nk} represents the kernel weights as in (26.9) and $n(k)$ is defined in (26.8). This matrix is assuredly positive definite with the right choice of kernel, noting that

$$\boldsymbol{\alpha}'\hat{\boldsymbol{\Sigma}}_n\boldsymbol{\alpha} = \mathbf{x}'\boldsymbol{W}\boldsymbol{x}$$

for arbitrary $\boldsymbol{\alpha}$, where $\boldsymbol{W}$ $(n \times n)$ is the Toeplitz matrix with elements w_{nk} on the k^{th} diagonals as in (26.10) and $\boldsymbol{x} = (\boldsymbol{X}_1'\boldsymbol{\alpha}, \dots, \boldsymbol{X}_n'\boldsymbol{\alpha})'$.

This is an appropriate place to mention that the most important application of the Cramér theorem (**23.14**) is to multivariate convergence. If $\boldsymbol{T}_n \to_{\mathrm{d}} \boldsymbol{T}$ and $\boldsymbol{A}_n$ is a conformable stochastic matrix satisfying $\boldsymbol{A}_n \to_{\mathrm{pr}} \boldsymbol{A}$ where $\boldsymbol{A}$ is a nonstochastic matrix, then **23.14** implies $\boldsymbol{A}_n\boldsymbol{T}_n \to_{\mathrm{d}} \boldsymbol{A}\boldsymbol{T}$. If the distributional limit in question is multivariate Gaussian with $\boldsymbol{T}_n \to_{\mathrm{d}} \mathrm{N}(\mathbf{0}, \boldsymbol{\Omega})$ then

$$\boldsymbol{A}_n\boldsymbol{T}_n \xrightarrow{\mathrm{d}} \mathrm{N}(\mathbf{0}, \mathbf{A}\boldsymbol{\Omega}\mathbf{A}'). \tag{26.37}$$

26.5 The Delta Method

If $\bar{X}_n = n^{-1}\sum_{t=1}^n X_t$ then $\sqrt{n}(\bar{X}_n - \mu) \to_{\mathrm{d}} \mathrm{N}(0, \sigma^2)$ whenever the sequence $\{X_t\}$ satisfies suitable regularity conditions, as detailed in Chapters 24 and 25. According to the continuous mapping theorem **23.11**, the limiting distribution of $n(\bar{X}_n - \mu)^2$ is $\sigma^2\chi^2(1)$, with a mean of σ^2 and variance of $2\sigma^4$. The general rule, given a random sequence $\{T_n\}$ and a measurable function $g(\cdot)$ that is continuous except at a set of points with measure 0, is that if $T_n \to_{\mathrm{d}} T$ then $g(T_n) \to_{\mathrm{d}} g(T)$.

However, by contrast,

$$\sqrt{n}(\bar{X}_n^2 - \mu^2) \xrightarrow{\text{d}} \text{N}(0, 4\mu^2\sigma^2). \tag{26.38}$$

To demonstrate the truth of this interesting fact, consider generally a function $g(\cdot)$ that is continuously differentiable at least in an open neighbourhood of μ and let g' denote this derivative. The mean value theorem gives the result

$$g(\bar{X}_n) - g(\mu) = g'(\mu + \lambda(\bar{X}_n - \mu))(\bar{X}_n - \mu).$$

for $0 \le \lambda \le 1$. However, $\bar{X}_n - \mu = O_p(n^{-1/2})$ by assumption and hence Theorem **19.8**(ii) implies that

$$\sqrt{n}\big(g(\bar{X}_n) - g(\mu)\big) = g'(\mu)\sqrt{n}(X_n - \mu) + o_p(1). \tag{26.39}$$

This is the basic idea of the delta method. The normal limit is preserved under continuous transformations that admit a Taylor expansion at least to first order, with a scale change in the variance. While the assumption $g'(\mu) \neq 0$ would be a natural one to require under the circumstances, it is worth noting that the result remains valid in the case of (26.38) even when $\mu = 0$. In this case, $\bar{X}_n^2 = O_p(n^{-1})$ and (26.38) returns the correct result that the distribution of $\sqrt{n}\bar{X}_n^2$ is degenerate.

The basic delta method can be generalized in several directions. Consider a vector-valued function $\boldsymbol{g}$ of a vector-valued argument $\bar{\boldsymbol{X}}_n$, and also let $\boldsymbol{g}$ might be replaced by a sequence of stochastic functions $\{\boldsymbol{g}_n, n \in \mathbb{N}\}$ provided these converge in an appropriate manner. The general result is as follows.

26.7 Theorem On a probability space $(\Omega, \mathcal{F}, P)$ let $\boldsymbol{g}_n = (g_{n1}, \dots, g_{nm})' : \mathbb{R}^k \times \Omega \mapsto \mathbb{R}^m$ be an m-vector of $\mathcal{F}/\mathcal{B}^m$-measurable continuously differentiable random functions, where $\boldsymbol{G}_n : \mathbb{R}^k \times \Omega \mapsto \mathbb{R}^{mk}$ with $\{\boldsymbol{G}_n\}_{ij} = \partial g_i / \partial x_j$ is the corresponding $m \times k$ Jacobian matrix. Let $\bar{\boldsymbol{X}}_n$ $(k \times 1)$ satisfy $\sqrt{n}(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu}) \to_{\text{d}} \text{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ and let $\boldsymbol{G}_n(\cdot) \to_{\text{pr}} \boldsymbol{G}(\cdot)$ uniformly in an open set containing $\boldsymbol{\mu}$. Then

$$\sqrt{n}\big(\boldsymbol{g}_n(\bar{\boldsymbol{X}}_n) - \boldsymbol{g}_n(\boldsymbol{\mu})\big) \to_{\text{d}} \text{N}\big(\boldsymbol{0}, \boldsymbol{G}(\boldsymbol{\mu})\boldsymbol{\Sigma}\mathbf{G}(\boldsymbol{\mu})'\big).$$

Proof For each $i = 1, \dots, m$ the mean value theorem gives the relation

$$g_{ni}(\bar{\boldsymbol{X}}_n) - g_{ni}(\boldsymbol{\mu}) = \nabla g_{ni}(\boldsymbol{\mu} + \lambda_i(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu}))'(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu})$$

where ∇g_{ni} ($k \times 1$) denotes the vector of partial derivatives of g_{ni} (row i of $\boldsymbol{G}_n$) with respect to its arguments and $0 \leq \lambda_i \leq 1$. Stacking these equations gives

$$\sqrt{n}\big(\boldsymbol{g}_n(\bar{\boldsymbol{X}}_n) - \boldsymbol{g}_n(\boldsymbol{\mu})\big) = \boldsymbol{G}_n^* \sqrt{n}(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu})$$

where $\boldsymbol{G}_n^*$ is the matrix whose i^{th} row is $\nabla g_{ni}(\boldsymbol{\mu} + \lambda_i(\bar{\boldsymbol{X}}_n - \boldsymbol{\mu}))'$. Since $\bar{\boldsymbol{X}}_n - \boldsymbol{\mu} = O(n^{-1/2})$ it follows by assumption and Theorem **22.6** that $\boldsymbol{G}_n^* \to_{\text{pr}} \boldsymbol{G}(\boldsymbol{\mu})$. The conclusion now follows by the extension of the Cramér theorem **23.14** shown in (26.37). ∎

This result is valid whatever the rank of $\boldsymbol{G}(\boldsymbol{\mu})$ provided it is understood that a singular Gaussian limit is a possibility.

Here are two well-known applications of Theorem **26.7**. The first is to the distribution of the sample variance.

26.8 Example Assume that $\{Y_t\}$ is a sequence of i.i.d. random variables having mean of zero and finite fourth moment, with $\text{E}(Y_t^j) = \mu_j$ for $j = 2, 3$, and 4. Consider the vector $\boldsymbol{t}_n = (t_{1n}, t_{2n})' = \left(n^{-1}\sum_{t=1}^n Y_t, n^{-1}\sum_{t=1}^n Y_t^2\right)'$. Then, application of **26.5** to $\boldsymbol{t}_n$ gives the result $\sqrt{n}(\boldsymbol{t}_n - \boldsymbol{\alpha}) \to_{\text{d}} \text{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$ where $\boldsymbol{\alpha} = \text{E}(\boldsymbol{t}_n) = (0, \mu_2)'$ and

$$\boldsymbol{\Sigma} = \lim_{n\to\infty} n\text{E}(\boldsymbol{t}_n - \boldsymbol{\alpha})(\boldsymbol{t}_n - \boldsymbol{\alpha})' = \begin{bmatrix} \mu_2 & \mu_3 \\ \mu_3 & \mu_4 - \mu_2^2 \end{bmatrix}.$$

The variance of Y_t is estimated consistently by $s_n^2 = t_{2n} - t_{1n}^2$ where $\text{E}(s_n^2) = \mu_2$. To apply the delta method to determine the asymptotic distribution of s_n^2 after normalization, calculate the Jacobian matrix as

$$\boldsymbol{G}_n = \partial s_n^2 / \partial \boldsymbol{t}_n' = (-2t_1, 1) \overset{\text{pr}}{\to} (0, 1).$$

Applying **26.7**, the conclusion is

$$\sqrt{n}(s_n^2 - \mu_2) \overset{\text{d}}{\to} \text{N}(0, \mu_4 - \mu_2^2). \qquad \square$$

The second example is a fundamental result in estimation theory, giving the limit distribution for the solution of sets of nonlinear equations.

26.9 Example Let $\{\boldsymbol{Y}_t(\boldsymbol{\theta})\}$ be a sequence of i.i.d. random vectors where $\boldsymbol{Y}_t : \Theta \times \Omega \mapsto \mathbb{R}^m$ with $\Theta \subseteq \mathbb{R}^m$ is an $\Omega/\mathcal{B}$-measurable function of parameters $\boldsymbol{\theta}$ ($m \times 1$). Suppose that $\boldsymbol{\theta}_0 \in \Theta$ is the unique point defined by $\text{E}(\boldsymbol{Y}_t(\boldsymbol{\theta}_0)) = \boldsymbol{0}$. Let $\boldsymbol{q}_n(\boldsymbol{\theta}) = n^{-1}\sum_{t=1}^n \boldsymbol{Y}_t(\boldsymbol{\theta})$ so that $\text{E}(\boldsymbol{q}_n(\boldsymbol{\theta}_0)) = \boldsymbol{0}$ and hence, under suitable regularity conditions, $\sqrt{n}\boldsymbol{q}_n(\boldsymbol{\theta}_0) \to_{\text{d}} \text{N}(\boldsymbol{0}, \boldsymbol{\Sigma})$. In applications, $\boldsymbol{q}_n$ is typically the score (gradient) of a log-likelihood function or comparable estimation criterion.

Let $\hat{\theta}_n$ denote the solution, assumed unique, to the equations

$$\boldsymbol{q}_n(\hat{\theta}_n) = \mathbf{0}. \tag{26.40}$$

Write the inverted relation as $\hat{\theta}_n = \boldsymbol{g}_n(\mathbf{0})$ where $\boldsymbol{g}_n = \boldsymbol{q}_n^{-1}$ and let $\boldsymbol{G}_n = \partial \boldsymbol{g}_n / \partial \boldsymbol{q}'$ $(m \times m)$. In view of (26.40) the mean value theorem delivers the relation

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -\boldsymbol{G}_n^* \sqrt{n} \boldsymbol{q}_n(\theta_0)$$

where $\boldsymbol{G}_n^*$ is the matrix whose i^{th} row is evaluated at a point $(\lambda_i \boldsymbol{q}_n(\theta_0))$ for $0 \leq \lambda_i \leq 1$. Since $\boldsymbol{q}_n(\theta_0) = O_p(n^{-1/2})$ by hypothesis, **22.6** implies that $\boldsymbol{G}_n^* \to_{\text{pr}} \boldsymbol{G}(\mathbf{0})$ and hence

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \xrightarrow{\text{d}} \mathrm{N}\big(\mathbf{0}, \boldsymbol{G}(\mathbf{0})\boldsymbol{\Sigma}\mathbf{G}(\mathbf{0})'\big). \qquad \square \tag{26.41}$$

The striking feature of this last result is that its application does not require an analytic formula for $\hat{\theta}$. The gradient $\boldsymbol{q}_n$ may be obtained by numerical differentiation of a criterion function and the solution of (26.40) calculated numerically. Likewise, $\boldsymbol{G}_n(\mathbf{0})$ may be evaluated as $\boldsymbol{Q}_n^{-1}(\hat{\theta}_n)$ where $\boldsymbol{Q}_n = \partial \boldsymbol{q}_n / \partial \theta'$, and this matrix may also be computed numerically. In other words the covariance matrix in (26.41) may be estimated consistently by $\boldsymbol{Q}_n^{-1}(\hat{\theta}_n)\hat{\boldsymbol{\Sigma}}_n \boldsymbol{Q}_n^{-1}(\hat{\theta}_n)'$ where $\hat{\boldsymbol{\Sigma}}_n = n^{-1}\sum_{t=1}^n \boldsymbol{Y}_t(\hat{\theta}_n)\boldsymbol{Y}_t(\hat{\theta}_n)'$.

26.6 Law of the Iterated Logarithm

The sum of n independent and identically distributed random variables with mean 0 and finite variance σ^2, divided by $\sqrt{n}$, converges weakly to a normal limit supported on the entire real line. The same sum divided by n converges almost surely to zero. These two facts immediately provoke the question: for what precise function of n does the transition between these modes of convergence take place? The answer to this question goes under the name of the *law of the iterated logarithm* (LIL). Originally due to Khinchine and Kolmogorov, the best-known of numerous contributions in this vein are those of Hartman and Wintner ([97]) and Strassen ([177]). Here is the basic form of the result.

26.10 Theorem If $\{X_t\}$ is i.i.d. with mean 0 and variance σ^2, then

$$\limsup_{n\to\infty} \frac{\sum_{t=1}^n X_t}{\sigma(2n \log\log n)^{1/2}} = 1 \text{ a.s.} \qquad \square \tag{26.42}$$

Replacing X_t by $-X_t$ in (26.42), it is easy to see that the same relation holds and hence that the liminf of the sequence in question is almost surely equal to -1.

Notice the extraordinary delicacy of (26.42), being an almost sure equality. It is equivalent to the condition that for any $\delta > 0$, both

$$P\left(\frac{\sum_{t=1}^{n} X_t}{\sigma(2n\log\log n)^{1/2}} \geq 1+\delta, \text{ i.o.}\right) = 0 \tag{26.43}$$

and

$$P\left(\frac{\sum_{t=1}^{n} X_t}{\sigma(2n\log\log n)^{1/2}} \geq 1-\delta, \text{ i.o.}\right) = 1. \tag{26.44}$$

In words, this says that the normalized partial sum will come arbitrarily close to 1 infinitely many times as n increases without limit, but on no more than a finite number of occasions will it exceed 1, with exceptions to this rule arising only with probability zero. A feel for the precision involved can be grasped by trying some numbers: $(2\log\log 10^{99})^{1/2} = 3.3$. A check with the tabulation of the standard Gaussian probabilities shows that 3.3 is far enough into the tail that the probability of exceeding it is arbitrarily close to zero. What the LIL reveals is that for the normalized partial sums $\sum_{t=1}^{n} X_t/\sigma\sqrt{n}$ this probability *is* zero for some n not exceeding 10^{99}, although not for yet larger n. Be careful to note how this is true even if the X_t variables have the whole real line as their support.

The classical versions of the LIL are for the i.i.d. case, but here is an extension to nonstationary sequences due to Chung ([34] th. 7.5.1).

26.11 Theorem Let $\{X_t\}$ be independent and L_3-bounded with variance sequence $\{\sigma_t^2\}$, and let $s_n^2 = \sum_{t=1}^{n} \sigma_t^2$. Then (26.42) holds if for $\varepsilon > 0$,

$$\sum_{t=1}^{n} E|X_t|^3 = O\big(s_n^3/(\log s_n)^{1+\varepsilon}\big). \qquad \square$$

Generalizations to martingale differences also exist; see [176] and [88] *inter alia* for further details.

The proof of **26.10** entails combining the operation of the central limit theorem for the normalized sum with the properties of the normal distribution, specifically the exponential decay of the tails. The following result for normally distributed sequences captures the key second part of this argument. The proof the LIL itself benefits from an application of stochastic process theory and this part of the argument is given in §28.5; see page 626.

26.12 Theorem If $S_n = \sum_{t=1}^{n} X_t$ where the X_t are i.i.d. distributed as N(0, 1), then

$$\limsup_{n\to\infty} \frac{S_n}{\sqrt{2n\log\log n}} = 1 \text{ a.s.}$$

Proof Write $u(n) = (2n\log\log n)^{1/2}$ for $n \geq 16 > \mathrm{e}^{\mathrm{e}}$. The proof falls into two parts, showing for an arbitrary $\delta > 0$ that, respectively,

$$\limsup_{n\to\infty} S_n/u(n) < 1 + \delta \text{ a.s.} \tag{26.45}$$

and

$$\limsup_{n\to\infty} S_n/u(n) > 1 - \delta \text{ a.s.} \tag{26.46}$$

For the first part, the approach is to consider for some $0 < \varepsilon < 1$ the subsequence $n_k = [(1+\varepsilon)^k] \in \mathbb{N}$, $k = 1, 2, \ldots$. Applying (8.22) yields

$$\begin{aligned} P\big(S_{n_k}/u(n_k) \geq 1 + \delta\big) &= P\big(S_{n_k}/\sqrt{n_k} \geq (1+\delta)u(n_k)/\sqrt{n_k}\big) \\ &\leq \exp\{-(1+\delta)^2 \log\log n_k\} \\ &\leq k^{-(1+\delta)^2} \log(1+\varepsilon)^{-(1+\delta)^2}. \end{aligned}$$

For $\delta > 0$ the sequence $k^{-(1+\delta)^2}\log(1+\varepsilon)^{-(1+\delta)^2}$ is summable over k. It follows by the Borel–Cantelli lemma (**19.2**(i)) that

$$P\big(S_{n_k}/u(n_k) \geq 1 + \delta \text{ i.o.}\big) = 0$$

or equivalently that

$$\limsup_{k\to\infty} S_{n_k}/u(n_k) < 1 + \delta \text{ a.s.} \tag{26.47}$$

The subsequence $\{n_k : k \geq 1, n_k \geq 16\}$ in (26.47) can be replaced by $\{n \geq 16\}$ if the maximum variation of S_n over the ranges $n_k \leq n \leq n_{k+1}$ is suitably limited. The sequence $S_n - S_{n_k}$ for $n_k < n \leq n_{k+1}$ is Gaussian with independent increments and $\mathrm{E}(S_n - S_{n_k})^2 = n - n_k$. The reflection principle **13.16** and then (8.22) again gives, for large enough k, using the fact that $(n_{k+1} - n_k)/n_k\varepsilon \to 1$ as $k \to \infty$,

$$\begin{aligned} P\Big(\max_{n_k < n \leq n_{k+1}} |S_n - S_{n_k}| \geq \sqrt{2\varepsilon}u(n_k)\Big) &= 4P\big(S_{n_{k+1}} - S_{n_k} \geq \sqrt{2\varepsilon}u(n_k)\big) \\ &\leq 4\exp\{-\varepsilon u(n_k)^2/(n_{k+1} - n_k)\} \\ &\leq 4k^{-2}\log(1+\varepsilon)^{-2}. \end{aligned} \tag{26.48}$$

The sequence of probabilities in (26.48) is summable over k and hence

$$\max_{n_k\le n\le n_{k+1}} |S_n - S_{n_k}| < \sqrt{2\varepsilon} u(n_k) \text{ a.s.} \tag{26.49}$$

for all but a finite collection of k, by the Borel–Cantelli lemma. Letting $\varepsilon \downarrow 0$ as $k \to \infty$, (26.49) together with (26.47) imply (26.45). Since the distribution of S_n is symmetric about zero, this also implies

$$\liminf_{n\to\infty} S_n/u(n) > -(1+\delta) \text{ a.s.} \tag{26.50}$$

For the second part of the proof, consider for $0 < \delta < 1$, fixed n, and $j \geq 1$ the events

$$\begin{aligned} E_j &= \{\omega : S_{n^j} - S_{n^{j-1}} > (1-\delta/2)u(n^j - n^{j-1})\} \\ &= \{\omega : (S_{n^j} - S_{n^{j-1}})/(n^j - n^{j-1})^{1/2} > (1-\delta/2)(2\log\log(n^j - n^{j-1}))^{1/2}\}. \end{aligned} \tag{26.51}$$

(The superscripts here denote powers.) If E_j occurs and n is large enough then

$$\begin{aligned} S_{n^j} &> (1-\delta/2)u(n^j - n^{j-1}) + S_{n^{j-1}} \\ &> (1-\delta/2)u(n^j - n^{j-1}) - (1+\delta)u(n^{j-1}) \end{aligned} \tag{26.52}$$

where the second inequality is by (26.50). It is easily verified that as $j \to \infty$, $u(n^j - n^{j-1})/u(n^j) \to (1-1/n)^{1/2}$ and $u(n^{j-1})/u(n^j) \to n^{-1/2}$. Choose n large enough that $(1-\delta/2)(1-1/n)^{1/2} - (1+\delta)n^{-1/2} > 1-\delta$. Then, for j large enough (26.52) implies

$$S_{n^j}/u(n^j) > 1-\delta. \tag{26.53}$$

The increments $(S_{n^j} - S_{n^{j-1}})/(n^j - n^{j-1})^{1/2}$ are standard normal r.v.s and the successive events $E_1, E_2, \ldots$ relate to non-overlapping segments of the sequence $\{S_n\}$ and hence are independent. Applying (8.21) to (26.51),

$$\frac{\sqrt{2\pi}(1-\delta/2)\sqrt{2\log\log(n^j - n^{j-1})}}{\exp\{-(1-\delta/2)^2\log\log(n^j - n^{j-1})\}} P(E_j) \to 1$$

where

$$\exp\{-(1-\delta/2)^2\log\log(n^j - n^{j-1})\} = (j\log n + \log(1-1/n))^{-(1-\delta/2)^2}.$$

The implication is that $P(E_j) = O\big(j^{-(1-\delta/2)^2}\big)$ and so the sequence of probabilities is non-summable. The Borel–Cantelli lemma **19.2**(ii) implies that $P(E_j \text{ i.o.}) = 1$. Inequality (26.53) therefore holds for infinitely many j and (26.46) follows. Noting that $\delta > 0$ is arbitrary, putting (26.45) and (26.46) together and letting $\delta \downarrow 0$ completes the proof. ■

26.7 Berry–Esséen Bounds

Suppose $\{F_n\}$ is a sequence of c.d.f.s of standardized random sums and $F_n \Rightarrow \Phi$ where Φ denotes the standard Gaussian c.d.f. The setting for this result is the integer-moment case of the Liapunov CLT; see (24.36). The Berry–Esséen theorems set limits on the largest deviation of F_n from Φ. The following result was shown independently in [17] and [69].

26.13 Theorem Let $\{X_t\}$ be a zero-mean, independent, L_3-bounded random sequence, with variances $\{\sigma_t^2\}$; let $s_n^2 = \sum_{t=1}^n \sigma_t^2$; and let F_n be the c.d.f. of $S_n = \sum_{t=1}^n X_t/s_n$. There exists a constant $C > 0$ such that, for all n,

$$\sup_x |F_n(x) - \Phi(x)| \le C\sum_{t=1}^{n} E|X_t|^3/s_n^3. \qquad \square \tag{26.54}$$

The proof of **26.13** involves a comparison of the characteristic functions and is quite lengthy and mechanical. It will not be given here, although treatments such as that of Gut ([87]) and Durrett ([64]) can be recommended for interested readers.

The measure of distance between functions F_n and Φ appearing on the left-hand side of (26.54) is the uniform metric (see §5.5). Different authors have established different values for C, although Esséen ([70]) showed that it cannot be smaller than 0.4097 and the best bound shown to date appears to be $C \le 0.4784$, due to Korolev and Shevtsova ([117]). The Berry–Esséen bounds represent the 'worst case' scenario, in other words the slowest rate of uniform convergence of the c.d.f.s to be expected over all distributions having third absolute moments. As was noted in §24.1, convergence to the Gaussian limit can be very rapid with a favourable choice of F_n. For the uniformly L_3-bounded case in which $s_n^2 = O(n)$, the bound specified by (26.54) is of $O(n^{-1/2})$.

PART VI

THE FUNCTIONAL CENTRAL LIMIT THEOREM

27
Measures on Metric Spaces

The first two sections of this chapter deal with measures on metric spaces in a general way, while §§27.3–27.6 are concerned with the special case of function spaces, which are the ones that matter for the econometric results that motivate this part of the book. While this material is indispensable to a complete understanding of the issues, readers in a hurry may wish to jump straight to §27.3 and return to study the general case as their time and interests dictate.

27.1 Separability and Measurability

In any topological space $\mathbb{S}$ for which open and closed subsets are defined, the Borel field of $\mathbb{S}$ is defined as the smallest σ-field containing the open sets (and hence also the closed sets) of $\mathbb{S}$. This chapter is concerned with the properties of measurable spaces $(\mathbb{S}, \mathcal{S})$ where $\mathbb{S}$ is a metric space endowed with a metric d and $\mathcal{S}$ will always be taken to be the Borel field of $\mathbb{S}$.

Assigning a probability measure μ to the elements of $\mathcal{S}$ defines a probability space $\big((\mathbb{S}, d), \mathcal{S}, \mu\big)$. An element $x \in \mathbb{S}$ is then referred to as a *random element*. As in the theory of random variables it is often convenient to specify an underlying probability space $(\Omega, \mathcal{F}, P)$ and let $\big((\mathbb{S}, d), \mathcal{S}, \mu\big)$ be a derived space with the property $\mu(A) = P\big(x^{-1}(A)\big)$ for each $A \in \mathcal{S}$, where

$$x : \Omega \mapsto (\mathbb{S}, d)$$

is a measurable mapping. Often $\mathbb{S}$ is written in place of $(\mathbb{S}, d)$ when the choice of metric is understood, but it is important to keep in mind that d matters in this theory because $\mathcal{S}$ is not invariant to the choice of metric unless d_1 and d_2 are equivalent metrics; the open sets of $(\mathbb{S}, d_1)$ are not the same as those of $(\mathbb{S}, d_2)$.

A property of measure spaces that it is sometimes useful to assume is *regularity* (yet another usage of an overworked word, not to be confused with regularity of sequences etc.). $(\mathbb{S}, \mathcal{S}, \mu)$ is called a regular measure space (or μ a regular measure with respect to $(\mathbb{S}, \mathcal{S})$) if for each $A \in \mathcal{S}$ and each $\varepsilon > 0$ there exists an open set O_ε and a closed set C_ε such that

$$C_\varepsilon \subseteq A \subseteq O_\varepsilon \tag{27.1}$$

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0027

and

$$\mu(O_\varepsilon - C_\varepsilon) < \varepsilon. \tag{27.2}$$

Happily, as the following theorem shows, this condition can be relied upon when $\mathbb{S}$ is a metric space.

27.1 Theorem On a metric space $((\mathbb{S}, d), \mathcal{S})$ every measure is regular.

Proof Call a set $A \in \mathcal{S}$ regular if it satisfies (27.1) and (27.2). The first step is to show that any closed set is regular. Let $A_n = \{x : d(A, x) < 1/n\}$, $n = 1, 2, 3, \ldots$ denote a family of open sets. (Think of A with a 'halo' of width $1/n$.) When A is closed, $A = \bigcap_{n=1}^{\infty} A_n$ and $A_n \downarrow A$ as $n \to \infty$. By continuity of the measure this means $\mu(A_n - A) \to 0$. For any $\varepsilon > 0$ there therefore exists N such that $\mu(A_N - A) < \varepsilon$. Choosing $O_\varepsilon = A_N$ and $C_\varepsilon = A$ shows that A is regular.

Since $\mathbb{S}$ is both open and closed, it is clearly regular. If a set A is regular, so is its complement, since O_ε^c is closed, C_ε^c is open, $O_\varepsilon^c \subseteq A^c \subseteq C_\varepsilon^c$, and $C_\varepsilon^c - O_\varepsilon^c = O_\varepsilon - C_\varepsilon$. If the class of regular sets is also closed under countable unions then every Borel set is regular, which is the required result. Let $A_1, A_2, \ldots$ be regular sets and define $A = \bigcup_{n=1}^{\infty} A_n$. Fixing $\varepsilon > 0$, let $O_{n\varepsilon}$ and $C_{n\varepsilon}$ be open and closed sets respectively, satisfying

$$C_{n\varepsilon} \subseteq A_n \subseteq O_{n\varepsilon} \tag{27.3}$$

and

$$\mu(O_{n\varepsilon} - C_{n\varepsilon}) < \varepsilon/2^{n+1}. \tag{27.4}$$

Let $O_\varepsilon = \bigcup_{n=1}^{\infty} O_{n\varepsilon}$, which is open and $A \subseteq O_\varepsilon$. Also let $C_\varepsilon = \bigcup_{n=1}^{\infty} C_{n\varepsilon}$, where the latter set is not necessarily closed, but $C_\varepsilon^k = \bigcup_{n=1}^{k} C_{n\varepsilon}$ where k is finite *is* closed and $C_\varepsilon^k \subseteq A$; and since $C_\varepsilon^k \uparrow C_\varepsilon$, continuity of the measure implies that k can be chosen large enough that $\mu(C_\varepsilon - C_\varepsilon^k) < \varepsilon/2$. For such a k,

$$\begin{aligned}\mu(O_\varepsilon - C_\varepsilon^k) &\le \mu(O_\varepsilon - C_\varepsilon) + \mu(C_\varepsilon - C_\varepsilon^k)\\ &\le \sum_{n=1}^{\infty} \mu(O_{n\varepsilon} - C_{n\varepsilon}) + \mu(C_\varepsilon - C_\varepsilon^k) < \varepsilon. \end{aligned} \tag{27.5}$$

It follows that A is regular and this completes the proof. ∎

Often the theory of random variables has a straightforward generalization to the case of random elements. Consider the properties of mappings, for example. If $(\mathbb{S}, d)$ and $(\mathbb{T}, \rho)$ are metric spaces with Borel fields $\mathcal{S}$ and $\mathcal{T}$ and $f : \mathbb{S} \mapsto \mathbb{T}$ is a function, there is a natural extension of **3.39**(i), as follows.

27.2 Theorem If f is continuous, it is Borel-measurable.

Proof Direct from **5.19** and **3.29** and the fact that $\mathcal{S}$ and $\mathcal{T}$ contain the open sets of $\mathbb{S}$ and $\mathbb{T}$ respectively. ■

Let $((\mathbb{S}, d), \mathcal{S})$ and $((\mathbb{T}, \rho), \mathcal{T})$ be two measurable spaces and let $h : \mathbb{S} \mapsto \mathbb{T}$ define a measurable mapping, such that $A \in \mathcal{T}$ implies that $h^{-1}(A) \in \mathcal{S}$; then each measure μ on $\mathbb{S}$ has the property that μh^{-1} defined by

$$\mu h^{-1}(A) = \mu\big(h^{-1}(A)\big),\ A \in \mathcal{T} \tag{27.6}$$

is a measure on $((\mathbb{T}, \rho), \mathcal{T})$. This is just an application of **3.28**, which does not use topological properties of the spaces and deals solely with the set mappings involved.

However, the theory also presents some novel difficulties. A fundamental one concerns measurability. It is not always possible to assign probabilities to the Borel sets of a metric space—not, at least, without violating the axiom of choice.

27.3 Example Consider the space $(D_{[0,1]}, d_U)$, the case of **5.27** with $a = 0$ and $b = 1$. Recall that the random elements f_θ specified by (5.43) are at a mutual distance of 1 from one another. Hence, the spheres $B(f_\theta, \frac{1}{2})$ are all disjoint and any union of them is an open set (**5.4**). This means that the Borel field $\mathcal{D}_{[0,1]}$ on $(D_{[0,1]}, d_U)$ contains all of these sets. Imagine trying to construct a probability space on $((D_{[0,1]}, d_U), \mathcal{D}_{[0,1]})$ which assigns a uniform distribution to the f_θ, such that $\mu(\{f_\theta : a < \theta \le b\}) = b - a$ for $0 \le a < b \le 1$. Superficially this appears to be a perfectly reasonable project. The problem is formally identical to that of constructing the uniform distribution on $[0, 1]$. But there is one crucial difference: here, sets of f_θ functions corresponding to *every* subset of the interval are elements of $\mathcal{D}_{[0,1]}$. There are subsets of $[0, 1]$ that are not Lebesgue-measurable unless the axiom of choice is violated; see **3.24**. Hence, there is no consistent way of constructing the probability space $((D_{[0,1]}, d_U), \mathcal{D}_{[0,1]}, \mu)$, where μ assigns the uniform measure to sets of f_θ elements. This is merely a simple case, but any other scheme for assigning probabilities to these events would founder in a similar way. □

There is no reason why probabilities should not be assigned consistently to smaller σ-fields that exclude such odd cases, and in the case of $(D_{[0,1]}, d_U)$ the so-called *projection* σ-field will serve this purpose (see §30.1 below for details). The point is that with spaces like this, avoiding contradictions must entail moving beyond the familiar intuitions of the random variable case.

The space $(D_{[0,1]}, d_U)$ is of course nonseparable and nonseparability is the source of the difficulty encountered in the last example. The characteristic of a separable metric space that matters most in the present theory is the following.

27.4 Theorem In a separable metric space there exists a countable collection $\mathcal{V}$ of open spheres such that $\sigma(\mathcal{V})$ is the Borel field.

Proof This is direct from **5.6**, $\mathcal{V}$ being any collection of spheres $S(x,r)$ where x ranges over a countable dense subset of $\mathbb{S}$ and r over the positive rationals. ∎

The possible failure of the extension of a p.m. to $(\mathbb{S}, \mathcal{S})$ is avoided when there is a countable set which functions as a determining class for the space. Measurability difficulties on $\mathbb{R}$ were avoided in Chapter 3 by sticking to the Borel sets which (recall from **1.21**) are generated from countable collections of intervals. This dictum extends to other metric spaces so long as they are separable.

Another situation where separability is a useful property is the construction of product spaces. Some aspects of measures on product spaces were discussed in §3.4 but now the theory can be extended in the light of the additional structure contributed by the product topology. Let $(\mathbb{S}, \mathcal{S})$ and $(\mathbb{T}, \mathcal{T})$ be a pair of measurable topological spaces with $\mathcal{S}$ and $\mathcal{T}$ the respective Borel fields and let $\mathcal{R}$ denote the set of open rectangles of $\mathbb{S} \times \mathbb{T}$ and $\mathcal{S} \otimes \mathcal{T} = \sigma(\mathcal{R})$.

27.5 Theorem If $\mathbb{S}$ and $\mathbb{T}$ are separable spaces, $\mathcal{S} \otimes \mathcal{T}$ is the Borel field of $\mathbb{S} \times \mathbb{T}$ with the product topology.

Proof Under the product topology, $\mathcal{R}$ is a base for the open sets (see §6.5). Since $\mathbb{S} \times \mathbb{T}$ is separable by **6.16** any open set of $\mathbb{S} \times \mathbb{T}$ can be generated as a countable union of $\mathcal{R}$-sets. It follows that any σ-field containing $\mathcal{R}$ also contains the open sets of $\mathbb{S} \times \mathbb{T}$, and in particular $\mathcal{S} \otimes \mathcal{T}$ contains the Borel field. Since the sets of $\mathcal{R}$ are open, it is also true that any σ-field containing the open sets of $\mathbb{S} \times \mathbb{T}$ also contains $\mathcal{R}$ and it follows likewise that the Borel field contains $\mathcal{S} \otimes \mathcal{T}$. ∎

If either $\mathbb{S}$ or $\mathbb{T}$ is nonseparable, the last result does not generally hold. A counterexample is easily exhibited.

27.6 Example Consider the space $(D_{[0,1]} \times D_{[0,1]}, \rho_U)$, where ρ_U is the maximum metric defined by (6.13) with d_U for each of the component metrics. Let E denote the union of the open balls $B\big((x_\theta, y_\theta), \frac{1}{2}\big)$ over $\theta \in [0,1)$, where x_θ and y_θ are functions of the form f_θ in (5.43). In this metric the sets $B\big((x_\theta, y_\theta), \frac{1}{2}\big)$ are mutually disjoint rectangles, of which E is the uncountable union; if $\mathcal{R}$ denotes the open rectangles of $(D_{[0,1]} \times D_{[0,1]}, \rho_U)$, $E \notin \sigma(\mathcal{R})$ even though E is in the Borel field of $D_{[0,1]} \times D_{[0,1]}$, being an open set. □

The importance of this last result is shown by the following case. Given a probability space $(\Omega, \mathcal{F}, P)$, let x and y be random elements of derived probability spaces $\big((\mathbb{S}, d), \mathcal{S}, \mu_x\big)$ and $\big((\mathbb{S}, d), \mathcal{S}, \mu_y\big)$. Implicitly, the pair (x, y) can always be thought of as a random element of a product space of which the μ_x and μ_y are the marginal measures. Since x and y are points in the same metric space, for given $\omega \in \Omega$ a distance $d\big(x(\omega), y(\omega)\big)$ is a well-defined non-negative real number. The question of obvious interest is whether d is also a measurable function on $(\Omega, \mathcal{F})$.

27.7 Theorem If $(\mathbb{S}, d)$ is a separable space, $d(x, y)$ is a random variable.

Proof The inverse image of a rectangle $A \times B$ under the mapping

$$(x, y) : \Omega \mapsto \mathbb{S} \times \mathbb{S}$$

lies in $\mathcal{F}$, being the intersection of the $\mathcal{F}$-sets $x^{-1}(A)$ and $y^{-1}(B)$. The mapping is therefore $\mathcal{F}/\mathcal{S} \otimes \mathcal{S}$-measurable by **3.29**. But under separability, $\mathcal{S} \otimes \mathcal{S}$ is the Borel field of $\mathbb{S} \times \mathbb{S}$ according to **27.5**. Hence $(x, y)(\omega) = \big(x(\omega), y(\omega)\big)$ is a $\mathcal{F}$/Borel-measurable random element of $\mathbb{S} \times \mathbb{S}$. If the space $\mathbb{S} \times \mathbb{S}$ is endowed with the product topology, the function

$$d : \mathbb{S} \times \mathbb{S} \mapsto \mathbb{R}^+$$

is continuous by construction and this mapping is also Borel-measurable. The composite mapping

$$d \circ (x, y) : \Omega \mapsto \mathbb{R}$$

is therefore $\mathcal{F}/\mathcal{B}$-measurable and the theorem follows. ∎

27.2 Measures and Expectations

On top of measurability problems, a further challenge is the unavailability of various analytical tools that proved useful in the study of random variables. The c.d.f. and ch.f. can no longer be called on as handy representations of a distribution. However, if $U_{\mathbb{S}}$ is the set of bounded, uniformly continuous real functions $f : \mathbb{S} \mapsto \mathbb{R}$, the expectations

$$\mathrm{E}(f) = \int_{\mathbb{S}} f \mathrm{d}\mu,\ f \in U_{\mathbb{S}} \tag{27.7}$$

are always well defined.[1] The theory makes use of this family of expectations to fingerprint a distribution uniquely, a device that works regardless of the nature of the underlying space. While there is no single all-purpose function that will do this job, like $e^{i\lambda X}$ in the case $X \in \mathbb{R}$, the expectations in (27.7) play a role in this theory analogous to that of the ch.f. in the earlier theory.

As a preliminary, here are a pair of lemmas that establish the unique representation of a measure on $(\mathbb{S}, \mathcal{S})$ in terms of expectations of real functions on $\mathbb{S}$. The first establishes the uniqueness of the representation by integrals.

27.8 Lemma If μ and ν are measures on $\big((\mathbb{S}, d), \mathcal{S}\big)$ ($\mathcal{S}$ the Borel field) and

$$\int f \mathrm{d}\mu = \int f \mathrm{d}\nu, \text{ all } f \in U_{\mathbb{S}} \tag{27.8}$$

then $\mu = \nu$.

Proof $U_{\mathbb{S}}$ contains an element for which (27.8) yields the conclusion directly. Let $B \in \mathcal{S}$ be closed and define $B_n = \{x : d(x, B) < 1/n\}$. Think of B_n as B with an open halo of width $1/n$. $B_n \downarrow B$ as $n \to \infty$, B and B_n^c are closed and mutually disjoint and $\inf_{x \in B_n^c, y \in B} d(x, y) \geq 1/n$ for each n. Let $g_{B_n^c, B} \in U_{\mathbb{S}}$ be a separating function such that $g_{B_n^c, B}(x) = 0$ for $x \in B_n^c$ and 1 for $x \in B$ (see **6.13**). Then

$$\mu(B) \leq \int g_{B_n^c, B} \mathrm{d}\mu = \int g_{B_n^c, B} \mathrm{d}\nu = \int_{B_n} g_{B_n^c, B} \mathrm{d}\nu \leq \nu(B_n) \tag{27.9}$$

where the last inequality is because $g_{B_n^c, B}(x) \leq 1$. On letting $n \to \infty$ these relations give $\mu(B) \leq \nu(B)$, but μ and ν can be interchanged so $\mu(B) = \nu(B)$. This holds for all closed sets, which form a determining class for the space, so the theorem follows. ∎

Since $U_{\mathbb{S}} \subseteq C_{\mathbb{S}}$, the set of all continuous functions on $\mathbb{S}$, this result remains true after substituting $C_{\mathbb{S}}$ for $U_{\mathbb{S}}$; the point is that $U_{\mathbb{S}}$ is the *smallest* class of general functions for which it holds, by virtue of the fact that it contains the required separating functions for each closed set.

The second result, although intuitively very plausible, is considerably deeper. Given a p.m. μ on a space $\mathbb{S}$, define $\Lambda(f) = \int f \mathrm{d}\mu$ for $f \in U_{\mathbb{S}}$. Λ is a functional on $U_{\mathbb{S}}$ with the following properties:

$$f(x) \geq 0, \text{ all } x \in \mathbb{S} \Rightarrow \Lambda(f) \geq 0 \tag{27.10}$$

[1] From now on the domain of integration is understood to be $\mathbb{S}$ unless otherwise specified.

$$f(x) = 1,\ \text{all}\ x \in \mathbb{S} \Rightarrow \Lambda(f) = 1 \qquad (27.11)$$

$$\Lambda(af_1 + bf_2) = a\Lambda(f_1) + b\Lambda(f_2),\ f_1, f_2 \in U_{\mathbb{S}},\ a, b \in \mathbb{R} \qquad (27.12)$$

where (27.11) holds since $\int d\mu = 1$ and (27.12) is by the linearity property of integrals. The following lemma states that on compact spaces the implication also runs the other way.

27.9 Lemma Let $\mathbb{S}$ be a compact metric space and let $\Lambda(f) : U_{\mathbb{S}} \to \mathbb{R}$ define a functional satisfying (27.10)–(27.12). There exists a unique p.m. μ on $(\mathbb{S}, \mathcal{S})$ satisfying $\int f d\mu = \Lambda(f)$, each $f \in U_{\mathbb{S}}$. □

In other words, functionals Λ and measures μ are uniquely paired. At a later stage this result is used to establish the existence of a measure (the limit of a sequence) by exhibiting the corresponding Λ functional. The proof of this result is lengthy and will not be given here, but the details can be found in Parthasarathy ([141]) ch. 2.5. Note that because $\mathbb{S}$ is compact, $U_{\mathbb{S}}$ and $C_{\mathbb{S}}$ coincide here; see **5.21**.

27.3 Function Spaces

The leading examples of non-Euclidean metric spaces arising in econometrics are the spaces of real-valued functions having as domain an interval of the real line. This may in some contexts be the positive half-line $[0, \infty)$, but more frequently it is the unit interval $[0, 1]$ that plays a special role in the theory of weak convergence.

The notation $R_{[0,1]}$ will be used to denote the space of all such functions on $[0, 1]$, often abbreviated to R when the context is clear. Following convention the symbols x, y, etc. are now used to denote functions and t, s, etc. to denote their arguments, instead of f, g and x, y respectively as was the usage in §5.6 and elsewhere. This reflects the fact that the objects under consideration are often to be interpreted as empirical processes in the time domain. Thus,

$$x : [0, 1] \mapsto \mathbb{R}$$

is the function that assumes the value $x(t)$ at the point $t \in [0, 1]$.

Since in what follows the element x is typically stochastic, a measurable mapping from a probability space $(\Omega, \mathcal{F}, P)$ may legitimately be written

$$x : \Omega \mapsto R$$

assigning $x(\omega)$ as the image of the element ω; but also

$$x : \Omega \times [0, 1] \mapsto \mathbb{R}$$

where $x(\omega, t)$ denotes the value of x at (ω, t). Often, $x(t)$ is written to denote the ordinate at t where dependence on ω is left implicit. The potential ambiguity is generally resolved by the context. Some authors customarily write x_t to denote the random ordinate where $x_t(\omega) = x(\omega, t)$ but this usage is avoided here as far as possible, reserving the subscript notation for sequences with countable domain.

The notion of evaluating the function at a point of the domain is formalized as a projection mapping. The *coordinate projections* are the mappings

$$\pi_t : R_{[0,1]} \to \mathbb{R}$$

where $\pi_t(x) = x(t)$. The projections define cylinder sets in R; for example, the set $\pi_t^{-1}(a)$, $a \in \mathbb{R}$, is the collection of all functions on $[0, 1]$ which pass through the point of the plane with coordinates (a, t). This sort of thing is familiar from §12.3 and the union or intersection of a collection of k such cylinders with different coordinates is a k-dimensional cylinder; what is different here is that the number of coordinates available to choose from is uncountable.

Let $\{t_1, \ldots, t_k\}$ be any finite collection of points in $[0, 1]$ and let

$$\pi_{t_1,\ldots,t_k}(x) = \bigl(\pi_{t_1}(x), \ldots, \pi_{t_k}(x)\bigr) \in \mathbb{R}^k \tag{27.13}$$

denote the k-vector of projections from these coordinates. The sets of the collection

$$\mathcal{H} = \bigl\{\pi^{-1}_{t_1,\ldots,t_k}(B) \subseteq R_{[0,1]} : B \in \mathcal{B}^k, t_1, \ldots, t_k \in [0, 1], k \in \mathbb{N}\bigr\} \tag{27.14}$$

are called the *finite-dimensional sets* of $R_{[0,1]}$. It is easy to verify that $\mathcal{H}$ is a field. The *projection σ-field* is defined as $\mathcal{P} = \sigma(\mathcal{H})$.

Figure 27.1 shows a few of the elements of a rather simple $\mathcal{H}$-set, with $k = 1$ and B an interval $[a, b]$ of $\mathbb{R}$. The set $H = \pi_{t_1}^{-1}([a, b]) \in \mathcal{H}$ consists of all those functions that pass through the hole of width $b - a$ in a barrier erected at the point t_1 of the interval. Similarly, the set of all the functions passing through holes in two such barriers, at t_1 and t_2, is the image under $\pi^{-1}_{t_1,t_2}$ of a rectangle in the plane; and so forth. If the domain of the function had been countable, the projection σ-field $\mathcal{P}$ would effectively be the same collection as $\mathcal{B}^\infty$ of **12.3**. But since the domain is uncountable, $\mathcal{P}$ is strictly smaller than the Borel field of R. The sets of example **27.3**

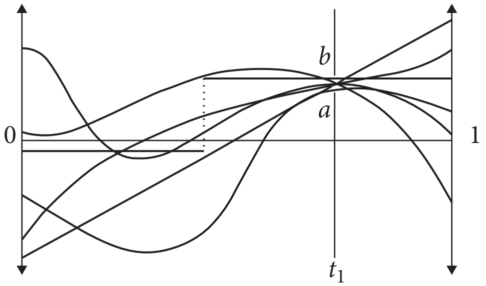

Figure 27.1

are Borel sets but are not in $\mathcal{P}$, since their elements are restricted at uncountably many points of the interval. As that example showed the Borel sets of R are not generally measurable, but $(R, \mathcal{P})$ is a measurable space, as can be shown as follows.

Define for $k = 1, 2, 3, \ldots$ the family of finite-dimensional p.m.s $\mu_{t_1,\ldots,t_k}$ on $(\mathbb{R}^k, \mathcal{B}^k)$, indexed on the collection of all the k-vectors of indices

$$\big\{(t_1, \ldots, t_k) : t_j \in [0,1], j = 1, \ldots, k\big\}.$$

This family will be required to satisfy two *consistency properties*. The first is

$$\mu_{t_1,\ldots,t_k}(E) = \mu_{t_1,\ldots,t_m}(E \times \mathbb{R}^{m-k}) \tag{27.15}$$

for $E \in \mathcal{B}^k$ and all $m > k > 0$. In other words, a k-dimensional distribution can be obtained from an m-dimensional distribution with $m > k$ by the usual operation of marginalization. This is simply the generalization to arbitrary collections of coordinates of condition (12.7). The second is

$$\mu_{t_1,\ldots,t_k} = \mu_{t_{p(1)},\ldots,t_{p(k)}}\phi^{-1} \tag{27.16}$$

where $p(1), \ldots, p(k)$ is a permutation of the integers $1, \ldots, k$ and $\phi : \mathbb{R}^k \mapsto \mathbb{R}^k$ denotes the (measurable) transformation which reorders the elements of a k-vector according to the inverse permutation; that is, $\phi(x_{p(1)}, \ldots, x_{p(k)}) = x_1, \ldots, x_k$. This condition basically means that reordering the vector elements would transform the measure in the expected way if the indices were $1, \ldots, k$ instead of $t_1, \ldots, t_k$.

The following extends the consistency theorem **12.4**.

27.10 Theorem For any family of finite-dimensional p.m.s $\{\mu_{t_1,\ldots,t_k}\}$ satisfying conditions (27.15) and (27.16), there exists a unique p.m. μ on $(R, \mathcal{P})$ such that $\mu_{t_1,\ldots,t_k} = \mu\pi^{-1}_{t_1,\ldots,t_k}$ for each finite collection of indices.

Proof Let T denote the set of countable sequences of real numbers from $[0,1]$; that is, $\tau \in T$ if $\tau = \{s_j \in [0,1], j \in \mathbb{N}\}$. Define the projections $\pi_\tau : R \mapsto \mathbb{R}^\infty$ by

$$\pi_\tau(x) = \big(x(s_1), x(s_2), \ldots\big). \tag{27.17}$$

For any τ, write $\nu^\tau_n = \mu_{s_1,\ldots,s_n}$ for $n = 1, 2, \ldots$. Then by **12.4**, which applies thanks to (27.15), there exist p.m.s ν^τ on $(\mathbb{R}^\infty, \mathcal{B}^\infty)$ such that $\nu^\tau_n = \nu^\tau\pi_n^{-1}$, where $\pi_n(y)$ is the projection of the first n coordinates of y, for $y \in \mathbb{R}^\infty$. Consistency requires that $\nu^\tau_n = \nu^{\tau'}_n$ if sequences τ and τ' have their first n coordinates the same. Since evidently $\mathcal{P} \subseteq \{\pi_\tau^{-1}(B) : B \in \mathcal{B}^\infty, \tau \in T\}$, define a p.m. μ on $(R, \mathcal{P})$ by setting

$$\mu\big(\pi_\tau^{-1}(B)\big) = \nu^\tau(B) \tag{27.18}$$

for each $B \in \mathcal{B}^\infty$. No extension is necessary here, since the measure is uniquely defined for each element of $\mathcal{P}$.

It remains to be shown that the family $\{\mu_{t_1,\dots,t_k}\}$ corresponds to the finite dimensional distributions of μ. For any $\mu_{t_1,\dots,t_k}$ there exists $\tau \in T$ such that $\{t_1,\dots,t_k\} \subseteq \{s_1,\dots,s_n\}$ for some n large enough. Construct a mapping $\Psi : \mathbb{R}^n \mapsto \mathbb{R}^k$ as follows: first apply to the indices $s_1,\dots,s_n$ the permutation p that sets $p(s_i) = t_i$ for $i = 1,\dots,k$ and then project from $\mathbb{R}^n$ to $\mathbb{R}^k$ by suppressing the indices $s_{p(k+1)},\dots,s_{p(n)}$. The consistency properties imply that

$$\mu_{t_1,\dots,t_k} = \mu_{s_1,\dots,s_n}\Psi^{-1} = \nu_n^\tau\Psi^{-1} = \nu^\tau(\Psi \circ \pi_n)^{-1} = \mu(\Psi \circ \pi_n \circ \pi_\tau)^{-1}. \quad (27.19)$$

Since $\Psi \circ \pi_n \circ \pi_\tau = \pi_{t_1,\dots,t_k}$ is a projection, $\mu_{t_1,\dots,t_k}$ is a finite-dimensional distribution of μ. ∎

A scheme for assigning a joint distribution to any finite collection of coordinate functions $\{x(t_1),\dots,x(t_k)\}$ with rational coordinates can be extended, according to the theorem, to define a unique measure on $(R,\mathcal{P})$. These p.m.s are called the finite-dimensional distributions of the stochastic process x. The sets generated by considering this vector of real r.v.s are elements of $\mathcal{H}$ and hence there is a corollary which exactly parallels **12.5**.

27.11 Corollary $\mathcal{H}$ is a determining class for $(R,\mathcal{P})$. □

27.4 The Space C

Visualize an element of $C_{[0,1]}$, the space of continuous real-valued functions on $[0,1]$, as a curve drawn by the pen of a seismograph or similar instrument as it traverses at a fixed rate a sheet of paper of unit width, making arbitrary movements up and down but never lifted from the paper. Since $[0,1]$ is a compact set the elements of $C_{[0,1]}$ are actually uniformly continuous.

To get an idea why distributions on $C_{[0,1]}$ might be of interest, imagine observing a realization of a stochastic sequence $\{S_j(\omega)\}_1^n$ from a probability space $(\Omega,\mathcal{F},P)$, for some finite n. A natural way to study these data is to display them graphically on a page or a computer screen; specifically, to construct a graph of S_j against the integer values of j from 1 to n on the abscissa and join the discrete points with ruled lines to produce a 'time plot', the kind of thing shown in Figure 27.2.

This operation does rather more than just drawing a picture; connecting the points defines a random continuous function, a random drawing (the word here operates in both its senses!) from the space $C[1,n]$. It is convenient and there is obviously no loss of generality if instead of plotting the points at unit intervals these

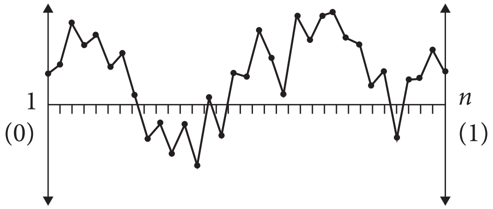

Figure 27.2

are plotted at intervals of $1/(n-1)$; in other words, let the width of the paper or computer screen be set at unity by choice of units of measurement. Also relocating the origin at 0 an element of $C_{[0,1]}$ is obtained by this means, a member of the subclass of *piecewise linear* functions, with formula

$$x(t) = (i - tm)x\big((i-1)/m\big) + (1 + tm - i)x(i/m) \tag{27.20}$$

for $t \in \big[(i-1)/m, i/m\big]$ and $i = 1, \ldots, m$, $m = n - 1$. The points $x(i/m) \in \mathbb{R}$ for $i = 0, \ldots, m$ are the $m + 1$ *vertices* of the function.

In effect, a measurable mapping has been defined between points of $\mathbb{R}^n$ and elements of $C_{[0,1]}$, derived from $(\Omega, \mathcal{F}, P)$ and indexed on n. The specific problem to be studied is the distribution of these graphs as n tends to infinity, under particular assumptions about the sequence $\{S_j\}$. When $\{S_j\}$ is a sequence of scaled partial sums of independent or asymptotically independent random variables, a useful generalization of the central limit theorem is obtained.

As in §5.6, let $C_{[0,1]}$ be metrized with the uniform metric

$$d_U(x,y) = \sup_t |x(t) - y(t)|. \tag{27.21}$$

Imagine tying two pens to a rod so that moving the rod up and down as it traverses the sheet of paper draws a band of fixed width. The uniform distance $d_U(x,y)$ between two elements of $C_{[0,1]}$ is the width of the narrowest such band that will contain both curves at all points. Henceforth, C will be written for $(C_{[0,1]}, d_U)$ when the context is clear.

C is a complete space by **5.24** and since $[0,1]$ is compact it is also separable by **5.26**(ii). An approximating function for any element of C is available in the form of a piecewise linear function as in Figure 27.2, fully determined by its values at a finite number of points of the interval (compare **5.25**). A set $\Pi_m = \{t_1, \ldots, t_m\}$ satisfying $0 = t_0 < t_1 < \ldots < t_m = 1$ is called a *partition* of $[0,1]$. This is a slight abuse of language, an abbreviated way of saying that the collection *defines* such a partition into subintervals, say $A_i = [t_{i-1}, t_i)$ for $i = 1, \ldots, m-1$ together with $A_m = [t_{m-1}, 1]$. The norm

$$||\Pi_m|| = \max_{1 \leq i \leq m}\{t_i - t_{i-1}\} \tag{27.22}$$

is called the *fineness* of the partition and a *refinement* of Π_m is any partition of which Π_m is a proper subset. Similarly, $\min_{1\leq i\leq m}\{t_i - t_{i-1}\}$ can be called the *coarseness* of Π_m.

The following approximation lemma specializes **5.25**, with the partition $\Pi_{2^n} = \{i/2^n, i = 1, \ldots, 2^n\}$ for $n \geq 1$ playing the role of the δ-net on the domain, with in this case $\delta < 2/2^n$.

27.12 Theorem Given $x \in C$ and $\varepsilon > 0$, let $y_n \in C$ be piecewise linear having $2^n + 1$ vertices with

$$\max_{1\leq i\leq 2^n}\left\{|x(2^{-n}i) - y_n(2^{-n}i)|\right\} < \tfrac{1}{2}\varepsilon. \tag{27.23}$$

Then there exists n large enough that $d_U(x, y_n) < \varepsilon$.

Proof Write $A_i = [2^{-n}(i-1), 2^{-n}i], i = 1, \ldots, 2^n$. (Inclusion of both endpoints is innocuous here). Applying (27.20), $y_n(t) = \lambda y_n(t') + (1-\lambda)y_n(t'')$ for $t \in A_i$ where $t' = 2^{-n}(i-1)$, $t'' = 2^{-n}i$, and $\lambda = i - 2^n t$. Therefore

$$\begin{aligned}|x(t) - y_n(t)| \leq \lambda|x(t) - x(t')| + (1-\lambda)|x(t) - x(t'')|\\ + \lambda|x(t') - y_n(t')| + (1-\lambda)|x(t'') - y_n(t'')|.\end{aligned} \tag{27.24}$$

For n large enough, $\sup_{s,t\in A_i}|x(s) - x(t)| < \frac{1}{2}\varepsilon$ by continuity and it follows by (27.23) that for such n,

$$d_U(x, y_n) = \max_{1\leq i\leq 2^n}\left\{\sup_{t\in A_i}|x(t) - y_n(t)|\right\} < \varepsilon. \quad ∎ \tag{27.25}$$

Note that as $n \to \infty$, $\Pi_{2^n} \to \mathbb{D}$ (the dyadic rationals). There is the following important implication.

27.13 Theorem If $x, y \in C$ and $x(t) = y(t)$ whenever $t \in \mathbb{D}$, then $x = y$.

Proof Let z_n be piecewise linear with $z_n(t) = x(t) = y(t)$ for $t \in \Pi_{2^n}$. By assumption, such a z_n exists for every $n \in \mathbb{N}$. Fixing ε and by taking n large enough that $\max\{d_U(x, z_n), d_U(y, z_n)\} < \frac{1}{2}\varepsilon$, as is possible by **27.12**, $d_U(x, y) < \varepsilon$ by the triangle inequality. Since ε is arbitrary it follows that $d_U(x, y) = 0$ and hence $x = y$ since d_U is a metric. ∎

The continuity of certain elements of R, particularly the limits of sequences of functions, is a crucial feature of several of the limit arguments to follow. An important tool is the *modulus of continuity* of a function $x \in R$, the monotone function $w_x : (0,1] \mapsto \mathbb{R}^+$ defined by

$$w_x(\delta) = \sup_{|s-t|<\delta} \left|x(s) - x(t)\right|. \tag{27.26}$$

w_x has already been encountered in the more general context of the Arzelà–Ascoli theorem **5.28**. It measures how rapidly x may change over intervals of width δ. Setting $\delta = 1$, for example, defines the range of x. But in particular, the fact that the x are uniformly continuous functions implies that, for every $x \in C$,

$$w_x(\delta) \downarrow 0 \text{ as } \delta \downarrow 0. \tag{27.27}$$

For fixed δ, think of $w_x(\delta)$ as a function on the domain C and write it as $w(x,\delta)$. Since $|w(x,\delta) - w(y,\delta)| \le 2d_U(x,y)$ (see **5.29**), $w(x,\delta)$ is continuous on C and hence a measurable function of x.

The following is the version of the Arzelà–Ascoli theorem relevant to C.

27.14 Theorem A set $A \subset C$ is relatively compact if

$$\sup_{x\in A}\left|x(0)\right| < \infty, \tag{27.28}$$

$$\lim_{\delta\to 0}\sup_{x\in A} w_x(\delta) = 0. \quad \square \tag{27.29}$$

These conditions together impose total boundedness and uniform equicontinuity on A. Consider for some $t \in [0,1]$ and $k \in \mathbb{N}$,

$$\left|x(t)\right| \le \left|x(0)\right| + \sum_{i=1}^{k}\left|x\Big(\frac{i}{k}t\Big) - x\Big(\frac{i-1}{k}t\Big)\right|. \tag{27.30}$$

Equality (27.29) implies that for large enough k, $\sup_{x\in A} w_x(1/k) < \infty$. Therefore (27.28) and (27.29) together imply that

$$\sup_t \sup_{x\in A}\left|x(t)\right| < \infty. \tag{27.31}$$

In other words, all the elements of A must be contained in a band of finite width around 0. This theorem is therefore a straightforward corollary of **5.28**.

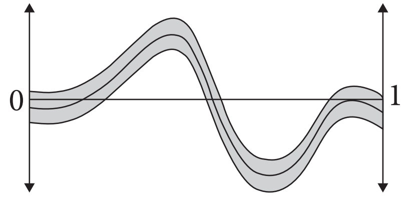

Figure 27.3

27.5 Measures on C

Theorem **27.10** is specialized when the class of functions under consideration is restricted to the members of C. The open spheres of C are sets with the form

$$S(x,r) = \{y \in C : d_U(x,y) < r\} \tag{27.32}$$

for $x \in C$. Such sets can be visualized as a bundle of continuous graphs, with radius r and the function x at the core, traversing the unit interval—for example all the functions lying within the shaded band in Figure 27.3. The Borel field of C is written $\mathcal{B}_C$. Since (C, d_U) is separable each open set has a countable covering by open spheres and $\mathcal{B}_C$ can be thought of as the σ-field generated by the open spheres of C. Each open sphere can be represented as a countable union of closed spheres,

$$S(x,r) = \bigcup_{n=1}^{\infty} \bar{S}(x, r - 1/n) \tag{27.33}$$

and hence $\mathcal{B}_C$ is also the σ-field generated from the closed spheres.

Now consider the coordinate projections on C. Happily these are continuous (see **6.15**) and hence the image of an open (closed) finite-dimensional rectangle under the inverse projection mapping is an open (closed) element of $\mathcal{P}$. Let

$$\mathcal{H}_C = \{H \cap C : H \in \mathcal{H}\} \tag{27.34}$$

with $\mathcal{H}$ defined in (27.14) and so define $\mathcal{P}_C = \sigma(\mathcal{H}_C)$.

27.15 Theorem $\mathcal{B}_C = \mathcal{P}_C$.

Proof Let

$$H_k(x,\alpha) = \left\{y \in C : \max_{1\le i\le 2^k} \left|y(2^{-k}i) - x(2^{-k}i)\right| < \alpha\right\} \in \mathcal{H}_C \tag{27.35}$$

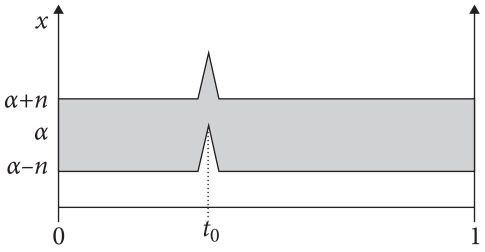

Figure 27.4

and so let

$$H(x,\alpha)=\bigcap_{k=1}^{\infty}H_k(x,\alpha)=\left\{y\in C:\sup_{t\in\mathbb{D}}\left|y(t)-x(t)\right|\le\alpha\right\}\in\mathcal{P}_C \tag{27.36}$$

where $\mathbb{D}$ denotes the dyadic rationals. The inequality in (27.35) may not remain strict in the limit, but

$$H(x,\alpha)=\bar{S}(x,\alpha) \tag{27.37}$$

by **27.13**, where $\bar{S}$ is the closure of S. Using (27.33),

$$S(x,r)=\bigcup_{n=1}^{\infty}H(x,r-1/n). \tag{27.38}$$

It follows that the open spheres of C lie in $\mathcal{P}_C$ and so $\mathcal{B}_C\subseteq\mathcal{P}_C$.

To show $\mathcal{P}_C\subseteq\mathcal{B}_C$, consider for $\alpha\in\mathbb{R}$ and $t_0\in[0,1]$ functions $x_n\in C$ defined by the restriction to $[0,1]$ of the functions on $\mathbb{R}$,

$$x_n(t)=\begin{cases}\alpha+n(n+1/n)(t+1/n-t_0), & t_0-1/n\le t<t_0\\ \alpha+n(n+1/n)(t_0+1/n-t), & t_0\le t<t_0+1/n\\ \alpha, & \text{otherwise.}\end{cases} \tag{27.39}$$

Every element y of the set $S(x_n,n)\in\mathcal{B}_C$ has the property $y(t_0)>\alpha$. (This is the shaded region in Figure 27.4.) Note that

$$G(\alpha,t_0)=\{y\in C:\pi_{t_0}(y)>\alpha\}=\bigcup_{n=1}^{\infty}S(x_n,n)\in\mathcal{B}_C. \tag{27.40}$$

Now, $G(\alpha, t_0)$ is an element of the collection $\mathcal{H}_{Ct_0}$ where for general t define

$$\mathcal{H}_{Ct} = \{\pi_t^{-1}(B), B \in \mathcal{B}\}. \tag{27.41}$$

In words, the elements of $\mathcal{H}_{Ct}$ are the sets of continuous functions x having $x(t) \in B$, for $B \in \mathcal{B}$. In view of **1.2**(ii) and (iii) and the fact that $\mathcal{B}$ can be generated by the collection of open half-lines (α, ∞), $\mathcal{H}_{Ct}$ is the σ-field generated from the sets of the form $G(\alpha, t)$ for fixed t and $\alpha \in \mathbb{R}$. Moreover, $\mathcal{H}_C$ defined in (27.34) is the σ-field generated by the collection $\{\mathcal{H}_{Ct}, t \in [0,1]\}$. Since $G(\alpha, t) \in \mathcal{B}_C$ for any α and t by (27.40) it follows that $\mathcal{H}_C \subseteq \mathcal{B}_C$. Since $\mathcal{P}_C = \sigma(\mathcal{H}_C)$, the smallest σ-field containing $\mathcal{H}_C$, it further follows that $\mathcal{P}_C \subseteq \mathcal{B}_C$. ∎

It will be noted that the limit $x_\infty(t)$ of (27.39) is not an element of C, taking the value α at all points except t_0 and $+\infty$ at t_0. Of course, $\{x_n\}$ is not a Cauchy sequence. However, the countable union of open spheres in (27.40) is an open set (the inverse projection of the open half line) and omits this point.

$\mathcal{P}_C$ is the projection σ-field on C with respect to arbitrary points of the continuum $[0,1]$, but consider the collection $\mathcal{P}_C' = \{H \cap C : H \in \mathcal{P}'\}$, where $\mathcal{P}'$ is the collection of cylinder sets of $R_{[0,1]}$ having rational coordinates as a base. In other words, the sets of $\mathcal{P}'$ contain functions whose values $x(t)$ are unrestricted except at rational t. Since elements of C which agree on the rational coordinates agree everywhere by **27.13**,

$$\mathcal{P}_C' = \mathcal{P}_C. \tag{27.42}$$

This argument is just an alternative route to the conclusion (from **6.22**) that C is homeomorphic to a subset of $\mathbb{R}^\infty$. However, it is *not* true that $\mathcal{P} = \mathcal{P}'$, because $\mathcal{P}$ is generated from the projections of every point of the continuum $[0,1]$ and arbitrary functions can be distinct in spite of agreeing on rational t.

Evidently $(C, \mathcal{B}_C)$ is a measurable space and, according to **27.11**, and **27.15**, $\mathcal{H}_C$ is a determining class for the space. In other words, the finite-dimensional distributions of a space of continuous functions uniquely determine a p.m. on the space. Every p.m. on $R_{[0,1]}$ must satisfy the consistency conditions, but the elements of C have the special property that $x(t_1)$ and $x(t_2)$ are close together whenever t_1 and t_2 are close together and this puts a further restriction on the class of finite-dimensional distributions that can generate distributions on C. Such distributions must have the property that for any $\varepsilon > 0$, $\exists\ \delta > 0$ such that

$$|t_1 - t_2| < \delta \Rightarrow \mu(\{x : |x(t_1) - x(t_2)| < \varepsilon\}) = 1. \tag{27.43}$$

The class of p.m.s in $(C, \mathcal{B}_C)$ whose finite-dimensional distributions satisfy this requirement will be denoted $\mathbb{M}_C$. Thanks to a key result—Theorem **29.7**, that is to appear in §29.3—$\mathbb{M}_C$ can be treated in later developments as a separable metric space. This fact will turn out to be most important in the sequel.

27.6 Wiener Measure

This is the original and best-known example of a probability measure on the space C, named after its discoverer Norbert Wiener ([188]). Wiener measure is the distribution that matters most in the theory of weak convergence on function spaces since it plays the role of the attractor measure analogous to that which the Gaussian distribution plays on the line. It is the natural generalization of that distribution to function spaces.

27.16 Definition Wiener measure W is the p.m. on $(C, \mathcal{B}_C)$ such that for a random function $x : [0,1] \mapsto \mathbb{R}$, $x \sim_{\mathrm{d}} W$ if

(a) $W(x(0) = 0) = 1$;

(b) $W(x(t) \leq a) = \dfrac{1}{\sqrt{2\pi t}} \displaystyle\int_{-\infty}^{a} \mathrm{e}^{-\xi^2/2t} \mathrm{d}\xi,\ 0 < t \leq 1,\ a \in \mathbb{R}$;

(c) for every partition $\{t_1, \ldots, t_m\}$ of $[0,1]$ and every $m > 0$ the increments $x(t_1) - x(0), x(t_2) - x(t_1), \ldots, x(t_m) - x(t_{m-1})$ are totally independent. □

A more formal notation is $x(\omega, t) : \Omega \times [0,1] \mapsto \mathbb{R}$, although the first argument is typically omitted. While the definition specifies $(C, \mathcal{B}_C)$ as the measurable space on which W is defined, it will transpire that a random element of $R_{[0,1]}$ satisfying conditions (a), (b), and (c) is an element of C almost surely. Such an element has distribution W.

Conditions (a) and (b) give the marginal distributions of the coordinate functions while condition (c) fixes their joint distribution. Any finite collection of process coordinates $\{x(t_i), i = 1, \ldots, k\}$ has the multivariate Gaussian distribution with $x(t_j) \sim_{\mathrm{d}} \mathrm{N}(0, t_j)$, and condition (c) implies that $\mathrm{E}(x(t_j)x(t_{j'})) = \min\{t_j, t_{j'}\}$, from which it further follows that $x(t_1) - x(t_2) \sim_{\mathrm{d}} \mathrm{N}(0, |t_1 - t_2|)$. The fact that the variance of an increment is equal to its width shows that a.s. continuity is implicit in the definition. The existence of the variances, implicit in condition (b), has the implication that the process does not jump except with probability zero, which agrees with the requirements of continuity.

However, while Definition **27.16** specifies a mathematical model of a random function, in practice it is only feasible to describe the finite-dimensional distributions and the existence of such an object (with positive probability) remains moot. In particular, reconciling continuity with the independence of adjacent increments (of arbitrary width) requires a stretch of the imagination. The consistency theorem

27.10 establishes the existence of a measure on $(C, \mathcal{B}_C)$ whose finite-dimensional distributions satisfy conditions (a)–(c) of **27.16**, so a continuous process having these properties might be constructed.

Consider a collection of random variables $\{U_1, \dots, U_n\}$, where $U_i \sim_{\mathrm{d}} \mathrm{N}(0,1)$ for $i = 1, \dots, n$ and the members are totally independent. The partial sums of such a sequence do define a Gaussian process with variance proportional to the number of terms, but it is not continuous. Instead, for given $\omega \in \Omega$, define the construction

$$Y_n(\omega, t) = n^{-1/2}\left(\sum_{i=1}^{[nt]} U_i(\omega) + (nt - [nt])U_{[nt]+1}(\omega)\right) \tag{27.44}$$

where $[x]$ denotes the largest integer below x. This is a piecewise linear function of the type sketched in Figure 27.2, with $Y_n(\omega, 0) = 0$, and it is an element of C, with the U_i representing the vertical distances from one vertex to the next. The random variable $Y_n(t)$ is Gaussian with mean 0 and variance

$$\begin{aligned} \mathrm{E}\big(Y_n(t)^2\big) &= n^{-1}\big([nt] + (nt - [nt])^2\big) \\ &= t + n^{-1}\big([nt] - nt + (nt - [nt])^2\big) \\ &= t + K(n,t)/n \end{aligned} \tag{27.45}$$

(say) where $|K(n,t)| \le 2$. Moreover, the Gaussian pair $Y_n(t)$ and $Y_n(t + s + n^{-1}) - Y_n(t + n^{-1})$ for $s > 0$ are independent. Extrapolating the same argument to general collections of non-overlapping increments, it becomes clear that $Y_n(t) \to_{\mathrm{d}} \mathrm{N}(0,t)$ and more generally that if $Y_n \to_{\mathrm{d}} Y$, then Y is a stochastic process whose finite-dimensional distributions match those of W. This argument does not show that the measure on $(C, \mathcal{B}_C)$ corresponding to Y actually is W, because there are attributes of the sample paths of the process which are not specified by the finite-dimensional distributions. According to the continuous mapping theorem, $Y_n \to_{\mathrm{d}} W$ would imply that $h(Y_n) \to_{\mathrm{d}} h(W)$ for any a.s. continuous function h. For example, $\sup_t|Y_n(t)|$ is such a function and there are no grounds from the arguments considered above for supposing that $\sup_t|Y_n(t)| \to_{\mathrm{d}} \sup_t|W(t)|$.

However, if the sequence of measures corresponding to Y_n converges to a unique limit, this can only be W, since the finite-dimensional cylinder sets of C are a determining class for distributions on $(C, \mathcal{B}_C)$. This is what can be concluded from **27.15**, in view of **27.10**. This question is taken up in the following chapters. Proof of the existence of W emerges finally as a corollary to the main weak convergence result in §29.6—see page 664.

28
Stochastic Processes in Continuous Time

28.1 Adapted Processes

The previous chapter focused on the geometric properties of stochastic processes as families of continuous or discontinuous curves. A different emphasis is provided by constructing them as elements of a filtered probability space. Parallels with the analysis of discrete series are close and natural, see in particular the discussion in §16.1.

In continuous time, a filtration $\boldsymbol{F} = \{\mathcal{F}(t), t \in [0,\infty)\}$ in a complete probability space $(\Omega, \mathcal{F}, P)$ is an uncountable collection of sub-σ-fields of $\mathcal{F}$ with the nested property

$$\mathcal{F}(t) \subseteq \mathcal{F}(s) \text{ when } t \leq s. \tag{28.1}$$

The filtration $\boldsymbol{F}$ is said to be *right-continuous* if

$$\mathcal{F}(t) = \mathcal{F}(t+) = \bigcap_{s>t} \mathcal{F}(s). \tag{28.2}$$

The probability space equipped with filtration $\boldsymbol{F}$ may be written $(\Omega, \mathcal{F}, \boldsymbol{F}, P)$, corresponding to the discrete time case of §16.1. Again paralleling that development, a continuous-time stochastic process $X = \{X(t), t \in [0,\infty)\}$ is said to be adapted to $\boldsymbol{F}$ if $X(t)$ is an $\mathcal{F}(t)$-measurable random variable for each t. Right-continuity of the filtration does not imply continuity of X, but if X is continuous the adaptation of $X(t)$ to $\mathcal{F}(t)$ implies adaptation to $\mathcal{F}(t+)$. There is typically no loss of generality in assuming (28.2) even when considering processes that are not continuous almost surely.

In the previous chapter the functions under consideration were invariably defined on domain $[0,1]$, reflecting the universal applications of the theory to the limits of discrete partial sums. The process of time aggregation is most naturally represented as compressing increasing numbers of observations into a fixed interval, which can have width of unity with no loss of generality. In this chapter there is less emphasis on these applications, and to avoid specifying a terminal point, stochastic processes are most conveniently defined on the nonnegative real line. The bounded domain $[0,1]$ remains the key special case.

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0028

The processes arising in asymptotic analysis are usually of a fairly special type. Essentially, attention is confined to cases for which a plausible generating mechanism can be imagined, ruling out completely arbitrary functions. The following definition is rather general but introduces a number of important ideas.

28.1 Definition A *Lévy process* is a random function $X : [0,\infty) \mapsto \mathbb{R}$ with the following properties.

(a) Sample paths are càdlàg a.s.

(b) $X(0) = 0$ a.s.

(c) Increments are independent: for any $r \in \mathbb{N}$ and $0 \leq t_1 < t_2 < \cdots < t_r < \infty$ the collection $\{X(t_1),\ X(t_2) - X(t_1), \ldots, X(t_t) - X(t_{r-1})\}$ comprises totally independent random variables.

(d) Increments are stationary: the distribution of $X(t+s) - X(t)$ for $s \geq 0$ and $t \geq 0$ does not depend on t.

(e) The process is continuous in probability: $\lim_{s\to 0} P(|X(t+s) - X(t)| > \varepsilon) = 0$, all $t \geq 0$, and $\varepsilon > 0$. □

More properly, write $X(\omega,t) : \Omega \times [0,\infty) \mapsto \mathbb{R}$, although the first argument is generally taken as understood and is not shown explicitly. The special case of a Lévy process on the interval $[0,1]$ requires very natural modifications of these conditions; specifically let $t_r = 1$ in (c), $0 < s \leq 1$ and $0 \leq t \leq 1 - s$ in (d), and $0 \leq t < 1$ in (e).

Consider each of properties **28.1**(a)–(e) in turn. The word càdlàg was defined in Example **5.27** and also see Example **27.3**. These functions may exhibit discontinuities but are right-continuous, with a limit existing to the left of every point. Imagining the progress of a stochastic process through time, there must exist the possibility of a jump (discontinuous change) occurring at a point in time. However, a process with *isolated* coordinates, which would entail both a jump and another opposing jump occurring simultaneously, is a much less plausible scenario in the context of a process in the time domain. If such extreme behaviour is ruled out, only one thing remains to be decided: where should a jump be assigned its value, at take-off, or on landing? 'Càdlàg' designates the landing as the point at which to measure the function, hence the right continuity, although note that this is merely a convention and left-continuity might equally be specified. The space of càdlàg functions on the unit interval is designated $D_{[0,1]}$ and their properties are reviewed in detail in §30.1. Be careful not to confuse the right-continuity of a function with the distinct concept of right-continuity of a filtration.

Condition (b) of the definition is self-evidently a simplifying choice that can always be imposed by a suitable translation. Condition (c), on the other hand, is rather mysterious and even counter-intuitive. Real-world processes embody trends and cycles that would offer some local predictability, but here the history of past movements may reveal nothing about the direction of future movements.

This may not always be realistic but it facilitates the use of such powerful statistical methods that it is worth entertaining, perhaps induced by some clever transformation but also as a natural consequence of time aggregation.

Condition (d) is not so mysterious but is a simplifying assumption that might be a natural feature or be effected by some suitable transformation. Conditions (c) and (d) jointly imply the important property of infinite divisibility of the distribution of the process increments, as defined in §23.7. The distribution of $X(t)$ for any $t > 0$ is also the distribution of the sum of the increments $X(2^{-n}k) - X(2^{-n}(k-1))$ for $k = 1, \ldots, 2^n t$ for any $n \in \mathbb{N}$. If these increments are independent by (c) and identically distributed by (d), the role of infinitely divisible distributions in the analysis is easy to appreciate.

Concerning condition (e), note that continuity in probability merely assigns the probability zero to any particular point of the domain being a jump point of the càdlàg function. The fact that the jump points of a càdlàg process on [0,1] are at most countable and hence of Lebesgue measure 0 is shown in §30.1.

28.2 Example The Poisson process defined in §13.2 is a Lévy process. Consider the conditions. **28.1**(a) and (b) hold since $N(t)$ defined in (13.4) is right-continuous and $N(0) = 0$ by construction. The increment $N(s) - N(t)$ has the Poisson distribution with parameter $\lambda(s-t)$, not depending on $N(t)$ or on t and hence conditions **28.1**(c) and (d) also hold. Condition (e) follows by the property of càdlàg functions just noted. □

The leading example of a Lévy process is undoubtedly *Brownian motion B*, which will not be cited as a numbered example because its properties are examined in detail in the remainder of this chapter. B satisfies the five conditions listed under **28.1** but is a special case because it also has the property of being continuous almost surely. The leading characteristic of Brownian motion is Gaussianity, although interestingly enough, adding continuity and finite variance of the increments to the list of attributes is sufficient to specify Brownian motion uniquely. This remarkable fact is proved in §28.7.

Therefore, since a continuous process with finite variance must be Brownian, the leading examples of Lévy processes other than the Poisson and Brownian cases are those whose variances are infinite.

28.3 Definition A standard α-stable Lévy motion on $[0,1]$ is a Lévy process $X :$ $[0,1] \mapsto \mathbb{R}$ with increments having the property

$$X(t) - X(s) \overset{\mathrm{d}}{\sim} S_\alpha((t-s)^{1/\alpha}, \beta, 0),\ 0 \leq s < t \leq 1$$

with $\alpha < 2$ and $\beta = 0$ if $\alpha = 1$. □

See §23.7 for details of these distributions. To interpret the scale parameter, note that the distribution of $(X(t)-X(s))/(t-s)^{1/\alpha}$ does not depend on $t-s$ and in this sense is standardized. Increments of fixed width form a stationary sequence and the process is *self-similar*, meaning that for any choice of $s \in [0,1)$ and $k \in (0,1-s]$, $X^*(t) = k^{-1/\alpha}(X(s+kt)-X(s))$ has the same distribution as $X(t)$ for $t \in [0,1]$. Varying s and k can be thought of as 'zooming in' on the portion of the process from s to $s+k$. That the case with $\alpha = 1$ and $\beta \neq 0$ is exceptional is evident by consideration of formula (23.67) and the remarks following.

28.2 Diffusions and Martingales

A *Markov process* is an adapted process $\{X(t), \mathcal{F}(t)\}$ having the property

$$P\big(X(t+s) \in A|\mathcal{F}(t)\big) = P\big(X(t+s) \in A|\sigma(X(t))\big) \text{ a.s.}[P] \tag{28.3}$$

for $A \in \mathcal{B}$ and $t,s \geq 0$. This means that all the information capable of predicting the future path of a Markov process is contained in its current realized value. An adapted Lévy process is a Markov process by the independence of increments property. A *diffusion process* is a Markov process having continuous sample paths. Brownian motion is the leading example of all these classes of object. It is a Lévy process that has continuous sample paths almost surely and is therefore both a Markov process and a diffusion process. However, while the sample paths of a diffusion process must be describable in terms of a stochastic mechanism generating infinitesimal increments, these need not be independent or identically distributed, nor for that matter Gaussian. Sources such as Cox and Miller ([35]) or Karatzas and Shreve ([111]) might be consulted for further details.

The idea of a martingale extends very naturally from discrete to continuous time. The adapted pair $\{X(t), \mathcal{F}(t), t \in [0,\infty)\}$ is said to be a martingale if

$$\mathrm{E}|X(t)| < \infty, \ \ t \geq 0 \tag{28.4}$$

and for $0 \leq t \leq s$,

$$\mathrm{E}\big(X(s)|\mathcal{F}(t)\big) = X(t) \text{ a.s.}[P]. \tag{28.5}$$

It is called a semimartingale (sub- or super-) if (28.4) plus one of the inequalities

$$\mathrm{E}\big(X(s)|\mathcal{F}(t)\big)\left\{ \begin{matrix} \geq \\ \leq \end{matrix} \right\}X(t) \text{ a.s.}[P] \tag{28.6}$$

holds for $0 \leq t \leq s$.

28.4 Example The centred Poisson process $N(t) - \lambda t$ is a martingale. For $s > t$ the increment $N(s) - N(t)$ is an independent Poisson variate (see **28.2**) with mean $\lambda(s - t)$ and hence $\mathrm{E}(N(s)|\mathcal{F}(t)) = N(t) + \lambda(s - t)$ a.s. Therefore,

$$\mathrm{E}(N(s) - \lambda s|\mathcal{F}(t)) = N(t) - \lambda t \text{ a.s.} \qquad \Box$$

One way to generate a continuous-time martingale is by mapping a discrete-time martingale $\{S_j, \mathcal{F}_j\}_1^\infty$ into $[0, \infty)$. Set $S_0 = 0$ and for some $n > 0$ let $X(t) = S_j$ for $j/n \leq t < (j + 1)/n$, so that $j = [nt]$. This is a right-continuous simple function that jumps at the points $t = j/n$. Similarly let $\mathcal{F}(t) = \mathcal{F}_j$ for $j/n \leq t < (j + 1)/n$. Then $X(t)$ is $\mathcal{F}(t)$-measurable and the collection $\{\mathcal{F}(t)\}$ can be verified to be right-continuous. X is a càdlàg function following from the fact that the intervals over which $X(t)$ is constant are closed below and open above.

The following result extends the maximal inequalities of **16.20** and **16.21** to the continuous case.

28.5 Theorem Let $\{X(t), \mathcal{F}(t)\}$ be a martingale. Then

(i) $P\big(\sup_{s\in[0,t]}|X(s)| > \varepsilon\big) \leq \dfrac{\mathrm{E}|X(t)|^p}{\varepsilon^p}$, $p \geq 1$. (Kolmogorov inequality)

(ii) $\mathrm{E}\big(\sup_{s\in[0,t]}|X(s)|^p\big) \leq \Big(\dfrac{p}{p-1}\Big)^p \mathrm{E}|X(t)|^p$, $p > 1$. (Doob inequality)

Proof Define a discrete martingale by setting

$$(S_k, \mathcal{F}_k) = \big(X(2^{-n}k), \mathcal{F}(2^{-n}k)\big),\ k = 1, \ldots, [2^n t] \qquad (28.7)$$

for some $n \in \mathbb{N}$. According to the definition, S_k is $\mathcal{F}_k$-measurable and $\mathrm{E}(S_k|\mathcal{F}_{k-1}) = S_{k-1}$. Then define the function $X_{(n)} : [0, t] \mapsto \mathbb{R}$ by $X_{(n)}(s) = S_k$ for $s = 2^{-n}k \leq s < 2^{-n}(k + 1)$ and $k = 1, \ldots, [2^n t] - 1$ with $X_{(n)}(t) = S_{[2^n t]}$. Inequalities (i) and (ii) hold for $X_{(n)}$ by **16.20** and **16.21**, noting that

$$\sup_{s\in[0,t]} |X_{(n)}(s)|^p = \max_{1\leq k\leq[2^n t]} |S_k|^p. \qquad (28.8)$$

$X_{(n)}$ is a right-continuous function by construction and $X_{(n)}(s) \downarrow X(s)$ for each $s \in [0, t]$ as $n \to \infty$. ∎

Since time is a continuum the moment at which some event in the evolution of X occurs is a real random variable. $\tau(\omega)$ is called a *stopping time* of the filtration $\{\mathcal{F}(t)\}$ if

$$\{\omega : \tau(\omega) \le t\} \in \mathcal{F}(t). \tag{28.9}$$

This may be compared with the definition in §16.2. For example, the first time $\{X(t)\}$ exceeds some positive constant M in absolute value is

$$\tau(\omega) = \inf\{t : |X(\omega,t)| > M\}. \tag{28.10}$$

Also paralleling the discrete case, the *Doob–Meyer* (DM) *decomposition* of an integrable submartingale, when it exists, is the unique decomposition

$$X(t) = M(t) + A(t) \tag{28.11}$$

where M is a martingale and A a nondecreasing integrable process. The DM decomposition has been shown to exist with M uniformly integrable if the set $\{X(\tau), \tau \in \mathcal{T}\}$ is uniformly integrable, where $\mathcal{T}$ denotes the set of stopping times of $\{\mathcal{F}(t)\}$. For the details see for example Karatzas and Shreve [111]: th. 1.4.10.

The case of particular interest is the square of a martingale. A martingale X is said to be *square-integrable* if $\mathrm{E}(X(t)^2) < \infty$ for each $t \ge 0$, and for such processes the inequality

$$\mathrm{E}\big(X(s)^2|\mathcal{F}(t)\big) = X(t)^2 + \mathrm{E}\big((X(s)-X(t))^2|\mathcal{F}(t)\big) \ge X(t)^2 \tag{28.12}$$

holds a.s.$[P]$ for $s \ge t$ in view of (28.5). It follows that X^2 is a submartingale. Given a martingale $\{X(t), \mathcal{F}(t)\}$, an increasing, adapted stochastic process $\{\langle X\rangle(t), \mathcal{F}(t)\}$ whose conditionally expected variations match those of X^2 almost surely is called the *quadratic variation process* of X. By definition $\{\langle X\rangle(t), \mathcal{F}(t)\}$ satisfies the relation

$$\mathrm{E}\big(\langle X\rangle(s)|\mathcal{F}(t)\big) - \langle X\rangle(t) = \mathrm{E}\big(X(s)^2|\mathcal{F}(t)\big) - X(t)^2 \text{ a.s.}[P] \tag{28.13}$$

for $s \ge t$. See §16.4 for the corresponding analysis of the discrete case. The increments of $\langle X\rangle(t)$ are the innovations of the process $X(t)^2$, and rearranging (28.13) gives

$$\mathrm{E}\big(X(s)^2 - \langle X\rangle(s)|\mathcal{F}(t)\big) = X(t)^2 - \langle X\rangle(t) \text{ a.s.}[P], \tag{28.14}$$

showing that $\{X(t)^2 - \langle X\rangle(t), \mathcal{F}(t)\}$ is a martingale. This process accordingly defines the DM decomposition of X^2.

28.3 Brownian Motion

The term Brownian motion refers historically to the microscopic random movements of small objects such as pollen grains suspended in water, first studied by the botanist Robert Brown in 1827. That the phenomenon was due to the thermal agitation of water molecules and in particular the Gaussian distribution of the movements was shown theoretically in the famous 1905 paper of Albert Einstein ([67]). The mathematical model describing the random paths followed by the Brownian particles in idealized form is Wiener measure. With variations restricted to one dimension, Wiener measure is defined in **27.16** as a distribution on the elements of the space of continuous functions $C_{[0,1]}$. The general case of a process having the non-negative half-line as domain can be specified as follows.

28.6 Definition Brownian motion $B : [0,\infty) \mapsto \mathbb{R}$ is the real-valued random process satisfying the following properties:

(a) $P(B(0)=0)=1$;

(b) $P(B(t) \le a) = \dfrac{1}{\sqrt{2\pi t}} \displaystyle\int_{-\infty}^{a} \mathrm{e}^{-\xi^2/2t}\mathrm{d}\xi,\ t>0,\ a \in \mathbb{R}$;

(c) For every $r \in \mathbb{N}$ and $0 < t_1 < \cdots < t_r < \infty$, the increments $B(t_1)-B(0)$, $B(t_2)-B(t_1),\dots,B(t_r)-B(t_{r-1})$ are totally independent. □

Putting $0 \le t \le 1$ in (b) and $t_r = 1$ in (c) gives in effect $B : [0,1] \mapsto \mathbb{R}$ where $B \sim_{\mathrm{d}} W$ from Definition **27.16**. Given property **28.6**(c), the variance of an increment $B(s)-B(t)$ for $s>t$ is $s-t$ according to property **28.6**(b) and therefore, by considering increments small enough, conditions (b) and (c) jointly imply that the process is continuous with probability 1. Whether the fact of exceptions of probability 0 needs to be formally acknowledged is no more than a matter of preference, and in this sense the definitions are equivalent. In common usage, the terms Wiener process and Brownian motion process are virtually synonymous.

B is evidently a Lévy process with the additional distinction of a.s. continuous sample paths. This property replaces **28.1**(a) and is a different and stronger condition than the continuity in probability specified as **28.1**(e). That the conditions $\mathrm{E}(B(t))=0$ and $\mathrm{E}(B(t)^2)=t$ could replace **27.16**(b) under a.s. continuity, so that Gaussianity does not need to be asserted separately, is proved in **28.21**.

The graph of a random element of Brownian motion, $B(\omega,\cdot)$ for $\omega \in \Omega$, is quite a remarkable object. Figure 28.1 shows the typical appearance of a realization on $[0,1]$. Many people find the fact that such an object has independent increments to be quite counter-intuitive. These curves belong to the class of geometrical forms named fractals by Mandelbrot ([129]), and like the Levy motion of **28.3** are self-similar, meaning that their appearance is invariant to scaling operations. If B is a Brownian motion so is B^*, where

$$B^*(\omega,t) = k^{-1/2}\big(B(\omega,s+kt)-B(\omega,s)\big) \tag{28.15}$$

for $s \ge 0$ and $k > 0$.

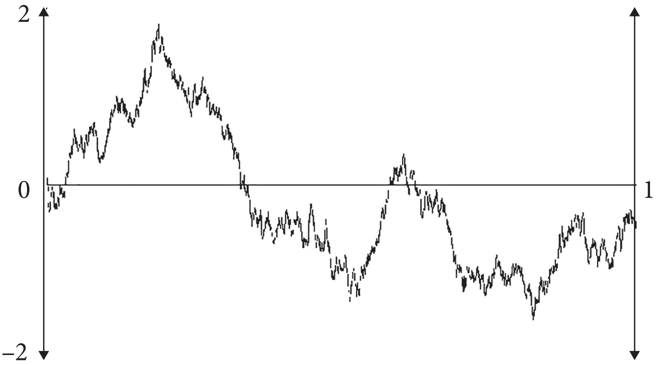

Figure 28.1

The key property is **28.6**(c), that of independent increments. A little thought is required to see what this means. In the definition, the points $t_1, \ldots, t_r$ may be arbitrarily close together. Considering a pair of points t and $t + \Delta$, the increment $B(\omega, t+\Delta) - B(\omega, t)$ is Gaussian with variance Δ and independent of $B(\omega, t)$. Symmetry of the Gaussian density implies that

$$P\big(\{\omega : \big(B(\omega, t+\Delta) - B(\omega, t)\big)\big(B(\omega, t) - B(\omega, t-\Delta)\big) < 0\}\big) = \tfrac{1}{2}$$

for $\Delta \leq t \leq 1 - \Delta$ and *every* $\Delta > 0$. This property is compatible with continuity, but completely rules out smoothness; in any realization of the process, almost every point of the graph is a corner and has no tangent. The property is also apparent when attempting to differentiate B. Note from the definition that

$$\frac{B(t+h) - B(t)}{h} \overset{\mathrm{d}}{\sim} \mathrm{N}(0, 1/h). \tag{28.16}$$

The sequence of measures defined by letting $h \to 0$ in (28.16) is not uniformly tight and fails to converge to any limit. This is another way of saying that the sample path $B(\omega)$ is everywhere non-differentiable, almost surely.

One way to think about Brownian motion is as the limit as $n \to \infty$ of the sequence of partial sums of n independent standard Gaussian r.v.s, scaled by $n^{-1/2}$. Define

$$\xi_j(\omega) = n^{1/2}\big(B(\omega, j/n) - B(\omega, (j-1)/n)\big) \tag{28.17}$$

for $j = 1, \ldots, [nt]$ where $[nt]$ denotes the integer part of nt, or equivalently write

$$B(\omega, j/n) = \frac{1}{\sqrt{n}} \sum_{i=1}^{j} \xi_i(\omega) \tag{28.18}$$

and note that $\xi_i \sim_{\mathrm{d}} \mathrm{N}(0,1)$ and independent of one another by assumption. By taking n large enough, $B(\omega, t)$ may be expressed in this form for any rational t, and by a.s. continuity of the process,

$$B(\omega,t) = \lim_{n\to\infty} \frac{1}{\sqrt{n}} \sum_{j=1}^{[nt]} \xi_j(\omega) \text{ a.s.} \tag{28.19}$$

for any $t > 0$. Figure 28.1 was constructed by plotting the partial sums of a sequence of 10,000 computer-generated random increments. Since 10,000 far exceeds the horizontal resolution of the graphic, self-similarity implies that it cannot be distinguished from a plot of the continuous limit process.

Consider the expected sum of the absolute values of the increments contributing to $B(t)$ in (28.19). According to **9.15**, $|\xi_j|$ has mean $(2/\pi)^{1/2}$ and variance $1 - 2/\pi$ and so by independence the r.v. $A_n(t) = n^{-1/2}\sum_{i=1}^{[nt]}|\xi_i|$ has mean equal to $[nt](2/n\pi)^{1/2}$ and variance of $[nt](1-2/\pi)/n$. Call these quantities $m(t,n)$ and $\nu(t,n)$ respectively. Applying Chebyshev's inequality, for $t > 0$

$$P\big(A_n(t) > \tfrac{1}{2}m(t,n)\big) \geq P\big(|A_n(t) - m(t,n)| \leq \tfrac{1}{2}m(t,n)\big) \geq 1 - \frac{4\nu(t,n)}{m(t,n)^2}. \tag{28.20}$$

Since $m(t,n) = O(n^{1/2})$, $A_n(t) \to \infty$ a.s.$[P]$ for all $t > 0$. This means that the random element $B(\omega)$ is a function of unbounded variation, almost surely. Since $\lim_{n\to\infty} A_n(t)$ is the total distance supposedly travelled by a Brownian particle as it traverses the interval from 0 to t and this turns out to be infinite for $t > 0$, Brownian motion cannot be taken as a literal description of such things as particles undergoing thermal agitation. Rather, it provides a simple limiting approximation to actual behaviour when the increments are small.

28.4 Properties of Brownian Motion

It is apparent that Brownian motion is both a Markov process and a diffusion. It is also a square-integrable martingale with respect to the filtration $\mathcal{F}(t) = \sigma(B(s), s \leq t)$. The martingale property is an obvious consequence of the independence of the increments of B. A further special feature of B is that the quadratic variation process is deterministic. Definition **28.6** implies that, for $s \geq t$,

$$\begin{aligned} \mathrm{E}\big(B(s)^2|\mathcal{F}(t)\big) - B(t)^2 &= \mathrm{E}\big([B(s)-B(t)]^2|\mathcal{F}(t)\big) \\ &= s - t \text{ a.s.}[P] \end{aligned} \tag{28.21}$$

and rearrangement of the equality shows that $B(t)^2 - t$ is a martingale; that is, $\langle B\rangle(t) = t$.

The first important result of this section is the *reflection principle*. This was shown for discrete random walks in §13.4 but extends directly to the continuous case; the supremum of a Brownian motion path over a given interval is easily shown to be distributed as half-normal (see **8.20**) with variance matching the width of the interval.

28.7 Theorem $P\big(\sup_{0\le s\le t} B(s) > x\big) = 2P(B(t) > x)$.

Proof Consider for $\{B(s), 0 \le s \le t\}$ the discrete-time process $\{U_i = B(2^{-n}i) - B(2^{-n}(i-1)), i = 1, \ldots, [2^n t] - 1\}$. Since $B(0) = 0$, $B(2^{-n}j) = \sum_{i=1}^{j} U_i$. Since these increments are Gaussian with mean 0, equality

$$P\Big(\max_{1\le j\le [2^n t]} B(2^{-n}j) > x\Big) = 2P(B(2^{-n}[2^n t]) > x)$$

holds by **13.15**. Letting $n \to \infty$, this equality becomes

$$P\Big(\sup_{s\in\mathbb{D}\cap[0,t]} B(s) > x\Big) = 2P(B(t) > x) \tag{28.22}$$

where $\mathbb{D}$ denotes the dyadic rationals. Since B is almost surely continuous, the left side of (28.22) is arbitrarily close to $P\big(\sup_{0\le s\le t} B(s) > x\big)$. ∎

This is a neat demonstration of the technique of extending a limit result from a special case to a general case, using an invariance principle. For any empirical process converging to Brownian motion, the half-normal distribution of the supremum is often a useful approximation.

The next result is the so-called *strong Markov* property of Brownian motion.

28.8 Theorem If τ is a stopping time with respect to filtration $\{\mathcal{F}(t), t \in [0,\infty)\}$, the process $B^{(\tau)}$ where

$$B^{(\tau)}(t) = B(t+\tau) - B(\tau),\ t \ge 0 \tag{28.23}$$

is a Brownian motion with $\mathrm{E}(B^{(\tau)}(t)) = 0$ and $\mathrm{E}(B^{(\tau)}(t)^2) = t$ and is distributed independently of $\mathcal{F}(\tau)$. □

The qualifier 'strong' here denotes the case where the Brownian process is restarted as in (28.23) at a random time. With the random variable τ in (28.23) replaced by a fixed value $u > 0$, the independence of the processes $\{B(t), 0 < t \leq u\}$ and $\{B^{(u)}(t), t > 0\}$ follows directly from Definition **28.6**. This property may be characterized in terms of the finite-dimensional distributions. Consider vectors $(B(s_1), \ldots, B(s_k))' \in \mathbb{R}^k$ for arbitrary $s_1 \leq s_2 \leq \cdots \leq s_k \leq u$ and arbitrary finite k. This random vector defines an element of the set $\mathcal{G}_u$ of inverse projections, similarly to (27.14) but for the space $C_{[0,u]}$ instead of $R_{[0,1]}$ and $\sigma(\mathcal{G}_u) = \mathcal{F}(u)$ by **27.15**. Similarly consider $(B^{(u)}(t_1), \ldots, B^{(u)}(t_m))' \in \mathbb{R}^m$ for arbitrary $m > 0$ and $t_1 \leq \cdots \leq t_m$, where

$$B^{(u)}(t) = B(t+u) - B(u),\ t \geq 0 \tag{28.24}$$

and let $\mathcal{H}_u$ denote the counterpart of $\mathcal{G}_u$ for these vectors. There are two important facts about these collections. The first is that the same probability measures apply since the processes B and $B^{(u)}$ are Brownian motions. If $m = k$ and $G \in \mathcal{G}_u$ and $H \in \mathcal{H}_u$ are the inverse projections of matching Borel sets of $\mathbb{R}^k$ defined for matching coordinates $t_j = s_j$ for $j = 1, \ldots, k$, then $P(H) = P(G)$. The second important fact is that $\mathcal{G}_u$ and $\mathcal{H}_u$ have the independence property since they relate to non-overlapping segments of Brownian motion. Since these are π-systems the σ-fields $\mathcal{F}(u)$ and $\sigma(\mathcal{H}_u)$ also have the independence property, by **7.7**. The strong Markov property extends these properties from fixed u to random stopping time τ.

Proof of 28.8 τ is a stopping time if $\{\tau \leq u\} \in \mathcal{F}(u)$ for $u > 0$ and the same is true of the event $\{\tau = u\} = \{t \leq u\} - \bigcup_n \{\tau \leq u - 1/n\}$. Suppose first that the distribution of τ is discrete, with a countable set of outcomes $u_n \in M_n$ for $n \in \mathbb{N}$, where $M_n = \{2^{-n}k,\ k \in \mathbb{N}\}$. Let the event $G_\tau \in \mathcal{G}_\tau$ be defined for the joint distribution of the random pair $\{(B(s_1), \ldots, B(s_k))', \tau\}$ so that the collection of subsets $\{G_\tau \cap \{\tau = u_n\},\ u_n \in M_n\}$ is a partition of G_τ, disjoint with union G_τ. Similarly let $H_\tau \in \mathcal{H}_\tau$ be defined for a set of vectors $(B^{(\tau)}(t_1), \ldots, B^{(\tau)}(t_k))'$. $B^{(\tau)}$ is distributed like $B^{(u)}$ in (28.24) and accordingly like B. Hence note that

$$\begin{aligned} P(H_\tau \cap G_\tau) &= \sum_{u_n \in M_n} P\big(H_\tau \cap G_\tau \cap \{\tau = u_n\}\big) \\ &= P(H_\tau) \sum_{u_n \in M_n} P\big(G_\tau \cap \{\tau = u_n\}\big) \\ &= P(H_\tau)P(G_\tau) \end{aligned} \tag{28.25}$$

where the second equality of (28.25) holds because for fixed u_n the sets $G_\tau \cap \{\tau = u_n\}$ and H_τ relate to non-overlapping Brownian segments and are independent.

For general τ, define $\tau_n = 2^{-n}k$ if $2^{-n}k \leq \tau < 2^{-n}(k+1)$. τ_n is a stopping time because $\{\tau_n \leq t\} \in \mathcal{F}(t)$ holds for each t that satisfies the condition $\{\tau \leq t\} \in \mathcal{F}(t)$. Hence, (28.25) holds for the discrete random variable τ_n. Letting $n \to \infty$, M_n converges to the set of dyadic rationals on each unit interval of $\mathbb{R}^+$ and **27.13** implies that $G_{\tau_n} \to G_\tau$ and $H_{\tau_n} \to H_\tau$. It follows by Corollary **3.10** that $P(G_{\tau_n} \cap H_{\tau_n}) \to P(G_\tau \cap H_\tau)$ and $P(G_{\tau_n})P(H_{\tau_n}) \to P(G_\tau)P(H_\tau)$. ∎

Finally, here is a further interesting and useful fact about the stopping times of Brownian motion.

28.9 Theorem If $\tau \in [0,t]$ is a stopping time, $\mathrm{E}(B(\tau)) = 0$, $\mathrm{E}(B(\tau)^2) = \mathrm{E}(\tau)$ and $\mathrm{E}(B(\tau)^4) \geq \frac{1}{4}\mathrm{E}(\tau^2)$.

Proof Consider a martingale $\{Z(t), \mathcal{F}(t)\}$ having the property $\mathrm{E}(Z(t)|\mathcal{F}(s)) = Z(s)$, $0 \leq s \leq t$ and hence $\mathrm{E}(Z(t)) = Z(0) = 0$. Letting $\tau \leq t$ be a stopping time, $\mathrm{E}(Z(\tau)) = 0$ is shown as follows. First suppose that the distribution of τ has finite range $\tau_1 < \tau_2 < \cdots < \tau_m = t$. Then, since $\mathrm{E}(Z(\tau_j)) = \mathrm{E}(\mathrm{E}(Z(t)|\tau_j)) = \mathrm{E}(Z(t))$ by the LIE,

$$\begin{aligned}\mathrm{E}(Z(\tau)) &= \sum_{j=1}^{m} \mathrm{E}\big(Z(\tau)|\tau = \tau_j\big)P(\tau = \tau_j) \\ &= \sum_{j=1}^{m} \mathrm{E}\big(Z(\tau_j)\big)P(\tau = \tau_j) \\ &= \mathrm{E}\big(Z(t)\big)\sum_{j=1}^{m} P(\tau = \tau_j) = 0. \end{aligned} \tag{28.26}$$

Now consider the general case of a distribution of τ on $[0,t]$ and for $n \in \mathbb{N}$ define the stopping time $\tau_n = 2^{-n}jt$ when $\tau \in [2^{-n}(j-1)t, 2^{-n}jt]$ for $1 \leq j \leq 2^n$. The above argument shows $\mathrm{E}(Z(\tau_n)) = 0$. Letting $n \to \infty$, the dominated convergence theorem (**4.16**) implies $\mathrm{E}(Z(\tau_n)) \to \mathrm{E}(Z(\tau)) = 0$ if it can be shown that $\mathrm{E}(\sup_{s\leq t}|Z(s)|) < \infty$.

The three cases of interest for which this result applies are $Z(t) = B(t)$, $Z(t) = B(t)^2 - t$, and $Z(t) = B(t)^4 - 6tB(t)^2 + 3t^2$. Simple calculations show that $\mathrm{E}(Z(t)|\mathcal{F}(s)) = Z(s)$ for $t > s$ and hence $\mathrm{E}(Z(\tau)) = 0$ for each of these cases. The fact that $\mathrm{E}(\sup_{s\leq t}|Z(s)|) < \infty$ is readily deduced from **28.7**, the probability of $|Z(t)|$ exceeding a finite bound declining exponentially in each case. Considering the three cases of $Z(t)$: $\mathrm{E}(B(\tau)) = 0$ and $\mathrm{E}(B(\tau)^2) = \mathrm{E}(\tau)$ are immediate; for the third case, $\mathrm{E}(Z(\tau)) = 0$ and the Cauchy–Schwarz inequality give

$$\mathrm{E}(B(\tau)^4) = 6\mathrm{E}(\tau B(\tau)^2) - 3\mathrm{E}(\tau^2) \leq 6\sqrt{\mathrm{E}(\tau^2)}\sqrt{\mathrm{E}(B(\tau)^4)} - 3\mathrm{E}(\tau^2).$$

Completing the square shows that this inequality is violated unless $\mathrm{E}(\tau^2) \leq 4\mathrm{E}(B(\tau)^4)$. ∎

28.5 Skorokhod Embedding

It is a curious fact that it is possible to generate drawings of any r.v. X having zero mean and finite variance by means of a stopped Brownian motion. The idea is to observe the path of a Brownian motion B and specify a stopping time τ derived from the distribution of X. If this procedure is repeated many times with independent drawings, with suitable τ the distribution of $B(\tau)$ so generated can match that of X. As originally shown by A. V. Skorokhod ([172]), any such distribution can be 'embedded' in Brownian motion, in this sense.

This turns out to be a useful device for implementing an invariance principle. To show how the trick is accomplished is conveniently done in two stages; first to show how it works for two-point distributions, and then to show the generalization.

28.10 Theorem Let a r.v. X have a two-point distribution with atoms $-u$ and v for $u, v > 0$ and zero mean. If $B : [0,\infty) \mapsto \mathbb{R}$ denotes a standard Brownian motion there exists a stopping time τ with $\mathrm{E}(\tau) < \infty$ such that $B(\tau) \sim_{\mathrm{d}} X$.

Proof The distribution of X is completely determined by the three stated conditions since the probabilities $P(X = -u) = v/(u+v)$ and $P(X = v) = u/(u+v)$ uniquely imply $\mathrm{E}(X) = 0$. Let

$$\tau = \inf\{t : \text{either } B(t) = -u \text{ or } B(t) = v\}. \tag{28.27}$$

To show that $\mathrm{E}(\tau) < \infty$, define the events $A_j = \{|B(j) - B(j-1)| \leq u + v\}$ for $j = 1, 2, \ldots$. Since $B(j) - B(j-1) \sim_{\mathrm{d}} \mathrm{N}(0,1)$, $P(A_j) < 1$ and since these increments are distributed independently there exists $\gamma > 0$ such that $P(A_1 \cap \ldots \cap A_m) = P(A_1) \cdots P(A_m) < (1-\gamma)^m \to 0$ as $m \to \infty$. If $|B(j) - B(j-1)| > u + v$ then at least one of the conditions $B(j-1) \leq -u$, $B(j) \leq -u$, $B(j-1) \geq v$, and $B(j) \geq v$ holds, any one of which implies $\tau \leq j$ since B has continuous sample paths. The probability of this event for some $j \leq m$ converges geometrically to 1 as $m \to \infty$, hence $\tau < \infty$ a.s. and $\mathrm{E}(\tau) < \infty$. A finite t therefore exists w.p.1 such that $\tau < t$ and it follows by **28.9** that $\mathrm{E}(B(\tau)) = 0$. Either $B(\tau) = -u$ or $B(\tau) = v$ and the distribution of $B(\tau)$ therefore satisfies the same three conditions as that of X. Its distribution is likewise uniquely determined. ∎

Note the additional implication of Theorem **28.9** that $\mathrm{E}(\tau) = \mathrm{E}(X^2) = uv$. It may surprise the reader to discover that the simple rule in (28.27) dictates the respective probabilities of the two possible stops, but this is what the distribution of the Brownian motion dictates, for the reasons elucidated in the proof of **28.9**.

Since two-point distributions have a limited range of application, this result might be found to be a mere curiosity. The fact that it can be generalized to any random variable with zero mean and finite variance is the remarkable conclusion of Skorokhod's first embedding theorem, which is as follows.

28.11 Theorem If X is a r.v. with p.m. μ having zero mean and finite variance, there exists a stopping time τ such that $B(\tau) \sim_{\mathrm{d}} \mu$ with $\mathrm{E}(\tau) = \mathrm{E}(X^2)$ and $\mathrm{E}(\tau^2) \leq 4\mathrm{E}(X^4)$.

Proof Consider a martingale sequence $\{S_n, \mathcal{F}_n\}_0^\infty$ where $S_n = \mathrm{E}(X|\mathcal{F}_n)$. By suitably defining the filtration $\{\mathcal{F}_n\}$ this can be constructed so that $S_n|\mathcal{F}_{n-1}$ has a two-point distribution for each n. The construction is detailed in Example **16.4**. The embedding procedure is to draw these coordinates sequentially from the conditional distributions. In view of the strong Markov property of B (**28.8**) the draw at step n can be made from the process $B^*(t) = B(t + \tau_{n-1}) - B(\tau_{n-1})$, yielding the stopping time $\tau_n^* = \tau_n - \tau_{n-1}$ so that $\tau_n = \tau_1 + \tau_2^* + \cdots + \tau_n^*$. According to Theorem **28.10**, S_n uniquely matches $B(\tau_n)$ in distribution for each n. $S_n \to_{\text{a.s.}} X$ by martingale convergence so that $\tau_n \to_{\text{a.s.}} \tau$ where path-continuity of Brownian motion implies that $B(\tau_n) \to_{\text{a.s.}} B(\tau)$. The theorem now follows by Corollary **3.10** applied to sets such as $A_n = \{S_n \leq x\}$ for $x \in \mathbb{R}$ and also **28.9**, since $\mathrm{E}(X^2) = \mathrm{E}(B(\tau)^2)$ and $\mathrm{E}(X^4) = \mathrm{E}(B(\tau)^4)$ where either the two parts of the latter equality are finite and equal or both are infinite. ∎

This result is surely intriguing and unexpected—an arbitrary distribution is conjured out of a Brownian motion path by the judicious choice of stopping times. That it is more than just a curiosity emerges from the second Skorokhod embedding theorem.

28.12 Theorem Consider $S_n = X_1 + \cdots + X_n$ where the X_j are identically and independently distributed random variables with mean zero and finite variance σ^2. There exists a stopping time of Brownian motion τ_n such that $B(\tau_n)$ is distributed like S_n and $\tau_n = \sum_{j=1}^n (\tau_j - \tau_{j-1})$ with $\tau_0 = 0$ where the sequence $\{\tau_j - \tau_{j-1}, 1 \leq j \leq n\}$ is identically and independently distributed with mean σ^2.

Proof The fact that $X_1 \sim_{\mathrm{d}} B(\tau_1)$, with $\mathrm{E}(\tau_1) = \mathrm{E}(X_1^2) = \sigma^2$ follows from Theorem **28.11**. Now form the Brownian motion $B^*(t) = B(\tau_1 + t) - B(\tau_1)$ and find stopping time τ_2 such that $X_2 \sim_{\mathrm{d}} B^*(\tau_2 - \tau_1)$ and hence $X_1 + X_2 \sim_{\mathrm{d}} B(\tau_1) + B^*(\tau_2 - \tau_1) = B(\tau_2)$. In view of the strong Markov property and the fact that X_2 and X_1 are independent with the same distribution, the distribution of the random variable $\tau_2 - \tau_1$ is independent of and identical to that of τ_1, with $\mathrm{E}(\tau_2 - \tau_1) = \sigma^2$ by further application of **28.11**. Proceeding by induction, the same argument establishes successively for $m = 1, \dots, n$ that $\sum_{j=1}^{m} X_j$ is distributed like $B(\tau_m)$ with the sequence $\{\tau_j - \tau_{j-1}, j \le m\}$ having the asserted properties. ∎

The interesting implication of Theorem **28.12** is that whereas the distribution of S_n drives the distribution of $B(\tau_n)$, as n get large the matching of distributions becomes a two-way street. By the strong law of large numbers and Theorem **28.11**, $n^{-1}\tau_n \to_{\text{a.s.}} \mathrm{E}(\tau_1) = \sigma^2$. This means that with n large, S_n must be distributed approximately like $B(n\sigma^2)$. However, this distribution is Gaussian with variance $n\sigma^2$. This fact gives an alternative proof of the central limit theorem for independent and identically distributed random variables. Comparing this proof with Theorem **24.3** (Lindeberg–Lévy), it is notable that the arguments in each case barely intersect, with no mention here of characteristic functions.

28.13 Theorem If $S_n = X_1 + \cdots + X_n$ where the X_j are independently and identically distributed with mean 0 and variance σ^2, then $n^{-1/2}S_n/\sigma \to_{\mathrm{d}} \mathrm{N}(0,1)$.

Proof According to the second embedding theorem **28.12**, $B(\tau_n)$ and S_n have the same distribution, for any n. Establishing the limiting distribution of the first of these gives that of the second.

For real numbers a, b, x, and $\varepsilon > 0$ note the following implications, with each set of cases contained in the preceding one:

$$\{a - b \le x\} \Rightarrow \{a \le x + |b|\} \Rightarrow \{a \le x + \varepsilon\} \cup \{|b| \ge \varepsilon\}.$$

With the obvious substitutions of a and b by random variables, these imply by subadditivity the inequality

$$P\Big(\frac{B(\tau_n)}{\sigma\sqrt{n}} \le x\Big) \le P\Big(\frac{B(n\sigma^2)}{\sigma\sqrt{n}} \le x + \varepsilon\Big) + P\Big(\frac{|B(\tau_n) - B(n\sigma^2)|}{\sigma\sqrt{n}} \ge \varepsilon\Big). \tag{28.28}$$

Similarly, the implications

$$\{a + \varepsilon \le x\} \Rightarrow \{a + |b| \le x\} \cup \{|b| \ge \varepsilon\} \Rightarrow \{a + b \le x\} \cup \{|b| \ge \varepsilon\}$$

translate into the inequality

$$P\Big(\frac{B(n\sigma^2)}{\sigma\sqrt{n}} \le x - \varepsilon\Big) \le P\Big(\frac{B(\tau_n)}{\sigma\sqrt{n}} \le x\Big) + P\Big(\frac{|B(\tau_n) - B(n\sigma^2)|}{\sigma\sqrt{n}} \ge \varepsilon\Big). \tag{28.29}$$

Noting that $B(n\sigma^2)/\sigma\sqrt{n} \sim_{\mathrm{d}} \mathrm{N}(0,1)$, the theorem is proved by showing that the second term appearing on the right-hand sides of (28.28) and (28.29) vanishes as $n \to \infty$ for all $\varepsilon > 0$.

Since τ_n is the sum of independently and identically distributed increments with means of σ^2 by **28.12**, the weak law of large numbers implies that $P(|\tau_n - n\sigma^2| > n\delta) \to 0$ as $n \to \infty$ for all $\delta > 0$. It is therefore possible to choose a sequence $\{\delta_n\}$ such that $\delta_n \downarrow 0$ as $n \to \infty$ and $P(|\tau_n - n\sigma^2| > n\delta_n) < \delta_n$ for each n. If $\tau_n > n\sigma^2$, consider the Brownian motion $B^*(t) = B(t + n\sigma^2) - B(n\sigma^2)$; otherwise take the case $B^*(t) = B(t + \tau_n) - B(\tau_n)$. Apply the reflection principle **28.7** to obtain, by Chebyshev's inequality,

$$\begin{aligned} P\Big(\sup_{0 \le t \le n\delta_n} |B^*(t)| > \sigma\sqrt{n}\varepsilon\Big) &= 4P\big(|B^*(n\delta_n)| > \sigma\sqrt{n}\varepsilon\big) \\ &\le 4\frac{n\delta_n}{n\sigma^2\varepsilon^2} = O(\delta_n). \end{aligned}$$

Thus, note that

$$\begin{aligned} P\big(|B(\tau_n) - B(n\sigma^2)| \ge \sigma\sqrt{n}\varepsilon\big) &\le P\Big(\sup_{0 \le t \le |\tau_n - n\sigma^2|} |B^*(t)| \ge \sigma\sqrt{n}\varepsilon\Big) \\ &\le P(|\tau_n - n\sigma^2| > n\delta_n) \\ &\quad + P\Big(\sup_{0 \le t \le n\delta_n} |B^*(t)| \ge \sigma\sqrt{n}\varepsilon\Big) \\ &= O(\delta_n) \qquad (28.30) \end{aligned}$$

where the second inequality holds because the occurrence of the event in the second member implies the occurrence of at least one of the events in the third member. This completes the proof. ∎

A second useful application of the embedding approach is a proof of Theorem **26.10**, the law of the iterated logarithm.

Proof of 26.10 Put $\sigma^2 = 1$ without loss of generality, since this is equivalent to replacing X_t in (26.42) by X_t/σ. Theorem **26.12** deals with the case of **26.10** where the increments of the partial sum process S_n are independent $\mathrm{N}(0,1)$-distributed, so that S_n is distributed like $B(n)$. On the other hand, Theorem **28.12** shows that in the present case S_n is distributed like $B(\tau_n)$, a Gaussian process with variance $\mathrm{E}(\tau_n)$ by **28.9**. The result therefore follows if it can be shown that $B(\tau_n)$ and $B(n)$ are close under suitable normalization when n is large.

Since τ_n is a sum of i.i.d. increments with mean σ^2 the strong law of large numbers applies (see e.g. **14.6**) and $\tau_n/n \to_{\mathrm{a.s.}} 1$. Letting $u(n) = \sqrt{2n\log\log n}$, write $u(\tau_n) = u((\tau_n/n)n)$ and apply **19.12**(i) to the case $g_n(\cdot) = u((\cdot)n)/u(n)$ to show that $u(\tau_n)/u(n) \to_{\text{a.s.}} 1$. Let $n_k = (1+\varepsilon)^k$ for $0 < \varepsilon < 1$. Given ε and noting $n_{k+1} - n_k = n_k\varepsilon$, there exists n large enough that $n_k \le n \le n_{k+1}$ and $n_k \le \tau_n \le n_{k+1}$. The reflection principle **13.16**, subadditivity, and (8.22) give, similarly to (26.48),

$$\begin{aligned} P\Big(|B(\tau_n) - B(n)| \ge \sqrt{2\varepsilon}u(n_k)\Big) &\le P\Big(\max_{n_k<n\le n_{k+1}} |B(n) - B(n_k)| \ge \sqrt{2\varepsilon}u(n_k)\Big) \\ &\quad + P\Big(\max_{n_k<\tau_n\le n_{k+1}} |B(\tau_n) - B(n_k)| \ge \sqrt{2\varepsilon}u(n_k)\Big) \\ &\le 8k^{-2}\log(1+\varepsilon)^{-2}. \end{aligned}$$

Since these probabilities are summable over k, the Borel–Cantelli lemma **19.2**(i) implies that there is k large enough and hence n large enough that

$$\frac{|B(\tau_n) - B(n)|}{u(n_k)} < \sqrt{2\varepsilon} \text{ a.s.} \tag{28.31}$$

Since ε is arbitrary, let $\varepsilon \downarrow 0$ to complete the proof. ∎

28.6 Processes Derived from Brownian Motion

Standard Brownian motion is the leading member of an extensive family of a.s. continuous processes on $[0,1]$ having Gaussian characteristics.

Multiplying B by a constant $\sigma > 0$ produces a Brownian motion with variance σ^2. Then, adding the deterministic function μt to the process defines a Brownian motion with *drift* μ. Thus, $X(t) = \sigma B(t) + \mu t$ represents a family of processes having independent increments of the form

$$X(t) - X(s) \overset{\mathrm{d}}{\sim} \mathrm{N}\big(\mu(t-s), \sigma^2|t-s|\big).$$

In the presence of drift, Brownian motion is semimartingale.

More elaborate generalizations include the following. In each of the following examples, B denotes a standard Brownian motion on $[0,\infty)$ unless specified otherwise. The restriction to $[0,1]$ is in most cases transparent. In the usual way, the argument $\omega \in \Omega$ where it appears denotes a point in the probability space $(\Omega, \mathcal{F}, \mathcal{P})$.

28.14 Example Let $X(t) = B(t^\theta)$ for $\theta > 0$. X is a Brownian motion that has been subjected to stretching and squeezing of the time domain. Like B, it is a.s.

continuous with independent Gaussian increments. It can be thought of as the limit of a partial sum process whose increments have trending variance. Suppose $\xi_i \sim_{\mathrm{d}} \mathrm{N}(0, \sigma_i^2)$ for $i = 1, \ldots, n$ are independently distributed shocks with variances $\sigma_i^2 = \theta i^{\theta-1}$ for $\theta > 0$, tending to 0 if $\theta < 1$, or to infinity if $\theta > 1$. The integral approximation argument of **2.17** yields $n^{-\theta} \sum_{i=1}^{[nt]} \sigma_i^2 \to t^\theta$ and hence

$$n^{-\theta/2} \sum_{i=1}^{[nt]} \xi_i \to B(t^\theta) \text{ a.s.} \tag{28.32}$$

This is a case of what is called *transformed* (or variance-transformed) Brownian motion. □

28.15 Example If $Y \sim_{\mathrm{d}} \mathrm{N}(0, \sigma^2)$ then using (9.27) note that

$$\begin{aligned} \mathrm{E}(\mathrm{e}^Y) &= 1 + \sum_{k=1}^{\infty} \frac{1}{(2k)!} \frac{(2k)!\sigma^{2k}}{2^k k!} \\ &= 1 + \sum_{k=1}^{\infty} \frac{1}{k!} \left(\frac{\sigma^2}{2}\right)^k = \mathrm{e}^{\sigma^2/2}. \end{aligned}$$

Extending this property to Brownian motion, $\mathrm{E}(\mathrm{e}^{B(t)}) = \mathrm{e}^{t/2}$. It follows that $X(t) = \mathrm{e}^{B(t)-t/2}$ is a continuous-time stochastic process with zero mean and the adapted process $(X(t), \mathcal{F}(t))$ is a martingale, since for $s \geq t$ one can write $X(s) = \mathrm{e}^{B(s)-B(t)-(s-t)/2}\mathrm{e}^{B(t)-t/2}$ and hence

$$\begin{aligned} \mathrm{E}(X(s)|\mathcal{F}(t)) &= \mathrm{E}(\mathrm{e}^{B(s)-B(t)-(s-t)/2}|\mathcal{F}(t))\mathrm{e}^{B(t)-t/2} \\ &= \mathrm{e}^{B(t)-t/2} = X(t) \text{ a.s.} \end{aligned}$$

This is called *exponential Brownian motion*. The process $\mathrm{e}^{B(t)}$, without a mean correction, is a submartingale. □

28.16 Example Let $X(t) = \theta(t)B(t)$ where $\theta : [0,1] \mapsto \mathbb{R}$ is any continuous deterministic function. This is a Gaussian diffusion but it has dependent increments and is not a martingale, noting that for $s > t$,

$$\mathrm{E}(X(s)|\mathcal{F}(t)) = X(t) + (\theta(s) - \theta(t))B(t). \tag{28.33}$$

It can be visualized as the almost sure limit as $n \to \infty$ of a double partial sum process,

$$X_n(t) = n^{-1/2}\sum_{i=1}^{[nt]}\Big(\theta(i/n)\xi_i(\omega) + \big(\theta(i/n) - \theta((i-1)/n)\big)\sum_{j=1}^{i-1}\xi_j(\omega)\Big) \tag{28.34}$$

where $\xi_i \sim_{\mathrm{d}} \mathrm{N}(0,1)$. □

28.17 Example Let

$$X(t) = \mathrm{e}^{-\beta t}B(\mathrm{e}^{2\beta t}) \tag{28.35}$$

for fixed $\beta > 0$. This is a zero-mean Gaussian process having dependent increments similarly to **28.16**. The remarkable feature of this process is that it is stationary, with $X(t) \sim_{\mathrm{d}} \mathrm{N}(0,1)$ for all $t > 0$ and

$$\mathrm{E}\big(X(t)X(s)\big) = \mathrm{e}^{\beta(2\min\{t,s\}-t-s)} = \mathrm{e}^{-\beta|t-s|}. \tag{28.36}$$

This is the *Ornstein–Uhlenbeck* process. □

28.18 Example The *Brownian bridge* is the process $B^\circ \in C_{[0,1]}$ where

$$B^\circ(t) = B(t) - tB(1),\ t \in [0,1]. \tag{28.37}$$

This is a Brownian motion tied down at both ends and has $\mathrm{E}\big(B^\circ(t)B^\circ(s)\big) = \min\{t,s\} - ts$. A natural way to think about B° is as the limit of the partial sums of a mean-deviation process; that is

$$B^\circ(t,\omega) = \lim_{n\to\infty}\frac{1}{\sqrt{n}}\sum_{i=1}^{[nt]}\Big(\xi_i(\omega) - \frac{1}{n}\sum_{j=1}^{n}\xi_j(\omega)\Big)\ \text{a.s.} \tag{28.38}$$

where $\xi_i(\omega) \sim_{\mathrm{d}} \mathrm{N}(0,1)$ and i.i.d. □

The next cases are distinguished by being defined as stochastic integrals. The technical underpinnings of such processes can be found detailed in §32.2, but these are conceptually somewhat simpler than the objects defined in that section and are best understood in the framework of expression (28.19). There, Brownian motion is represented as the limiting sum of independent Gaussian increments and a logical change of notation allows the representation

$$B(\omega,t) = \lim_{n\to\infty} \frac{1}{\sqrt{n}} \sum_{i=1}^{[nt]} \xi_i(\omega) = \int_0^t \mathrm{d}B(\omega,r) \text{ a.s.} \tag{28.39}$$

where the i.i.d. Gaussian terms $\xi_i/\sqrt{n}$ are tending to the infinitesimal contributions dB. In the next examples, the expression in (28.39) is modified by replacing the simple sum with a weighted moving average.

28.19 Example Consider for $\beta > 0$,

$$X(t) = \sqrt{2\beta} \int_0^t \mathrm{e}^{-\beta(t-r)} \mathrm{d}B(r).$$

This is a variant of the Ornstein-Uhlenbeck process **28.17**. Using the Itô isometry (32.35) and the orthogonality of the non-overlapping Brownian segments, the covariance function is

$$\begin{aligned}
\mathrm{E}(X(t)X(s)) &= 2\beta \mathrm{e}^{-\beta(t+s)} \mathrm{E}\Big(\int_0^t \mathrm{e}^{\beta r} \mathrm{d}B(r) \int_0^s \mathrm{e}^{\beta r} \mathrm{d}B(r) \Big) \\
&= 2\beta \mathrm{e}^{-\beta(t+s)} \int_0^{\min\{t,s\}} \mathrm{e}^{2\beta r} \mathrm{d}r \\
&= \mathrm{e}^{-\beta|t-s|} - \mathrm{e}^{-\beta(t+s)},
\end{aligned}$$

approaching the stationary case when $t+s$ is large. □

28.20 Example Consider, for $-\frac{1}{2} < d < \frac{1}{2}$ and $t \in [0,1]$,

$$X(t) = \frac{1}{\Gamma(d+1)} \Big(\int_{-\infty}^t (t-r)^d \mathrm{d}B(r) - \int_{-\infty}^0 (-r)^d \mathrm{d}B(r) \Big). \tag{28.40}$$

Defined originally by Mandelbrot and van Ness ([130]), this process is known as *fractional Brownian motion*. The notable difference from **28.19** is that the moving average weights decline hyperbolically rather than geometrically and in particular the dependence on the infinitely remote past cannot be ignored. Sometimes the formula in (28.40) is truncated at $r = 0$ to define what is known as type 2 fractional Brownian motion, but this process has nonstationary increments, unlike the type 1 case shown. The covariances have the form

$$\mathrm{E}(X(t)X(s)) = \tfrac{1}{2} V_d \big(t^{2d+1} + s^{2d+1} - |t-s|^{2d+1} \big) \tag{28.41}$$

for $t, s > 0$ where

$$V_d = \frac{\Gamma(1-2d)}{(2d+1)\Gamma(1-d)\Gamma(1+d)} \tag{28.42}$$

(see e.g. [47]). The formula gives $\mathrm{E}(X(t) - X(s))^2 = V_d|t-s|^{2d+1}$, confirming that the increments are stationary. Also, setting $\delta > 0$, the covariance of adjacent increments is given as

$$\mathrm{E}\big((X(t+\delta) - X(t))(X(t) - X(t-\delta))\big) = V_d(2^{2d} - 1)\delta^{2d+1} \tag{28.43}$$

which has the sign of d. □

28.7 Independent Increments and Continuity

To conclude this chapter, a surprising fact about Wiener measure is proved by way of a theorem of Billingsley ([21] th. 19.1). Definition **27.16** is actually redundant; if part (b) of that definition is replaced by a.s. continuity and the specification of just the first two moments of $x(t)$, Gaussianity must follow.

28.21 Theorem Let $X : [0,1] \mapsto \mathbb{R}$ be a random function. $X \sim_{\mathrm{d}} W$ if

(a) $\mathrm{E}(X(t)) = 0$, $\mathrm{E}(X^2(t)) = t$, $0 \le t \le 1$

(b) $P(X \in C) = 1$

(c) For any partition $\{t_1, \dots, t_k\}$ of $[0,1]$ the increments $X(t_2) - X(t_1)$, $X(t_3) - X(t_2), \dots, X(t_k) - X(t_{k-1})$ are totally independent. □

The apparent triviality of these conditions is remarkable, but of course the clue is in the independence. If $X(t)$ can be decomposed as the sum of an infinity of tiny increments and these are all totally independent, a central limit property is implicit. The essential insight is that continuity of the sample paths corresponds to the Lindeberg condition being satisfied by the increments.

The proof of **28.21** falls into two parts of which the second part, which shows that the third absolute moment of the increments is bounded, is tricky and cumbersome. The first part, which derives the ch.f. of $X(t)$ directly subject to this condition and is fairly transparent, is conveniently given as a lemma.

28.22 Lemma Let the conditions of **28.21** hold and in addition let $\mathrm{E}|X(t+h) - X(t)|^3 = o(h)$ for $0 \le t < 1$. Then, $X \sim_{\mathrm{d}} W$.

Proof Let the characteristic function of $X(t)$ be

$$\phi(t,\lambda) = \mathrm{E}(\mathrm{e}^{i\lambda X(t)}). \tag{28.44}$$

By (11.32),

$$\mathrm{e}^{iu} = 1 + iu - \frac{1}{2}u^2 + r(u) \tag{28.45}$$

where $|r(u)| \le |u|^3$. To denote $X(s) - X(t)$ for $0 \le t \le s \le 1$ one or other of the notations $\Delta_{s,t}$ or $\Delta(s,t)$ will be used, as is most convenient. It is easily shown applying conditions (a) and (c) of **28.21** that $\mathrm{E}(\Delta^2_{t+h,t}) = h$. Therefore, by condition (c),

$$\begin{aligned}\phi(t+h,\lambda) - \phi(t,\lambda) &= \mathrm{E}\big(\mathrm{e}^{i\lambda X(t)}(\mathrm{e}^{i\lambda\Delta t+h,t} - 1)\big)\\ &= \mathrm{E}\big(\mathrm{e}^{i\lambda X(t)}\big)\mathrm{E}\big(i\lambda\Delta_{t+h,t} - \frac{1}{2}\lambda^2\Delta^2_{t+h,t} + r(\lambda\Delta_{t+h,t})\big)\\ &= \phi(t,\lambda)\big(-\tfrac{1}{2}\lambda^2 h + \mathrm{E}(r(\lambda\Delta_{t+h,t}))\big).\end{aligned} \tag{28.46}$$

Since $\mathrm{E}\big(r(\lambda\Delta_{t+h,t})\big) \le \lambda^3\mathrm{E}|\Delta_{t+h,t}|^3$, it follows that

$$\left|\frac{\phi(t+h,\lambda) - \phi(t,\lambda)}{h} + \frac{1}{2}\lambda^2\phi(t,\lambda)\right| \le \frac{\phi(t,\lambda)\lambda^3\mathrm{E}|\Delta_{t+h,t}|^3}{h}. \tag{28.47}$$

By the assumption of the lemma, the majorant side of (28.47) is $o(1)$ so ϕ possesses a right-hand derivative

$$\lim_{h\downarrow 0}\frac{\phi(t+h,\lambda) - \phi(t,\lambda)}{h} = -\tfrac{1}{2}\lambda^2\phi(t,\lambda) \tag{28.48}$$

for all $0 \le t \le 1$. For $h > 0$ and $h \le t \le 1$ (28.47) holds at the point $t - h$, so considering a path to the limit through such points shows there is also a left-hand derivative,

$$\lim_{h\downarrow 0}\frac{\phi(t,\lambda) - \phi(t-h,\lambda)}{h} = -\tfrac{1}{2}\lambda^2\phi(t-,\lambda). \tag{28.49}$$

ϕ is continuous in t by condition (b) of **28.21** and hence $\phi(t-,\lambda) = \phi(t,\lambda)$. It follows that ϕ is differentiable on $(0,1)$ and

$$\frac{\partial\phi}{\partial t} = -\tfrac{1}{2}\lambda^2\phi(t,\lambda). \tag{28.50}$$

Differentiating $\log \phi$ with respect to t shows that this differential equation has the solution

$$\phi(t,\lambda) = \phi(0,\lambda)\mathrm{e}^{-t\lambda^2/2},\ t \geq 0 \tag{28.51}$$

where $\phi(0,\lambda) = 1$ since $P(X(0) = 0) = 1$, as a consequence of condition **28.21**(a) with $t = 0$. Applying the inversion theorem **11.18** yields $X(t) \sim_{\mathrm{d}} \mathrm{N}(0,t)$ for each $t \in (0,1)$. The result extends to $t = 1$ by continuity of ϕ at 1. The conditions of Definition **27.16** are therefore satisfied. ∎

It is necessary to assume a.s. continuity (condition (b)) to ensure that the characteristic function is differentiable and obtain (28.50), but it is easy to see that conditions (a), (b), and (c) of **27.16** imply a.s. continuity in turn.

The assumed property of the third absolute moment under the conditions of **28.21** remains to be proved. This calls for two technical lemmas and in the second case the proof is rather lengthy; the reader might prefer to take this on trust initially. If $\zeta_1,\ldots,\zeta_m$ is a random sequence and $S_j = \sum_{i=1}^{j} \zeta_i$ for $1 \leq j \leq m$ and $S_0 = 0$, the requirement is to bound the probability of $|S_m|$ exceeding a given value, and the lemmas work together to this end. The first is non-probabilistic and the increments can be arbitrary.

28.23 Lemma

$$|S_m| \leq 2 \max_{0\leq j\leq m} \min\{|S_j|, |S_m - S_j|\} + \max_{0\leq j\leq m} |\zeta_j|.$$

Proof Let $I \subseteq \{0,\ldots,m\}$ denote the set of integers k for which $|S_k| \leq |S_m - S_k|$. If $S_m = 0$ the lemma holds and if $S_m \neq 0$ then $m \notin I$. On the other hand, $0 \in I$. It follows that there is a $k \notin I$ such that $k - 1 \in I$. For this choice of k,

$$\begin{aligned} |S_m| &\leq |S_m - S_k| + |S_k| \\ &\leq |S_m - S_k| + |S_{k-1}| + |\zeta_k| \\ &\leq 2 \max_{0\leq j\leq m} \min\{|S_j|, |S_m - S_j|\} + \max_{0\leq j\leq m} |\zeta_j|. \quad \blacksquare \end{aligned} \tag{28.52}$$

The second lemma is a variation on the maximal inequality for partial sums. Here it is assumed that the increments are independent random variables with finite variances.

28.24 Lemma If

$$\mathrm{E}\big((S_j - S_i)^2(S_k - S_j)^2\big) \le \bigg(\sum_{l=i+1}^{k} b_l\bigg)^2,\ j = i, \dots, k \tag{28.53}$$

for each pair i, k with $0 \le i \le k \le m$, where $\{b_1, \dots, b_m\}$ is a collection of positive numbers, then $\exists\ K > 0$ such that, for all $\alpha > 0$ and all m,

$$P\Big(\max_{0\le j\le m} \min\{|S_j|, |S_m - S_j|\} \ge \alpha\Big) \le \frac{KB^2}{\alpha^4} \tag{28.54}$$

where $B = \sum_{j=1}^{m} b_j$.

Proof For $0 \le i \le k \le m$ and $\alpha > 0$,

$$\begin{aligned} P\big(\min\{|S_j - S_i|, |S_k - S_j|\} \ge \alpha\big) &= P\big(\{|S_j - S_i| \ge \alpha\} \cap \{|S_k - S_j| \ge \alpha\}\big) \\ &\le P\big(|S_j - S_i||S_k - S_j| \ge \alpha^2\big) \\ &\le \frac{1}{\alpha^4}\bigg(\sum_{l=i+1}^{k} b_l\bigg)^2,\ j = i, \dots, k \end{aligned} \tag{28.55}$$

where Chebyshev's inequality and (28.53) give the final inequality. If $m = 1$, the minorant side of (28.54) is zero. If $m = 2$, (28.55) with $i = 0$ and $k = 2$ yields

$$P\big(\min\{|S_1|, |S_2 - S_1|\} \ge \alpha\big) \le \frac{(b_1 + b_2)^2}{\alpha^4} \tag{28.56}$$

so that (28.54) holds for $K = 1$ and hence for any $K \ge 1$.

The proof now proceeds by induction. If there is a K for which (28.54) holds when m is replaced by any integer between 1 and $m - 1$, then it holds for m itself with the same K. To show this the basic idea is to split the sum into two parts, each with fewer than m terms, obtain valid inequalities for each part, and combine these. Choose h to be the largest integer such that $\sum_{j=1}^{h-1} b_j \le B/2$ (the sum is 0 if $h = 1$); it is easy to see that $\sum_{j=h+1}^{m} b_j \le B/2$ also (the sum being 0 if $h = m$). First define

$$U_1 = \max_{0\le j\le h-1} \min\big\{|S_j|, |S_{h-1} - S_j|\big\} \tag{28.57}$$

$$D_1 = \min\big\{|S_{h-1}|, |S_m - S_{h-1}|\big\}. \tag{28.58}$$

Evidently,

$$P(U_1 \geq \alpha) \leq \frac{K}{\alpha^4}\left(\sum_{j=1}^{h-1} b_j\right)^2 \leq \frac{KB^2}{4\alpha^4} \tag{28.59}$$

by the induction hypothesis. Also, by (28.55) with $i = 0$ and $k = m$,

$$P(D_1 \geq \alpha) \leq \frac{B^2}{\alpha^4}. \tag{28.60}$$

The object is now to show that

$$\min\{|S_j|, |S_m - S_j|\} \leq U_1 + D_1,\ 0 \leq j \leq h-1. \tag{28.61}$$

If $|S_j| \leq U_1$ then (28.61) holds. Hence, suppose $|S_{h-1} - S_j| \leq U_1$, the only other possibility according to (28.57). If $D_1 = |S_{h-1}|$, then

$$\min\{|S_j|, |S_m - S_j|\} \leq |S_j| \leq |S_{h-1} - S_j| + |S_{h-1}| \leq U_1 + D_1$$

and if $D_1 = |S_m - S_{h-1}|$ then again,

$$\min\{|S_j|, |S_m - S_j|\} \leq |S_m - S_j| \leq |S_{h-1} - S_j| + |S_m - S_{h-1}| \leq U_1 + D_1.$$

Hence (28.61) holds in all cases. Now, for $0 \leq \mu \leq 1$,

$$\begin{aligned} P(U_1 + D_1 \geq \alpha) &\leq P\big(\{U_1 \geq \mu\alpha\} \cup \{D_1 \geq (1-\mu)\alpha\}\big) \\ &\leq P(U_1 \geq \mu\alpha) + P\big(D_1 \geq (1-\mu)\alpha\big) \\ &\leq \frac{KB^2}{4\alpha^4\mu^4} + \frac{B^2}{\alpha^4(1-\mu)^4}. \end{aligned} \tag{28.62}$$

Choosing μ to minimize $K/4\mu^4 + 1/(1-\mu)^4$ yields $\mu = (\frac{1}{4}K)^{1/5}/\big(1 + (\frac{1}{4}K)^{1/5}\big)$. Back-substituting for μ and simplifying yields, for $K \geq 2\big(1 - (\frac{1}{2})^{1/5}\big)^{-5} \approx 55{,}021$,

$$P(U_1 + D_1 \geq \alpha) \leq \frac{B^2\big((\frac{1}{4}K)^{1/5} + 1\big)^5}{\alpha^4} \leq \frac{KB^2}{2\alpha^4}. \tag{28.63}$$

According to (28.61), $\min\{|S_j|, |S_m - S_j|\}$ is bounded in the range $0 \leq j \leq h-1$. To do the same for the range $h \leq j \leq m$, define

$$U_2 = \max_{h \leq j \leq m} \min\{|S_j - S_h|, |S_m - S_j|\} \tag{28.64}$$

and

$$D_2 = \min\{|S_h|, |S_m - S_h|\}. \tag{28.65}$$

It can be verified by variants of the previous arguments that

$$\min\{|S_j|, |S_m - S_j|\} \le U_2 + D_2,\ h \le j \le m \tag{28.66}$$

and also that

$$P(U_2 + D_2 \ge \alpha) \le \frac{KB^2}{2\alpha^4} \tag{28.67}$$

for the same choice of K. Combining (28.67) with (28.63),

$$\begin{aligned} P\Big(\max_{0\le j\le m} \min\{|S_j|, |S_m - S_j|\} \ge \alpha\Big) &\le P\big(\max\{U_1 + D_1, U_2 + D_2\} \ge \alpha\big) \\ &= P\big(\{U_1 + D_1 \ge \alpha\} \cup \{U_2 + D_2 \ge \alpha\}\big) \\ &\le P(U_1 + D_1 \ge \alpha) + P(U_2 + D_2 \ge \alpha) \\ &\le \frac{KB^2}{\alpha^4}. \quad \blacksquare \end{aligned} \tag{28.68}$$

Proof of 28.21 Given **28.22**, the remaining task is to prove

$$\lim_{h\downarrow 0} \frac{1}{h} \mathrm{E}|\Delta_{t+h,t}|^3 = 0. \tag{28.69}$$

As in **28.22** let $\Delta(s,t) = X(s) - X(t)$ and for some finite m let $\zeta_j = \Delta\big(t + hj/m, t + h(j-1)/m\big)$ for $j = 1, \dots, m$. By condition (c) the ζ_j are independent r.v.s with variances of h/m. If $S_j = \sum_{i=1}^{j} \zeta_i = \Delta(t + jh/m, t)$, then

$$\mathrm{E}\big((S_j - S_i)^2(S_k - S_j)^2\big) = (j-i)(k-j)h^2/m^2 \le h^2. \tag{28.70}$$

By **28.24**, setting $b_j = h/m$,

$$P\Big(\max_{0\le j\le m} \min\Big\{\Big|\Delta(t + \frac{j}{m}h, t)\Big|, \Big|\Delta(t + h, t + \frac{j}{m}h)\Big|\Big\} \ge \alpha\Big) \le \frac{Kh^2}{\alpha^4}. \tag{28.71}$$

Hence, by **28.23** and subadditivity,

$$\begin{aligned}P(|\Delta(t+h,t)|\ge\alpha)&\le P\Big(2\max_{0\le j\le m}\min\Big\{\Big|\Delta(t+\frac{j}{m}h,t)\Big|,\Big|\Delta(t+h,t+\frac{j}{m}h)\Big|\Big\}\\&\qquad+\max_{0\le j\le m}\Big|\Delta(t+\frac{j}{m}h,t+\frac{j-1}{m}h)\Big|\ge\alpha\Big)\\&\le\frac{K^*h^2}{\alpha^4}+P\Big(\max_{0\le j\le m}\Big|\Delta(t+\frac{j}{m}h,t+\frac{j-1}{m}h)\Big|\ge\tfrac{1}{2}\alpha\Big)\end{aligned}\tag{28.72}$$

where $K^*=4^4K$. Letting $m\to\infty$, the second term of the majorant member must go to zero by condition (b), so (switching to the subscript notation)

$$P\big(|\Delta_{t+h,t}|\ge\alpha\big)\le\frac{K^*h^2}{\alpha^4}.\tag{28.73}$$

Now use **9.23** to give

$$\begin{aligned}\mathrm{E}|\Delta_{t+h,t}|^3&=\int_0^\varepsilon|\Delta_{t+h,t}|^3\mathrm{d}F+\varepsilon P\big(|\Delta_{t+h,t}|^3\ge\varepsilon\big)+\int_\varepsilon^\infty P\big(|\Delta_{t+h,t}|^3>\zeta\big)\mathrm{d}\zeta\\&\le2\varepsilon+\int_\varepsilon^\infty P\big(|\Delta_{t+h,t}|\ge\zeta^{1/3}\big)\mathrm{d}\zeta\\&\le2\varepsilon+K^*h^2\int_\varepsilon^\infty\frac{1}{\zeta^{4/3}}\mathrm{d}\zeta=2\varepsilon+\frac{3K^*h^2}{\varepsilon^{1/3}}.\end{aligned}\tag{28.74}$$

Choosing $\varepsilon=(K^*/2)^{3/4}h^{3/2}$ to minimize the last member above yields

$$\mathrm{E}|\Delta_{t+h,t}|^3\le3K^{*3/4}h^{3/2}.\tag{28.75}$$

This condition verifies (28.69) and completes the proof. ∎

Notice how (28.73) is a substantial strengthening of the Chebyshev inequality, which gives merely $P\big(|\Delta_{t+h,t}|\ge\alpha\big)\le h/\alpha^2$. The existence of the third moment has not been assumed at the outset; this emerges, along with the Gaussianity, from the assumption of independent increments of arbitrarily small width, which allows (28.72) to be taken to the limit.

29
Weak Convergence

As with Chapter 27, this chapter contains a general discussion of weak convergence (in §29.1–§29.4) and then in the remaining sections focuses on the leading case of partial sum processes and the functional CLT. Skipping direct to this latter material may be the best strategy on the first pass. That said, to make sense of the tightness issue in particular a general perspective is recommended.

29.1 Weak Convergence in Metric Spaces

Consider $\mathbb{M}$, the set of all probability measures defined on $((\mathbb{S},d),\mathcal{S})$. As a matter of fact, the results of this chapter hold for all finite measures and there are a couple of cases in the sequel where they are applied to measures μ where $\int \mathrm{d}\mu \neq 1$. However, the modifications required for the extension are trivial. It is helpful in the proofs to have an agreed normalization and $\int \mathrm{d}\mu = 1$ is as good as any, so let $\mathbb{M}$ be the p.m.s, while keeping the possibility of generalization in mind.

Weak convergence concerns the properties of sequences in $\mathbb{M}$ and it is mathematically convenient to approach this problem by treating $\mathbb{M}$ as a topological space. The natural way of doing this is to define a collection of real-valued functions on $\mathbb{M}$ and adopt the weak topology that they induce. In view of (27.7), a natural class to consider are the integrals of bounded, continuous, real-valued functions with respect to the elements of $\mathbb{M}$.

For a point $\mu \in \mathbb{M}$ define the base sets

$$V_\mu(k,f_1,\dots,f_k,\varepsilon) = \left\{\nu : \nu \in \mathbb{M}, \left|\int f_i \mathrm{d}\nu - \int f_i \mathrm{d}\mu\right| < \varepsilon,\ i = 1,\dots,k\right\} \qquad (29.1)$$

where $f_i \in U_{\mathbb{S}}$ for each i and $\varepsilon > 0$. By ranging over all the possible $f_1,\dots,f_k$ and ε, for each $k \in \mathbb{N}$, (29.1) defines a collection of open neighbourhoods of μ. The base collection $V_\mu(k,f_1,\dots,f_k,\varepsilon)$ for $k \in \mathbb{N}$ and $\mu \in \mathbb{M}$ defines the weak topology on $\mathbb{M}$.

The idea is that two measures are close to one another when the expectations of various elements of $U_{\mathbb{S}}$ are close to one another. The more functions this applies to and the closer they are, the closer are the measures. This is not the consequence of some more fundamental notion of closeness, but is the defining property itself. This simple yet remarkable application illustrates the power of the topological

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0029

ideas developed in Chapter 6. The weak topology is the basic trick that allows distributions on general metric spaces to be handled by a single theory.

Given a concept of closeness there is immediately a companion concept of convergence. A sequence of measures $\{\mu_n, n \in \mathbb{N}\}$ is said to converge to a limit μ in the weak topology on $\mathbb{M}$ if for every neighbourhood V_μ, $\exists N$ such that $\mu_n \in V_\mu$ for all $n \geq N$. This is called *weak convergence* and is often written as $\mu_n \Rightarrow \mu$. If x_n is a random element from a probability space $(\mathbb{S}, \mathcal{S}, \mu_n)$ and $\mu_n \Rightarrow \mu$, x_n may be said to converge in distribution to x, written $x_n \to_{\mathrm{d}} x$ where x represents a random element from $(\mathbb{S}, \mathcal{S}, \mu)$. Essentially, the same caveats noted in §23.1 apply in the use of this terminology.

The following theorem shows that there are several ways to characterize weak convergence.

29.1 Theorem The following conditions are equivalent to one another.

(a) $\mu_n \Rightarrow \mu$.

(b) $\int f \mathrm{d}\mu_n \to \int f \mathrm{d}\mu$ for every $f \in U_{\mathbb{S}}$.

(c) $\limsup_n \mu_n(C) \leq \mu(C)$ for every closed set $C \in \mathcal{S}$.

(d) $\liminf_n \mu_n(B) \geq \mu(B)$ for every open set $B \in \mathcal{S}$.

(e) $\lim_n \mu_n(A) = \mu(A)$ for every $A \in \mathcal{S}$ for which $\mu(\partial A) = 0$. □

The equivalences of (a) and (b) and of (a) and (e) were proved for the case of measures on the line as **23.8** and **23.1** respectively; in that case weak convergence was identified with the convergence of the sequence of c.d.f.s, but this characterization has no counterpart here. A noteworthy consequence of the theorem is the fact that the sets (29.1) are not the only way to generate the topology of weak convergence. The alternative corresponding to part (e) of the theorem, for example, is the system of neighbourhoods

$$V'_\mu(k, A_1, \ldots, A_k, \varepsilon) = \big\{\nu : \nu \in \mathbb{M}, \ |\nu(A_i) - \mu(A_i)| < \varepsilon, \ i = 1, \ldots, k\big\} \tag{29.2}$$

where $A_i \in \mathcal{S}$, $i = 1, \ldots, k$ and $\mu(\partial A_i) = 0$.

Proof of 29.1 This theorem is proved by showing the circular set of implications, (a) ⇒ (b) ⇒ (c) ⇒ (c), (d) ⇒ (e) ⇒ (a). The first is by definition. The device in the proof of **27.8** can be used to show that (b) ⇒ (c); let B be any closed set in $\mathcal{S}$ and put $B_m = \{x : d(x, B) < 1/m\}$ so that B and B_m^c are closed and $\inf_{x \in B_m^c, y \in B} d(x, y) \geq 1/m$. Letting $g_{B_m^c, B} \in U_{\mathbb{S}}$ be the separating function defined above (27.9),

$$\limsup_{n \to \infty} \mu_n(B) \leq \limsup_{n \to \infty} \int g_{B_m^c, B} \mathrm{d}\mu_n = \int g_{B_m^c, B} \mathrm{d}\mu = \int_{B_m} g_{B_m^c, B} \mathrm{d}\mu \leq \mu(B_m) \tag{29.3}$$

where the first equality is by (b). (c) now follows on letting $m \to \infty$. (c) $\Rightarrow$ (d) is immediate since every closed set is the complement of an open set relative to $\mathbb{S}$ and $\mu(\mathbb{S}) = 1$.

To show (c) and (d)$\Rightarrow$ (e): for any $A \in \mathcal{S}$, $A^o \subseteq A \subseteq \bar{A}$, where A^o is open and $\bar{A}$ is closed and $\partial A = \bar{A} - A^o$. From (c),

$$\limsup_{n\to\infty} \mu_n(A) \le \limsup_{n\to\infty} \mu_n(\bar{A}) \le \mu(\bar{A}) = \mu(A) \tag{29.4}$$

and from (d),

$$\liminf_{n\to\infty} \mu_n(A) \ge \liminf_{n\to\infty} \mu_n(A^o) \ge \mu(A^o) = \mu(A), \tag{29.5}$$

hence $\lim_n \mu_n(A) = \mu(A)$.

The one relatively tricky step is to show (e) $\Rightarrow$ (a), which means in effect (e) $\Rightarrow$ (b). Let $f \in U_{\mathbb{S}}$ and define (what is easily verified to be) a measure μ^f on the real line $(\mathbb{R}, \mathcal{B})$ by

$$\mu^f(B) = \mu\big(\{x : f(x) \in B\}\big),\ B \in \mathcal{B}. \tag{29.6}$$

f is bounded so there exists an interval (a, b) such that $a < f(x) < b$, all $x \in \mathbb{S}$. Recall that a distribution on $(\mathbb{R}, \mathcal{B})$ has at most a countable number of atoms. Also, a finite interval can be divided into a finite collection of disjoint subintervals of width not exceeding ε, for any $\varepsilon > 0$. Therefore it is possible to choose m points t_j, with $a = t_0 < t_1 < \ldots < t_m = b$, such that $t_j - t_{j-1} < \varepsilon$ and $\mu^f(\{t_j\}) = 0$, for each j. Use these to construct a simple r.v.

$$g_m(x) = \sum_{j=1}^{m} t_{j-1} 1_{A_j}(x) \tag{29.7}$$

where $A_j = \{x : t_{j-1} \le f(x) \le t_j\}$ and note that $\sup_x |f(x) - g_m(x)| < \varepsilon$. Thus,

$$\begin{aligned}\left|\int f\mathrm{d}\mu_n - \int f\mathrm{d}\mu\right| &\le \int |f - g_m|\mathrm{d}\mu_n + \int |f - g_m|\mathrm{d}\mu + \left|\int g_m\mathrm{d}\mu_n - \int g_m\mathrm{d}\mu\right| \\ &\le 2\varepsilon + \sum_{j=1}^{m} |t_{j-1}|\left|\mu_n(A_j) - \mu(A_j)\right|. \end{aligned} \tag{29.8}$$

Since $\mu(\partial A_j) = 0$ by the choice of t_j, so that $\lim_n \mu_n(A_j) = \mu(A_j)$ for each j by (e),

$$\limsup_{n\to\infty} \left|\int f\mathrm{d}\mu_n - \int f\mathrm{d}\mu\right| \le 2\varepsilon. \tag{29.9}$$

Since ε can be chosen arbitrarily small, (b) follows and the proof is complete. ∎

A *convergence-determining class* for $(\mathbb{S}, \mathcal{S})$ is a class of sets $\mathcal{U} \subseteq \mathcal{S}$ that satisfy the following condition for every $\mu \in \mathbb{M}$: if $\mu_n(A) \to \mu(A)$ for every $A \in \mathcal{U}$ with $\mu(\partial A) = 0$, then $\mu_n \Rightarrow \mu$. This notion may be helpful for establishing weak convergence in cases where the conditions of **29.1** are difficult to show directly. The following theorem illustrates this with an example.

29.2 Theorem Let $\mathcal{U}$ be a π-system such that every open set of $\mathcal{S}$ is a finite or countable union of $\mathcal{U}$-sets. If $\mu_n(A) \to \mu(A)$ for every $A \in \mathcal{U}$ then $\mu_n \Rightarrow \mu$.

Proof Consider a finite union of $\mathcal{U}$-sets $A_1, \dots, A_m$. Applying the inclusion–exclusion formula (see **3.6**),

$$\mu_n\left(\bigcup_{j=1}^{m} A_j\right) = \sum_{k=1}^{2^m-1} \pm\mu_n(C_k) \tag{29.10}$$

where the sets C_k consist of the A_j and all their mutual intersections and '±' indicates that the sign of the term is given in accordance with (3.4). Since $\mathcal{U}$ is a π-system these are all elements of $\mathcal{U}$ and hence by hypothesis,

$$\mu_n\left(\bigcup_{j=1}^{m} A_j\right) \to \mu\left(\bigcup_{j=1}^{m} A_j\right). \tag{29.11}$$

To extend to a countable union $B = \bigcup_{j=1}^{\infty} A_j$, note that continuity of μ implies $\mu(\bigcup_{j=1}^{m} A_j) \uparrow \mu(B)$ as $m \to \infty$. For any $\varepsilon > 0$ a finite m may be chosen large enough that $\mu(B) - \mu(\bigcup_{j=1}^{m} A_j) < \varepsilon$. Then

$$\liminf_{n\to\infty} \mu_n(B) \geq \liminf_{n\to\infty} \mu_n\left(\bigcup_{j=1}^{m} A_j\right) = \mu\left(\bigcup_{j=1}^{m} A_j\right) > \mu(B) - \varepsilon. \tag{29.12}$$

Since ε is arbitrary and (29.12) holds for any open $B \in \mathcal{S}$ by hypothesis on $\mathcal{U}$, condition (d) of **29.1** is satisfied. ∎

These conditions may be stronger than necessary, noting that condition $\mu(\partial A) = 0$ is not invoked in the requirement $\mu_n(A) \to \mu(A)$. More inclusive ways to specify $\mathcal{U}$ can be found when $\mathbb{S}$ is separable.

A convergence-determining class must also be a determining class for the space (see §3.2). However, caution is necessary since the converse does not hold, as the following counterexample given by Billingsley ([21], [22]) shows.

29.3 Example Consider the family of p.m.s $\{\mu_n\}$ on the half-open unit interval [0,1) with μ_n assigning unit measure to the singleton set $\{1 - 1/n\}$. That is, $\mu_n(\{1 - 1/n\}) = 1$. Evidently, $\{\mu_n\}$ does not have a weak limit. The collection $\mathcal{C}$ of half-open intervals $[a, b)$ for $0 < a < b < 1$ generates the Borel field of $[0, 1)$ and so is a determining class. However, even though $\mu_n \not\Rightarrow \mu$, $\mu_n([a, b)) \to 0$ for every fixed a and b. Since the p.m. μ for which $\mu(\{0\}) = 1$ has the property that $\mu([a, b)) = 0$ whenever $a > 0$, it is valid to write

$$\mu_n(A) \to \mu(A), \text{ all } A \in \mathcal{C}. \tag{29.13}$$

Hence, $\mathcal{C}$ is not convergence-determining. □

The last topic to be considered in this section is the preservation of weak convergence under mappings from one metric space to another. Since $\mu_n \Rightarrow \mu$ means $\int f\mathrm{d}\mu_n \to \int f\mathrm{d}\mu$ for any $f \in U_{\mathbb{S}}$, it is clear since $f \circ h \in U_{\mathbb{S}}$ when h is continuous that $\int f(h(x))\mathrm{d}\mu_n(x) \to \int f(h(x))\mathrm{d}\mu(x)$. Writing y for $h(x)$, this becomes

$$\int f(y)\mathrm{d}\mu_n h^{-1}(y) \to \int f(y)\mathrm{d}\mu h^{-1}(y). \tag{29.14}$$

The foregoing is direct and relatively trivial, but what can also be shown, often much more usefully, is that mappings that are 'almost' continuous have the same property. This is the continuous mapping theorem proper, the desired generalization of **23.11**.

29.4 Theorem Let $h : \mathbb{S} \mapsto \mathbb{T}$ be a measurable function and let $D_h \subseteq \mathbb{S}$ be the set of discontinuity points of h. If $\mu_n \Rightarrow \mu$ and $\mu(D_h) = 0$, then $\mu_n h^{-1} \Rightarrow \mu h^{-1}$.

Proof Let C be a closed subset of $\mathbb{T}$. Recalling that $(A)^-$ denotes the closure of A,

$$\begin{aligned}
\limsup_{n\to\infty} \mu_n h^{-1}(C) &= \limsup_{n\to\infty} \mu_n\big((h^{-1}(C)\big) \\
&\leq \limsup_{n\to\infty} \mu_n\big((h^{-1}(C))^-\big) \\
&\leq \mu\big((h^{-1}(C))^-\big) \\
&\leq \mu\big(h^{-1}(C) \cup D_h\big) \\
&\leq \mu\big(h^{-1}(C)\big) + \mu(D_h) = \mu\big(h^{-1}(C)\big) = \mu h^{-1}(C).
\end{aligned} \tag{29.15}$$

The second inequality is by **29.1**(c) and the third inequality uses the fact that $(h^{-1}(C))^- \subseteq h^{-1}(C) \cup D_h$. That is to say, since C is closed a closure point of $h^{-1}(C)$ is either in $h^{-1}(C)$ or is not a continuity point of h. ∎

29.2 Skorokhod's Representation

Considering a sequence of random elements, here is a generalization of some familiar ideas from the theory of random variables. Recall from **27.7** that separability ensures that the distance functions in the following definitions are r.v.s.

Let $\{x_n\}$ be a sequence of random elements and x a given random element of a separable space $(\mathbb{S}, \mathcal{S})$. The statement

$$d\big(x_n(\omega), x(\omega)\big) \to 0 \text{ for } \omega \in E, \text{ with } P(E) = 1 \tag{29.16}$$

says that x_n converges almost surely to x, written for brevity as $x_n \to_{\text{a.s.}} x$. Similarly,

$$P\big(d(x_n, x) \geq \varepsilon\big) \to 0, \text{ all } \varepsilon > 0 \tag{29.17}$$

says that x_n converges in probability to x, written $x_n \to_{\text{pr}} x$. Convergence a.s. is sufficient for convergence in probability, which in turn is sufficient for $x_n \to_{\text{d}} x$. A case subsumed in the above definition is where $x = a$ with probability 1, a being a fixed element of $\mathbb{S}$.

The following result is the generalization to metric spaces of **23.18**.

29.5 Theorem Given a probability space $(\Omega, \mathcal{F}, P)$, let $\{x_n(\omega)\}$ and $\{y_n(\omega)\}$ be random sequences on a separable space $(\mathbb{S}, \mathcal{S})$. If $x_n \to_{\text{d}} x$ and $d(x_n, y_n) \to_{\text{pr}} 0$, then $y_n \to_{\text{d}} x$.

Proof Let $A \in \mathcal{S}$ be a closed set and for $\varepsilon > 0$ put $A_\varepsilon = \{x : d(x, A) \leq \varepsilon\} \in \mathcal{S}$, also a closed set for each ε and $A_\varepsilon \downarrow A$ as $\varepsilon \downarrow 0$. From the inclusion

$$\{\omega : y_n(\omega) \in A\} \subseteq \{\omega : x_n(\omega) \in A_\varepsilon\} \cup \{\omega : d\big(x_n(\omega), y_n(\omega)\big) \geq \varepsilon\}$$

it follows that

$$P(y_n \in A) \leq P(x_n \in A_\varepsilon) + P\big(d(x_n, y_n) \geq \varepsilon\big) \tag{29.18}$$

and, also on letting $n \to \infty$, that

$$\limsup_{n\to\infty} P(y_n \in A) \leq \limsup_{n\to\infty} \mu_n(A_\varepsilon) \leq \mu(A_\varepsilon) \tag{29.19}$$

where μ_n is the measure associated with x_n, μ the measure associated with x, and the second inequality of (29.19) is by hypothesis on $\{x_n\}$ and **29.1**(c). Since this inequality holds for every $\varepsilon > 0$ it follows by continuity of the measure that

$$\limsup_{n\to\infty} P(y_n \in A) \leq \mu(A). \tag{29.20}$$

This is sufficient for the result by **29.1**. ∎

It was shown in §23.2 that the weak convergence of a sequence of distributions on the line implies the a.s. convergence of a sequence of random variables. That result is a special case of the following, the Skorokhod representation of weak convergence.

29.6 Theorem ([170]: 3.1) Let $\{\mu_n\}$ be a sequence of measures on the separable, complete metric space $(\mathbb{S}, \mathcal{S})$. There exists a sequence of $\mathcal{B}_{[0,1]}/\mathcal{S}$-measurable functions

$x_n : [0,1] \mapsto \mathbb{S}$

such that $\mu_n(A) = m(\{\omega : x_n(\omega) \in A\})$ for each $A \in \mathcal{S}$, where m is Lebesgue measure. If $\mu_n \Rightarrow \mu$, there exists a function $x(\omega)$ such that $\mu(A) = m(\{\omega : x(\omega) \in A\})$ for each $A \in \mathcal{S}$ and $d(x_n(\omega), x(\omega)) \to 0$ a.s. $[m]$ as $n \to \infty$.

Proof This is by construction of the functions $x_n(\omega)$. For some $k \in \mathbb{N}$ let $\{x_i^{(k)}, i \in \mathbb{N}\}$ denote a countable collection of points in $\mathbb{S}$ such that, for every $x \in \mathbb{S}$, $d(x, x_i^{(k)}) \leq 1/2^{k+1}$ for some i. Such sequences exist for every k by separability. Let $S(x_i^{(k)}, r_k)$ for $1/2^{k+1} < r_k < 1/2^k$ denote a system of spheres in $\mathbb{S}$ having the property $\mu(\partial S(x_i^{(k)}, r_k)) = 0$ for every i. An r_k satisfying this condition exists, since there can be at most a countable number of points r such that $\mu(\partial S(x_i^{(k)}, r)) > 0$ for one or more i; this fact follows from **7.4**.

For given k, the system $\{S(x_i^{(k)}, r_k), i \in \mathbb{N}\}$ covers $\mathbb{S}$ and accordingly the sets

$$D_i^k = S(x_i^{(k)}, r_k) - \bigcup_{j=1}^{i-1} S(x_j^{(k)}, r_k),\ i \in \mathbb{N} \tag{29.21}$$

form a partition of $\mathbb{S}$. By letting each of the k integers $i_1, \ldots, i_k$ range independently over $\mathbb{N}$, define the countable collection of Borel sets

$$S_{i_1,\ldots,i_k} = D_{i_1}^1 \cap D_{i_2}^2 \cap \ldots \cap D_{i_k}^k \in \mathcal{S}. \tag{29.22}$$

Each $S_{i_1,\ldots,i_k}$ is a subset of a sphere of radius $r_k < 1/2^k$ and $\mu(\partial S_{i_1,\ldots,i_k}) = 0$. By construction, any pair $S_{i_1,\ldots,i_k}$ and $S_{i'_1,\ldots,i'_k}$ are disjoint unless $i_k = i'_k$. Fixing $i_1, \ldots, i_{k-1}$,

$$\bigcup_{i_k=1}^{\infty} S_{i_1,\ldots,i_k} = S_{i_1,\ldots,i_{k-1}} \tag{29.23}$$

and in particular,

$$\bigcup_{i_1=1}^{\infty} S_{i_1} = \mathbb{S}. \tag{29.24}$$

That is to say, for any k the collection $\{S_{i_1,\ldots,i_k}\}$ forms a partition of $\mathbb{S}$, which gets finer as k increases. These sets are not all required to be nonempty.

Next, for any $n \in \mathbb{N}$ and $k \in \mathbb{N}$ define a partition of [0,1] into intervals $\Delta^{(n)}_{i_1,\ldots,i_k}$ where it is understood that $\Delta^{(n)}_{i_1,\ldots,i_k}$ lies to the left of $\Delta^{(n)}_{i'_1,\ldots,i'_k}$ if $i_j = i'_j$ for $j = 1,\ldots,r-1$ and $i_r < i'_r$ for some r, and the lengths of the segments equal the probabilities $\mu_n(S_{i_1,\ldots,i_k})$.

To define a measurable mapping from [0,1] to $\mathbb{S}$, choose an element $\bar{x}_{i_1,\ldots,i_k}$ from each nonempty $S_{i_1,\ldots,i_k}$ and for $\omega \in [0,1]$ put

$$x^k_n(\omega) = \bar{x}_{i_1,\ldots,i_k} \text{ if } \omega \in \Delta^{(n)}_{i_1,\ldots,i_k}. \tag{29.25}$$

Note that by construction $d(x^k_n(\omega), x^{k+m}_n(\omega)) \leq 1/2^k$ for $m \geq 1$ and taking $k = 1,2,\ldots$ defines a Cauchy sequence in $\mathbb{S}$ which is convergent since $\mathbb{S}$ is a complete space by assumption. Write $x_n(\omega) = \lim_{k\to\infty} x^k_n(\omega)$. To show that $x_n(\omega)$ is a random element with distribution defined by μ_n, it is sufficient to verify that for (at least) all $A \in \mathcal{S}$ such that $\mu_n(\partial A) = 0$,

$$\mu_n(A) = P(x_n \in A) = m\big(\{\omega : x_n(\omega) \in A\}\big). \tag{29.26}$$

Letting $A^{(k)}$ denote the union of all $S_{i_1,\ldots,i_k} \subseteq A$ and $A'^{(k)}$ the union of all $S_{i_1,\ldots,i_k} \nsubseteq A^c$, it is clear that $A^{(k)} \subseteq A \subseteq A'^{(k)}$ and that (29.26) holds in respect of $A^{(k)}$ and $A'^{(k)}$. Let

$$E^{(k)} = \{x : d(x,\partial A) \leq 1/2^k\} \tag{29.27}$$

so that $A'^{(k)} - A^{(k)} \subseteq E^{(k)}$. Since $\mu_n(E^{(k)}) \to \mu_n(\partial A) = 0$ as $k \to \infty$, it follows that $\mu_n(A'^{(k)} - A^{(k)}) \to 0$ and hence $\mu_n(A^{(k)}) \to \mu_n(A)$. This proves (29.26).

It remains to be shown that if $\mu_n \Rightarrow \mu$ then $x_n \to x$ a.s. $[m]$. Since the length of $\Delta^{(n)}_{i_1,\ldots,i_k}$ equals $\mu_n(S_{i_1,\ldots,i_k})$, the sequence of intervals $\{\Delta^{(n)}_{i_1,\ldots,i_k}\}$ has a limit $\Delta_{i_1,\ldots,i_k}$ as $n \to \infty$. Pick an interior point ω of $\Delta_{i_1,\ldots,i_k}$ and note that x^k meets the condition $x^k(\omega) \in S_{i_1,\ldots,i_k}$ by definition. Then for N large enough, $\omega \in \Delta^{(n)}_{i_1,\ldots,i_k}$ for $n \geq N$ and hence $d(x^k_n(\omega), x^k(\omega)) \leq 1/2^{k-1}$. Letting $k \to \infty$, it follows that $d(x_n(\omega), x(\omega)) \leq \varepsilon$ for any $\varepsilon > 0$ whenever n is large enough. It is not possible to say this about the boundary points of the $\Delta_{i_1,\ldots,i_k}$ but these are at most countable even as $k \to \infty$ and have Lebesgue measure 0. This completes the proof. ∎

The essence of this proof is to set up a mapping between sets of the abstract space $\mathbb{S}$ and intervals of $[0,1]$ measuring their probabilities. Given any partition of $\mathbb{S}$ by $\mathcal{S}$-sets it is possible to assign an interval to each one that collectively partition $[0,1]$. The way the intervals are arranged is arbitrary and unrelated to any ordering that may exist on $\mathbb{S}$, but the insight of the theorem is that this doesn't matter. The construction of §23.2 where $\mathbb{S} = \mathbb{R}$ is revealed as the particularly elegant special case where the mapping from [0,1] to $\mathbb{S}$ is the inverse of the c.d.f.

In his 1956 paper, Skorokhod goes on to use this theorem to prove convergence results in spaces such as $D_{[0,1]}$. While the present development uses a different approach, this is a useful trick with a variety of potential applications just as in the case of $\mathbb{R}$. One of these will be encountered in §32.3.

29.3 Metrizing the Space of Measures

The problem of determining the weak limit of a sequence of measures $\{\mu_n\}$ on $(\mathbb{S}, \mathcal{S})$ falls into two parts. One of these is to determine the limits of the sequences $\{\mu_n(A)\}$ for each $A \in \mathcal{A}$, where $\mathcal{A}$ is a determining class for the space. This part of the programme is specific to the particular space under consideration. The other part, which is quite general, is to verify conditions under which the sequence of measures as a whole has a weak limit. Without this reassurance, the convergence of measures of elements of $\mathcal{A}$ is not generally sufficient to ensure that the extensions to $\mathcal{S}$ also converge. It is this second aspect of the problem that is the focus here.

It is sufficient if every sequence of measures on the space is shown to have a cluster point. If a subsequence converges to a limit, this must agree with the unique ordinary limit established (by assumption) for the determining class. The goal can be achieved by finding conditions under which the relevant topological space of measures is sequentially compact (see §6.2). This is similar to what Billingsley ([21], [22]) calls 'relative' compactness and the required results can be derived in his framework. The alternative approach to be adopted here follows Prokhorov ([153]) and Parthasarathy ([141]) in making $\mathbb{M}$ a metric space which will under appropriate circumstances be compact. The following theorem shows that this project is feasible; the basic idea is an application of the embedding theorem (**6.20**/**6.22**).

29.7 Theorem ([141]: th. II.6.2) Iff $(\mathbb{S}, d)$ is separable, $\mathbb{M}$ can be metrized as a separable space and embedded in $[0,1]^\infty$.

Proof Assume $(\mathbb{S}, d)$ is separable. The first task is to show that $U_{\mathbb{S}}$ is also separable. $\mathbb{S}$ can be metrized as a totally bounded space $(\mathbb{S}, d')$ where d' is equivalent to d; see for example the proof of **6.22**. Let $\bar{\mathbb{S}}$ denote the completion of $\mathbb{S}$ under d'

(including the limits of all Cauchy sequences in $\mathbb{S}$) and then $\bar{\mathbb{S}}$ is a compact space (**5.12**). The space of continuous functions $C_{\bar{\mathbb{S}}}$ is accordingly separable under the uniform metric (**5.26**(ii)). The spaces $C_{\bar{\mathbb{S}}}$ and $U_{\mathbb{S}}$ are isometric (see §5.6) and if the former is separable so is the latter. In fact, $U_{\mathbb{S}} = C_{\mathbb{S}}$ when $\bar{\mathbb{S}}$ is compact by **5.21**.

Let $\{g_m, m \in \mathbb{N}\}$ be a dense subset of $U_{\mathbb{S}}$ and define the mapping $T : \mathbb{M} \mapsto \mathbb{R}^\infty$ by

$$T(\mu) = (\textstyle\int g_1 \mathrm{d}\mu, \int g_2 \mathrm{d}\mu, \ldots). \tag{29.28}$$

The object is to show that T embeds $\mathbb{M}$ in $\mathbb{R}^\infty$. Suppose $T(\mu) = T(\nu)$, so that $\int g_m \mathrm{d}\mu = \int g_m \mathrm{d}\nu$ for all m. Since $\{g_m\}$ is dense in U_S, $f \in U_{\mathbb{S}}$ implies that

$$\left|\int f \mathrm{d}\mu - \int g_m \mathrm{d}\mu\right| \leq \int |f - g_m| \mathrm{d}\mu \leq d_U(f, g_m) < \varepsilon \tag{29.29}$$

for some m and every $\varepsilon > 0$. (The second inequality is because $\int \mathrm{d}\mu = 1$, note.) The same inequalities hold for ν and hence $\int f \mathrm{d}\mu = \int f \mathrm{d}\nu$ for all $f \in U_{\mathbb{S}}$. It follows by **27.8** that $\mu = \nu$, so T is 1–1.

Continuity of T follows from the equivalence of (a) and (b) in **29.1**. To show that T^{-1} is continuous, let $\{\mu_n\}$ be a sequence of measures and assume $T(\mu_n) \to T(\mu)$. For $f \in U_{\mathbb{S}}$ and any $m \geq 1$,

$$\begin{aligned}\left|\int f \mathrm{d}\mu_n - \int f \mathrm{d}\mu\right| &= \left|\int (f - g_m) \mathrm{d}\mu_n + \int (g_m - f) \mathrm{d}\mu + \int g_m \mathrm{d}\mu_n - \int g_m \mathrm{d}\mu\right| \\ &\leq 2d_U(f, g_m) + \left|\int g_m \mathrm{d}\mu_n - \int g_m \mathrm{d}\mu\right|.\end{aligned} \tag{29.30}$$

Since the second term of the majorant side converges to zero by assumption,

$$\limsup_{n\to\infty}\left|\int f \mathrm{d}\mu_n - \int f \mathrm{d}\mu\right| \leq 2d_U(f, g_m) < 2\varepsilon \tag{29.31}$$

for some m and $\varepsilon > 0$, by the right-hand inequality of (29.29). Hence $\lim_n |\int f \mathrm{d}\mu_n - \int f \mathrm{d}\mu| = 0$ and $\mu_n \Rightarrow \mu$ by **29.1**(b).

This shows that $\mathbb{M}$ is homeomorphic with the set $T(\mathbb{M}) \subseteq \mathbb{R}^\infty$ and $\mathbb{R}^\infty$ is homeomorphic to $[0,1]^\infty$ as noted in **5.22**. The distances d_∞ between the images of points of $\mathbb{M}$ under T define a metric on $\mathbb{M}$ that induces the weak topology. The space $T(\mathbb{M})$ with the product topology is separable (see **6.16**) so applying **6.9**(i) to T^{-1} yields the result that $\mathbb{M}$ is separable. This completes the sufficiency part of the proof.

The necessity part requires a lemma that will be needed again later on. Let $p_x \in \mathbb{M}$ be the degenerate p.m. with unit mass at x, that is, $p_x(\{x\}) = 1$ and $p_x(\mathbb{S} - \{x\}) = 0$ and so let $D = \{p_x : x \in \mathbb{S}\} \subseteq \mathbb{M}$.

29.8 Lemma The topological spaces $\mathbb{S}$ and D are homeomorphic.

Proof The mapping $p : \mathbb{S} \mapsto D$ taking points $x \in \mathbb{S}$ to points $p_x \in D$ is clearly 1–1, onto. For $f \in C_{\mathbb{S}}$, $\int f\mathrm{d}p_x = f(x)$ and $x_n \to x$ implies $f(x_n) \to f(x)$ and hence $p_{x_n} \Rightarrow p_x$ by **29.1**, establishing continuity of p. Conversely, suppose $x_n \nrightarrow x$. There is then an open set A containing x such that for every $N \in \mathbb{N}$, $x_n \in \mathbb{S} - A$ for some $n \geq N$. Let f be a separating function such that $f(x) = 0, f(y) = 1$ for $y \in \mathbb{S} - A$ and $0 \leq f \leq 1$. Then $\int f\mathrm{d}p_{x_n} = f(x_n) = 1$ and $\int f\mathrm{d}p_x = 0$, so $p_{x_n} \not\Rightarrow p_x$. This establishes continuity of p^{-1} and p is a homeomorphism, as required. ∎

Proof of 29.7, continued. Now suppose $\mathbb{M}$ is a separable metric space. It can be embedded in a subset of $[0,1]^\infty$ and the subsets of $\mathbb{M}$ are homeomorphic to their images in $[0,1]^\infty$ under the embedding, which are separable sets and hence are themselves separable (again, by **6.16** and **6.9**(i)). Since $D \subseteq \mathbb{M}$, D is separable and hence $\mathbb{S}$ must be separable since it is homeomorphic to D by **29.8**. This proves necessity. ∎

The last theorem showed that $\mathbb{M}$ is metrizable, but did not exhibit a specific metric on $\mathbb{M}$. Note that different collections of functions $\{g_m\}$ yield different metrics, given how d_∞ is defined. Another approach to the problem is to construct such a metric directly and one such was proposed by Prokhorov ([153]). For a set $A \in \mathcal{S}$, define the open set $A^\delta = \{x : d(x,A) < \delta\}$, that is, '$A$ with a δ-halo'. The *Prokhorov distance* between measures $\mu,\nu \in \mathbb{M}$ is

$$L(\mu,\nu) = \inf\{\delta > 0 : \mu(A^\delta) + \delta \geq \nu(A), \text{all } A \in \mathcal{S}\}. \tag{29.32}$$

Since $\mathcal{S}$ contains complements and $\mu(\mathbb{S}) = \nu(\mathbb{S}) = 1$, it must be the case unless $\mu = \nu$ that $\mu(A) \geq \nu(A)$ for some sets $A \in \mathcal{S}$ and $\mu(A) < \nu(A)$ for others. The idea of the Prokhorov distance is to focus on the latter cases and see how much has to be added to both the sets *and* their μ-measures, to reverse all the inequalities. When the measures are close this amount should be small, but it is worth taking some time to convince oneself that both the adjustments are necessary to get the desired properties. As shown below, L is a metric and hence is symmetric in μ and ν.

The properties are most easily appreciated in the case of measures on the real line, in which case the metric has the representation in terms of the c.d.f.s,

$$L^*(F_1,F_2) = \inf\{\delta > 0 : F_2(x-\delta) - \delta \leq F_1(x) \leq F_2(x+\delta) + \delta, \forall x \in \mathbb{R}\} \tag{29.33}$$

for c.d.f.s F_1 and F_2. This is also known as *Lévy's metric.* Figure 29.1 sketches this case and F_2 has been given a discontinuity, so that the form of the bounding

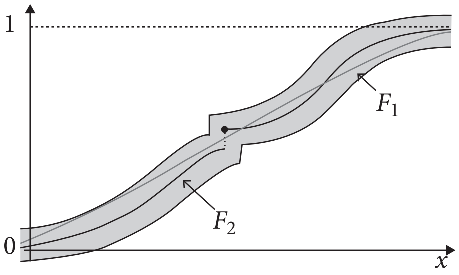

Figure 29.1

functions $F_2(x+\delta)+\delta$ and $F_2(x-\delta)-\delta$ can be easily discerned. Any c.d.f. lying wholly within the region defined by these extremes, such as the one shown, is within δ of F_2 in the L^* metric.

29.9 Theorem L is a metric.

Proof $L(\mu,\nu) = L(\nu,\mu)$ is not obvious from the definition; but for any $\delta > 0$ consider $B = (A^\delta)^c$. If $x \in A$, then $d(x,y) \geq \delta$ for each $y \in B$, whereas if $x \in B^\delta$, $d(x,y) < \delta$ for some $y \in B$; in other words, $B^\delta = A^c$. If $L(\mu,\nu) \leq \delta$, then

$$\mu(A^c) + \delta = \mu(B^\delta) + \delta \geq \nu(B). \tag{29.34}$$

Subtracting both sides of (29.34) from 1 gives

$$\mu(A) \leq \nu(B^c) + \delta = \nu(A^\delta) + \delta \tag{29.35}$$

and hence $L(\nu,\mu) \leq \delta$. This means there is no δ for which $L(\mu,\nu) > \delta \geq L(\nu,\mu)$, nor, by symmetry, for which $L(\nu,\mu) > \delta \geq L(\mu,\nu)$ and equality follows.

It is immediate that $L(\mu,\nu) = 0$ if $\mu = \nu$. To show the converse holds, note that if $L(\mu,\nu) = 0$, $\mu(A^{1/n}) + 1/n \geq \nu(A)$ for $A \in \mathcal{S}$ and any $n \in \mathbb{N}$. If A is closed, $A^{1/n} \downarrow A$ as $n \to \infty$. By continuity of μ, $\mu(A) = \lim_n(\mu(A^{1/n}) + 1/n) \geq \nu(A)$ and by symmetry, $\nu(A) = \lim_n(\nu(A^{1/n}) + 1/n) \geq \mu(A)$ likewise. It follows that $\mu(A) = \nu(A)$ for all closed A. Since the closed sets are a determining class, $\mu = \nu$.

Finally, for measures μ, ν, and τ let $L(\mu,\nu) = \delta$ and $L(\nu,\tau) = \eta$. Then for any $A \in \mathcal{S}$,

$$\mu(A) \leq \nu(A^\delta) + \delta \leq \tau\big((A^\delta)^\eta\big) + \delta + \eta \leq \tau(A^{\delta+\eta}) + \delta + \eta \tag{29.36}$$

where the last inequality holds because

$$(A^\delta)^\eta = \{x : d(x, A^\delta) < \eta\} \subseteq \{x : d(x,A) < \delta + \eta\} = A^{\delta+\eta}. \tag{29.37}$$

The inclusion is valid since d satisfies the triangle inequality. Hence $L(\mu,\tau) \le \delta + \eta = L(\mu,\nu) + L(\nu,\tau)$. ∎

L induces the topology of weak convergence.

29.10 Theorem If $\{\mu_n\}$ is a sequence of measures in $\mathbb{M}$, $\mu_n \Rightarrow \mu$ iff $L(\mu_n,\mu) \to 0$.

Proof To show 'if', suppose $L(\mu_n,\mu) \to 0$. For each closed set $A \in \mathcal{S}$,

$$\limsup_{n\to\infty} \mu_n(A) \le \mu(A^\delta) + \delta$$

for every $\delta > 0$ and hence, letting $\delta \downarrow 0$, $\limsup_n \mu_n(A) \le \mu(A)$ by continuity. Weak convergence follows by (c) of **29.1**. To show 'only if' consider for $A \in \mathcal{S}$ and fixed δ the bounded function

$$f_A(x) = \max\left\{0, 1 - \frac{d(x,A)}{\delta}\right\}. \tag{29.38}$$

Note that $f_A(x) = 1$ for $x \in A$, $0 < f_A(x) \le 1$ for $x \in A^\delta$, and $f_A(x) = 0$ for $x \notin A^\delta$. Since

$$|f_A(x) - f_A(y)| \le \frac{|d(x,A) - d(y,A)|}{\delta} \le \frac{d(x,y)}{\delta} \tag{29.39}$$

independent of A, the family $\{f_A, A \in \mathcal{S}\}$ is uniformly equicontinuous (see §5.6) and so is a subset of $U_{\mathbb{S}}$. If $\mu_n \Rightarrow \mu$, then by **29.1**(b),

$$\Delta_n = \sup_{A\in\mathcal{S}} \left| \int f_A \mathrm{d}\mu_n - \int f_A \mathrm{d}\mu \right| \to 0. \tag{29.40}$$

Hence, n can be chosen large enough that $\Delta_n \le \delta$, for any $\delta > 0$. For this n or larger,

$$\mu_n(A) \le \int f_A \mathrm{d}\mu_n \le \int f_A \mathrm{d}\mu + \Delta_n \le \int f_A \mathrm{d}\mu + \delta \le \mu(A^\delta) + \delta \tag{29.41}$$

or, equivalently, $L(\mu_n,\mu) \le \delta$. It follows that $L(\mu_n,\mu) \to 0$. ∎

It is possible to establish the theory of convergence in $\mathbb{M}$ by working explicitly in the metric space $(\mathbb{M}, L)$. By contrast, the approach of Varadarajan ([183]) is to work in the equivalent space derived in **29.7** and this is the one taken here. The treatment in this section and the following one draws principally on Parthasarathy ([141]). The Prokhorov metric has an application in a different context, in §30.5.

The next theorem leads on from **29.7** by answering the crucial question: when is $\mathbb{M}$ compact?

29.11 Theorem ([141]: th. II.6.4) $\mathbb{M}$ is compact iff $\mathbb{S}$ is compact.

Proof First, let $\mathbb{S}$ be compact and recall that in this case $C_{\mathbb{S}} = U_{\mathbb{S}}$ (**5.21**) and $C_{\mathbb{S}}$ is separable (**5.26**(ii)). For simplicity of notation write just C for $C_{\mathbb{S}}$ and write 0 for that element of C which takes the value 0 everywhere in $\mathbb{S}$. Let $\bar{S}_C(0,1)$ denote the closed unit sphere around 0 in C, such that $\sup_t |f(t)| \leq 1$ for all $f \in \bar{S}_C(0,1)$, and let $\{g_m, m \in \mathbb{N}\}$ be a sequence that is dense in $\bar{S}_C(0,1)$. For this sequence of functions, the map T defined in (29.28) is a homeomorphism taking $\mathbb{M}$ into $T(\mathbb{M})$, a subset of the compact space $[-1,1]^\infty$. This follows by the argument used in **29.7**. It must be shown that $T(\mathbb{M})$ is closed and therefore compact. Let $\{\mu_n\}$ be a sequence of measures in $\mathbb{M}$ such that $T(\mu_n) \to y \in [-1,1]^\infty$. To prove sufficiency, the next step is to show that $y \in T(\mathbb{M})$. Since the mapping T^{-1} onto $\mathbb{M}$ is continuous, this would imply by **6.9**(ii) that $\mathbb{M}$ itself is compact.

Write $\Lambda_n(f) = \int f \mathrm{d}\mu_n$ and note, since $|\int f \mathrm{d}\mu_n| \leq \sup_t |f(t)| \leq 1$, that this defines a functional

$$\Lambda_n(f) : \bar{S}_C(0,1) \mapsto [-1,1]. \tag{29.42}$$

In this notation $T(\mu_n) = (\Lambda_n(g_1), \Lambda_n(g_2), \ldots)$. Since $\bar{S}_C(0,1)$ is compact and $\{g_m\}$ is dense in it, choose for every $f \in \bar{S}_C(0,1)$ a subsequence $\{g_{m_k}, k \in \mathbb{N}\}$ converging to f. Then for some k and for n, $n' \in \mathbb{N}$,

$$|\Lambda_n(f) - \Lambda_{n'}(f)| \leq 2d_U(f, g_{m_k}) + |\Lambda_n(g_{m_k}) - \Lambda_{n'}(g_{m_k})| \tag{29.43}$$

as in (29.30). The second term of the majorant side contains a coordinate of $T(\mu_n) - T(\mu_{n'})$ and converges to 0 as n and $n' \to \infty$ by assumption. Letting $k \to \infty$,

$$\lim_{n,n' \to \infty} |\Lambda_n(f) - \Lambda_{n'}(f)| = 0 \tag{29.44}$$

similarly to (29.31). This says that $\{\Lambda_n\}$ is a Cauchy sequence of real functionals on $[-1,1]$ and so must have a limit Λ; in particular, $y = (\Lambda(g_1), \Lambda(g_2), \ldots)$.

It is easy to verify that each $\Lambda_n(f)$ and hence also $\Lambda(f)$ satisfies conditions (27.10)–(27.12) for $f \in \bar{S}_C(0,1)$. Since for every $f \in C$ there is a constant $c > 0$ such that $cf \in \bar{S}_C(0,1)$, it follows from (27.12) that $\Lambda(f) = c\Lambda^*(f/c)$ where $\Lambda^*(\cdot)$ is a functional on C that must also satisfy (27.10)–(27.12). From **27.9** there exists a unique $\mu \in \mathbb{M}$ such that $\Lambda^*(f) = \int f \mathrm{d}\mu$, $f \in C$. Hence, $y = T(\mu)$. It follows that $T(\mathbb{M})$ contains its limit points and being also bounded is compact; and since T^{-1} is a homeomorphism, $\mathbb{M}$ is also compact. This completes the proof of sufficiency.

To prove necessity, consider $D = \{p_x : x \in \mathbb{S}\} \subseteq \mathbb{M}$, the set shown to be homeomorphic to $\mathbb{S}$ in **29.8**. If D is compact, then so is $\mathbb{S}$. D is totally bounded when $\mathbb{M}$ is compact, so by **5.12** it suffices to show completeness. Every sequence in D is the image of a sequence $\{x_n \in \mathbb{S}\}$ and can be written as $\{p_{x_n}\}$, so suppose $p_{x_n} \Rightarrow q \in \mathbb{M}$. If $x_n \to x \in \mathbb{S}$, then $q = p_x \in D$ by **29.8**, so it suffices to show that $x_n \nrightarrow x$ is impossible.

The possibility that $\{x_n\}$ has two or more distinct cluster points in $\mathbb{S}$ is ruled out by the assumption $p_{x_n} \Rightarrow q$, so $x_n \nrightarrow x$ means that the sequence has no cluster points in $\mathbb{S}$. Assuming this leads to a contradiction. Let $E = \{x_1, x_2, \dots\} \subseteq \mathbb{S}$ be the set of the sequence coordinates and let E_1 be any infinite subset of E. If the sequence has no cluster points, every point $y \in E_1$ is isolated, in that $E_1 \cap S(y, \varepsilon) - \{y\}$ is empty for some $\varepsilon > 0$. Otherwise, there would have to exist a sequence $\{y_n \in E_1\}$ such that $y_n \in S(y, 1/n)$ for every n and, contrary to assumption, y would be a cluster point of $\{x_n\}$. A set containing only isolated points is closed, so E_1 is closed and by **29.1**(c),

$$q(E_1) \geq \limsup_{n\to\infty} p_{x_n}(E_1) = 1 \tag{29.45}$$

where the equality must obtain since E_1 contains x_n for some $n \geq N$, for every $N \in \mathbb{N}$. Since $q \in \mathbb{M}$, this has to mean $q(E_1) = 1$. But clearly another subset from E can be chosen, say E_2, such that E_1 and E_2 are disjoint and the same logic would give $q(E_2) = 1$, which is impossible. The contradiction is shown, concluding the proof. ∎

29.4 Tightness and Convergence

In §23.5 the idea of a tight probability measure was introduced as one whose mass is concentrated on a compact subset of the sample space. Formally, a measure μ on a space $(\mathbb{S}, \mathcal{S})$ is said to be tight if for every $\varepsilon > 0$ there exists a compact set $K_\varepsilon \in \mathcal{S}$ such that $\mu(K_\varepsilon^c) \leq \varepsilon$. Let $\Pi \subseteq \mathbb{M}$ denote any family of measures. The family Π is said to be *uniformly tight* if $\sup_{\mu\in\Pi} \mu(K_\varepsilon^c) \leq \varepsilon$.

Although the present focus is on p.m.s, tightness is a property of general measures. In the applications below Π typically represents the sequence of p.m.s associated with a stochastic sequence $\{X_n\}_1^\infty$. If a p.m. μ is tight then of course $\mu(K_\varepsilon) > 1 - \varepsilon$ for compact K_ε. In §23.5 uniform tightness of a sequence of p.m.s on the line was shown to be a necessary condition for weak convergence of the sequence and here the same result is obtained for any metric space that is separable and complete. The first result needed is the following.

29.12 Theorem ([141]: th. II.3.2) When $\mathbb{S}$ is separable and complete, every p.m. on the space is tight. □

Notice that this proves the earlier assertion that every measure on $(\mathbb{R}, \mathcal{B})$ is tight, given that $\mathbb{R}$ is a separable, complete space. Another lemma is needed for the proof and also subsequently.

29.13 Lemma Let $\mathbb{S}$ be a complete space and let

$$K = \bigcap_{n=1}^{\infty}\left(\bigcup_{i=1}^{j_n} \bar{S}_{ni}\right) \tag{29.46}$$

where S_{ni} is a sphere of radius $1/n$ in $\mathbb{S}$, $\bar{S}_{ni}$ is its closure, and j_n is a finite integer for each n. Then K is compact.

Proof Being covered by a finite collection of the $\bar{S}_{ni}$ for each n, K is totally bounded. If $\{x_j, j \in \mathbb{N}\}$ is a Cauchy sequence in K, completeness of $\mathbb{S}$ implies that $x_j \to x \in \mathbb{S}$. For each n, since $K \subseteq \bigcup_{i=1}^{j_n} \bar{S}_{ni}$, infinitely many of the sequence coordinates must lie in $K_n = K \cap \bar{S}_{nk}$ for some k, $1 \le k \le j_n$. Since $\bar{S}_{nk}$ has radius $1/n$, taking n to the limit leads to the conclusion that $\bigcap_n K_n = \{x\}$ and hence $x \in K$; K is therefore complete and the lemma follows by **5.12**. ∎

Proof of 29.12 By separability, a covering of $\mathbb{S}$ by $1/n$-balls $S_n = S(x, 1/n)$, $x \in \mathbb{S}$, has a countable subcover, say $\{S_{ni}, i \in \mathbb{N}\}$ for each $n = 1,2,\ldots$. Fix n. Letting $A_n = \bigcup_{i=1}^{j_n} S_{ni}$, for any $\varepsilon > 0$ there must exist j_n large enough that $\mu(A_n) \ge 1 - \varepsilon/2^n$, since otherwise $\mu(\bigcup_{i=1}^{\infty} S_{ni}) = \mu(\mathbb{S}) < 1 - \varepsilon/2^n$, which is a contradiction since μ is a p.m.

Given ε, choose A_n in this manner for each n and let $K_\varepsilon = \bigcap_{n=1}^{\infty} \bar{A}_n$ where $\bar{A}_n = \bigcup_{i=1}^{j_n} \bar{S}_{ni}$, note. Then K_ε is compact by **29.13**. Further, since

$$K_\varepsilon^c = \left(\bigcap_{n=1}^{\infty} \bar{A}_n\right)^c = \bigcup_{n=1}^{\infty} \bar{A}_n^c \tag{29.47}$$

and noting that $\mu(\bar{A}_n^c) = 1 - \mu(\bar{A}_n) \le 1 - \mu(A_n) \le \varepsilon/2^n$,

$$\mu(K_\varepsilon^c) \le \sum_{n=1}^{\infty} \mu(\bar{A}_n^c) \le \varepsilon \sum_{n=1}^{\infty} 1/2^n = \varepsilon \tag{29.48}$$

or, in other words, $\mu(K_\varepsilon) > 1 - \varepsilon$. ∎

Before moving on, note that the promised proof of **12.6** can be obtained as a corollary of **29.12**.

29.14 Corollary Let $(\mathbb{S},\mathcal{S},\mu)$ be a separable complete probability space. For any $E \in \mathcal{S}$ there is for any $\varepsilon > 0$ a compact subset K of E such that $\mu(E-K) < \varepsilon$.

Proof Let the compact set $\Delta \in \mathcal{S}$ satisfy $\mu(\Delta) > 1-\varepsilon/2$, as is possible by **29.12** and let $(\Delta,\mathcal{S}_\Delta,\mu_\Delta)$ denote the trace of $(\mathbb{S},\mathcal{S},\mu)$ on Δ. This is a compact space, such that every set in $\mathcal{S}_\Delta$ is totally bounded. By regularity of the measure (**27.1**) there exists for any $A \in \mathcal{S}_\Delta$ an open set $A' \in \mathcal{S}_\Delta$ such that $A' \supseteq A$ and $\mu_\Delta(A'-A) < \varepsilon/2$. Now move to complements with respect to Δ, defining $A^c = \Delta - A$. A'^c is a closed and hence compact set contained in A^c. Note that $A^c - A'^c = A' - A$ and hence $\mu(A^c - A'^c) = \mu_\Delta(A'-A)\mu(\Delta) < \varepsilon/2$.

Now for any set $E \in \mathcal{S}$ let $A = (E \cap \Delta)^c \in \mathcal{S}_\Delta$ and so let $K = A'^c$. This is a compact subset of $E \cap \Delta$, and hence of E, and by the foregoing argument $\mu(E \cap \Delta - K) < \varepsilon/2$. Since $E - K = (E \cap \Delta - K) \cup (E \cap \Delta^c - K)$ where the union is disjoint and $E \cap \Delta^c - K \subseteq \Delta^c$ where $\mu(\Delta^c) \le \varepsilon/2$, $\mu(E-K) < \varepsilon$ as required. ∎

Lemma **12.6** follows from this result on noting that $\mathbb{R}^k$ is a separable complete space.

Theorem **29.12** tells us that on a separable complete space, every member of a sequence of measures $\{\mu_n\}$ is tight. It remains to be established whether the same property applies to the weak limit of any such sequence. Here the reader should review examples **23.19** and **23.20** to appreciate how this need not be the case. The next theorem is a partial parallel of **23.22**, although the latter result goes further in giving sufficient conditions for a weak limit to exist. What is done here is merely to establish the possibility of weak convergence, via an application of theorems **5.10** and **5.11** by showing the link between uniform tightness and compactness.

29.15 Theorem ([141]: th. II.6.7) Let $(\mathbb{S},d)$ be a separable complete space and let $\Pi \subseteq \mathbb{M}$ be a family of p.m.s on $(\mathbb{S},\mathcal{S})$. Π is compact iff it is uniformly tight.

Proof Since $(\mathbb{S},d)$ is separable, it is homeomorphic to a subset of $[0,1]^\infty$ by **6.22**. Accordingly, there exists a metric d' equivalent to d such that $(\mathbb{S},d')$ is relatively compact. In this metric, let $\hat{\mathbb{S}}$ be a compact space containing $\mathbb{S}$ and let $\hat{\mathcal{S}}$ be the Borel field on $\hat{\mathbb{S}}$. It cannot be assumed that $\mathbb{S} \in \hat{\mathcal{S}}$, but $\mathcal{S}$, the Borel field of $\mathbb{S}$, is the trace of $\hat{\mathcal{S}}$ on $\mathbb{S}$.

Define a family of measures $\hat{\Pi}$ on $\hat{\mathbb{S}}$ such that, for $\hat{\mu} \in \hat{\Pi}$, $\hat{\mu}(A) = \mu(A \cap \mathbb{S})$ for $\mu \in \Pi$ and each $A \in \hat{\mathcal{S}}$. To prove that Π is compact it is shown that a sequence of measures $\{\mu_n, n \in \mathbb{N}\}$ from Π has a cluster point in Π. Consider the counterpart sequence $\{\hat{\mu}_n, n \in \mathbb{N}\}$ in $\hat{\Pi}$. Since $\hat{\mathbb{S}}$ is compact, $\hat{\Pi}$ is compact by **29.11** so this sequence has one or more cluster points in $\hat{\Pi}$. Let ν be such a cluster point. The object is to show that there exists a p.m. $\mu \in \Pi$ such that $\hat{\mu} = \nu$.

Tightness of Π means that for every integer r there is a compact set $K_r \in \mathcal{S}$ such that $\mu(K_r) \geq 1 - 1/r$ for all $\mu \in \Pi$. Being closed in $\hat{\mathbb{S}}$, $K_r \in \hat{\mathcal{S}}$ and $\hat{\mu}(K_r) = \mu(K_r \cap \mathbb{S}) = \mu(K_r)$, all $\mu \in \Pi$. Since K_r is closed,

$$\nu(K_r) \geq \limsup_{k\to\infty} \hat{\mu}_{n_k}(K_r) \geq 1 - 1/r \tag{29.49}$$

for some subsequence $\{n_k, k \in \mathbb{N}\}$, by **29.1**(c). Since $\bigcup_r K_r \in \hat{\mathcal{S}}$, $\nu(\bigcup_r K_r) = 1$. Now, suppose $\nu^*(\mathbb{S})$ denotes the outer measure of $\mathbb{S}$ in terms of coverings by $\hat{\mathcal{S}}$-sets. Since $\bigcup_r K_r \subseteq \mathbb{S}$,

$$\nu^*(\mathbb{S}) \geq \nu^*\Big(\bigcup_r K_r\Big) = \nu\Big(\bigcup_r K_r\Big) = 1.$$

Applying **3.17**, note that $\mathbb{S}$ is ν-measurable since the inequality in (3.19) becomes

$$\nu^*(B \cap \mathbb{S}) \leq \nu^*(B) \tag{29.50}$$

which holds for all $B \in \hat{\mathcal{S}}$. Since $\mathcal{S}$ is the trace of $\hat{\mathcal{S}}$ on $\mathbb{S}$, all the sets of $\mathcal{S}$ are accordingly ν-measurable and there exists a p.m. $\mu \in \Pi$ such that $\hat{\mu} = \nu$, as required. For any closed subset C of $\mathbb{S}$, there exists a closed $D \in \hat{\mathcal{S}}$ such that $C = D \cap \mathbb{S}$. The assertions $\limsup_k \hat{\mu}_{n_k}(D) \leq \hat{\mu}(D)$ and $\limsup_k \mu_{n_k}(C) \leq \mu(C)$ are equivalent and hence $\mu_{n_k} \Rightarrow \mu$ by **29.1**. This means that $\{\mu_n\}$ has a convergent subsequence, proving sufficiency. Notice that completeness of $\mathbb{S}$ is not needed for this part of the proof.

To prove necessity, assume Π is compact. Letting $\{S_{ni}, i \in \mathbb{N}\}$ be a countable covering of $\mathbb{S}$ by $1/n$-spheres and $\{j_n, n \in \mathbb{N}\}$ be any increasing subsequence of integers, define $\bar{A}_n = \bigcup_{i=1}^{j_n} \bar{S}_{ni}$. The assumption that there exists $\mu \in \Pi$ such that

$$\mu(\bar{A}_n) \leq 1 - \delta, \text{ all } n, \delta > 0 \tag{29.51}$$

leads to a contradiction, and so has to be false. If (29.51) is true for at least one element of (compact) Π, there is a convergent sequence $\{\mu_k, k \in \mathbb{N}\}$ in Π, with $\mu_k \Rightarrow \mu$, such that it holds for all μ_k. (Even if there is only one such element, $\mu_k = \mu$ for all k meets the requirement). By **29.1**,

$$\limsup_{k\to\infty} \mu_k(\bar{A}_n) \leq \mu(\bar{A}_n) \leq 1 - \delta. \tag{29.52}$$

However, now letting $n \to \infty$ and hence $j_n \to \infty$ yields $\mu(\mathbb{S}) \leq 1 - \delta$, which is the asserted contradiction. Putting $\delta = \varepsilon/2^n$ for $\varepsilon > 0$, it must be the case that

$$\mu(\bar{A}_n) > 1 - \varepsilon/2^n, \text{ all } n, \text{ all } \mu \in \Pi. \tag{29.53}$$

Letting $K_\varepsilon = \bigcap_{n=1}^{\infty} \bar{A}_n$, this set is compact by **29.13** ($\mathbb{S}$ being complete) and it follows as in (29.48) above that $\mu(K_\varepsilon) > 1 - \varepsilon$. Since μ is an arbitrary element of Π, the family is uniformly tight. ∎

This section is concluded with a useful result for measures on product spaces. See §7.4 for a discussion of the marginal measures.

29.16 Theorem A p.m. μ on the space $(\mathbb{S} \times \mathbb{T}, \mathcal{S} \otimes \mathcal{T})$ with the product topology is tight iff the marginal p.m.s μ_x and μ_y are tight.

Proof For a set $K \in \mathbb{S} \times \mathbb{T}$, let $K_x = \pi_x(K)$ denote the projection of K onto $\mathbb{S}$. Since the projection is continuous (see §6.5), K_x is compact if K is compact (**5.20**). Since

$$\mu_x(K_x) = \mu(K_x \times \mathbb{T}) \geq \mu(K) \tag{29.54}$$

tightness of μ implies tightness of μ_x. Repeating the argument for μ_y proves the necessity. It is sufficient if there exists a compact set $K \in \mathcal{S} \otimes \mathcal{T}$, having measure exceeding $1 - \varepsilon$. Consider the set $K = A \times B$ where $A \in \mathcal{S}$ and $\mu_x(A) > 1 - \varepsilon/2$ and $B \in \mathcal{T}$ where $\mu_y(B) > 1 - \varepsilon/2$. Note that

$$K^c = (A \times B^c) \cup (A^c \times B) \cup (A^c \times B^c) \tag{29.55}$$

where the sets of the union on the right are disjoint. Thus,

$$\begin{aligned} \mu(K^c) &\leq \mu(A^c \times B) + \mu(A \times B^c) + 2\mu(A^c \times B^c) \\ &= \mu(A^c \times \mathbb{T}) + \mu(\mathbb{S} \times B^c) \\ &= \mu_x(A^c) + \mu_y(B^c) \leq \varepsilon. \end{aligned} \tag{29.56}$$

If A and B are compact they are separable in the relative topologies generated from $\mathbb{S}$ and $\mathbb{T}$ (**5.7**) and hence K is compact by **6.17**. ∎

This result generalizes by iteration to products of any finite order.

29.5 Weak Convergence in *C*

As in §27.4 the symbol $C_{[0,1]}$, also written C for brevity when the context is clear, represents the space of continuous functions on the unit interval. C is a separable space and the Borel field $\mathcal{B}_C$ matches the projection σ-field. Further recall from §27.5 (see page 608) the class $\mathbb{M}_C$ of probability measures on $(C, \mathcal{B}_C)$ whose finite dimensional distributions satisfy (27.43). As was pointed out there, thanks to **29.7** this class of p.m.s can be metrized as a separable space.

Now let $\{\mu_n\}$ be a sequence of elements from $\mathbb{M}_C$. For example, consider the distributions associated with a sequence like $\{Y_n, n \in \mathbb{N}\}$ whose elements are defined in (27.44). According to **29.15**, the necessary and sufficient condition for the family $\{\mu_n\}$ to be compact and hence to possess (by **5.10**) a cluster point in $\mathbb{M}_C$ is that it is uniformly tight. The next result is essentially a stochastic variant of the Arzelà–Ascoli theorem. The message is that uniform tightness of measures on C is a matter of assigning a high enough probability to a compact set, with Theorem **27.14** providing the relevant compactness criteria.

29.17 Theorem ([21] th. 8.2; [22] th. 7.3) $\{\mu_n\}$ is uniformly tight iff

(a) for each $\eta > 0 \;\exists\; M < \infty$ such that

$$\mu_n\big(\{x \in C : |x(0)| > M\}\big) \le \eta, \quad n \ge 1; \tag{29.57}$$

(b) for each $\varepsilon > 0$ and $\eta > 0 \;\exists\; \delta \in (0,1)$ and $N \in \mathbb{N}$ such that

$$\mu_n\big(\{x \in C : w_x(\delta) \ge \varepsilon\}\big) \le \eta, \quad n \ge N \tag{29.58}$$

where w_x is defined in (27.26). □

Condition (a) must hold for a tight probability measure for any finite n and the issue is only whether it holds in the limit. Condition (b) is a form of the stochastic equicontinuity defined in §22.3. Asymptotic equicontinuity is sufficient in this application; the conditions need only be shown for $n \ge N$ for some finite N since C is a separable complete space and each individual member of $\{\mu_n\}$ is tight by **29.12**. For uniform tightness it suffices to show that the conditions hold 'in the tail'.

Proof of 29.17 To prove the necessity, let $\{\mu_n\}$ be uniformly tight and for $\eta > 0$ choose a compact set $K \in \mathcal{B}_C$ with $\mu_n(K) > 1 - \eta$. By **27.14**, for any $\varepsilon > 0$ there exist $M < \infty$ and $\delta \in (0,1)$ such that

$$K \subseteq \big\{x : |x(0)| \le M\big\} \cap \big\{x : w_x(\delta) < \varepsilon\big\}. \tag{29.59}$$

Applying the De Morgan law,

$$\begin{aligned}\eta &\ge \mu_n(K^c)\\ &\ge \mu_n\big(\{x : |x(0)| > M\} \cup \{x : w_x(\delta) \ge \varepsilon\}\big)\\ &\ge \max\big\{\mu_n(\{x : |x(0)| > M\}), \mu_n(\{x : w_x(\delta) \ge \varepsilon\})\big\}.\end{aligned} \tag{29.60}$$

Hence (29.57) and (29.58) hold, for all $n \in \mathbb{N}$.

Write $\mu^*(\cdot)$ as shorthand for $\sup_{n\geq N}\mu_n(\cdot)$. To prove sufficiency, consider for $k = 1,2,\ldots$ the sets

$$A_k = \{x : w_x(\delta_k) < 1/k\} \tag{29.61}$$

where $\{\delta_k\}$ is a sequence chosen so that $\mu^*(A_k) > 1 - \theta/2^{k+1}$, for $\theta > 0$. This is possible by condition (b). Also set $B = \{x : |x(0)| \leq M\}$ where M is chosen so that $\mu^*(B) > 1 - \theta/2$, which is possible by condition (a). Then define a closed set $K = (\bigcap_{k=1}^{\infty} A_k \cap B)^-$ and note that conditions (27.28) and (27.29) hold for the case $A = K$. Hence by **27.14**, K is compact. But

$$\begin{aligned}\mu^*(K^c) &\leq \mu^*\left(\bigcup_{k=1}^{\infty} A_k^c \cup B^c\right) \\ &\leq \sum_{k=1}^{\infty} \mu^*(A_k^c) + \mu^*(B^c) \\ &\leq 2\theta \sum_{k=1}^{\infty} 1/2^{k+2} + \theta/2 = \theta.\end{aligned} \tag{29.62}$$

This last inequality is to be read as $\sup_{n\geq N}\mu_n(K^c) \leq \theta$, or equivalently $\inf_{n\geq N}\mu_n(K) > 1 - \theta$. Since θ is arbitrary and every individual μ_n is tight by **29.12**, in particular for $1 \leq n < N$, it follows that the sequence $\{\mu_n\}$ is uniformly tight. ∎

The following lemma adapted from [21] th. 8.3 is a companion to the last result, supplying in conjunction with it a relatively primitive sufficient condition for uniform tightness.

29.18 Lemma If for some $\delta \in (0,1)$ and $N \geq 1$

$$\sup_{0\leq t\leq 1-\delta} \mu_n\left(\left\{\sup_{t\leq s\leq t+\delta} |x(s) - x(t)| \geq \tfrac{1}{2}\varepsilon\right\}\right) \leq \tfrac{1}{2}\eta\delta \tag{29.63}$$

for $\varepsilon > 0$, $\eta > 0$, and $n \geq N$ then, in the same case,

$$\mu_n(\{w_x(\delta) \geq \varepsilon\}) \leq \eta. \tag{29.64}$$

Proof Fixing δ, consider the partition $\{t_1, \ldots, t_r\}$ of $[0,1]$, for $r = 1 + [1/\delta]$, where $t_i = i\delta$ for $i = 1, \ldots, r-1$ and $t_r = 1$. Thus, for $\frac{1}{2} < \delta < 1$, $r = 2$ and the partition is $\{\delta, 1\}$; for $\frac{1}{3} < \delta \leq \frac{1}{2}$, $r = 3$ and the partition is $\{\delta, 2\delta, 1\}$; and so on. The width of these intervals is at most δ. A given interval $[t, t']$ with $|t' - t| \leq \delta$ must either lie within an interval of the partition, or at most overlap two adjoining intervals;

it cannot span three or more. In the event that $|x(t')-x(t)|\geq\varepsilon$, x must change absolutely by at least $\frac{1}{2}\varepsilon$ in at least one of the interval(s) overlapping $[t,t']$ and the probability of the latter event is at least that of the former. In other words, considering all such intervals,

$$\begin{aligned}\mu_n(\{x : w_x(\delta)\geq\varepsilon\}) &\leq \mu_n\Bigg(\bigcup_{i=1}^{r}\Big\{x : \sup_{s,s'\in[t_{i-1},t_i]}|x(s')-x(s)|\geq\tfrac{1}{2}\varepsilon\Big\}\Bigg)\\ &\leq \sum_{i=1}^{r}\mu_n\Bigg(\Big\{x : \sup_{s,s'\in[t_{i-1},t_i]}|x(s')-x(s)|\geq\tfrac{1}{2}\varepsilon\Big\}\Bigg)\\ &\leq \tfrac{1}{2}r\eta\delta\\ &\leq \eta \end{aligned}\tag{29.65}$$

where the third of these inequalities applies (29.63) and the final one follows because $r\delta\leq 2$. ∎

The condition in (29.64) is identical with (29.58) for elements of C and provides a more convenient way of specifying the required equicontinuity property.

These results provoke a technical query over measurability. In §22.1 difficulties with standard measure theory arose over showing that functions such as $\sup_{t\leq s\leq t+\delta}|x(s)-x(t)|$ in (29.63) and $w_x(\delta)$ in (29.58) are random variables. However, it is possible to show that sets such as the one in (29.63) are $\mathcal{F}$-analytic and hence nearly measurable. In other words, complacency about this issue can be justified. The same qualification can be taken as implicit wherever such sets arise below.

29.6 An FCLT for Martingale Differences

Let $S_{n0}=0$ and $S_{nj}=\sum_{i=1}^{j}U_{ni}$ for $j=1,\ldots,n$, where $\{U_{ni}\}$ is a zero-mean stochastic array, normalized so that $\mathrm{E}(S_{nn}^2)=1$. As in the previous applications of array notation in Part V and elsewhere, the leading example is $U_{ni}=U_i/s_n$ where $\{U_i\}$ is a zero-mean sequence and $s_n^2=\mathrm{E}(\sum_{i=1}^{n}U_i)^2$. Define an element Y_n of $C_{[0,1]}$, somewhat as in (27.44), as follows:

$$Y_n(t)=S_{n,[nt]}+(nt-[nt])U_{n,[nt]+1},\ 0\leq t<1 \tag{29.66a}$$

$$Y_n(1)=S_{nn}. \tag{29.66b}$$

This is the type of process sketched in Figure 27.2. An equivalent representation of (29.66a) is

$$Y_n(t) = S_{n,j-1} + (nt - j + 1)U_{nj} \text{ for } (j-1)/n \leq t < j/n, j = 1, \ldots, n.$$

The question to be addressed is whether the distribution of Y_n possesses a weak limit as $n \to \infty$.

The interpolation terms in $Y_n(t)$ are necessary to generate a continuous function, but from an algebraic point of view they are a nuisance; dropping them gives

$$X_n(t) = S_{n,[nt]} = S_{n,j-1} \text{ for } (j-1)/n \leq t < j/n,\ j = 1, \ldots, n \tag{29.67a}$$

$$X_n(1) = S_{nn}. \tag{29.67b}$$

If conditions of the type discussed in Chapters 24 and 25 are imposed on $\{U_{ni}\}$, $X_n(1) \to_{\mathrm{d}} \mathrm{N}(0,1)$ as $n \to \infty$. If for example $U_i \sim_{\mathrm{d}}$ i.i.d.$(0,\sigma^2)$ so that $U_{ni} = U_i/s_n$ where $s_n^2 = n\sigma^2$, this is just the Lindeberg–Lévy theorem. However, the Lindeberg–Lévy theorem yields additional conclusions that are less often remarked; it is easy to verify that for each distinct pair $t_1, t_2 \in [0,1]$,

$$X_n(t_2) - X_n(t_1) \xrightarrow{\mathrm{d}} \mathrm{N}(0, |t_2 - t_1|). \tag{29.68}$$

Since non-overlapping partial sums of independent variates are independent, for any $0 \leq t_1 < t_2 < t_3 \leq 1$, the increments $X_n(t_2) - X_n(t_1)$ and $X_n(t_3) - X_n(t_2)$ converge to a pair of independent Gaussian variates with variances $t_2 - t_1$ and $t_3 - t_2$, so that their sum $X_n(t_3) - X_n(t_1)$ is asymptotically Gaussian with variance $t_3 - t_1$, as required. Under the assumptions,

$$|Y_n(t) - X_n(t)| = (nt - [nt])|U_{n,[nt]+1}| \xrightarrow{\mathrm{pr}} 0 \tag{29.69}$$

so that $Y_n(t)$ and $X_n(t)$ have the same asymptotic distribution.

Since $Y_n(0) = 0$ the finite-dimensional distributions of Y_n converge under the same conditions to those of a Brownian motion process as $n \to \infty$. As noted in §27.6, this is not a sufficient condition for the convergence of the p.m.s of Y_n to Wiener measure. But with the aid of **29.17**, $\{Y_n\}$ can be proved uniformly tight and hence the sequence has at least one cluster point in $\mathbb{M}_C$. Since any such cluster point must have the finite-dimensional distributions of W and the finite-dimensional cylinders are a determining class for $(C, \mathcal{B}_C)$, W must be the weak limit of the sequence. This convergence will be expressed either by writing $\mu_n \Rightarrow W$, or, more commonly in what follows, by $Y_n \to_{\mathrm{d}} B$.

This type of result is called a functional central limit theorem (FCLT), although the term *invariance principle* is also used in this context, to capture the idea of a distribution arising in the limit that is invariant to that of the constituent random variables. The original FCLT for i.i.d. increments (the generalization of the Lindeberg–Lévy theorem) is known as Donsker's theorem ([59]). Using the results of previous chapters, in particular **25.3**, the theorem can be generalized to the case of a heterogeneously distributed martingale difference, although the basic idea is the same.

29.19 Theorem Let Y_n be defined by (29.66a) and (29.66b), where $\{U_{ni}, \mathcal{F}_{ni}\}$ is a martingale difference array with variance array $\{\sigma_{ni}^2\}$ and $\sum_{i=1}^{n} \sigma_{ni}^2 = 1$. If

(a) $\displaystyle\sum_{i=1}^{n} U_{ni}^2 \xrightarrow{\text{pr}} 1$

(b) $\max_{1 \leq i \leq n} |U_{ni}| \xrightarrow{\text{pr}} 0$

(c) $\displaystyle\lim_{n\to\infty} \sum_{i=1}^{[nt]} \sigma_{ni}^2 = t$ for all $t \in [0,1]$

then $Y_n \xrightarrow{\text{d}} B$. □

Conditions (a) and (b) reproduce the corresponding conditions of **25.3** and their role is to establish the finite-dimensional distributions of the process, via the conventional CLT. Condition (c) is a global stationarity condition (see §13.1) which has no counterpart in the CLT conditions of Chapter 25. Its effect is to rule out cases such as **25.8** and **25.9**. By simple subtraction, the condition is sufficient for

$$\lim_{n\to\infty} \sum_{[nt]+1}^{[ns]} \sigma_{ni}^2 = s - t \tag{29.70}$$

for $0 \leq t < s \leq 1$. Clearly, without this restriction condition (29.68) could not hold for all t_1 and t_2.

Proof of 29.19 Conditions **25.3**(a) and **25.3**(b) are satisfied, writing U_{ni} for X_{nt}. In view of the last remarks and (29.69), the finite-dimensional distributions of Y_n converge to those of W. Therefore, it has only to be shown that the sequence of p.m.s of the $\{Y_n\}$ is uniformly tight.

For positive integers k and m with $k + m \leq n$, $p > 2$, and $\lambda > 0$, the maximal inequality for martingales **16.20** gives

$$P\Big(\max_{1\leq j\leq m} |S_{n,k+j} - S_{nk}| > \lambda\Big) \leq \frac{\mathrm{E}|S_{n,k+m} - S_{nk}|^p}{\lambda^p}. \tag{29.71}$$

Note that $S_{n,k+m} - S_{nk} = \sum_{i=k+1}^{k+m} U_{ni}$ and so define

$$s_{nkm}^2 = \mathrm{E}(S_{n,k+m} - S_{nk})^2 = \sum_{i=k+1}^{k+m} \sigma_{ni}^2. \tag{29.72}$$

Since $(S_{n,k+m} - S_{nk})/s_{nkm} \to_{\mathrm{d}} \mathrm{N}(0,1)$ by **25.3**, $\mathrm{E}|S_{n,k+m} - S_{nk}|^p/s_{nkm}^p \to \mu_p$ where μ_p is the p^{th} absolute moment of the standard normal. Setting $k = [nt]$ and $m = [n\delta]$ for fixed $\delta \in (0,1)$ to be chosen and $t \in [0, 1-\delta]$, $\lim_{n\to\infty} s_{nkm}^2 = \delta$ according to (29.70) and hence $\mathrm{E}|S_{n,k+m} - S_{nk}|^p \to \delta^{p/2}\mu_p$. For $N_0 \geq 1$, large enough that the Gaussian approximation is sufficiently close, (29.71) implies

$$P\Big(\max_{1\leq j\leq m} |S_{n,k+j} - S_{nk}| > \lambda\Big) \leq \frac{\delta^{p/2}\mu_p}{\lambda^p} \text{ for } n \geq N_0. \tag{29.73}$$

For given positive numbers η and ε, put $\lambda = \varepsilon/4$ and then for any $p > 2$, δ can be chosen to satisfy the inequality

$$4^p \frac{\delta^{p/2}\mu_p}{\varepsilon^p} \leq \tfrac{1}{4}\eta\delta.$$

For example, for the case $p = 3$, (29.73) with $\mu_3 = \sqrt{8/\pi}$ implies that

$$P\Big(\max_{1\leq j\leq m} |S_{n,k+j} - S_{nk}| > \tfrac{1}{4}\varepsilon\Big) \leq \tfrac{1}{4}\eta\delta,\ n \geq N_0 \tag{29.74}$$

for $\sqrt{\delta} \leq \min\{1, \varepsilon^3\eta/408\}$.

Now,

$$Y_n(s) - Y_n(t) = S_{n,[ns]} - S_{n,[nt]} + R_n(s,t) \tag{29.75}$$

for $s > t$ from (29.66a) and (29.66b), where

$$R_n(s,t) = (ns - [ns])U_{n,[ns]+1} - (nt - [nt])U_{n,[nt]+1}. \tag{29.76}$$

For $t \in [0, 1-\delta]$ there exists $s' \in [t, t+\delta]$ such that

$$|Y_n(s') - Y_n(t)| = \sup_{t\leq s\leq t+\delta} |Y_n(s) - Y_n(t)|. \tag{29.77}$$

There also exists n large enough (say, $n \geq N_1$) that, for any such t and s', $[nt] \leq [ns'] \leq [nt] + [n\delta]$ and hence

$$|S_{n,[ns']} - S_{n,[nt]}| \leq \max_{1 \leq j \leq [n\delta]} |S_{n,[nt]+j} - S_{n,[nt]}|. \tag{29.78}$$

It follows (also invoking the triangle inequality) that for $n \geq N_1$,

$$\begin{aligned} |Y_n(s') - Y_n(t)| &= |S_{n,[ns']} - S_{n,[nt]} + R_n(s',t)| \\ &\leq \max_{1 \leq j \leq [n\delta]} |S_{n,[nt]+j} - S_{n,[nt]}| + |R_n(s',t)|. \end{aligned} \tag{29.79}$$

By condition (b) of the theorem, $P\big(\sup_{s,t} |R_n(s,t)| \geq \frac{1}{4}\varepsilon\big) \to 0$ as $n \to \infty$ and hence there exists $N_2 \geq 1$ such that, for $n \geq N_2$,

$$P(R_n(s',t)| \geq \tfrac{1}{4}\varepsilon) \leq \tfrac{1}{4}\eta\delta. \tag{29.80}$$

Inequalities (29.80) and (29.74) jointly imply that for $n \geq N^* = \max\{N_0, N_1, N_2\}$,

$$\begin{aligned} &P(|Y_n(s') - Y_n(t)| \geq \tfrac{1}{2}\varepsilon) \\ &\quad \leq P\Big(\max_{1 \leq j \leq |n\delta]} |S_{n,[nt]+j} - S_{n,[nt]}| + |R_n(s',t)| \geq \tfrac{1}{2}\varepsilon\Big) \\ &\quad \leq P\Big(\Big\{\max_{1 \leq j \leq [n\delta]} |S_{n,[nt]+j} - S_{n,[nt]}| \geq \tfrac{1}{4}\varepsilon\Big\} \cup \Big\{|R_n(s',t)| \geq \tfrac{1}{4}\varepsilon\Big\}\Big) \\ &\quad \leq P\Big(\max_{1 \leq j \leq [n\delta]} |S_{n,[nt]+j} - S_{n,[nt]}| \geq \tfrac{1}{4}\varepsilon\Big) + P\big(|R_n(s',t)| \geq \tfrac{1}{4}\varepsilon\big) \\ &\quad \leq \tfrac{1}{2}\eta\delta. \end{aligned} \tag{29.81}$$

This inequality holds for all $t \in [0, 1-\delta]$, leading to the conclusion

$$\sup_{0 \leq t \leq 1-\delta} P\Big(\sup_{t \leq s \leq t+\delta} |Y_n(s) - Y_n(t)| \geq \tfrac{1}{2}\varepsilon\Big) \leq \tfrac{1}{2}\eta\delta,\ n \geq N^*. \tag{29.82}$$

Note that (29.63) is identical to (29.82) for the case $\mu_n(A) = P(Y_n \in A)$ and that η and ε are arbitrary. Therefore, uniform tightness of the corresponding sequence of measures follows by **29.17** and **29.18**. ∎

It would not be unreasonable, reading this proof, to ask: Why include the terms R_n in the sum, only to show that they are negligible and make no contribution to the limit distribution? In other words, why not work with X_n in (29.67) instead of Y_n in (29.66) from the outset? The answer is, of course, that the tightness criteria are specified for elements of C so the formality of making the empirical process continuous must be gone through, though it is no more than a formality given that the limit process does lie in C with probability 1. It is tidier to obtain the tightness criteria for the space of processes to which X_n belongs and this is precisely what is done in Chapter 30 to follow. In the results of Chapters 31 and 32 the R_n terms do not appear. However, the additional complications with these techniques are such that it is useful to know how to work wholly in C should this prove a convenient option.

This section concludes with the result promised in §27.6:

29.20 Corollary Wiener measure exists. □

The existence is actually proved in **29.19** which derived a unique limiting distribution that satisfied the specifications of **27.16**. The points that are conveniently highlighted by a separate statement are that the tightness argument developed to prove **29.19** holds independently of the existence of W as such and that the central limit theorem plays no role in the proof of existence.

Proof of 29.20 Consider the process Y_n of (27.44), having Gaussian increments. Setting $U_{ni} = n^{-1/2}U_i$ in **29.19**, if conditions (a), (b), and (c) of the theorem are satisfied it follows by the reasoning of **29.17** and **29.18** that the associated sequence of measures is uniformly tight and possesses a limit. Since the Gaussian distribution is stable and the increments are independent by construction, this limit must have the finite-dimensional distributions specified by **27.16**.

The proof now consists in showing that the conditions of **29.19** hold for the Gaussian case. Condition **29.19**(c) holds by construction. Condition **29.19**(a) follows from an application of the weak law of large numbers (e.g. **24.5**) recalling that since U_i is Gaussian $\{U_i^2\}$ is an independent sequence possessing all its moments. Finally, condition **29.19**(b) holds by **24.16** if the collection $\{U_1, \dots, U_n\}$ satisfies the Lindeberg condition, which is obvious given their Gaussianity and **24.10**. ∎

29.7 The Multivariate Case

To extend these results to vector-valued processes, there is no difficulty in extending the approach of §26.4. Define the space $(C_{[0,1]})^m$, written as C^m for brevity, to be the space of continuous vector functions

$$\boldsymbol{x} = (x_1, \dots, x_m)' : [0,1] \mapsto \mathbb{R}^m.$$

C^m is the product of m copies of C and can be endowed with a metric such as

$$d_U^m(\boldsymbol{x}, \boldsymbol{y}) = \max_{1 \le j \le m} \{d_U(x_j, y_j)\}. \tag{29.83}$$

This induces the product topology and the coordinate projections remain continuous. Since C is separable, C^m is also separable by **6.16** and $\mathcal{B}_C^m = \mathcal{B}_C \otimes \mathcal{B}_C \otimes \ldots \otimes \mathcal{B}_C$ (the σ-field generated by the open rectangles of C^m) is the Borel field of C^m by m-fold iteration of **27.5**. $(C^m, \mathcal{B}_C^m)$ is therefore a measurable space.

Let the finite-dimensional sets of C^m be denoted by

$$\mathcal{H}_C^m = \{\pi_{t_1,\ldots,t_k}^{-1}(B) \subseteq C^m : B \in \mathcal{B}^{mk}, t_1, \ldots, t_k \in [0,1], k \in \mathbb{N}\}. \tag{29.84}$$

Again thanks to the product topology, $\mathcal{H}_C^m$ is the field generated from the sets in the product of m copies of $\mathcal{H}_C$ defined in (27.34). Define $\mathcal{P}_C^m = \sigma(\mathcal{H}_C^m)$.

29.21 Theorem $\mathcal{H}_C^m$ is a determining class for $(C^m, \mathcal{B}_C^m)$.

Proof An open sphere in $\mathcal{B}_C^m$ is a set

$$\begin{aligned} S(\boldsymbol{x},\alpha) &= \{\boldsymbol{y} \in C^m : d_U^m(\boldsymbol{x},\boldsymbol{y}) < \alpha\} \\ &= \left\{\boldsymbol{y} \in C^m : \max_{1\le j\le m} \sup_t |y_j(t) - x_j(t)| < \alpha\right\}. \end{aligned} \tag{29.85}$$

The set

$$H_k(\boldsymbol{x},\alpha) = \left\{\boldsymbol{y} \in C^m : \max_{1\le j\le m} \max_{1\le i\le 2^k} \left|y_j(2^{-k}i) - x_j(2^{-k}i)\right| < \alpha\right\} \tag{29.86}$$

is an element of $\mathcal{H}_C^m$. It follows by the argument of **27.15** that

$$S(\boldsymbol{x},r) = \bigcup_{n=1}^{\infty} \bigcap_{k=1}^{\infty} H_k(\boldsymbol{x}, r - 1/n) \in \mathcal{P}_C^m \tag{29.87}$$

and hence that $\mathcal{B}_C^m \subseteq \mathcal{P}_C^m$. Since $\mathcal{H}_C^m$ is a field, the result follows by the extension theorem. ∎

It is also straightforward to show that $\mathcal{B}_C^m = \mathcal{P}_C^m$ by a similar generalization from **27.15**, but the above is all that is required for the present purpose.

A leading example of a measure on $(C^m, \mathcal{B}_C^m)$ is W^m, the p.m. of m-dimensional standard Brownian motion. An m-vector $\boldsymbol{B} = (B_1, \ldots, B_m)'$ distributed according to W^m has as its elements m mutually independent Brownian motions, such that

$$\boldsymbol{B}(t) \overset{\mathrm{d}}{\sim} \mathrm{N}(\boldsymbol{0}, t\boldsymbol{I}_m) \tag{29.88}$$

where $\boldsymbol{I}_m$ is the $m \times m$ identity matrix and the process has independent increments with

$$\mathrm{E}\big((\boldsymbol{B}(s)-\boldsymbol{B}(t))(\boldsymbol{B}(s)-\boldsymbol{B}(t))'\big) = (s-t)\boldsymbol{I}_m \tag{29.89}$$

for $0 \le t < s \le 1$. The following general result can now be proved.

29.22 Theorem Let $\{\boldsymbol{U}_{ni}, \mathcal{F}_{ni}\}$ be an m-vector martingale difference array with variance matrix array $\{\boldsymbol{\Sigma}_{ni}\}$, such that $\sum_{i=1}^n \boldsymbol{\Sigma}_{ni} = \boldsymbol{I}_m$. Also let

$$\boldsymbol{Y}_n(t) = \boldsymbol{S}_{n,j-1} + (nt-j+1)\boldsymbol{U}_{nj} \text{ for } (j-1)/n \le t < j/n \tag{29.90}$$

for $j = 1, \dots, n$ and $\boldsymbol{Y}_n(1) = \boldsymbol{S}_{nn}$, where $\boldsymbol{S}_{n0} = \boldsymbol{0}$ and

$$\boldsymbol{S}_{nj} = \sum_{i=1}^{j} \boldsymbol{U}_{ni},\ j = 1, \dots, n. \tag{29.91}$$

$\boldsymbol{Y}_n \to_{\mathrm{d}} \boldsymbol{B}$ if

(a) $\sum_{i=1}^n \boldsymbol{U}_{ni}\boldsymbol{U}_{ni}' \to_{\mathrm{pr}} \boldsymbol{I}_m$

(b) $\max_{1\le i\le n} \boldsymbol{U}_{ni}'\boldsymbol{U}_{ni} \to_{\mathrm{pr}} \boldsymbol{0}$

(c) $\lim_{n\to\infty} \sum_{i=1}^{[nt]} \boldsymbol{\Sigma}_{ni} = t\boldsymbol{I}_m$, for all $t \in [0,1]$.

Proof Consider for an m-vector $\boldsymbol{\lambda}$ of unit length the scalar process $\boldsymbol{\lambda}'\boldsymbol{Y}_n$, having increments $\boldsymbol{\lambda}'\boldsymbol{U}_{ni}$ By definition, $\{\boldsymbol{\lambda}'\boldsymbol{U}_{ni}, \mathcal{F}_{ni}\}$ is a scalar martingale difference array with variance sequence $\boldsymbol{\lambda}'\boldsymbol{\Sigma}_{ni}\boldsymbol{\lambda}$. It is easily verified that all the conditions of **29.19** are satisfied and so $\boldsymbol{\lambda}'\boldsymbol{Y}_n \to_{\mathrm{d}} \mathrm{N}(0,t)$ with similar conclusions regarding all the finite-dimensional distributions of the process.

It follows by the Cramér–Wold theorem **26.5** that

$$\boldsymbol{Y}_n(t) \xrightarrow{\mathrm{d}} \mathrm{N}(\boldsymbol{0}, t\boldsymbol{I}_m) \tag{29.92}$$

with similar conclusions regarding all the finite-dimensional distributions of the process; these are identical to the finite-dimensional distributions of W^m. Since $\mathcal{H}_C^m$ is a determining class for $(C^m, \mathcal{B}_C^m)$, any weak limit of the p.m.s of $\{\boldsymbol{Y}_n\}$ can only be W^m. It remains to be shown that these p.m.s are uniformly tight. But by **29.16**, this is true provided the marginal p.m.s of the process are uniformly tight. Picking $\boldsymbol{\lambda}$ to be the j^{th} column of $\boldsymbol{I}_m$ for $j = 1, \dots, m$ and applying the argument of **29.19** shows that this condition holds and completes the proof. ∎

The arguments of §26.4 can be extended to convert **29.22** into a powerful general limit result. Replacing **29.22**(c) by

(c′) $\lim_{n\to\infty}\sum_{i=1}^{[nt]}\boldsymbol{\Sigma}_{ni} = t\boldsymbol{\Sigma}$

where $\boldsymbol{\Sigma}$ is an arbitrary variance matrix and defining $\boldsymbol{L}^{-}$ such that $\boldsymbol{L}^{-}\boldsymbol{\Sigma}\boldsymbol{L}^{-'} = \boldsymbol{I}_p$ as in **26.6**, **29.22** holds for the transformed vector process $\boldsymbol{Z}_n = \boldsymbol{L}^{-}\boldsymbol{Y}_n$. The limit of the process $\boldsymbol{Y}_n$ itself can then be determined by applying the continuous mapping theorem. This is a linear combination of independent Brownian motions, the finite-dimensional distributions of which are jointly Gaussian by **11.19**. This is a p-dimensional correlated Brownian motion having covariance matrix $\boldsymbol{\Sigma}$ and denoted $\boldsymbol{B}(\boldsymbol{\Sigma})$. The result is written in the form

$$\boldsymbol{Y}_n \xrightarrow{\mathrm{d}} \boldsymbol{B}(\boldsymbol{\Sigma}). \tag{29.93}$$

An invariance principle can be used in this way to convert propositions about dependence between stochastic processes converging to Brownian motion into more tractable results about correlation in large samples. Given an arbitrarily related set of such processes, there always exist linear combinations of the set that are asymptotically independent of one another.

30
Càdlàg Functions

30.1 The Space D

The space C is an immensely important object due to its convenient properties, in particular the separability and completeness under the uniform metric. This gives rise to a mathematical framework that is pleasingly similar to the well-understood theory of real random variables. The downside is that the discrete-time empirical processes arising in econometrics cannot very conveniently be treated as elements of C. Consider again Figure 27.2 (page 603), where a finite sequence of observations is accommodated in C as a piecewise linear function by the visually appealing procedure of 'joining the dots'. Algebraically, on the other hand, the trick introduces needless complications. Compare Figure 27.2 with Figure 30.1 which shows the same set of observations (apart from the omission of the point $X_n(1)$, to be explained below) in the form of a step function with discontinuities. The object illustrated is an element of the space $D_{[0,1]}$ of càdlàg functions on the unit interval (see **5.27**) of which $C_{[0,1]}$ is a subset. Henceforth, $D_{[0,1]}$ will be denoted D for brevity when there is no risk of confusion with other usages.

As shown in **27.3**, D is not a separable space under the uniform metric which means that the convergence theory of Chapter 29 will not apply to (D, d_U). d_U is not the only metric that can be defined on D and it is worth investigating alternatives because, once the theory can be shown to work on D in the same kind of way that it does on C, a great simplification is achieved.

Abandoning d_U is not the only way of overcoming measurability problems. Another approach is simply to agree to exclude the pathological cases from the

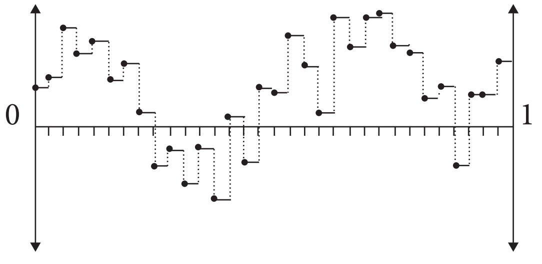

Figure 30.1

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0030

field of events under consideration. This can be achieved by working with the σ-field $\mathcal{P}_{D,}$ the restriction to D of the projection σ-field (see §27.3). In contrast with the case of C, $\mathcal{P}_D \subset \mathcal{B}_D$ (compare **27.15**) and all the awkward cases such as uncountable discrete subsets are excluded from $\mathcal{P}_D$, while all the ones likely to arise in econometric analysis (which most often concerns convergence to limit points lying in C) are included. Studying measures on the space $\big((D, d_U), \mathcal{P}_D\big)$ is an interesting line of attack proposed originally by Dudley ([61], [62]) and described in detail in Pollard ([147]).

While this approach represents a large potential simplification (much of the present chapter could be dispensed with) an early decision has to be made about which line to adopt; there is little overlap between this theory and the methods pioneered by Skorokhod ([170], [171]), Prokhorov ([153]), and Billingsley ([21], [22]) which involve metrizing D as a separable complete space. Although the technical overheads of the latter approach are greater, it has the advantage that, once the investment is made, the probabilistic environment is an analogue of Euclidean space for which all sorts of useful topological and metric properties are known to hold. There is scope for debate on the relative merits of the two approaches, but the present treatment follows the majority of subsequent authors, who take their cue from Billingsley's work.

The possibility of metrizing $D_{[0,1]}$ as a separable space depends crucially on the fact that the permitted departures from continuity are of a limited type. The only possible discontinuities are jumps (also called 'discontinuities of the first kind'), points t at which $|x(t) - x(t-)| > 0$. There is no possibility of isolated discontinuity points t at which both $|x(t) - (t-)|$ and $|x(t) - x(t+)|$ are positive, because that would contradict right-continuity. There is, however, the possibility that $x(1)$ is isolated; it is necessary to discard this point and let $x(1) = x(1-)$. This is a little unfortunate but will not affect anything material.

30.1 Definition $D_{[0,1]}$ is the space of functions $x : [0,1] \mapsto \mathbb{R}$ satisfying the following conditions:

(a) $x(t+)$ exists for $t \in [0,1)$;

(b) $x(t-)$ exists for $t \in (0,1]$;

(c) $x(t) = x(t+)$, $t < 1$, and $x(1) = x(1-)$. □

The first theorem shows how under these conditions the maximum number of jumps is limited.

30.2 Theorem There exists for all $x \in D_{[0,1]}$ and every $\varepsilon > 0$ a finite partition $\{t_1, \ldots, t_r\}$ of $[0,1]$ with the property

$$\sup_{s,t\in[t_{i-1},t_i)} \left|x(t)-x(s)\right| < \varepsilon \tag{30.1}$$

for each $i = 1, \ldots, r$.

Proof This is by showing that $t_r = 1$ for a collection $\{t_1, \ldots, t_r\}$ satisfying (30.1), with $t_0 = 0$. For given x and ε let $\tau = \sup\{t_r\}$, the supremum being taken over all these collections. Noting the intervals of the partition are open above and $x(t-)$ exists for all $t > 0$, τ belongs to the set; that is, there exists r such that $\tau = t_r$.

Suppose $t_r < 1$ and consider the point $t_r + \delta \le 1$, for some $\delta > 0$. By definition of t_r, $\left|x(t_r+\delta) - x(t_{r-1})\right| \ge \varepsilon$. Hence consider the interval $[t_r, t_r + \delta)$. By right continuity, δ can be chosen so that $\left|x(t_r+\delta) - x(t_r)\right| < \varepsilon$. Hence there exists an $(r+1)$-fold collection satisfying the conditions of the theorem. It must be the case that $\tau \ge t_{r+1} = t_r + \delta$ and the assertion that $t_r = \tau$ is contradicted. It follows that $t_r = 1$. ∎

This elementary but slightly startling result shows that the number of jump points at which $\left|x(t) - x(t-)\right|$ exceeds any given positive number is at most finite. The number of jumps such that $\left|x(t) - x(t-)\right| > 1/n$ is finite for every n and the entire set of discontinuities is a countable union of finite sets, hence countable. Further,

$$\sup_t |x(t)| < \infty \tag{30.2}$$

since for any $t \in [0,1]$, $x(t)$ is expressible according to (30.1) as a finite sum of finite increments.

The modulus of continuity $w_x(\delta)$ in (27.26) provides a means of discriminating between functions in $C_{[0,1]}$ and functions outside the space. For just the same reasons, it is helpful to have a means of discriminating between càdlàg functions and those with arbitrary discontinuities. First write

$$w_x[t_1, t_2) = \sup_{t_1 \le s < t < t_2} |x(t) - x(s)| \tag{30.3}$$

to denote the largest change in x over the interval $[t_1, t_2)$. Then for $\delta \in (0,1)$ let Π_δ denote a partition $\{t_1, \ldots, t_r\}$ with $r \le [1/\delta]$ and $\min_i\{t_i - t_{i-1}\} > \delta$ where $t_0 = 0$ and $t_r = 1$ and then define

$$w'_x(\delta) = \inf_{\Pi_\delta}\left\{\max_{1\le i\le r} w_x[t_{i-1}, t_i)\right\}. \tag{30.4}$$

In words, $w'_x(\delta)$ is the smallest value over all partitions of $[0,1]$ coarser than δ of the largest change in x within an interval of the partition. This notion differs from and weakens that of $w_x(\delta)$ in (27.26), since $w'_x(\delta)$ can be small even if the points t_i are jump points such that $w_x(\delta)$ would be large. Taking the infimum with respect to choice of partition means that the points t_i are chosen as far as possible to be jump points, and the smaller δ is, the more such points can be accommodated and the smaller the modulus can be. For $\delta < \frac{1}{2}$ there is always a partition Π_δ in which $t_i - t_{i-1} < 2\delta$ for some i, so that

$$w'_x(\delta) \leq w_x(2\delta) \tag{30.5}$$

for any $x \in D$. The property

$$\lim_{\delta \to 0} w'_x(\delta) = 0 \tag{30.6}$$

holds for all $x \in D$, including obviously the elements of C as a special case, but not for more general functions.

30.3 Theorem Iff $x \in D$, $\exists\ \delta$ such that $w'_x(\delta) < \varepsilon$, for any $\varepsilon > 0$.

Proof Sufficiency is immediate from **30.2**. Necessity follows from the fact that if $x \notin D$ there is a point other than 1 at which x is not right-continuous; in other words, a point t at which $|x(t) - x(t+)| \geq \varepsilon$ for some $\varepsilon > 0$. Choose arbitrary δ and consider (30.4). If $t \neq t_i$ for any i, then $w'_x(\delta) \geq \varepsilon$ by definition. Also, if $t = t_i$ for some i then $t_i \in [t_i, t_{i+1})$ and $|x(t_i) - x(t_i+)| \geq \varepsilon$ and again $w'_x(\delta) \geq \varepsilon$. ∎

An alternative modulus of continuity that may in certain situations be more convenient to work with is

$$w''_x(\delta) = \sup_{t \in [t_1, t_2], t_2 - t_1 \leq \delta} \min\{|x(t) - x(t_1)|, |x(t_2) - x(t)|\}. \tag{30.7}$$

Every interval of the line of width not exceeding δ is divided into two at each possible point and the smaller of the two changes in the function over the subintervals is supped. If an interval contains a jump this can always appear in the 'other' subinterval by choice of t, so the function can jump without affecting $w''_x(\delta)$. Such jumps can only be at interior points of $[0, 1]$ since $x(0) = x(0+)$ and $x(1) = x(1-)$ by Definition **30.1**. The following result parallels **30.3**.

30.4 Theorem $\lim_{\delta \to 0} w''_x(\delta) = 0$ iff $x \in D$.

Proof Set $\delta = 1/k$ for $k = 1, 2, \ldots$ and let $\{t_{1k}, t_k, t_{2k}\}$ denote the points at which the supremum in (30.7) is attained at step k. If $x \in D$, $t_{1k} > 0$ and $t_{2k} < 1$ when k is large enough, thanks to right-continuity at 0 and left-continuity at 1. Suppose

$t_k \to t$ as $k \to \infty$, considering a convergent subsequence if necessary. Then, $t_{1k} \to t$ and $t_{2k} \to t$. Since $x(t) = x(t+)$ for every $t < 1$, this implies $|x(t_{2k}) - x(t_k)| \to 0$, proving sufficiency.

Now suppose $w''_x(\delta) \to w''_x(0) > 0$. There then exists a $t \in (0,1)$ such that

$$w''_x(0) = \min\{|x(t) - x(t-)|, |x(t) - x(t+)|\} > 0.$$

It follows that $x(t)$ is an isolated discontinuity and $x \notin D$, proving necessity. ∎

To find the relationship between w'_x and w''_x consider an interval of Π_δ in (30.4), say $s_i - s_{i-1} \geq \delta$. The intervals $[t_1, t_2]$ defining w''_x, of width at most δ, either fall inside $[s_{i-1}, s_i)$ for some i or straddle adjacent intervals $[s_{i-1}, s_i)$ and $[s_i, s_{i+1})$. In the first case $w_x[s_{i-1}, s_i) \geq |x(t_2) - x(t_1)|$, while in the second case $w_x[s_{i-1}, s_i) \geq |x(t) - x(t_1)|$ for $t_1 \leq t < s_i$ and $w_x[s_i, s_{i+1}) \geq |x(t_2) - x(t)|$ for $s_i \leq t \leq t_2$. Since $w'_x(\delta) \geq w_x[s_{i-1}, s_i)$ for every i, it follows that

$$w''_x(\delta) \leq w'_x(\delta). \tag{30.8}$$

However, the reverse inequality is not in general true since by construction w''_x only constrains jumps over intervals with interior bounds. For example, if x_k is defined for $t \in [0,1]$ by

$$x_k(t) = 1_{\{t<1/k\}}(t) \tag{30.9}$$

or by

$$x_k(t) = 1_{\{t>1-1/k\}}(t) \tag{30.10}$$

then $w'_{x_k}(\delta) = 1$ when $k > 1/\delta$ whereas $w''_{x_k}(\delta) = 0$. While noting that the elements $\lim_{k\to\infty} x_k$ have discontinuities at 0 and 1 and hence are not elements of D, cases such as (30.9) and (30.10) for finite k need to be ruled out by additional restrictions. Adopting the notation of (30.3) let these restrictions be bounds on $w_x[0,\delta)$ and $w_x[1-\delta, 1)$. Write

$$w^*_x(\delta) = \max\{w''_x(\delta),\ w_x[0,\delta),\ w_x[1-\delta,1)\}. \tag{30.11}$$

Then, the following can be shown:

30.5 Lemma If $w^*_x(\delta) < \varepsilon$ then $w'_x(\delta/2) < 6\varepsilon$.

Proof Consider points $0 \leq t_1 \leq s \leq t \leq t_2 < 1$ where $t_2 - t_1 \leq \delta$, and suppose $|x(s) - x(t_1)| > \varepsilon$. Then the condition $w''_x(\delta) < \varepsilon$ implies that $|x(t) - x(s)| < \varepsilon$ and also that $|x(t_2) - x(s)| < \varepsilon$. Therefore, $|x(t_2) - x(t)| > 2\varepsilon$ is impossible and

$$\min\{|x(s) - x(t_1)|, |x(t_2) - x(t)|\} \le 2\varepsilon. \tag{30.12}$$

The argument is symmetric in the case $|x(t_2) - x(t)| > \varepsilon$. With $w''_x(\delta) < \varepsilon$ there cannot be two jumps, both of size 2ε or greater, falling within an interval of width δ.

By similar reasoning, if $w_x[0,\delta) < \varepsilon$ such that $|x(t) - x(s)| < \varepsilon$ whenever $0 \le s \le t \le \delta$, then $|x(t) - x(0)| < \varepsilon$ by construction and $w''_x(\delta) < \varepsilon$ implies $|x(\delta) - x(t)| < \varepsilon$ also, hence $|x(\delta) - x(0)| < 2\varepsilon$. Likewise, $|x(1-\delta) - x(1-)| < 2\varepsilon$.

It is therefore possible to construct a partition $\Pi_{\delta/2} = \{s_1, \dots, s_r\}$ of $[0,1]$ with $\frac{1}{2}\delta \le s_i - s_{i-1} \le \delta$ where any jump exceeding 2ε occurs at a point s_i for some $1 \le i \le r-1$. For such an i let $s_{i-1} \le t_1 < t_2 < s_i$ so that $t_2 - t_1 < \delta$. Let $v_1 \ge t_1$ be the largest $v \le t_2$ such that $\sup_{t_1 \le u \le v} |x(u) - x(t_1)| \le 2\varepsilon$ and let $v_2 \le t_2$ be the smallest $v \ge t_1$ such that $\sup_{v \le u \le t_2} |x(t_2) - x(u)| \le 2\varepsilon$. Suppose $v_1 < v_2$. This would imply a jump exceeding 2ε at a point above v_1 and also below v_2, which is ruled out by (30.12) and the assumed jump at s_i. Therefore, $v_1 \ge v_2$. By construction, $|x(v_1-) - x(t_1)| \le 2\varepsilon$ and likewise $|x(t_2) - x(v_1)| \le |x(t_2) - x(v_2)| \le 2\varepsilon$. A jump at v_1 is not ruled out but also cannot exceed 2ε. Hence,

$$|x(t_2) - x(t_1)| \le |x(v_1-) - x(t_1)| + |x(v_1) - x(v_1-)| + |x(t_2) - x(v_1)| \le 6\varepsilon$$

and by definition of v_1 there is no alternative point of the interval giving a larger value. These relations hold for any choice of t_1 and t_2 and therefore $\sup_{s_{i-1} \le t_1 < t_2 \le s_i} |x(t_2) - x(t_1)| \le 6\varepsilon$. This is also true for any i and for any such partition $\Pi_{\delta/2}$. Hence, $w'_x(\delta/2) \le 6\varepsilon$. ∎

The relationship between the two moduli of continuity definitions can now be summarily expressed as follows.

30.6 Lemma For $x \in D$ and any $\varepsilon > 0$, if there exists $\delta > 0$ so that $w''_x(\delta) < \varepsilon$ then there also exists $\delta > 0$ so that $w'_x(\delta) < \varepsilon$.

Proof Because x is right continuous at 0, there exists δ small enough that $|x(\delta) - x(0)| < \varepsilon$. For any $t \le \delta$, suppose $|x(t) - x(0)| > \varepsilon$. If $w''_x(\delta) < \varepsilon$ then $|x(\delta) - x(t)| < \varepsilon$ and hence $|x(t) - x(0)| < 2\varepsilon$. By the same reasoning, if $0 \le s \le t \le \delta$ then $|x(s) - x(0)| < 2\varepsilon$ and so $|x(t) - x(s)| < 4\varepsilon$. Given (30.3) the conclusion is that

$$\{x : |x(\delta) - x(0)| < \varepsilon/4\} \cap \{x : w''_x(\delta) < \varepsilon\} \subseteq \{x : w_x[0,\delta) < \varepsilon\}.$$

By the same argument, since x is right-continuous at $1 - \delta$,

$$\{x : |x(1-\delta) - x(1)| < \varepsilon/4\} \cap \{x : w''_x(\delta) < \varepsilon\} \subseteq \{x : w_x[1-\delta, 1) < \varepsilon\}.$$

It now follows by (30.11) and Lemma **30.5** that

$$\begin{aligned}\{x : w_x''(2\delta) < \varepsilon/6,\ |x(2\delta) - x(0)| < \varepsilon/24,\ |x(1-2\delta) - x(1-)| < \varepsilon/24\}& \\ \subseteq \{x : w_x^*(2\delta) < \varepsilon/6\}& \\ \subseteq \{x : w_x'(\delta) < \varepsilon\}.& \end{aligned} \tag{30.13}$$

For arbitrary $\varepsilon > 0$, right-continuity of x at 0 and left-continuity at 1 and the assumption of the lemma implies that there exists $\delta > 0$ small enough to validate the event defined on the left side of (30.13). ∎

30.2 Metrizing *D*

Recall from **5.27** the difficulty presented by the existence of uncountable discrete sets in (D, d_U) such as the sets of functions

$$x_\theta(t) = \begin{cases} 0, & 0 \le t < \theta \\ 1, & \theta \le t \le 1. \end{cases} \tag{30.14}$$

This is the case of (5.43) with $a = 0$ and $b = 1$. What is needed is a topology in which x_θ and $x_{\theta'}$ are regarded as close when $|\theta - \theta'|$ is small. Skorokhod ([170]) devised a metric conferring this property.

Let Λ denote the collection of all homeomorphisms $\lambda : [0,1] \mapsto [0,1]$ with $\lambda(0) = 0$ and $\lambda(1) = 1$; think of these as the set of increasing graphs connecting the opposite corners of the unit square (see Figure 30.2). The Skorokhod J1 metric is defined as

$$d_S(x,y) = \inf_{\lambda \in \Lambda} \{\varepsilon > 0 : \sup_t |\lambda(t) - t| \le \varepsilon, \sup_t |x(t) - y(\lambda(t))| \le \varepsilon\}. \tag{30.15}$$

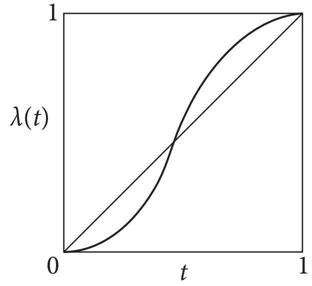

Figure 30.2

In his 1956 paper Skorokhod proposes four metrics, denoted J1, J2, M1, and M2, but only the J1 case plays a role here and d_S will be referred to as is customary as 'the' Skorokhod metric.

It is easy to verify that d_S is a metric after noting that $\sup_t |\lambda(t) - t| = \sup_t |t - \lambda^{-1}(t)|$ and $\sup_t |x(t) - y(\lambda(t))| = \sup_t |x(\lambda^{-1}(t)) - y(t)|$, where $\lambda^{-1} \in \Lambda$ if $\lambda \in \Lambda$. While in the uniform metric two functions are close only if their vertical separation is uniformly small, the Skorokhod metric also takes into account the possibility that the *horizontal* separation is small. If x is uniformly close to y except that it jumps slightly before or slightly after y, the functions would be considered close as measured by d_S, if not by d_U.

Consider x_θ in (30.14) and another element $x_{\theta+\delta}$. The uniform distance between these elements is 1, as noted above. To calculate the Skorokhod distance, note that the quantity in braces in (30.15) will be 1 for any λ for which $\lambda(\theta) \neq \theta + \delta$. Confining consideration to the subclass of Λ with $\lambda(\theta) = \theta + \delta$, choose a case where $|\lambda(t) - t| \leq \delta$; for example the graph $\{t, \lambda(t)\}$ obtained by joining the three points $(0,0)$, $(\theta, \theta + \delta)$, and $(1,1)$ with straight lines will fulfil the definition. Hence,

$$d_S(x_\theta, x_{\theta+\delta}) = \delta. \tag{30.16}$$

This distance approaches zero smoothly as $\delta \downarrow 0$, which conforms better to the intuitive idea of 'proximity' than the uniform metric in these circumstances. Figure 30.3 illustrates, although note that $x_{\theta+\delta}$ has been displaced slightly upwards in the interests of clarity.

30.7 Theorem On C, d_S and d_U are equivalent metrics.

Proof Obviously $d_S(x,y) \leq d_U(x,y)$, since the latter corresponds to the case where λ is the identity function in (30.15), so for any $\varepsilon > 0$, $d_U(x,y) < \varepsilon/2$ implies $d_S(x,y) < \varepsilon/2 < \varepsilon$. On the other hand, for any $\lambda \in \Lambda$,

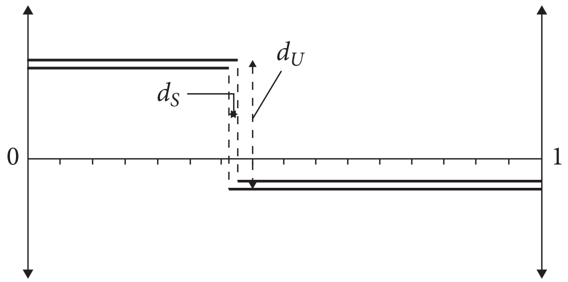

Figure 30.3

$$d_U(x,y) \leq \sup_t |x(t) - y(\lambda(t))| + \sup_t |y(\lambda(t)) - y(t)|. \tag{30.17}$$

If y is uniformly continuous, for every $\eta > 0$ there must exist $\delta > 0$ small enough that if $d_S(x,y) < \delta$ and hence $\sup_t |\lambda(t) - t| < \delta$ then $\sup_t |y(\lambda(t)) - y(t)| < \eta$. For the case $\eta = \varepsilon/2$, (30.17) implies that $d_U(x,y) < \varepsilon$ when $d_S(x,y) < \varepsilon/2$. Thus, the criteria of (5.6) and (5.7) are satisfied with $\delta = \varepsilon/2$. Uniform continuity is equivalent to continuity on [0,1] and so the stated inequalities hold for all $y \in C$. ∎

The following result helps to explain the interest in the Skorokhod metric.

30.8 Theorem (D, d_S) is separable.

Proof As usual, this is shown by exhibiting a countable dense subset. The counterpart in D of the piecewise linear function defined for C is the piecewise constant function (as in Figure 30.1) defined as

$$y(t) = y(t_i),\ t \in [t_i, t_{i+1}),\ i = 0, \ldots, m-1 \tag{30.18}$$

where the $y(t_i)$ are specified real numbers. For some $n \in \mathbb{N}$ define the set A_n as the countable collection of the piecewise constant functions of form (30.18), with $t_i = i/2^n$ for $i = 0, \ldots, 2^n - 1$, and $y(t_i)$ assuming *rational* values for each i. Letting A denote the limit of the sequence $\{A_n\}$, A is a set of functions taking rational values at a set of points indexed on the dyadic rationals $\mathbb{D}$ and hence is countable by **1.5**.

According to **30.2** there exists for $x \in D$ and $\varepsilon > 0$ a finite partition $\{t_1, \ldots, t_m\}$ of $[0,1]$ such that, for each i,

$$\sup_{s,t \in [t_i, t_{i+1})} |x(s) - x(t)| < \varepsilon.$$

Let y be a piecewise constant function constructed on the same intervals, assuming rational values $y_1, \ldots, y_m$ where y_i differs by no more than ε from a value assumed by x on $[t_i, t_{i+1})$. Then $d_S(x,y) < 2\varepsilon$. Now, given $n \geq 1$, choose $z \in A_n$ such that $z_j = y_i$ when $j/2^n \in [t_i, t_{i+1})$. Since $\mathbb{D}$ is dense in $[0,1]$, $d_S(y,z) \to 0$ as $n \to \infty$. Hence, $d_S(x,z) \leq d_S(x,y) + d_S(y,z)$ is tending to a value not exceeding 2ε. Since by taking m large enough ε can be made as small as desired, x is a closure point of A. Since x is arbitrary it follows that A is dense in D. ∎

Notice how this argument would fail under the uniform metric in the cases where x has discontinuities at one or more of the points t_i. Then, $d_U(y,z)$ will be small only if the two sets of intervals overlap precisely, such that $t_i = j/2^n$ for some j.

If t_i were irrational, this would not occur for any finite m, since $j/2^n$ is rational. Under these circumstances x would fail to be a closure point of A. This shows why the Skorokhod topology (that is, the topology induced by the Skorokhod metric) is needed to ensure separability.

Working with d_S nonetheless complicates matters somewhat. For one thing, d_S does not generate the Tychonoff topology and the coordinate projections are not in general continuous mappings. The fact that x and y are close in the Skorokhod metric does not imply that $x(t)$ is close to $y(t)$ for every t, the examples of x_θ and $x_{\theta+\delta}$ cited above being a case in point. What can be said is that projections are continuous mappings at and only at points of continuity of the process. If $\{x_n \in D\}$ is a sequence that converges to x in the Skorokhod topology, then according to (30.15) there must exist a sequence $\{\lambda_n \in \Lambda\}$ with $\lambda_n(t) \to t$ uniformly in t. Noting that $|x_n(t) - x(\lambda_n(t))| \leq |x_n(t) - x(t)| + |x(t) - x(\lambda_n(t))|$, if t is a continuity point of x so that the second of the majorant terms goes to zero as $n \to \infty$, it must be the case that $x_n(t) \to x(t)$. On the other hand, if $x(t) \neq x(t-)$ then if $x(\lambda_n(t)) \to x(t-)$, as is possible, the projection at t fails to converge.

A further more critical problem is that the space (D, d_S) is not complete. This is easily seen by considering the sequence of elements $\{x_n\}$ where

$$x_n(t) = \begin{cases} 1, & t \in [\frac{1}{2}, \frac{1}{2} + 1/n) \\ 0, & \text{otherwise} \end{cases} \tag{30.19}$$

(see Figure 30.4). The limit of this sequence is a function having an isolated point of discontinuity at $\frac{1}{2}$ and hence is not in D. However, to calculate $d_S(x_n, x_m)$, λ must be chosen so that $\lambda(\frac{1}{2}) = \frac{1}{2}$ and $\lambda(\frac{1}{2} + 1/n) = \frac{1}{2} + 1/m$. The distance is 1 for any other choice. The piecewise-linear graph with vertices at $(0,0)$, $(\frac{1}{2}, \frac{1}{2})$, $(\frac{1}{2} + 1/n, \frac{1}{2} + 1/m)$, and $(1,1)$ fulfils the definition and satisfies (30.15). It follows that $d_S(x_n, x_m) = |1/n - 1/m|$ and so $\{x_n\}$ is a Cauchy sequence.

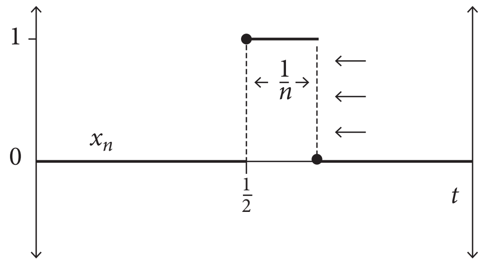

Figure 30.4

30.3 Billingsley's Metric

The solution to the completeness problem is to devise a metric that is equivalent to d_S (in the sense of generating the same topology and hence a separable space) but in which sequences such as the one in (30.19) are not Cauchy sequences. The following construction is due to Billingsley ([21], [22]), from which sources the results of this section and the next are adapted.

Let Λ be the collection of homeomorphisms λ from $[0,1]$ to $[0,1]$ with $\lambda(0)=0$ and $\lambda(1)=1$ and satisfying $\|\lambda\|<\infty$ where

$$\|\lambda\| = \sup_{t\neq s}\left|\log\frac{\lambda(t)-\lambda(s)}{t-s}\right|. \tag{30.20}$$

Here, $\|\lambda\| : \Lambda \mapsto \mathbb{R}^+$ is a functional measuring the maximum deviation of the gradient of λ from 1, so that in particular $\|\lambda\|=0$ for the case $\lambda(t)=t$. The set Λ is like the one defined for the Skorokhod metric with the added proviso that $\|\lambda\|$ be finite; both λ and λ^{-1} must be strictly increasing functions. Then, what will be called the 'Billingsley metric' is defined as

$$d_B(x,y) = \inf_{\lambda\in\Lambda}\left\{\varepsilon>0 : \|\lambda\|\leq\varepsilon,\ \sup_t|x(t)-y(\lambda(t))|\leq\varepsilon\right\}. \tag{30.21}$$

Here are the essential properties of d_B.

30.9 Theorem d_B is a metric.

Proof $d_B(x,y)=0$ iff $x=y$ is immediate. $d_B(x,y)=d_B(y,x)$ is also easy once it is noted that $\|\lambda^{-1}\|=\|\lambda\|$. To show the triangle inequality, define the composite mapping $\lambda_1\circ\lambda_2(t)=\lambda_1(\lambda_2(t))$ which is an element of Λ. Then note that

$$\begin{aligned}\|\lambda_1\circ\lambda_2\| &= \sup_{t\neq s}\left|\log\frac{\lambda_1(\lambda_2(t))-\lambda_1(\lambda_2(s))}{t-s}\right|\\ &\leq \sup_{t\neq s,t'\neq s'}\left|\log\frac{(\lambda_1(t')-\lambda_1(s'))(\lambda_2(t)-\lambda_2(s))}{(t'-s')(t-s)}\right|\\ &\leq \sup_{t\neq s,t'\neq s'}\left\{\left|\log\frac{\lambda_1(t')-\lambda_1(s')}{t'-s'}\right|+\left|\log\frac{\lambda_2(t)-\lambda_2(s)}{t-s}\right|\right\}\\ &= \|\lambda_1\|+\|\lambda_2\|.\end{aligned} \tag{30.22}$$

By the ordinary triangle inequality,

$$\begin{aligned}\sup_t |x(t) - z(\lambda_1 \circ \lambda_2(t))| &\le \sup_t\{|x(t) - y(\lambda_1(t))| + |y(\lambda_1(t)) - z(\lambda_1 \circ \lambda_2(t))|\} \\ &\le \sup_t |x(t) - y(\lambda_1(t))| + \sup_t |y(t) - z(\lambda_2(t))|. \qquad (30.23)\end{aligned}$$

It follows from definition (30.21) that $d_B(x,z) \le d_B(x,y) + d_B(y,z)$. ∎

d_S and d_B are equivalent metrics and hence confer the same topology on D. Inequalities going in both directions can be derived provided the distances are sufficiently small. To show this, first consider functions x and y for which $d_B(x,y) = \varepsilon \le \frac{1}{2}$ and suppose $\lambda \in \Lambda$ satisfies the definition of d_B for this pair such that $||\lambda|| \le \varepsilon$. Since $\lambda(0) = 0$, there evidently must exist $t \in (0,1]$ such that $|\log(\lambda(t)/t)| \le ||\lambda||$, or

$$te^{-\varepsilon} \le \lambda(t) \le te^{\varepsilon}. \qquad (30.24)$$

Noting that $e^{\varepsilon} - 1 \le 2\varepsilon$ and $e^{-\varepsilon} - 1 \ge -2\varepsilon$ when $\varepsilon \le \frac{1}{2}$, it follows that $-2\varepsilon \le t(e^{-\varepsilon} - 1) \le \lambda(t) - t \le t(e^{\varepsilon} - 1) \le 2\varepsilon$, or $|\lambda(t) - t| \le 2\varepsilon$. Since $\sup_t |x(t) - y(\lambda(t))| \le \varepsilon$ for the specified λ, it follows that $d_S(x,y)$ cannot exceed 2ε. In other words,

$$d_S(x,y) \le 2d_B(x,y) \qquad (30.25)$$

whenever $d_B(x,y) \le \frac{1}{2}$.

To go the other way, consider a function $\mu \in \Lambda$ which is piecewise-linear with vertices at the points of a partition Π_δ, as defined above (30.4), for $0 < \delta \le \frac{1}{4}$. If $\sup_t |\mu(t) - t| \le \delta^2$, then

$$\left|\frac{\mu(t_i) - \mu(t_{i-1})}{t_i - t_{i-1}} - 1\right| \le \frac{|\mu(t_i) - t_i| + |\mu(t_{i-1}) - t_{i-1}|}{\delta} \le 2\delta \le \tfrac{1}{2}. \qquad (30.26)$$

Since $|\log(1 + x)| \le 2|x|$ if $|x| \le \frac{1}{2}$, it follows letting $|x|$ be the minorant side of (30.26) that

$$||\mu|| \le 4\delta. \qquad (30.27)$$

Now suppose that $d_S(x,y) \le \delta^2$. This means there exists $\lambda \in \Lambda$ satisfying $\sup_t\{\lambda(t) - t\} \le \delta^2$ and $\sup_t\{x(t) - y(\lambda(t))\} \le \delta^2$, the latter inequality being equivalent to $\sup_t\{x(\lambda^{-1} \circ \mu(t)) - y(\mu(t))\} \le \delta^2$. Choose μ to have $\mu(t_i) = \lambda(t_i)$ for $i = 0, \ldots, r$, so that the function $\lambda^{-1} \circ \mu$ is 'tied down' to the diagonal at the points of the partition, with $\lambda^{-1} \circ \mu(t) \in [t_{i-1}, t_i)$ if and only if $t \in [t_{i-1}, t_i)$. Therefore, choosing Π_δ to correspond to the definition of $w'_x(\delta)$,

$$\begin{aligned}|x(t)-y(\mu(t))| &\le |x(t)-x(\lambda^{-1}\circ\mu(t))|+|x(\lambda^{-1}\circ\mu(t))-y(\mu(t))|\\ &\le w'_x(\delta)+\delta^2. \end{aligned} \tag{30.28}$$

Putting this together with (30.27) gives the inequality

$$d_B(x,y)\le\max\{4\delta, w'_x(\delta)+\delta^2\}\le w'_x(\delta)+4\delta. \tag{30.29}$$

For $x\in D$, $w'_x(\delta)$ can be made arbitrarily small by choice of δ, and hence

$$d_B(x,y)\le 4d_S(x,y)^{1/2} \tag{30.30}$$

whenever $d_S(x,y)$ is sufficiently small. These arguments lead to the following conclusion.

30.10 Theorem In D, metrics d_B and d_S are equivalent.

Proof Given $\varepsilon>0$, choose $\delta\le\frac{1}{4}$ and also small enough that $w'_x(\delta)+4\delta\le\varepsilon$. Then, for $\eta\le\min\{\delta^2,\frac{1}{2}\varepsilon\}$,

$$d_B(x,y)<\eta\Rightarrow d_S(x,y)<\varepsilon \tag{30.31}$$

$$d_S(x,y)<\eta\Rightarrow d_B(x,y)<\varepsilon \tag{30.32}$$

by (30.25) and (30.30) respectively. The criteria of (5.6) and (5.7) are therefore satisfied. ∎

Given a sequence of elements $\{x_n\}$, $d_B(x_n,x)\to 0$ if and only if $d_S(x_n,x)\to 0$, whenever $x\in D$. However, the common topology the two metrics induce on D (the Skorokhod topology) does *not* imply that $\{x_n\}$ is a Cauchy sequence in (D,d_B) whenever it is a Cauchy sequence in (D,d_S). For example, the sequence of functions in (30.19) is not a Cauchy sequence in (D,d_B). To define $d_B(x_n,x_m)$ (for $n\ge 3, m\ge 4$) it is necessary to find an element of Λ for which $\lambda(\frac{1}{2})=\frac{1}{2}$ and $\lambda(\frac{1}{2}+1/n)=\frac{1}{2}+1/m$. $||\lambda||$ is minimized by the same piecewise-linear function defined for d_S with vertices at the points $(0,0)$, $(\frac{1}{2}+1/n,\frac{1}{2}+1/m)$, and $(1,1)$. However, $||\lambda||\ge|\log(n/m)|$, which does not necessarily approach zero as n and m increase; consider for example $m=2n$.

30.11 Theorem The space (D,d_B) is complete.

Proof Let $\{y_k,k\in\mathbb{N}\}$ be a Cauchy sequence in (D,d_B) satisfying $d_B(y_k,y_{k+1})<1/2^k$, implying the existence of a sequence of functions $\{\mu_k\in\Lambda\}$ with $||\mu_k||<1/2^k$ and

$$\sup_t |y_k(t) - y_{k+1}(\mu_k(t))| < 1/2^k. \tag{30.33}$$

It follows from (30.25) that $\sup_t |\mu_{k+m}(t) - t| \leq 2/2^{k+m}$ for $m > 0$. Define $\mu_{km} = \mu_{k+m} \circ \mu_{k+m-1} \circ \cdots \circ \mu_k$, which is also an element of Λ for each finite m; the sequence $\{\mu_{km}, m = 1, 2, \ldots\}$ is a Cauchy sequence in (C, d_U) because

$$\sup_t |\mu_{k,m+1}(t) - \mu_{km}(t)| = \sup_s |\mu_{k+m+1}(s) - s| \leq 1/2^{k+m}. \tag{30.34}$$

Since (C, d_U) is complete there exists a limit function $\lambda_k = \lim_{m\to\infty} \mu_{km}$. To show that $\lambda_k \in \Lambda$, it is sufficient to show that $||\lambda_k|| < \infty$. But by (30.22),

$$||\mu_{km}|| \leq ||\mu_k \circ \mu_{k+1} \circ \ldots \circ \mu_{k+m}|| \leq \sum_{j=0}^{m} ||\mu_{k+j}|| < \sum_{j=0}^{m} \frac{1}{2^{k+j}} \leq \frac{1}{2^{k-1}} \tag{30.35}$$

for any m, so $||\lambda_k|| \leq 1/2^{k-1}$.

Note that $\lambda_k = \lambda_{k+1} \circ \mu_k$, so that $\lambda_{k+1}^{-1} = \mu_k \circ \lambda_k^{-1}$ and hence, by (30.33),

$$\sup_t |y_k(\lambda_k^{-1}(t)) - y_{k+1}(\lambda_{k+1}^{-1}(t))| = \sup_s |y_k(s) - y_{k+1}(\mu_k(s))| < 1/2^k. \tag{30.36}$$

Defining $y_k \circ \lambda_k^{-1}(t) = y_k(\lambda_k^{-1}(t))$, consider the sequence $\{y_k \circ \lambda_k^{-1} \in D, k \in \mathbb{N}\}$. According to (30.36) this is a Cauchy sequence in (D, d_U). Therefore, for each $t \in [0, 1]$, $|y_k(\lambda_k^{-1}(t)) - y_{k+1}(\lambda_{k+1}^{-1}(t))| < 1/2^k$ which defines a Cauchy sequence in $\mathbb{R}$, having a real limit $y(t)$. It follows that $y_k \circ \lambda_k^{-1}$ converges to a limit $y \in D$ uniformly in $[0, 1]$. This implies that $\sup_t |y_k(t) - y(\lambda_k(t))| = \sup_t |y_k(\lambda_k^{-1}(t)) - y(t)| \to 0$ and also that $||\lambda_k|| = ||\lambda_k^{-1}|| \to 0$, which together imply that $d_B(y_k, y) \to 0$. In other words $\{y_k\}$ has a limit y in (D, d_B).

The assumption that $\{y_k\}$ is a Cauchy sequence with $d_B(y_k, y_{k+1}) < 1/2^k$ involves no loss of generality, because it suffices to show that any Cauchy sequence $\{x_n, n \in \mathbb{N}\}$ contains a convergent subsequence $\{y_k = x_{n_k}, k \in \mathbb{N}\}$. A cluster point of a Cauchy sequence is necessarily a limit point. If $d_B(x_n, x_{n+1}) < 1/g(n) \to 0$ for some increasing function g, choosing $n_k \geq g^{-1}(2^k)$ defines a subsequence with the required property. ∎

30.4 Measures on D

Henceforth D denotes (D, d_B), the metric being specified explicitly only if different from d_B, and the Borel field of (D, d_B) is denoted $\mathcal{B}_D$. The basic property the measurable space $(D, \mathcal{B}_D)$ must possess is that measures can be fully specified by the finite-dimensional sets. An argument analogous to **27.15** is called for, although some elaboration will be needed. In particular it must be shown without appealing to continuity of the projections that the finite-dimensional distributions are well-defined and that there are finite-dimensional sets that constitute a determining class for $(D, \mathcal{B}_D)$.

Begin with a lemma. Define the field of finite-dimensional sets of D as $\mathcal{H}_D = \{H \cap D : H \in \mathcal{H}\}$, where $\mathcal{H}$ was defined in (27.14).

30.12 Lemma Given $x \in D$, $\alpha > 0$, and any $t_1, \ldots, t_m \in [0,1]$, let

$$H_m(x,\alpha) = \{y \in D : \exists\, \lambda \in \Lambda \text{ s.t. } \|\lambda\| < \alpha,\ \max_{1\le i\le m} |y(t_i) - x(\lambda(t_i))| < \alpha\}. \tag{30.37}$$

Then $H_m(x,\alpha) \in \mathcal{H}_D$.

Proof Since $H_m(x,\alpha) \subseteq D$, what has to be shown according to (27.14) is that $\pi_{t_1,\ldots,t_m}(H_m(x,\alpha)) \in \mathcal{B}^m$. This is the set whose elements are $(y(t_1), \ldots, y(t_m))$ for each $y \in H_m(x,\alpha)$. To identify these, first define the set

$$A_m(x,\alpha) = \{x(\lambda(t_1)), \ldots, x(\lambda(t_m)) : \lambda \in \Lambda,\ \|\lambda\| < \alpha\} \subseteq \mathbb{R}^m. \tag{30.38}$$

Then it is apparent that

$$\begin{aligned} &\pi_{t_1,\ldots,t_m}(H_m(x,\alpha)) \\ &= \{b_1, \ldots, b_m : \max_{1\le i\le m} |a_i - b_i| < \alpha,\ (a_1, \ldots, a_m) \in A_m(x,\alpha)\} \subseteq \mathbb{R}^m. \end{aligned} \tag{30.39}$$

In words, this is the set $A_m(x,\alpha)$ with an open α-halo and it is an open set. It therefore belongs to $\mathcal{B}^m$. ∎

To compare the present situation with that for C described in §27.5, it may be helpful to look at the case $m = 1$. Consider the one-dimensional projection $\pi_t(H_t(x,\alpha))$ where

$$H_t(x,\alpha) = \{y \in D : \exists\, \lambda \in \Lambda \text{ s.t. } \|\lambda\| < \alpha,\ |y(t) - x(\lambda(t))| < \alpha\}. \tag{30.40}$$

Recalling the discussion on page 676, if t is a continuity point of x the difference between $\pi_t(H_t(x,\alpha))$ and $S(x(t),\alpha)$ (the open interval of width 2α centred on $x(t)$) can be made arbitrarily small by taking α small enough and $\pi_t(H_t(x,1/n)) \to x(t)$ as $n \to \infty$. The exceptional cases of t where this result does not apply are at most countable by **30.2** and hence of Lebesgue measure zero in $[0,1]$. This means that given a distribution for elements of D, the finite-dimensional projections have the status of random vectors. This is shown as follows.

30.13 Theorem The finite-dimensional projections are $\mathcal{B}_D/\mathcal{B}^k$-measurable.

Proof It is sufficient to consider the case $k = 1$ where the object is to show that the projection mapping from elements of D to points of $\mathbb{R}$ is continuous. For $t < 1$, define the mapping $h_{mt}(x) = m\int_t^{t+1/m} x(s)\mathrm{d}s$ for $m > (1-t)^{-1}$. If $x_n \to x$ in the Skorokhod topology then $x_n(s) \to x(s)$ at the points s where x is continuous, everywhere except in a set of Lebesgue measure zero. Therefore, in view of (30.2), $h_{mt}(x_n) \to h_{mt}(x)$ as $n \to \infty$ so h_{mt} is continuous in the Skorokhod topology. Since x is right-continuous it is also the case that $h_{mt}(x) \to \pi_t(x)$ as $m \to \infty$. As the limit of this sequence of continuous mappings, $\pi_t : D \mapsto \mathbb{R}$ is therefore continuous and hence measurable by **3.39**(i). Since $t = 1$ is always a continuity point, this completes the proof. ∎

To reconcile this result with the previously noted failure of the projections to converge at points of discontinuity, note how this mapping is constructed always to identify $x(t)$ and not $x(t-)$ with the limit of the sequence. This makes sense, given that in a random drawing x the value observed at point t is $x(t)$, not $x(t-)$ should this be different.

The next key result is to validate the extension of a measure from the finite-dimensional distributions to $(D, \mathcal{B}_D)$. It is easily verified that $\mathcal{H}_D$, like $\mathcal{H}$, is a field. The counterpart for D of Theorem **27.15** for the case of C is to show that $\mathcal{H}_D$ is a determining class for $(D, \mathcal{B}_D)$.

30.14 Theorem $\mathcal{B}_D = \sigma(\mathcal{H}_D)$.

Proof An open sphere in (D, d_B) is a set of the form

$$\begin{aligned} S(x,\alpha) &= \{y \in D : d_B(y,x) < \alpha\} \\ &= \{y \in D : \exists\, \lambda \in \Lambda \text{ s.t.} \|\lambda\| < \alpha,\ \sup_t |y(t) - x(\lambda(t))| < \alpha\} \end{aligned} \tag{30.41}$$

for $x \in D$, $\alpha > 0$. Since these sets generate $\mathcal{B}_D$, it will suffice to show they can be constructed by countable unions and complements (and hence also countable intersections) of sets in $\mathcal{H}_D$. Let $H(x,\alpha) = \bigcap_{k=1}^{\infty} H_k(x,\alpha)$ where $H_k(x,\alpha)$ is a set with the form of H_m defined in (30.37), lying in $\mathcal{H}_D$ by **30.12** and with $m = 2^k - 1$ and $t_i = i/2^k$ so that the set $\{t_1, \ldots, t_{2^k-1}\}$ converges on $\mathbb{D}$ (the dyadic rationals) as $k \to \infty$. Consider $y \in H(x,\alpha)$. Since $y \in H_k(x,\alpha)$ for every k choose a sequence $\{\lambda_k\}$ such that, for each $k \geq 1$,

$$\|\lambda_k\| < \alpha \tag{30.42}$$

and

$$\max_{1\le i\le 2^k-1}\left|y(2^{-k}i)-x(\lambda_k(2^{-k}i))\right|<\alpha. \tag{30.43}$$

Making use of the fortuitous fact that λ_k has the properties of a c.d.f. on [0,1], Helly's theorem (**23.21**) may be applied to show that there is a subsequence $\{\lambda_{k_n}, n\in\mathbb{N}\}$ converging to a limit function λ which is increasing on [0,1] with $\lambda(0)=0$ and $\lambda(1)=1$. λ is necessarily in Λ, satisfying

$$\|\lambda\|\le\alpha \tag{30.44}$$

according to (30.42). In view of (30.43) and the facts that $\lambda_k(t)\to\lambda(t)$ and x is right-continuous on [0,1], it must also satisfy either $|y(t)-x(\lambda(t))|\le\alpha$ or $|y(t)-x(\lambda(t)-)|\le\alpha$ for every $t\in\mathbb{D}$. Since $\mathbb{D}$ is dense in [0,1], this is equivalent to

$$\sup_t\left|y(t)-x(\lambda(t))\right|\le\alpha. \tag{30.45}$$

The limiting inequalities (30.44) and (30.45) cannot be relied on to be strict but comparing with (30.41), $y\in\bar{S}(x,\alpha)$. Since $y\in H(x,\alpha)$, by assumption it follows that $H(x,\alpha)\subseteq\bar{S}(x,\alpha)$. Put $\alpha=r-1/n$ and take the countable union to give

$$\bigcup_{n=1}^{\infty}H(x,r-1/n)\subseteq\bigcup_{n=1}^{\infty}\bar{S}(x,r-1/n)=S(x,r). \tag{30.46}$$

It is also evident on comparing (30.37) with (30.41) that $S(x,\alpha)\subseteq H_k(x,\alpha)$ for $\alpha>0$. Again, put $\alpha=r-1/n$ and

$$S(x,r)=\bigcup_{n=1}^{\infty}S(x,r-1/n)\subseteq\bigcup_{n=1}^{\infty}H(x,r-1/n). \tag{30.47}$$

It follows that for any $x\in D$ and $r>0$, $S(x,r)=\bigcup_{n=1}^{\infty}\bigcap_{k=1}^{\infty}H_k(x,r-1/n)$ where $H_k(x,r-1/n)\in\mathcal{H}_D$. This completes the proof. ∎

The defining of measures on $(D,\mathcal{B}_D)$ is now possible by arguments that broadly parallel those for C. There is one caveat to keep in mind. In a stochastic context it is important to distinguish between discontinuity points of a càdlàg process that may arise in a systematic manner from those that are purely random. As an example of the former case consider the empirical step function (29.67), where the points $t=j/n$ for $j=1,\dots,n$ are jump points for every finite n, by construction. These would be problematic if they were to persist in the limit process, although fortunately they shrink to zero under the required normalization. When considering the finite-dimensional distributions it will be necessary to exclude from consideration points of $[0,1]$ where jumps occur with positive probability.

This issue is formalized in the analysis of §30.7 and an important requisite is the following corollary of **30.14**. Letting $T \subseteq [0,1]$ be a set whose complement is countable, define the class of finite-dimensional sets

$$\mathcal{H}_{DT} = \{\pi^{-1}_{t_1,\dots,t_m}(B) \subseteq D_{[0,1]} : B \in \mathcal{B}^m, t_1, \dots, t_m \in T, m \in \mathbb{N}\}. \tag{30.48}$$

30.15 Corollary $\mathcal{B}_D = \sigma(\mathcal{H}_{DT})$.

Proof In the proof of **30.14** let $\mathcal{H}_{DT}$ replace $\mathcal{H}_D$ so that, in particular, $t_i = 2^{-k}i$ for $i = 1, \dots, m$ and $m = 2^k - 1$, if $2^{-k}i \in T$. Otherwise, t_i is omitted from the set and $m < 2^k - 1$ accordingly. Letting $k \to \infty$, the collections $\{t_1, \dots, t_m\}$ converge on $\mathbb{D} \cap T$. Since T^c is countable by assumption, $\mathbb{D} \cap T$ is dense in $[0,1] \cap T$ and hence in $[0,1]$ by **1.39** and the argument leading to (30.45) holds for $\mathbb{D} \cap T$ in the same way as for $\mathbb{D}$. The proof therefore goes through as for **30.14**. ∎

There is no presumption here that the elements of T^c are rational numbers. In applications $\mathbb{D} \cap T = \mathbb{D}$ is the most probable case.

30.5 Prokhorov's Metric

The material of this section is not essential to the development and can be omitted. Billingsley's metric is all that is needed to work successfully with càdlàg functions, but it is nonetheless interesting to compare it with the alternative approach due to Prokhorov ([153]). Define for $z \in \mathbb{R}$ the function

$$\bar{w}_x(z) = \begin{cases} w_x^*(\mathrm{e}^z+), & z < 0, \\ w_x^*(1), & z \geq 0. \end{cases} \tag{30.49}$$

where w_x^* is defined in (30.11). This is non-decreasing, right-continuous, and bounded below by 0 and above by $w_x^*(1)$. It therefore defines a finite measure on $\mathbb{R}$, just as a c.d.f. defines a p.m. on $\mathbb{R}$. By defining a family of measures in this way (indexed by x) on a separable space, the fact explored in §29.3 that a space of measures is metrizable can be exploited using Lévy's metric L^* in (29.33). The space (D, d_P) is endowed with the Prokhorov metric

$$d_P(x,y) = d_H(\Gamma_x, \Gamma_y) + L^*(\bar{w}_x, \bar{w}_y) \tag{30.50}$$

where Γ_x and Γ_y are the graphs of x and y and d_H is the Hausdorff metric.

The idea here should be clear. The first term alone confers a property similar to that of the Skorokhod metric; if $d(x(t), \Gamma_y) = \inf_{t'} d_E(x(t), y(t'))$ then

$$d_H(\Gamma_x, \Gamma_y) = \max\left\{\sup_t d(x(t), \Gamma_y), \sup_{t'} d(\Gamma_x, y(t'))\right\}. \tag{30.51}$$

In words, the smallest Euclidean distances between $x(t)$ and a point of y and $y(t)$ and a point of x are supped over t. For comparison, the Skorokhod metric minimizes the greater of the horizontal and vertical distances separating points on Γ_x and Γ_y in the plane, subject to the constraints imposed on the choice of λ such as continuity. In cases such as the functions x_θ of (30.14), x_θ and $x_{\theta+\delta}$ are close in (D, d_H) when δ is small. (Think in terms of the distances the graphs would have to be moved to fit over one another.)

The purpose of the second term is to ensure completeness. By **30.4**, $\lim_{z\to-\infty} \bar{w}_x(z) = 0$ if and only if $x \in D$; otherwise this limit will be strictly positive and $L^*(\bar{w}_x, \bar{w}_y)$ cannot be close to zero according to (29.33) unless both $x \in D$ and $y \in D$. Therefore, unlike the case of (D, d_H), it is not possible to have a Cauchy sequence in (D, d_P) approaching a point outside the space. It can be shown that d_P is equivalent to d_S and hence of course also to d_B, and that the space (D, d_P) is complete. The proofs of these propositions can be found in ([141]). For practical purposes there is nothing to choose between d_P and d_B.

30.6 Compactness and Tightness in *D*

The next task is to characterize the compact sets of D in parallel with the earlier application of the Arzelà–Ascoli theorem for C.

30.16 Theorem A set $A \subset D$ is relatively compact in (D, d_B) iff

$$\sup_{x\in A} \sup_t |x(t)| < \infty \tag{30.52}$$

and

$$\lim_{\delta\to 0} \sup_{x\in A} w'_x(\delta) = 0. \qquad \square \tag{30.53}$$

This theorem obviously parallels **27.14** but there are significant differences in the conditions. The modulus of continuity w'_x appears in place of w_x which is a weakening of the previous conditions, but on the other hand (30.52) replaces (27.28). It is no longer sufficient to bound the elements at one point of the interval to ensure that they are bounded everywhere: the whole element must be bounded. Replacing $\sup_t|x(t)|$ with $d_B(|x|,0)$ where 0 denotes the element of D which is identically zero everywhere on [0,1] gives an equivalent condition.

A feature of the proof that follows, which is basically similar to that of **5.28**, is that it avoids invoking completeness of the space until, so to speak, the last moment. The sufficiency argument establishing total boundedness of A is couched in terms of the more tractable Skorokhod metric and then the equivalence of d_S with a complete metric such as d_B is exploited to get the compactness of $\bar{A}$. The argument for necessity also uses d_S to prove upper semicontinuity of $w'_x(\delta)$, a property that implies (30.53) when the space is compact.

Proof of 30.16 Let $\sup_{x\in A}\sup_t|x(t)| = M$. To show sufficiency, fix $\varepsilon > 0$ and choose m as the smallest integer such that both $1/m < \frac{1}{2}\varepsilon$ and $\sup_{x\in A} w'_x(1/m) < \frac{1}{2}\varepsilon$. Such an m exists by (30.53). Construct the finite collection E_m of piecewise *constant* functions, whose values at the discontinuity points $t = j/m$ for $j = 0,\dots,m-1$ are drawn from the set $\{M(2u/v-1), u = 0,1,\dots,v\}$ where v is an integer exceeding $2M/\varepsilon$; hence, E_m has $(v+1)^m$ different elements. This set is shown to be an ε-net for A.

Given the definition of m, one can choose for $x \in A$ a partition $\Pi_{1/m} = \{t_1,\dots,t_r\}$, defined as above (30.4) with $\min_i\{t_i - t_{i-1}\} > 1/m$ and to satisfy

$$\max_{1\leq i\leq r}\left\{\sup_{s,t\in[t_{i-1},t_i)}|x(t)-x(s)|\right\} < \tfrac{1}{2}\varepsilon. \tag{30.54}$$

For $i = 0,\dots,r-1$ let j_i be the integer such that $j_i/m \leq t_i < (j_i+1)/m$, noting that, since the t_i are at a distance of more than $1/m$, there is at most one of them in any one of the intervals $[j/m,(j+1)/m)$, $j = 0,\dots,m-1$. Choose a piecewise linear function $\lambda \in \Lambda$ with vertices $\lambda(j_i/m) = t_i$, $i = 0,\dots,r$. Since $|t_i - j_i/m| \leq 1/m < \frac{1}{2}\varepsilon$ it follows that $\max_{0\leq i\leq r}|\lambda(j_i/m) - j_i/m| \leq \frac{1}{2}\varepsilon$, and the linearity of λ between these points then implies

$$\sup_t|\lambda(t) - t| \leq \tfrac{1}{2}\varepsilon. \tag{30.55}$$

By construction, λ maps points in $[j/m,(j+1)/m)$ into $[t_i,t_{i+1})$ whenever $j_i \leq j \leq j_{i+1}$, and since x varies by at most $\frac{1}{2}\varepsilon$ over intervals $[t_i,t_{i+1})$ the composite function $x\circ\lambda$ can vary by at most $\frac{1}{2}\varepsilon$ over intervals $[j/m,(j+1)/m)$. An example of λ with $m = 10$, and $r = 4$ is sketched in Figure 30.5; here, $j_1 = 2$, $j_2 = 4$, and $j_3 = 6$. The points $t_0,\dots,t_4$ must be more than a distance of 1/10 apart in this instance.

One can therefore choose $y \in E_m$ such that

$$|y(j/m) - x(\lambda(j/m))| < \tfrac{1}{2}\varepsilon,\ j = 0,\dots,m-1. \tag{30.56}$$

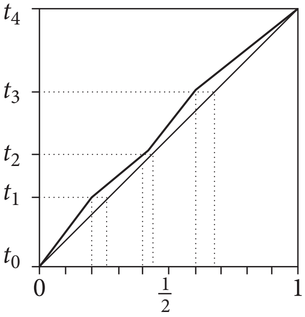

Figure 30.5

Since $y(t) = y(j/m)$ for $t \in \big[j/m, (j+1)/m\big)$, (30.54) and (30.56) imply

$$\sup_t \big|y(t) - x(\lambda(t))\big| \le \max_{0 \le j \le m-1} \Big\{ \big|y(j/m) - x(\lambda(j/m))\big| + \sup_{t \in [j/m,(j+1)/m)} \big|x(\lambda(j/m)) - x(\lambda(t))\big| \Big\} < \varepsilon. \tag{30.57}$$

Together, (30.57) and (30.55) imply $d_S(x,y) \le \varepsilon$, showing that E_m is an ε-net for A as required. This proves that A is totally bounded in (D, d_S). But since d_S and d_B are equivalent by **30.10**, A is also totally bounded in (D, d_B); in particular, if E_m is an ε-net for A in (D, d_S), there exists η such that it is also an η-net for A in (D, d_B) according to (30.31) and (30.32), where η can be set arbitrarily small. Since (D, d_B) is complete $\bar{A}$ is therefore compact by **5.13**, proving sufficiency.

When A is totally bounded it is bounded, proving the necessity of (30.52). The necessity of (30.53) follows because the functions $w'_x(1/m)$ are upper semicontinuous in (D, d_S) for each m. This means that the sets $B_m = \{x : w'_x(1/m) < \varepsilon\}$ are open in (D, d_S) for each $\varepsilon > 0$. By equivalence of the metrics, they are also open in (D, d_B). In this case, for any such ε, the sets $\{B_m, m \in \mathbb{N}\}$ are an open covering for D by **30.3**. Any compact subset of D then has a finite subcovering, or in other words, if $\bar{A}$ is compact there is an m such that $\bar{A} \subseteq B_m$. By definition of B_m, this implies that (30.53) holds.

To show upper semicontinuity, fix $\varepsilon > 0$, $\delta > 0$, and $x \in D$ and choose a partition Π_δ satisfying

$$\max_{1 \le i \le r} \Big\{ \sup_{s,t \in [t_{i-1}, t_i)} \big|x(t) - x(s)\big| \Big\} < w'_x(\delta) + \tfrac{1}{2}\varepsilon. \tag{30.58}$$

Then, choose $\eta < \frac{1}{4}\varepsilon$ and also small enough that

$$\max_{1\le i\le r}\{t_i - t_{i-1}\} > \delta + 2\eta. \tag{30.59}$$

The object is to show, after (5.33), that if $y \in D$ and $d_S(x,y) < \eta$ then

$$w'_y(\delta) < w'_x(\delta) + \varepsilon. \tag{30.60}$$

If $d_S(x,y) < \eta$ there is $\lambda \in \Lambda$ such that

$$\sup_t \left|y(\lambda(t)) - x(t)\right| < \eta \tag{30.61}$$

and

$$\sup_t \left|\lambda(t) - t\right| < \eta. \tag{30.62}$$

Letting $s_i = \lambda(t_i)$, (30.59) and (30.62) and the triangle inequality imply that

$$\max_{1\le i\le r}\{s_i - s_{i-1}\} > \max_{1\le i\le r}\{t_i - t_{i-1}\} - 2\eta > \delta. \tag{30.63}$$

If both s and t lie in $[t_{i-1}, t_i)$, $\lambda(s)$, and $\lambda(t)$ must both lie in $[s_{i-1}, s_i)$. It follows by (30.58), (30.61), and the choice of η that

$$\max_{1\le i\le r}\left\{\sup_{s,t\in[s_{i-1},s_i)} \left|y(t) - y(s)\right|\right\} < w'_x(\delta) + \varepsilon. \tag{30.64}$$

In view of (30.63) and the definition in (30.4) this shows that (30.60) holds and since ε and x are arbitrary the proof is complete. ∎

Theorem **30.16** can be restated in terms of the alternative continuity modulus (30.7) augmented to allow for end effects, as in (30.11).

30.17 Corollary A set $A \subset D$ is relatively compact in (D, d_B) iff

$$\sup_{x\in A}\sup_t \left|x(t)\right| < \infty \tag{30.65}$$

and

$$\lim_{\delta\to 0}\sup_{x\in A} w^*_x(\delta) = 0. \tag{30.66}$$

Proof If condition (30.66) holds, then by Lemma **30.5** so does (30.53). Hence, sufficiency follows directly from Theorem **30.16**. Necessity follows if (30.66) is implied by (30.53) and this is immediate from (30.8), noting that $w_x[0,\delta) \leq w'_x(\delta)$ and $w_x[1-\delta,1) \leq w'_x(\delta)$. ∎

These results are used to characterize uniform tightness of a sequence in D. The next theorem directly parallels **29.17**. Completeness is needed for this argument to avoid having to prove tightness of every μ_n, so it is necessary to specify an appropriate metric. Without loss of generality d_B, can be cited where required.

30.18 Theorem A sequence $\{\mu_n\}$ of p.m.s on the space $\big((D,d_B),\mathcal{B}_D\big)$ is uniformly tight iff the following conditions hold:

(a) for each $\eta > 0 \;\exists\; M < \infty$ such that for $n \geq 1$,

$$\mu_n\big(\{x : \sup_t |x(t)| > M\}\big) \leq \eta \tag{30.67}$$

(b) for each $\varepsilon > 0$ and $\eta > 0 \;\exists\; \delta \in (0,1)$ and $N \in \mathbb{N}$ such that for $n \geq N$,

$$\mu_n\big(\{x : w'_x(\delta) \geq \varepsilon\}\big) \leq \eta. \tag{30.68}$$

Proof Let $\{\mu_n\}$ be uniformly tight and for $\eta > 0$ choose a compact set K with $\mu_n(K) > 1-\eta$. By **30.16** there exist $M < \infty$, $\delta \in (0,1)$, and $n \geq N$ such that

$$K \subseteq \{x : \sup_t |x(t)| \leq M\} \cap \{x : w'_x(\delta) < \varepsilon\} \tag{30.69}$$

for any $\varepsilon > 0$. Inequalities (30.67) and (30.68) follow, proving necessity.

The object is now to find a set A satisfying the conditions of **30.16**, whose closure K satisfies $\sup_{n\geq N}\mu_n(K) > 1-\theta$ for some $N \in \mathbb{N}$ and all $\theta > 0$. Because (D,d_B) is a complete separable space, each μ_n is tight by **29.12** so this last condition suffices for uniform tightness. As in the proof of **29.17**, let μ^* stand for $\sup_{n\geq N}\mu_n$. For $\theta > 0$ define

$$A_k = \{x : w'_x(\delta_k) < 1/k\} \tag{30.70}$$

where $\{\delta_k\}$ is chosen so that $\mu^*(A_k) > 1-\theta/2^{k+1}$, possible by condition (b). Also set $B = \{x : \sup_t |x(t)| \leq M\}$ such that $\mu^*(B) > 1-\theta/2$, possible by condition (a). Let $K = \big(\bigcap_{k=1}^{\infty} A_k \cap B\big)^{-}$ and note that K satisfies the conditions in (30.52) and (30.53) and hence is compact by **30.16**. With these definitions, the argument follows that of **29.17** word for word. ∎

The last result of this section concerns an issue of obvious relevance to the functional CLT; how to characterize a sequence in D that is converging to a limit in

C. The modulus of continuity w_x is the natural medium for expressing this property of a sequence. Essentially, the following theorem amounts to the result that the sufficiency part of **29.17** holds in (D, d_B) just as in (C, d_U).

30.19 Theorem ([21] th. 15.5) Let $\{\mu_n\}$ be a sequence of measures on $((D, d_B), \mathcal{B}_D)$. If

(a) for each $\eta > 0 \; \exists \; M < \infty$ such that

$$\mu_n\big(\{x : |x(0)| > M\}\big) \leq \eta, \quad n \geq 1 \tag{30.71}$$

(b) for each $\varepsilon > 0$ and $\eta > 0 \; \exists \; \delta \in (0, 1)$ and $N \in \mathbb{N}$ such that

$$\mu_n\big(\{x : w_x(\delta) \geq \varepsilon\}\big) \leq \eta, \quad n \geq N \tag{30.72}$$

then $\{\mu_n\}$ is uniformly tight and if μ is any cluster point of the sequence, $\mu(C) = 1$.

Proof By (30.5), if (30.72) holds for a given δ then (30.68) holds for $\delta/2$. Let $k = [1/\delta] + 1$ (so that $k\delta > 1$) where $\delta > 0$ is specified by condition (b). Then according to (30.72), $\mu_n\big(\{x : |x(ti/k) - x(t(i-1)/k)| \geq \varepsilon\}\big) \leq \eta$ for $i = 1, \ldots, k$, and $t \in [0, 1]$. It was noted previously that

$$|x(t)| \leq |x(0)| + \sum_{i=1}^{k} \left| x\Big(\frac{i}{k}t\Big) - x\Big(\frac{i-1}{k}t\Big) \right| \tag{30.73}$$

where each of the k intervals indicated has width less than δ. It follows by (30.72) and (30.71) that

$$\mu_n\big(\{x : \sup_t |x(t)| > M + k\varepsilon\}\big) \leq \mu_n\big(\{x : |x(0)| > M\}\big) \leq \eta \tag{30.74}$$

so that (30.67) also holds for finite M. The conditions of **30.18** are therefore satisfied, proving uniform tightness.

Let μ be a cluster point such that $\mu_{n_k} \Rightarrow \mu$ for some subsequence $\{n_k, k \in \mathbb{N}\}$. Defining $A = \{x : w_x(\delta) \geq \varepsilon\}$, consider the open set A^o, the interior of A; for example, $x \in A^o$ if $w_x(\delta/2) \geq 2\varepsilon$. Then, by (d) of **29.1** and (30.72),

$$\mu(A^o) \leq \text{liminf}_{k\to\infty} \mu_{n_k}(A^o) \leq \eta. \tag{30.75}$$

Hence $\mu(B) \leq \eta$ for any set $B \subseteq A^o$. Since ε and η are arbitrary here it is possible to choose a decreasing sequence $\{\delta_j\}$ such that $\mu(B_j) \leq 1/j$, where $B_j = \{x : w_x(\delta_j) \geq 1/j\}$. For each $m \geq 1$, $\mu(\bigcap_{j=m}^{\infty} B_j) = 0$ and so, by subadditivity and **3.12**(ii), $\mu(B) = 0$ where $B = \liminf B_j$. But suppose $x \in B^c$, where $B^c = \bigcap_{m=1}^{\infty} \bigcup_{j=m}^{\infty} B_j^c$ is the set

$$\{x : w_x(\delta_j) < 1/j, \text{ some } j \geq m : \text{ all } m \in \mathbb{N}\}.$$

Since $\{\delta_j\}$ is monotonic, it must be the case that $\lim_{\delta \to 0} w_x(\delta) = 0$ for this x. Hence $B^c \subseteq C$ and since $\mu(B^c) = 1$, $\mu(C) = 1$ follows. ∎

30.7 Weak Convergence in *D*

This section shows how the conditions of Theorem **30.18** might be verified in practice by saying something about the behaviour of increments of the stochastic process $\{X_n \in D\}$ associated with the sequence of measures. Conditions sufficient for $X_n \to_{\mathrm{d}} X$ follow from an assumption about the limiting finite-dimensional distributions.

Write $\mu_n(A) = P(X_n \in A)$ for $A \in \mathcal{B}_D$ where $\{\mu_n\}$ is a sequence of probability measures on the space $((D, d_B), B_D)$. The first step is to establish conditions for uniform tightness of the sequence. It turns out to be most convenient in this context to work with the equivalent (by **30.6**) continuity modulus w_x''.

30.20 Theorem For $n \in \mathbb{N}$ let there exist $\theta > 1$, $\rho > 0$, and $K < \infty$ such that

$$\mu_n\big(\min\{|x(s) - x(r)|, |x(t) - x(s)|\} \geq \lambda\big) \leq K\frac{(t-r)^{\theta}}{\lambda^{\rho}} \tag{30.76}$$

for each $0 \leq r < s < t \leq 1$, $\lambda > 0$. Then, for any $\varepsilon > 0$ and $\eta > 0$, $\exists\ \delta > 0$ such that

$$\mu_n\big(w_x''(\delta) \geq \varepsilon\big) < \eta. \tag{30.77}$$

Proof Let t and r be given. For a triple (s_1, s_2, s_3) where $r \leq s_1 < s_2 < s_3 \leq t$, write for brevity

$$m(s_1, s_2, s_3) = \min\{|x(s_2) - x(s_1)|, |x(s_3) - x(s_2)|\}.$$

Note in particular that the event whose probability appears in (30.76) is $m(r, s, t) \geq \lambda$. For $k = 0, 1, 2, \ldots$ define the collection

$$D_k = \{r + 2^{-k}i(t-r) : i = 0, \ldots, 2^k\}$$

so that $D_k \to \mathbb{D} \cap [r, t]$ as $k \to \infty$. Then for $k \geq 1$ define

$$B_k = \max_{s_1, s_2, s_3 \in D_k} m(s_1, s_2, s_3)$$

with $B_0 = 0$. Since $\mathbb{D}$ is dense in $[0,1]$,

$$B_k \to \sup_{r \le s_1 < s_2 < s_3 \le t} \min\{|x(s_2) - x(s_1)|, |x(s_3) - x(s_2)|\} = w_x''(t-r)$$

as $k \to \infty$, where the equality defines $w_x''(t-r)$. Also, for $k \ge 1$ define

$$A_k = \max_{1 \le i \le 2^k - 1} E_{ki} \tag{30.78}$$

where

$$E_{ki} = m(r + 2^{-k}(t-r)(i-1), r + 2^{-k}(t-r)i, r + 2^{-k}(t-r)(i+1)).$$

Thus, A_k is the maximum of m over triplets of *consecutive* points of D_k.

Given any three points $s_1, s_2, s_3 \in D_k$, define points $s_1', s_2', s_3' \in D_{k-1}$ in the following way. Since the partition is doubled at each step, either $s_1 \in D_{k-1}$, or $s_1 - 2^{-k} \in D_{k-1}$ and $s_1 + 2^{-k} \in D_{k-1}$. In the first case, put $s_1' = s_1$. Otherwise, put $s_1' = s_1 - 2^{-k}$ if $|x(s_1) - x(s_1 - 2^{-k})| \le |x(s_1) - x(s_1 + 2^{-k})|$ and $s_1' = s_1 + 2^{-k}$ if $|x(s_1) - x(s_1 - 2^{-k})| > |x(s_1) - x(s_1 + 2^{-k})|$. The same construction defines s_2' given s_2 and s_3' given s_3. Note that the pairs (s_1, s_1'), (s_2, s_2'), and (s_3, s_3') each define one of the E_{ki} whose maximum is A_k. Therefore

$$\begin{aligned} |x(s_2) - x(s_1)| &\le |x(s_2) - x(s_2')| + |x(s_2') - x(s_1')| + |x(s_1) - x(s_1')| \\ &\le |x(s_2') - x(s_1')| + 2A_k \end{aligned} \tag{30.79}$$

and

$$\begin{aligned} |x(s_3) - x(s_2)| &\le |x(s_3) - x(s_3')| + |x(s_3') - x(s_2')| + |x(s_2) - x(s_2')| \\ &\le |x(s_3') - x(s_2')| + 2A_k. \end{aligned} \tag{30.80}$$

Taking the minimum of the minorants of (30.79) and (30.80), note the implication

$$\begin{aligned} m(s_1, s_2, s_3) &\le \min\{|x(s_2') - x(s_1')|, |x(s_3') - x(s_2')|\} + 2A_k \\ &\le B_{k-1} + 2A_k. \end{aligned}$$

Since this inequality holds for any $s_1, s_2, s_3 \in D_k$, it implies $B_k \le B_{k-1} + 2A_k$ and hence by induction, $B_k \le 2(A_1 + A_2 + \cdots + A_k)$ and

$$w_x''(t-r) \le 2\sum_{k=1}^{\infty} A_k.$$

The next step is to bound $\mu_n(w''_x(t-r) \geq \varepsilon)$. Choose a positive constant ζ to satisfy $2^{(1-\theta)/\rho} < \zeta < 1$ where ρ and θ are given by (30.76) and let $C = \frac{1}{2}(1/\zeta - 1)\varepsilon$ so that $C\sum_{k=1}^{\infty}\zeta^k = \frac{1}{2}\varepsilon$. Then, subadditivity gives

$$\begin{aligned}\mu_n(w''_x(t-r) \geq \varepsilon) &\leq \mu_n\Bigl(2\sum_{k=1}^{\infty} A_k \geq \varepsilon\Bigr).\\ &\leq \mu_n\Bigl(\bigcup_{k=1}^{\infty}\{A_k \geq C\zeta^k\}\Bigr)\\ &\leq \sum_{k=1}^{\infty}\mu_n(A_k \geq C\zeta^k). \qquad (30.81)\end{aligned}$$

Note from (30.78) that

$$\begin{aligned}\mu_n(A_k \geq C\zeta^k) &\leq \mu_n\Bigl(\bigcup\nolimits_{1\leq i\leq 2^k-1}\{E_{ki} > C\zeta^k\}\Bigr)\\ &\leq \sum_{i=1}^{2^k-1}\mu_n(E_{ki} \geq C\zeta^k)\\ &\leq 2^k\frac{K}{C^\rho\zeta^{\rho k}}\Bigl(\frac{2(t-r)}{2^k}\Bigr)^\theta = K\frac{2^\theta(t-r)^\theta}{C^\rho(\zeta^\rho 2^{\theta-1})^k} \qquad (30.82)\end{aligned}$$

where the second inequality is by subadditivity and the third is according to (30.76). With $\zeta > 2^{(1-\theta)/\rho}$ the sequence so defined is summable and so from (30.81) and the definition of C,

$$\mu_n(w''_x(t-r) \geq \varepsilon) \leq K\frac{2^\theta(t-r)^\theta}{C^\rho}\sum_{k=1}^{\infty}(\zeta^\rho 2^{\theta-1})^{-k} \leq K^*\frac{(t-r)^\theta}{\varepsilon^\rho} \qquad (30.83)$$

where

$$K^* = \frac{2^{\theta+\rho}K}{(1/\zeta - 1)^\rho(\zeta^\rho 2^{\theta-1} - 1)}.$$

Choosing ζ to minimize this expression gives K^* as a function of K, θ, and ρ alone.

The final step is to show what inequality (30.83) implies for $w''_x(\delta)$. Take $\nu = [1/\delta]$ and set $t_i = i\delta$ for $0 \leq i \leq \nu - 1$ with $t_\nu = 1$. If $t - r < \delta$ then for some i between 1 and $\nu - 1$, $[r,t] \subseteq [t_{i-1}, t_{i+1}]$ and

$$w_x''(t-r) \le w_x''(t_{i+1}-t_{i-1}).$$

Since $[0,1] \subseteq \bigcup_{i=1}^{\nu-1}[t_{i-1},t_{i+1}]$, it follows by another application of subadditivity and then of (30.76) that

$$\begin{aligned}\mu_n(w_x''(\delta) \ge \varepsilon) &\le \max_{1\le i\le \nu-1} \mu_n(w_x''(t_{i+1}-t_{i-1}) \ge \varepsilon)\\ &\le \sum_{i=1}^{\nu-1} \mu_n(w_x''(t_{i+1}-t_{i-1}) \ge \varepsilon)\\ &\le \frac{K^*}{\varepsilon^\rho}\sum_{i=1}^{\nu-1}(t_{i+1}-t_{i-1})^\theta\\ &\le \frac{K^*}{\varepsilon^\rho}\max_i(t_{i+1}-t_{i-1})^{\theta-1}\sum_{i=1}^{\nu-1}(t_{i+1}-t_{i-1})\\ &\le \frac{2K^*}{\varepsilon^\rho}\delta^{\theta-1}.\end{aligned}$$

Given $\eta > 0$, choosing $\delta > 0$ small enough that $2K^*\delta^{\theta-1}\varepsilon^{-\rho} < \eta$ shows (30.77) and completes the proof. ∎

The next step of the weak convergence argument is to specify the finite-dimensional distributions. Recall that $T_\mu \subseteq [0,1]$ is the collection of almost sure continuity points of measure μ, such that $\mu(x(t) \ne x(t-)) = 0$ for $t \in T_\mu$. Points belonging to T_μ include 0 and 1 and by **30.2** the complement of T_μ is at most countable. It follows by Theorem **3.12**(ii) that if points $t_1,\dots,t_k$ are elements of T_μ then the projection $\pi_{t_1,\dots,t_k}$ is continuous except for a set of μ-measure zero. Also, recall the definition of the class of sets $\mathcal{H}_{DT}$ from (30.48) where in the present case $T = T_\mu$.

30.21 Theorem If $\{\mu_n, n \in \mathbb{N}\} \subseteq \mathbb{M}_D$ and $\mu \in \mathbb{M}_D$ where $\mathbb{M}_D$ is the space of probability measures on the space $\big((D,d_B),\mathcal{B}_D\big)$, and

(a) $\mu_n(E) \to \mu(E)$ for $E \in \mathcal{H}_{DT_\mu}$

(b) for any $\varepsilon > 0$ and $\eta > 0$, $\exists\ \delta > 0$ and $N \in \mathbb{N}$ such that

$$\mu_n\big(x : w_x''(\delta) \ge \varepsilon\big) < \eta \text{ for } n > N$$

then $\mu_n \Rightarrow \mu$.

Proof The first step is to show that conditions (a) and (b) imply that the sequence $\{\mu_n\}$ is uniformly tight, by proving the conditions of **30.18**. Condition (a) implies that the sequences of finite-dimensional distributions are tight. Consider E defined, as in (30.48), by a collection $0 = t_1 < \ldots < t_k = 1$ with $\max_i(t_i - t_{i-1}) < \delta$ and $t_i \in T_\mu$. Such collections always exist since T_μ is dense in $[0,1]$. Tightness implies

$$\mu_n\big(\{x : \max_{1\le i\le k}|x(t_i)| \ge M_0\}\big) < \eta/2 \tag{30.84}$$

for any $\eta > 0$, $M_0 < \infty$, and $n \ge 1$. By right continuity, for arbitrary $\varepsilon > 0$ the probabilities $\mu(\{x : |x(\delta) - x(0)| < \varepsilon\})$ and $\mu(\{x : |x(1-\delta) - x(1-)| < \varepsilon\})$ can be made arbitrarily small by choice of δ and when N is large enough condition (a) extends these properties to μ_n for $n > N$. Condition (b) and Lemma **30.6** now show, again for arbitrary ε and any choice of η, that there is δ small enough that for $n > N$,

$$\mu_n(\{x : w'_x(\delta) \ge \varepsilon\}) \le \eta/2. \tag{30.85}$$

$w'_x(\delta)$ shows the largest variation in x over a set of intervals of $[0,1]$ of width at least δ. Since $\max(t_i - t_{i-1}) < \delta$, all of these intervals contain one or more of the t_i and it follows that $\sup_t|x(t)| \le \max_{1\le i\le k}|x(t_i)| + w'_x(\delta)$. Hence,

$$\{x : \sup_t|x(t)| < M_0 + \varepsilon\} \subseteq \big\{x : \max_{1\le i\le k}|x(t_i)| < M_0\big\} \cap \{x : w'_x(\delta) < \varepsilon\}.$$

Moving to complements and applying subadditivity (30.84) and (30.85) therefore give for $M \ge M_0 + \varepsilon$, and $n > N$,

$$\begin{aligned}\mu_n\big(\{x : \sup_t|x(t)| \ge M\}\big) &\le \mu_n\big(\{x : \max_{1\le i\le k}|x(t_i)| \ge M_0\}\big) + \mu_n\big(\{x : w'_x(\delta) \ge \varepsilon\}\big)\\ &\le \eta.\end{aligned} \tag{30.86}$$

According to **30.18,** this argument establishes tightness in the tail of the sequence for $n > N$. For the cases $n \le N$ it follows directly by **29.12**.

Therefore, according to Theorem **29.15** the sequence $\{\mu_n\}$ is compact and by **5.12** it contains a cluster point. In other words there exists a subsequence $\{n_m, m \in \mathbb{N}\}$ such that $\mu_{n_m} \Rightarrow \nu$ for some $\nu \in \mathbb{M}$, meaning in particular that $\mu_{n_m}(E) \to \nu(E)$ for $E \in \mathcal{H}_{DT_\nu}$ where $T_\nu \subseteq [0,1]$ is the set of almost sure continuity points under the measure ν, which like T_μ includes 0 and 1 and has countable complement. Writing $T^* = T_\mu \cap T_\nu$, the assumption of the theorem then implies that $\nu(E) = \mu(E)$ for $E \in \mathcal{H}_{DT^*}$. However, T^* has countable complement if this is true of both T_μ and T_ν and hence $\mu = \nu$ by **30.15**. ∎

Notice how the appeal to **29.15** is a valid step for (D, d_B) thanks to the separability and completeness established in §30.3. This application further motivates those developments.

Formally, this completes the proof of weak convergence in D, but it is helpful to consolidate the various arguments developed by presenting the implicit conclusion as conditions on the sequence of stochastic processes $\{X_n \in D\}$, associated with the sequence of measures, to a limit process $X \in D$. Let T_X denote the set of points t of the interval $[0,1]$ where $P(X(t) \neq X(t-)) = 0$. Points belonging to T_X include 0 and 1 and T_X has countable complement by **30.2**.

30.22 Theorem If, for a sequence $\{X_n \in D\}_{n=1}^{\infty}$,

(a) for $k \in \mathbb{N}$ and $(t_1, \dots, t_k) \in T_X$, $\{X_n(t_1), \dots, X_n(t_k)\} \to_{\mathrm{d}} \{X(t_1), \dots, X(t_k)\}$

(b) $\exists\, \theta > 1, \rho > 0, K < \infty$, and $N \in \mathbb{N}$ such that for each $0 \leq r < s < t \leq 1, \lambda > 0$, and $n \geq N$,

$$P\big(|X_n(s) - X_n(r)| \geq \lambda, |X_n(t) - X_n(s)| \geq \lambda\big) \leq K\frac{(t-r)^\theta}{\lambda^\rho} \qquad (30.87)$$

then $X_n \to_{\mathrm{d}} X$.

Proof This is by verifying the conditions of **30.21** for measures $\{\mu_n\}$ where $\mu_n(A) = P(X_n \in A)$ for $A \in \mathcal{B}_D$. Condition (a) restates condition **30.21**(a) while condition (b) specifies the condition of **30.20** which is sufficient for condition **30.21**(b). ∎

Theorem **30.22** is the key result in terms of the conditions imposed on empirical distributions with finite sample size and the key condition of the theorem is that $\theta > 1$. Since $\{x \geq \lambda, y \geq \lambda\} \subseteq \{xy > \lambda^2\}$ it is easy to see that (30.87) holds if there is a finite $K > 0$ such that for each $0 \leq r < s < t \leq 1$,

$$\mathrm{E}|X_n(s) - X_n(r)|^{\rho/2}|X_n(t) - X_n(s)|^{\rho/2} \leq K(t-r)^\theta. \qquad (30.88)$$

However, existence of moments is not a requirement. The theorem given here imposes stationarity of the increments, but at the cost of some elaboration of the proof the bounding function can be generalized from t to $F(t)$ where F is a continuous monotone nondecreasing function—effectively a distribution function. Billingsley ([21], [22]) gives a variant of **30.22** in this form. The extension allows the distribution of $X(t)$ to depend systematically on t, for example by trends in volatility of the increments. This type of phenomenon is examined for the Gaussian case in §31.4.

30.23 Example Let $X_n(t) = n^{-1/\alpha}\sum_{i=1}^{[nt]} U_i$ where $\{U_i\}_{i=1}^n$ is an i.i.d. sequence whose coordinates are in the domain of attraction of an α-stable distribution, $S_\alpha(\beta)$ for $0 < \alpha < 2$ where either $\alpha \neq 1$ or $\beta = 0$. With large enough n, the normalized increments $(X_n(t) - X_n(s))/(t-s)^{1/\alpha}$ for $0 \leq s < t \leq 1$ are in the same domain of attraction. According to Theorem **24.23**,

$$\begin{aligned} P(|X_n(t) - X_n(s)| \geq \lambda) &= P\left(\left|\frac{X_n(t) - X_n(s)}{(t-s)^{1/\alpha}}\right| \geq \frac{\lambda}{(t-s)^{1/\alpha}}\right) \\ &= (t-s)\lambda^{-\alpha}L_n(\lambda) \end{aligned}$$

where L_n is a slowly varying function of λ. There therefore exists $K < \infty$ large enough that $L_n(\lambda)^2 \leq K\lambda^\eta$ for $0 < \eta < 2\alpha$. If $0 \leq r < s < t \leq 1$, independence of the increments implies

$$\begin{aligned} P\big(|X_n(t) - X_n(s)| &\geq \lambda,\, |X_n(s) - X_n(r)| \geq \lambda\big) \\ &= P\big(|X_n(t) - X_n(s)| \geq \lambda\big)P\big(|X_n(s) - X_n(r)| \geq \lambda\big) \\ &= \frac{(t-s)(s-r)}{\lambda^{2\alpha}}L_n(\lambda)^2 \\ &\leq K\frac{(t-r)^2}{\lambda^{2\alpha-\eta}}. \end{aligned} \tag{30.89}$$

The tightness condition (30.87) is therefore satisfied. By Theorem **30.22**, X_n converges to the Lévy motion on $[0, 1]$ defined in **28.3**. □

31
FCLTs for Dependent Variables

This chapter contrasts two very different approaches to proving the FCLT. One of them turns out to dominate the other under certain relatively extreme conditions, but the comparison of two different styles of argument is perhaps of sufficient interest to justify including both for consideration, as the reader may judge. The objects of interest are generally stochastic processes $X : [0,1] \mapsto \mathbb{R}$ for which $\mathrm{E}(X(1)^2) = 1$, although the normalization is merely conventional. Unless otherwise noted, it is understood that in a case $\mathrm{E}(X(1)^2) = \sigma^2$ the process under consideration has the form X/σ.

31.1 Asymptotic Independence

Let $\{X_n\}_1^\infty$ denote a stochastic sequence in $(D, \mathcal{B}_D)$. X_n has asymptotically independent increments if, for any collection of points $\{s_i, t_i, i = 1, \dots, r\}$ such that

$$0 \leq s_1 \leq t_1 < s_2 \leq t_2 < \dots < s_r \leq t_r \leq 1$$

and all collections of linear Borel sets $B_1, \dots, B_r \in \mathcal{B}$,

$$\left| P\big(X_n(t_i) - X_n(s_i) \in B_i, i = 1, \dots, r\big) - \prod_{i=1}^{r} P\big(X_n(t_i) - X_n(s_i) \in B_i\big) \right| \to 0 \quad (31.1)$$

as $n \to \infty$. In this definition gaps of positive width are allowed to separate the increments. This will be essential to establish asymptotic independence in the partial sums of mixing sequences, although the gaps can be arbitrarily small, and continuity allows them to be ignored.

Theorem **28.21** shows that in a continuous process with a finite variance, independent increments imply distribution W. The problem is that a partial-sum process with a finite number of terms, even if continuous in the manner of (29.66), clearly cannot have independent increments of width less than $1/n$. However, a condition of asymptotic independence arising as a consequence of time aggregation of observations having limited memory provides a natural path to an invariance principle. In the following result, be careful to note that $w(\cdot, \delta)$ in condition (b) is the modulus of continuity of (27.26), *not* w' of (30.4).

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0031

31.1 Theorem Let $\{X_n\}_{n=1}^{\infty}$ have the following properties:

(a) The increments are asymptotically independent.

(b) For any $\varepsilon > 0$ and $\eta > 0$, $\exists\, \delta > 0$ such that $\limsup_{n\to\infty} P\big(w(X_n, \delta) \geq \varepsilon\big) \leq \eta$.

(c) $\big\{X_n^2(t)\big\}_{n=1}^{\infty}$ is uniformly integrable for each $t \in [0,1]$.

(d) $\mathrm{E}\big(X_n(t)\big) \to 0$ and $\mathrm{E}\big(X_n^2(t)\big) \to t$ as $n \to \infty$, for each $t \in [0,1]$.

Then $X_n \to_{\mathrm{d}} B$.

Proof Condition (b) and the fact that $\mathrm{E}|X_n(0)| \to 0$ by (d) imply by **30.19** that the associated sequence of p.m.s is uniformly tight. Theorem **29.15** then implies that the latter sequence is compact and so has one or more cluster points. Consider the properties these cluster points must possess. Writing X for the random element, **30.19** also gives $P(X \in C) = 1$. Applying **23.16**, uniform integrability of $X_n^2(t)$ and hence also of $X_n(t)$ by condition (c) implies by condition (d) that $\mathrm{E}(X(t)) = 0$ and $\mathrm{E}(X^2(t)) = t$. By condition (a), the increments $X(t_1) - X(s_1), \dots, X(t_r) - X(s_r)$ are totally independent according to (31.1). Specifically, consider increments $X(t_i) - X(s_i)$ and $X(t_{i+1}) - X(s_{i+1})$ for the case where $s_{i+1} = t_i + 1/m$. By a.s. continuity,

$$\lim_{m\to\infty} \big(X(t_{i+1}) - X(t_i + 1/m)\big) = X(t_{i+1}) - X(t_i) \text{ w.p.1} \tag{31.2}$$

so that asymptotic independence extends to contiguous increments. Therefore, the conditions of **28.21** are all satisfied by X. All cluster points have the characteristics of Wiener measure and hence the sequence has this p.m. as its unique weak limit. ∎

The aim is now to get a FCLT for partial-sum processes by linking up the asymptotic independence idea with the established characterization of a dependent increment process; that is to say, as a near-epoch dependent function of a mixing process. Making this connection is perhaps the biggest difficulty to be surmounted. An approach comparable to the 'blocking' argument used in the CLTs of §25.4 is needed, which in the present context means mapping an infinite sequence into $[0,1]$ and identifying the increments with asymptotically independent blocks of summands. This is a particularly elegant route to the result. However, a property of the type exploited in **25.6** where a mixingale behaves asymptotically like a martingale difference is not going to work in this approach to the problem. While the terms of a mixing process (of suitable size) can be 'blocked' so that the blocks are asymptotically independent (more or less by definition of mixing), mixingale theory will not serve here; near-epoch dependent functions can be dealt with only by a direct approximation argument.

If the difference between two stochastic processes is $o_p(1)$ and one of them exhibits asymptotic independence then so must the other, in a sense to be defined. Near-epoch dependent functions can be approximated in the required way by their near-epoch conditional expectations, where the latter are functions of mixing variables. This result is established in the following lemma in terms of the independence of a pair of sequences. In the application these will be adjacent increments of a partial sum process. The following lemma is from Wooldridge and White [194].

31.2 Lemma If $\{Y_{jn}\}$ and $\{Z_{jn}\}$ are real stochastic sequences and

(a) $Y_{jn} - Z_{jn} \to_{\text{pr}} 0$, for $j = 1, 2$

(b) $Y_{jn} \to_{\text{d}} Y_j$ for $j = 1, 2$

(c) for any $A_1, A_2 \in \mathcal{B}$

$$\big|P\big(\{Z_{1n} \in A_1\} \cap \{Z_{2n} \in A_2\}\big) - P(Z_{1n} \in A_1)P(Z_{2n} \in A_2)\big| \to 0 \qquad (31.3)$$

as $n \to \infty$

then for all Y_j-continuity sets (sets $B_j \in \mathcal{B}$ such that $P(Y_j \in \partial B_j) = 0$) for $j = 1, 2$,

$$P\big(\{Y_{1n} \in B_1\} \cap \{Y_{2n} \in B_2\}\big) \to P(Y_1 \in B_1)P(Y_2 \in B_2). \qquad (31.4)$$

Proof Considering (Z_{1n}, Z_{2n}) and (Y_{1n}, Y_{2n}) as points of $\mathbb{R}^2$ with the Euclidean metric, (a) implies $d_E((Z_{1n}, Z_{2n}), (Y_{1n}, Y_{2n})) \to_{\text{pr}} 0$, and by an application of **29.5** (b) implies $(Z_{1n}, Z_{2n}) \to_{\text{d}} (Y_1, Y_2)$. Letting μ_n be the measure associated with the element (Z_{1n}, Z_{2n}), write

$$P\big(\{Z_{1n} \in B_1\} \cap \{Z_{2n} \in B_2\}\big) = \mu_n(B_1 \times B_2). \qquad (31.5)$$

If μ is the measure associated with (Y_1, Y_2), define the marginal measures μ^j by $\mu^j(B_j) = P(Y_j \in B_j)$. Then $\mu^j(\partial B_j) = 0$ for $j = 1, 2$ implies $\mu(\partial(B_1 \times B_2)) = 0$ in view of the fact that

$$\partial(B_1 \times B_2) \subseteq (\partial B_1 \times \mathbb{R}) \cup (\mathbb{R} \times \partial B_2). \qquad (31.6)$$

Applying (e) of **29.1**, it follows from the weak convergence of the joint distribution that for all Y_j-continuity sets B_j,

$$\begin{aligned} P\big(\{Z_{1n} \in B_1\} \cap \{Z_{2n} \in B_2\}\big) &= \mu_n(B_1 \times B_2) \\ &\to \mu(B_1 \times B_2) \\ &= P\big(\{Y_1 \in B_1\} \cap \{Y_2 \in B_2\}\big). \end{aligned} \qquad (31.7)$$

By the weak convergence of both sets of marginal distributions it follows that for the same B_j,

$$P(Z_{1n} \in B_1)P(Z_{2n} \in B_2) \to P(Y_1 \in B_1)P(Y_2 \in B_2). \tag{31.8}$$

This completes the proof, since the limits of the left-hand sides of (31.7) and (31.8) are the same by condition (c). ∎

31.2 NED Functions of Mixing Processes 1

From **31.1** to a general invariance principle for dependent sequences is only a short step, even though some of the details in the following version of the result are quite fiddly. This is basically the one given by [194].

31.3 Theorem Let $\{U_{ni}\}$ be a zero-mean triangular stochastic array and $\{c_{ni}\}$ be an array of positive constants. Also define $X_n(t) = \sum_{i=1}^{[nt]} U_{ni}$. If

(a) $\sup_{i,n} \|U_{ni}/c_{ni}\|_r < \infty$, where either $r > 2$ or $r = 2$ and $\{U_{ni}^2/c_{ni}^2\}$ is uniformly integrable

(b) $\{U_{ni}\}$ is L_2-NED of size $-\frac{1}{2}$ with respect to constant array $\{c_{ni}\}$ on either an α-mixing array $\{\boldsymbol{V}_{ni}\}$ of size $-r/(r-2)$ for $r > 2$ or a ϕ-mixing array $\{\boldsymbol{V}_{ni}\}$ of size $-r/(2r-2)$ for $r \geq 2$

(c) $\sup_n n \max_{1 \leq i \leq n}\{c_{ni}^2\} < \infty$

(d) $\sup_{t\in[0,1],\delta\in(0,1-t]} \limsup_{n\to\infty} v_n^2(t,\delta)/\delta < \infty$, where $v_n^2(t,\delta) = \sum_{i=[nt]+1}^{[n(t+\delta)]} c_{ni}^2$

(e) $\mathrm{E}\big(X_n^2(t)\big) \to t$ as $n \to \infty$, for each $t \in [0,1]$

then $X_n \to_{\mathrm{d}} B$. □

A standard case is $U_{ni} = U_i/s_n$, where

$$s_n^2 = \mathrm{E}\Big(\sum_{i=1}^{n} U_i\Big)^2 = \sum_{i=1}^{n} \sigma_{ii} + 2\sum_{i=1}^{n-1}\sum_{m=1}^{n-i} \sigma_{i,i+m} \tag{31.9}$$

with $\sigma_{ii} = \mathrm{Var}(U_i)$ and $\sigma_{i,i+m} = \mathrm{Cov}(U_i, U_{i+m})$. If $\sup_i \|U_i\|_r < \infty$, $r > 2$ then choose $c_{ni} = 1/s_n$. Condition **31.3**(c) is satisfied if the autocovariances are summable and $s_n^2 = O(n)$. Then, condition **31.3**(d) reduces to the requirement that $s_n^2/n > 0$ uniformly in n. If in addition $s_n^2/n \to \sigma^2 < \infty$ then $\mathrm{E}(X_n(t)^2) = s_{[nt]}^2/s_n^2 \to t$ and **31.3**(e) also holds. These conclusions are summarized in the following corollary.

31.4 Corollary Let the sequence $\{U_i\}$ have mean zero, be uniformly L_r-bounded, and be L_2-NED of size $-\frac{1}{2}$ on either an α-mixing process of size $-r/(r-2)$ for $r > 2$, or a ϕ-mixing process of size $-r/(2r-2)$ for $r \geq 2$. If $r = 2$, then $\{U_i^2\}$ is uniformly integrable. Let $X_n(t) = n^{-1/2}\sum_{i=1}^{[nt]} U_i$. If

$$n^{-1}\mathrm{E}\left(\sum_{i=1}^{n} U_i\right)^2 \to \sigma^2,\ 0 < \sigma^2 < \infty$$

then $X_n \to_{\mathrm{d}} \sigma^2 B$. □

Be careful to note that $\sigma^2 = \bar{\sigma}^2 + 2\sum_{m=1}^{\infty}\lambda_m$, where $\bar{\sigma}^2 = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n}\sigma_{ii}$ and $\lambda_m = \lim_{n\to\infty} n^{-1}\sum_{i=1}^{n-m}\sigma_{i,i+m}$. It is σ^2 and not $\bar{\sigma}^2$ that is the variance of the limiting Brownian motion, notwithstanding the fact that B has independent increments. The condition $s_n^2/n \to \sigma^2$ has two parts. The first is that the limits $\bar{\sigma}^2$ and λ_m for $m = 1, 2, 3, \ldots$ all exist, which is the condition of global wide-sense stationarity discussed in §13.1. Examples where this condition is violated are provided by **25.8** and **25.9**. The second is that $\sum_{m=1}^{\infty}\lambda_m < \infty$, for which it is sufficient that $\sum_{m=1}^{\infty}|\sigma_{i,i+m}| < \infty$ for each i. According to **17.16** this follows from condition **31.3**(b).

It is instructive to compare the conditions of **31.3** with those of **25.12**. Since $X_n(1) \to_{\mathrm{d}} B(1) \sim_{\mathrm{d}} \mathrm{N}(0,1)$, the two theorems give alternative sets of conditions for the central limit theorem, and although they are stated in very different terms, conditions **31.3**(c) and (d) clearly have a role analogous to **25.12**(c). However, **31.3**(e) has no counterpart in the CLT conditions.

Proof of 31.3. The conditions of **31.1** are shown for the sequence $\{X_n\}$. Condition **31.1**(d) holds directly, by the present condition (e). Condition (b) implies by **18.7**(i) in the α-mixing case or **18.7**(ii) in the ϕ-mixing case that $\{U_{ni}, \mathcal{F}_{ni}\}$ is an L_2-mixingale of size $-\frac{1}{2}$ with respect to the scaling constants $\{c_{ni}\}$, where $\mathcal{F}_{ni} = \sigma(V_{n,i-j}, j \geq 0)$. By condition (a) the array $\{U_{ni}^2/c_{ni}^2\}$ is uniformly integrable.

Let $k = [nt]$, $m = [n(t+\delta)] - [nt]$ for $\delta \in [0, 1-t)$ and $S_{nk} = \sum_{i=1}^{k} U_{ni}$. It follows by **17.15** that the set

$$\left\{\max_{1\leq j\leq m}\frac{(S_{n,k+j} - S_{nk})^2}{v_n^2(t,\delta)},\ n \geq 1\right\} \tag{31.10}$$

is uniformly integrable. This is true for any $t \in [0,1)$ and any $\delta \in (0, 1-t]$. Because of condition (d), for a positive constant $M < \infty$ there exists $N(t,\delta) \geq 1$ with the property $v_n^2(t,\delta)/\delta \leq M$ for $n \geq N(t,\delta)$, for any such t and δ. Therefore the set

$$\left\{\max_{1\leq j\leq m} \frac{(S_{n,k+j} - S_{nk})^2}{\delta},\ n \geq N(t,\delta)\right\} \tag{31.11}$$

is also uniformly integrable. Letting $N^* = \sup_{t,\delta} N(t,\delta)$, condition (d) implies that N^* is finite. Take the case $t = 0$ and hence $k = 0$ and $m = [n\delta]$ but then, putting t in place of δ and $X_n(t)$ in place of $S_{n,[nt]}$ in (31.11), what has been shown is that $\{X_n^2(t)\}_{n=1}^{\infty}$ is uniformly integrable for any $t \in (0,1]$. In other words, condition **31.1**(c) holds.

By **9.17**, $\lambda^2 P(|X| > \lambda) \leq \mathrm{E}(X^2 1_{\{|X|>\lambda\}})$ for any r.v. X that is square-integrable. The uniform integrability of set (31.11) implies that for $\delta \in (0,1)$ and $t \leq 1 - \delta$ and any $\varepsilon > 0$ and $\eta > 0$, there exists $\lambda > 0$ large enough that for $n \geq N^*$, with k and m defined as before,

$$P\Big(\max_{1\leq j\leq m} |S_{n,k+j} - S_{nk}| \geq \lambda\sqrt{\delta}\Big) \leq \frac{\eta\varepsilon^2}{8\lambda^2}. \tag{31.12}$$

For the case $\delta = \varepsilon^2/4\lambda^2$, the condition that (31.12) holds for $0 \leq t \leq 1 - \delta$ and $n \geq N^*$can be written

$$\sup_{0\leq t\leq 1-\delta} P\Big(\sup_{t\leq s\leq t+\delta} |X_n(s) - X_n(t)| \geq \tfrac{1}{2}\varepsilon\Big) \leq \tfrac{1}{2}\eta\delta,\ n \geq N^*. \tag{31.13}$$

This is identical to (29.63) in the present notation, and condition **31.1**(b) now follows by **29.18**.

The final step is to show asymptotic independence. Whereas the theorem requires that (31.1) hold for any r, since the argument is based on the mixing property and the linear separation of the increments it suffices to show independence for adjacent pairs of increments indexed i and $i + 1$, having $t_i < s_{i+1}$. The extension to the general case is easy in principle if tedious to write out.

Hence, without loss of generality consider the pair of variables

$$Y_{jn} = X_n(t_j) - X_n(s_j) = \sum_{i=[ns_j]+1}^{[nt_j]} U_{ni}, j = 1 \text{ and } 2 \tag{31.14}$$

where $0 \leq s_1 < t_1 < s_2 < t_2 \leq 1$. Asymptotic independence of Y_{1n} and Y_{2n} cannot be shown directly because the increment process need not be mixing, but there is an approximation argument direct from the NED property. Defining $\mathcal{F}_{n,j}^k = \sigma(V_{nj}, \ldots, V_{nk})$, the r.v. $\mathrm{E}(Y_{1n}|\mathcal{F}_{n,-\infty}^{[nt_1]})$ is $\mathcal{F}_{n,-\infty}^{[nt_1]}$-measurable and similarly

$\mathrm{E}(Y_{2n}|\mathcal{F}^{\infty}_{n,[ns_2]})$ is $\mathcal{F}^{\infty}_{n,[ns_2]}$-measurable. By assumption (b),

$$\sup_{A\in\mathcal{F}^{[nt_1]}_{n,-\infty},B\in\mathcal{F}^{\infty}_{n,[ns_2]}} |P(A\cap B)-P(A)P(B)| = \alpha([ns_2]-[nt_1]) \to 0 \text{ as } n\to\infty \qquad (31.15)$$

whenever $t_1 < s_2$, where the events A include those of the form $\{\mathrm{E}(Y_{1n}|\mathcal{F}^{[nt_1]}_{n,-\infty}) \in E\}$ for $E \in \mathcal{B}$ and similarly events B include those of the type $\{\mathrm{E}(Y_{2n}|\mathcal{F}^{\infty}_{n,[ns_2]}) \in E\}$. Since $\phi(m) \geq \alpha(m)$, (31.15) also holds in the ϕ-mixing case. These conditional expectations are asymptotically independent r.v.s and it remains to be shown that Y_{1n} and Y_{2n} share the same property.

The conditions of **31.2** are satisfied when $Z_{1n} = \mathrm{E}(Y_{1n}|\mathcal{F}^{[nt_1]}_{n,-\infty})$ and $Z_{2n} = \mathrm{E}(Y_{2n}|\mathcal{F}^{\infty}_{n,[ns_2]})$. This is sufficient in view of the fact that the Y_j-continuity sets are a convergence-determining class for the sequences $\{Y_{jn}\}$, by **29.1**(e). The argument of the preceding paragraph has already established condition **31.2**(c). To show condition **31.2**(a), note

$$\begin{aligned}\mathrm{E}\left\|Y_{1n} - \mathrm{E}(Y_{1n}|\mathcal{F}^{[nt_1]}_{n,-\infty})\right\|_2 &\leq \sum_{i=[ns_1]+1}^{[nt_1]} \left\|U_{ni} - \mathrm{E}(U_{ni}|\mathcal{F}^{[nt_1]}_{n,-\infty})\right\|_2 \\ &\leq 2\sum_{i=[ns_1]+1}^{[nt_1]} \left\|U_{ni} - \mathrm{E}(U_{ni}|\mathcal{F}^{[nt_1]}_{n,2i-[nt_1]})\right\|_2 \\ &\leq 2\sum_{i=[ns_1]+1}^{[nt_1]} c_{ni}\nu_{[nt_1]-i} \\ &\leq 2\max_{[ns_1]<i\leq[nt_1]} c_{ni} \sum_{m=0}^{[nt_1]-[ns_1]-1} \nu_m \\ &\to 0 \text{ as } n\to\infty \end{aligned} \qquad (31.16)$$

where ν_m is defined in **18.1** and the inequalities are successively Minkowski's, **10.29** and (18.2). Under assumption (b) and **2.17**(i) the sum of NED numbers is $o(n^{1/2})$, while $\max_i c_{ni} = O(n^{-1/2})$ by assumption (c). This result implies that $Y_{1n} - \mathrm{E}(Y_{1n}|\mathcal{F}^{[nt_1]}_{n,-\infty}) \to_{\mathrm{pr}} 0$. By the same reasoning, $Y_{2n} - \mathrm{E}(Y_{2n}|\mathcal{F}^{\infty}_{n,[ns_2]}) \to_{\mathrm{pr}} 0$ also.

If conditions **31.1**(b) and **31.1**(d) hold, the sequence of measures associated with $\{X_n\}$ is uniformly tight and so contains at least one convergent subsequence. Let this be $\{n_k, k\in\mathbb{N}\}$ such that $X_{n_k} \to_{\mathrm{d}} X$ (say) as $k\to\infty$ where $P(X\in C) = 1$. It follows that the continuous mapping theorem applies to the coordinate projections $\pi_t(X) = X(t)$ and $X_{n_k}(t) \to_{\mathrm{d}} X(t)$. Confining attention to this subsequence, condition **31.2**(b) is satisfied for the case $Y_{j,n_k} = X_{n_k}(t_j) - X_{n_k}(s_j)$. All the conditions of

31.2 have now been confirmed, so these increments are asymptotically independent in the sense of (31.4). But since this is true for every convergent subsequence $\{n_k\}$, the weak limit of $\{X_n\}$ has asymptotically independent increments whenever it exists. All the conditions of **31.1** are therefore fulfilled by $\{X_n\}$ and the proof is complete. ∎

31.3 NED Functions of Mixing Processes 2

Here is an alternative approach, proving the same result by a method that is closer in spirit to **29.19**. This starts with the central limit theorem **25.12** to establish the finite-dimensional distributions of the process and then adds a condition sufficient to ensure stochastic equicontinuity, so that the limit process is both Gaussian and almost surely continuous. The dependence permitted under the CLT is compatible with the limit process having independent increments. The following is a somewhat simplified version of [54] Theorem 3.1.

31.5 Theorem Let $\{U_{ni}\}$ be a zero-mean stochastic triangular array and $\{c_{ni}\}$ an array of positive constants. Also define $X_n(t) = \sum_{i=1}^{[nt]} U_{ni}$. If

(a) $\sup_{i,n} \|U_{ni}/c_{ni}\|_r < \infty$, where either $r > 2$ or $r = 2$ and $\{U_{ni}^2/c_{ni}^2\}$ is uniformly integrable

(b) $\{U_{ni}\}$ is L_2-NED of size $-\frac{1}{2}$ with respect to constant array $\{c_{ni}\}$ on either an α-mixing array $\{V_{ni}\}$ of size $-r/(r-2)$ for $r > 2$ or a ϕ-mixing array $\{V_{ni}\}$ of size $-r/(2r-2)$ for $r \geq 2$

(c) $\sup_n n \max_{1\leq i\leq n}\{c_{ni}^2\} < \infty$

(d) $\lim_{\delta\to 0} \sup_{t\in[0,1-\delta]} \limsup_{n\to\infty} \nu_n^2(t,\delta) = 0$ where $\nu_n^2(t,\delta) = \sum_{i=[nt]+1}^{[n(t+\delta)]} c_{ni}^2$

(e) $\mathrm{E}\big(X_n^2(t)\big) \to t$ as $n \to \infty$, for each $t \in [0,1]$

then, $X_n \to_{\mathrm{d}} B$. □

It is of interest to compare the assumptions of Theorems **31.5** and **31.3**. Assumptions (a), (b), (c), and (e) are the same in each theorem, the distinguishing feature being the respective assumptions (d). **31.5**(d) is generally milder than **31.3**(d), although in the case of Corollary **31.4** where $c_{ni} = 1/s_n$ for each i, they are clearly equivalent. This issue is revisited in §31.4.

Proof of 31.5 Assumptions (a), (b), and (c) are sufficient for the assumptions of Theorem **25.12**, including (25.35) which as noted is sufficient for **25.6**(c) by Corollary **25.7**. These assumptions ensure that the central limit theorem holds for any finite collection of coordinates $t_1, \dots, t_k \in [0,1]$ so that

$$\big(X_n(t_1), \dots, X_n(t_k)\big) \xrightarrow{\mathrm{d}} \big(X(t_1), \dots, X(t_k)\big)$$

where by the Cramér–Wold theorem (**26.5**) the limit distribution of the indicated k-vector is multivariate Gaussian. To show that the limit process X is identical with B, the procedure is first to show that under assumption (d) it is almost surely continuous, and second to show that it has independent increments, while assumption (e) specifies the variance profile of Brownian motion.

Define the modulus of continuity as

$$w_n(\delta) = \sup_{t\in[0,1]} \sup_{\{s:\,0<|t-s|<\delta\}} |X_n(t) - X_n(s)|.$$

Since $X_n(0) = 0$ by construction, uniform tightness and almost sure continuity follow by **30.19** if $P(w_n(\delta) > \varepsilon)$ can be made arbitrarily close to zero for any $\varepsilon > 0$ by taking δ small enough and n large enough. Begin by setting up some notation. In what follows, set $X_n(t) = X_n(1)$ for $t > 1$. Fix $\delta > 0$ and for $j = 0, \dots, [1/\delta]$ and any pair t, s such that $|t-s| < \delta$, let $j_\delta(t,s)$ denote the maximal value of j such that $t \geq j\delta$ and $s \geq j\delta$. Note that $0 \leq t - j_\delta(t,s)\delta < 2\delta$ and also $0 \leq s - j_\delta(t,s)\delta < 2\delta$. Then,

$$\begin{aligned}
&P(w_n(\delta) > \varepsilon)\\
&= P\Big(\sup_{t\in[0,1]} \sup_{\{s:\,|t-s|<\delta\}} |X_n(t) - X_n(j_\delta(t,s)\delta) + X_n(j_\delta(t,s)\delta) - X_n(s)| > \varepsilon\Big)\\
&\leq 2P\Big(\max_{j=0,\dots,[1/\delta]} w_{nj}(\delta) > \tfrac{1}{2}\varepsilon\Big) \qquad (31.17)
\end{aligned}$$

where

$$w_{nj}(\delta) = \sup_{\{t:\,0<|t-j\delta|<2\delta\}} |X_n(t) - X_n(j\delta)|.$$

The inequality in (31.17) is a consequence of subadditivity and the stylized fact that for variables a, b, and c,

$$\{|a-b| > \varepsilon\} \subseteq \{|a-c| > \tfrac{1}{2}\varepsilon\} \cup \{|c-b| > \tfrac{1}{2}\varepsilon\}.$$

For brevity let $\nu_{nj}(\delta) = \nu_n(j\delta, 2\delta)$ where ν_n is defined in condition (d). Continuing the development of (31.17), letting $1(A)$ denote the indicator of event A,

$$\begin{aligned}
2P\Big(\max_{j=0,\dots,[1/\delta]} w_{nj}(\delta) > \tfrac{1}{2}\varepsilon\Big)
&\le 2\sum_{j=0}^{[1/\delta]} P\big(w_{nj}(\delta) > \tfrac{1}{2}\varepsilon\big)\\
&= 2\sum_{j=0}^{[1/\delta]} P\big(w_{nj}^2(\delta)1(w_{nj}^2(\delta) > \tfrac{1}{4}\varepsilon^2) > \tfrac{1}{4}\varepsilon^2\big)\\
&\le 2\sum_{j=0}^{[1/\delta]} \frac{4}{\varepsilon^2}\mathrm{E}\big(w_{nj}^2(\delta)1(w_{nj}^2(\delta) > \tfrac{1}{4}\varepsilon^2)\big)\\
&= 2\sum_{j=0}^{[1/\delta]} \frac{4\nu_{nj}^2(\delta)}{\varepsilon^2}\mathrm{E}\bigg(\frac{w_{nj}^2(\delta)}{\nu_{nj}^2(\delta)}1\bigg(\frac{w_{nj}^2(\delta)}{\nu_{nj}^2(\delta)} > \frac{\varepsilon^2}{4\nu_{nj}^2(\delta)}\bigg)\bigg)\\
&\le 4\bigg(\sum_{i=1}^{n} c_{ni}^2\bigg)\max_{j=0,\dots,[1/\delta]}\frac{4}{\varepsilon^2}\mathrm{E}\bigg(\frac{w_{nj}^2(\delta)}{\nu_{nj}^2(\delta)}1\bigg(\frac{w_{nj}^2(\delta)}{\nu_{nj}^2(\delta)} > \frac{\varepsilon^2}{4\nu_{nj}^2(\delta)}\bigg)\bigg)\\
&\le \frac{C}{\varepsilon^2}\max_{j=0,\dots,[1/\delta]}\mathrm{E}\bigg(\frac{w_{nj}^2(\delta)}{\nu_{nj}^2(\delta)}1\bigg(\frac{w_{nj}^2(\delta)}{\nu_{nj}^2(\delta)} > \frac{\varepsilon^2}{4\max_{k=0,\dots,[1/\delta]}\nu_{nk}^2(\delta))}\bigg)\bigg) \qquad (31.18)
\end{aligned}$$

for a finite constant $C > 0$. Here, the first inequality is by subadditivity, the second is the Markov inequality **9.17**, the third substitutes from the definition of $\nu_{nj}(\delta)$, and the last uses condition (c).

Next, note that by the mixingale property and Corollary **17.15**, for any t and δ the sequences

$$\sup_{\{s:0<s-t<\delta\}}\frac{(X_n(s)-X_n(t))^2}{\nu_n^2(t,\delta)} \qquad (31.19)$$

are uniformly integrable. Assumptions (a) and (b) imply that this property is independent of the choice of t and δ. In other words,

$$\begin{aligned}
\limsup_{n\to\infty}\max_{j=0,\dots,[1/\delta]}\mathrm{E}\bigg(\frac{w_{nj}^2(\delta)}{\nu_{nj}^2(\delta)}1\bigg(\frac{w_{nj}^2(\delta)}{\nu_{nj}^2(\delta)} > K\bigg)\bigg)&\\
= \max_{j=0,\dots,[1/\delta]}\limsup_{n\to\infty}\mathrm{E}\bigg(\frac{w_{nj}^2(\delta)}{\nu_{nj}^2(\delta)}1\bigg(\frac{w_{nj}^2(\delta)}{\nu_{nj}^2(\delta)} > K\bigg)\bigg) &= f(K)
\end{aligned}$$

where $f(K)$ does not depend on δ and $f(K) \to 0$ as $K \to \infty$. Thus, in view of (31.17) and (31.18),

$$\lim_{\delta\to 0}\limsup_{n\to\infty} P(w_n(\delta) > \varepsilon) = 0$$

for any $\varepsilon > 0$ if

$$\lim_{\delta\to 0}\limsup_{n\to\infty} \max_{k=0,\dots,[1/\delta]} v_{nk}^2(\delta) = 0. \tag{31.20}$$

Since the 'max' and the 'limsup' in equation (31.20) can similarly be interchanged, this holds by assumption (d).

Letting X denote the limit process, the second step is to show that it has independent increments. In contrast to the approach of §31.1, since Gaussianity is established via the finite-dimensional distributions it is only necessary to prove that increments $X(t_1) - X(s_1)$ and $X(t_2) - X(s_2)$ are uncorrelated for any $0 \le s_1 < t_1 < s_2 < t_2 \le 1$. This uncorrelatedness follows because

$$\mathrm{E}(X(t_1) - X(s_1))(X(t_2) - X(s_2)) = \lim_{n\to\infty} \mathrm{E}(X_n(t_1) - X_n(s_1))(X_n(t_2) - X_n(s_2))$$

and for any fixed $\delta > 0$,

$$\begin{aligned}
&\left|\mathrm{E}(X_n(t_1) - X_n(s_1))(X_n(t_2) - X_n(s_2))\right| \\
&\quad \le \left|\sum_{i=[ns_1]+1}^{[nt_1]} \sum_{k=[ns_2]+1}^{[n(s_2+\delta)]} \mathrm{E}U_{ni}U_{nk}\right| + \left|\sum_{i=[ns_1]+1}^{[nt_1]} \sum_{k=[n(s_2+\delta)]+1}^{[nt_2]} \mathrm{E}U_{ni}U_{nk}\right| \\
&\quad \le \left\|\sum_{i=[ns_1]+1}^{[nt_1]} U_{ni}\right\|_2 \left\|\sum_{k=[ns_2]+1}^{[n(s_2+\delta)]} U_{nk}\right\|_2 \\
&\qquad\qquad + \sum_{i=1}^{n}\sum_{k=1}^{n} |\mathrm{E}U_{ni}U_{nk}| 1\big(|i-k| \ge [n(s_2+\delta)] - [nt_1]\big).
\end{aligned} \tag{31.21}$$

The first term of the majorant of (31.21) can be made as close to zero as desired as $n \to \infty$ by choice of δ, because

$$\lim_{\delta\to 0}\lim_{n\to\infty} \left\|\sum_{k=[ns_2]+1}^{[n(s_2+\delta)]} U_{nk}\right\|_2 = 0$$

by assumption (d), whereas the second term vanishes as $n \to \infty$ for any $\delta > 0$ by Corollary **17.18**. Since δ is arbitrary, this completes the proof. ∎

31.4 Nonstationary Increments

To develop a fully general theory of weak convergence of partial sum processes, permitting global heterogeneity of the increments with possibly trending moments and particularly to accommodate the multivariate case, the class of limit processes must be extended beyond ordinary Brownian motion. Consider processes X_n whose limits X are Gaussian and almost surely continuous, but lack the Brownian property $\mathrm{E}(X^2(t)) = t$. This phenomenon may arise if the increment processes are nonstationary.

The desired generalization has already been introduced as Example **28.14**, but now consider the theory of these processes a little more formally. *A transformed* (or *variance-transformed*) *Brownian motion* B_η will be defined as a stochastic process on $[0,1]$ with finite-dimensional distributions given by

$$B_\eta(t) \overset{\mathrm{d}}{\sim} B\big(\eta(t)\big), t \in [0,1] \tag{31.22}$$

where B is a Brownian motion and η is an increasing homeomorphism on $[0,1]$ with $\eta(0) = 0$. The increments of this process, $B_\eta(t) - B_\eta(s)$ for $0 \leq t < s \leq 1$, are therefore independent and Gaussian with mean 0 and variance $\eta(t) - \eta(s)$. Since $\eta(1)$ must be finite, the condition $\eta(1) = 1$ is achieved by a trivial normalization.

To appreciate the relevance of these processes, consider as was done in §28.3 the characterization of Brownian motion B as the limit of a partial-sum process with independent and identically distributed Gaussian summands. Here, by contrast, the variance of the terms of the sum changes with time. Consider for simplicity an independent sequence $\zeta_i \sim_{\mathrm{d}} \mathrm{N}(0, \sigma_i^2)$ for $i = 1, \dots, n$ where with $s_n^2 = \sum_{i=1}^n \sigma_i^2$,

$$\frac{s_{[nt]}^2}{s_n^2} \to \eta(t) \text{ as } n \to \infty \tag{31.23}$$

and the limit function $\eta : [0,1] \mapsto [0,1]$ is continuous and strictly increasing everywhere. Then,

$$\frac{1}{s_n}\sum_{i=1}^{[nt]} \zeta_i \xrightarrow{\mathrm{d}} B_\eta(t) \tag{31.24}$$

for each $t \in [0,1]$.

What mode of evolution of the variances might satisfy (31.23) and give rise to this limiting result? In the globally stationary case where the sequence $\{\sigma_i^2\}_{i=1}^\infty$ is Cesàro-summable and the Cesàro limit is strictly positive, $\eta(t) = t$ is the only possible limit for (31.23). This conclusion extends to any case where the variances are uniformly bounded and the limit exists; however, the fact that uniform boundedness of the variances is not sufficient for the limit to exist is illustrated by **25.9**. (Try evaluating the sequence in (31.23) for this case.)

Alternatively, consider the example in **28.14**. Here, surprisingly enough, the partial sums of independent increments have a well-defined limit process even when the Cesàro limit of the variances is either zero or infinity, as in the respective cases $0 < \theta < 1$ and $\theta > 1$. In principle, the process B_η can be defined for any increasing function η with the prescribed properties, but the fact that it is defined as the limiting case of a partial-sum process means that **28.14** is a more general formulation than it may appear. If s_n^2 satisfies (31.23), it is regularly varying at infinity according to **2.30** and $\eta(t) = t^\theta$ for some $\theta > 0$. The logic underlying these features of the limit is easy to grasp. If the function s_n^2 were to contain an inflection point at any finite value of its argument, since new increments are added only at the end this would be necessarily be 'sucked towards the origin' as n increased without limit. If it does not grow linearly with n its slope must eventually either increase or decrease monotonically. With inflection points arising at every sample size in the manner of **25.9**, no fixed limit function η can exist and there is no convergence. This does not mean that B_η processes with more general forms of η cannot exist; simply, that they cannot be generated as the limit of partial sums.

The calculation in **28.14** is easy thanks to the assumption of independent increments. The same conclusion is obtained by a somewhat more elaborate argument in the presence of autocorrelation, subject to summability of the autocovariances.

31.6 Theorem Let $\{U_i\}_{i=1}^n$ denote a sequence satisfying the conditions of **31.4**. For the càdlàg process

$$X_n(t) = \frac{\sqrt{\theta}}{\sigma n^{\theta/2}} \sum_{j=1}^{[nt]} j^{(\theta-1)/2} U_j \tag{31.25}$$

with $\theta > 0$,

$$\mathrm{E}\big(X_n^2(t)\big) \to t^\theta \text{ as } n \to \infty. \tag{31.26}$$

Proof Choose a monotone sequence $\{b_n \in \mathbb{N}\}$ such that $b_n \to \infty$ but $b_n/n \to 0$; $b_n = [n^{1/2}]$ will serve. Putting $r_n = [nt/b_n]$ for $t \in (0,1]$ with n large enough that $r_n \geq 1$, (31.25) can be written as

$$X_n(t) = \frac{\sqrt{\theta}}{\sigma n^{\theta/2}} \left(\sum_{i=1}^{r_n} \left(\sum_{j=(i-1)b_n+1}^{ib_n} j^{(\theta-1)/2} U_j \right) + \sum_{j=r_n b_n+1}^{[nt]} j^{(\theta-1)/2} U_j \right). \tag{31.27}$$

The r_n blocks of this sum have the decomposition

$$\sum_{j=(i-1)b_n+1}^{ib_n} j^{(\theta-1)/2}U_j = i^{(\theta-1)/2}b_n^{\theta/2}S_{ni} + i^{(\theta-1)/2}b_n^{\theta/2}S_{ni}^* \tag{31.28}$$

in which

$$S_{ni} = b_n^{-1/2}\sum_{j=(i-1)b_n+1}^{ib_n} U_j$$

and

$$S_{ni}^* = b_n^{-1/2}a_{ni}^*\sum_{j=(i-1)b_n+1}^{ib_n} \frac{a_{nij}}{a_{ni}^*}U_j$$

where

$$a_{nij} = \frac{j^{(\theta-1)/2} - (b_n i)^{(\theta-1)/2}}{i^{(\theta-1)/2}b_n^{\theta/2}},$$

$a_{n1}^* = b_n^{-1/2}$, and for $i > 1$, a_{ni}^* is calculated as follows. In the case $\theta > 1$, $a_{nij} \le 0$ and by the mean value theorem,

$$\begin{aligned}|a_{nij}| \le \frac{i^{(\theta-1)/2} - (i-1)^{(\theta-1)/2}}{i^{(\theta-1)/2}b_n^{1/2}} &= \frac{(\theta-1)(1-\gamma/i)^{(\theta-3)/2}}{2ib_n^{1/2}},\ 0 \le \gamma \le 1 \\ &\le \frac{\theta-1}{2ib_n^{1/2}}\begin{cases}(1-1/i)^{(\theta-3)/2}, & 1<\theta<3 \\ 1, & \theta \ge 3\end{cases} \\ &= a_{ni}^*\end{aligned}$$

where the last equality defines a_{ni}^*. Similarly when $0 < \theta < 1$, $a_{nij} \ge 0$ and

$$|a_{nij}| \le a_{ni}^* = \frac{|\theta-1|}{2ib_n^{1/2}}(1-1/i)^{(\theta-3)/2}.$$

In either case the weights a_{nij}/a_{ni}^* are all of the same sign (depending on θ) and $0 \le |a_{nij}/a_{ni}^*| \le 1$. Similar results apply to the term S_{n,r_n+1}, corresponding to the residual sum in (31.27).

Now consider $\mathrm{E}(X_n^2(t))$. Multiplying out the square of (31.27) after substituting (31.28), there are three types of summand: those involving squares and products of the S_{ni} ($(r_n+1)^2$ terms), those involving squares and products of the S_{ni}^* ($(r_n+1)^2$ terms), and those involving products of S_{ni}^* with S_{ni} ($2(r_n+1)^2$ terms). The sum of the first type has the form

$$\frac{1}{n^{\theta}}\mathrm{E}\Bigg(\sum_{i=1}^{r_n+1} i^{(\theta-1)/2} b_n^{\theta/2} S_{ni}\Bigg)^2 = \frac{r_n^{\theta} b_n^{\theta}}{n^{\theta}}\Bigg(\frac{1}{r_n^{\theta}}\sum_{i=1}^{r_n+1} i^{\theta-1}\mathrm{E}(S_{ni}^2) + \frac{2}{r_n^{\theta}}\sum_{i=2}^{r_n+1}\sum_{m=1}^{i-1} i^{(\theta-1)/2}(i-m)^{(\theta-1)/2}\mathrm{E}(S_{ni}S_{n,i-m})\Bigg). \tag{31.29}$$

Note that $r_n b_n/n \to t$ and that $r_n^{-\theta}\sum_{i=1}^{r_n+1} i^{\theta-1} \to 1/\theta$. Letting $\{\gamma_p\}_0^{\infty}$ denote the autovariance sequence the assumptions imply, by **17.16** and the discussion following, that $\gamma_p = O(p^{-1-\delta})$ for $\delta > 0$. Hence, $\mathrm{E}(S_{ni}^2) \to \sigma^2 < \infty$ for all i. For $m \geq 1$, $|\gamma_{p+b_n m}| = O(b_n^{-1-\delta} m^{-1-\delta})$ when $p < b_n$ and it follows that for all i,

$$|\mathrm{E}(S_{ni}S_{n,i-m})| \leq \frac{2}{b_n}\sum_{p=0}^{b_n-1}(b_n - p)|\gamma_{p+b_n m}| = O(m^{-1-\delta} b_n^{-\delta}).$$

Using **2.19** and **2.17** gives

$$\frac{2}{b_n^{\delta} r_n^{\theta}}\sum_{i=2}^{r_n+1} i^{(\theta-1)/2}\sum_{m=1}^{i-1}(i-m)^{(\theta-1)/2} m^{-1-\delta} = O(b_n^{-\delta}).$$

Hence the cross-product terms vanish and expression (31.29) converges to $t^{\theta}\sigma^2/\theta$.

Considering the terms of the second type, note that $\mathrm{E}(S_{ni}^{*2}) \leq \mathrm{E}(S_{ni}^2)$ since every term in the double sum of the minorant has a weight from $[0,1]$. Therefore

$$\frac{1}{n^{\theta}}\sum_{i=1}^{r_n+1} i^{\theta-1} a_{ni}^{*2}\mathrm{E}(S_{ni}^{*2}) = O(n^{-\theta} r_n^{\theta-2} b_n^{-1}) = o(1).$$

Noting $|\mathrm{E}(S_{ni}S_{ni}^{*})| \leq |\mathrm{E}(S_{ni}^2)|$, the contemporaneous products of the third type are similarly bounded by

$$\frac{1}{n^{\theta}}\sum_{i=1}^{r_n+1} i^{\theta-1} a_{ni}^{*}|\mathrm{E}(S_{ni}S_{ni}^{*})| = O(n^{-\theta} r_n^{\theta-1} b_n^{-1/2}) = o(1).$$

Hence, all the components involving the S_{ni}^{*} vanish asymptotically and (31.26) follows. ∎

The results obtained in §28.7 and §31.1 are now found to have generalizations from B to the class B_{η}. For **28.21** there is the following corollary.

31.7 Corollary Let the conditions of Theorem **28.21** hold, except that condition (a) is replaced by (a′) $\mathrm{E}(X(t)) = 0$, $\mathrm{E}(X^2(t)) = \eta(t)$, $0 \le t \le 1$.

Then $X \sim_{\mathrm{d}} B_\eta$.

Proof Define $X^*(t) = X\big(\eta^{-1}(t)\big)$ and apply **28.21** to X^*. $\eta^{-1}(\cdot)$ is continuous, so condition **28.21**(b) continues to hold. Strict monotonicity ensures that if $\{t_1, \dots, t_m\}$ define arbitrary non-overlapping intervals, so also do $\{\eta^{-1}(t_1), \dots, \eta^{-1}(t_m)\}$, so **28.21**(c) continues to hold. Finally, $\mathrm{E}(X^*(t)^2) = \eta^{-1}(\eta(t)) = t$ and condition (a) holds. Therefore $X^* \sim_{\mathrm{d}} B$ and the argument is completed by noting $X(t) = X^*(\eta(t))$. ∎

Similarly, for **31.1** there is the following corollary.

31.8 Corollary Let the conditions (a), (b), and (c) of **31.1** hold, and instead of condition **31.1**(d) assume

(d) $\mathrm{E}\big(X_n(t)\big) \to 0$ and $\mathrm{E}\big(X_n^2(t)\big) \to \eta(t)$ as $n \to \infty$, for each $t \in [0, 1]$.

Then $X_n \to_{\mathrm{d}} B_\eta$.

Proof The argument in the proof of **31.1** shows that the conditions of **31.7** hold for X. ∎

However, proving the generalized FCLT introduces some new difficulties. Consider the case $\eta(t) = t^\theta$ for $0 < \theta < 1$. Applying **31.3** to the sequence $X_n(t) = n^{-\theta/2} \sum_{i=1}^{[nt]} i^{(\theta-1)/2} U_i$ requires setting $c_{ni} = n^{-\theta/2} i^{(\theta-1)/2}$. It is easy to see that conditions **31.3**(c) and **31.5**(c) are not satisfied since $\max_{1 \le i \le n} c_{ni} = O(n^{-\theta/2})$. Theorems **31.3** and **31.5** therefore fail. A specific problem with **31.3** is in the convergence in (31.16) where condition (b) of the theorem permits

$$\sum_{m=0}^{[nt_1]-[ns_1]-1} \nu_m = o(n^{1/2})$$

which has to be balanced by $\max_{[ns_1] < i \le [nt_1]} c_{ni} = O(n^{-1/2})$. This step might evidently be rescued by requiring a stronger restriction on the NED size. However, there is also a problem with condition **31.3**(d). Note that by (2.18),

$$\limsup_{n\to\infty} \frac{1}{\delta} \sum_{i=[nt]+1}^{[n(t+\delta)]} c_{ni}^2 = \frac{1}{\theta} \frac{(t+\delta)^\theta - t^\theta}{\delta} = t^{\theta-1} + O(\delta) \tag{31.30}$$

whose supremum over $t \in [0,1]$ diverges as $\delta \to 0$ when $\theta < 1$.

Theorem **31.5** fails in the same way, through the violation of condition (c). However, in this case there is a remedy since this theorem uses a different argument to establish asymptotic independence, thanks to the direct derivation of the finite-dimensional distributions. Theorem **25.12** contains a less restrictive condition than condition (c), that is, condition **25.6**(c) which for the purpose of proving convergence to Brownian motion has been left unexploited. The simplest way to see the effect of letting this assumption replace **31.5**(c) is by reference to Lemma **25.10**. Here, the specification in **31.6** is matched by setting $\beta = \gamma = \theta - 1$, which satisfies (25.36). All that is required is that in the application of the central limit theorem, the block dimension function b_n defined in **25.6**(c) must be set to $[n^{1-\alpha}]$ for $\alpha > 1 - \theta$ according to (25.37), which is feasible for any $\theta > 0$.

As noted previously, partial-sum processes must give rise to the class of functions $\eta(t) = t^{\theta}$ in the limit so these conclusions are quite general. The upshot of this is that a modified form of **31.5** provides a proof of the generalized FCLT, as follows.

31.9 Theorem For a stochastic triangular array $\{U_{ni}\}_{i=1}^{n}$ where

$$U_{ni} = \sqrt{\theta}\sigma^{-1}n^{-\theta/2}i^{(\theta-1)/2}U_i$$

and U_i satisfies the conditions of Corollary **31.4**, let $X_n(t) = \sum_{i=1}^{[nt]} U_{ni}$. Then $X_n \to_{\mathrm{d}} B_{\eta}$ as $n \to \infty$.

Proof The proof of Theorem **31.5** proceeds by deriving the finite-dimensional distributions, showing that under conditions (a), (b), and (c) of the theorem the conditions of Theorem **25.12** are satisfied. Uniform tightness is then shown to hold by condition (d) and the form of the limit distribution follows from condition (e). This chain of reasoning is now modified as follows.

Define $c_{ni} = \sqrt{\theta}\sigma^{-1}n^{-\theta/2}i^{(\theta-1)/2}$, so that $U_{ni}/c_{ni} = U_i$. It is immediate that conditions (a) and (b) of **31.5** hold and hence also conditions (a) and (b) of **25.12** with suitable changes of notation. Consider condition (c) of Theorem **25.6**. It is easy to verify using **25.10** that (25.21) and (25.22) hold for $\alpha > 1 - \theta$. In conjunction with conditions (a) and (b), these are sufficient for the application of Theorem **25.12** to the finite-dimensional distributions.

To show that condition (d) of **31.5** holds is a matter of verifying that the numerator in the second member of (31.30) vanishes as $\delta \to 0$ for all $t \in [0,1]$, an unproblematic requirement for $\theta > 0$. The argument of Theorem **31.5** therefore establishes uniform tightness under the present assumptions. Finally, Theorem **31.6** has shown that $\mathrm{E}(X_n^2(t)) \to t^{\theta}$ under the stated assumptions and this fact can replace condition (e) of **31.5**, whose only role is to identify the form of the limit process. ∎

Perhaps the most important conclusion to be drawn from **31.9** is that convergence to B_η holds under effectively the same conditions as convergence to B, which is no more than the special case for which $\eta(t) = t$. The extra generality may be useful only occasionally, but is free of penalties.

However, there is one remaining question: could there exist processes for which **25.6**(c) holds but **31.5**(d) fails? It turns out to be a challenge to construct a plausible example with this property. The following case is sufficiently exotic as to support the view that this is not a question to be very concerned about. Let **31.5**(a) hold for $c_{ni}^2 = n^{-1/3} i^{1/4} 1_{\{i \le n^{4/15}\}}$, implying that $U_{ni} = 0$ for $i > n^{4/15}$. However, $\max_{1 \le i \le n} c_{ni} = o(1)$ and by **2.17**(i),

$$\limsup_{n\to\infty} \sum_{i=1}^{n} c_{ni}^2 = \limsup_{n\to\infty} \frac{1}{n^{1/3}} \sum_{i=1}^{n^{4/15}} i^{1/4} = \frac{4}{5}.$$

Setting $b_n = 1$ satisfies **25.6**(c). However, compare **31.5**(d) with

$$\begin{aligned} \sup_{t\in[0,1.\delta]} \limsup_{n\to\infty} \sum_{i=[nt]+1}^{[n(t+\delta)]} c_{ni}^2 &\ge \limsup_{n\to\infty} \frac{1}{n^{1/3}} \sum_{i=1}^{[n\delta]} i^{1/4} 1_{\{i \le n^{4/15}\}} \\ &= \min\left\{ \limsup_{n\to\infty} \frac{1}{n^{1/3}} \sum_{i=1}^{[n\delta]} i^{1/4}, \frac{4}{5} \right\}. \end{aligned}$$

Since the first term in the last member is divergent, letting $\delta \to 0$ does not cause it to vanish and **31.5**(d) does fail. Note however that the limit process in question has a discontinuity at the origin, with $X(0) = 0$ and $X(t) = Y$ a.s. for $0 < t \le 1$, where Y is distributed as $\mathrm{N}(0, 4/5)$.

31.5 Generalized Partial Sums

The nonstationarity issue can be addressed in a different way by forming the partial sums of the increments according to a modified rule. Let $\{K_n(t), n \in \mathbb{N}\}$ represent a sequence of integer-valued, right-continuous, increasing functions of t with $K_n(0) = 0$ for all n and $K_n(t) - K_n(s) \to \infty$ as $n \to \infty$ if $t > s$. Then the partial-sum process

$$X_n^K(t) = \sum_{i=1}^{K_n(t)} U_{ni},\ 0 \le t \le 1 \tag{31.31}$$

is constructed by accumulating increments at different rates in different parts of the sample. By choice of K_n, it is possible both to obtain convergence to B instead of to a different limit process and to obtain convergence which otherwise does not occur at all. If X_n is replaced by X_n^K and $[nt]$ is replaced everywhere by $K_n(t)$ in formulae, then Theorems **31.3** and **31.5** hold without other modification to the proofs.

Two examples illustrate how this might work.

31.10 Example Let the sequence $\{U_i\}$ be serially uncorrelated (just for simplicity) and have variances $\sigma_i^2 = \sigma^2 i^{\theta-1}$ for $\theta > 0$. Then, $s_n^2 = O(n^\theta)$. However, choosing $K_n(t) = [(nt)^{1/\theta}]$ yields

$$\mathrm{E}\biggl(\frac{1}{n^{1/2}}\sum_{i=1}^{K_n(t)} U_i\biggr)^2 = \frac{\sigma^2}{n}\sum_{i=1}^{K_n(t)} i^{\theta-1} = \frac{\sigma^2}{\theta}t + o(1).$$

In this case, Brownian motion is the limit process even when the processes are globally nonstationary as indicated. Of course, θ would need to be known in order to implement this procedure. □

31.11 Example Let a sequence $\{U_i\}$ have the property

$$U_i = \begin{cases} 0 \text{ a.s.}, & 2^{2k} \le i < 2^{2k+1}, k = 0,1,2,3,\ldots \\ \mathrm{N}(0,\sigma^2), & \text{otherwise.} \end{cases}$$

Thus, $U_1 = 0$, $U_4 = U_5 = U_6 = U_7 = 0, U_{16} = U_{17} = \ldots = U_{31} = 0$, and so forth. Let $U_{ni} = U_i/s_n$ as before. If $X_n(t) = \sum_{=1}^{[nt]} U_{ni}$, X_n does not possess a limit in distribution. For example, when $n = 2^k - 1$ for k even, $X_n(t) = X_n(\frac{1}{2})$ for $\frac{1}{2} < t \le 1$ with probability 1, but when $n = 2^k - 1$ for k odd, this occurrence has probability 0. This 'cycling' behaviour is present however large n is. However, let $K_n(t)$ be the integer that satisfies

$$\sum_{i=2}^{K_n(t)} 1(2^{2k-1} \le i < 2^{2k}, k \in \mathbb{N}) = [nt] \tag{31.32}$$

where the indicator is equal to 1 when i is in the indicated range and 0 otherwise. With this arrangement, n counts the actual number of increments in the sum, while $K_n(1)$ counts the nominal number, including the zeros. $K_{n+1}(1) = K_n(1) + 1$, except when $K_n(1) = 2^{2k}$ and in the latter case, $K_{n+1}(1) = 2^{2k+1}$. □

Incidentally, since conditions **31.3**(e) and **31.5**(e) impose

$$\mathrm{E}\bigl(X_n^K(1)^2\bigr) = \mathrm{E}\biggl(\sum_{i=1}^{K_n(1)} U_{ni}\biggr)^2 \to 1 \tag{31.33}$$

one might expect that $K_n(1)/n \to 1$. The last example shows that this is not necessarily the case.

31.6 The Multivariate Case

To extend the FCLT to vector processes requires an approach similar in principle to that of §29.7. To go from univariate to multivariate limits necessitates an appeal to the Cramér–Wold theorem **26.5** and so to consider the finite dimensional sets of D, generalizing the results of §30.4. This section draws on [145].

Define D^m as the space of m-vectors of càdlàg functions endowed with the metric

$$d_B^m(\boldsymbol{x},\boldsymbol{y}) = \max_{1\le j\le m}\{d_B(x_j,y_j)\} \tag{31.34}$$

where d_B is the Billingsley metric (30.21). Be particularly careful to notice that D^m is different from the space $D\times D\times\cdots\times D$ where the Ds represent the space of scalar càdlàg functions. The difference is that in D^m the jump points are synchronized, so that a single $\lambda\in\Lambda$ can serve to operate the Skorokhod class of metrics on the vector elements. The analysis would be different and substantially more complex if it were necessary to define d_B^m in terms of different λ_j functions for each element of the product. Happily, in the case of empirical processes for finite samples the observed jump times are coordinated in precisely this way since they represent common dates, with the limit processes lying in C^m.

d_B^m induces the product topology and the separability of (D,d_S) implies both separability of (D^m,d_B^m) and also that $\mathcal{B}_D^m$, the σ-field generated from the measurable rectangles of (D^m,d_B^m), is the Borel field of (D^m,d_B^m). Also let

$$\mathcal{H}_D^m = \{\pi^{-1}_{t_1,\ldots,t_k}(B)\subseteq D^m \,:\, B\in\mathcal{B}^{mk}, t_1,\ldots,t_k\in[0,1],\ k\in\mathbb{N}\} \tag{31.35}$$

be the finite-dimensional sets of D^m, the field generated from the product of m copies of $\mathcal{H}_D$. The following theorem extends **30.14** in a way that closely parallels the extension of **27.15** to **29.21**.

31.12 Theorem $\mathcal{H}_D^m$ is a determining class for $(D^m,\mathcal{B}_D^m)$.

Proof An open sphere in $\mathcal{B}_D^m$ is

$$\begin{aligned} S(\boldsymbol{x},\alpha) &= \{\boldsymbol{y}\in D^m \,:\, d_B^m(\boldsymbol{x},\boldsymbol{y})<\alpha\} \\ &= \{\boldsymbol{y}\in D^m \,:\, \exists\,\lambda\in\Lambda \text{ s.t.} \|\lambda\|<\alpha,\ \max_{1\le j\le m}\sup_t |y_j(t)-x_j(\lambda(t))|<\alpha\}. \end{aligned} \tag{31.36}$$

For $t_i = i/2^k$, $i = 1, \dots, 2^k - 1$, $k \in \mathbb{N}$, let

$$H_k(\boldsymbol{x}, \alpha) = \{\boldsymbol{y} \in D^m : \exists\, \lambda \in \Lambda \text{ s.t. } \|\lambda\| < \alpha, \\ \max_{1 \le j \le m} \max_{1 \le i \le 2^k - 1} |y_j(t_i) - x_j(\lambda(t_i))| < \alpha\} \in \mathcal{H}_D^m. \tag{31.37}$$

A direct generalization of the argument of **30.14** gives the result for any $\boldsymbol{x} \in D^m$ and $r > 0$,

$$S(\boldsymbol{x}, r) = \bigcup_{n=1}^{\infty} \bigcap_{k=1}^{\infty} H_k(\boldsymbol{x}, r - 1/n) \in \sigma(\mathcal{H}_D^m). \tag{31.38}$$

Hence, $\mathcal{B}_D^m \subseteq \sigma(\mathcal{H}_D^m)$ as required. ∎

The following can be thought of as a generic multivariate convergence theorem, in that the weak limit specified need only be a.s. continuous. It is not necessarily vector Brownian motion.

31.13 Theorem Let $\boldsymbol{X}_n \in D^m$ be an m-vector of random elements. $\boldsymbol{X}_n \to_{\mathrm{d}} \boldsymbol{X}$, where $P(\boldsymbol{X} \in C^m) = 1$, iff $\boldsymbol{\lambda}'\boldsymbol{X}_n \to_{\mathrm{d}} \boldsymbol{\lambda}'\boldsymbol{X}$ for every fixed $\boldsymbol{\lambda}$ $(m \times 1)$ with $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$ where $P(\boldsymbol{\lambda}'\boldsymbol{X} \in C) = 1$.

Proof If $x_j \in D, j = 1, \dots, m$, $\sum_{j=1}^m \lambda_j x_j$ possesses a left limit and is continuous on the right since for $t \in [0, 1)$,

$$\lim_{\varepsilon \downarrow 0} \sum_{j=1}^m \lambda_j x_j(t + \varepsilon) = \sum_{j=1}^m \lambda_j \lim_{\varepsilon \downarrow 0} x_j(t + \varepsilon) = \sum_{j=1}^m \lambda_j x_j(t). \tag{31.39}$$

Hence, $\boldsymbol{x} = (x_1, \dots, x_m)' \in D^m$ implies $\boldsymbol{\lambda}'\boldsymbol{x} \in D$. It follows that $\boldsymbol{\lambda}'\boldsymbol{X}_n$ is a random element of D. Similarly, $\boldsymbol{x} \in C^m$ implies $\boldsymbol{\lambda}'\boldsymbol{x} \in C$ and hence $P(\boldsymbol{\lambda}'\boldsymbol{X} \in C) = 1$.

Let μ_n denote the p.m. of $\boldsymbol{X}_n$ and μ the p.m. of $\boldsymbol{X}$. To prove sufficiency let μ_n^λ denote the sequence of measures corresponding to $\boldsymbol{\lambda}'\boldsymbol{X}_n$ and assume $\mu_n^\lambda \Rightarrow \mu^\lambda$. Fix $t_1, \dots, t_k \in [0, 1]$ for finite k. Noting that $\pi_{t_1,\dots,t_k}^{-1}(B) \cap D \in \mathcal{H}_D \subseteq \mathcal{B}_D$ for each $B \in \mathcal{B}^k$ (see **30.14**), the projections are measurable and $\nu_n^\lambda = \mu_n^\lambda \pi_{t_1,\dots,t_k}^{-1}$ is a measure on $(\mathbb{R}^k, \mathcal{B}^k)$. Although $\pi_{t_1,\dots,t_k}$ is not continuous (see the discussion in §30.2), the stipulation $\mu^\lambda(C) = 1$ implies that the discontinuity points have μ^λ-measure 0 and hence $\nu_n^\lambda \Rightarrow \nu^\lambda$ by the continuous mapping theorem (**29.4**). Since ν_n^λ is the p.m. of a k-vector of r.v.s and $\boldsymbol{\lambda}$ is arbitrary, the Cramér–Wold theorem (**26.5**) implies that $\nu_n \Rightarrow \nu$ where $\nu_n = \mu_n \pi_{t_1,\dots,t_k}^{-1}$ is the p.m. of an mk-vector, the distribution of $\boldsymbol{X}_n(t_1), \dots, \boldsymbol{X}_n(t_k)$. Since $t_1, \dots, t_k$ are arbitrary, the finite-dimensional distributions of $\boldsymbol{X}_n$ converge.

Showing $\{\mu_n\}$ to be uniformly tight completes the proof of sufficiency. Choose $\boldsymbol{\lambda} = \boldsymbol{e}_j$, the vector with 1 in position j and 0 elsewhere, to show that $X_{nj} \to_{\mathrm{d}} X_j$; this means the marginal p.m.s are uniformly tight and so $\{\mu_n\}$ is uniformly tight by **29.16**. Then $\boldsymbol{X}_n \to_{\mathrm{d}} \boldsymbol{X}$ by **31.12**.

To show necessity, simply apply the continuous mapping theorem **29.4** to the continuous functional $h(\boldsymbol{x}) = \boldsymbol{\lambda}'\boldsymbol{x}$. ∎

Although this is a general result, note the importance of the requirement $\mu(C^m) = 1$. It is easy to devise a counterexample where this condition is violated, in which case convergence fails.

31.14 Example Suppose μ is the p.m. on $(D, \mathcal{B}_D)$ which assigns probability 1 to elements x with

$$x(t) = \begin{cases} 0, \; t < \frac{1}{2} \\ 1, \; t \geq \frac{1}{2}. \end{cases}$$

Also, let μ_n assign probability 1 to elements with

$$x(t) = \begin{cases} 0, \; t < \frac{1}{2} + 1/n \\ 1, \; t \geq \frac{1}{2} + 1/n. \end{cases}$$

If $X_{1n} \sim_{\mathrm{d}} \mu$ all n and $X_{2n} \sim_{\mathrm{d}} \mu_n$, then clearly $(X_{1n}, X_{2n}) \to_{\mathrm{d}} (X_1, X_2) = (x, x)$ w.p.1. But $X_{2n} - X_{1n}$ is equal w.p.1 to the function in (30.19), which does not converge in (D, d_B). □

The next step is to review how a multivariate process might satisfy the conditions of **31.13**.

31.15 Theorem Let $\{\boldsymbol{U}_{ni}\}$ be an array of zero-mean stationary stochastic m-vectors, where $\boldsymbol{U}_{ni} = n^{-1/2}\boldsymbol{U}_i$ with $\mathrm{E}(\boldsymbol{U}_i\boldsymbol{U}_i') = \boldsymbol{\Sigma}$ and $\mathrm{E}(\boldsymbol{U}_i\boldsymbol{U}_{i-k}') = \boldsymbol{\Lambda}_k$ for $k \geq 1$, so that $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\Lambda} + \boldsymbol{\Lambda}'$ where $\boldsymbol{\Lambda} = \sum_{k=1}^{\infty} \boldsymbol{\Lambda}_k$. Let $\boldsymbol{X}_n(t) = \sum_{i=1}^{[nt]} \boldsymbol{U}_{ni}$. If

(a) for each fixed m-vector $\boldsymbol{\lambda}$ satisfying $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$ there exists a scalar array $\{c_{ni}^{\lambda}\}$ such that the conditions of **31.5** hold for the arrays $\{\boldsymbol{\lambda}'\boldsymbol{U}_{ni}\}$ and $\{c_{ni}^{\lambda}\}$

(b) $\mathrm{E}\big(\boldsymbol{X}_n(t)\boldsymbol{X}_n(t)'\big) \to t\boldsymbol{\Omega}$ as $n \to \infty$

then $\boldsymbol{X}_n \to_{\mathrm{d}} \boldsymbol{B}(\boldsymbol{\Omega})$, an m-vector of Brownian motions with covariance matrix $\boldsymbol{\Omega}$.

Proof By **31.13**. ∎

The point not to be overlooked is that $\boldsymbol{\Omega}$ and not $\boldsymbol{\Sigma}$ is the relevant covariance matrix, notwithstanding that $\boldsymbol{B}$ has independent increments.

To permit nonstationary increments following the results of §31.4, there is little sacrifice of generality in considering the increment process

$$\boldsymbol{U}_{ni} = \boldsymbol{J}_{ni}\boldsymbol{U}_i \quad (m \times 1) \tag{31.40}$$

where $\boldsymbol{U}_i$ is as before and

$$\boldsymbol{J}_{ni} = \operatorname{diag}\{n^{-\theta_1/2} i^{(\theta_1-1)/2}, \dots, n^{-\theta_m/2} i^{(\theta_m-1)/2}\}$$

where $(\theta_1, \dots, \theta_m)'$ is a vector of trend coefficients. Define $\widetilde{\boldsymbol{\Sigma}} = \mathrm{E}(\boldsymbol{U}_i^2)$ and $\widetilde{\boldsymbol{\Lambda}}_k = \mathrm{E}(\boldsymbol{U}_i\boldsymbol{U}_{i-k}')$ and hence $\widetilde{\boldsymbol{\Omega}} = \widetilde{\boldsymbol{\Sigma}} + \widetilde{\boldsymbol{\Lambda}} + \widetilde{\boldsymbol{\Lambda}}'$ where $\widetilde{\boldsymbol{\Lambda}} = \sum_{k=1}^{\infty} \widetilde{\boldsymbol{\Lambda}}_k$. Then,

$$\mathrm{E}(\boldsymbol{U}_{ni}\boldsymbol{U}_{ni}') = \boldsymbol{J}_{ni}\widetilde{\boldsymbol{\Sigma}}\boldsymbol{J}_{ni}$$

and

$$\mathrm{E}(\boldsymbol{U}_{ni}\boldsymbol{U}_{n,i-k}') = \boldsymbol{J}_{ni}\widetilde{\boldsymbol{\Lambda}}_k\boldsymbol{J}_{n,i-k}.$$

Noting that $\boldsymbol{J}_{n,i-k} = \boldsymbol{J}_{ni} + O(n^{-\min_j\{\theta_j/2\}})$ and that the autocovariances are summable, the variance matrix is

$$\begin{aligned}
\mathrm{E}\big(\boldsymbol{X}_n(t)\boldsymbol{X}_n(t)'\big) &= \sum_{i=1}^{[nt]} \boldsymbol{J}_{ni}\widetilde{\boldsymbol{\Sigma}}\boldsymbol{J}_{ni} + \sum_{i=1}^{[nt]}\sum_{k=1}^{[nt]-i} (\boldsymbol{J}_{ni}\widetilde{\boldsymbol{\Lambda}}_k\boldsymbol{J}_{n,i-k} + \boldsymbol{J}_{n,i-k}\widetilde{\boldsymbol{\Lambda}}_k'\boldsymbol{J}_{ni}) \\
&= \sum_{i=1}^{[nt]} \boldsymbol{J}_{ni}\widetilde{\boldsymbol{\Omega}}\boldsymbol{J}_{ni} + o(1) \\
&\to \boldsymbol{H}(t)\boldsymbol{\Omega}\mathbf{H}(t)
\end{aligned} \tag{31.41}$$

where $\boldsymbol{H}(t) = \operatorname{diag}\{t^{\theta_1/2}, \dots, t^{\theta_m/2}\}$ and the matrix $\boldsymbol{\Omega}$ has $(j,l)^{\text{th}}$ element $\omega_{jl} = 2\tilde{\omega}_{jl}/(\theta_j + \theta_l)$ where $\tilde{\omega}_{jl}$ is the corresponding element of $\widetilde{\boldsymbol{\Omega}}$. Specifically, applying Theorem **2.17**(i) the convergence for the (j,l) element in the final step of (31.41) has the form

$$\begin{aligned}
\tilde{\omega}_{jl}\sum_{i=1}^{[nt]} \frac{i^{(\theta_j-1)/2} i^{(\theta_l-1)/2}}{n^{\theta_j/2} n^{\theta_l/2}} &= \frac{\tilde{\omega}_{jl}}{n^{(\theta_j+\theta_l)/2}} \sum_{i=1}^{[nt]} i^{(\theta_j+\theta_l)/2-1} \\
&= \frac{2\tilde{\omega}_{jl}[nt]^{(\theta_j+\theta_l)/2}}{(\theta_j+\theta_l)n^{(\theta_j+\theta_l)/2}} \to \omega_{jl}t^{(\theta_j+\theta_l)/2}.
\end{aligned}$$

Since $\boldsymbol{H}(1) = \boldsymbol{I}_m$ it should be noted that $\boldsymbol{S}_n = \mathrm{E}\big(\sum_{i=1}^n \boldsymbol{U}_{ni}\big)\big(\sum_{i=1}^n \boldsymbol{U}'_{ni}\big) \to \boldsymbol{\Omega}$. This is the matrix that would be estimated by the usual type of kernel estimator, as in (26.36). $\widetilde{\boldsymbol{\Omega}} = \boldsymbol{\Omega}$ in stationary series with $\theta_1 = \cdots = \theta_m = 1$.

Let $\boldsymbol{B}_\eta(\boldsymbol{\Omega})$ denote the family of $m \times 1$ vector a.s. continuous Gaussian processes on $[0,1]$ with independent increments and covariance matrix $\boldsymbol{H}(t)\boldsymbol{\Omega}\mathbf{H}(t)$ at t. The following is the multivariate generalization of Theorem **31.9**.

31.16 Theorem If $\boldsymbol{X}_n(t) = \sum_{i=1}^{[nt]} \boldsymbol{U}_{ni}$ where $\boldsymbol{U}_{ni}$ $(m \times 1)$ is defined by (31.40) and the elements of $\boldsymbol{U}_i$ satisfy the conditions of Corollary **31.4**, then $\boldsymbol{X}_n \to_{\mathrm{d}} \boldsymbol{B}_\eta(\boldsymbol{\Omega})$ having the limiting covariance matrix in (31.41). □

Proof In view of **31.13** it is sufficient to verify conditions for the convergence of the arrays $\{\boldsymbol{\lambda}'\boldsymbol{U}_{ni}\}$ for each fixed m-vector $\boldsymbol{\lambda}$ satisfying $\boldsymbol{\lambda}'\boldsymbol{\lambda} = 1$. Relating the argument to that of **31.9** the point of interest in particular is the form of the scale constants $c_{ni}^{\lambda} = (\boldsymbol{\lambda}'\boldsymbol{J}_{ni}\widetilde{\boldsymbol{\Omega}}\boldsymbol{J}_{ni}\boldsymbol{\lambda})^{1/2}$. Suppose without loss of generality, relabelling elements as required, that $\theta_1 \le \theta_j$ for every j. Also assume $\lambda_1 \ne 0$ since otherwise X_{1n} drops out. For this case note that

$$(c_{ni}^{\lambda})^2 = \sum_{j=1}^{m}\sum_{l=1}^{m} \lambda_j\lambda_l\tilde{\omega}_{jl}\frac{i^{(\theta_j+\theta_l)/2-1}}{n^{(\theta_j+\theta_l)/2}} = \frac{i^{\theta_1-1}}{n^{\theta_1}}B_{ni}^{\lambda} \tag{31.42}$$

where

$$B_{ni}^{\lambda} = \sum_{j=1}^{m}\sum_{l=1}^{m} \lambda_j\lambda_l\tilde{\omega}_{jl}\Big(\frac{i}{n}\Big)^{(\theta_j+\theta_l)/2-\theta_1}. \tag{31.43}$$

B_{ni}^{λ} is strictly positive and also bounded, since for any choice of n and $i = 1, \ldots, n$ it is a positive definite quadratic form in $\widetilde{\boldsymbol{\Omega}}$ with respect to weights $\lambda_j(i/n)^{(\theta_j-\theta_1)/2}$ for $j = 1, \ldots m$. These may vary between 0 and λ_j, but also note that $B_{n0}^{\lambda} = \lambda_1^2\tilde{\omega}_{11} > 0$ as well as $B_{nn}^{\lambda} = \boldsymbol{\lambda}'\widetilde{\boldsymbol{\Omega}}\boldsymbol{\lambda}$.

Therefore, consider the argument of Theorem **31.9**. (The normalization to unit variance adopted in that case is of course an irrelevant detail.) Replacing $\theta\sigma^{-2}$ by B_{ni}^{λ} and θ by $\min\{\theta_1, \ldots, \theta_n\}$ in the definition of c_{ni}^2, the array $\{\boldsymbol{\lambda}'\boldsymbol{U}_{ni}/c_{ni}^{\lambda}\}$ satisfies the conditions of **31.4** by construction. Since conditions (25.21) and (25.22) relate to orders of magnitude, their validity is not affected by the fact that B_{ni}^{λ} may vary over a positive and finite range. Also, following the formulation of (31.41), note that

$$\limsup_{n\to\infty} \sum_{i=[nt]+1}^{[n(t+\delta)]} (c_{ni}^{\lambda})^2 = \sum_{j=1}^{m}\sum_{l=1}^{m} \lambda_j\lambda_l\tilde{\omega}_{kl}\big((t+\delta)^{(\theta_j+\theta_l)/2} - t^{(\theta_j+\theta_l)/2}\big). \tag{31.44}$$

This expression converges to zero with δ, uniformly with respect to $t \in [0,1]$, so that condition **31.5**(d) is satisfied. The argument of **31.9** therefore holds for this case, completing the proof. ∎

The conclusion is that the conditions for convergence to variance-transformed Brownian motions are satisfied provided this is the case for convergence to regular Brownian motions when trends are absent.

32
Weak Convergence to Stochastic Integrals

32.1 Weak Limit Results for Random Functionals

The main task of this chapter is to examine an important corollary to the functional central limit theorem: convergence of a particular class of partial sums to a limit distribution that can be identified with a stochastic integral with respect to Brownian motion, or possibly another Gaussian diffusion process. As a preliminary, another class of results involving integrals is reviewed, superficially similar to what follows but actually different and rather more straightforward. There will turn out to be an unexpected correspondence in certain cases between the results obtained by each approach.

While for a probability space $(\Omega, \mathcal{F}, P)$ the notion of a measurable mapping $f: \Omega \mapsto C$ is familiar where C is $C_{[0,1]}$ as usual, measurability now has to be extended to functionals on C and especially to integrals. Let $F(t) = \int_0^t f \mathrm{d}s : C \mapsto \mathbb{R}$ denote the ordinary Riemann integrals of f over $[0, t]$.

32.1 Theorem If f is $\mathcal{F}/\mathcal{B}_C$-measurable, the composite mapping

$F(t) \circ f : \Omega \mapsto \mathbb{R}$

is $\mathcal{F}/\mathcal{B}$-measurable for $t \in [0,1]$.

Proof It is sufficient by **3.39** to show that $F(t)$ is continuous on (C, d_U). This follows, since for $G(t) = \int_0^t g \mathrm{d}s$, $g \in C$, and $0 \leq t \leq 1$,

$$|F(t) - G(t)| \leq \int_0^t |f - g| \mathrm{d}s \leq \sup_s |f(s) - g(s)|. \quad \blacksquare \tag{32.1}$$

This shows that $F(t)$ is a random variable for any t. Now, writing

$F : C \mapsto C$

as the mapping whose domain is the set of functions $f: [0,1] \mapsto \mathbb{R}$ and range the set of functions $\int_0^t f \mathrm{d}s : [0,1] \mapsto \mathbb{R}$, it can further be shown that F is a new random function whose distribution is uniquely found by extension from the

Stochastic Limit Theory: An Introduction for Econometricians. Second Edition. James Davidson, Oxford University Press.
© James Davidson 2021. DOI: 10.1093/oso/9780192844507.003.0032

finite-dimensional distributions, just as for f. The same reasoning extends to $F^2(t) = \int_0^t F\mathrm{d}s$, to F^3, and so on. Other important examples of measurable functionals under d_U include the extrema $\sup_t\{f(t)\}$ and $\inf_t\{f(t)\}$, whose distribution for the case of Brownian motion is derived in **28.7**.

Limit results for the integrals (i.e. sample averages) of partial-sum processes, or continuous functions thereof, are obtained by the straightforward method of teaming a functional central limit theorem with the continuous mapping theorem.

32.2 Theorem Let $S_{n0} = 0$, $S_{nj} = \sum_{i=1}^{j} U_{ni}$ for $j = 1, \dots, n-1$, and $X_n(t) = S_{n,[nt]}$. If $X_n \to_{\mathrm{d}} B$ then for any continuous function $g : \mathbb{R} \mapsto \mathbb{R}$,

$$\frac{1}{n}\sum_{j=0}^{n-1} g(S_{nj}) \xrightarrow{\mathrm{d}} \int_0^1 g(B)\mathrm{d}t. \tag{32.2}$$

Proof Formally,

$$g(S_{nj}) = ng(S_{nj})\int_{j/n}^{(j+1)/n} \mathrm{d}t = n\int_{j/n}^{(j+1)/n} g(X_n(t))\mathrm{d}t. \tag{32.3}$$

Hence,

$$\frac{1}{n}\sum_{j=0}^{n-1} g(S_{nj}) = \sum_{j=0}^{n-1}\int_{j/n}^{(j+1)/n} g(X_n(t))\mathrm{d}t = \int_0^1 g(X_n(t))\mathrm{d}t. \tag{32.4}$$

Since $\int_0^1 g(x(t))\mathrm{d}t$ for $x \in C$ is a continuous mapping from C to $\mathbb{R}$, the result follows by the continuous mapping theorem **29.4**. ∎

Note how $g(S_{nn})$ is omitted from these sums in accordance with the convention that elements of D are right-continuous. Since the limit process is continuous almost surely, its inclusion would change nothing material.

The leading cases of $g(\cdot)$ include the identity function and the square. For the former case, observe that $\sum_{j=1}^{n-1} S_{nj} = \sum_{i=1}^{n-1}(n-i)U_{ni}$. Theorem **32.2** should be compared with **31.6** for the case $\theta = 3$. Assuming $\{U_i\}$ is i.i.d. with mean 0 and variance σ^2, let $U_{ni} = U_i/(n^{1/2}\sigma)$ and reverse the order of summation. Then **31.6** in combination with **31.8** shows that $n^{-1}\sum_{j=1}^{n-1} S_{nj} = n^{-3/2}\sum_{j=1}^{n-1} jU_{n-j}/\sigma \to_{\mathrm{d}} B_\eta(1)/\sqrt{3}$ for the case $\eta(t) = t^3$. Since $B(1)$ and $B_\eta(1)$ have the same distribution, what this amounts to is a demonstration that $\int_0^1 B\mathrm{d}t \sim_{\mathrm{d}} \mathrm{N}(0, \frac{1}{3})$.

However, there is no such simple equivalence for the functional $\int_0^1 B^2\mathrm{d}t$, the limit for the case $g(\cdot) = (\cdot)^2$. These limit results do not generally yield closed formulae for the c.d.f. so there are no exact tabulations of the percentage points such as are available for the Gaussian case. Their main practical value is in showing that the limits *exist*. Applications in statistical inference usually involve estimating the percentiles of the distributions by Monte Carlo simulation; in other words, tabulating random variates generated as the averages of large but finite samples of g evaluated at a Gaussian drawing, to approximate integrals of $g(B)$. Knowledge of the weak convergence provides assurance that such approximations can be made as close as desired by taking n large enough.

Given a basic repertoire of limit results, it is not difficult to find the distributions of other limit processes and random variables in the same manner. To take again the case of an i.i.d. sequence, if $S_{[nt]} = \sum_{i=1}^{[nt]} U_i$ and $n^{-1/2}S_{[nt]}/\sigma \to_{\mathrm{d}} B(t)$, the continuous mapping theorem implies that the partial sums of the sample mean deviations converge to the Brownian bridge (**28.18**). That is, if $\bar{U}_n = n^{-1}\sum_{i=1}^{n} U_i$ then

$$\frac{1}{n^{1/2}\sigma}\sum_{i=1}^{[nt]}(U_i - \bar{U}_n) \xrightarrow{\mathrm{d}} B(t) - tB(1) = B^{\circ}(t). \tag{32.5}$$

On the other hand, if the partial sum process *itself* is expressed in mean deviations as $S_j - \bar{S}_n$ where $\bar{S}_n = n^{-1}\sum_{j=0}^{n-1} S_j$, convergence takes the form

$$\frac{1}{n^{1/2}\sigma}(S_{[nt]} - \bar{S}_n) \xrightarrow{\mathrm{d}} B(t) - \int_0^1 B\mathrm{d}s. \tag{32.6}$$

The limit process on the right-hand side of (32.6) is the *de-meaned Brownian motion*. Take care to distinguish the last two cases. The integral of the latter process over $[0, 1]$ is identically zero. The mean square of the mean deviations converges similarly, according to

$$\frac{1}{n^2\sigma^2}\sum_{j=0}^{n-1}(S_j - \bar{S}_n)^2 \xrightarrow{\mathrm{d}} \int_0^1 B^2\mathrm{d}s - \left(\int_0^1 B\mathrm{d}s\right)^2. \tag{32.7}$$

There is also an easy generalization of these results to vector processes. The following is the vector counterpart of the leading cases of **32.2**, the details of whose proof the reader can readily supply.

32.3 Corollary Let $\{\boldsymbol{U}_{ni}\}$ satisfy the conditions of Theorem **31.15**. If $\boldsymbol{S}_{nj} = \sum_{i=1}^{j} \boldsymbol{U}_{ni}$, then

$$\frac{1}{n}\sum_{j=1}^{n-1}\boldsymbol{S}_{nj} \xrightarrow{\mathrm{d}} \int_0^1 \boldsymbol{B}\mathrm{d}t \tag{32.8}$$

$$\frac{1}{n}\sum_{j=1}^{n-1} \boldsymbol{S}_{nj}\boldsymbol{S}_{nj}' \xrightarrow{\mathrm{d}} \int_0^1 \boldsymbol{B}\boldsymbol{B}'\mathrm{d}t. \quad \square \tag{32.9}$$

For the m-dimensional standard Brownian motion it is easily deduced that $\int_0^1 \boldsymbol{B}\mathrm{d}t \sim_{\mathrm{d}} \mathrm{N}(\mathbf{0}, \frac{1}{3}\boldsymbol{I}_m)$. Under the conditions of Theorem **31.16**, these limits apply with $\boldsymbol{B}$ replaced by $\boldsymbol{B}_\eta$.

The same approach of applying the continuous mapping theorem yields an important result involving the product of the partial-sum process with its increment. The limits obtained do not appear at first sight to involve stochastic integrals, although there will turn out to be an intimate connection.

32.4 Theorem Let the assumptions of **32.2** hold with $\mathrm{E}(S_{nn}^2) = 1$ and $\mathrm{E}(U_{ni}^2) = \sigma_{ni}^2$. Then

$$\sum_{j=1}^{n-1} S_{nj}U_{n,j+1} \xrightarrow{\mathrm{d}} \tfrac{1}{2}\left(\chi^2(1) - \bar{\sigma}^2\right) \tag{32.10}$$

where $\bar{\sigma}^2 = \lim_{n\to\infty}\sum_{i=1}^n \sigma_{ni}^2$.

Proof Squaring both sides of $S_{n,j+1} = S_{nj} + U_{n,j+1}$ gives the identity

$$S_{n,j+1}^2 = S_{nj}^2 + 2S_{nj}U_{n,j+1} + U_{n,j+1}^2. \tag{32.11}$$

Setting $S_{n0} = 0$ and summing from 0 to $n-1$ yields

$$S_{nn}^2 = \sum_{j=0}^{n-1}(S_{n,j+1}^2 - S_{nj}^2) = 2\sum_{j=1}^{n-1} S_{nj}U_{n,j+1} + \sum_{j=1}^{n} U_{nj}^2 \tag{32.12}$$

or

$$\sum_{j=1}^{n-1} S_{nj}U_{n,j+1} = \frac{1}{2}\left(S_{nn}^2 - \sum_{j=1}^{n} U_{nj}^2\right). \tag{32.13}$$

Under the assumptions, $S_{nn} \to_{\mathrm{d}} B(1) \sim_{\mathrm{d}} \mathrm{N}(0,1)$ and $\sum_{i=1}^n U_{ni}^2 \to_{\mathrm{pr}} \bar{\sigma}^2$. The result follows on applying the continuous mapping theorem and **23.14**(i). ∎

This is an unexpectedly general result that does not depend on the FCLT at all for its validity. It is true so long as $\{U_{ni}\}$ satisfies the conditions for a CLT. Since $\bar{\sigma}^2 = 1 - 2\lambda$ where

$$\lambda = \lim_{n\to\infty}\sum_{i=2}^{n}\sum_{m=1}^{i-1} \mathrm{E}(U_{n,i-m}U_{ni})$$

the sum in (32.10) has a mean of zero in the limit if and only if the sequence $\{U_{ni}\}$ is serially uncorrelated.

There is a generalization to the vector case, although only in a restricted sense. Let $\boldsymbol{S}_{nj} = \sum_{i=1}^{j} \boldsymbol{U}_{ni}$, and then generalizing (32.11) gives the identity

$$\boldsymbol{S}_{n,j+1}\boldsymbol{S}_{n,j+1}' = \boldsymbol{S}_{nj}\boldsymbol{S}_{nj}' + \boldsymbol{S}_{nj}\boldsymbol{U}_{n,j+1}' + \boldsymbol{U}_{n,j+1}\boldsymbol{S}_{nj}' + \boldsymbol{U}_{n,j+1}\boldsymbol{U}_{n,j+1}'. \tag{32.14}$$

Summing and taking limits in the same manner as before putting $\boldsymbol{\Sigma} = \lim_{n\to\infty} \sum_{i=1}^{n} \boldsymbol{U}_{nj}\boldsymbol{U}_{nj}'$ leads to the following result.

32.5 Theorem Let $\{\boldsymbol{U}_{ni}\}$ satisfy the conditions of **31.15**. Then

$$\sum_{j=1}^{n-1} \boldsymbol{S}_{nj}\boldsymbol{U}_{n,j+1}' + \sum_{j=1}^{n-1} \boldsymbol{U}_{n,j+1}\boldsymbol{S}_{nj}' \xrightarrow{\mathrm{d}} \boldsymbol{B}(1)\boldsymbol{B}(1)' - \boldsymbol{\Sigma}. \quad \square \tag{32.15}$$

Details of the proof are left to the reader. The peculiarity of this result is that it does *not* lead to a limiting distribution for the matrix $\sum_{j=1}^{n-1} \boldsymbol{S}_{nj}\boldsymbol{U}_{n,j+1}'$. This must be obtained by an entirely different approach, which is explored in §32.3.

32.2 Stochastic Integrals

This section introduces a class of stochastic integrals on $[0,1]$. To understand how these objects are constructed requires some additional theory for continuous stochastic processes and the starting point for this development is §28.1, a review of which is recommended at this point. Adaptation of a process $\{X(t)\}$ to a filtration $\boldsymbol{F}$=$\{\mathcal{F}(t)\}$ in the probability space $(\Omega, \mathcal{F}, \boldsymbol{F}, P)$ has been defined, but for stochastic integrals of the X process a stronger notion of measurability is needed. $\{X(t)\}$ is said to be *progressively measurable* with respect to $\boldsymbol{F}$ if the mappings

$$X(\cdot,\cdot) : \Omega \times [0,t] \mapsto \mathbb{R}$$

are $\mathcal{F}(t) \otimes \mathcal{B}_{[0,t]}/\mathcal{B}$-measurable, for each $t \in [0,1]$. Every progressively measurable process is adapted; just consider the rectangles $E \times [0,t]$ for any $E \in \mathcal{F}(t)$. The converse is not always true since measurability problems can arise with arbitrary functions. However,

32.6 Theorem An adapted càdlàg process is progressively measurable.

Proof For an adapted process $X \in D$ and any $t \in (0,1]$, define for $\omega \in \Omega$ the simple process on $[0,t]$:

$$X_{(n)}(\omega, s) = X(\omega, 2^{-n}k),\ s \in [2^{-n}(k-1), 2^{-n}k),\ k = 1, \dots, [2^n t] \tag{32.16}$$

with $X_{(n)}(\omega,t) = X(\omega,t)$. $X_{(n)}$ need not be adapted but it is a right-continuous function with respect to the second argument on $\Omega \times [0,t]$. If

$$E_k^x = \{\omega : X(\omega, 2^{-n}k) \le x\} \in \mathcal{F}(t)$$

then

$$\begin{aligned} A_x &= \left\{(\omega,s) : X_{(n)}(\omega,s) \le x\right\} \\ &= \left(\bigcup_{k=1}^{[2^n t]} E_k^x \times \left[2^{-n}(k-1), 2^{-n}k\right)\right) \cup \left(E_{[2^n t]+1}^x \times \{t\}\right) \end{aligned} \tag{32.17}$$

is a finite union of measurable rectangles and so $A_x \in \mathcal{F}(t) \otimes \mathcal{B}_{[0,t]}$. This is true for each $x \in \mathbb{R}$ and hence $X_{(n)}$ is $\mathcal{F}(t) \otimes \mathcal{B}_{[0,t]}/\mathcal{B}$-measurable. Fix ω and s and note that for each n

$$X_{(n)}(\omega,s) = X(\omega,u) \tag{32.18}$$

where $u > s$ and $u \downarrow s$ as $n \to \infty$. Since $X(\omega,u) \to X(\omega,s)$ by right-continuity, it follows that $X_{(n)}(\omega,s) \to X(\omega,s)$ everywhere on $\Omega \times [0,t]$ and hence X is $\mathcal{F}(t) \otimes \mathcal{B}_{[0,t]}/\mathcal{B}$-measurable (apply **3.33**). This holds for any t and the theorem follows. ∎

It is a simple exercise to show that if X is progressively measurable, so is the stopped process X^T where $X^T(t) = X(t \wedge T)$ and T is a stopping time of the filtration.

Let $\{M(t), \mathcal{F}(t)\}$ denote a martingale in continuous time having a deterministic quadratic variation process, $\langle M \rangle$. For a function $f \in D$ satisfying a prescribed set of properties to be detailed below, a stochastic process on $[0,1]$ will be represented by

$$I(\omega,t) = \int_0^t f(\omega,\tau)\mathrm{d}M(\omega,\tau),\ t \in [0,1] \tag{32.19}$$

more compactly written as $I(t) = \int_0^t f\mathrm{d}M$. For fixed ω the notation appears to correspond to the Riemann–Stieltjes integral of $f(\omega,\cdot)$ over $[0,t]$ with respect to $M(\omega,\cdot)$, but this appearance is deceptive. In fact, for almost every ω this Riemann–Stieltjes integral *does not exist*, since $M(\omega,\cdot)$ is not required to be of bounded variation. The example of Brownian motion shows that this requirement could fail for almost all ω. Hence, a different interpretation of the process $I(t)$ is called for. The results to be obtained hold for a larger class of integrator functions, including martingales whose quadratic variation is a stochastic process, but it is substantially easier to prove the existence of the integral for the case of deterministic $\langle M\rangle$ and this covers the applications of interest.

On a filtered probability space $(\Omega, \mathcal{F}, \mathrm{F}, P)$ let

$$\alpha : [0,1] \mapsto \mathbb{R}$$

be a positive, increasing element of D and let $\alpha(0) = 0$ and $\alpha(1) = 1$, with no loss of generality as it turns out. For any $t \in (0,1]$, the restriction of α to $[0,t]$ induces a finite Lebesgue–Stieltjes measure. That is to say, α is a c.d.f. and the function $\int_B \mathrm{d}\alpha(s)$ assigns a measure to each $B \in \mathcal{B}_{[0,t]}$. Accordingly the product measure μ_α can be defined on the product space $\left(\Omega \times [0,t], \mathcal{F}(t) \otimes \mathcal{B}_{[0,t]}\right)$ where for each $A \in \mathcal{F}(t) \otimes \mathcal{B}_{[0,t]}$,

$$\mu_\alpha(A) = \int_\Omega \int_0^t 1_A(\omega, s)\mathrm{d}\alpha(s)\mathrm{d}P(\omega) = \mathrm{E}\left(\int_0^t 1_A(\omega, s)\mathrm{d}\alpha(s)\right). \tag{32.20}$$

Let $\mathbb{L}_\alpha$ denote the class of functions

$$f : \Omega \mapsto R_{[0,1]}$$

which are progressively measurable (and hence adapted to $\{\mathcal{F}(t)\}$) and also square-integrable with respect to μ_α; that is to say, $||f||_\alpha < \infty$ where

$$||f||_\alpha = \mathrm{E}\left(\int_0^1 f^2 \mathrm{d}\alpha\right)^{1/2}. \tag{32.21}$$

It is easy to verify that $||f - g||_\alpha$ is a pseudo-metric on $\mathbb{L}_\alpha$. While $||f - g||_a = 0$ does not guarantee that $f(\omega) = g(\omega)$ for every $\omega \in \Omega$, it does imply that the integrals of f and g with respect to α are equal almost surely $[P]$. In this case functions f and g are said to be *equivalent.*

The chief technical result needed is that a class of simple functions is dense in $\mathbb{L}_\alpha$. Let $\mathbb{E}_\alpha \subseteq \mathbb{L}_\alpha$ denote the class such that $f(t) = f(t_k)$ for $t \in [t_k, t_{k+1}), k = 0, \ldots, m-1$ and $f(1) = f(1-)$, where $\{t_1, \ldots, t_m\} = \Pi_m$ is a partition of $[0,1]$ for some $m \in \mathbb{N}$. The following is adapted from Kopp ([116]).

32.7 Lemma For each $f \in \mathbb{L}_a$, there exists a sequence $\{f_{(n)} \in \mathbb{E}_a, n \in \mathbb{N}\}$ with $||f_{(n)} - f||_\alpha \to 0$ as $n \to \infty$.

Proof Let the domain of f be extended to $\mathbb{R}$ by setting $f(t) = 0$ for $t \notin [0,1]$. By square-integrability, $\int_{-\infty}^{+\infty} f(\omega, t)^2 d\alpha(t) < \infty$ a.s.$[P]$ and

$$\int_{-\infty}^{+\infty} \big(f(\omega, t+h) - f(\omega, t)\big)^2 \mathrm{d}\alpha(t) \to 0 \text{ a.s.}[P], \text{ as } h \to 0.$$

Hence

$$\lim_{h\to 0} \mathrm{E}\Big(\int_{-\infty}^{+\infty} \big(f(s+h) - f(s)\big)^2 \mathrm{d}\alpha(s)\Big) = 0 \tag{32.22}$$

by the bounded convergence theorem. This holds for any sequence of points going to 0, so given a partition $\Pi_{m(n)}$ such that $||\Pi_{m(n)}|| \to 0$ as $n \to \infty$ and $t \in [0,1]$, consider the case $h = k_n(t) - t$ where

$$k_n(t) = t_i,\ t \in [t_i, t_{i+1}),\ i = 0, \dots, m(n) - 1 \tag{32.23}$$

$$k_n(1) = t_{m(n)-1}. \tag{32.24}$$

Clearly, $k_n(t) \to t$. Hence, (32.22) implies that

$$\begin{aligned} &\int_{-\infty}^{+\infty} \mathrm{E}\Big(\int_{-1}^{1} \big(f(s+k_n(t)) - f(s+t)\big)^2 \mathrm{d}\alpha(t)\Big) \mathrm{d}\alpha(s) \\ &\quad = \int_{-1}^{1} \mathrm{E}\Big(\int_{-\infty}^{+\infty} \big(f(s+k_n(t)) - f(s+t)\big)^2 \mathrm{d}\alpha(s)\Big) \mathrm{d}\alpha(t) \\ &\quad = \int_{-1}^{1} \mathrm{E}\Big(\int_{-\infty}^{+\infty} \big(f(s+k_n(t)-t) - f(s)\big)^2 \mathrm{d}\alpha(s)\Big) \mathrm{d}\alpha(t) \\ &\quad \to 0 \text{ as } n \to \infty \end{aligned} \tag{32.25}$$

where the first equality applies Fubini's theorem (**4.27**). Since the inner integral on the left-hand side is non-negative, (32.25) implies

$$\lim_{n\to\infty} \mathrm{E}\Big(\int_{-1}^{1} \big(f(s+k_n(t)) - f(s+t)\big)^2 \mathrm{d}\alpha(t)\Big) = 0 \tag{32.26}$$

for almost all $s \in \mathbb{R}$. Fixing $s \in [0,1]$ and making a change of variable from t to $t - s$ gives

$$\lim_{n\to\infty} \mathrm{E}\Big(\int_{s-1}^{1+s} \big(f(t+l_n(t)) - f(t)\big)^2 \mathrm{d}\alpha(t)\Big) = 0 \tag{32.27}$$

where $l_n(t) = k_n(t-s) - (t-s)$, noting that $k_n(t-s) = t_i$ when $t \in [t_i + s, t_{i+1} + s)$. Define a function

$$f_{(n)}(t) = \begin{cases} f(t + l_n(t)), & t + l_n(t) \in [0,1] \\ 0, & \text{otherwise} \end{cases} \tag{32.28}$$

and note that $f_{(n)}(t) = f(t_i + s)$ for $t \in [t_i + s, t_{i+1} + s) \cap [0,1]$ and hence $f_{(n)} \in \mathbb{E}_\alpha$. In view of (32.27), the proof is completed by noting that $[0,1] \subseteq [s-1, 1+s]$ and hence

$$\begin{aligned} \mathrm{E}\Big(\int_0^1 \big(f_{(n)}(t) - f(t)\big)^2 \mathrm{d}\alpha(t)\Big) &= \mathrm{E}\Big(\int_{s-1}^{1+s} \big(f_{(n)}(t) - f(t)\big)^2 \mathrm{d}\alpha(t)\Big) \\ &\leq \mathrm{E}\Big(\int_{s-1}^{1+s} \big(f(t + l_n(t)) - f(t)\big)^2 \mathrm{d}\alpha(t)\Big). \qquad (32.29) \end{aligned}$$

The final inequality uses the fact that outside $[0,1]$, $f_{(n)}(t) = f(t) = 0$, whereas $f(t + l_n(t)) \neq 0$ is possible. ∎

Be careful to note the role of the assumptions here. Fubini's theorem can be used in (32.25) because the function α is nonstochastic and does not depend on ω, and hence μ_α in (32.20) is a product measure. Without this property, more roundabout arguments are needed.

The construction of $I(t)$ for a martingale M having a deterministic quadratic variation process $\langle M \rangle$ proceeds by applying **32.7** for the case $\alpha = \langle M \rangle$. The integral is first defined for simple functions and then a limit argument gives the extension to $\mathbb{L}_{\langle M \rangle}$. Let $f \in \mathbb{E}_{\langle M \rangle}$ be defined on a partition Π_m and then for $t \in [t_{k-1}, t_k)$, $k = 1, \dots, m$, and $\omega \in \Omega$ let

$$\begin{aligned} I(\omega, t) &= \sum_{j=0}^{k-1} f(\omega, t_j)\big(M(\omega, t_{j+1}) - M(\omega, t_j)\big) + f(\omega, t_k)\big(M(\omega, t) - M(\omega, t_k)\big) \\ &= \sum_{j=0}^{k} f(\omega, t_j)\big(M(\omega, t \wedge t_{j+1}) - M(\omega, t_j)\big) \qquad (32.30) \end{aligned}$$

where $t \wedge t_j$ denotes $\min\{t, t_j\}$. The stochastic process $\{I(\omega, t), t \in [0,1]\}$ is an element of D. For $t \in [t_k, t_{k+1})$ and $s \in (t, t_{k+1})$,

$$I(\omega, s) - I(\omega, t) = f(\omega, t_k)\big(M(\omega, s) - M(\omega, t)\big)$$

so that right-continuity is shared with the process M. It is easily verified that

$$\mathrm{E}\big(I(s)|\mathcal{F}(t)\big) = I(t) \text{ a.s.}[P], \text{ for } 0 \leq t \leq s \leq 1. \qquad (32.31)$$

Since $f \in \mathbb{E}_{\langle M \rangle}$ is a simple function with $f(\tau) = f(t_k)$ for $\tau \in [t_k, t_{k+1})$, the orthogonality of the martingale increments, the law of iterated expectations, and (28.13) yield

$$
\begin{aligned}
\mathrm{E}\big(I(t)^2\big) &= \mathrm{E}\Big(\sum_{j=0}^{k-1} f(t_j)^2 \mathrm{E}\big((M(t \wedge t_{j+1}) - M(t_j))^2 | \mathcal{F}(t_j)\big)\Big) \\
&= \mathrm{E}\Big(\sum_{j=0}^{k-1} f(t_j)^2 \big(\langle M\rangle(t \wedge t_{j+1}) - \langle M\rangle(t_j)\big)\Big) \\
&= \mathrm{E}\Big(\sum_{j=0}^{k-1} f(t_j)^2 \int_{t_j}^{t\wedge t_{j+1}} \mathrm{d}\langle M\rangle(\tau)\Big) \\
&= \mathrm{E}\Big(\int_0^t f(\tau)^2 \mathrm{d}\langle M\rangle(\tau)\Big) = \|1_{[0,t]}f\|^2_{\langle M\rangle}. \qquad (32.32)
\end{aligned}
$$

(See (32.21).) The last member is finite by assumption on the class $\mathbb{L}_{\langle M\rangle}$ and hence $I(t)$ is found to be itself a square-integrable martingale.

Equation (32.30) is an adequate definition of the integral in the sense that $\mathbb{E}_{\langle M\rangle}$ is dense in $\mathbb{L}_{\langle M\rangle}$; every $f \in \mathbb{L}_{\langle M\rangle}$ is arbitrarily close to an element of $\mathbb{E}_{\langle M\rangle}$. Given $f \in \mathbb{L}_{\langle M\rangle}$, let a sequence of functions $\{f_{(n)} \in \mathbb{E}_{\langle M\rangle}\}$ be defined with respect to partitions $\Pi_{m(n)}$, such that $\|\Pi_{m(n)}\| \to 0$ and $\|f_{(n)} - f\|_{\langle M\rangle} \to 0$. For example, setting $m(n) = 2^n$ and $t_i = i/2^n$ means that the intervals of the partition are bisected each time n is incremented and $\Pi_{2^n} \to \mathbb{D}$. For $f_{(n)}$, (32.28) with $s = 0$ will serve. The integrals of $f_{(n)}$, say $\{I_n(t),\ n \in \mathbb{N}\}$ for fixed t, form a real stochastic sequence. According to (32.32),

$$
\mathrm{E}\big(I_{n+m}(t) - I_n(t)\big)^2 = \|(f_{(n+m)} - f_{(n)})1_{[0,t]}\|^2_{\langle M\rangle} \to 0 \text{ as } n \to \infty \qquad (32.33)
$$

for any $m > 0$, and for each $t \in [0,1]$. It follows that the sequence $\{I_n(t)\}$ converges in mean square as $f_{(n)}$ approaches f. Moreover, for any n and m the process $\{I_{n+m}(t) - I_n(t),\ t \in [0,1]\}$ is a martingale, and applying the Doob inequality (**28.5**(ii)),

$$
\begin{aligned}
\mathrm{E}\Big(\sup_{t\in[0,1]} \big(I_{n+m}(t) - I_n(t)\big)^2\Big) &\leq 4\mathrm{E}\big(I_{n+m}(1) - I_n(1)\big)^2 \\
&= 4\|f_{(n+m)} - f_{(n)}\|^2_{\langle M\rangle} \to 0. \qquad (32.34)
\end{aligned}
$$

The mean square convergence is therefore uniform in t.

In view of **19.6**, L_2 convergence implies that there exists a subsequence $\{n_k, k \in \mathbb{N}\}$ on which the convergence occurs with probability 1. Since the sequence of partitions $\{\Pi_{m(n)}\}$ specified in the construction of $f_{(n)}$ is required only to converge in the sense $\|\Pi_{m(n)}\| \to 0$ as $n \to \infty$, the sequence $\{\Pi_{m(n_k)}\}$ can be used in the construction with no loss of generality and so $\lim_{m,n\to\infty} d_B(I_{n+m}, I_n) = 0$, a.s.$[P]$. Since the space (D, d_B) is complete, a limit function I exists in D almost surely.

If f and g are equivalent in the sense defined following (32.21), then $\int_0^t f\mathrm{d}M = \int_0^t g\mathrm{d}M$ a.s.[P]. Moreover, mean squared convergence implies weak convergence so that the distribution of $I(t)$ can be characterized as the weak limit of the sequence of distributions of the $I_n(t)$. Note the characteristic property of the integral, applying the limit to (32.32):

$$\mathrm{E}\big(I(t)^2\big) = \mathrm{E}\Big(\int_0^t f\mathrm{d}M\Big)^2 = \mathrm{E}\Big(\int_0^t f^2\mathrm{d}\langle M\rangle\Big). \tag{32.35}$$

For the case $M = B$, Brownian motion, $\langle M\rangle = t$. I is commonly known as the *Itô integral* and the relation (32.35) is the *Itô isometry*.

The so-called fundamental theorem of stochastic calculus, or *Itô's rule*, shows that these objects are quite different from the Riemann–Stieltjes integrals which they superficially resemble. A form of Itô's rule holds for a general class of continuous semimartingales; see for example Karatzas and Shreve ([111]), McKean ([123]), or Protter ([154]) for details. Here the result is given for the basic case of Brownian motion B, allowing a relatively straightforward proof. Let $g : \mathbb{R} \mapsto \mathbb{R}$ be a twice-continuously differentiable function with g' and g'' denoting the first and second derivatives.

32.8 Theorem For $0 < t \le 1$,

$$g(B(t)) - g(0) = \int_0^t g'(B)\mathrm{d}B + \tfrac{1}{2}\int_0^t g''(B)\mathrm{d}s \text{ a.s.}[P]. \tag{32.36}$$

Proof Let Π_n denote the partition of $[0, t]$ in which $t_j = tj/n$ for $j = 1, \ldots, n$, with $t_n = t$, so that $\|\Pi_n\| \to 0$ as $n \to \infty$. For economy of notation write B_j to denote $B(t_j)$ and $\mathcal{F}_j$ for $\mathcal{F}(t_j)$ similarly. Consider the telescoping sum

$$\begin{aligned} g(B(t)) - g(0) &= \sum_{j=0}^{n-1}\big(g(B_{j+1}) - g(B_j)\big) \\ &= \sum_{j=0}^{n-1} g'(B_j)(B_{j+1} - B_j) + \tfrac{1}{2}\sum_{j=0}^{n-1} g''(\eta_j)(B_{j+1} - B_j)^2 \end{aligned} \tag{32.37}$$

where the second equality takes the Taylor expansions of the terms to second order and $\eta_j = B_j + \theta_j(B_{j+1} - B_j)$ for $0 \le \theta_j \le 1$. Let $\mathbb{L}_m$ denote the case of $\mathbb{L}_\alpha$ where α is Lebesgue measure and let $\mathbb{E}_m \subseteq \mathbb{L}_m$ be the associated class of simple functions. Since $g'(\cdot)$ is continuous, $g'(B) \in \mathbb{L}_m$. In a construction similar to that used in the proof of **32.7**, define $p_n \in \mathbb{E}_m$ by

$$p_n(s) = g'(B_j),\ s \in [t_j, t_{j+1}),\ j = 0, \dots, n-1. \tag{32.38}$$

Then write

$$\sum_{j=0}^{n-1} g'(B_j)(B_{j+1} - B_j) = \sum_{j=0}^{n-1} g'(B_j) \int_{t_j}^{t_{j+1}} \mathrm{d}B(s) = \int_0^t p_n(s)\mathrm{d}B(s). \tag{32.39}$$

Making this substitution and using (32.35),

$$\begin{aligned} \mathrm{E}\left(\sum_{j=0}^{n-1} g'(B_j)(B_{j+1} - B_j) - \int_0^t g'(B)\mathrm{d}B\right)^2 &= \mathrm{E}\left(\int_0^t (p_n - g'(B))\mathrm{d}B\right)^2 \\ &= \|(p_n - g'(B))1_{[0,t]}\|_m \to 0. \end{aligned} \tag{32.40}$$

Now consider the second sum on the right-hand side of (32.37). Write

$$\sum_{j=0}^{n-1} g''(\eta_j)(B_{j+1} - B_j)^2 = R_{1n} + R_{2n} + R_{3n} \tag{32.41}$$

where

$$\begin{aligned} R_{1n} &= \sum_{j=0}^{n-1} g''(B_j)(t_{j+1} - t_j), \\ R_{2n} &= \sum_{j=0}^{n-1} (g''(\eta_j) - g''(B_j))(B_{j+1} - B_j)^2, \\ R_{3n} &= \sum_{j=0}^{n-1} g''(B_j)\big((B_{j+1} - B_j)^2 - (t_{j+1} - t_j)\big). \end{aligned}$$

Consider the limit of each of these terms as $\|\Pi_n\| \to 0$. First, similarly to p_n define

$$q_n(s) = g''(B_j),\ s \in [t_j, t_{j+1}),\ j = 0, \dots, n-1.$$

Since g'' is continuous, $\{q_n(\omega)\}$ for $\omega \in C \in \mathcal{F}$ with $P(C) = 1$ is a sequence of bounded functions on $[0, t]$ converging to $g''(\omega)$. Applying the bounded convergence theorem elementwise yields

$$R_{1n} = \sum_{j=0}^{n-1} g''(B_j) \int_{t_j}^{t_{j+1}} \mathrm{d}s = \int_0^t q_n(s)\mathrm{d}s \to \int_0^t g''(B)\mathrm{d}s \text{ a.s.}[P]. \tag{32.42}$$

Second, by the modulus and Cauchy–Schwarz inequalities,

$$\begin{aligned}\mathrm{E}|R_{2n}| &\le \mathrm{E}\Big(\max_{0\le j\le n-1}|g''(\eta_j)-g''(B_j)|\sum_{j=0}^{n-1}(B_{j+1}-B_j)^2\Big)\\ &\le \Big\|\max_{0\le j\le n-1}|g''(\eta_j)-g''(B_j)|\Big\|_2\Big\|\sum_{j=0}^{n-1}(B_{j+1}-B_j)^2\Big\|_2\end{aligned}$$

where

$$\Big\|\max_{0\le j\le n-1}|g''(\eta_j)-g''(B_j)|\Big\|_2 \to 0 \tag{32.43}$$

since $\max_{0\le j\le n-1}|\eta_j - B_j| \to 0$ a.s.$[P]$, while the Minkowski inequality gives

$$\Big\|\sum_{j=0}^{n-1}(B_{j+1}-B_j)^2\Big\|_2 \le \sum_{j=0}^{n-1}\|(B_{j+1}-B_j)^2\|_2 = \sqrt{3}\sum_{j=0}^{n-1}|t_{j+1}-t_j| = \sqrt{3}t.$$

Here, the equality applies the Gaussianity of the increments together with **9.14** to compute the fourth moments. Third,

$$\begin{aligned}\mathrm{E}(R_{3n}^2) &= \sum_{j=0}^{n-1}\mathrm{E}\Big(g''(B_j)\mathrm{E}\big((B_{j+1}-B_j)^2-(t_{j+1}-t_j)|\mathcal{F}_j\big)\Big)^2\\ &\le 2\max_{0\le j\le n-1}\{\mathrm{E}(g''(B_j))^2\}\sum_{j=0}^{n-1}(t_{j+1}-t_j)^2\\ &\le 2\max_{0\le j\le n-1}\{\mathrm{E}(g''(B_j))^2\}t\max_{0\le j\le n-1}\{t_{j+1}-t_j\}\to 0\end{aligned} \tag{32.44}$$

where the equality uses the fact that the cross-product terms vanish since

$$\mathrm{E}\big((B_{j+1}-B_j)^2-(t_{j+1}-t_j)|\mathcal{F}_j\big) = 0 \text{ a.s.} \tag{32.45}$$

and the first inequality uses (32.45) and the LIE.

According to **19.6**, L_2-convergence and L_1-convergence imply convergence with probability 1 on a subsequence $\{n_k\}$ and the choice of partitions Π_{n_k} is arbitrary so long as $\|\Pi_{n_k}\| \to 0$ as $k \to \infty$. Therefore, combining (32.40) with (32.42), (32.43), and (32.44) completes the proof. ∎

32.9 Example In the case $g(\cdot) = (\cdot)^2$, (32.36) has the form

$$B(t)^2 = 2\int_0^t B\mathrm{d}B + t \text{ a.s.}[P]. \tag{32.46}$$

Since $B(t) \sim_{\mathrm{d}} \mathrm{N}(0,t)$, (32.46) and the continuous mapping theorem give the result

$$\int_0^1 B\mathrm{d}B \overset{\mathrm{d}}{\sim} \tfrac{1}{2}\big(\chi^2(1)-1\big). \tag{32.47}$$

Comparing with **32.4**, it is apparent that the limit in (32.10) can be expressed as $\int_0^1 B\mathrm{d}B + \lambda$ where $\lambda = \frac{1}{2}(1-\bar{\sigma}^2)$, vanishing in the case where the process increments are uncorrelated. □

32.3 Convergence to Stochastic Integrals

Let $\{U_{nj}\}$ and $\{W_{nj}\}$ be a pair of stochastic arrays, and let $X_n(t) = \sum_{i=1}^{[nt]} U_{ni}$ and $Y_n(t) = \sum_{j=1}^{[nt]} W_{nj}$. Then (X_n, Y_n) is an element of the separable, complete metric space (D^2, d_B^2) and under the assumptions of Theorem **31.15**,

$$(X_n, Y_n) \xrightarrow{\mathrm{d}} (X, Y) \tag{32.48}$$

where X and Y are a pair of Brownian motions. The problem to be considered in this section is the weak convergence of the partial sums

$$\begin{aligned} G_n &= \sum_{j=1}^{n-1}\left(\sum_{i=1}^{j} U_{ni}\right) W_{n,j+1} \\ &= \sum_{j=1}^{n-1} X_n(j/n)\big(Y_n((j+1)/n) - Y_n(j/n)\big). \end{aligned} \tag{32.49}$$

This problem differs from those of §31.1 because it cannot be deduced merely by combining the functional CLT with the continuous mapping theorem. The aim is to show that

$$G_n \xrightarrow{\mathrm{d}} \int_0^1 X\mathrm{d}Y + \Lambda_{XY} \tag{32.50}$$

where

$$\Lambda_{XY} = \lim_{n\to\infty} \sum_{j=1}^{n-1}\sum_{i=1}^{j} \mathrm{E}(U_{ni}W_{n,j+1}). \tag{32.51}$$

An admissible case is $W_{nj} = U_{nj}$, in which case the relevant joint distribution is singular. The results of this section stem mainly from those of Chan and Wei

([30]) and their extension due to Phillips ([144]); see *inter alia* Strasser ([178]), Phillips ([143]), Kurtz and Protter ([118]), and Hansen ([96]) for some alternative approaches.

The main difficulty with proving results such as (32.50) is to deal with the possible cross-autocorrelation represented by Λ_{XY}. This is not unlike the difficulties with the estimation of long-run variances encountered in §26.1. A version of the convergence in (32.50) can be proved under the assumptions sufficient for **31.15** if a partially aggregated statistic is substituted for G_n. This can be constructed as follows. Choose an increasing integer subsequence $\{k_n, n \in \mathbb{N}\}$ so that $k_n/n \to 0$, but $k_n \uparrow \infty$. Define $n_j = [nj/k_n]$ for $j = 0, \dots, k_n$ so that $\min_{1\le j\le k_n}\{n_j - n_{j-1}\} \to \infty$. Then define partitions $\Pi_n = (t_1, \dots, t_{k_n}\}$ of $[0,1]$ where $t_j = n_j/n$, such that $||\Pi_n|| \to 0$, and finally let

$$G_n^* = \sum_{j=1}^{k_n} X_n(t_{j-1})\big(Y_n(t_j) - Y_n(t_{j-1})\big). \tag{32.52}$$

In particular, note how the increments $Y_n(t_j) - Y_n(t_{j-1}) = \sum_{j=n_{j-1}+1}^{n_j} W_{nj}$ grow with sample size and are converging to Brownian segments. This makes it possible to prove the following.

32.10 Theorem Under the assumptions of Theorem **31.15**,

$$G_n^* \xrightarrow{\text{d}} \int_0^1 X\text{d}Y.$$

Proof The main ingredient of the proof is the Skorokhod representation theorem **29.6**, by which the a.s. convergence of a random sequence can be deduced from its weak convergence. The convergence in (32.48) implies the existence of a sequence $\{(X^n, Y^n) \in D^2,\ n \in \mathbb{N}\}$ such that (X^n, Y^n) is distributed like (X_n, Y_n) and $d_B^2\big((X^n, Y^n), (X, Y)\big) \to_{\text{a.s.}} 0$. Let G^{*n} represent the same expression as (32.52) except that the Skorokhod variables X^n and Y^n are substituted for X_n and Y_n. In view of **23.18** and the fact that G^{*n} and G_n^* have the same distribution, to prove the theorem it suffices to prove that

$$\left|G^{*n} - \int_0^1 X\text{d}Y\right| \xrightarrow{\text{pr}} 0. \tag{32.53}$$

Note first that $|P_n - \int_0^1 X\text{d}Y| \to_{L_2} 0$ where

$$P_n = \sum_{j=1}^{k_n} X(t_{j-1})\big(Y(t_j) - Y(t_{j-1})\big) = \sum_{j=1}^{k_n}\int_{t_{j-1}}^{t_j} X(t_{j-1})\text{d}Y(t).$$

This convergence follows because

$$\begin{aligned}\mathrm{E}\Big(P_n-\int_0^1 X\mathrm{d}Y\Big)^2&=\mathrm{E}\Big(\sum_{j=1}^{k_n}\int_{t_{j-1}}^{t_j}\big(X(t_{j-1})-X(t)\big)\mathrm{d}Y(t)\Big)^2\\&=\sum_{j=1}^{k_n}\int_{t_{j-1}}^{t_j}\big(t-t_{j-1}\big)\mathrm{d}t\\&\ll\sum_{j=1}^{k_n}(t_j-t_{j-1})^2\\&=O\Big(\max_{1\le j\le k_n}\{t_j-t_{j-1}\}\Big)\to 0\end{aligned}\qquad(32.54)$$

applying the Itô isometry (32.35) and Fubini's theorem to get the second equality. The proof is therefore completed by showing $|G^{*n}-P_n|\to_{L_2}0$.

The trick is to apply 'summation by parts'. The following identity is closely related to Abel's summation formula (**2.25**) and can be verified by inspection.

32.11 Lemma For arbitrary real numbers $a_j,b_j,\alpha_j,\beta_j,j=1,\dots,k$,

$$\begin{aligned}&\sum_{j=1}^{k}a_{j-1}(b_j-b_{j-1})-\sum_{j=1}^{k}\alpha_{j-1}(\beta_j-\beta_{j-1})\\&\quad=\sum_{j=1}^{k}(a_{j-1}-\alpha_{j-1})(b_j-b_{j-1})-\sum_{j=1}^{k}(\alpha_j-\alpha_{j-1})(b_j-\beta_j)\\&\qquad\qquad+\alpha_k(b_k-\beta_k)-\alpha_0\,(b_0-\beta_0).\quad\Box\end{aligned}\qquad(32.55)$$

To apply **32.11** put $k=k_n$ and then set $a_j=X^n(\omega,t_j)$, $b_j=Y^n(\omega,t_j)$, $\alpha_j=X(\omega,t_j)$, and $\beta_j=Y(\omega,t_j)$. Then the left-hand side of (32.55) corresponds to $G^{*n}-P_n$. The proof is completed by showing that the right-hand side vanishes in mean square, noting $a_0=b_0=\alpha_0=\beta_0=0$.

According to Egorov's theorem (**19.4**) and the equivalence of d_S and d_B in D, (32.48) implies that for a set $C_\varepsilon\in\mathcal{F}$ with $P(C_\varepsilon)\ge 1-\varepsilon$,

$$\sup_{\omega\in C_\varepsilon}d_S^2\big((X^n(\omega),Y^n(\omega)),(X(\omega),Y(\omega))\big)\to 0\qquad(32.56)$$

for each $\varepsilon>0$. Since X is a.s. continuous, there exists a set E_X with $P(E_X)=1$ and the following property: if $\omega\in E_X$ then for any $\eta>0$ there is a constant $\delta>0$ such that, if $d_S\big(X^n(\omega),X(\omega)\big)\le\delta$,

$$\begin{aligned}\sup_t |X^n(\omega,t) - X(\omega,t)| &\le \sup_t |X^n(\omega,t) - X(\omega,\lambda(t))| \\ &\qquad + \sup_t |X(w,t) - X(\omega,\lambda(t))| \\ &\le \delta + \eta \end{aligned} \tag{32.57}$$

where $\lambda(\cdot)$ is the function from (30.15). The same result holds for Y in respect of a set E_Y with $P(E_Y) = 1$. It follows from (32.56) that, for $\omega \in C_\varepsilon^* = C_\varepsilon \cap E_X \cap E_Y$,

$$d_U^2\big((X^n(\omega), Y^n(\omega)), (X(\omega), Y(\omega))\big) = \delta_n \to 0 \tag{32.58}$$

where the equality defines δ_n. Note too that $P(C_\varepsilon^*) = P(C_\varepsilon)$. Choose the sequence $\{k_n\}$ to increase slowly enough that $k_n\delta_n^2 \to 0$. The Cauchy–Schwarz inequality and (32.58) give, for each $\omega \in C_\varepsilon^*$,

$$\begin{aligned}&\left(\sum_{j=1}^{k_n}\big(X^n(\omega,t_{j-1}) - X(\omega,t_{j-1})\big)\big(Y^n(\omega,t_j) - Y^n(\omega,t_{j-1})\big)\right)^2 \\ &\quad\le \sum_{j=1}^{k_n}\big(X^n(\omega,t_{j-1}) - X(\omega,t_{j-1})\big)^2 \sum_{j=1}^{k_n}\big(Y^n(\omega,t_j) - Y^n(\omega,t_{j-1})\big)^2 \\ &\quad\le k_n\delta_n^2 \sum_{j=1}^{k_n}\big(Y^n(\omega,t_j) - Y^n(\omega,t_{j-1})\big)^2. \end{aligned} \tag{32.59}$$

Equivalence of the distributions and assumption **31.5**(e) in respect of W_{ni} (holding in view of **31.15**(a)) imply that

$$\begin{aligned}\mathrm{E}\big(Y^n(t_j) - Y^n(t_{j-1})\big)^2 &= \mathrm{E}\left(\sum_{i=n_{j-1}+1}^{n_j} W_{ni}\right)^2 \\ &\to t_j - t_{j-1}. \end{aligned} \tag{32.60}$$

Hence from (32.59),

$$\begin{aligned}&\mathrm{E}\left(\sum_{j=1}^{k_n}\big(X^n(t_{j-1}) - X(t_{j-1})\big)\big(Y^n(t_j) - Y^n(t_{j-1})\big)1_{C_\varepsilon^*}\right)^2 \\ &\qquad\le k_n\delta_n^2 \sum_{j=1}^{k_n}\mathrm{E}\big(Y^n(t_j) - Y^n(t_{j-1})\big)^2 \to 0 \end{aligned} \tag{32.61}$$

noting from (32.60) that the sum in the second member is bounded by $||\Pi_n||$ when n is large enough. This accounts for the first right-hand-side sum of (32.55). Closely similar arguments give

$$\mathrm{E}\Bigl(\sum_{j=1}^{k_n}\bigl(Y^n(t_j)-Y(t_j)\bigr)\bigl(X(t_j)-X(t_{j-1})\bigr)1_{C_\varepsilon^*}\Bigr)^2 \to 0 \qquad (32.62)$$

to account for the second right-hand-side sum and for the final term,

$$\mathrm{E}\bigl(X(1)\bigl(Y^n(1)-Y(1)\bigr)1_{C_\varepsilon^*}\bigr)^2 \le \delta_n^2 \to 0. \qquad (32.63)$$

This completes the proof of (32.53) and hence of the theorem. ∎

To show that (32.50) holds, the remaining problem is to show that $G_n - G_n^* \to_{\text{pr}} \Lambda_{XY}$. Note that

$$Y_n(t_j)-Y_n(t_{j-1}) = \sum_{i=n_{j-1}}^{n_j-1} W_{n,i+1}$$

and formally setting $U_{n0}=0$, define X_{ni} by

$$X_n(i/n)-X_n(t_{j-1}) = \sum_{l=n_{j-1}}^{i} U_{nl} = X_{ni}.$$

Then from the definitions of G_n in (32.49) and G_n^* in (32.52),

$$\begin{aligned} G_n - G_n^* &= \sum_{j=1}^{k_n}\Biggl(\Biggl(\sum_{i=n_{j-1}}^{n_j-1} X_n(i/n)\bigl(Y_n((i+1)/n)-Y_n(i/n)\bigr)\Biggr) \\ &\qquad\qquad - X_n(t_{j-1})\bigl(Y_n(t_j)-Y_n(t_{j-1})\bigr)\Biggr) \\ &= \sum_{j=1}^{k_n}\Biggl(\sum_{i=n_{j-1}}^{n_j-1} X_{ni}W_{n,i+1}\Biggr). \end{aligned} \qquad (32.64)$$

The convergence in probability to Λ_{XY} of the sum in (32.64) under the assumptions of **31.15** is proved in §32.4, but the argument is lengthy and complex. In the context of this section, attention is confined to cases where plausible restrictions on the distributions allow simpler arguments to prevail.

Consider first an assumption on the integrator process under which $\Lambda_{XY} = 0$. Define arrays $\{c_{ni}^U\}$ and $\{c_{ni}^W\}$ to correspond to those specified by **31.15** for $\boldsymbol{\lambda} = (1,0)'$ and $\boldsymbol{\lambda} = (0,1)'$ respectively. These satisfy condition **31.5**(c) by assumption, which implies that

$$\max\Big\{\sup_l\{c_{nl}^U\}, \sup_i\{c_{ni}^W\}\Big\} = O(n^{-1/2}). \tag{32.65}$$

32.12 Theorem Under the assumptions of Theorem **31.15**, let $\{W_{nj}, \mathcal{F}_{nj}\}$ be a martingale difference array where $\mathcal{F}_{nj} = \sigma(\boldsymbol{V}_{nk}, k \le j)$ and

$$\mathrm{E}(W_{n,j+1}^2|\mathcal{F}_{nj}) \ll (c_{n,j+1}^W)^2 < \infty \text{ a.s.}[P]. \tag{32.66}$$

Then, $G_n - G_n^* \to_{L_2} 0$. □

Condition (32.66) is a rather mild condition imposing a limited form of conditional homoscedasticity on the W_{ni} process, at worst ruling out some relatively exotic cases. The assumptions are certainly satisfied by an i.i.d. process, to take the simplest example.

Proof of 32.12 By assumption, $\mathrm{E}(X_{ni}W_{n,i+1}X_{nk}W_{n,k+1}) = 0$ unless $i = k$, and

$$\mathrm{E}(X_{ni}^2 W_{n,i+1}^2) = \mathrm{E}\big(X_{ni}^2\mathrm{E}(W_{n,i+1}^2|\mathcal{F}_{ni})\big) \ll \mathrm{E}(X_{ni}^2)(c_{n,i+1}^W)^2. \tag{32.67}$$

Since U_{nl} is an L_2-mixingale of size $-\frac{1}{2}$ by **18.7**, it follows by **17.21** that

$$\mathrm{E}(X_{ni}^2) \ll (i - n_{j-1})\Big(\max_{n_{j-1}\le l\le i} c_{nl}^U\Big)^2. \tag{32.68}$$

Therefore, using (32.65), (32.68), and (32.66),

$$\begin{aligned}
\mathrm{E}(G_n - G_n^*)^2 &= \sum_{j=1}^{k_n}\sum_{i=n_{j-1}}^{n_j-1} \mathrm{E}(X_{ni}^2 W_{n,i+1}^2) \\
&\ll \frac{1}{n^2}\sum_{j=1}^{k_n}\sum_{i=n_{j-1}}^{n_j-1} (i - n_{j-1}) \\
&\ll \frac{1}{n^2}\sum_{j=1}^{k_n}(n_j - n_{j-1})^2 \\
&= O\Big(\max_{1\le j\le k_n}\{t_j - t_{j-1}\}\Big) = o(1). \quad ■
\end{aligned} \tag{32.69}$$

While this argument works it is worth noting that an L_p-convergence law such as **20.11** does not work here, even though $\{X_{n,i-1}W_{ni}, \mathcal{F}_{ni}\}$ is a m.d., since the summability conditions are violated.

Another case for which the argument is comparatively straightforward is where W_{ni} is a linear process.

32.13 Theorem Under the assumptions of Theorem **31.15**, let

$$W_{ni} = \sum_{k=0}^{\infty} \theta_k V_{1n,i-k} \tag{32.70}$$

where V_{1ni} is a zero-mean, L_r-bounded, and serially independent element of $\boldsymbol{V}_{ni}$ and $\{\theta_k\}$ is a constant sequence with $\sum_{j=0}^{\infty} j|\theta_j| < \infty$. Then, $G_n - G_n^* \to_{\text{pr}} \Lambda_{XY}$.

Proof Apply the Beveridge–Nelson decomposition of (26.19) in the form

$$W_{ni} = M_{ni} + Z_{n,i-1} - Z_{ni} \tag{32.71}$$

where

$$M_{ni} = \left(\sum_{k=0}^{\infty} \theta_k\right) V_{1ni}$$

is a serially independent process and hence a m.d. and

$$Z_{ni} = \sum_{j=1}^{\infty} \theta_j^* V_{1,n,i-j+1}$$

where $\theta_j^* = \sum_{k=j+1}^{\infty} \theta_k$. (Also compare Example **17.8**). Note that

$$||Z_{ni} - E_{i-k}^i Z_{ni}||_2 \leq \max_{j \geq k} ||V_{1,n,i-j}||_2 \sum_{j=k+1}^{\infty} |\theta_j^*|$$

and hence Z_{ni} is L_2-NED on $\{\boldsymbol{V}_{ni}\}$ with $c_{ni}^W = \max_{j\geq 0} ||V_{1,n,i-j}||_2 \sum_{j=1}^{\infty} |\theta_j^*|$, noting that the sequence $\{\theta_j^*\}$ is absolutely summable by assumption and also that $\sup_{n,i} c_{ni}^W = O(n^{-1/2})$ is imposed by the choice of normalization for V_1. Hence write

$$G_n - G_n^* = A_{n1} + A_{n2}$$

where

$$A_{n1} = \sum_{j=1}^{k_n} \sum_{i=n_{j-1}}^{n_j-1} X_{ni} M_{n,i+1} \tag{32.72}$$

and

$$A_{n2} = \sum_{j=1}^{k_n} \sum_{i=n_{j-1}}^{n_j-1} X_{ni}(Z_{ni} - Z_{n,i+1}). \tag{32.73}$$

First, $A_{n1} \to_{L_2} 0$ by **32.12** whose assumptions are satisfied here. Second, setting $X_{n0} = X_{nn} = 0$ and noting $U_{ni} = X_{ni} - X_{n,i-1}$, Abel's partial summation formula **2.25** and (32.73) give

$$\begin{aligned} A_{n2} &= \sum_{j=1}^{k_n} \sum_{i=n_{j-1}+1}^{n_j} U_{ni} Z_{ni} - X_{n,n_j} Z_{n,n_j} + X_{n,n_{j-1}} Z_{n,n_{j-1}} \\ &= \sum_{i=1}^{n-1} U_{ni} Z_{ni}. \end{aligned} \tag{32.74}$$

By the assumptions and **18.9**, $U_{ni}Z_{ni}$ is L_1-NED on $\{\boldsymbol{V}_{ni}\}$ and by **18.7** the array $\{U_{ni}Z_{ni} - \mathrm{E}(U_{ni}Z_{ni}), \mathcal{F}^i_{n,-\infty}\}$ is therefore an L_1-mixingale with respect to the constant array $\{c^U_{ni} c^W_{ni}\}$ and Theorem **20.15** (for which the mixingale size is arbitrary) can be applied to show that $|A_{n2} - \mathrm{E}(A_{n2})| \to_{L_1} 0$.

That the conditions of **20.15** are met for A_{n2} is shown as follows. Uniform integrability (**20.15**(a)) follows because, with $r > 2$, making use successively of the Loève c_r inequality and the Jensen and Hölder inequalities,

$$\begin{aligned} \sup_{i,n} \mathrm{E}\left|\frac{U_{ni}Z_{ni} - \mathrm{E}(U_{ni}Z_{ni})}{c^U_{ni} c^W_{ni}}\right|^{r/2} &\le 2^{r/2} \sup_{i,n} \mathrm{E}\left|\frac{U_{ni}Z_{ni}}{c^U_{ni} c^W_{ni}}\right|^{r/2} \\ &\le 2^{r/2} \sup_{i,n} \left\|\frac{U_{ni}}{c^U_{ni}}\right\|_r^{r/2} \left\|\frac{Z_{ni}}{c^W_{ni}}\right\|_r^{r/2} < \infty. \end{aligned} \tag{32.75}$$

Also, by (32.65),

$$\sum_{i=1}^{n} c^U_{ni} c^W_{ni} = O(1) \tag{32.76}$$

$$\sum_{i=1}^{n} (c^U_{ni} c^W_{ni})^2 = O(1/n) \tag{32.77}$$

so that conditions **20.15**(b) and **20.15**(c) are also satisfied. It follows that $G_n - G_n^* - \mathrm{E}(A_{n2}) \to_{\mathrm{pr}} 0$. Finally, it can be verified that

$$Z_{ni} = \sum_{k=1}^{\infty} \mathrm{E}(W_{n,i+k}|\mathcal{F}_{ni})$$

and hence

$$\mathrm{E}(A_{n2}) = \sum_{i=1}^{n-1}\sum_{k=1}^{n-i} \mathrm{E}(U_{ni}\mathrm{E}(W_{n,i+k}|\mathcal{F}_{ni})) \to \Lambda_{XY} \tag{32.78}$$

and the proof is therefore complete. ∎

The conclusion of these arguments may be summarized in the following theorem. The joint convergence specified is a requisite for various arguments arising in the theory of asymptotic inference, allowing the application of the continuous mapping theorem **29.4** to functionals of the components.

32.14 Theorem If the array $\{U_{ni}, W_{ni}\}$ satisfies the assumptions of Theorem **31.5** and also those of either **32.12** or **32.13**, then

$$(X_n, Y_n, G_n) \xrightarrow{\mathrm{d}} \left(X, Y, \int_0^1 X\mathrm{d}Y + \Lambda_{XY}\right).$$

Proof Given the cited results, the remaining issue is the joint convergence of the triple (X_n, Y_n, G_n). This follows from **31.13**. To accommodate the style of that argument to the present case it suffices to map G_n into an element of $D_{[0,1]}$, which is constant over $[0,1]$ and hence continuous. ∎

Theorem **32.14** is the result whose proof is contained in this section. The assumptions can be replaced by those of Theorem **31.5** alone by applying the argument to be developed in §32.4.

The analysis of this section has dealt with the scalar integrals formed from pairs of scalar processes. Since the setting is inherently multivariate, it is of interest to fit the results into the general analysis of random vectors. Let $\{\boldsymbol{U}_{ni}\}$ $(m \times 1)$ be a vector array satisfying the conditions of **31.15**. Since **32.14** holds for each element paired with each element, including itself, the argument may be generalized in the following manner.

32.15 Theorem Let $\boldsymbol{X}_n(t) = \sum_{i=1}^{[nt]} \boldsymbol{U}_{ni}$ satisfy the conditions of **31.15**. Then, $\boldsymbol{X}_n \to_{\mathrm{d}} \boldsymbol{X} \sim_{\mathrm{d}} \boldsymbol{B}(\boldsymbol{\Omega})$ and

$$\sum_{j=1}^{n-1} \boldsymbol{X}_{nj}\boldsymbol{U}'_{n,j+1} \xrightarrow{\mathrm{d}} \int_0^1 \boldsymbol{X}\mathrm{d}\boldsymbol{X}' + \boldsymbol{\Lambda}\,(m\times m) \tag{32.79}$$

where $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\Lambda} + \boldsymbol{\Lambda}'$.

Proof For arbitrary m-vectors $\boldsymbol{\lambda}$ and $\boldsymbol{\mu}$ of unit length, the arrays $\{\boldsymbol{\lambda}'\boldsymbol{X}_{nj}\}$ and $\{\boldsymbol{\mu}'\boldsymbol{U}_{n,j+1}\}$ converge to scalar Brownian motions. Letting $\boldsymbol{G}_n$ denote the matrix on the left-hand side of (32.79) and $\boldsymbol{G}$ the matrix on the right-hand side, the result $\boldsymbol{\lambda}'\boldsymbol{G}_n\boldsymbol{\mu} \to_{\mathrm{d}} \boldsymbol{\lambda}'\boldsymbol{G}\boldsymbol{\mu}$ is therefore given. A well-known matrix formula (see e.g. Magnus and Neudecker [128] th. 2.2) yields

$$\boldsymbol{\lambda}'\boldsymbol{G}_n\boldsymbol{\mu} = (\boldsymbol{\mu}' \otimes \boldsymbol{\lambda}')\,Vec\,\boldsymbol{G}_n \tag{32.80}$$

where $\boldsymbol{\mu}' \otimes \boldsymbol{\lambda}'$ is the Kronecker product **??** of the vectors, the row vector with elements $(\mu_1\lambda_1, \ldots, \mu_1\lambda_m, \mu_2\lambda_1, \ldots, \mu_m\lambda_m)$ $(1\times m^2)$, and $Vec\,\boldsymbol{G}_n$ $(m^2\times 1)$ is the vector consisting of the columns of $\boldsymbol{G}_n$ stacked one above the other. $\boldsymbol{\mu}' \otimes \boldsymbol{\lambda}'$ is of unit length, and applying the Cramér–Wold theorem **26.5** in respect of (32.80) implies that $\boldsymbol{G}_n \to_{\mathrm{d}} \boldsymbol{G}$, as asserted in (32.79). ∎

This result is to be compared with **32.5**. Between them they provide the intriguing incidental information that

$$\int_0^1 \boldsymbol{B}\mathrm{d}\boldsymbol{B}' + \int_0^1 \mathrm{d}\boldsymbol{B}\boldsymbol{B}' \overset{\mathrm{d}}{\sim} \boldsymbol{B}(1)\boldsymbol{B}(1)' - \boldsymbol{\Omega}. \tag{32.81}$$

Note that the stochastic matrix on the right-hand side has rank 1.

Between them, **32.3** and **32.15** provide the basic theoretical tools necessary to analyse the linear regression model in variables generated as partial-sum processes. See Phillips and Durlauf ([145]) and Park and Phillips ([139], [140]) among many related references.

The extension to non-Brownian processes B_η is straightforward, paralleling the extension to **31.9**. The modification of proofs is a matter of replacing increments of the line representing variances with their transformed values under the mapping η. The expressions involved include (32.54), (32.60), and (32.69). Also note that the long-run covariance matrix is defined like $\boldsymbol{\Omega}$, not $\widetilde{\boldsymbol{\Omega}}$, in (31.41). The conclusion may then be stated formally as follows.

32.16 Theorem If $(U_{ni}, W_{ni})'$ satisfies the conditions of **31.16** and either **32.12** or **32.13** then

$$(X_n, Y_n, G_n) \xrightarrow{\mathrm{d}} \left(X, Y, \int_0^1 X\mathrm{d}Y + \Lambda_{XY}\right)$$

where $(X,Y)' \sim_{\mathrm{d}} \boldsymbol{B}_\eta(\boldsymbol{\Omega})$ with $\boldsymbol{\Omega} = \boldsymbol{\Sigma} + \boldsymbol{\Lambda} + \boldsymbol{\Lambda}'$ and Λ_{XY} is the (2,1) element of $\boldsymbol{\Lambda}$. □

32.4 Convergence in Probability to $\boldsymbol{\Lambda_{XY}}$

This section treats the surprisingly difficult problem of showing the convergence in probability to Λ_{XY} of the random variable $G_n - G_n^*$, with no stronger assumptions than are needed for the FCLT and hence for the integral convergence of **32.10**. Unlike the relatively straightforward special cases of **32.12** and **32.13**, the proof adapted from [54] to be set out in this section is quite lengthy, although the effort is rewarded by the various nice techniques using mixingales to be exhibited along the way.

32.17 Theorem If the pair $\{U_{ni}, W_{ni}\}$ satisfy the assumptions of Theorem **31.5**, $X_n(t) = \sum_{i=1}^{[nt]} U_{ni}$, $Y_n(t) = \sum_{j=1}^{[nt]} W_{nj}$, and G_n is defined by (32.49), then

$$(X_n, Y_n, G_n) \xrightarrow{\mathrm{d}} \left(X, Y, \int_0^1 X\mathrm{d}Y + \Lambda_{XY}\right).$$

Proof Let the definitions of sequences k_n and n_j for $j = 1, \ldots, k_n$ given in the paragraph before Theorem **32.10** be retained. Given **32.10** and **32.14** it remains to be shown that $G_n - G_n^* \to_{\mathrm{pr}} \Lambda_{XY}$ under the assumptions stated, where G_n^* is defined in (32.52).

Recalling $U_{n0} = 0$, note that by reassigning subscripts,

$$G_n - G_n^* = \sum_{j=1}^{k_n} \sum_{i=n_{j-1}}^{n_j - 1} \sum_{m=0}^{i-n_{j-1}} U_{n,i-m} W_{n,i+1}. \tag{32.82}$$

For a sequence $\{q_n\}$ to be chosen with $q_n \to \infty$ as $n \to \infty$, there exists the decomposition

$$\Lambda_{XY} = \mathrm{E}(G_n - G_n^*) + R_{1n} + R_{2n}$$

where

$$R_{1n} = \sum_{j=1}^{k_n} \sum_{i=n_{j-1}}^{n_{j-1}+q_n} \sum_{m=i-n_{j-1}+1}^{q_n} \mathrm{E}(U_{n,i-m} W_{n,i+1})$$

$$R_{2n} = \sum_{j=1}^{k_n} \sum_{i=n_{j-1}+q_n+1}^{n_j - 1} \sum_{m=q_n+1}^{i-1} \mathrm{E}(U_{n,i-m} W_{n,i+1}).$$

These remainders contain the autocovariances across sequence coordinates falling in different blocks. R_{1n} contains those terms for which which $m \leq q_n$, which are $k_n q_n^2$ in number. Therefore, according to Assumption **31.5**(c), choosing $q_n = o(n^{1/2}/k_n^{1/2})$ gives

$$R_{1n} = O\Big(k_n q_n^2 \max_{1\leq i\leq n}\{\max\{(c_{ni}^U)^2, (c_{ni}^W)^2\}\}\Big) = o(1).$$

On the other hand, $R_{2n} = O(q_n^{-\delta})$ for $\delta > 0$ by **17.19** under the assumptions.

The remaining task is to show the convergence in probability to zero of

$$\begin{aligned} G_n - G_n^* - \mathrm{E}(G_n - G_n^*) & \\ &= \sum_{j=1}^{k_n} \sum_{i=n_{j-1}}^{n_j - 1} \sum_{l=n_{j-1}}^{i} (U_{nl}W_{n,i+1} - \mathrm{E}(U_{nl}W_{n,i+1})) \end{aligned} \tag{32.83}$$

reverting here to writing U_{nl} in place of $U_{n,i-m}$. The first step is to break each of the random variables into truncated and tail components, where the latter are negligible and the former have all their moments. To do this define

$$g(a,x) = (x-a)1_{\{x>a\}} + (x+a)1_{\{x<-a\}}$$

$$h(a,x) = x1_{\{|x|\leq a\}} + a1_{\{x>a\}} - a1_{\{x<-a\}}$$

so that $x = g(a,x) + h(a,x)$. Then for some $K > 0$ to be chosen let

$$\tilde{U}_{ni} = g(Kc_{ni}^U, U_{ni}) - \mathrm{E}(g(Kc_{ni}^U, U_{ni}))$$

$$\bar{U}_{ni} = h(Kc_{ni}^U, U_{ni}) - \mathrm{E}(h(Kc_{ni}^U, U_{ni}))$$

noting that since it has zero mean, $U_{ni} = \tilde{U}_{ni} + \bar{U}_{ni}$. The important thing to note is that by Theorem **18.12** (compare **18.14**), $\bar{U}_{ni}$ and $\tilde{U}_{ni}$ are both L_2-NED on the same mixing process, of the same size as U_{ni} and with the same scale constants. Hence, by assumptions **31.5**(a) and (b) both are L_2-mixingales of size $-\frac{1}{2}$. In particular,

$$\|\mathrm{E}_{t-m}\tilde{U}_{ni}\|_2 \ll c_{ni}^U m^{-1/2-\mu}$$

for $\mu > 0$. However, it is also possible to write, using the Jensen inequality and law of iterated expectations and the definition of $\tilde{U}_{ni}$,

$$\|\mathrm{E}_{t-m}\tilde{U}_{ni}\|_2 \leq \|\tilde{U}_{ni}\|_2 \leq c_{ni}^U \sup_{i,n} \|(U_{ni}/c_{ni}^U)1_{\{U_{ni}/c_{ni}^U > K\}}\|_2 \leq c_{ni}^U f(K)$$

where $f(K)$ does not depend on i or n and $f(K) \to 0$ as $K \to \infty$ as a consequence of the uniform square-integrability assumption that is either implicit or explicit by assumption **31.5**(a). Combining both of these representations gives

$$||\mathrm{E}_{t-m}\tilde{U}_{ni}||_2 \ll c^U_{ni}(m^{-1/2-\mu})^{1-\mu}f(K)^\mu = c^U_{ni}m^{-1/2-\mu/2+\mu^2}f(K)^\mu$$

so that with $\mu < \frac{1}{2}$, $\tilde{U}_{ni}$ is an L_2-mixingale of size $-\frac{1}{2}$ with respect to scale constants $c^U_{ni}f(K)^\mu$. Exactly the same development can of course be applied to W_{ni} to define $\bar{W}_{ni}$ and $\tilde{W}_{ni}$ with respect to indices c^W_{ni}.

Now, note that for $l = 1, \ldots, i$ and $i = 1, \ldots, n-1$,

$$U_{nl}W_{n,i+1} = \bar{U}_{nl}\bar{W}_{n,i+1} + \tilde{U}_{nl}\bar{W}_{n,i+1} + \bar{U}_{nl}\tilde{W}_{n,i+1} + \tilde{U}_{nl}\tilde{W}_{n,i+1}$$

and likewise

$$\mathrm{E}(U_{nl}W_{n,i+1}) = \mathrm{E}(\bar{U}_{nl}\bar{W}_{n,i+1}) + \mathrm{E}(\tilde{U}_{nl}\bar{W}_{n,i+1}) + \mathrm{E}(\bar{U}_{nl}\tilde{W}_{n,i+1}) + \mathrm{E}(\tilde{U}_{nl}\tilde{W}_{n,i+1}).$$

Substitute these decompositions into (32.83) and apply the Minkowski inequality to obtain a sum of L_1 norms, using the modulus inequality to deal with the expectations. Applying Corollary **17.21** to these terms, it is possible to write

$$||G_n - G_n^* - \mathrm{E}(G_n - G_n^*)||_1 = ||\bar{A}_n||_1 + O(f(K)^\mu) \tag{32.84}$$

where

$$\bar{A}_n = \sum_{j=1}^{k_n}\sum_{i=n_{j-1}}^{n_j-1}\sum_{l=n_{j-1}}^{i}(\bar{U}_{nl}\bar{W}_{n,i+1} - \mathrm{E}(\bar{U}_{nl}\bar{W}_{n,i+1})) \tag{32.85}$$

and the remainder term can be made as small as desired by choosing K large enough.

The next step is to write $\bar{A}_n = \sum_{j=1}^{k_n} Y_{nj}$ where

$$Y_{nj} = \sum_{i=n_{j-1}}^{n_j-1}\sum_{l=n_{j-1}}^{i}(\bar{U}_{nl}\bar{W}_{n,i+1} - \mathrm{E}(\bar{U}_{nl}\bar{W}_{n,i+1})) \tag{32.86}$$

and consider the dependence of these blocks of observations. Define $\mathrm{E}_{j-m}^{j+m}(\cdot) = \mathrm{E}(\cdot|\boldsymbol{V}_{n,n_{j-m-1}+1}, \ldots, \boldsymbol{V}_{n,n_{j+m}})$. Then, applying the schematic decomposition in (18.25) and the Minkowski inequality, noting that $n_j - n_{j-1} = O(n/k_n)$,

$$
\begin{aligned}
\|Y_{nj} - \mathrm{E}_{j-m}^{j+m} Y_{nj}\|_1 \le{}& \sum_{i=n_{j-1}}^{n_j-1} \|\bar{W}_{n,i+1} - \mathrm{E}_{j-m}^{j+m} \bar{W}_{n,i+1}\|_2 \Bigl\| \sum_{l=n_{j-1}}^{i} \bar{U}_{nl} \Bigr\|_2 \\
&+ \sum_{i=n_{j-1}}^{n_j-1} \|\mathrm{E}_{j-m}^{j+m} \bar{W}_{n,i+1}\|_2 \sum_{l=n_{j-1}}^{i} \|\bar{U}_{nl} - \mathrm{E}_{j-m}^{j+m} \bar{U}_{nl}\|_2 \\
&+ \sum_{i=n_{j-1}}^{n_j-1} \|\bar{W}_{n,i+1} - \mathrm{E}_{j-m}^{j+m} \bar{W}_{n,i+1}\|_2 \sum_{l=n_{j-1}}^{i} \|\bar{U}_{nl} - \mathrm{E}_{j-m}^{j+m} \bar{U}_{nl}\|_2 \\
\ll{}& \sum_{i=n_{j-1}}^{n_j-1} c_{ni}^{W} \Bigl(\frac{mn}{k_n}\Bigr)^{-1/2-\mu} \Bigl(\sum_{i=n_{j-1}}^{n_j-1} (c_{ni}^{U})^2 \Bigr)^{1/2} \\
&+ \Bigl(\sum_{i=n_{j-1}}^{n_j-1} (c_{ni}^{W})^2 \Bigr)^{1/2} \sum_{l=n_{j-1}}^{i} c_{nl}^{U} \Bigl(\frac{mn}{k_n}\Bigr)^{-1/2-\mu} \\
&+ \sum_{i=n_{j-1}}^{n_j-1} c_{ni}^{W} \sum_{l=n_{j-1}}^{i} c_{nl}^{U} \Bigl(\frac{mn}{k_n}\Bigr)^{-1-2\mu} \\
\ll{}& \Bigl(\sum_{i=n_{j-1}}^{n_j-1} (c_{ni}^{U})^2 \sum_{i=n_{j-1}}^{n_j-1} (c_{ni}^{W})^2 \Bigr)^{1/2} \Bigl(\frac{n}{k_n}\Bigr)^{-\mu} m^{-1/2-\mu} \qquad (32.87)
\end{aligned}
$$

for $\mu > 0$. The bound on the first term is obtained since $\{\bar{W}_{ni}\}$ is L_2-NED of size $-\frac{1}{2}$ and by using Theorem **17.10** to bound the second factor. The bounds on the second and third terms are obtained similarly and the final inequality follows on using (2.24) with $p = 2$ to bound the sums of indices. For the case $m = 0$ the terms depending on mn/k_n in these formulae can be replaced by 1.

Letting $\mathcal{H}_{nj} = \sigma(\boldsymbol{V}_{nl}, l \le n_j)$, the array $\{Y_{nj}, \mathcal{H}_{nj}\}$ is thus shown to be L_1-NED of size $-\frac{1}{2}$ on the mixing array $\{\boldsymbol{V}_{nl}\}$ with scale constants

$$
c_{nj}^{Y} = \Bigl(\sum_{i=n_{j-1}}^{n_j-1} (c_{ni}^{U})^2 \sum_{i=n_{j-1}}^{n_j-1} (c_{ni}^{W})^2 \Bigr)^{1/2}. \qquad (32.88)
$$

Hence, it is also an L_1-mixingale of size $-\frac{1}{2}$ with respect to the c_{nj}^{Y} by application of **18.7** with $r = \infty$, noting that Y_{nj} possesses all its moments since it depends only on $\bar{W}_{ni}$ and $\bar{U}_{nl}$. The fact that $\sum_{j=1}^{k_n} c_{nj}^{Y} < \infty$ follows from (32.88) and the assumption $\sup_i \max\{c_{ni}^{U}, c_{ni}^{W}\} = O(n^{-1/2})$ from **31.5**(c).

To prove the theorem, the object must now be to show that

$$\sum_{j=1}^{k_n} Y_{nj} \xrightarrow{L_1} 0. \tag{32.89}$$

However, Theorem **20.15** cannot be applied at this point because the uniform integrability of the array $\{Y_{nj}/c_{nj}^Y\}$ cannot be established directly. Instead, note for any choice of m the decomposition

$$Y_{nj} = (Y_{nj} - \mathrm{E}_{j+m}Y_{nj}) + \sum_{q=-m}^{m-1} (\mathrm{E}_{j-q}Y_{nj} - \mathrm{E}_{j-q-1}Y_{nj}) + \mathrm{E}_{j-m}Y_{nj} \tag{32.90}$$

where $\mathrm{E}_j = \mathrm{E}(\cdot|\mathcal{H}_{nj})$. Thanks to the mixingale property, by taking m large enough the L_1-norms of the first and last terms in (32.90) can be made as small as desired. Hence, to show (32.89) it suffices to show that for each q,

$$\sum_{j=1}^{k_n} (\mathrm{E}_{j-q}Y_{nj} - \mathrm{E}_{j-q-1}Y_{nj}) \xrightarrow{L_1} 0. \tag{32.91}$$

To show this, let $\mathrm{E}_i(\cdot) = \mathrm{E}(\cdot|\mathcal{F}_{ni})$ where $\mathcal{F}_{ni} = \sigma(\boldsymbol{V}_{nl}, l \le i)$, and write $P_{nlm}^U = \mathrm{E}_{l+m}\bar{U}_{nl} - \mathrm{E}_{l-m}\bar{U}_{nl}$, so that

$$\bar{U}_{nl} - P_{nlm}^U = \mathrm{E}_{l-m}\bar{U}_{nl} + (\bar{U}_{nl} - \mathrm{E}_{l+m}\bar{U}_{nl}) \tag{32.92}$$

and similarly let $P_{nim}^W = \mathrm{E}_{i+m}\bar{W}_{ni} - \mathrm{E}_{i-m}\bar{W}_{ni}$ where

$$\bar{W}_{ni} - P_{nim}^W = \mathrm{E}_{i-m}\bar{W}_{ni} + (\bar{W}_{ni} - \mathrm{E}_{i+m}\bar{W}_{ni}). \tag{32.93}$$

It follows from from these equalities that $\{\bar{U}_{nl} - P_{nlm}^U, \mathcal{F}_{nl}\}$ and $\{\bar{W}_{ni} - P_{nim}^W, \mathcal{F}_{ni}\}$ are L_2-mixingales. By the definitions and application of **10.27**, note that

$$\begin{aligned} ||\mathrm{E}_{l-p}(\bar{U}_{nl} - P_{nlm}^U)||_2 &\le c_{nl}^U \zeta(\max\{m,p\}) \\ ||(\bar{U}_{nl} - P_{nlm}^U) - \mathrm{E}_{l+p}(\bar{U}_{nl} - P_{nlm}^U)||_2 &\le c_{nl}^U \zeta(\max\{m,p\}) \end{aligned}$$

and similarly

$$\begin{aligned} ||\mathrm{E}_{i-p}(\bar{W}_{ni} - P_{nim}^W)||_2 &\le c_{ni}^W \zeta(\max\{m,p\}) \\ ||(\bar{W}_{ni} - P_{nim}^W) - \mathrm{E}_{i+p}(\bar{W}_{ni} - P_{nim}^W)||_2 &\le c_{ni}^W \zeta(\max\{m,p\}) \end{aligned}$$

where ζ denotes the mixingale index. Write Y_{nj}^m to represent the random variable defined by the variant of (32.86) where P_{nlm}^U replaces $\bar{U}_{nl}$ and P_{nim}^W replaces $\bar{W}_{n,i+1}$. Then, according to (32.92) and (32.93),

$$\begin{aligned}\left\|\sum_{j=1}^{k_n}(Y_{nj}-Y_{nj}^m)\right\|_1 &\leq \left\|\sum_{j=1}^{k_n}\sum_{i=n_{j-1}}^{n_j-1}\sum_{l=n_{j-1}}^{i}(\bar{U}_{nl}-P_{nlm}^U)\bar{W}_{ni}\right\|_1 \\ &\quad+\left\|\sum_{j=1}^{k_n}\sum_{i=n_{j-1}}^{n_j-1}\sum_{l=n_{j-1}}^{i}P_{nlm}^U(\bar{W}_{ni}-P_{nim}^W)\right\|_1 \\ &\ll \left(\zeta(m)^2\sum_{p=1}^{m}(\log p)^2+\sum_{p=m+1}^{\infty}\zeta(p)^2(\log p)^2\right)^{1/2} \\ &= O(m^{-\mu}) \end{aligned} \tag{32.94}$$

for $\mu > 0$, noting that in each of the terms defined following the first inequality the variables in the sums are L_2-mixingales and the second inequality is got by applying **17.21** to each L_1 norm. Since m is arbitrary, it follows that if $\left\|\sum_{j=1}^{k_n} Y_{nj}^m\right\|_1 \to 0$ as $n \to \infty$ then (32.89) also holds and Y_{nj}^m can be substituted for Y_{nj} in (32.90) and (32.91). The theorem is equivalently proved by showing

$$\sum_{j=1}^{k_n}(\mathrm{E}_{j-q}Y_{nj}^m-\mathrm{E}_{j-q-1}Y_{nj}^m)\xrightarrow{L_1}0 \tag{32.95}$$

for each q. To this end, let

$$\mathrm{E}_{l+m}\bar{U}_{nl}-\mathrm{E}_{l-m}\bar{U}_{nl}=\sum_{g=-m}^{m-1}(\mathrm{E}_{l+g+1}\bar{U}_{nl}-\mathrm{E}_{l+g}\bar{U}_{nl})$$

and

$$\mathrm{E}_{i+m}\bar{W}_{n,i+1}-\mathrm{E}_{i-m}\bar{W}_{n,i+1}=\sum_{h=-m}^{m-1}(\mathrm{E}_{i+h+1}\bar{W}_{n,i+1}-\mathrm{E}_{i+h}\bar{W}_{n,i+1})$$

and so define

$$Z_{nj}^{gh}=\sum_{i=n_{j-1}}^{n_j-1}\sum_{l=n_{j-1}}^{i}(\mathrm{E}_{l+g+1}\bar{U}_{nl}-\mathrm{E}_{l+g}\bar{U}_{nl})(\mathrm{E}_{i+h+1}\bar{W}_{n,i+1}-\mathrm{E}_{i+h}\bar{W}_{n,i+1}).$$

This is the variant of (32.86) in which $\mathrm{E}_{l+g+1}\bar{U}_{nl}-\mathrm{E}_{l+g}\bar{U}_{nl}$ replaces $\bar{U}_{nl}$ and $\mathrm{E}_{i+h+1}\bar{W}_{n,i+1}-\mathrm{E}_{i+h}\bar{W}_{n,i+1}$ replaces $\bar{W}_{n,i+1}$, and

$$\sum_{j=1}^{k_n}(\mathrm{E}_{j-q}Y_{nj}^m - \mathrm{E}_{j-q-1}Y_{nj}^m) = \sum_{g=-m}^{m-1}\sum_{h=-m}^{m-1}\sum_{j=1}^{k_n}(\mathrm{E}_{j-q}Z_{nj}^{gh} - \mathrm{E}_{j-q-1}Z_{nj}^{gh}). \tag{32.96}$$

Consider any one of the $4m^2$ right-hand-side sums over k_n. Since this is a sum of uncorrelated and L_2-bounded random variables, a sufficient condition for

$$\sum_{j=1}^{k_n}(\mathrm{E}_{j-q}Z_{nj}^{gh} - \mathrm{E}_{j-q-1}Z_{nj}^{gh}) \xrightarrow{L_2} 0 \tag{32.97}$$

is

$$\sum_{j=1}^{k_n}\mathrm{E}(Z_{nj}^{gh})^2 \to 0. \tag{32.98}$$

However,

$$\begin{aligned}\mathrm{E}(Z_{nj}^{gh})^2 = \sum_{i_1=n_{j-1}}^{n_j-1}\sum_{i_2=n_{j-1}}^{n_j-1}\sum_{l_1=n_{j-1}}^{i_1}\sum_{l_2=n_{j-1}}^{i_2} &\mathrm{E}(\mathrm{E}_{l_1+g+1}\bar{U}_{nl_1} - \mathrm{E}_{l_1+g}\bar{U}_{nl_1}) \\ &\times(\mathrm{E}_{i_1+h+1}\bar{W}_{n,i_1+1} - \mathrm{E}_{i_1+h}\bar{W}_{n,i_1+1})(\mathrm{E}_{l_2+g+1}\bar{U}_{nl_2} - \mathrm{E}_{l_2+g}\bar{U}_{nl_2}) \\ &\qquad\times(\mathrm{E}_{i_2+h+1}\bar{W}_{n,i_2+1} - \mathrm{E}_{i_2+h}\bar{W}_{n,i_2+1}).\end{aligned}$$

Observe that these expectations vanish except in the cases where the two largest conditional expectation indices match, for example when $l_1+g+1 = i_1+h+1$. In other cases the increments are orthogonal by the martingale difference property. Given a particular (g,h) pairing, this matching can happen only for the single value of l_1 satisfying this equality. Hence, in this case,

$$\begin{aligned}\mathrm{E}(Z_{nj}^{gh})^2 = \mathrm{E}\biggl(\sum_{i_1=n_{j-1}}^{n_j-1}&(\mathrm{E}_{i_1+h+1}\bar{U}_{n,i_1+h-g} - \mathrm{E}_{i_1+h}\bar{U}_{n,i_1+h-g}) \\ &\times(\mathrm{E}_{i_1+h+1}\bar{W}_{n,i_1+1} - \mathrm{E}_{i_1+h}\bar{W}_{n,i_1+1}) \\ &\times\sum_{i_2=n_{j-1}}^{n_j-1}\sum_{l_2=n_{j-1}}^{i_2}(\mathrm{E}_{l_2+g+1}\bar{U}_{nl_2} - \mathrm{E}_{l_2+g}\bar{U}_{nl_2}) \\ &\qquad\times(\mathrm{E}_{i_2+h+1}\bar{W}_{n,i_2+1} - \mathrm{E}_{i_2+1}\bar{W}_{n,i_2+1})\biggr)\end{aligned} \tag{32.99}$$

(it being understood that $\bar{U}_{n,i_1+h-g} = 0$ if the time subscript falls outside the set $1,\dots,n$). The expectation on the right-hand side of (32.99) is dominated by the corresponding L_1-norm and so

$$
\begin{aligned}
\sum_{j=1}^{k_n} \mathrm{E}(Z_{nj}^{gh})^2 \leq \sum_{j=1}^{k_n} \Bigg\| \sum_{i_1=n_{j-1}}^{n_j-1} & (\mathrm{E}_{i_1+h+1}\bar{U}_{n,i_1+h-g} - \mathrm{E}_{i_1+h}\bar{U}_{n,i_1+h-g}) \\
& \times (\mathrm{E}_{i_1+h+1}\bar{W}_{n,i_1+1} - \mathrm{E}_{i_1+h}\bar{W}_{n,i_1+1}) \\
& \times \sum_{i_2=n_{j-1}}^{n_j-1} \sum_{l_2=n_{j-1}}^{i_2} (\mathrm{E}_{l_2+g+1}\bar{U}_{nl_2} - \mathrm{E}_{l_2+g}\bar{U}_{nl_2}) \\
& \times (\mathrm{E}_{i_2+h+1}\bar{W}_{n,i_2+1} - \mathrm{E}_{i_2+h}\bar{W}_{n,i_2+1}) \Bigg\|_1 \\
\leq \sum_{j=1}^{k_n} \Bigg\| \sum_{i_1=n_{j-1}}^{n_j-1} & (\mathrm{E}_{l_1+g+1}\bar{U}_{nl_1} - \mathrm{E}_{l_1+g}\bar{U}_{nl_1}) \\
& \times (\mathrm{E}_{l_1+g+1}\bar{W}_{n,i_1+1} - \mathrm{E}_{l_1+g}\bar{W}_{n,i_1+1}) \Bigg\|_\infty \\
& \times \Bigg\| \sum_{i_2=n_{j-1}}^{n_j-1} \sum_{l_2=n_{j-1}}^{i_2} (\mathrm{E}_{l_2+g+1}\bar{U}_{nl_2} - \mathrm{E}_{l_2+g}\bar{U}_{nl_2}) \\
& \times (\mathrm{E}_{i_2+h+1}\bar{W}_{n,i_2+1} - \mathrm{E}_{i_2+h}\bar{W}_{n,i_2+1}) \Bigg\|_1 \qquad (32.100)
\end{aligned}
$$

where the Hölder inequality **9.29** with $p = 1$ has been applied. Note that the L_∞-norm (essential supremum) in the majorant is the sum of the product of the bounds Kc_{nl}^U and Kc_{ni}^W where K is fixed and finite. Applying successively **2.22** to the first factor and **17.21** to the second one yields

$$
\begin{aligned}
\sum_{j=1}^{k_n} \mathrm{E}(Z_{nj}^{gh})^2 &= O\Bigg(\sum_{j=1}^{k_n} \sum_{i_1=n_{j-1}}^{n_j-1} c_{nl}^U c_{ni}^W \Bigg(\sum_{l=n_{j-1}}^{n_j-1} (c_{nl}^U)^2 \sum_{i=n_{j-1}}^{n_j-1} (c_{ni}^W)^2 \Bigg)^{1/2} \Bigg) \\
&= O\Bigg(\sum_{j=1}^{k_n} \Bigg(\sum_{l=n_{j-1}}^{n_j-1} (c_{nl}^U)^2 \sum_{i=n_{j-1}}^{n_j-1} (c_{ni}^W)^2 \Bigg) \Bigg) \\
&= O\Bigg(\max_{1\leq j\leq k_n} \Bigg\{ \Bigg(\sum_{l=n_{j-1}}^{n_j-1} (c_{nl}^U)^2 \sum_{i=n_{j-1}}^{n_j-1} (c_{ni}^W)^2 \Bigg)^{1/2} \Bigg\} \Bigg) \\
&= o(1). \qquad (32.101)
\end{aligned}
$$

Here, the penultimate equality of (32.101) is validated by assumption **31.5**(c) which implies the sums of squares are each of $O(1/k_n)$ and the last one is by assumption **31.5**(d) in respect of each variable. Identical reasoning can be applied to prove (32.98) and hence (32.97) for each of the pairs (g, h), which in turn proves the convergence to zero of (32.96), which since m can be chosen arbitrarily large in (32.94) proves $\bar{A}_n \to_{L_1} 0$. In view of (32.84), this proves the theorem. ∎

Bibliography

[1] Abramowitz, M. and I. A. Stegun (1965), *Handbook of Mathematical Functions*, Dover Publications, New York.

[2] Aldous, D. J. and G. K. Eagleson (1978), 'On mixing and stability of limit theorems', *Annals of Probability* 6(2), 325–331.

[3] Amemiya, T. (1985), *Advanced Econometrics*, Basil Blackwell, Oxford.

[4] Andrews, D. W. K. (1984), 'Non-strong mixing autoregressive processes', *Journal of Applied Probability* 21, 930–4.

[5] Andrews, D. W. K. (1987), 'Consistency in nonlinear econometric models: a generic uniform law of large numbers', *Econometrica* 55, 1465–71.

[6] Andrews, D. W. K. (1988), 'Laws of large numbers for dependent non-identically distributed random variables', *Econometric Theory* 4, 458–67.

[7] Andrews, D. W. K. (1991), 'Heteroscedasticity and autocorrelation consistent covariance matrix estimation', *Econometrica* 59, 817–58.

[8] Andrews, D. W. K. (1992), 'Generic uniform convergence', *Econometric Theory* 8, 241–57.

[9] Apostol, T. M. (1974), *Mathematical Analysis* (2nd edn.) Addison-Wesley, Menlo Park, CA.

[10] Ash, R. B. (2000), *Probability and Measure Theory,* Academic Press/Harcourt,San Diego.

[11] Athreya, K. B. and S. L. Lahiri (2006) *Measure Theory and Probability Theory*, Springer, New York.

[12] Athreya, K. B. and S. G. Pantula (1986a), 'Mixing properties of Harris chains and autoregressive processes', *Journal of Applied Probability* 23, 880–92.

[13] Athreya, K. B. and S. G. Pantula (1986b), 'A note on strong mixing of ARMA processes', *Statistics and Probability Letters* 4, 187–90.

[14] Azuma, K. (1967), 'Weighted sums of certain dependent random variables', *Tohoku Mathematical Journal* 19, 357–67.

[15] Bates, C. and H. White (1985), 'A unified theory of consistent estimation for parametric models', *Econometric Theory* 1, 151–78.

[16] Bernstein, S. (1927), 'Sur l'extension du theoreme du calcul des probabilities aux sommes de quantites dependantes', *Mathematische Annalen* 97, 1–59.

[17] Berry, A. C. (1941), 'The accuracy of the Gaussian approximation to the sum of independent variates'. *Transactions of the American Mathematical Society* 49(1), 122–136.

[18] Beveridge, S. and C. R. Nelson (1981), 'A new approach to decomposition of economic time series into permanent and transitory components with particular attention to the measurement of the business cycle'. *Journal of Monetary Economics* 7, 151–74.

[19] Bierens, H. (1983), 'Uniform consistency of kernel estimators of a regression function under generalized conditions', *Journal of the American Statistical Association* 77, 699–707.

[20] Bierens, H. (1991), 'Least squares estimation of linear and nonlinear ARMAX models under data heterogeneity', *Annales d'Économie et de Statistique* No. 20/21, L'Hétérogénéité en Économétrie / Heterogeneity in Econometrics (Oct. 1990–Mar. 1991), 143–69.

[21] Billingsley, P. (1968), *Convergence of Probability Measures*, John Wiley & Sons, New York.

[22] Billingsley, P. (1999), *Convergence of Probability Measures* (2nd edn.), John Wiley & Sons, New York.

[23] Billingsley, P. (2012), *Probability and Measure* (anniversary edn.), John Wiley & Sons, New York.

[24] Bradley, R. C. (1988), 'On some results of M. I. Gordin: a clarification of a misunderstanding', *Journal of Theoretical Probability* 1(2), 115–19.

[25] Bradley, R. C., W. Bryc, and S. Janson, (1987), 'On domination between measures of dependence', *Journal of Multivariate Analysis* 23, 312–29.

[26] Breiman, L. (1968), *Probability*, Addison-Wesley, Reading, MA.

[27] Brown, B. M. (1971), 'Martingale central limit theorems', *Annals of Mathematical Statistics* 42, 59–66.

[28] Burkholder, D. L. (1966), 'Martingale transforms', *Annals of Mathematical Statistics* 37(6), 1494–504.

[29] Burkholder, D. L. (1973), 'Distribution function inequalities for martingales', *Annals of Probability* 1, 19–42.

[30] Chan, N. H. and C. Z. Wei, (1988), 'Limiting distributions of least squares estimates of unstable autoregressive processes', *Annals of Statistics,* 16, 367–401.

[31] Chanda, K. C. (1974), 'Strong mixing properties of linear stochastic processes', *Journal of Applied Probability* 11, 401–8.

[32] Chow, Y. S. (1971), 'On the L_p convergence for $n^{-1/p}S_n, 0 < p < 2$', *Annals of Mathematical Statistics* 36, 393–4.

[33] Chow, Y. S. and Teicher, H. (1978), *Probability Theory: Independence, Interchangeability and Martingales,* Springer-Verlag, Berlin.

[34] Chung, K. L. (1974), *A Course in Probability Theory* (2nd edn.), Academic Press, Orlando. FA.

[35] Cox, D. R. and H. D. Miller (1965), *The Theory of Stochastic Processes*, Methuen, London.

[36] Cramér, H. (1946), *Mathematical Methods of Statistics*, Princeton University Press, Princeton, NJ.

[37] Csörgő, M. and P. Révész (1981), *Strong Approximations in Probability and Statistics*, Academic Press, New York.

[38] Davidson, J. (1992), 'A central limit theorem for globally nonstationary near-epoch dependent functions of mixing processes', *Econometric Theory,* 8, 313–29.

[39] Davidson, J. (1993a), 'An L_1-convergence theorem for heterogeneous mixingale arrays with trending moments', *Statistics and Probability Letters* 16, 301–4.

[40] Davidson, J. (1993b), 'The central limit theorem for globally nonstationary near-epoch dependent functions of mixing process: the asymptotically degenerate case', *Econometric Theory* 9, 402–12.

[41] Davidson, J. (2000), *Econometric Theory*, Blackwell Publishers, Oxford.

[42] Davidson, J. (2002a), 'Establishing conditions for the functional central limit theorem in nonlinear and semiparametric time series processes', *Journal of Econometrics* 106(2), 243–69.

[43] Davidson, J. (2002b), 'Establishing conditions for the functional central limit theorem in nonlinear and semiparametric time series processes: Corrigendum', *Journal of Econometrics* 110(1), 103–4.

[44] Davidson, J. (2004), 'Moment and Memory Properties of Linear Conditional Heteroscedasticity Models, and a New Model', *Journal of Business & Economic Statistics* 22, 16–29.

[45] Davidson, J. and R. M. de Jong (1997), 'Strong laws of large numbers for dependent heterogeneous processes: a synthesis of recent and new results', *Econometric Reviews* 16(3) 251–79.

[46] Davidson, J. and R. M. de Jong (2002), 'Consistency of kernel variance estimators for sums of semiparametric linear processes', *Econometrics Journal* 5, 160–75.

[47] Davidson, J. and N. Hashimzade (2009), 'Type I and type II fractional Brownian motions: a reconsideration', *Computational Statistics and Data Analysis* 53(6), 2089–106.

[48] Davidson, J. and X. Li (2016), 'Strict stationarity, persistence and volatility forecasting in ARCH(∞) processes', *Journal of Empirical Finance* 38B, 534–47.

[49] de Jong, R. M. (1995), 'Laws of large numbers for dependent heterogeneous processes', *Econometric Theory* 11, 347–58.

[50] de Jong, R. M. (1996), 'A strong law of large numbers for triangular mixingale arrays', *Statistics and Probability Letters* 27, 1–9.

[51] de Jong, R. M. (1997), 'Central limit theorems for dependent heterogeneous random variables', *Econometric Theory* 13, 353–67.

[52] de Jong, R. M. (2000), 'A strong consistency proof for heteroskedasticity and autocorrelation consistent covariance matrix estimators', *Econometric Theory* 16(2), 262–68.

[53] de Jong, R. M. and J. Davidson (2000a), 'Consistency of kernel estimators of heteroskedastic and autocorrelated covariance matrices', *Econometrica* 68, 407–24.

[54] de Jong, R. M. and J. Davidson (2000b), 'The functional central limit theorem and weak convergence to stochastic integrals I: weakly dependent processes'. *Econometric Theory* 16(5), 621–42.

[55] Dellacherie, C. and P.-A. Meyer (1978), *Probabilities and Potential,* North–Holland, Amsterdam.

[56] Dhrymes, P. J. (1989), *Topics in Advanced Econometrics,* Springer-Verlag, New York.

[57] Dieudonné, J. (1969), *Foundations of Modern Analysis,* Academic Press, New York and London.

[58] Domowitz, I. and H. White (1982), 'Misspecified models with dependent observations', *Journal of Econometrics* 20, 35–58.

[59] Donsker, M. D. (1951), 'An invariance principle for certain probability limit theorems', *Memoirs of the American Mathematical Society*, 6, 1–12.

[60] Doob, J. L. (1953), *Stochastic Processes,* John Wiley, New York; Chapman & Hall, London.

[61] Dudley, R. M. (1966), 'Weak convergence of probabilities on nonseparable metric spaces and empirical measures on Euclidean spaces', *Illinois Journal of Mathematics* 10, 109–26.

[62] Dudley, R. M. (1967), 'Measures on non-separable metric spaces', *Illinois Journal of Mathematics* 11, 109–26.

[63] Dudley, R. M. (1989), *Real Analysis and Probability,* Wadsworth and Brooks/Cole, Pacific Grove, CA.

[64] Durrett, R. (2019), *Probability: Theory and Examples*, Cambridge University Press, Cambridge.

[65] Dvoretsky, A. (1972), 'Asymptotic normality of sums of dependent random variables', in *Proceedings of the Sixth Berkeley Symposium on Mathematical Statistics and Probability*, ii, University of California Press, Berkeley, CA 513–35.

[66] Eberlein, E. and M. S. Taqqu (eds.) (1986), *Dependence in Probability and Statistics: a Survey of Recent Results,* Birkhauser, Boston.

[67] Einstein, A. (1956), *Investigations on the Theory of Brownian Movement*, Dover Publications Inc., New York.

[68] Engle, R. F., D. F. Hendry, and J.-F. Richard (1983), 'Exogeneity', *Econometrica* 51, 277–304.

[69] Esséen, C.-G. (1942), 'On the Liapounoff limit of error in the theory of probability'. *Arkiv för matematik, astronomi och fysik* 28A No. 9, 1–19.

[70] Esséen, C.-G. (1956), 'A moment inequality with an application to the central limit theorem', *Skand. Aktuar. Tidskr*. XXXIX, 160–70.

[71] Esséen, C.-G. and S. Janson (1985), 'On moment conditions for normed sums of independent variables and martingale differences', *Stochastic Processes and their Applications* 19, 173–82.

[72] Etemadi, N (1981), 'An elementary proof of the strong law of large numbers', *Z. Wahrscheinlichkeitstheorie verw. Gebiete* 55, 119–22.

[73] Feller, W. (1968), *An Introduction to Probability Theory and its Applications*, Vol. I, John Wiley, New York.

[74] Feller, W. (1971), *An Introduction to Probability Theory and its Applications*, Vol. II, John Wiley, New York.

[75] Gallant, A. R. (1987), *Nonlinear Statistical Models,* John Wiley, New York.

[76] Gallant, A. R. and H. White (1988), *A Unified Theory of Estimation and Inference for Nonlinear Dynamic Models*, Basil Blackwell, Oxford.

[77] Gastwirth, J. L. and H. Rubin (1975), 'The asymptotic distribution theory of the empiric CDF for mixing stochastic processes', *Annals of Statistics* 3, 809–24.

[78] Gikhman, I. I. and A. V. Skorokhod (1969), *Introduction to the Theory of Random Processes,* W. B. Saunders Co., Philadelphia.

[79] Gikhman, I. I. and A. V. Skorokhod (1974), *Theory of Stochastic Processes* (3 vols.) Springer-Verlag, Berlin.

[80] Gnedenko, B. V. (1967), *The Theory of Probability* (4th edn.), Chelsea Publishing, New York.

[81] Gnedenko, B. V. and A. Ya. Khinchine (2013), *An Elementary Introduction to the Theory of Probability* (5th edn.), Dover Publications Inc., New York.

[82] Gnedenko, B. V. and A. N. Kolmogorov (1954), *Limit Distributions for Sums of Independent Random Variables*, Addison-Wesley Publishing Co., Reading, MA.

[83] Gordin, M. I. (1973), 'Central limit theorems for stationary processes without the assumption of finite variance'. *Abstracts of Communications* T. I:A-K. 173–74, International Conference on Probability Theory and Mathematical Statistics, June 25–30, 1973, Vilnius (in Russian).

[84] Gorodetskii, V. V. (1978), ' On the strong mixing property for linear sequences', *Theory of Probability and its Applications* 22, 411–13.

[85] Gradshteyn, I. S. and I. M. Ryzhik (2007), *Table of Integrals, Series and Products* (7th edn.), Academic Press, San Diego.

[86] Gray, R. M. (2006), *Toeplitz and Circulant Matrices: A Review.* Foundations and Trends in Communications and Information Theory, Now Publishers.

[87] Gut, A. (2013), *Probability: a Graduate Course* (2nd edn.), Springer, New York.

[88] Hall, P. and C. C. Heyde (1980), *Martingale Limit Theory and its Application,* Academic Press, New York and London.

[89] Halmos, P. R. (1956), *Lectures in Ergodic Theory,* Chelsea Publishing, New York.

[90] Halmos, P. R. (1960), *Naive Set Theory,* Van Nostrand Reinhold, New York.

[91] Halmos, P. R. (1974), *Measure Theory,* Springer-Verlag, New York.

[92] Hannan, E. J. (1970), *Multiple Time Series,* John Wiley, New York.

[93] Hansen, B. E. (1991), 'Strong laws for dependent heterogeneous processes', *Econometric Theory* 7, 213–21.

[94] Hansen, B. E. (1992a), 'Errata', *Econometric Theory* 8, 421–2.

[95] Hansen, B. E. (1992b), 'Consistent covariance matrix estimation for dependent heterogeneous processes', *Econometrica* 60, 967–72.

[96] Hansen, B. E. (1992c), 'Convergence to Stochastic integrals for dependent heterogeneous processes', *Econometric Theory* 8, 489–500.

[97] Hartman, P. and A. Wintner (1941), 'On the law of the iterated logarithm', *American Journal of Mathematics* 63, 169–76.

[98] Herrndorf, N. (1983), 'Stationary strongly mixing sequences not satisfying the central limit theorem, *Annals of Probability* 11(3), 809–13.

[99] Herrndorf, N. (1984), 'A functional central limit theorem for weakly dependent sequences of random variables', *Annals of Probability* 12, 141–53.

[100] Herrndorf, N. (1985), 'A functional central limit theorem for strongly mixing sequences of random variables', *Z. Wahrscheinlichkeitstheorie verw. Gebiete* 69, 540–50.

[101] Hoadley, B. (1971), 'Asymptotic properties of maximum likelihood estimators for the independent not identically distributed case', *Annals of Mathematical Statistics* 42, 1977–91.

[102] Hoeffding, W. (1963), 'Probability inequalities for sums of bounded random variables', *Journal of the American Statistical Association* 58, 13–30.

[103] Ibragimov, I. A. (1962), 'Some limit theorems for stationary processes', *Theory of Probability and its Applications* 7, 349–82.

[104] Ibragimov, I. A. (1965), 'On the spectrum of stationary Gaussian sequences satisfying the strong mixing condition. I: Necessary conditions', *Theory of Probability and its Applications* 10, 85–106.

[105] Ibragimov, I. A. and Yu. V. Linnik (1971), *Independent and Stationary Sequences of Random Variables,* Wolters–Noordhoff, Groningen.

[106] Iosifescu, M. and R. Theodorescu (1969), *Random Processes and Learning,* Springer-Verlag, Berlin.

[107] Jacod, J. and A. N. Shiryaev (2002), *Limit Theorems for Stochastic Processes* (2nd edn.), Springer, Berlin.

[108] Jansson, M. (2002), 'Consistent Covariance Matrix Estimation for Linear Processes', *Econometric Theory* 18(6), 1449–59.

[109] Kallenberg, O. (2001), *Foundations of Modern Probability* Springer-Verlag, New York.

[110] Karamata, J. (1930), 'Sur une mode de croissance régulière des fonctions', *Mathematica* (Cluj) 4, 38–53.

[111] Karatzas, I. and S. E. Shreve (1991), *Brownian Motion and Stochastic Calculus* (2nd edn.), Springer-Verlag, New York.

[112] Kelley, J. L. (1955), *General Topology*, Springer-Verlag, New York.

[113] Kingman, J. F. C. and S. J. Taylor (1966), *Introduction to Measure and Probability*, Cambridge University Press, London and New York.

[114] Kolmogorov, A. N. (1950), *Foundations of the Theory of Probability*, Chelsea Publishing, New York (published in German as *Grundbegriffe der Wahrscheinlichkeitsrechnung*, Springer-Verlag, Berlin, 1933).

[115] Kolmogorov, A. N. and Yu. A. Rozanov (1960), 'On strong mixing conditions for stationary Gaussian processes', *Theory of Probability and its Applications* 5, 204–8.

[116] Kopp, P. E. (1984), *Martingales and Stochastic Integrals*, Cambridge University Press, Cambridge.

[117] Korolev, V. and I. Shevtsova (2012), 'An improvement of the Berry–Esséen inequality with applications to Poisson and mixed Poisson random sums', *Scandinavian Actuarial Journal* 2012(2), 81–105.

[118] Kurtz, T. G. and P. Protter (1991), 'Weak limit theorems for stochastic integrals and stochastic differential equations', *Annals of Probability* 19, 1035–70.

[119] L'Ecuyer, P. (2012), 'Random Number Generation', *Handbook of Computational Statistics: Concepts and Methods* (2nd edn.), J. E. Gentle, W. Haerdle, and Y. Mori (eds.), Springer, Heidelberg. 35–71.

[120] LePage, R., M. Woodroofe, and J. Zinn (1981), 'Convergence to a stable distribution via order statistics', *Annals of Probability* 9(4), 624–32.

[121] Loève, M. (1977), *Probability Theory,* (4th edn.), Springer-Verlag, New York.

[122] Lukacs, E. (1975), *Stochastic Convergence* (2nd edn.), Academic Press, New York.

[123] McKean, H. P., Jr. (1969), *Stochastic Integrals*, Academic Press, New York.

[124] McLeish, D. L. (1974), 'Dependent central limit theorems and invariance principles', *Annals of Probability* 2(4) 620–8.

[125] McLeish, D. L. (1975a), 'A maximal inequality and dependent strong laws', *Annals of Probability* 3,5, 329–39.

[126] McLeish, D. L. (1975b), 'Invariance principles for dependent variables', *Z. Wahrscheinlichkeitstheorie verw. Gebiete* 32, 165–78.

[127] McLeish, D. L. (1977), 'On the invariance principle for nonstationary mixingales', *Annals of Probability* 5(4), 616–21.

[128] Magnus, J. R., and H. Neudecker (1988), *Matrix Differential Calculus with Applications in Statistics and Econometrics,* John Wiley, Chichester.

[129] Mandelbrot, B. B. (1983), *The Fractal Geometry of Nature,* W. H. Freeman, New York.

[130] Mandelbrot, B. B. and J. van Ness (1968), 'Fractional Brownian motions, fractional noises and applications', *SIAM Review* 10(4), 422–37.

[131] Mann, H. B. and A. Wald (1943a), 'On the statistical treatment of linear stochastic difference equations', *Econometrica* 11, 173–220.

[132] Mann, H. B. and A. Wald (1943b), 'On stochastic limit and order relationships', *Annals of Mathematical Statistics* 14, 390–402.

[133] Marcinkiewicz, J. and A. Zygmund (1938), Quelques theoremes sur les fonctions independantes. *Studia Mathematica* 7, 104–20.

[134] Nagaev, S. V. and A. Kh. Fuk (1971), 'Probability inequalities for sums of independent random variables', *Theory of Probability and its Applications* 6, 643–60.

[135] Nelson, D. (1990), 'Stationarity and persistence in the GARCH(1,1) model', *Econometric Theory* 6, 318–34.

[136] Newey, W. K. (1991), 'Uniform convergence in probability and stochastic equicontinuity', *Econometrica* 59, 1161–8.

[137] Newey, W. K. and West, K. (1987), 'A simple positive definite heteroskedasticity and correlation consistent covariance matrix', *Econometrica* 55, 703–8.

[138] Novikov, A. A. (1971), 'On moment inequalities for stochastic integrals', *Theory of Probability and its Applications* 16(3), 538–41.

[139] Park, J. Y. and P. C. B. Phillips (1988), 'Statistical inference in regressions with integrated processes, Part 1', *Econometric Theory* 4, 468–97.

[140] Park, J. Y. and P. C. B. Phillips (1989), 'Statistical inference in regressions with integrated processes, Part 2', *Econometric Theory* 5, 95–132.

[141] Parthasarathy, K. R. (1967), *Probability Measures on Metric Spaces*, Academic Press, New York and London.

[142] Pham, T. D. and L. T. Tran (1985), 'Some mixing properties of time series models', *Stochastic Processes and their Applications* 19, 297–303.

[143] Phillips, P. C. B. (1988a), 'Weak convergence to the matrix stochastic integral $\int_0^1 BdB'$', *Journal of Multivariate Analysis* 24, 252–64.

[144] Phillips, P. C. B. (1988b), 'Weak convergence of sample covariance matrices to stochastic integrals via martingale approximations', *Econometric Theory* 4, 528–33.

[145] Phillips, P. C. B. and S. N. Durlauf (1986), 'Multiple time series regression with integrated processes', *Review of Economic Studies* 53, 473–95.

[146] Phillips, P. C. B. and V. Solo (1992), 'Asymptotics for linear processes', *Annals of Statistics* 20(2), 971–1001.

[147] Pollard, D. (1984), *Convergence of Stochastic Processes,* Springer-Verlag, New York.

[148] Pötscher, B. M. and I. R. Prucha (1989), 'A uniform law of large numbers for dependent and heterogeneous data processes', *Econometrica* 57, 675–84.

[149] Pötscher, B. M. and I. R. Prucha (1991a), 'Basic structure of the asymptotic theory in dynamic nonlinear econometric models, Part I: Consistency and approximation concepts', *Econometric Reviews* 10, 125–216.

[150] Pötscher, B. M. and I. R. Prucha (1991b), 'Basic structure of the asymptotic theory in dynamic nonlinear econometric models, Part II: Asymptotic normality', *Econometric Reviews* 10, 253–325.

[151] Pötscher, B. M. and I. R. Prucha (1994), 'Generic uniform convergence and equicontinuity concepts for random functions: an exploration of the basic structure', *Journal of Econometrics* 60, 23–63.

[152] Priestley, M. B. (1988), *Non-Linear and Non-Stationary Time Series Analysis*. Academic Press, London.

[153] Prokhorov, Yu. V. (1956), 'Convergence of random processes and limit theorems in probability theory', *Theory of Probability and its Applications* 1, 157–213.

[154] Protter, P. E. (2005), *Stochastic Integration and Differential Equations*, (2nd edn.) Springer Verlag, Berlin.

[155] Rao, C. R. (1973), *Linear Statistical Inference and its Applications* (2nd edn.), John Wiley, New York.

[156] Rényi, A. (1963), 'On stable sequences of events', *Sankhya* Series A, 25(3), 293–302.

[157] Rényi, A. and P. Révész (1958), 'On mixing sequences of random variables', *Acta Mathematica* (Hungarian Academy of Sciences) 9, 389–94.

[158] Révész, Pál (1968), *The Laws of Large Numbers,* Academic Press, New York.

[159] Rosenblatt, M. (1956), 'A central limit theorem and a strong mixing condition', *Proceedings of the National Academy of Science, USA,* 42, 43–7.

[160] Rosenblatt, M. (1972), 'Uniform ergodicity and strong mixing', *Z. Wahrscheinlichkeitstheorie verw. Gebiete* 24, 79–84.

[161] Rosenblatt, M. (1978), 'Dependence and asymptotic independence for random processes', in *Studies in Probability Theory* (ed. M. Rosenblatt), Mathematical Association of America, Washington, DC.

[162] Royden, H. L. (1988), *Real Analysis* (3rd edn*)* Macmillan, New York.

[163] Samorodnitsky, G. and M. S. Taqqu (1994), *Stable Non-Gaussian Random Processes: Stochastic Models with Infinite Variance,* Chapman & Hall/CRC, Boca Raton FL.

[164] Seneta, E. (1976), *Regularly Varying Functions*, Springer-Verlag, Berlin.

[165] Serfling, R. J. (1968), 'Contributions to central limit theory for dependent variables', *Annals of Mathematical Statistics* 39, 1158–75.

[166] Serfling, R. J. (1980), *Approximation Theorems of Mathematical Statistics*, John Wiley, New York.

[167] Shafer, G. (1988), 'The St Petersburg Paradox', in *Encyclopaedia of the Statistical Sciences,* viii (ed. S. Kotz and N. L. Johnson), John Wiley, New York.

[168] Shevtsova, I. G. (2011), 'On the asymptotically exact constants in the Berry–Esséen–Katz inequality'. *Theory of Probability and Its Applications* 55(2), 225–52.

[169] Shiryaev, A. N. (1984), *Probability*, Springer-Verlag, New York.

[170] Skorokhod, A. V. (1956), 'Limit theorems for stochastic processes', *Theory of Probability and its Applications* 1, 261–90.

[171] Skorokhod, A. V. (1957), 'Limit theorems for stochastic processes with independent increments', *Theory of Probability and its Applications* 2, 138–71.

[172] Skorokhod, A. V. (1965), *Studies in the Theory of Random Processes*. Addison-Wesley Publishing Co., Reading, MA.

[173] Slutsky, E. (1925), 'Über stochastiche Asymptoter und Grenzwerte', *Metron* 5(3), 3–89.

[174] Stinchcombe. M. B. and H. White (1992), 'Some measurability results for extrema of random functions over random sets', *Review of Economic Studies* 59, 495–514.

[175] Stone, C. (1963), 'Weak convergence of stochastic processes defined on semi-infinite time intervals', *Proceedings of the American Mathematical Society* 14, 694–6.

[176] Stout, W. F. (1974), *Almost Sure Convergence,* Academic Press, New York.

[177] Strassen, V. (1964), 'An invariance principle for the law of the iterated logarithm', *Z. Wahrscheinlichkeitstheorie verw. Gebiete* 3, 211–26.

[178] Strasser, H. (1986), 'Martingale difference arrays and stochastic integrals', *Probability Theory and Related Fields* 72, 83–98.

[179] Tong, H. (1990), *Non-Linear Time Series. A Dynamical System Approach*, Clarendon Press, Oxford.

[180] Uchaikin, V. V. and V. M. Zolotarev (1999), *Chance and Stability: Stable Distributions and their Applications,* de Gruyter, Berlin.

[181] van der Vaart, A. W. (1998), *Asymptotic Statistics,* Cambridge University Press, Cambridge.

[182] van der Vaart, A. W. and J. A. Wellner (1996), *Weak Convergence and Empirical Processes*, Springer Verlag, New York.

[183] Varadarajan, V. S. (1958), 'Weak convergence of measures on separable metric spaces', *Sankhya* 19, 15–22.

[184] von Bahr, B. and C.-G. Esséen (1965), 'Inequalities for the rth absolute moment of a sum of random variables, $1 \leq r \leq 2$', *Annals of Mathematical Statistics* 36, 299–303.

[185] White, H. (1984), *Asymptotic Theory for Econometricians*, Academic Press, New York.

[186] White, H. (2001), *Asymptotic Theory for Econometricians (revised edn.)* Academic Press, New York.

[187] White, H. and I. Domowitz (1984), 'Nonlinear regression with dependent observations', *Econometrica* 52, 143–62.

[188] Wiener, N. (1923), 'Differential space', *Journal of Mathematical Physics* 2, 131–74.

[189] Willard, S. (1970), *General Topology,* Addison-Wesley, Reading, MA.

[190] Williams, D. (1991), *Probability with Martingales*, Cambridge University Press, Cambridge.

[191] Withers, C. S. (1981a), 'Conditions for linear processes to be strong-mixing', *Z. Wahrscheinlichkeitstheorie verw. Gebiete* 57, 477–80.

[192] Withers, C. S. (1981b), 'Central limit theorems for dependent variables, I', *Z. Wahrscheinlichkeitstheorie verw. Gebiete* 57, 509–34.

[193] Wold, H. (1938), *A Study in the Analysis of Stationary Time Series*, Almqvist and Wiksell, Uppsala Almqvist and Wiksell.

[194] Wooldridge, J. M. and H. White (1988), 'Some invariance principles and central limit theorems for dependent heterogeneous processes', *Econometric Theory* 4, 210–30.

[195] Zolotarev, V. M. (1986), *One-dimensional Stable Distributions*, American Mathematical Society, Providence, RI.

[196] Zygmund, A. (2002), *Trigonometric Series, Vol. 1* (3rd edn.), Cambridge University Press, Cambridge.

Index