David Pollard

Convergence of Stochastic Processes

With 36 Illustrations

Springer-Verlag
New York Berlin Heidelberg Tokyo

David Pollard
Department of Statistics
Yale University
New Haven, CT 06520
U.S.A.

AMS Subject Classifications: 60F99, 60G07, 60H99, 62M99

Library of Congress Cataloging in Publication Data
Pollard, David
Convergence of stochastic processes.
(Springer series in statistics)
Bibliography: p.
Includes index.
1. Stochastic processes. 2. Convergence. I. Title.
II. Series.
QA274.P64 1984 519.2 84-1401

Typeset by Composition House Ltd., Salisbury, England.
Printed and bound by R. R. Donnelley & Sons, Harrisonburg, Virginia.
Printed in the United States of America.

9 8 7 6 5 4 3 2 1

ISBN 0-387-90990-7 Springer-Verlag New York Berlin Heidelberg Tokyo
ISBN 3-540-90990-7 Springer-Verlag Berlin Heidelberg New York Tokyo

To Barbara Amato

Preface

A more accurate title for this book might be: An Exposition of Selected Parts of Empirical Process Theory, With Related Interesting Facts About Weak Convergence, and Applications to Mathematical Statistics. The high points are Chapters II and VII, which describe some of the developments inspired by Richard Dudley's 1978 paper. There I explain the combinatorial ideas and approximation methods that are needed to prove maximal inequalities for empirical processes indexed by classes of sets or classes of functions. The material is somewhat arbitrarily divided into results used to prove consistency theorems and results used to prove central limit theorems. This has allowed me to put the easier material in Chapter II, with the hope of enticing the casual reader to delve deeper.

Chapters III through VI deal with more classical material, as seen from a different perspective. The novelties are: convergence for measures that don't live on borel σ-fields; the joys of working with the uniform metric on $D[0, 1]$; and finite-dimensional approximation as the unifying idea behind weak convergence. Uniform tightness reappears in disguise as a condition that justifies the finite-dimensional approximation. Only later is it exploited as a method for proving the existence of limit distributions.

The last chapter has a heuristic flavor. I didn't want to confuse the martingale issues with the martingale facts.

My introduction to empirical processes came during my 1977–8 stay with Peter Gaenssler and Winfried Stute at the Ruhr University in Bochum, while I was supported by an Alexander von Humboldt Fellowship. Peter and I both spent part of 1982 at the University of Washington in Seattle, where we both gave lectures and absorbed the empirical process wisdom of Ron Pyke and Galen Shorack. The published lecture notes (Gaenssler 1984) show how closely our ideas have evolved in parallel since Bochum. I also

had the privilege of seeing a draft manuscript of a book on empirical processes by Galen Shorack and Jon Wellner.

At Yale I have been helped by a number of friends. Dan Barry read and criticized early drafts of the manuscript. Deb Nolan did the same for the later drafts, and then helped with the proofreading. First Jeanne Boyce, and then Barbara Amato, fed innumerable versions of the manuscript into the DEC-20. John Hartigan inspired me to think.

The National Science Foundation has supported my research and writing over several summers.

I am most grateful to everyone who has encouraged and aided me to get this thing finished.

Contents

Notation xiii

CHAPTER I
Functionals on Stochastic Processes 1
1. Stochastic Processes as Random Functions 1
Notes 4
Problems 4

CHAPTER II
Uniform Convergence of Empirical Measures 6
1. Uniformity and Consistency 6
2. Direct Approximation 8
3. The Combinatorial Method 13
4. Classes of Sets with Polynomial Discrimination 16
5. Classes of Functions 24
6. Rates of Convergence 30
Notes 36
Problems 38

CHAPTER III
Convergence in Distribution in Euclidean Spaces 43
1. The Definition 43
2. The Continuous Mapping Theorem 44
3. Expectations of Smooth Functions 48
4. The Central Limit Theorem 50
5. Characteristic Functions 54
6. Quantile Transformations and Almost Sure Representations 57
Notes 61
Problems 62

CHAPTER IV
Convergence in Distribution in Metric Spaces 64
1. Measurability 64
2. The Continuous Mapping Theorem 66
3. Representation by Almost Surely Convergent Sequences 71
4. Coupling 76
5. Weakly Convergent Subsequences 81
Notes 85
Problems 86

CHAPTER V
The Uniform Metric on Spaces of Cadlag Functions 89
1. Approximation of Stochastic Processes 89
2. Empirical Processes 95
3. Existence of Brownian Bridge and Brownian Motion 100
4. Processes with Independent Increments 103
5. Infinite Time Scales 107
6. Functionals of Brownian Motion and Brownian Bridge 110
Notes 117
Problems 118

CHAPTER VI
The Skorohod Metric on $D[0, \infty)$ 122
1. Properties of the Metric 122
2. Convergence in Distribution 130
Notes 136
Problems 137

CHAPTER VII
Central Limit Theorems 138
1. Stochastic Equicontinuity 138
2. Chaining 142
3. Gaussian Processes 146
4. Random Covering Numbers 149
5. Empirical Central Limit Theorems 155
6. Restricted Chaining 160
Notes 165
Problems 167

CHAPTER VIII
Martingales 170
1. A Central Limit Theorem for Martingale-Difference Arrays 170
2. Continuous Time Martingales 176
3. Estimation from Censored Data 182
Notes 185
Problems 186

APPENDIX A
Stochastic-Order Symbols 189

APPENDIX B
Exponential Inequalities 191
Notes 193
Problems 193

APPENDIX C
Measurability 195
Notes 200
Problems 200

References 201

Author Index 209

Subject Index 211

Notation

Integrals and expectations are written in linear functional notation; sets are identified with their indicator functions. Thus, instead of $\int_A f(x)\mathbb{P}(dx)$ write $\mathbb{P}(fA)$. When the variable of integration needs to be identified, as in iterated integrals, I return to the traditional notation. And orthodoxy constrains me to write $\int f(x)\,dx$ for the lebesgue integral, in whatever dimension is appropriate. If unspecified, the domain of integration is the whole space.

Abbreviations can stand for a probability measure or a random variable distributed according to that probability measure:

- Bin(n, p) = binomial distribution for n trials with success probability p.
- $N(\mu, \sigma^2)$ = normal distribution with mean μ and variance σ^2.
- $N(\mu, V)$ = multivariate normal distribution with mean vector μ and variance matrix V.
- Uniform(a, b) = uniform distribution on the open interval (a, b); square brackets, as in Uniform$[0, 1]$, indicate closed intervals.
- Poisson(λ) = poisson distribution with mean λ.

The symbol □ denotes end of proof, end of definition, and so on—something to indicate resumption of the main text. Product measures, product spaces, and product σ-fields share the product symbol $\otimes$. Maxima and minima are $\vee$ and $\wedge$. Set-theoretic difference is $\backslash$; symmetric difference is $\triangle$. If $a_n/b_n \to \infty$, for sequences $\{a_n\}$ and $\{b_n\}$, then write $a_n \gg b_n$.

Invariably $\mathbb{R}$ denotes the real line, and $\mathbb{R}^k$ denotes k-dimensional euclidean space. The borel σ-field on a metric space $\mathscr{X}$ is always $\mathscr{B}(\mathscr{X})$. The symbol $\mathbb{P}$

denotes a probability measure on a (sometimes unspecified) measurable space $(\Omega, \mathscr{E})$; miscellaneous random variables live on this space.

An $\rightsquigarrow$, a cross between $\sim$ (the sign for "is distributed according to") and an ordinary arrow $\rightarrow$ (for convergence), is used for convergence in distribution and weak convergence.

A result stated and proved in the text is always referred to with initial letters capitalized. Thus the Multivariate Central Limit Theorem is numbered III.30, but Taylor's theorem and dominated convergence are not reproved.

The letters B, U, B_P, E_P usually denote the gaussian processes: brownian motion, brownian bridge, P-motion, and P-bridge. The letters U_n and E_n denote empirical processes, with U_n generated by observations on Uniform(0, 1). Usually P_n is the empirical measure.

The set of all square-integrable functions with respect to a measure μ is written $\mathscr{L}^2(\mu)$; the corresponding space of equivalence classes is $L_2(\mu)$. A similar distinction holds for $\mathscr{L}^1(\mu)$ and $L^1(\mu)$. Often ρ_P denotes the $\mathscr{L}^2(P)$ seminorm; $\|\cdot\|$ is the supremum norm on a space of functions. The symbols π_t, π_S, and so on, are usually projection maps on function spaces.

Expressions like $N(\delta)$, $N_1(\delta, d, T)$, and $N_2(\delta, P, \mathscr{F})$ represent various covering numbers; $J(\delta)$, $J_1(\delta, d, T)$, and $J_2(\delta, P, \mathscr{F})$ are the corresponding covering integrals.

CHAPTER I

Functionals on Stochastic Processes

... which introduces the idea of studying random variables determined by the whole sample path of a stochastic process.

I.1. Stochastic Processes as Random Functions

Functions analyzed as points of abstract spaces of functions appear in many branches of mathematics. Geometric intuitions about distance (or approximation, or convergence, or orthogonality, or any other ideas learned from the study of euclidean space) carry over to those abstract spaces, lending familiarity to operations carried out on the functions. We enjoy similar benefits in the study of stochastic processes if we analyze them as random elements of spaces of functions.

Remember that a stochastic process is a collection $\{X_t: t \in T\}$ of real random variables, all defined on a common probability space $(\Omega, \mathscr{E}, \mathbb{P})$. Often T will be an interval of the real line (which makes the temptation to think of the process as evolving in time almost irresistible), but we shall also meet up with fancier index sets—subsets of higher-dimensional euclidean spaces, and collections of functions. The random variable X_t depends on both t and the point ω in Ω at which it is evaluated. To emphasize its role as a function of two variables, write it as $X(\omega, t)$. For fixed t, the function $X(\cdot, t)$ is, by assumption, a measurable map from Ω into $\mathbb{R}$. For fixed ω, the function $X(\omega, \cdot)$ is called a sample path of the stochastic process. If all the sample paths lie in some fixed collection $\mathscr{X}$ of real-valued functions on T, the process X can be thought of as a map from Ω into $\mathscr{X}$, a random element of $\mathscr{X}$. For example, if a process indexed by $[0, 1]$ has continuous sample paths it defines a random element of the space $C[0, 1]$ of all continuous, real-valued functions on $[0, 1]$. (In Chapter IV we shall formalize the definition by adding a measurability requirement.)

Each sample path of X is a single point in $\mathscr{X}$. Each random variable Z for which $Z(\omega)$ depends only on the sample path $X(\omega, \cdot)$, such as the maximum of $X(\omega, t)$ over all t, can be expressed by means of a functional on $\mathscr{X}$. That is, the value $Z(\omega)$ is found by applying to $X(\omega, \cdot)$ a map H that sends functions in $\mathscr{X}$ onto points of the real line. The name functional serves to distinguish functions on spaces of functions from functions on other abstract sets, an outmoded distinction, but one that can help us to remember where H lives.

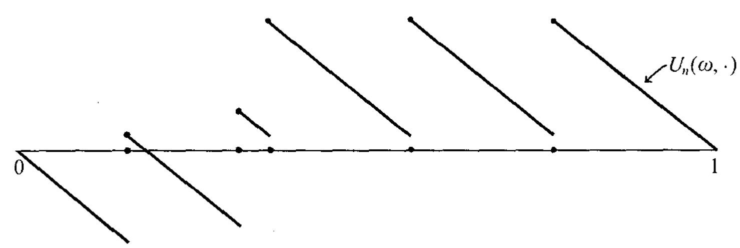
By breaking Z into the composition of a functional with a map from Ω into $\mathscr{X}$ we also break any analysis of Z into two parts: calculations involving only the random element X; and calculations involving only the functional H. This allows us to study many different Z's simultaneously, just by varying the choice of H. Of course we only gain by this if most of the hard work can be disposed of once and for all in the analysis of X.

The idea can be taken further. Suppose that a second stochastic process $\{Y_t: t \in T\}$ puts all its sample paths in the same function space $\mathscr{X}$. Suppose we want to study the same functional H of both processes; we want to show that HX and HY have distributions that are close, perhaps. Break the problem into its two parts: show that the distributions of X and Y (the probability measures they induce on $\mathscr{X}$) are close; then show that H has a continuity property ensuring that closeness of the distributions of X and Y implies closeness of the distributions of HX and HY. Such an approach would make the analysis easier for other functionals with the same sort of continuity property; for a different H only the second part of the analysis would need repeating.

1 Example. Goodness-of-fit test statistics can often be expressed as functionals on a suitably standardized empirical distribution function. Consider the basic case of an independent sample $\xi_1, \ldots, \xi_n$ from the Uniform(0, 1) distribution. Define the uniform empirical process U_n by

$$U_n(\omega, t) = n^{-1/2} \sum_{i=1}^{n} (\{\xi_i(\omega) \leq t\} - t) \quad \text{for} \quad 0 \leq t \leq 1.$$

This has the standardized binomial distribution, $n^{-1/2}(\text{Bin}(n, t) - nt)$, for fixed t. Each sample path is piecewise linear with jump discontinuities at the n points $\xi_1(\omega), \ldots, \xi_n(\omega)$.



The process U_n defines a random element of any function space $\mathscr{X}$ that contains all sample paths of this form. We could take $\mathscr{X}$ as the smallest such set of functions, but there is some advantage to choosing a space that can accommodate the sample paths of other important stochastic processes.

That way the analysis for this particular problem will carry over to many other limit and approximation problems. The usual choice is the space $D[0, 1]$ of all real-valued functions on $[0, 1]$ that are right continuous at each point of $[0, 1)$ with left limits existing at each point of $(0, 1]$. The catchy French acronym *cadlag* (continue à droite, limites à gauche) offers a most convenient way to avoid repeating this mouthful of a description. From now on it's cadlag. The processes U_n define random elements of the space $D[0, 1]$ of all (real-valued) cadlag functions on $[0, 1]$.

Numerous statistics have been suggested as candidates for testing whether the observations really do come from the uniform distribution, amongst them:

$$(H_1 U_n)(\omega) = \sup_t U_n(\omega, t),$$

$$(H_2 U_n)(\omega) = \sup_t |U_n(\omega, t)|,$$

$$(H_3 U_n)(\omega) = \int_0^1 U_n(\omega, t)^2 \, dt.$$

How do these statistics behave as n tends to infinity? We can try to answer the question by first determining how $\{U_n\}$, as a sequence of stochastic processes indexed by $[0, 1]$, behaves as n tends to infinity.

Suppress the argument ω. At each fixed t the sequence of real random variables $\{U_n(t)\}$ converges in distribution to a $N(0, t - t^2)$ distribution, by virtue of the Central Limit Theorem; at each fixed pair (s, t) the sequence of bivariate random vectors $\{(U_n(s), U_n(t))\}$ converges to a bivariate normal distribution with covariance the same as the covariance between $U_n(s)$ and $U_n(t)$, by virtue of the bivariate form of the Central Limit Theorem; and similarly for the higher finite-dimensional distributions. That is, the joint distributions of the random variables obtained by sampling U_n at any fixed, finite set of index points converge to the corresponding distributions of the so-called brownian bridge (also known as the tied-down brownian motion), the gaussian process $U(t)$ characterized by:

(i) U has continuous sample paths (it is a random element of $C[0, 1]$) between the two fixed points $U(0) = U(1) = 0$;
(ii) For fixed $t_1, \ldots, t_k$ the random vector $(U(t_1), \ldots, U(t_k))$ has a multivariate normal distribution with zero means and $\text{cov}(U(t_i), U(t_j)) = \min(t_i, t_j) - t_i t_j$.

The process U need not live on the same probability space as the empirical processes. It must, however, have sample paths in $C[0, 1]$. The reason we need this apparently irrelevant property will emerge in Chapter V, when we make the argument rigorous.

When n is large, the process U_n behaves something like U. On this basis, Doob (1949) made the bold heuristic assumption that the distributions of statistics calculated from U_n should be close to the distributions of the corresponding statistics for U; he conjectured that functionals of U_n should converge in distribution to the analogous functionals of U. In his words (he wrote x_n and x instead of U_n and U):

> We shall assume, until a contradiction frustrates our devotion to heuristic reasoning, that *in calculating asymptotic $x_n(t)$ process distributions when $n \to \infty$ we may simply replace the $x_n(t)$ processes by the $x(t)$ process*. It is clear that this cannot be done in all possible situations, but let the reader who has never used this sort of reasoning exhibit the first counter example.

Happily, the heuristic does give the right answer for each of the three functionals mentioned above. □

This example will guide us in our definition of convergence in distribution for stochastic processes treated as random elements of function spaces. We shall explore one possible meaning for closeness of two processes in a distributional sense. The empirical process application will point us towards a theory slightly different from the one normally espoused in the weak convergence literature.

Our progress towards the theory will begin in Chapter III with another look at the classical results for convergence of real random variables and random vectors, partly as a refresher course in the basic techniques, and partly as a way to get a running start on the theory for random elements of general metric spaces in Chapter IV.

The momentum of the development will carry us through to Chapter V, where Doob's heuristic argument will receive its rigorous justification, as a weak convergence result for $D[0, 1]$. Chapter VII will extend the theory to empirical processes with index sets more complicated than $[0, 1]$, building upon methods that will be introduced and developed in Chapter II.

Notes

Stochastic processes have frequently been treated as random functions in the probability literature. Doob (1953) pioneered. Gihman and Skorohod (1974, Chapter III) have collected together criteria for existence of stochastic processes with sample paths in $C[0, 1]$ and $D[0, 1]$. Breiman (1968, Chapters 12 to 14) is a good place to start.

Problems

[1] Verify that U has the same covariance function as U_n, by writing $(U_n(s), U_n(t))$ as a normalized sum of the independent random vectors $(\{\xi_i \leq s\}, \{\xi_i \leq t\})$.

[2] Show that the random variables $H_3 U_n$ and $H_3 U$ both have expectation $\frac{1}{6}$. [Remember that U_n and U have the same first and second moment structures.]

[3] Show that $H_3 U_n$ can be written in terms of the order statistics $\xi_{(i)}$ as

$$\tfrac{1}{6}(n+1)^{-1} + \sum_{i=1}^{n} (\xi_{(i)} - \mathbb{P}\xi_{(i)})^2 + \sum_{i=1}^{n} (\xi_{(i)} - \mathbb{P}\xi_{(i)}) \frac{n - 2i + 1}{n(n+1)}.$$

[Complicated functions of the observations can sometimes be represented by simple functionals of a suitably constructed random process.]

CHAPTER II
Uniform Convergence of Empirical Measures

... in which uniform analogues of the strong law of large numbers are proved by two methods. These generalize the classical Glivenko–Cantelli theorem, which concerns only empirical measures indexed by intervals on the real line, to uniform convergence over classes of sets and uniform convergence over classes of functions. The results are applied to prove consistency for statistics expressible as continuous functionals of the empirical measure. A refinement of the second method gives rates of convergence.

II.1. Uniformity and Consistency

For independent sampling from a distribution function F, the strong law of large numbers tells us that the proportion of points in an interval $(-\infty, t]$ converges almost surely to $F(t)$. The classical Glivenko–Cantelli theorem strengthens the result by adding that the convergence holds uniformly over all t. The strong law also tells us that the proportion of points in any fixed set converges almost surely to the probability of that set. The strengthening of this result, to give uniform convergence over classes of sets more interesting than intervals on the real line, and its further generalization to classes of functions, will be the main concern of this chapter.

For the most part we shall consider only independent sampling from a fixed distribution P on a set S. The probability measure P_n that puts equal mass at each of the n observations $\xi_1, \ldots, \xi_n$ will be called the empirical measure. It captures everything we might need to know about the observations, except for the order in which they were taken. Averages over the observations can be written as expectations with respect to P_n:

$$n^{-1} \sum_{i=1}^{n} f(\xi_i) = P_n f.$$

If $P|f| < \infty$, the average converges almost surely to its expected value, Pf. We shall be finding conditions under which the convergence is uniform over a class $\mathscr{F}$ of functions.

Of course we should not expect uniform convergence over all classes of functions, except in trivial cases. Unless P is a discrete distribution, the difference $P_n D - PD$ cannot even converge to zero uniformly over all sets;

there always exists a countable set with P_n measure one. But there are non-trivial classes over which the convergence is uniform. When we have such a class $\mathscr{F}$ we can deduce consistency results for statistics that depend on the observations only through the values $P_n f$, for f in $\mathscr{F}$.

1 Example. The median of a distribution P on the real line can be defined as the smallest value of m for which $P(-\infty, m] \geq \frac{1}{2}$. If $P(-\infty, t] > \frac{1}{2}$ for each $t > m$ then the median is a continuous functional, in the sense that

$$|\text{median}(Q) - \text{median}(P)| \leq \varepsilon$$

whenever the distribution Q is close enough to P. Close means

$$\sup_t |Q(-\infty, t] - P(-\infty, t]| < \delta,$$

where the tiny δ is chosen so that

$$P(-\infty, m - \varepsilon] < \tfrac{1}{2} - \delta,$$
$$P(-\infty, m + \varepsilon] > \tfrac{1}{2} + \delta.$$

The argument goes: if Q has median m' then

$$P(-\infty, m'] > Q(-\infty, m'] - \delta \geq \tfrac{1}{2} - \delta,$$

so certainly $m' > m - \varepsilon$. Similarly, for every $m'' < m'$,

$$P(-\infty, m''] < Q(-\infty, m''] + \delta < \tfrac{1}{2} + \delta,$$

which implies $m'' < m + \varepsilon$, and hence $m' \leq m + \varepsilon$.

Next comes the probability theory. If the empirical measure P_n is constructed from a sample of independent observations on P, the Glivenko–Cantelli theorem tells us that

$$\sup_t |P_n(-\infty, t] - P(-\infty, t]| \to 0 \quad \text{almost surely.}$$

From this we deduce that, almost surely,

$$|\text{median}(P_n) - \text{median}(P)| \leq \varepsilon \quad \text{eventually.}$$

The sample median is strongly consistent as an estimator of the population median. □

For this example we didn't have to prove the uniformity result; the Glivenko–Cantelli theorem is the oldest and best-known uniform strong law of large numbers in the literature. But as we encounter new functions (usually called functionals) of the empirical measure, new uniform convergence theorems will be demanded. We shall be exploring two methods for proving these theorems.

The first method is simpler in concept, but harder in execution. It involves direct approximation of functions in an infinite class $\mathscr{F}$ by a finite collection of functions. Classical convergence results, such as the strong law of large numbers or the ergodic theorem, ensure uniform convergence for the finite collection; the form of approximation carries the uniformity over to $\mathscr{F}$. Section 2 deals with direct approximation.

The second method depends heavily upon symmetry properties implied by independence. It uses simple combinatorial arguments to identify classes satisfying uniform strong laws of large numbers under independent sampling. Sections 3 to 5 assemble the ideas behind this method.

II.2. Direct Approximation

Throughout the section $\mathscr{F}$ will be a class of (measurable) functions on a set S with a σ-field that carries a probability measure P. The empirical measure P_n is constructed by sampling from P. Assume $P|f| < \infty$ for each f in $\mathscr{F}$. If $\mathscr{F}$ were finite, the convergence of $P_n f$ to Pf assured by the strong law of large numbers would, for trivial reasons, be uniform in f. If $\mathscr{F}$ can be approximated by a finite class (not necessarily a subclass of $\mathscr{F}$) in such a way that the errors of approximation are uniformly small, the uniformity carries over to $\mathscr{F}$. The direct method achieves this by requiring that each member of $\mathscr{F}$ be sandwiched between a pair of approximating functions taken from the finite class.

2 Theorem. *Suppose that for each $\varepsilon > 0$ there exists a finite class $\mathscr{F}_\varepsilon$ containing lower and upper approximations to each f in $\mathscr{F}$, for which*

$$f_{\varepsilon,L} \le f \le f_{\varepsilon,U} \quad \textit{and} \quad P(f_{\varepsilon,U} - f_{\varepsilon,L}) < \varepsilon.$$

Then $\sup_{\mathscr{F}} |P_n f - Pf| \to 0$ *almost surely.*

Proof. Break the asserted convergence into a pair of one-sided results:

$$\liminf \inf_{\mathscr{F}} (P_n f - Pf) \ge 0$$

and

$$\limsup \sup_{\mathscr{F}} (P_n f - Pf) \le 0$$

or, equivalently,

$$\liminf \inf_{\mathscr{F}} (P_n(-f) - P(-f)) \ge 0.$$

Then two applications of the next theorem will complete the proof. □

3 Theorem. *Suppose that for each $\varepsilon > 0$ there exists a finite class $\mathscr{F}_\varepsilon$ of functions for which: to each f in $\mathscr{F}$ there exists an f_ε in $\mathscr{F}_\varepsilon$ such that $f_\varepsilon \le f$ and $Pf_\varepsilon \ge Pf - \varepsilon$. Then*

$$\liminf \inf_{\mathscr{F}} (P_n f - Pf) \ge 0 \quad \textit{almost surely.}$$

PROOF. For each $\varepsilon > 0$,

$$\begin{aligned}\liminf \inf_{\mathscr{F}} (P_n f - Pf) &\geq \liminf \inf_{\mathscr{F}} (P_n f_\varepsilon - Pf) \quad \text{because } f_\varepsilon \leq f \\ &\geq \liminf \inf_{\mathscr{F}} (P_n f_\varepsilon - Pf_\varepsilon) + \inf_{\mathscr{F}} (Pf_\varepsilon - Pf) \\ &\geq 0 + -\varepsilon \quad \text{almost surely, as } \mathscr{F}_\varepsilon \text{ is finite.}\end{aligned}$$

Throw away an aberrant null set for each positive rational ε to arrive at the asserted result. □

You might have noticed that independence enters only as a way of guaranteeing the almost sure convergence of $P_n f_\varepsilon$ to Pf_ε for each approximating f_ε. Weaker assumptions, such as stationarity and ergodicity, could substitute for independence.

4 Example. The method of k-means belongs to the host of ad hoc procedures that have been suggested as ways of partitioning multivariate data into groups somehow indicative of clusters in the underlying population. We can prove a consistency theorem for the procedure by application of the one-sided uniformity result of Theorem 3.

For purposes of illustration, consider only the simple case where observations $\xi_1, \ldots, \xi_n$ from a distribution P on the real line are to be partitioned into two groups. The method prescribes that the two groups be chosen to minimize the within-groups sum of squares. Equivalently, we may choose optimal centers a_n and b_n to minimize

$$\sum_{i=1}^{n} |\xi_i - a|^2 \wedge |\xi_i - b|^2,$$

then allocate each ξ_i to its nearest center. The optimal centers must lie at the mean of those observations drawn into their clusters, hence the name k-means (or 2-means, in the present case). In terms of the empirical measure P_n, the method seeks to minimize

$$W(a, b, P_n) = P_n f_{a,b},$$

where

$$f_{a,b}(x) = |x - a|^2 \wedge |x - b|^2.$$

As the sample size increases, $W(a, b, P_n)$ converges almost surely to

$$W(a, b, P) = Pf_{a,b}$$

for each fixed (a, b). This suggests that (a_n, b_n), which minimizes $W(\cdot, \cdot, P_n)$, might converge to the (a^*, b^*) that minimizes $W(\cdot, \cdot, P)$. Given a few obvious conditions, that is indeed what happens.

To ensure finiteness of $W(\cdot, \cdot, P)$, assume that $P|x|^2 < \infty$. Assume also that there exists a unique (a^*, b^*) minimizing W. Adopt the convention that

$a \le b$ in order that (b^*, a^*) be ruled out as a distinct minimizing pair. Without uniqueness the consistency statement needs a slight reinterpretation (Problem 1).

The continuity argument lurking behind the consistency theorem does depend on one-sided uniform convergence of $W(\cdot, \cdot, P_n)$ to $W(\cdot, \cdot, P)$, but not uniformly over all possible choices for the centers. We must first force (a_n, b_n) into a region

$$C = [-M, M] \otimes \mathbb{R} \quad \cup \quad \mathbb{R} \otimes [-M, M]$$

for some suitably large M, then prove

$$\liminf \inf_C (P_n f_{a,b} - P f_{a,b}) \ge 0 \quad \text{almost surely.}$$

We need at least one of the centers within a bounded region $[-M, M]$ to get the uniformity. Determine how large M needs to be by invoking optimality of (a_n, b_n).

$$\begin{aligned} W(a_n, b_n, P_n) &\le W(0, 0, P_n) \\ &\to W(0, 0, P) \quad \text{almost surely} \\ &= P|x|^2. \end{aligned}$$

If both a_n and b_n lay outside $[-M, M]$ then

$$\begin{aligned} W(a_n, b_n, P_n) &\ge (\tfrac{1}{2}M)^2 P_n[-\tfrac{1}{2}M, \tfrac{1}{2}M] \\ &\to (\tfrac{1}{2}M)^2 P[-\tfrac{1}{2}M, \tfrac{1}{2}M] \quad \text{almost surely.} \end{aligned}$$

If we choose M so that $P|x|^2 < (\frac{1}{2}M)^2 P[-\frac{1}{2}M, \frac{1}{2}M]$ then there must eventually be at least one of the optimal centers within $[-M, M]$, almost surely. We shall later also need M so large that (a^*, b^*) belongs to C.

Explicit construction of the finite approximating class demanded by Theorem 3 is straightforward, but a trifle messy. That is one of the disadvantages of brute-force methods. First note that

$$f_{a,b}(x) \le (x - M)^2 + (x + M)^2 \quad \text{for } (a, b) \text{ in } C.$$

Write $F(x)$ for the upper bound. Because $PF < \infty$, there exists a constant D, larger than M, for which $PF[-D, D]^c < \varepsilon$. We have only to worry about the approximation to $f_{a,b}$ on $[-D, D]$.

We may assume that both a and b lie in the interval $[-3D, 3D]$. For if, say, $|b| > 3D$ then

$$f_{a,b}(x)\{|x| \le D\} = |x - a|^2 = f_{a,a}(x)\{|x| \le D\}$$

because $|a| \le M$; the lower approximation for $f_{a,a}$ on $[-D, D]$ will also serve for $f_{a,b}$.

Let C_ε be a finite subset of $[-3D, 3D]^2$ such that each (a, b) in that square has an (a', b') with $|a - a'| < \varepsilon/D$ and $|b - b'| < \varepsilon/D$. Then for each x in $[-D, D]$,

$$\begin{aligned} |f_{a,b}(x) - f_{a',b'}(x)| &\le |(x-a)^2 - (x-a')^2| + |(x-b)^2 - (x-b')^2| \\ &\le 2|a-a'|\,|x - \tfrac{1}{2}(a+a')| + 2|b-b'|\,|x - \tfrac{1}{2}(b+b')| \\ &\le 2(\varepsilon/D)(D+3D) + 2(\varepsilon/D)(D+3D) \\ &= 16\varepsilon. \end{aligned}$$

The class $\mathscr{F}_{33\varepsilon}$ consists of all functions $(f_{a',b'}(x) - 16\varepsilon)\{|x| \le D\}$ for (a', b') ranging over C_ε.

From Theorem 3,

$$\liminf \inf_C (P_n f_{a,b} - P f_{a,b}) \ge 0.$$

Eventually the optimal centers (a_n, b_n) lie in C. Thus

$$\liminf (W(a_n, b_n, P_n) - W(a_n, b_n, P)) \ge 0 \quad \text{almost surely.}$$

Since

$$\begin{aligned} W(a_n, b_n, P_n) &\le W(a^*, b^*, P_n) && \text{because } (a_n, b_n) \text{ is optimal for } P_n \\ &\to W(a^*, b^*, P) && \text{almost surely} \\ &\le W(a_n, b_n, P) && \text{because } (a^*, b^*) \text{ is optimal for } P, \end{aligned}$$

we then deduce that

$$W(a_n, b_n, P) \to W(a^*, b^*, P) \quad \text{almost surely.}$$

Notice what happened. The uniformity allowed us to transfer optimality of (a_n, b_n) for P_n to a sort of asymptotic optimality for P; the processes $W(\cdot, \cdot, P_n)$ have disappeared, leaving everything in terms of the fixed, non-random function $W(\cdot, \cdot, P)$.

We have assumed that $W(\cdot, \cdot, P)$ achieves its unique minimum at (a^*, b^*). Complete the argument by strengthening this to: for each neighborhood U of (a^*, b^*),

$$\inf_{C\setminus U} W(a, b, P) > W(a^*, b^*, P).$$

Continuity of $W(\cdot, \cdot, P)$ takes care of the infimum over bounded regions of $C\setminus U$. If there were an unbounded sequence (α_i, β_i) in C with

$$W(\alpha_i, \beta_i, P) \to W(a^*, b^*, P),$$

we could extract a subsequence along which, say, $\alpha_i \to -\infty$ and $\beta_i \to \beta$, with $|\beta| \le M$. Dominated convergence would give

$$W(a^*, b^*, P) = P|x - \beta|^2,$$

which would contradict uniqueness of (a^*, b^*): for every a, the pair (a, β) would minimize $W(\cdot, \cdot, P)$. The pair (a_n, b_n), by seeking out the unique minimum of $W(\cdot, \cdot, P)$ over the region C, must converge to (a^*, b^*). □

The k-means example typifies consistency proofs for estimators defined by optimization of a random criterion function. By ad hoc arguments one forces the optimal solution into a restricted, often compact, region. That is usually the hardest part of the proof. (Problem 2 describes one particularly nice ad hoc argument.) Then one appeals to a uniform strong law over the restricted region, to replace the random criterion function by a deterministic limit function. Global properties of the limit function force the optimal solution into desired neighborhoods. If one wants consistency results that apply not just to independent sequences but also, for example, to stationary ergodic sequences, one is stuck with cumbersome direct approximation arguments; but for independent sampling, slicker methods are available for proving the uniform strong laws. We shall return to the k-means problem in Section 5 (Example 29 to be precise) after we have developed these methods.

5 Example. Let θ be the parameter of a stationary autoregressive process

$$y_{n+1} = \theta y_n + u_n$$

for independent, identically distributed innovations $\{u_n\}$. Stationarity requires $|\theta| \le 1$. A generalized M-estimator for θ is any value θ_n for which the random function

$$H_n(\theta) = (n-1)^{-1} \sum_{i=1}^{n-1} g(y_i)\phi(y_{i+1} - \theta y_i)$$

takes the value zero. We would hope that θ_n converges to the θ^* at which the deterministic function

$$H(\theta) = \mathbb{P} g(y_1)\phi(y_2 - \theta y_1)$$

takes the value zero. If $|g| \le 1$ and $|\phi| \le 1$ and ϕ is continuous, we can go part of the way towards proving this by means of a uniform strong law for a bivariate empirical measure.

Write Q_n for the probability measure that puts equal mass $(n-1)^{-1}$ on each of the pairs $(y_1, y_2), \ldots, (y_{n-1}, y_n)$. For fixed (integrable) $f(\cdot, \cdot)$,

$$Q_n f \to Qf \quad \text{almost surely,}$$

where Q denotes the joint distribution of (y_1, y_2). This follows from the ergodic theorem for the stationary bivariate process $\{(y_n, y_{n+1})\}$.

Check the approximation conditions of Theorem 2, with Q in place of P, for the class of functions

$$f(x_1, x_2, \theta) = g(x_1)\phi(x_2 - \theta x_1) \quad \text{for} \quad -1 \le \theta \le 1.$$

First, choose an integer K so large that

$$\mathbb{P}\{|y_1| \le K, |y_2| \le K\} > 1 - \varepsilon.$$

Then appeal to uniform continuity of ϕ on the compact interval $[-2K, 2K]$ to find a $\delta > 0$ such that $|\phi(a) - \phi(b)| < \varepsilon$ whenever $|a - b| \leq \delta$ and $|a| \leq 2K$ and $|b| \leq 2K$. For θ in the interval $[k\delta/K, (k+1)\delta/K]$,

$$|f(x_1, x_2, \theta) - f(x_1, x_2, k\delta/K)| \leq \varepsilon + 2\{|x_1| > K\} + 2\{|x_2| > K\}.$$

With the integer k running over the finite range needed for these intervals to cover $[-1, 1]$, the functions

$$f(x_1, x_2, k\delta/K) \pm \varepsilon \pm 2\{|x_1| > K\} \pm 2\{|x_2| > K\}$$

provide the upper and lower approximations required by Theorem 2.

As noted following Theorem 3, the uniform strong laws also apply to empirical measures constructed from stationary ergodic sequences. Accordingly,

$$\sup_{|\theta| \leq 1} |Q_n f(\cdot, \cdot, \theta) - Qf(\cdot, \cdot, \theta)| \to 0 \quad \text{almost surely,} \tag{6}$$

that is,

$$\sup_{|\theta| \leq 1} |H_n(\theta) - H(\theta)| \to 0 \quad \text{almost surely.}$$

Provided θ_n lies in the range $[-1, 1]$, we can deduce from (6) that $H(\theta_n) \to 0$ almost surely. It would be a sore embarrassment if the estimate of the autoregressive parameter were not in this range. Usually one avoids the embarrassment by insisting only that $H_n(\theta_n) \to 0$, with θ_n in $[-1, 1]$. Such a θ_n always exists because $H_n(\theta^*) \to 0$ almost surely.

Convergence results for θ_n depend upon the form of $H(\cdot)$. We know θ_n gets forced eventually into the set $\{|H| < \varepsilon\}$ for each $\varepsilon > 0$. If this set shrinks to θ^* as $\varepsilon \downarrow 0$ then θ_n must converge to θ^*, which necessarily would have to be the unique zero of $H(\cdot)$. If we assume that H does have these properties we get the consistency result for the generalized M-estimator. □

II.3. The Combinatorial Method

Since understanding of general methods grows from insights into simple special cases, let us begin with the best-known example of a uniform strong law of large numbers, the classical Glivenko–Cantelli theorem. This asserts that, for every distribution P on the real line,

$$\sup_t |P_n(-\infty, t] - P(-\infty, t]| \to 0 \quad \text{almost surely,} \tag{7}$$

when the empirical measure P_n comes from independent sampling on P. The ideas that will emerge from the treatment of this special case will later be expanded into methods applicable to other classes of functions. To facilitate back reference, break the proof into five steps.

Keep the notation tidy by writing $\|\cdot\|$ to denote the supremum over the class $\mathscr{I}$ of intervals $(-\infty, t]$, for $-\infty < t < \infty$. We could restrict the supremum to rational t to ensure measurability.

FIRST SYMMETRIZATION.

Instead of matching P_n against its parent distribution P, look at the difference between P_n and an independent copy, P'_n say, of itself. The difference $P_n - P'_n$ is determined by a set of $2n$ points (albeit random) on the real line; it can be attacked by combinatorial methods, which lead to a bound on deviation probabilities for $\|P_n - P'_n\|$. A symmetrization inequality converts this into a bound on $\|P_n - P\|$ deviations.

8 Symmetrization Lemma. *Let* $\{Z(t): t \in T\}$ *and* $\{Z'(t): t \in T\}$ *be independent stochastic processes sharing an index set* T. *Suppose there exist constants* $\beta > 0$ *and* $\alpha > 0$ *such that* $\mathbb{P}\{|Z'(t)| \le \alpha\} \ge \beta$ *for every* t *in* T. *Then*

$$\mathbb{P}\left\{\sup_t |Z(t)| > \varepsilon\right\} \le \beta^{-1}\mathbb{P}\left\{\sup_t |Z(t) - Z'(t)| > \varepsilon - \alpha\right\}. \tag{9}$$

PROOF. Select a random τ for which $|Z(\tau)| > \varepsilon$ on the set $\{\sup |Z(t)| > \varepsilon\}$. Since τ is determined by Z, it is independent of Z'. It behaves like a fixed index value when we condition on Z:

$$\mathbb{P}\{|Z'(\tau)| \le \alpha \,|\, Z\} \ge \beta.$$

Integrate out.

$$\begin{aligned}\beta\mathbb{P}\left\{\sup_t |Z(t)| > \varepsilon\right\} &\le \mathbb{P}\{|Z'(\tau)| \le \alpha, |Z(\tau)| > \varepsilon\}\\ &\le \mathbb{P}\{|Z(\tau) - Z'(\tau)| > \varepsilon - \alpha\}\\ &\le \mathbb{P}\left\{\sup_t |Z(t) - Z'(t)| > \varepsilon - \alpha\right\}.\end{aligned}$$ □

Close inspection of the proof would reveal a disregard for a number of measure-theoretic niceties. A more careful treatment may be found in Appendix C. For our present purpose it would suffice if we assumed T countable; the proof is impeccable for stochastic processes sharing a countable index set. We could replace suprema over all intervals $(-\infty, t]$ by suprema over intervals with a rational endpoint.

For fixed t, $P_n(-\infty, t]$ is an average of the n independent random variables $\{\xi_i \le t\}$, each having expected value $P(-\infty, t]$ and variance $P(-\infty, t] - (P(-\infty, t])^2$, which is less than one. By Tchebychev's inequality,

$$\mathbb{P}\{|P'_n(-\infty, t] - P(-\infty, t]| \le \tfrac{1}{2}\varepsilon\} \ge \tfrac{1}{2} \quad \text{if} \quad n \ge 8\varepsilon^{-2}.$$

Apply the Symmetrization Lemma with $Z = P_n - P$ and $Z' = P'_n - P$, the class $\mathscr{I}$ as index set, $\alpha = \frac{1}{2}\varepsilon$, and $\beta = \frac{1}{2}$.

(10) $$\mathbb{P}\{\|P_n - P\| > \varepsilon\} \leq 2\mathbb{P}\{\|P_n - P'_n\| > \tfrac{1}{2}\varepsilon\} \quad \text{if} \quad n \geq 8\varepsilon^{-2}.$$

Second Symmetrization.

The difference $P_n - P'_n$ depends on $2n$ observations. The double sample size creates a minor nuisance, at least notationally. It can be avoided by a second symmetrization trick, at the cost of a further diminution of the ε. Independently of the observations $\xi_1, \ldots, \xi_n, \xi'_1, \ldots, \xi'_n$ from which the empirical measures are constructed, generate independent sign random variables $\sigma_1, \ldots, \sigma_n$ for which $\mathbb{P}\{\sigma_i = +1\} = \mathbb{P}\{\sigma_i = -1\} = \frac{1}{2}$. The symmetric random variables $\{\xi_i \leq t\} - \{\xi'_i \leq t\}$, for $i = 1, \ldots, n$ and $-\infty < t < \infty$, have the same joint distribution as the random variables $\sigma_i[\{\xi_i \leq t\} - \{\xi'_i \leq t\}]$. (Consider the conditional distribution given $\{\sigma_i\}$.) Thus

$$\begin{aligned}\mathbb{P}\{\|P_n - P'_n\| > \tfrac{1}{2}\varepsilon\} &= \mathbb{P}\left\{\sup_t \left|n^{-1}\sum_{i=1}^n \sigma_i[\{\xi_i \leq t\} - \{\xi'_i \leq t\}]\right| > \tfrac{1}{2}\varepsilon\right\} \\ &\leq \mathbb{P}\left\{\sup_t \left|n^{-1}\sum_{i=1}^n \sigma_i\{\xi_i \leq t\}\right| > \tfrac{1}{4}\varepsilon\right\} \\ &\quad + \mathbb{P}\left\{\sup_t \left|n^{-1}\sum_{i=1}^n \sigma_i\{\xi'_i \leq t\}\right| > \tfrac{1}{4}\varepsilon\right\}.\end{aligned}$$

Write P_n° for the signed measure that places mass $n^{-1}\sigma_i$ at ξ_i. The two symmetrizations give, for $n \geq 8\varepsilon^{-2}$,

(11) $$\mathbb{P}\{\|P_n - P\| > \varepsilon\} \leq 4\mathbb{P}\{\|P_n^\circ\| > \tfrac{1}{4}\varepsilon\}.$$

To bound the right-hand side, work conditionally on the vector of observations $\boldsymbol{\xi}$, leaving only the randomness contributed by the sign variables.

Maximal Inequality.

Once the locations of the $\boldsymbol{\xi}$ observations are fixed, the supremum $\|P_n^\circ\|$ reduces to a maximum taken over a strategically chosen set of intervals $I_j = (-\infty, t_j]$, for $j = 0, 1, \ldots, n$. Of course the choice of these intervals depends on $\boldsymbol{\xi}$; we need one t_j between each pair of adjacent observations. (The t_0 and t_n are not really necessary.) With the number of intervals reduced so drastically, we can afford a crude bound for the supremum.

(12) $$\begin{aligned}\mathbb{P}\{\|P_n^\circ\| > \tfrac{1}{4}\varepsilon \mid \boldsymbol{\xi}\} &\leq \sum_{j=0}^n \mathbb{P}\{|P_n^\circ I_j| > \tfrac{1}{4}\varepsilon \mid \boldsymbol{\xi}\} \\ &\leq (n+1)\max_j \mathbb{P}\{|P_n^\circ I_j| > \tfrac{1}{4}\varepsilon \mid \boldsymbol{\xi}\}.\end{aligned}$$

This bound will be adequate for the present because the conditional probabilities decrease exponentially fast with n, thanks to an inequality of Hoeffding for sums of independent, bounded random variables.

EXPONENTIAL BOUNDS.

Let $Y_1, \ldots, Y_n$ be independent random variables, each with zero mean and bounded range: $a_i \le Y_i \le b_i$. For each $\eta > 0$, Hoeffding's Inequality (Appendix B) asserts

$$\mathbb{P}\{|Y_1 + \cdots + Y_n| \ge \eta\} \le 2 \exp\left[-2\eta^2 \Big/ \sum_{i=1}^{n} (b_i - a_i)^2\right].$$

Apply the inequality with $Y_i = \sigma_i\{\xi_i \le t\}$. Given $\boldsymbol{\xi}$, the random variable Y_i takes only two values, $\pm\{\xi_i \le t\}$, each with probability $\frac{1}{2}$.

$$\begin{aligned}\mathbb{P}\{|P_n^\circ(-\infty, t]| \ge \tfrac{1}{4}\varepsilon \,|\, \boldsymbol{\xi}\} &\le 2 \exp\left[-2(n\varepsilon/4)^2 \Big/ \sum_{i=1}^{n} 4\{\xi_i \le t\}\right] \\ &\le 2\exp(-n\varepsilon^2/32),\end{aligned}$$

because the indicator functions sum to at most n. Use this for each I_j in inequality (12).

$$\mathbb{P}\{\|P_n^\circ\| > \tfrac{1}{4}\varepsilon \,|\, \boldsymbol{\xi}\} \le 2(n+1)\exp(-n\varepsilon^2/32).$$

Notice that the right-hand side now does not depend on $\boldsymbol{\xi}$.

INTEGRATION.

Take expectations over $\boldsymbol{\xi}$.

$$\mathbb{P}\{\|P_n - P\| > \varepsilon\} \le 8(n+1)\exp(-n\varepsilon^2/32).$$

This gives very fast convergence in probability, so fast that

$$\sum_{n=1}^{\infty} \mathbb{P}\{\|P_n - P\| > \varepsilon\} < \infty$$

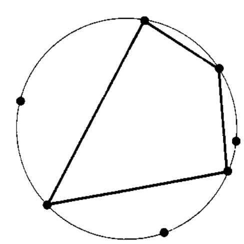
for each $\varepsilon > 0$. The Borel–Cantelli lemma turns this into the full almost sure convergence asserted by the Glivenko–Cantelli theorem.

II.4. Classes of Sets with Polynomial Discrimination

We made use of very few distinguishing properties of intervals for the proof of the Glivenko–Cantelli theorem in Section 3. The main requirement was that they should pick out at most $n + 1$ subsets from any set of n points. Other classes have a similar property. For example, quadrants of the form $(-\infty, \mathbf{t}]$ in $\mathbb{R}^2$ can pick out fewer than $(n + 1)^2$ different subsets from a set of n points in the plane—there are at most $n + 1$ places to set the horizontal boundary and at most $n + 1$ places to set the vertical boundary. (Problem 8 gives the precise upper bound.) With $(n + 1)^2$ replacing the $n + 1$ factor, we could repeat the arguments from Section 3 to get the bivariate analogue of the Glivenko–Cantelli theorem. The exponential bound would swallow up $(n + 1)^2$, just as it did the $n + 1$. Indeed, it would swallow up any polynomial. The argument works for intervals, quadrants, and any other class of sets that picks out a polynomial number of subsets.

13 Definition. Let $\mathscr{D}$ be a class of subsets of some space S. It is said to have polynomial discrimination (of degree v) if there exists a polynomial $\rho(\cdot)$ (of degree v) such that, from every set of N points in S, the class picks out at most $\rho(N)$ distinct subsets. Formally, if S_0 consists of N points, then there are at most $\rho(N)$ distinct sets of the form $S_0 \cap D$ with D in $\mathscr{D}$. Call $\rho(\cdot)$ the discriminating polynomial for $\mathscr{D}$. □

When the risk of confusion with the algebraic sort of polynomial is slight, let us shorten the name "class having polynomial discrimination" to "polynomial class," and adopt the usual terminology for polynomials of low degree. For example, the intervals on the real line have linear discrimination (they form a linear class) and the quadrants in the plane have quadratic discrimination (they form a quadratic class). Of course there are classes that don't have polynomial discrimination. For example, from every collection of N points lying on the circumference of a circle in $\mathbb{R}^2$ the class of closed, convex sets can pick out all 2^N subsets, and 2^N increases much faster than any polynomial.



The method of proof set out in Section 3 applies to any polynomial class of sets, provided measurability complications can be taken care of. Appendix C describes a general method for guarding against these complications. Classes satisfying the conditions described there are called permissible. Every specific class we shall encounter will be permissible. As the precise details of the method are rather delicate—they depend upon properties of analytic sets—let us adopt a naive approach. Ignore measurability problems from now on, but keep the term *permissible* as a reminder that some regularity conditions are needed if pathological examples (Problem 10) are to be excluded. Problems 3 through 7 describe a simpler approach, based on the more familiar idea of existence of countable, dense subclasses.

14 Theorem. *Let P be a probability measure on a space S. For every permissible class $\mathscr{D}$ of subsets of S with polynomial discrimination,*

$$\sup_{\mathscr{D}} |P_n D - PD| \to 0 \quad \textit{almost surely.}$$

PROOF. Go back to Section 3, change $\mathscr{I}$ to $\mathscr{D}$, replace the $n + 1$ multiplier by the polynomial appropriate to $\mathscr{D}$, and strike out the odd reference to interval and real line. □

Which classes have only polynomial discrimination? We already know about intervals and quadrants; their higher-dimensional analogues have the property too. Other classes can be built up from these.

15 Lemma. *If $\mathscr{C}$ and $\mathscr{D}$ have polynomial discrimination, then so do each of:*

(i) $\{D^c: D \in \mathscr{D}\}$;
(ii) $\{C \cup D: C \in \mathscr{C} \text{ and } D \in \mathscr{D}\}$;
(iii) $\{C \cap D: C \in \mathscr{C} \text{ and } D \in \mathscr{D}\}$.

PROOF. Write $c(\cdot)$ and $d(\cdot)$ for the discriminating polynomials. We may assume them both to be increasing functions of N. From a set S_0 consisting of N points, suppose $\mathscr{C}$ picks out subsets $S_1, \ldots, S_k$ with $k \leq c(N)$. Suppose S_i consists of N_i points. The class $\mathscr{D}$ picks out at most $d(N_i)$ distinct subsets from S_i. This gives the bound $d(N_1) + \cdots + d(N_k)$ for the size of the class in (iii). The sum is less than $c(N)\,d(N)$. That proves the assertion for (iii). The other two are just as easy. □

The lemma can be applied repeatedly to generate larger and larger polynomial classes. We must place a fixed limit on the number of operations allowed, though. For instance, the class of all singletons has only linear discrimination, but with arbitrary finite unions of singletons we can pick out any finite set.

Very quickly we run out of interesting new classes to manufacture by means of Lemma 15 from quadrants and the like. Fortunately, there are other systematic methods for finding polynomial classes.

Polynomials increase much more slowly than exponentials. For N large enough, a polynomial class must fail to pick out at least one of the 2^N subsets from each collection of N points. Surprisingly, this characterizes polynomial discrimination. Some picturesque terminology to describe the situation has become accepted in the literature. A class $\mathscr{D}$ is said to shatter a set of points F if it can pick out every possible subset (the empty subset and the whole of F included); that is, $\mathscr{D}$ shatters F if each of the subsets of F has the form $D \cap F$ for some D in $\mathscr{D}$. This conveys a slightly inappropriate image, in which F gets broken into tiny fragments, rather than an image of a diligent $\mathscr{D}$ trying to pick out all the different subsets of F; but at least it is vivid.

For example, the class of all closed discs in $\mathbb{R}^2$ can shatter each three-point set, provided the points are not collinear. But from no set of four points, no matter what its configuration, can the discs pick out more than 15 of the 16 possible subsets. The discs shatter some sets of three points; they shatter no set of four points.

16 Theorem. *Let S_0 be a set of N points in S. Suppose there is an integer $V \le N$ such that $\mathscr{D}$ shatters no set of V points in S_0. Then $\mathscr{D}$ picks out no more than $\binom{N}{0} + \binom{N}{1} + \cdots + \binom{N}{V-1}$ subsets from S_0.*

PROOF. Write $F_1, \ldots, F_k$ for the collection of all subsets of V elements from S_0. Of course $k = \binom{N}{V}$. By assumption, each F_i has a "hidden" subset H_i that $\mathscr{D}$ overlooks: $D \cap F_i \ne H_i$ for every D in $\mathscr{D}$. That is, all the sets of the form $D \cap S_0$, with D in $\mathscr{D}$, belong to

$$\mathscr{C}_0 = \{C \subseteq S_0 : C \cap F_i \ne H_i \text{ for each } i\}.$$

It will suffice to find an upper bound for the size of $\mathscr{C}_0$.

In one special case it is possible to count the number of sets in $\mathscr{C}_0$ directly. If $H_i = F_i$ for every i then no C in $\mathscr{C}_0$ can contain an F_i; no C can contain a set of V points. In other words, members of $\mathscr{C}_0$ consist of either $0, 1, \ldots,$ or $V - 1$ points. The sum of the binomial coefficients gives the number of sets of this form.

By playing around with the hidden sets we can reduce the general case to the special case just treated. Label the points of S_0 as $1, \ldots, N$. For each i define $H_i' = (H_i \cup \{1\}) \cap F_i$; that is, augment H_i by the point 1, provided it can be done without violating the constraint that the hidden set be contained in F_i. Define the corresponding class

$$\mathscr{C}_1 = \{C \subseteq S_0 : C \cap F_i \ne H_i' \quad \text{for each } i\}.$$

The class $\mathscr{C}_1$ has nothing much to do with $\mathscr{C}_0$. The only connection is that all its hidden sets, the sets it overlooks, are bigger. Let us show that this implies $\mathscr{C}_1$ has a greater cardinality than $\mathscr{C}_0$. (Notice: the assertion is not that $\mathscr{C}_0 \subseteq \mathscr{C}_1$.)

Check that the map $C \mapsto C \backslash \{1\}$ is one-to-one from $\mathscr{C}_0 \backslash \mathscr{C}_1$ into $\mathscr{C}_1 \backslash \mathscr{C}_0$. Start with any C in $\mathscr{C}_0 \backslash \mathscr{C}_1$. By definition, $C \cap F_i \ne H_i$ for every i, but $C \cap F_j = H_j'$ for at least one j. Deduce that $H_j \ne H_j'$, so 1 belongs to C and F_j and H_j', but not to H_j. The stripping of the point 1 does define a one-to-one map. Why should $C \backslash \{1\}$ belong to $\mathscr{C}_1 \backslash \mathscr{C}_0$? Observe that

$$(C \backslash \{1\}) \cap F_j = H_j' \backslash \{1\} = H_j,$$

which bars $C \backslash \{1\}$ from belonging to $\mathscr{C}_0$. Also, if F_i contains 1 then so must H_i', but $C \backslash \{1\}$ certainly cannot; and if F_i doesn't contain 1 then

$$(C \backslash \{1\}) \cap F_i = C \cap F_i \ne H_i = H_i'.$$

In either case $(C \backslash \{1\}) \cap F_i \ne H_i'$, so $C \backslash \{1\}$ belongs to $\mathscr{C}_1$.

Repeat the procedure, starting from the new hidden sets and with 2 taking over the role played by 1. Define $H_i'' = (H_i' \cup \{2\}) \cap F_i$ and

$$\mathscr{C}_2 = \{C \subseteq S_0 : C \cap F_i \neq H_i'' \text{ for each } i\}.$$

The cardinality of $\mathscr{C}_2$ is greater than the cardinality of $\mathscr{C}_1$. Another $N - 2$ repetitions would generate classes $\mathscr{C}_3, \mathscr{C}_4, \ldots, \mathscr{C}_N$ with increasing cardinalities. The hidden sets for $\mathscr{C}_N$ would fill out the whole of each F_i: the special case already treated. □

17 Corollary. *If a class shatters no set of V points, then it must have polynomial discrimination of degree no greater than $V - 1$.* □

All we lack now is a good method for identifying classes that have trouble picking out subsets from large enough sets of points.

18 Lemma. *Let $\mathscr{G}$ be a finite-dimensional vector space of real functions on S. The class of sets of the form $\{g \geq 0\}$, for g in $\mathscr{G}$, has polynomial discrimination of degree no greater than the dimension of $\mathscr{G}$.*

Proof. Write $V - 1$ for the dimension of $\mathscr{G}$. Choose any collection $\{s_1, \ldots, s_V\}$ of distinct points from S. (Everything reduces to triviality if S contains fewer than V points.) Define a linear map L from $\mathscr{G}$ into $\mathbb{R}^V$ by

$$L(g) = (g(s_1), \ldots, g(s_V)).$$

Since $L\mathscr{G}$ is a linear subspace of $\mathbb{R}^V$ of dimension at most $V - 1$, there exists in $\mathbb{R}^V$ a non-zero vector γ orthogonal to $L\mathscr{G}$. That is,

$$\sum_i \gamma_i g(s_i) = 0 \quad \text{for each } g \text{ in } \mathscr{G},$$

or

$$(19) \qquad \sum_{\{+\}} \gamma_i g(s_i) = \sum_{\{-\}} (-\gamma_i) g(s_i) \quad \text{for each } g.$$

Here $\{+\}$ stands for the set of those i for which $\gamma_i \geq 0$, and $\{-\}$ for those with $\gamma_i < 0$. Replacing γ by $-\gamma$ if necessary, we may assume that $\{-\}$ is non-empty.

Suppose there were a g for which $\{g \geq 0\}$ picked out precisely those points s_i with i in $\{+\}$. For this g, the left-hand side of (19) would be ≥ 0, but the right-hand side would be < 0. We have found a set that cannot be picked out. □

Many familiar classes of geometric objects fall within the scope of the lemma. For example, the class of subsets of the plane generated by the linear space of quadratic forms $ax^2 + bxy + cy^2 + dx + ey + f$ includes all closed discs, ellipsoids, and (as a degenerate case) half-spaces. More complicated regions, such as intersections of 257 closed or open half-spaces, can be built up from these by means of Lemma 15. You can feed them into

Theorem 14 to churn out a whole host of generalizations of the classical Glivenko–Cantelli theorem.

Uniform limit theorems for polynomial classes of sets have one thing in common: they hold regardless of the sampling distribution. This happens because the number $\Delta_n(\boldsymbol{\xi})$ of subsets picked out by the class from the sample $\{\xi_1, \ldots, \xi_n\}$ can be bounded above by a polynomial in n, independently of the configuration of that sample. Without the uniform bound the inequality (12) would be replaced by

$$(20) \qquad \mathbb{P}\{\|P_n^\circ\| > \tfrac{1}{4}\varepsilon \,|\, \boldsymbol{\xi}\} \le 2\Delta_n(\boldsymbol{\xi}) \exp(-n\varepsilon^2/32).$$

Write W_n for the minimum of 1 and the right-hand side of (20). Then the argument from the INTEGRATION step gives the sharper bound

$$\mathbb{P}\{\|P_n - P\| > \varepsilon\} \le 4\mathbb{P}W_n$$

for all n large enough. Thus a sufficient condition for $\|P_n - P\|$ to converge in probability to zero is: $\mathbb{P}W_n \to 0$ for each $\varepsilon > 0$. Equivalently, because $0 \le W_n \le 1$, we could check that $\log \Delta_n(\boldsymbol{\xi}) = o_p(n)$. Theorem 16 helps here.

$$\Delta_n(\boldsymbol{\xi}) \le B_n(V-1) = \binom{n}{0} + \cdots + \binom{n}{V-1},$$

where $V = V_n(\xi_1, \ldots, \xi_n)$ is the smallest integer such that $\mathscr{D}$ shatters no collection of V points from $\{\xi_1, \ldots, \xi_n\}$. Set $k = V - 1$. If $k \le \frac{1}{2}n$, all the terms in the sum for $B_n(k)$ are less than $\binom{n}{k}$:

$$n^{-1} \log B_n(k) \le n^{-1} \log[(k+1)n!/(n-k)!\,k!].$$

Three applications of Stirling's approximation and some tidying up reduce the right-hand side to

$$-(1 - k/n)\log(1 - k/n) - (k/n)\log(k/n) + o(1),$$

which tends to zero as $k/n \to 0$. It follows that both $n^{-1} \log \Delta_n \to 0$ and $\|P_n - P\| \to 0$ in probability, if $V/n \to 0$ in probability.

If we don't know how fast V/n converges to zero, we can't use the Borel–Cantelli lemma to deduce from these inequalities that $\|P_n - P\|$ converges almost surely to zero. But there is another reason why the convergence in probability implies the stronger result.

Symmetry properties would force $\|P_n - P\|$ to converge almost surely to some constant, no matter how V/n behaved. Given P_n, the unordered set $\{\xi_1, \ldots, \xi_n\}$ is uniquely determined, but there's no way of deciding the order in which the observations were generated. Given P_{n+1}, we know slightly less about $\{\xi_1, \ldots, \xi_n\}$; it could be any of the $(n+1)$ possible subsets of size n obtained by deleting one of the support points of P_{n+1}. (Count coincident observations as distinct support points.) The conditional distribution of P_n given P_{n+1} must be uniform on one of these $(n+1)$ subsets, each subset being chosen with probability $(n+1)^{-1}$. The conditional expectation of P_n given P_{n+1} (in the intuitive sense of the average over the $n+1$ possible choices for P_n) must be P_{n+1}. The extra information carried by $P_{n+2}, P_{n+3}, \ldots$ adds nothing more to our knowledge about P_n; the conditional expectation of P_n given the σ-field generated by $P_{n+1}, P_{n+2}, \ldots$ still equals P_{n+1}. That is, the sequence $\{P_n\}$ is a reversed martingale, in some wonderful measure-valued sense. Apply Jensen's inequality to the convex function that takes P_n onto $\|P_n - P\|$ to deduce that $\{\|P_n - P\|\}$ is a bounded, reversed submartingale. (Problem 11 arrives at the same conclusion in a slightly more rigorous manner.) Such a sequence must converge almost surely (Neveu 1975, Proposition V-3-13) to a limit random variable, W. Since W is unchanged by finite permutations of $\{\xi_i\}$, the zero-one law of Hewitt and Savage (Breiman 1968, Section 3.9) forces it to take on a constant value almost surely. The only question remaining for the proof of a uniform strong law of large numbers is whether the constant equals zero or not: convergence in probability to zero plus convergence almost surely to a constant gives convergence almost surely to zero.

21 Theorem. *Let $\mathscr{D}$ be a permissible class of subsets of S. A necessary and sufficient condition for*

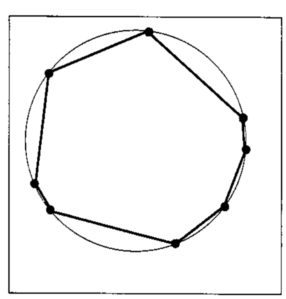
$$\sup_{\mathscr{D}} |P_n D - PD| \to 0 \quad \textit{almost surely}$$

is the convergence of $n^{-1}V_n$ to zero in probability, where $V_n = V_n(\xi_1, \ldots, \xi_n)$ is the smallest integer such that $\mathscr{D}$ shatters no collection of V_n points from $\{\xi_1, \ldots, \xi_n\}$.

PROOF. You can formalize the sufficiency argument outlined above; necessity is taken care of in Problem 12. □

Because $0 \le n^{-1}V_n \le 1$, convergence in probability of $n^{-1}V_n$ to zero is equivalent to $n^{-1}\mathbb{P}V_n \to 0$. This has an appealing interpretation. The uniform strong law of large numbers holds if and only if, on the average, the class of sets behaves as if it has polynomial discrimination with degree but a tiny fraction of the sample size.

22 Example. Let's see how easy it is to check the necessary and sufficient condition stated in Theorem 21. Consider the class $\mathscr{C}$ of all closed, convex subsets of the unit square $[0, 1]^2$. We know that there exist arbitrarily large collections of points shattered by $\mathscr{C}$. Were we sampling from a non-atomic



distribution concentrated around the rim of a disc inside $[0, 1]^2$, the class $\mathscr{C}$ could always pick out too many subsets from the sample. Indeed, there would always exist a convex C with $P_n C = 1$ and $PC = 0$. But such configurations of sample points should be thoroughly atypical for sampling from the uniform distribution on $[0, 1]^2$. Theorem 21 should say something useful in that case.

How large a subcollection of sample points can $\mathscr{C}$ shatter? Suppose it is larger than the size requested by Theorem 21. That is, for some $\varepsilon > 0$,

$$\mathbb{P}\{n^{-1}V_n \geq \varepsilon\} \geq \varepsilon \quad \text{infinitely often.}$$

This will lead us to a contradiction.

A set of k points is shattered by $\mathscr{C}$ if and only if none of the points can be written as a convex combination of the others; each must be an extreme point of their convex hull. So there exists a convex set whose boundary has empirical measure at least k/n, which seems highly unlikely because P puts zero measure around the boundary of every convex set. Be careful of this plausibility argument; it contains a hidden appeal to the very uniformity result we are trying to establish. An approximation argument will help us to avoid the trap.

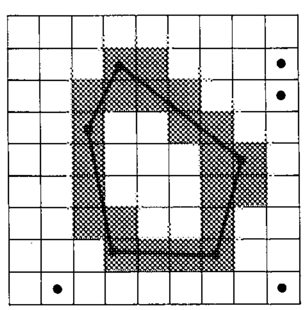
Divide $[0, 1]^2$ into a patchwork of m^2 equal subsquares, for some fixed m that will be specified shortly. Because the class $\mathscr{A}$ of all possible unions of these subsquares is finite,

$$\mathbb{P}\left\{\sup_{\mathscr{A}} |P_n A - PA| \geq \tfrac{1}{2}\varepsilon\right\} < \tfrac{1}{2}\varepsilon \quad \text{for all } n \text{ large enough.}$$

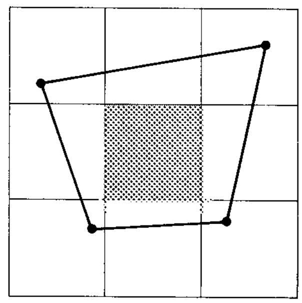
The $\frac{1}{2}\varepsilon$ here is chosen to ensure that, for some n,

$$\mathbb{P}\{n^{-1}V_n \geq \varepsilon \text{ and } \sup_{\mathscr{A}} |P_n A - PA| < \tfrac{1}{2}\varepsilon\} > \tfrac{1}{2}\varepsilon.$$

Since a set with positive probability can't be empty, there must exist a sample configuration for which $\mathscr{C}$ shatters some collection of at least $n\varepsilon$ sample points and for which $|P_n A - PA| < \frac{1}{2}\varepsilon$ for every A in $\mathscr{A}$. Write H for the convex hull of the shattered set, and A_H for the union of those subsquares that intersect the boundary of H. The set A_H contains all the extreme points of H, so $P_n A_H \geq \varepsilon$; it belongs to $\mathscr{A}$, so $|P_n A_H - PA_H| < \frac{1}{2}\varepsilon$. Consequently $PA_H > \frac{1}{2}\varepsilon$, which will give the desired contradiction if we make m large enough.



Experiment with values of m equal to a power of 3. No convex set can have boundary points in all nine of the subsquares; the middle subsquare would lie inside the convex hull of four points occupying each of the four corner squares. For every convex C the P measure of the union of those



subsquares intersecting its boundary must be less than $\frac{8}{9}$. Subdivide each of the nine subsquares into nine parts, then repeat the same argument eight times. This brings the measure of squares on the boundary down to $(\frac{8}{9})^2$. Keep repeating the argument until the power of $\frac{8}{9}$ falls below $\frac{1}{2}\varepsilon$. That destroys the claim made for A_H. □

II.5. Classes of Functions

The direct approximation methods of Section 2 gave us sufficient conditions for the empirical measure P_n to converge to the underlying P uniformly over a class of functions,

$$\sup_{\mathscr{F}} |P_n f - Pf| \to 0 \quad \text{almost surely.}$$

The conditions, though straightforward, can prove burdensome to check. In this section a transfusion of ideas from Sections 3 and 4 will lead to a more tractable condition for the uniform convergence. The method will depend heavily on the independence of the observations $\{\xi_i\}$, but the assumption of identical distribution could be relaxed (Problem 23).

Throughout the section write $\|\cdot\|$ to denote $\sup_{\mathscr{F}} |\cdot|$.

Let us again adopt a naive approach towards possible measurability difficulties, with only the word permissible (explained in Appendix C) to remind us that some regularity conditions are needed to exclude pathological examples.

A domination condition will guard against any complications that could be caused by $\mathscr{F}$ containing unbounded functions. Call each measurable F such that $|f| \leq F$, for every f in $\mathscr{F}$, an envelope for $\mathscr{F}$. Often F will be taken as the pointwise supremum of $|f|$ over $\mathscr{F}$, the natural envelope, but it will be convenient not to force this. We shall assume $PF < \infty$. With the proper centering, the natural envelope must satisfy this condition (Problem 14) if the uniform strong law holds.

The key to the uniform convergence will again be an approximation condition, but this time with distances calculated using the $\mathscr{L}^1$ seminorm for the empirical measures themselves. This allows us to drop the requirement that the approximating functions sandwich each member of $\mathscr{F}$.

23 Definition. Let Q be a probability measure on S and $\mathscr{F}$ be a class of functions in $\mathscr{L}^1(Q)$. For each $\varepsilon > 0$ define the covering number $N_1(\varepsilon, Q, \mathscr{F})$ as the smallest value of m for which there exist functions $g_1, \ldots, g_m$ (not necessarily in $\mathscr{F}$) such that $\min_j Q|f - g_j| \leq \varepsilon$ for each f in $\mathscr{F}$. For definiteness set $N_1(\varepsilon, Q, \mathscr{F}) = \infty$ if no such m exists. □

If $\mathscr{F}$ has envelope F we can require that the approximating functions satisfy the inequality $|g_j| \leq F$ without increasing $N_1(\varepsilon, Q, \mathscr{F})$: replace g_j by

$$\max\{-F, \min[F, g_j]\}.$$

We could also require g_j to belong to $\mathscr{F}$, at the cost of a doubling of ε: replace g_j by an f_j in $\mathscr{F}$ for which $Q|f_j - g_j| \leq \varepsilon$.

24 Theorem. *Let $\mathscr{F}$ be a permissible class of functions with envelope F. Suppose $PF < \infty$. If P_n is obtained by independent sampling from the probability measure P and if* $\log N_1(\varepsilon, P_n, \mathscr{F}) = o_p(n)$ *for each fixed $\varepsilon > 0$, then* $\sup_{\mathscr{F}} |P_n f - Pf| \to 0$ *almost surely.*

Proof. Problem 11 (or the slightly less formal symmetry argument leading up to Theorem 21 in Section 4) shows that $\{\|P_n - P\|\}$ is a reversed submartingale; it converges almost surely to a constant. It will suffice if we deduce from the approximation condition that $\{\|P_n - P\|\}$ converges in probability to zero.

Exploit integrability of the envelope to truncate the functions back to a finite range. Given $\varepsilon > 0$, choose a constant K so large that $PF\{F > K\} < \varepsilon$. Then

$$\begin{aligned}\sup_{\mathscr{F}} |P_n f - Pf| \leq{} & \sup_{\mathscr{F}} |P_n f\{F \leq K\} - Pf\{F \leq K\}| \\ & + \sup_{\mathscr{F}} P_n|f|\{F > K\} + \sup_{\mathscr{F}} P|f|\{F > K\}.\end{aligned}$$

Because $|f| \leq F$ for each f in $\mathscr{F}$, the last two terms sum to less than

$$P_n F\{F > K\} + PF\{F > K\}.$$

This converges almost surely to $2PF\{F > K\}$, which is less than 2ε. It remains for us to show that the supremum over the functions $f\{F \leq K\}$ converges in probability to zero. As truncation can only decrease the $\mathscr{L}^1(P_n)$ distance between two functions, the condition on log covering numbers also holds if each f is replaced by its truncation; without loss of generality we may assume that $|f| \leq K$ for each f in $\mathscr{F}$.

In the two SYMMETRIZATION steps of the proof of the Glivenko–Cantelli theorem (Section 3) we showed that

$$\mathbb{P}\{\|P_n - P\| > \varepsilon\} \leq 4\mathbb{P}\{\|P_n^\circ\| > \tfrac{1}{4}\varepsilon\} \quad \text{for} \quad n \geq 8\varepsilon^{-2},$$

where $\|\cdot\|$ denoted a supremum over intervals $(-\infty, t]$ of the real line. The signed measure P_n° put mass $\pm n^{-1}$ on each observation $\xi_1, \ldots, \xi_n$, the random $\pm$ signs being decided independently of the $\{\xi_i\}$. The argument works just as well if $\|\cdot\|$ denotes a supremum over $\mathscr{F}$, the interpretation adopted in the current section. The only property of the indicator function $(-\infty, t]$ needed in the SYMMETRIZATION steps was the boundedness, which implied $\operatorname{var}(P_n(-\infty, t]) \leq n^{-1}$. This time an extra factor of K^2 would appear in the lower bound for n.

With intervals we were able to reduce $\|P_n^\circ\|$ to a maximum over a finite collection; for functions the reduction will not be quite so startling. Given $\boldsymbol{\xi}$, choose functions $g_1, \ldots, g_M$, where $M = N_1(\frac{1}{8}\varepsilon, P_n, \mathscr{F})$, such that

$$\min_j P_n|f - g_j| \leq \tfrac{1}{8}\varepsilon \quad \text{for each } f \text{ in } \mathscr{F}.$$

Write f^* for the g_j at which the minimum is achieved.

Now we reap the benefits of approximation in the $\mathscr{L}^1(P_n)$ sense. For any function g,

$$|P_n^\circ g| = \left|n^{-1} \sum_{i=1}^n \pm g(\xi_i)\right| \leq n^{-1} \sum_{i=1}^n |g(\xi_i)| = P_n|g|.$$

Choose $g = f - f^*$ for each f in turn.

$$\begin{aligned}
\mathbb{P}\left\{\sup_{\mathscr{F}} |P_n^\circ f| > \tfrac{1}{4}\varepsilon \,\middle|\, \boldsymbol{\xi}\right\} &\leq \mathbb{P}\left\{\sup_{\mathscr{F}} \left[|P_n^\circ f^*| + P_n|f - f^*|\right] > \tfrac{1}{4}\varepsilon \,\middle|\, \boldsymbol{\xi}\right\} \\
&\leq \mathbb{P}\left\{\max_j |P_n^\circ g_j| > \tfrac{1}{8}\varepsilon \,\middle|\, \boldsymbol{\xi}\right\} \text{ because } P_n|f - f^*| \leq \tfrac{1}{8}\varepsilon \\
&\leq N_1(\tfrac{1}{8}\varepsilon, P_n, \mathscr{F}) \max_j \mathbb{P}\{|P_n^\circ g_j| > \tfrac{1}{8}\varepsilon \,|\, \boldsymbol{\xi}\}.
\end{aligned}$$

Once again Hoeffding's Inequality (Appendix B) gives an excellent bound on the conditional probabilities for each g_j.

$$\begin{aligned}
\mathbb{P}\{|P_n^\circ g_j| > \tfrac{1}{8}\varepsilon \,|\, \boldsymbol{\xi}\} &= \mathbb{P}\left\{\left|\sum_{i=1}^n \pm g_j(\xi_i)\right| > \tfrac{1}{8}n\varepsilon \,\middle|\, \boldsymbol{\xi}\right\} \\
&\leq 2\exp\left[-2(\tfrac{1}{8}n\varepsilon)^2 \Big/ \sum_{i=1}^n (2g_j(\xi_i))^2\right] \\
&\leq 2\exp(-n\varepsilon^2/128K^2) \quad \text{because } |g_j| \leq K.
\end{aligned}$$

When the logarithm of the covering number is less than $n\varepsilon^2/256K^2$, the inequality

$$\mathbb{P}\{\|P_n^\circ\| > \tfrac{1}{4}\varepsilon \,|\, \xi\} \leq 2\exp[\log N_1(\tfrac{1}{8}\varepsilon, P_n, \mathscr{F}) - n\varepsilon^2/128K^2]$$

will serve us well; otherwise use the trivial upper bound of 1. Integrate out.

$$\mathbb{P}\{\|P_n^\circ\| > \tfrac{1}{4}\varepsilon\} \leq 2\exp(-n\varepsilon^2/256K^2) + \mathbb{P}\{\log N_1(\tfrac{1}{8}\varepsilon, P_n, \mathscr{F}) > n\varepsilon^2/256K^2\}.$$

Both terms on the right-hand side of the inequality converge to zero. □

For some classes of functions the conditions of the theorem are easily met because $N_1(\varepsilon, P_n, \mathscr{F})$ remains bounded for each fixed $\varepsilon > 0$. This happens if the graphs of the functions in $\mathscr{F}$ form a polynomial class of sets. The graph of a real-valued function f on a set S is defined as the subset

$$G_f = \{(s, t): 0 \leq t \leq f(s) \quad \text{or} \quad f(s) \leq t \leq 0\}$$

of $S \otimes \mathbb{R}$. We learn something about the covering numbers of a class $\mathscr{F}$ by observing how its graphs pick out sets of points in $S \otimes \mathbb{R}$.

25 Approximation Lemma. *Let $\mathscr{F}$ be a class of functions on a set S with envelope F, and let Q be a probability measure on S with $0 < QF < \infty$. If the graphs of functions in $\mathscr{F}$ form a polynomial class of sets then*

$$N_1(\varepsilon QF, Q, \mathscr{F}) \leq A\varepsilon^{-W} \quad \textit{for} \quad 0 < \varepsilon < 1,$$

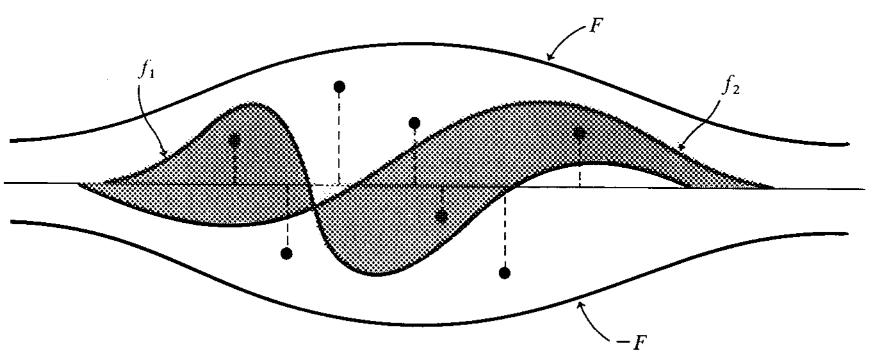
where the constants A and W depend upon only the discriminating polynomial of the class of graphs.

Proof. Let $f_1, \ldots, f_m$ be a maximal collection of functions in $\mathscr{F}$ for which

$$Q|f_i - f_j| > \varepsilon QF \quad \text{if} \quad i \neq j.$$

Maximality means that no larger collection has the same property; each f must lie within εQF of at least one f_j. Thus $m \geq N_1(\varepsilon QF, Q, \mathscr{F})$.

Choose independent points $(s_1, t_1), \ldots, (s_k, t_k)$ in $S \otimes \mathbb{R}$ by a two-step procedure. First sample s_i from the distribution $Q(\cdot F)/Q(F)$ on S. Given s_i, sample t_i from the conditional distribution Uniform$[-F(s_i), F(s_i)]$. The value of k, which depends on m and ε, will be specified soon.



The graphs G_1 and G_2, corresponding to f_1 and f_2, pick out the same subset from this sample if and only if every one of the k points lands outside the region $G_1 \triangle G_2$. This occurs with probability equal to

$$\begin{aligned}\prod_{i=1}^{k} [1 - \mathbb{P}\mathbb{P}\{(s_i, t_i) \in G_1 \triangle G_2 \mid s_i\}] &= [1 - \mathbb{P}(|f_1(s_1) - f_2(s_1)|/2F(s_1))]^k \\ &= [1 - Q|f_1 - f_2|/2Q(F)]^k \\ &\le (1 - \tfrac{1}{2}\varepsilon)^k \\ &\le \exp(-\tfrac{1}{2}k\varepsilon).\end{aligned}$$

Apply the same reasoning to each of the $\binom{m}{2}$ possible pairs of functions f_i and f_j. The probability that at least one pair of graphs picks out the same set of points from the k sample is less than

$$\binom{m}{2} \exp(-\tfrac{1}{2}k\varepsilon) \le \tfrac{1}{2} \exp(2 \log m - \tfrac{1}{2}k\varepsilon).$$

Choose k to be the smallest value that makes the upper bound strictly less than 1. Certainly $k \le (1 + 4 \log m)/\varepsilon$. With positive probability the graphs all pick different subsets from the k sample; there exists a set of k points in $S \otimes \mathbb{R}$ from which the polynomial class of graphs can pick out m distinct subsets. From the defining property of polynomial classes, there exist constants B and V such that $m \le Bk^V$ for all $k \ge 1$. Find n_0 so that $(1 + 4 \log n)^V \le n^{1/2}$ for all $n \ge n_0$. Then either $m < n_0$ or $m \le Bm^{1/2}\varepsilon^{-V}$. Set $W = 2V$ and $A = \max(B^2, n_0)$. □

To show that a class of graphs has only polynomial discrimination we can call upon the results of Section 4. We build up the graphs as finite unions and intersections (Lemma 15) of simpler classes of sets. We establish their discrimination properties by direct geometric argument (as for intervals and quadrants) or by exploitation of finite dimensionality (as in Lemma 18) of a generating class of functions.

26 Example. Define a center of location for a distribution P on $\mathbb{R}^m$ as any value θ minimizing the criterion function

$$H(\theta, P) = P\phi(|x - \theta|),$$

where $\phi(\cdot)$ is a continuous, non-decreasing function on $[0, \infty)$ and $|\cdot|$ denotes the usual euclidean distance. If $P\phi(|x|) < \infty$ and $\phi(\cdot)$ does not increase too rapidly, in the sense that there exists a constant C for which $\phi(2t) \le C\phi(t)$ for all t, then the function $H(\cdot, P)$ is well defined:

$$H(\theta, P) \le P[\phi(2|\theta|)\{|x| \le |\theta|\} + C\phi(|x|)\{|x| > |\theta|\}] < \infty.$$

If trivial cases are ruled out by the requirement

$$P\{x: \phi(|x|) < \phi(\infty -)\} > 0, \tag{27}$$

the minimizing value will be achieved (Problem 21); extra regularity conditions on P, which are satisfied by distributions such as the multivariate normal, ensure uniqueness (Problem 22). For this example, let us not get bogged down by the exact conditions needed; just assume that $H(\cdot, P)$ has a unique minimum at some θ_0.

Estimate θ_0 by any value θ_n that minimizes the sample criterion function $H(\cdot, P_n)$. To show that θ_n converges to θ_0 almost surely, it will suffice to prove that $H(\theta_n, P) \to H(\theta_0, P)$ almost surely, because $H(\theta, P)$ is bounded away from $H(\theta_0, P)$ outside each neighborhood of θ_0.

The argument follows the same pattern as for k-means (Example 4). First show that θ_n eventually stays within a large compact ball $\{|x| \le K\}$. Choose the K greater than $|\theta_0|$ and large enough to ensure that

$$\phi(\tfrac{1}{2}K)P\{|x| \le \tfrac{1}{2}K\} > P\phi(|x|),$$

which is possible by (27): as K tends to infinity the left-hand side converges to $\phi(\infty -)$. Such a K will suffice because $H(0, P_n) = P_n\phi(|x|)$ and

$$H(\theta, P_n) \ge \phi(\tfrac{1}{2}K)P_n\{|x| \le \tfrac{1}{2}K\}$$

for every θ with $|\theta| > K$.

If we prove uniform almost sure convergence of P_n to P over the class

$$\mathscr{F} = \{\phi(|\cdot - \theta|) : |\theta| \le K\},$$

then we can deduce almost surely that $H(\theta_n, P) \to H(\theta_0, P)$ from

$$H(\theta_n, P_n) - H(\theta_n, P) \to 0,$$
$$H(\theta_n, P_n) \le H(\theta_0, P_n) \to H(\theta_0, P) \le H(\theta_n, P).$$

Here's our chance to apply Theorem 24.

The class $\mathscr{F}$ has envelope $\phi(2K) + C\phi(|x|)$, which satisfies the first requirement of the theorem. Bound the covering numbers by showing that the graphs of functions in $\mathscr{F}$ have only polynomial discrimination. We may assume that $\phi(0) = 0$. The graph of $\phi(|\cdot - \theta|)$ contains a point (y, t), with $t \ge 0$, if and only if $|y - \theta| \ge \alpha(t)$, where $\alpha(t)$ denotes the smallest value of α for which $\phi(\alpha) \ge t$. From a collection of points $\{(y_i, t_i)\}$ the graph picks out those points satisfying $|y_i|^2 - 2y_i \cdot \theta + |\theta|^2 - \alpha(t_i)^2 \ge 0$. Construct from (y_i, t_i) a point $z_i = (y_i, |y_i|^2 - \alpha(t_i)^2)$ in $\mathbb{R}^{m+1}$. On $\mathbb{R}^{m+1}$ define a vector space $\mathscr{G}$ of functions

$$g_{\beta,\gamma,\delta}(x, s) = \beta \cdot x + \gamma s + \delta$$

with parameters β in $\mathbb{R}^m$ and γ, δ in $\mathbb{R}$. By Lemma 18, the sets $\{g \ge 0\}$, for g in $\mathscr{G}$, pick out only a polynomial number of subsets from $\{z_i\}$; those sets corresponding to functions in $\mathscr{G}$ with $\beta = -2\theta$, $\gamma = 1$, and $\delta = |\theta|^2$ pick out even fewer subsets from $\{z_i\}$. The graphs of functions $\phi(|\cdot - \theta|)$ have only polynomial discrimination. □

Buried within the argument of the last example lies a mini lemma relating finite dimensionality of a class of functions to discrimination properties of the graphs. It is perhaps worth noting.

28 Lemma. *Let $\mathscr{F}$ be a finite-dimensional vector space of real functions on S. The class of graphs of functions in $\mathscr{F}$ has polynomial discrimination.*

Proof. Define on $S \otimes \mathbb{R}$ the vector space $\mathscr{G}$ of real functions

$$g_{f,r}(s, t) = f(s) - rt.$$

Define

$$G_1 = \{g_{f,1} \geq 0\} = \{(s, t): f(s) \geq t\}$$
$$G_2 = \{g_{-f,-1} \geq 0\} = \{(s, t): f(s) \leq t\}$$

Observe that

$$G_f = G_1\{t \geq 0\} \cup G_2\{t \leq 0\}.$$

Invoke Lemmas 18 and 15. □

29 Example. Now let's have another try at the k-means problem introduced in Example 4. There we met the class of functions of the form

$$f_{a,b}(x) = |x - a|^2 \wedge |x - b|^2$$

with (a, b) ranging over the subset C of $\mathbb{R}^2$. We know that $\sup_C f_{a,b} \leq F$ for an F with $PF < \infty$, provided $P|x|^2 < \infty$.

The graphs of functions $|x - a|^2$ form a class with polynomial discrimination, by Lemma 28. Intersect pairs of such graphs in all possible ways to get the graphs of all functions $f_{a,b}$. Apply Lemma 15 (to handle the intersections), then the Approximation Lemma (to bound covering numbers), then Theorem 24:

$$\sup_C |P_n f_{a,b} - Pf_{a,b}| \to 0 \quad \text{almost surely.}$$

Compare this with the direct approximation argument of Example 4. □

II.6. Rates of Convergence

Theorem 24 imposed the condition $\log N_1(\varepsilon, P_n, \mathscr{F}) = o_p(n)$ on the rate of growth of the covering numbers. Many classes meet the condition easily. For example, if the graphs of functions from $\mathscr{F}$ have only polynomial discrimination, the covering numbers stay smaller than a fixed polynomial in ε^{-1}. The method of proof will deliver a finer result for such a class; we can get good bounds not just for a fixed ε deviation but also for an ε_n that decreases to zero as n increases. That is, we get a rate of convergence for the uniform strong law of large numbers. The method will also allow the class of functions to change with n, provided the covering numbers do not grow too rapidly. If the classes are uniformly bounded, and if the supremum of Pf^2 over the nth class tends to zero as n increases, this will speed the rate of convergence.

Consider the effect upon the two key steps of the argument for Theorem 24 if we let both ε and $\mathscr{F}$ depend on n. As before, replace $P_n - P$ by the signed measure P_n° that places mass $\pm n^{-1}$ at each of $\xi_1, \ldots, \xi_n$. The symmetrization inequality

$$\mathbb{P}\left\{\sup_{\mathscr{F}_n} |P_n f - Pf| > 8\varepsilon_n\right\} \le 4\mathbb{P}\left\{\sup_{\mathscr{F}_n} |P_n^\circ f| > 2\varepsilon_n\right\} \tag{30}$$

still holds provided $\operatorname{var}(P_n f)/(4\varepsilon_n)^2 \le \frac{1}{2}$ for each f in $\mathscr{F}_n$. The approximation argument and Hoeffding's Inequality still lead to

$$\mathbb{P}\left\{\sup_{\mathscr{F}_n} |P_n^\circ f| > 2\varepsilon_n \,\middle|\, \xi\right\} \le 2N_1(\varepsilon_n, P_n, \mathscr{F}_n) \exp\left[-\tfrac{1}{2}n\varepsilon_n^2 \Big/ \left(\max_j P_n g_j^2\right)\right], \tag{31}$$

where the maximum runs over all $N_1(\varepsilon_n, P_n, \mathscr{F}_n)$ functions $\{g_j\}$ in the approximating class.

If the supremum over $\mathscr{F}_n$ of Pf^2 tends to zero, one might expect that the maximum over the $\{P_n g_j^2\}$ should converge to zero at about the same rate. The next lemma will help us make the idea precise if the approximating $\{g_j\}$ are chosen from $\mathscr{F}_n$. As squares of functions are involved, covering numbers need to be calculated using $\mathscr{L}^2$ seminorms rather than the $\mathscr{L}^1$ seminorms of Definition 23.

32 Definition. Let Q be a probability measure on S and $\mathscr{F}$ be a class of functions in $\mathscr{L}^2(Q)$. For each $\varepsilon > 0$ define the covering number $N_2(\varepsilon, Q, \mathscr{F})$ as the smallest value of m for which there exist functions $g_1, \ldots, g_m$ (not necessarily in $\mathscr{F}$) such that $\min_j (Q(f - g_j)^2)^{1/2} \le \varepsilon$ for each f in $\mathscr{F}$. For definiteness set $N_2(\varepsilon, Q, \mathscr{F}) = \infty$ if no such m exists. □

As before, if $\mathscr{F}$ has envelope F we can require that $|g_j| \le F$; and we could require g_j to belong to $\mathscr{F}$, at the cost of a doubling of ε, by substituting for g_j an f_j in $\mathscr{F}$ such that $(Q(f_j - g_j)^2)^{1/2} \le \varepsilon$.

33 Lemma. *Let $\mathscr{F}$ be a permissible class of functions with $|f| \le 1$ and $(Pf^2)^{1/2} \le \delta$ for each f in $\mathscr{F}$. Then*

$$\mathbb{P}\left\{\sup_{\mathscr{F}} (P_n f^2)^{1/2} > 8\delta\right\} \le 4\mathbb{P}[N_2(\delta, P_n, \mathscr{F}) \exp(-n\delta^2) \wedge 1].$$

Proof. Let P_n' be an independent copy of P_n. Write $Z(f)$ for $(P_n f^2)^{1/2}$ and $Z'(f)$ for $(P_n' f^2)^{1/2}$. From the Symmetrization Lemma of Section 3,

$$\mathbb{P}\left\{\sup_{\mathscr{F}} |Z(f)| > 8\delta\right\} \le \tfrac{4}{3}\mathbb{P}\left\{\sup_{\mathscr{F}} |Z(f) - Z'(f)| > 6\delta\right\} \tag{34}$$

because, for each f in $\mathscr{F}$,

$$\mathbb{P}\{|Z'(f)| \leq 2\delta\} \geq 1 - \mathbb{P}(Z'(f)^2)/4\delta^2 = 1 - (Pf^2)/4\delta^2.$$

The intrusion of the square-root into the definition of Z and Z' would complicate reduction to the P_n° process. Instead, construct P_n and P_n' by a method that guarantees equal numbers of observations for both empirical measures. Sample $2n$ observations $X_1, \ldots, X_{2n}$ from P. Independently of the vector $\mathbf{X}$ of these observations, generate independent selection variables $\tau_1, \ldots, \tau_n$ with $\mathbb{P}\{\tau(i) = 1\} = \mathbb{P}\{\tau(i) = 0\} = \frac{1}{2}$. Use these to choose one observation from each of the pairs (X_{2i-1}, X_{2i}), for $i = 1, \ldots, n$. Construct P_n from these observations, and P_n' from the remaining observations. Formally, set $\xi_i = X_{2i-1+\tau(i)}$ and $\xi_i' = X_{2i-\tau(i)}$, then put mass n^{-1} on each point ξ_i for P_n, and put mass n^{-1} on each ξ_i' for P_n'. Set $S_{2n} = \frac{1}{2}(P_n + P_n')$. It has the same distribution as P_{2n}.

Temporarily write $\rho(\cdot)$ for the $\mathscr{L}^2(S_{2n})$ seminorm: $\rho(f) = (S_{2n}f^2)^{1/2}$. Given $\mathbf{X}$, find functions $g_1, \ldots, g_M$, where $M = N_2(\sqrt{2}\delta, S_{2n}, \mathscr{F})$, for which

$$\min_j \rho(f - g_j) \leq \sqrt{2}\delta \quad \text{for every } f \text{ in } \mathscr{F}.$$

We may assume that $|g_j| \leq 1$ for every j. The awkward $\sqrt{2}$ will disappear at the end when we convert to $\mathscr{L}^2(P_n)$ covering numbers.

From the triangle inequality for the $\mathscr{L}^2(P_n)$ seminorm, and the bound $2S_{2n}$ for P_n, deduce for each f and g that

$$|Z(f) - Z(g)| \leq Z(f - g) \leq (2S_{2n}(f - g)^2)^{1/2} = \sqrt{2}\rho(f - g)$$

and similarly for Z'. For f in $\mathscr{F}$ set g equal to the g_j that minimizes $\rho(f - g_j)$. Then

$$\begin{aligned}|Z(f) - Z'(f)| &\leq Z(f - g_j) + |Z(g_j) - Z'(g_j)| + Z'(g_j - f)\\ &\leq 4\delta + |Z(g_j) - Z'(g_j)|,\end{aligned}$$

whence

$$\begin{aligned}\mathbb{P}\left\{\sup_{\mathscr{F}} |Z(f) - Z'(f)| > 6\delta \,\middle|\, \mathbf{X}\right\} &\leq \mathbb{P}\left\{\max_j |Z(g_j) - Z'(g_j)| > 2\delta \,\middle|\, \mathbf{X}\right\}\\ &\leq M \max_j \mathbb{P}\{|Z(g_j) - Z'(g_j)| > 2\delta \,|\, \mathbf{X}\}.\end{aligned}$$

Fix a g with $|g| \leq 1$. Bound $|Z(g) - Z'(g)|$ by

$$|Z(g)^2 - Z'(g)^2|/[Z(g) + Z'(g)]$$

which is less than

$$|P_n g^2 - P_n' g^2|/(2S_{2n}g^2)^{1/2}$$

thanks to the inequality $a^{1/2} + b^{1/2} \geq (a+b)^{1/2}$, for $a, b \geq 0$. Apply Hoeffding's Inequality (Appendix B).

$$\begin{aligned}
&\mathbb{P}\{|Z(g) - Z'(g)| > 2\delta \,|\, \mathbf{X}\} \\
&\quad \leq \mathbb{P}\left\{\left|\sum_{i=1}^{n} \pm[g^2(X_{2i-1}) - g^2(X_{2i})]\right| > 2n\delta(2S_{2n}g^2)^{1/2} \,\middle|\, \mathbf{X}\right\} \\
&\quad \leq 2\exp\left[-16n^2\delta^2 S_{2n}g^2 \Big/ \sum_{i=1}^{n} 4[g^2(X_{2i-1}) - g^2(X_{2i})]^2\right] \\
&\quad \leq 2\exp(-2n\delta^2)
\end{aligned}$$

because the inequality $|g| \leq 1$ implies

$$\sum_{i=1}^{n} [g^2(X_{2i-1}) - g^2(X_{2i})]^2 \leq \sum_{i=1}^{n} g^2(X_{2i-1}) + g^2(X_{2i}) =' 2nS_{2n}g^2.$$

Notice how the $S_{2n}g^2$ factor cancelled. That happened because we symmetrized Z instead of P_n.

Setting g equal to each g_j in turn, we end up with

$$\mathbb{P}\left\{\sup_{\mathscr{F}} |Z(f) - Z'(f)| > 6\delta \,\middle|\, \mathbf{X}\right\} \leq 2N_2(\sqrt{2}\delta, S_{2n}, \mathscr{F}) \exp(-2n\delta^2)$$

Decrease the right-hand side to the trivial upper bound of 1, if necessary, then average out over $\mathbf{X}$:

(35)

$$\mathbb{P}\left\{\sup_{\mathscr{F}} |Z(f) - Z'(f)| > 6\delta\right\} \leq \mathbb{P}[2N_2(\sqrt{2}\delta, S_{2n}, \mathscr{F}) \exp(-2n\delta^2) \wedge 1]$$

The presence of the S_{2n} is aesthetically unpleasing, especially since both δ and $\mathscr{F}$ will always depend on n in applications. Problem 24 allows us to replace it by P_n, at a small cost:

$$\begin{aligned}
&\mathbb{P}[2N_2(\sqrt{2}\delta, S_{2n}, \mathscr{F}) \exp(-2n\delta^2) \wedge 1] \\
&\qquad \leq \mathbb{P}[2N_2(\delta, P_n, \mathscr{F}) N_2(\delta, P_n', \mathscr{F}) \exp(-2n\delta^2) \wedge 1] \\
&\qquad \leq \mathbb{P}[2N_2(\delta, P_n, \mathscr{F}) \exp(-n\delta^2) \wedge 1] \\
&\qquad\quad + \mathbb{P}[N_2(\delta, P_n', \mathscr{F}) \exp(-n\delta^2) \wedge 1]
\end{aligned}$$

by virtue of the inequality $xy \wedge 1 \leq (x \wedge 1) + (y \wedge 1)$ for $x, y \geq 0$. The empirical measures P_n and P_n' have the same distribution; the sum of expectations is less than

$$3\mathbb{P}[N_2(\delta, P_n, \mathscr{F}) \exp(-n\delta^2) \wedge 1].$$

Combine the last bound with (34) and (35) to complete the proof. □

The bounds we have for $\mathscr{L}^1$ covering numbers can be converted into bounds for $\mathscr{L}^2$ covering numbers. For the sake of future reference, consider a class $\mathscr{F}$ with envelope F. Set F equal to a constant to recover the inequalities for a uniformly bounded $\mathscr{F}$.

36 Lemma. *Let $\mathscr{F}$ be a class of functions with strictly positive envelope F, and Q be a probability measure with $QF^2 < \infty$. Define $P(\cdot) = Q(\cdot F^2)/Q(F^2)$ and $\mathscr{G} = \{f/F : f \in \mathscr{F}\}$. Then:*

(i) $N_2(\delta(QF^2)^{1/2}, Q, \mathscr{F}) \leq N_2(\delta, P, \mathscr{G}) \leq N_1(\frac{1}{2}\delta^2, P, \mathscr{G})$;
(ii) *if the class of graphs of functions in $\mathscr{F}$ has only polynomial discrimination then there exist constants A and W, not depending on Q and F, such that $N_2(\delta(QF^2)^{1/2}, Q, \mathscr{F}) \leq A\delta^{-W}$ for $0 < \delta \leq 1$.*

Proof. For every pair of functions f_1, f_2 with $|f_1| \leq F$ and $|f_2| \leq F$,

$$(QF^2)^{-1}Q|f_1 - f_2|^2 = P|f_1/F - f_2/F|^2 \leq 2P|f_1/F - f_2/F|.$$

Assertion (i) follows from these connections between the seminorms used in the definitions of the covering numbers.

The graph of f covers a point (x, t) if and only if the graph of f/F covers $(x, t/F(x))$; the graphs of functions in $\mathscr{G}$ also have polynomial discrimination. Invoke the Approximation Lemma (II.25) for classes with envelope 1.

$$N_1(\tfrac{1}{2}\delta^2, P, \mathscr{G}) \leq A(\tfrac{1}{2}\delta^2)^{-W}.$$

Rechristen $A2^W$ as A and $2W$ as W. □

It is now an easy matter to prove rate of convergence theorems for uniformly bounded classes of functions. As an example, here is a result tailored to classes whose graphs have polynomial discrimination. (Remember that the notation $x_n \gg y_n$ means $x_n/y_n \to \infty$.)

37 Theorem. *For each n, let $\mathscr{F}_n$ be a permissible class of functions whose covering numbers satisfy*

$$\sup_Q N_1(\varepsilon, Q, \mathscr{F}_n) \leq A\varepsilon^{-W} \quad \text{for} \quad 0 < \varepsilon < 1$$

with constants A and W not depending on n. Let $\{\alpha_n\}$ be a non-increasing sequence of positive numbers for which $n\delta_n^2\alpha_n^2 \gg \log n$. If $|f| \leq 1$ and $(Pf^2)^{1/2} \leq \delta_n$ for each f in $\mathscr{F}_n$, then

$$\sup_{\mathscr{F}_n} |P_n f - Pf| \ll \delta_n^2\alpha_n \quad \text{almost surely,}$$

Proof. Fix $\varepsilon > 0$. Set $\varepsilon_n = \varepsilon\delta_n^2\alpha_n$. Because

$$\operatorname{var}(P_n f)/(4\varepsilon_n)^2 \leq (16n\varepsilon^2\delta_n^2\alpha_n^2)^{-1} \ll (\log n)^{-1}$$

the symmetrization inequality (30) holds for all n large enough:

$$\mathbb{P}\left\{\sup_{\mathscr{F}_n} |P_n f - Pf| > 8\varepsilon_n\right\} \leq 4\mathbb{P}\left\{\sup_{\mathscr{F}_n} |P_n^\circ f| > 2\varepsilon_n\right\}.$$

Condition on ξ. Find approximating functions $\{g_j\}$ as in (31). We may assume that each g_j belongs to $\mathscr{F}_n$. (More formally, we could replace g_j by an f_j in $\mathscr{F}_n$ for which $P_n|f_j - g_j| \le \varepsilon$, then replace ε by 2ε throughout the argument.) From (31),

$$\mathbb{P}\left\{\sup_{\mathscr{F}_n} |P_n^\circ f| > 2\varepsilon_n\right\} \le 2A\varepsilon_n^{-W} \exp(-\tfrac{1}{2}n\varepsilon_n^2/64\delta_n^2) + \mathbb{P}\left\{\sup_{\mathscr{F}_n} P_n f^2 > 64\delta_n^2\right\}.$$

The first term on the right-hand side equals

$$2A\varepsilon^{-W} \exp[W \log(1/\delta_n^2\alpha_n) - n\varepsilon^2\delta_n^2\alpha_n^2/128],$$

which decreases faster than every power of n because $\log(1/\delta_n^2\alpha_n)$ increases more slowly than $\log n$, while $n\delta_n^2\alpha_n^2$ increases faster than $\log n$. Lemma 33 bounds the second term by

$$4A(\varepsilon\delta_n^2\alpha_n)^{-W} \exp(-n\delta_n^2)$$

which converges to zero even faster than the first term. An application of the Borel–Cantelli lemma completes the proof. □

Specialized to the case of constant $\{\alpha_n\}$, the constraint placed on $\{\delta_n\}$ by Theorem 37 becomes $\delta_n^2 \gg n^{-1} \log n$. This particular rate pops up in many limit theorems involving smoothing of the empirical measure because (Problem 25) it corresponds to the size of the largest ball containing no sample points.

38 Example. Let P be a probability measure on $\mathbb{R}^d$ having a bounded density $p(\cdot)$ with respect to d-dimensional lebesgue measure. One theoretically attractive method for estimating $p(\cdot)$ is kernel smoothing: convolution of the empirical measure P_n with a convenient density function to smear out the point masses into a continuous distribution. The estimate is

$$p_n(x) = P_n \sigma^{-d} K_{x,\sigma},$$

where

$$K_{x,\sigma}(y) = K[\sigma^{-1}(y - x)]$$

for some density function K on $\mathbb{R}^d$ and a scaling factor σ that depends on n. Note that the σ^{-d} is not part of $K_{x,\sigma}$.

The traditional method for analyzing p_n compares it with the corresponding smoothed form of p,

$$\bar{p}(x) = \mathbb{P}p_n(x) = P\sigma^{-d}K_{x,\sigma}.$$

The difference $p_n - p$ breaks into a sum of a random component, $p_n - \bar{p}$, and a bias component $\bar{p} - p$. The smaller the value of σ, the smaller the bias (Problem 26); the slower σ tends to zero, the faster $p_n - \bar{p}$ converges to zero. These two conflicting demands determine the rate at which $p_n - p$ can tend to zero.

If $p_n - p$ is to converge uniformly to zero, we must not allow σ to decrease too fast. Otherwise we might somewhere be smoothing P_n over a region where it puts too little mass, making $|p_n - p|$ too large. Theorem 37 will let σ^d decrease almost as fast as $n^{-1}\log n$, the best rate possible.

For concreteness take K to be the standard normal density, which enjoys the uniform bound $0 \leq K_{x,\sigma} \leq 1$ for all x and σ (the constant $(2\pi)^{-d/2}$ is too awkward for repeated use.) The class of graphs of all $K_{x,\sigma}$ functions has polynomial discrimination (Problem 28 proves this for a whole class of kernel functions). Assume also that the density $p(\cdot)$ is bounded, say $0 \leq p(\cdot) \leq 1$. In that case

$$\sup_x PK_{x,\sigma}^2 \leq \sigma^d$$

because

$$\begin{aligned} PK_{x,\sigma}^2 &\leq PK_{x,\sigma} \\ &= \int K((y-x)/\sigma)p(y)\,dy \\ &= \sigma^d \int K(t)p(x+\sigma t)\,dt. \end{aligned}$$

Everything is set up for Theorem 37. Put $\alpha_n = 1$ and $\delta_n^2 = \sigma^d$. Provided $\sigma^d \gg n^{-1}\log n$,

$$\sup_x |P_n K_{x,\sigma} - PK_{x,\sigma}| \ll \sigma^d \quad \text{almost surely,}$$

that is,

$$\sup_x |p_n(x) - \bar{p}(x)| \to 0 \quad \text{almost surely.}$$

Smoothness properties of $p(\cdot)$ determine the rate at which the bias term converges to zero (Problem 26). For example, one bounded derivative would give maximum bias of order $O(\sigma)$ in one dimension. We would then want something like $\sigma^3 \gg n^{-1}\log n$ to get a comparable rate of convergence from Theorem 37 for $p_n - \bar{p}$. □

Notes

Uniform strong laws of large numbers have a long history, which is described in the first section of the survey paper by Gaenssler and Stute (1979). Theorem 2 comes from DeHardt (1971), but the idea behind it is much older. Billingsley and Topsøe (1967) and Topsøe (1970, Sections 12 to 15) developed much deeper results for the closely related area of uniformity in weak convergence.

Hartigan (1975) is a good source for information about clustering. Hartigan (1978) and Pollard (1981b, 1982b, 1982c) have proved limit theorems for k-means. Engineers know the method of k-means by the name quantization. The March 1982 *IEEE Transactions on Information Theory* was devoted to the topic. Denby and Martin (1979) proposed the generalized M-estimator of Example 5.

It has long been appreciated that comparison of two independent empirical distribution functions transforms readily into a combinatorial problem. Gnedenko (1968, Section 68), for example, reduced the analysis of two-sample Smirnov statistics to a random walk problem. The method in the text has evolved from the ideas introduced by Vapnik and Cervonenkis (1971). Their method of conditioning turned calculations for a single fixed set into an application of an exponential bound for hypergeometric tail probabilities. Classes with polynomial discrimination are often called VC classes in the literature. The symmetrization described in Section 3 is a well-known method for proving central limit theorems in Banach spaces (Araujo and Giné 1980, Section 2; Giné and Zinn 1984, Section 1). Steele (1975, 1978) discovered subadditivity properties of empirical measures that strengthen the Vapnik–Cervonenkis convergence in probability results to necessary and sufficient conditions for almost sure convergence. Pollard (1981b) introduced the martingale tools and the randomization method described in Problem 12 to rederive Steele's conditions. Theorem 21 and Example 22 are based on Steele (1978); Flemming Topsøe explained to me the $\frac{8}{9}$-trick for convex sets. See Gaenssler and Stute (1979) for more about the history of this example.

The proof of Theorem 16, which is often called the Vapnik–Cervonenkis lemma, is adapted from Steele (1975). Sauer (1970) was the first to publish the inequality in precisely this form, although he suggested that Shelah had also established the result. (I am unable to follow the two papers of Shelah that Sauer cited.) Vapnik and Cervonenkis (1971) proved an insignificantly weaker version of the inequality. Dudley (1978, Section 7) has dug further into the history. Lemma 18 is due to Steele (1975) and Dudley (1978).

The sum of binomial coefficients in Theorem 16 and the randomization method of Problem 12 suggest that a direct probabilistic path might lead to the necessary and sufficient conditions of Theorem 21. Does there exist a set of n independent experiments that can be performed to decide whether a particular subset of a collection of n points can be picked out by a particular class of sets? Or maybe the experiments could be linked in some way. For a class that picks out only subsets with fewer than V points the solution is easy—it lies buried within the proof of Theorem 16.

The notes to Chapter VII will give more background to the concept of covering number.

Vapnik and Cervonenkis (1981) have found necessary and sufficient conditions for uniform almost sure convergence over bounded classes of functions. They worked with $\mathscr{L}^\infty$ and $\mathscr{L}^1$ distances between functions. Giné and Zinn (1984) applied chaining inequalities and gaussian-process methods (see Chapter VII) to deduce $\mathscr{L}^2$ necessary and sufficient conditions. The square-root trick in Lemma 33 comes from Le Cam (1983) via Giné and

Zinn. Kolchinsky (1982) and Pollard (1982c) independently introduced the symmetrization method used in Lemma 33. The Approximation Lemma is due to Dudley (1978), who proved it for classes of sets. The extension to classes of functions was proved ad hoc by Pollard (1982d), using an idea of Le Cam (1983).

The density estimation literature is enormous. Silverman (1978) and Stute (1982b) have found sharp results involving the $n^{-1}\log n$ critical rate. Bertrand-Retali (1974, 1978) proved that $\sigma^d \gg n^{-1}\log n$ is both necessary and sufficient for uniform consistency over the class of all uniformly continuous densities on $\mathbb{R}^d$.

Universal separability was mentioned in passing by Dudley (1978) as a way of avoiding measurability difficulties.

Most of the results in the chapter can be extended to independent, non-identically distributed observations (Alexander 1984a).

Problems

[1] In Example 4 relax the assumption that $W(\cdot, \cdot, P)$ has a unique minimum; assume the function achieves its minimum for each (a, b) in a region D. Prove that the distance from the optimal (a_n, b_n) to D converges to zero almost surely, provided P does not concentrate at a single point. [The condition rules out the possibility of a minimizing pair for $W(\cdot, \cdot, P)$ with one center off at infinity.]

[2] Here is an example of an ad hoc method to force an optimal solution into a restricted region. Suppose an estimator corresponds to the f_n that minimizes $P_n f$ over a class $\mathscr{F}$, and that we want to force f_n into a region $\mathscr{K}$. Write γ_0 for the infimum, assumed finite, of Pf over $\mathscr{F}$. Suppose there exists a positive function $b(\cdot)$ on $\mathscr{F}$ such that, for some $\varepsilon > 0$,

$$b(f) \geq \gamma_0 + \varepsilon \quad \text{for } f \text{ in } \mathscr{F}\backslash\mathscr{K}$$

$$P\left[\inf_{\mathscr{F}\backslash\mathscr{K}} f/b(f)\right] > |\gamma_0|/(|\gamma_0| + \varepsilon).$$

Show that $\liminf \inf_{\mathscr{F}\backslash\mathscr{K}} P_n f > \gamma_0$ almost surely. [Trivial if $\gamma_0 < 0$.] Deduce that f_n belongs to $\mathscr{K}$ eventually (almost surely). Now read the case A consistency proof of Huber (1967). Compare the last part of his argument with our Theorem 3.

[3] Call a class $\mathscr{F}$ of functions universally separable if there exists a countable subclass $\mathscr{F}_0$ such that each f in $\mathscr{F}$ can be written as a pointwise limit of a sequence in $\mathscr{F}_0$. If $\mathscr{F}$ has an envelope F for which $PF < \infty$, prove that universal separability implies measurability of $\|P_n - P\|$.

[4] For any finite-dimensional vector space $\mathscr{G}$ of real functions on S, the class $\mathscr{D}$ of sets of the form $\{g \geq 0\}$, for g in $\mathscr{G}$, is universally separable. [Express each g in $\mathscr{G}$ as a linear combination of some fixed finite collection of non-negative functions. Let $\mathscr{G}_0$ be the countable subclass generated by taking rational coefficients. For each g in $\mathscr{G}$ there exists a sequence $\{g_n\}$ in $\mathscr{G}_0$ for which $g_n \downarrow g$. Show that $\{g_n \geq 0\} \downarrow \{g \geq 0\}$ pointwise.]

[5] The operations in Lemma 15 preserve universal separability.

[6] For a universally separable class $\mathscr{D}$, the quantity V_n defined in Theorem 21 is unchanged if $\mathscr{D}$ is replaced by its countable subclass $\mathscr{D}_0$. Prove that V_n is measurable.

[7] Prove that the class of indicator functions of closed, convex subsets of $\mathbb{R}^d$ is universally separable. [Consider convex hulls of finite sets of points with rational coordinates.]

[8] Theorem 16 informs us that the class of quadrants in $\mathbb{R}^2$ picks out at most $1 + \frac{1}{2}N + \frac{1}{2}N^2$ subsets from any collection of N points. Find a configuration for which this bound is achieved.

[9] For $N \geq 2$ the sum of binomial coefficients singled out by Theorem 16 is bounded by N^V. [Count subsets of $\{1, \ldots, N\}$ containing fewer than V elements by arranging each subset into increasing order then padding it out with enough copies of the largest element to bring it up to a V-tuple. Don't forget the empty set.]

[10] Let M be a subset of $[0, 1]$ that has inner lebesgue measure zero and outer lebesgue measure one (Halmos 1969, Section 16). Define the probability measure μ as the trace of lebesgue measure on M (the measure defined in Theorem A of Halmos (1969), Section 17). Assuming the validity of the continuum hypothesis, put M into a one-to-one correspondence with the space $[0, U)$ of all ordinals less than the first uncountable ordinal U (Kelley 1955, Chapter 0). Define $\mathscr{D}$ as the class of subsets of $[0, 1]$ corresponding to the initial segments $[0, x]$ in $[0, U)$.
(a) Show that $\mathscr{D}$ has linear discrimination. [It shatters no two-point set.]
(b) Equip M^∞ with its product σ-field and product measure μ^∞. Generate observations $\xi_1, \xi_2, \ldots$ on $P = \text{Uniform}(0, 1)$ by taking them as the coordinate projection maps on M^∞. Construct empirical measures $\{P_n\}$ from these observations. Show that $\sup_{\mathscr{D}} |P_n D - PD|$ is identically one.
(c) Repeat the construction with the same $\mathscr{D}$, but replace (M^∞, μ^∞) by a countable product of copies of M^c equipped with the product measure λ^∞, where λ equals the trace of lebesgue measure on M^c. This time $\sup_{\mathscr{D}} |P_n D - PD|$ is identically zero.

[Funny things can happen when $\mathscr{D}$ has measurability problems. Argument adapted from Pollard (1981a) and Durst and Dudley (1981).]

[11] For independent and identically distributed random elements $\{\xi_i\}$, write $\mathscr{E}^n$ for the σ-field generated by all symmetric functions of $\xi_1, \ldots, \xi_N$ as N ranges over $n, n + 1, \ldots$. For a fixed function f, apply the usual reversed martingale argument (Ash 1972, page 310) to show that $\mathbb{P}(P_n f | \mathscr{E}^{n+1}) = P_{n+1} f$. If $P(\sup_{\mathscr{F}} |f|) < \infty$, deduce

$$\mathbb{P}\left(\sup_{\mathscr{F}} |P_n f - Pf| \,|\, \mathscr{E}^{n+1}\right) \geq \sup_{\mathscr{F}} |P_{n+1} f - Pf|$$

for every class of functions $\mathscr{F}$ that makes both suprema measurable.

[12] Here is one way to prove necessity in Theorem 21. Suppose $\|P_n - P\| \to 0$ almost surely. Construct μ_n^+ by placing mass n^{-1} at each ξ_i for which the sign variable σ_i equals $+1$; construct μ_n^- similarly from the remaining ξ_i's. Notice that $P_n^\circ = \mu_n^+ - \mu_n^-$. Let N be the number of sign variables $\sigma_1, \ldots, \sigma_n$ equal to $+1$.
(a) Prove that $(n/N)\mu_n^+$ has the same distributions as P_N. [What if $N = 0$?]
(b) Deduce that both $\|\mu_n^+ - \frac{1}{2}P\| \to 0$ and $\|\mu_n^- - \frac{1}{2}P\| \to 0$, in probability.

(c) Deduce that $\|\mu_n^+ - \mu_n^-\| \to 0$ in probability.

(d) Suppose $\mathscr{D}$ shatters a set F consisting of at least $n\eta$ of the points $\xi_1, \ldots, \xi_n$. Without loss of generality, at least $\frac{1}{2}n\eta$ of the points in F are allocated to μ_n^+. Choose a D to pick out just those points from F. Use independence properties of the $\{\sigma_i\}$ to show that, with high probability, $\mu_n^+(D\backslash F)$ and $\mu_n^-(D\backslash F)$ are nearly equal. [Argue conditionally on P_n and the σ_i for those ξ_i in F.]

(e) Show that $\mu_n^+(D) - \mu_n^-(D) \approx \frac{1}{2}\eta$ with high conditional probability. This contradicts (c).

[13] Rederive the uniform strong law for convex sets (Example 22) by the direct approximation method of Theorem 2.

[14] Let $\mathscr{F}$ be a permissible class with natural envelope $F = \sup_{\mathscr{F}} |f|$. If $\|P_n - P\| \to 0$ almost surely and if $\sup_{\mathscr{F}} |Pf| < \infty$ then $PF < \infty$. [The condition on $\sup_{\mathscr{F}} |Pf|$ excludes trivial cases such as $\mathscr{F}$ consisting of all constant functions. From $\|P_n - P\| < \varepsilon$ and $\|P_{n-1} - P\| < \varepsilon$ deduce $n^{-1}|f(\xi_n) - Pf| < 2\varepsilon$; almost sure convergence implies

$$\mathbb{P}\left\{\sup_{\mathscr{F}} |f(\xi_n) - Pf| \geq n \ \text{ infinitely often}\right\} = 0.$$

Invoke the non-trivial half of the Borel–Cantelli lemma, then replace each ξ_n by ξ_1 to get

$$\infty > \mathbb{P}\left(\sup_{\mathscr{F}} |f(\xi_1) - Pf|\right) \geq \mathbb{P}F(\xi_1) - \text{constant}.$$

Noted by Giné and Zinn (1984).]

[15] Here is an example of how Theorem 24 can go wrong if the envelope F has $PF = \infty$. Let P be the Uniform(0, 1) distribution and let $\mathscr{F}$ be the countable class consisting of the sequence $\{f_i\}$, where $f_i(x) = x^{-2}\{(i+1)^{-1} \leq x < i^{-1}\}$. Show that the graphs have polynomial discrimination and that $Pf_i = 1$ for every i. But $\sup_i P_n f_i \to \infty$ almost surely. [Find an α_n with $n\alpha_n^2 \to 0$, such that $[0, \alpha_n]$ contains at least one observation, for n large enough.]

[16] Let $\mathscr{F}$ be the class of all monotone increasing functions on $\mathbb{R}$ taking values in the range [0, 1]. The class of graphs does not have polynomial discrimination, but it does satisfy the conditions of Theorem 24 for every P. [If $\{x_i\}$ and $\{t_i\}$ are strictly increasing sequences, the graphs can shatter the set of points $(x_1, t_1), \ldots, (x_N, t_N)$.]

[17] For the $\mathscr{F}$ of the previous problem, rewrite $P_n f$ as $\int_0^1 P_n\{f \geq t\}\, dt$. Deduce uniform almost sure convergence from the classical result for intervals. [Suggested by Peter Gaenssler.]

[18] Let $\mathscr{F}$ and $\mathscr{G}$ be classes of functions on S with envelopes F and G. Write $\mathscr{S}$ for the class of all sums $f + g$ with f in $\mathscr{F}$ and g in $\mathscr{G}$. Prove that

$$N_i(\delta Q(F+G), Q, \mathscr{S}) \leq N_i(\delta QF, Q, \mathscr{F})N_i(\delta QG, Q, \mathscr{G}) \quad \text{for} \quad i = 1, 2.$$

[19] A condition involving only covering numbers for P would not be enough to give a uniform strong law of large numbers. Let P be Uniform(0, 1). Let $\mathscr{D}$ consist of all sets that are unions of at most n intervals each with length less than n^{-2}, for $n = 1, 2, \ldots$. Show that $\sup_{\mathscr{D}} |P_n D - PD| = 1$, even though $N_1(\varepsilon, P, \mathscr{D}) < \infty$ for each $\varepsilon > 0$.

[20] Deduce Theorem 2 from Theorem 24.

[21] Under the conditions set down in Example 26, the function $H(\cdot, P)$ achieves its minimum. [If $H(\theta_i, P)$ converges to the infimum as $i \to \infty$, use Fatou's lemma to show that the infimum is achieved at a cluster point of $\{\theta_i\}$; condition (27) rules out cluster points at infinity.] Notice that only left continuity of ϕ is needed for the proof. Find and overcome the extra complications in the argument that would be caused if ϕ were only left-continuous.

[22] This problem assumes familiarity with convexity methods, as described in Section 4.2 of Tong (1980). Suppose that the distribution P of Example 26 has a density $p(\cdot)$ whose high level sets $D_t = \{p \geq t\}$ are convex and symmetric about the origin. Prove that $H(\theta, P)$ has a minimum at $\theta = 0$. [By Fubini's theorem,

$$H(\theta, P) = \iiint \{0 \leq s \leq \phi(|x - \theta|)\}\{0 \leq t \leq p(x)\}\, ds\, dt\, dx$$

$$= \iint \text{volume}[B(\theta, \alpha(s))^c \cap D_t]\, ds\, dt,$$

where $B(\theta, r)$ denotes the closed ball of radius r centered at θ. The volume of $B(\theta, r) \cap D_t$ is maximized at $\theta = 0$.] When is the minimum unique? Show that a multivariate normal with zero means and non-singular variance matrix satisfies the condition for uniqueness.

[23] Suppose $\{\xi_i\}$ are independent, but that the distribution of ξ_i, call it Q_i, changes with i. Write $P^{(n)}$ for the average distribution of the first n observations, $P^{(n)} = n^{-1}(Q_1 + \cdots + Q_n)$. Show for a permissible polynomial class $\mathscr{D}$ that

$$\sup_{\mathscr{D}} |P_n D - P^{(n)} D| \to 0 \quad \text{almost surely.}$$

What difficulties occur in the extension to more general classes of sets, or functions? [Adapt the double-sample symmetrization method of Lemma 33: sample a pair (X_{2i-1}, X_{2i}) from Q_i; use the selection variable τ_i to choose which member of the pair is allocated to P_n, and which to P_n'.]

[24] Show that

$$N_2(\sqrt{2}\delta, \tfrac{1}{2}(Q_1 + Q_2), \mathscr{F}) \leq N_2(\delta, Q_1, \mathscr{F}) N_2(\delta, Q_2, \mathscr{F}).$$

[Let h_1 be the density of Q_1 with respect to $Q_1 + Q_2$. Consider the approximating functions $g_1\{h_1 > \tfrac{1}{2}\} + g_2\{h_1 \leq \tfrac{1}{2}\}$.]

[25] Let P be the uniform distribution on $[0, 1]^2$. For a sample of n independent observations on P show that

$$\mathbb{P}\{\text{some square of area } \alpha_n \text{ contains no observations}\} \to 1$$

if α_n is just slightly smaller than $n^{-1} \log n$. [Break $[0, 1]^2$ into N subsquares each with area slightly less than $n^{-1} \log n$. Set $A_i = \{i\text{th subsquare contains at least one observation}\}$. Show that $\mathbb{P}(A_{i+1} | A_1 \cap \cdots \cap A_i) \leq \mathbb{P}A_{i+1}$. The probability that each of these subsquares contains at least one point is less than $(\mathbb{P}A_1)^N$. Bertrand-Retali (1978).]

[26] In one dimension, write the bias term for the kernel density estimate as

$$\bar{p}(x) - p(x) = \int K(z)[p(x + \sigma z) - p(x)]\, dz.$$

Suppose p has a bounded derivative, and that $\int |z| K(z)\, dz < \infty$. Show that the bias is of order $O(\sigma)$. Generalize to higher dimensions. [If p has higher-order smooth derivatives, and K is replaced by a function orthogonal to low degree polynomials, the bias can be made to depend only on higher powers of σ.]

[27] The graphs of translated kernels $K_{x,\sigma}$ have polynomial discrimination for any K on the real line with bounded variation. [Break K into a difference of monotone functions.]

[28] Let K be a density on $\mathbb{R}^d$ of the form $h(|x|)$, where $h(\cdot)$ is a monotone decreasing function on $[0, \infty)$. Adapt the method of Example 26 to prove that the graphs of the functions $K_{x,\sigma}$ have polynomial discrimination.

[29] Modify the density estimate of Example 38 for distributions on the real line by choosing K as a function of bounded variation for which $\int K(z)\, dz = 0$ and $\int zK(z)\, dz = 1$ and $\int |zK(z)|\, dz < \infty$. Replace p_n by $q_n(x) = \sigma^{-2} P_n K_{x,\sigma}$. Show that $\mathbb{P}q_n(x)$ converges to the derivative of p. How fast can σ tend to zero without destroying the almost sure uniform convergence $\sup_x |q_n(x) - \mathbb{P}q_n(x)| \to 0$?

CHAPTER III

Convergence in Distribution in Euclidean Spaces

. . . which runs through some of the standard methods for proving convergence in distribution of sequences of random vectors, and for proving weak convergence of sequences of probability measures on euclidean spaces. These include: checking convergence for expectations of smooth functions of the random vectors; checking moment conditions for sums of independent random variables (the Central Limit Theorems); checking convergence of characteristic functions (the Continuity Theorems for characteristic functions); and reduction to analogous problems of almost sure convergence via quantile transformations.

III.1. The Definition

Convergence in distribution of a sequence $\{X_n\}$ of real random variables is traditionally defined to mean convergence of distribution functions at each continuity point of the limit distribution function:

$$\mathbb{P}\{X_n \le x\} \to \mathbb{P}\{X \le x\} \quad \text{whenever} \quad \mathbb{P}\{X = x\} = 0.$$

Although convenient for work with order statistics and quantiles, this definition does have some disadvantages. Distribution functions are not well suited to calculations involving sums of independent random variables. The simplest proofs of the Central Limit Theorem, for example, do not directly check pointwise convergence of distribution functions; they show that sequences of characteristic functions, or expectations of other smooth functions of the sums, converge. With the extensions to sequences of random vectors (measurable maps into multidimensional euclidean space $\mathbb{R}^d$), the difficulties multiply. And for random elements of more general spaces not equipped with a partial ordering, even the concept of distribution function disappears. With all this in mind, let us start afresh from an equivalent definition, which lends itself more readily to generalization.

1 Definition. A sequence of random vectors $\{X_n\}$ is said to converge in distribution to a random vector X, written $X_n \rightsquigarrow X$, if $\mathbb{P}f(X_n) \to \mathbb{P}f(X)$ for every f belonging to the class $\mathscr{C}(\mathbb{R}^k)$ of all bounded, continuous, real functions on $\mathbb{R}^k$. □

This notion of convergence does not specify the limit random vector uniquely. If X and Y have the same distribution, that is, if

$$\mathbb{P}\{X \in A\} = \mathbb{P}\{Y \in A\} \quad \text{for each borel set } A,$$

then $X_n \rightsquigarrow X$ means the same as $X_n \rightsquigarrow Y$. This invites slight abuses of notation. For example, it is convenient to write $X_n \rightsquigarrow N(0, 1)$, meaning that the sequence of real random variables $\{X_n\}$ converges in distribution to any random variable having a standard normal distribution. The symbol $N(0, 1)$ stands not only for a particular probability measure on the borel σ-field $\mathscr{B}(\mathbb{R})$ but also for any random variable having that distribution. Similarly, we can avoid much circumlocution by writing, for example,

$$n^{-1/2}[\text{Bin}(n, \tfrac{1}{2}) - \tfrac{1}{2}n] \rightsquigarrow N(0, \tfrac{1}{4}),$$

instead of:

> if X_n has a binomial distribution with parameters n and $\frac{1}{2}$, and X has a normal distribution with mean zero and variance $\frac{1}{4}$, then $n^{-1/2}(X_n - \frac{1}{2}n) \rightsquigarrow X$.

In general, for probability measures on the borel σ-field $\mathscr{B}(\mathbb{R}^k)$, define $P_n \rightsquigarrow P$ to mean convergence in distribution for random vectors having these distributions. This definition is equivalent to the requirement: $P_n f \to Pf$ for every f in $\mathscr{C}(\mathbb{R}^k)$. Most authors call this weak convergence.

III.2. The Continuous Mapping Theorem

Suppose $X_n \rightsquigarrow X$, as random vectors taking values in $\mathbb{R}^k$, and let H be a measurable map from $\mathbb{R}^k$ into $\mathbb{R}^s$. Does it follow that $HX_n \rightsquigarrow HX$? That is, does $\mathbb{P}f(HX_n) \to \mathbb{P}f(HX)$ for every f in $\mathscr{C}(\mathbb{R}^s)$? If H were continuous, $f \circ H$ would belong to $\mathscr{C}(\mathbb{R}^k)$ for every f in $\mathscr{C}(\mathbb{R}^s)$. The result would be trivially true. The convergence $HX_n \rightsquigarrow HX$ also holds under a slightly weaker assumption: it suffices that H be continuous at almost all points of the range of X. This will follow as a simple corollary to the next lemma.

2 Convergence Lemma. *Let h be a bounded, measurable, real-valued function on $\mathbb{R}^k$, continuous at each point of a measurable set C.*

(i) *Let $\{X_n\}$ be a sequence of random vectors converging in distribution to X. If $\mathbb{P}\{X \in C\} = 1$, then $\mathbb{P}h(X_n) \to \mathbb{P}h(X)$.*
(ii) *Let $\{P_n\}$ be a sequence of probability measures converging weakly to P. If $P(C) = 1$, then $P_n h \to Ph$.*

Proof. As the two assertions are similar, we need only prove (ii). Consider any increasing sequence $\{f_i\}$ of bounded, continuous functions for which $f_i \leq h$ everywhere and $f_i \uparrow h$ at each point of C. Accept for the moment that such a sequence exists. Then weak convergence of $\{P_n\}$ to P implies that

$$P_n f_i \to P f_i \quad \text{for each fixed } i. \tag{3}$$

On the left-hand side bound f_i by h.

$$\liminf P_n h \geq Pf_i \quad \text{for each fixed } i.$$

Invoke monotone convergence as i tends to infinity on the right-hand side.

$$\liminf P_n h \geq Ph. \tag{4}$$

The companion inequality obtained by substituting $-h$ for h combines with (4) to complete the proof.

Now to construct the functions $\{f_i\}$. They must be chosen from the family

$$\mathscr{F} = \{f \in \mathscr{C}(\mathbb{R}^k): f \leq h\}.$$

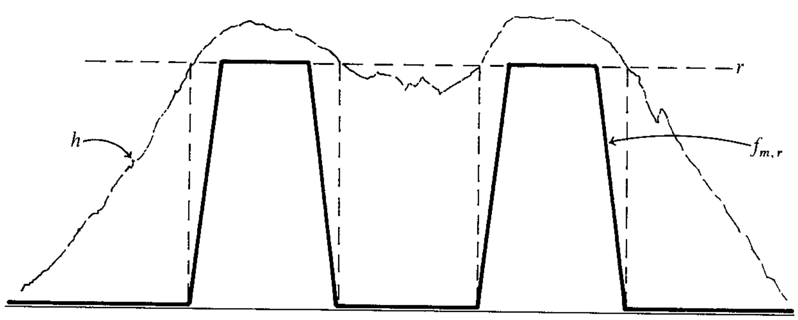
If we can find a countable subfamily of $\mathscr{F}$, say $\{g_1, g_2, \ldots\}$, whose pointwise supremum equals h at each point of C, then setting $f_i = \max\{g_1, \ldots, g_i\}$ will do the trick.

Without loss of generality suppose $h > 0$. (A constant could be added to h to achieve this.) For each subset A of $\mathbb{R}^k$ define the distance function $d(\cdot, A)$ by

$$d(x, A) = \inf\{|x - y| : y \in A\}.$$

It is a continuous function of x, for each fixed A. For positive integral m and positive rational r define

$$f_{m,r}(x) = r \wedge md(x, \{h \leq r\}).$$



Each $f_{m,r}$ is bounded and continuous; it is at most r if $h(x) > r$; it takes the value zero if $h(x) \leq r$: it belongs to $\mathscr{F}$. Given a point x in C and an $\varepsilon > 0$, choose a positive rational number r with $h(x) - \varepsilon < r < h(x)$. Continuity of h at x keeps its value greater than r in some neighborhood of x. Consequently, $d(x, \{h \leq r\}) > 0$ and $f_{m,r}(x) = r > h(x) - \varepsilon$ for all m large enough. □

Weak convergence of $\{P_n\}$ was needed only to establish the convergence (3) for the functions $\{f_i\}$. These functions were, however, not just continuous, but uniformly continuous. The functions from which they were constructed even satisfied a Lipschitz condition:

$$|f_{m,r}(x) - f_{m,r}(y)| \leq m|x - y|.$$

Thus the lemma could have been proved using only convergence of expectations of bounded, uniformly continuous functions of the random vectors. In particular, such a requirement would imply the convergence $P_n h \to Ph$ for each h in $\mathscr{C}(\mathbb{R}^k)$.

5 Corollary. *If $P_n f \to Pf$ for every bounded, uniformly continuous f then $P_n \rightsquigarrow P$. (And similarly for convergence in distribution of random vectors.)* □

The lemma also provides an answer to the question asked at the start of the section.

6 Continuous Mapping Theorem. *Let H be a measurable mapping from $\mathbb{R}^k$ into $\mathbb{R}^s$. Write C for the set of points in $\mathbb{R}^k$ at which H is continuous. If a sequence $\{X_n\}$ of random vectors taking values in $\mathbb{R}^k$ converges in distribution to a random vector X for which $\mathbb{P}\{X \in C\} = 1$, then $HX_n \rightsquigarrow HX$.*

PROOF. For each fixed f in $\mathscr{C}(\mathbb{R}^s)$, the bounded function $f \circ H$ is continuous at all points of C. □

Some authors seem to regard this result as trivial and obvious; they scarcely notice that, at least implicitly, they make use of it for many applications. It is better to recognize these covert appeals to the Continuous Mapping Theorem. That way the more general form of the result in Chapter IV will come as no surprise.

7 Example. If the real random variables $\{X_n\}$ converge in distribution to X then $\mathbb{P}\{X_n \le x\} \to \mathbb{P}\{X \le x\}$ at each x for which $\mathbb{P}\{X = x\} = 0$. That is, the sequence converges at each continuity point x of the distribution function of X. This holds because x is the only point of discontinuity of (the indicator function of) the set $(-\infty, x]$. Problem 1 shows you how to go the other way, from pointwise convergence of distribution functions to convergence in distribution of the random variables.

The same result is true in higher dimensions if the inequalities $X_n \le x$ and $X \le x$ are taken componentwise and $(-\infty, x]$ is interpreted as a multidimensional orthant with vertex at x. Continuity of the multidimensional distribution function of X at x requires that X lands on the boundary of $(-\infty, x]$ with zero probability. □

8 Example. Consider the multinomial distribution obtained by independent placement of n objects into k disjoint cells. Write $F_n = (F_{n1}, \ldots, F_{nk})$ for the column vector of observed frequencies—cell i receives F_{ni} objects—and $p = (p_1, \ldots, p_k)$ for the column vector of cell probabilities. Pearson's chi-square statistic is

$$Z_n = \sum_{i=1}^{k} (F_{ni} - np_i)^2 / np_i.$$

Write Z_n as a function of a standardized column vector, by setting

$$X_n = n^{-1/2}(F_n - np)$$

and

$$Z_n = X_n' \Delta^{-2} X_n,$$

where Δ denotes the diagonal matrix with $\sqrt{p_i}$ as its ith diagonal element. By the Multivariate Central Limit Theorem, which will be proved later in this chapter (Theorem 30), the random vectors $\{X_n\}$ converge to a $N(0, V)$ distribution whose variance matrix V has (i, j)th element $p_i - p_i^2$ if $i = j$ and $-p_i p_j$ otherwise. Manufacture a random vector with this limit distribution by applying a linear transformation to a column vector W of independent $N(0, 1)$ random variables: $X_n \rightsquigarrow \Delta(I_k - uu')W$, where u denotes the unit column vector $(\sqrt{p_1}, \ldots, \sqrt{p_k})$.

The mapping H from $\mathbb{R}^k$ into $\mathbb{R}$ defined by

$$Hx = x'\Delta^{-2}x = |\Delta^{-1}x|^2$$

is continuous. Apply the Continuous Mapping Theorem.

$$\begin{aligned} Z_n &= HX_n \\ &\rightsquigarrow H\Delta(I_k - uu')W \\ &= |(I_k - uu')W|^2 \\ &\sim \chi^2_{k-1}, \end{aligned}$$

because $I_k - uu'$ represents the projection orthogonal to the unit vector u. (The squared length of the projection of W onto any $(k-1)$-dimensional subspace has a chi-square distribution with $k - 1$ degrees of freedom.) □

9 Example. Suppose $\xi_1, \xi_2, \ldots$ are independent random variables each with a Uniform(0, 1) distribution. Neyman (1937) developed a goodness-of-fit test whose asymptotic properties depended on the behavior of the statistics

$$G_n = n^{-1} \sum_{j=1}^{k} \left[\sum_{i=1}^{n} \pi_j(\xi_i) \right]^2,$$

where $\pi_0, \pi_1, \ldots$ are given polynomials defined on $[0, 1]$, with the orthonormality property:

$$\int_0^1 \pi_i(y)\pi_j(y)\, dy = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

Explicitly, $\pi_0(y) = 1$, $\pi_1(y) = \sqrt{12}(y - \frac{1}{2})$, $\pi_2(y) = \sqrt{5}[6(y - \frac{1}{2})^2 - \frac{1}{2}]$, and so on. Define random column vectors

$$X_i = (\pi_1(\xi_i), \ldots, \pi_k(\xi_i)) \quad \text{for} \quad i = 1, 2, \ldots.$$

The statistic G_n can then be written as

$$\left| n^{-1/2} \sum_{i=1}^{n} X_i \right|^2.$$

The Multivariate Central Limit Theorem (Theorem 30) and the orthonormality properties of the polynomials ensure that

$$n^{-1/2}(X_1 + \cdots + X_n) \rightsquigarrow N(0, I_k).$$

The Continuous Mapping Theorem (applied to which map?) allows us to deduce that $G_n \rightsquigarrow \chi_k^2$. Neyman used this to determine the approximate critical region for his test. □

III.3. Expectations of Smooth Functions

There are two sorts of perturbation of a random vector X that don't affect the expectation $\mathbb{P}f(X)$ of a smooth, bounded function of X too greatly: changes, however gross, that occur with only small probability; and changes that might occur with high probability but which alter X by only small amounts. These effects are easy to quantify when the smooth function f is uniformly continuous. Suppose $|f(x) - f(z)| < \varepsilon$ whenever $|x - z| < \delta$. Write $\|f\|$ for the supremum of $f(\cdot)$. Then for any random vectors X and Y, whether dependent or not,

$$\begin{aligned}(10)\qquad & |\mathbb{P}f(X) - \mathbb{P}f(X+Y)| \\ &\leq \mathbb{P}\{|Y| < \delta\}|f(X) - f(X+Y)| \\ &\quad + \mathbb{P}\{|Y| \geq \delta\}(|f(X)| + |f(X+Y)|) \\ &\leq \varepsilon + 2\|f\|\mathbb{P}\{|Y| \geq \delta\}.\end{aligned}$$

The inequality lets us deduce convergence in distribution of a sequence of random vectors from convergence of slightly perturbed sequences.

11 Lemma. *Let $\{X_n\}$, X and Y be random vectors for which $X_n + \sigma Y \rightsquigarrow X + \sigma Y$ for each fixed positive σ. Then $X_n \rightsquigarrow X$.*

PROOF. Remember (Corollary 5) we have only to check that $\mathbb{P}f(X_n) \to \mathbb{P}f(X)$ for each bounded, uniformly continuous f. Apply inequality (10) with X replaced by X_n and Y replaced by σY.

$$\sup_n |\mathbb{P}f(X_n) - \mathbb{P}f(X_n + \sigma Y)| \leq \varepsilon + 2\|f\|\mathbb{P}\{|Y| \geq \delta\sigma^{-1}\}.$$

Similarly

$$|\mathbb{P}f(X) - \mathbb{P}f(X + \sigma Y)| \leq \varepsilon + 2\|f\|\mathbb{P}\{|Y| \geq \delta\sigma^{-1}\}.$$

Choose σ small enough to make both right-hand sides less than 2ε, then invoke the known convergence of $\{\mathbb{P}f(X_n + \sigma Y)\}$ to $\mathbb{P}f(X + \sigma Y)$ to deduce that $\limsup |\mathbb{P}f(X_n) - \mathbb{P}f(X)| \leq 4\varepsilon$. □

Now, instead of thinking of the σY as a perturbation of the random vectors, treat it as a means for smoothing the function f. This can be arranged by choosing independently of X and the $\{X_n\}$ a random vector Y having a smooth density function with respect to lebesgue measure—for convenience, take Y to have a $N(0, I_k)$ distribution. Integrate out first with respect to the distribution of Y then with respect to the distribution of X.

$$\mathbb{P}f(X + \sigma Y) = \mathbb{P}f_\sigma(X),$$

where

$$\begin{aligned} f_\sigma(x) &= \int (2\pi)^{-k/2} f(x + \sigma y) \exp(-\tfrac{1}{2}|y|^2)\, dy \\ &= \int (2\pi\sigma^2)^{-k/2} f(z) \exp(-\tfrac{1}{2}|z - x|^2/\sigma^2)\, dz. \end{aligned}$$

The function f has been smoothed by convolution. Dominated convergence justifies repeated differentiation under the last integral sign to prove that f_σ belongs to the class $\mathscr{C}^\infty(\mathbb{R}^k)$ of all bounded real functions on $\mathbb{R}^k$ having bounded, continuous partial derivatives of all orders.

12 Theorem. *If* $\mathbb{P}f(X_n) \to \mathbb{P}f(X)$ *for every* f *in* $\mathscr{C}^\infty(\mathbb{R}^k)$ *then* $X_n \rightsquigarrow X$.

PROOF. Convergence holds for every f_σ produced by convolution smoothing. Apply Lemma 11. □

For the remainder of the section assume that $k = 1$. That is, consider only real random variables. As the results of Section 5 will show, no great generality will be lost thereby—a trick with multidimensional characteristic functions will reduce problems of convergence of random vectors to their one-dimensional analogues.

For expectations of smooth functions of X, the effect of small perturbations can be expressed in terms of moments by applying Taylor's theorem. Suppose f belongs to $\mathscr{C}^\infty(\mathbb{R}^k)$. Then, ignoring the niceties of convergence, we can write

$$f(x + y) = f(x) + yf'(x) + \tfrac{1}{2}y^2 f''(x) + \cdots.$$

Suppose the random variable X is incremented by an independent amount Y. Then, again ignoring problems of convergence and finiteness, deduce

$$\text{(13)} \qquad \mathbb{P}f(X + Y) = \mathbb{P}f(X) + \mathbb{P}(Y)\mathbb{P}f'(X) + \tfrac{1}{2}\mathbb{P}(Y^2)\mathbb{P}f''(X) + \cdots.$$

Try to mimic the effect of the increment Y by a different increment W, also independent of X. As long as $\mathbb{P}(Y) = \mathbb{P}(W)$ and $\mathbb{P}(Y^2) = \mathbb{P}(W^2)$, the expectations $\mathbb{P}f(X + Y)$ and $\mathbb{P}f(X + W)$ should differ only by terms involving third or higher moments of Y and W. These higher-order terms should amount to very little provided both Y and W are small; the effect of substituting W for Y should be small in that case.

This method of substitution can be applied repeatedly for a random variable Z made up of a lot of little independent increments. We can replace the increments one after another by new independent random variables. If at each substitution we match up the first and second moments, as above, the overall effect on $\mathbb{P}f(Z)$ should involve only a sum of quantities of third or higher order. In the next section this approach, with normally distributed replacement increments, will establish the Liapounoff and Lindeberg forms of the Central Limit Theorem.

To make these approximation ideas more precise we need to bound the remainder terms in the informal expansion (13). Because only the first two moments of the increments are to be matched, a Taylor expansion to quadratic terms will suffice. Existence of third derivatives for f will help to control the error terms.

Assume f belongs to the class $\mathscr{C}^3(\mathbb{R})$ of all bounded real functions on $\mathbb{R}$ having bounded continuous derivatives up to third order. Then the remainder term in the Taylor expansion

$$(14) \qquad f(x+y) = f(x) + yf'(x) + \tfrac{1}{2}y^2f''(x) + R(x, y)$$

can be expressed as

$$R(x, y) = \tfrac{1}{6}y^3f'''(x + \theta_1 y)$$

with θ_1 (depending on x and y) between 0 and 1. Write $\|f'''\|$ for the supremum of $|f'''(\cdot)|$. Then

$$(15) \qquad |R(x, y)| \le \tfrac{1}{6}\|f'''\|\,|y|^3,$$

Set C equal to $\frac{1}{6}\|f'''\|$. Then from (14) and (15),

$$\begin{aligned}|\mathbb{P}f(X+Y) - \mathbb{P}f(X) - \mathbb{P}(Y)\mathbb{P}f'(X) - \tfrac{1}{2}\mathbb{P}(Y^2)\mathbb{P}f''(X)| &\\ \le \mathbb{P}|R(X, Y)| &\\ \le C\mathbb{P}(|Y|^3). &\end{aligned}$$

Apply the same argument with Y replaced by the increment W, which is also independent of X. Because $\mathbb{P}(Y) = \mathbb{P}(W)$ and $\mathbb{P}(Y^2) = \mathbb{P}(W^2)$, when the resulting expansion for $\mathbb{P}f(X + W)$ is subtracted from the expansion for $\mathbb{P}f(X + Y)$ most of the terms cancel out, leaving,

$$(16) \qquad \begin{aligned}|\mathbb{P}f(X+Y) - \mathbb{P}f(X+W)| &\le \mathbb{P}|R(X, Y)| + \mathbb{P}|R(X, W)| \\ &\le C\mathbb{P}(|Y|^3) + C\mathbb{P}(|W|^3).\end{aligned}$$

This inequality is sharp enough for the proof of a limit theorem for sums of independent random variables with third moments.

III.4. The Central Limit Theorem

A sum of a large number of small, independent random variables is approximately normally distributed—that is roughly what the Central Limit Theorem asserts. The rigorous formulations of the theorem set forth conditions for convergence in distribution of sums of independent random variables to a standard normal distribution. We shall prove two versions of the theorem.

To begin with, consider a sum $Z = \xi_1 + \cdots + \xi_k$ of independent random variables with finite third moments. Write σ_j^2 for $\mathbb{P}\xi_j^2$. Standardize, if necessary, to ensure that $\mathbb{P}\xi_j = 0$ for each j and $\sigma_1^2 + \cdots + \sigma_k^2 = 1$. Independently of the $\{\xi_j\}$, choose independent $N(0, \sigma_j^2)$-distributed random variables $\{\eta_j\}$, for $j = 1, \ldots, k$. Start replacing the $\{\xi_j\}$ by the $\{\eta_j\}$, beginning at the right-hand end. Define

$$S_j = \xi_1 + \cdots + \xi_{j-1} + \eta_{j+1} + \cdots + \eta_k.$$

Notice that $S_k + \xi_k = Z$ and that $S_1 + \eta_1$ has a $N(0, 1)$ distribution.

Choose a $\mathscr{C}^3(\mathbb{R})$ function f, as in Section 3. Theorem 12 has shown that convergence for expectations of infinitely differentiable functions of random variables is enough to establish convergence in distribution; convergence for functions in $\mathscr{C}^3(\mathbb{R})$ is more than enough. We need to show that $\mathbb{P}f(Z)$ is close to $\mathbb{P}f(N(0, 1))$.

Apply inequality (16) with $X = S_j$, $Y = \xi_j$, and $W = \eta_j$. Because $S_j + \xi_j = S_{j+1} + \eta_{j+1}$ for $j = 1, \ldots, k - 1$,

$$(17)\qquad \begin{aligned} |\mathbb{P}f(Z) - \mathbb{P}f(N(0, 1))| &\le \sum_{j=1}^{k} |\mathbb{P}f(S_j + \xi_j) - \mathbb{P}f(S_j + \eta_j)| \\ &\le \sum_{j=1}^{k} \mathbb{P}|R(S_j, \xi_j)| + \mathbb{P}|R(S_j, \eta_j)| \\ &\le C \sum_{j=1}^{k} \mathbb{P}|\xi_j|^3 + C \sum_{j=1}^{k} \mathbb{P}|\eta_j|^3. \end{aligned}$$

With this bound in hand, the proof of the first version of the Central Limit Theorem presents no difficulty.

18 Liapounoff Central Limit Theorem. *For each n let Z_n be a sum of independent random variables $\xi_{n1}, \xi_{n2}, \ldots, \xi_{nk(n)}$ with zero means and variances that sum to one. If the Liapounoff condition,*

$$(19)\qquad \sum_{j=1}^{k(n)} \mathbb{P}|\xi_{nj}|^3 \to 0 \quad \text{as } n \to \infty,$$

is satisfied, then $Z_n \rightsquigarrow N(0, 1)$.

PROOF. Choose and fix an f in $\mathscr{C}^3(\mathbb{R})$. Check that $\mathbb{P}f(Z_n) \to \mathbb{P}f(N(0, 1))$. The replacement normal random variables are denoted by $\eta_{n1}, \ldots, \eta_{nk(n)}$. The sum $\eta_{n1} + \cdots + \eta_{nk(n)}$ has a $N(0, 1)$ distribution. Write σ_{nj}^2 for the variance of ξ_{nj}, and λ_n for the sum on the left-hand side of (19). With subscripting n's attached, the bound (17) becomes

$$|\mathbb{P}f(Z_n) - \mathbb{P}f(N(0, 1))| \le C\lambda_n + C \sum_{j=1}^{k(n)} \sigma_{nj}^3 \mathbb{P}|N(0, 1)|^3.$$

By Jensen's inequality, $\sigma_{nj}^3 = (\mathbb{P}\xi_{nj}^2)^{3/2} \le \mathbb{P}|\xi_{nj}|^3$, which shows that the sum contributed by the normal increments is less than $\mathbb{P}|N(0,1)|^3\lambda_n$. Two calls upon (19) as $n \to \infty$ complete the proof. □

The Liapounoff condition (19) imposes the unnecessary constraint of finite third moments upon the summands. Liapounoff himself was able to weaken this to a condition on the $(2+\delta)$th moments, for some $\delta > 0$. The remainder term $R(x, y)$ in the Taylor expansion (14) does not increase as fast as $|y|^{2+\delta}$ though:

$$(20) \qquad \begin{aligned} |R(x,y)| &= |[f(x+y) - f(x) - yf'(x)] - \tfrac{1}{2}y^2 f''(x)| \\ &= |\tfrac{1}{2}y^2 f''(x+\theta_2 y) - \tfrac{1}{2}y^2 f''(x)| \\ &\le \|f''\|\,|y|^2 \quad \text{for all } x \text{ and } y. \end{aligned}$$

The new bound improves upon (15) for large $|y|$, but not for small $|y|$. To have it both ways, apply (15) if $|y| < \varepsilon$ and (20) otherwise. Increase C to the maximum of $\frac{1}{6}\|f'''\|$ and $\|f''\|$. Then the bound on the expected remainder is sharpened to

$$(21) \qquad \begin{aligned} \mathbb{P}|R(X,Y)| &\le \mathbb{P}C|Y|^3\{|Y| < \varepsilon\} + \mathbb{P}C|Y|^2\{|Y| \ge \varepsilon\} \\ &\le \varepsilon C\mathbb{P}(Y^2) + C\mathbb{P}Y^2\{|Y| \ge \varepsilon\}. \end{aligned}$$

22 Lindeberg Central Limit Theorem. *For each n let Z_n be a sum of independent random variables $\xi_{n1}, \xi_{n2}, \ldots, \xi_{nk(n)}$ with zero means and variances that sum to one. If, for each fixed $\varepsilon > 0$, the Lindeberg condition*

$$(23) \qquad \sum_{j=1}^{k(n)} \mathbb{P}\xi_{nj}^2\{|\xi_{nj}| \ge \varepsilon\} \to 0 \quad \text{as} \quad n \to \infty$$

is satisfied, then $Z_n \rightsquigarrow N(0, 1)$.

Proof. Use the same notation as in the proof of Theorem 18. Denote the left-hand side of (23) by $L_n(\varepsilon)$. Stop one line earlier than before in the application of (17):

$$|\mathbb{P}f(Z_n) - \mathbb{P}f(N(0,1))| \le \sum_{j=1}^{k(n)} \mathbb{P}|R(S_{nj}, \xi_{nj})| + \sum_{j=1}^{k(n)} \mathbb{P}|R(S_{nj}, \eta_{nj})|.$$

From (21), the first sum is less than

$$C\sum_{j=1}^{k(n)} [\varepsilon\mathbb{P}\xi_{nj}^2 + \mathbb{P}\xi_{nj}^2\{|\xi_{nj}| \ge \varepsilon\}]$$

which equals $C\varepsilon + CL_n(\varepsilon)$ because the variance sum to one. For the second sum retain the bound

$$C\sum_{j=1}^{k(n)} \sigma_{nj}^3 \mathbb{P}|N(0,1)|^3$$

but, in place of Jensen's inequality, use:

$$\begin{aligned}\left[\sum_{j=1}^{k(n)} \sigma_{nj}^3\right]^2 &\le \left(\max_j \sigma_{nj}\right)^2 \left[\sum_{j=1}^{k(n)} \sigma_{nj}^2\right]^2 \\ &= \max_j \sigma_{nj}^2 \\ &\le \max_j [\varepsilon^2 + \mathbb{P}\xi_{nj}^2\{|\xi_{nj}| \ge \varepsilon\}] \\ &\le \varepsilon^2 + L_n(\varepsilon). \qquad \square\end{aligned}$$

The strange-looking Lindeberg condition is not as artificial as one might think. For example, consider the standardized summands for a sequence $\{Y_n\}$ of independent, identically distributed random variables with zero means and unit variances: $\xi_{nj} = n^{-1/2}Y_j$ for $j = 1, 2, \ldots, n$. In this case

$$L_n(\varepsilon) = n\mathbb{P}n^{-1}Y_1^2\{|Y_1| \ge n^{1/2}\varepsilon\},$$

which tends to zero as n tends to infinity, by dominated convergence, because Y_1^2 is integrable. It is even more comforting to know that the Lindeberg condition comes very close to being a necessary condition (Feller 1971, Section XV.6) for the Central Limit Theorem to hold.

24 Example (A Central Limit Theorem for the Sample Median). Let M_n be the median of a sample $\{Y_1, Y_2, \ldots, Y_n\}$ from a distribution function G with median M. For simplicity, suppose the sample size n is odd, $n = 2N + 1$, so that M_n equals the $(N + 1)$st order statistic of the sample. Suppose also that the underlying distribution function G has a positive derivative γ at its median. To prove that

$$n^{1/2}(M_n - M) \rightsquigarrow N(0, \tfrac{1}{4}\gamma^{-2}),$$

it suffices to check pointwise convergence of the distribution functions.

$$\begin{aligned}(25)\qquad &\mathbb{P}\{n^{1/2}(M_n - M) \le x\} \\ &= \mathbb{P}\{M_n \le M + n^{-1/2}x\} \\ &= \mathbb{P}\ \{\text{at least } N + 1 \text{ observations} \le M + n^{-1/2}x\} \\ &= \mathbb{P}\left[N + 1 \le \sum_{j=1}^{n} \{Y_j \le M + n^{-1/2}x\}\right].\end{aligned}$$

Define

$$\begin{aligned}p_n &= \mathbb{P}\{Y_j \le M + n^{-1/2}x\} = G(M + n^{-1/2}x), \\ b_n &= [np_n(1 - p_n)]^{1/2}, \\ \xi_{nj} &= [\{Y_j \le M + n^{-1/2}x\} - p_n]/b_n, \\ t_n &= (N + 1 - np_n)/b_n.\end{aligned}$$

Check that $\mathbb{P}\xi_{nj} = 0$ and $\mathbb{P}\xi_{n1}^2 + \cdots + \mathbb{P}\xi_{nn}^2 = 1$. Continuity of G at M gives $p_n \to \frac{1}{2}$; differentiability of G at M gives $t_n \to -2x\gamma$. The last probability in (25) becomes

$$\mathbb{P}\left\{t_n \leq \sum_{j=1}^{n} \xi_{nj}\right\}. \tag{26}$$

As each $|\xi_{nj}|$ is less than b_n^{-1}, which converges to zero, both the Liapounoff and Lindeberg conditions are easy to check.

$$\sum_{j=1}^{n} \mathbb{P}|\xi_{nj}|^3 \leq b_n^{-1} \sum_{j=1}^{n} \mathbb{P}\xi_{nj}^2 = b_n^{-1}$$

$$\sum_{j=1}^{n} \mathbb{P}\xi_{nj}^2\{|\xi_{nj}| \geq \varepsilon\} = 0 \quad \text{if} \quad \varepsilon > b_n^{-1}.$$

By either of two routes,

$$\sum_{j=1}^{n} \xi_{nj} \rightsquigarrow N(0, 1).$$

Problem 13 shows that substitution of $-2x\gamma$ for t_n in (26) leads to the correct result:

$$\mathbb{P}\{n^{1/2}(M_n - M) \leq x\} \to \mathbb{P}\{N(0, 1) \geq -2x\gamma\}. \qquad \square$$

III.5. Characteristic Functions

The smoothing argument of Section 3 showed that only expectations of $\mathscr{C}^\infty(\mathbb{R}^k)$ functions need be considered when checking convergence in distribution of random vectors. In this section, further analysis of the method of smoothing will show that the class of functions can be narrowed a little more: it suffices to check the convergence

$$\mathbb{P}\exp(i\gamma \cdot X_n) \to \mathbb{P}\exp(i\gamma \cdot X)$$

for each fixed vector γ in $\mathbb{R}^k$. That is, pointwise convergence of characteristic functions implies convergence in distribution.

Start again from the convolution expression

$$\mathbb{P}f(X + \sigma Y) = \mathbb{P}\int (2\pi\sigma^2)^{-k/2} f(z) \exp(-\tfrac{1}{2}|z - X|^2/\sigma^2)\, dz,$$

derived under the assumption that Y has a $N(0, I_k)$ distribution independent of X. This holds for every bounded measurable f. Reverse the order of integration (Fubini) on the right-hand side to show that

$$\mathbb{P}f(X + \sigma Y) = \int f(z)J(z)\,dz,$$

where

$$J(z) = \mathbb{P}(2\pi\sigma^2)^{-k/2} \exp(-\tfrac{1}{2}|z - X|^2/\sigma^2). \tag{27}$$

The distribution of $X + \sigma Y$ has density function $J(\cdot)$ with respect to lebesgue measure. The dependence on σ, which will remain fixed for most of the argument, need not be made explicit.

The integrand appearing on the right-hand side of (27) comes from the density function of σY evaluated at $(z - X)$. It is also proportional to the characteristic function of Y evaluated at $(z - X)/\sigma$:

$$\exp(-\tfrac{1}{2}|z - X|^2/\sigma^2) = \int (2\pi)^{-k/2} \exp(iy\cdot(z - X)/\sigma - \tfrac{1}{2}|y|^2)\,dy.$$

Invoke Fubini's theorem.

$$\begin{aligned} J(z) &= \int (2\pi\sigma)^{-k}\mathbb{P} \exp(-iy\cdot X/\sigma) \exp(iy\cdot z/\sigma - \tfrac{1}{2}|y|^2)\,dy \\ &= \int (2\pi\sigma)^{-k}\phi(-y/\sigma) \exp(iy\cdot z/\sigma - \tfrac{1}{2}|y|^2)\,dy, \end{aligned} \tag{28}$$

where $\phi(\cdot)$ denotes the characteristic function of X.

Formula (28) shows that the characteristic function of X uniquely determines the distribution of $X + \sigma Y$ for each fixed σ. Because $X + \sigma Y$ converges almost surely to X as σ tends to zero, this proves that the characteristic function also uniquely determines the distribution of X. A stronger result can be proved by exploiting continuity properties of the dependence of J on ϕ.

29 Continuity Theorem. *A sequence of random vectors $\{X_n\}$ converges in distribution to a random vector X if and only if the corresponding sequence of characteristic functions converges pointwise:*

$$\mathbb{P} \exp(i\gamma\cdot X_n) \to \mathbb{P} \exp(i\gamma\cdot X)$$

for each fixed γ in $\mathbb{R}^k$.

PROOF. Prove necessity by splitting $\exp(ix\cdot\gamma)$ into its real and imaginary parts, both of which define functions in $\mathscr{C}(\mathbb{R}^k)$.

For the proof of sufficiency, denote the characteristic function of X_n by ϕ_n, and the density function of $X_n + \sigma Y$ by J_n. The domain of integration is $\mathbb{R}^k$. For fixed f and σ,

$$
\begin{aligned}
|\mathbb{P}f(X_n + \sigma Y) &- \mathbb{P}f(X + \sigma Y) \\
&= \left| \int f(z)J_n(z)\,dz - \int f(z)J(z)\,dz \right| \\
&\leq \|f\| \int |J_n(z) - J(z)|\,dz \\
&= \|f\| \left[\int (J(z) - J_n(z))^+\,dz + \int (J(z) - J_n(z))^-\,dz \right] \\
&= 2\|f\| \int (J(z) - J_n(z))^+\,dz,
\end{aligned}
$$

because

$$
\begin{aligned}
0 &= 1 - 1 \\
&= \int (J(z) - J_n(z))\,dz \\
&= \int (J(z) - J_n(z))^+\,dz - \int (J(z) - J_n(z))^-\,dz.
\end{aligned}
$$

Dominated convergence applied to (28) with ϕ replaced by ϕ_n shows that the functions $\{J_n\}$ converge pointwise to J. Thus $\{(J - J_n)^+\}$ converges pointwise to zero. Notice that $(J - J_n)^+ \leq J$ for each n. Invoke dominated convergence again.

$$
2\|f\| \int (J(z) - J_n(z))^+\,dz \to 0.
$$

Complete the proof by an appeal to Lemma 11. □

Perhaps this result should have been proved before we launched into the Central Limit Theorem proofs of Section 4. At least for identically distributed random variables, Theorems 18 and 22 could have been disposed of more rapidly; but the general cases would have required much the same level of effort, applied to $\exp(\gamma \cdot ix)$ in place of a general $\mathscr{C}^3(\mathbb{R})$ function. The advantages of working with characteristic functions come from the special multiplicative properties of the complex exponential. A central limit theorem for martingales in Chapter VIII will make use of this property.

Theorem 29 contains a bonus: convergence in distribution of random vectors, $X_n \rightsquigarrow X$, is equivalent to the convergence in distribution of all linear functions of the random vectors: $\gamma \cdot X_n \rightsquigarrow \gamma \cdot X$, for every γ. (For the non-trivial half of the assertion, notice that $\phi_n(t)$ is the expectation of a bounded continuous function of $t \cdot X_n$.) Convergence problems for random vectors reduce to collections of convergence problems for random variables; limit theorems for random vectors can be deduced from their univariate analogues.

30 Multivariate Central Limit Theorem. *For every sequence $\{\xi_n\}$ of independent, identically distributed random vectors with $\mathbb{P}\xi_j = 0$ and $\mathbb{P}(\xi_j \xi_j') = V$,*

$$n^{-1/2}(\xi_1 + \cdots + \xi_n) \rightsquigarrow N(0, V).$$

PROOF. For fixed γ, the random variables $\{\gamma \cdot \xi_n\}$ are independent and identically distributed with zero means and variances equal to $\gamma' V \gamma$. If $\gamma' V \gamma = 0$ the random variables $\gamma \cdot \xi_j$ and $\gamma \cdot N(0, V)$ degenerate to zero; the assertion then holds for trivial reasons. Otherwise standardize to unit variance. The standardized sequence satisfies the Lindeberg condition. By Theorem 22,

$$\begin{aligned}(\gamma' V \gamma)^{-1/2} \gamma \cdot n^{-1/2}(\xi_1 + \cdots + \xi_n) &\rightsquigarrow N(0, 1) \\ &= (\gamma' V \gamma)^{-1/2} \gamma \cdot N(0, V). \qquad \square\end{aligned}$$

III.6. Quantile Transformations and Almost Sure Representations

The definition of convergence in distribution by means of pointwise convergence of distribution functions does have its advantages, at least for real random variables. Behind some of these advantages lies a construction that reduces problems involving arbitrary distributions on the real line to the uniform case.

For each distribution function F define a quantile function

$$Q(u) = \inf\{t : F(t) \geq u\} \quad \text{for} \quad 0 < u < 1.$$

Right continuity of F shows that $Q(u)$ sits at the left endpoint of the closed interval of t values for which $F(t) \geq u$; in other words,

$$F(t) \geq u \quad \text{if and only if} \quad Q(u) \leq t.$$

If ξ has the Uniform(0, 1) distribution then

$$\mathbb{P}\{Q(\xi) \leq t\} = \mathbb{P}\{F(t) \geq \xi\} = F(t).$$

That is, $Q(\xi)$ has distribution function F.

The same construction gives a method for representing weakly convergent sequences of probability measures by sequences of random variables that converge almost surely.

31 Representation Theorem. *Let $\{P_n\}$ be a sequence of probability measures on the real line that converges weakly to P. There exist random variables $\{X_n\}$ with distributions $\{P_n\}$, and X, with distribution P, such that $X_n \to X$ almost surely.*

Proof. Write F_n for the distribution function of P_n, and F for the distribution function of P. Denote the quantile function corresponding to F_n by Q_n. Choose ξ with distribution Uniform(0, 1) then define $X_n = Q_n(\xi)$. It suffices to show that $\{Q_n(u)\}$ converges for each u outside a countable subset of (0, 1), for then we can define X (almost surely) as the limit of the $\{X_n\}$. (Problem 21 gives a more concrete representation for X.)

Recall that $F_n(t) \to F(t)$ for each t in the dense set T_0 of points for which $\mathbb{P}\{X = t\} = 0$. Each non-empty open interval contains points of T_0. In particular, if $\{Q_n(u)\}$ does not converge for some fixed u then there exist points t and s in T_0 for which

$$\liminf Q_n(u) < t < s < \limsup Q_n(u).$$

Since $Q_n(u) < t$ implies $F_n(t) \geq u$, the limit along a subsequence gives $F(t) \geq u$. Similarly, $Q_n(u) > s$ infinitely often implies that $F(s) \leq u$. Thus F takes the constant value u throughout the interval $[t, s]$. There can be at most countably many values of u (a set of zero lebesgue measure) for which F has such a flat spot: different u values produce disjoint flat spots, and the real line can accommodate only countably many disjoint intervals. This proves that $\liminf Q_n(u) = \limsup Q_n(u)$ for almost all values of u in (0, 1).

If X is to be defined as the pointwise limit of $\{X_n\}$ we should also exclude the possibility of infinite limits. Given u in (0, 1), choose t in T_0 such that $F(t) > u$. Convergence at points of T_0 gives $F_n(t) > u$ eventually, which implies $Q_n(u) \leq t$ eventually, and so $\limsup Q_n(u) < \infty$. The liminf can be handled similarly. □

Some weak convergence results, such as the one-dimensional case of the Continuous Mapping Theorem, reduce to trivialities when the Representation Theorem is invoked. Similar representations do obtain in higher dimensions and for probability measures on abstract metric spaces, but these require a completely different construction. More about that in Chapter IV.

32 Example. Given a probability measure P on $\mathscr{B}(\mathbb{R})$ and a bounded sequence of measurable functions $\{h_n\}$ converging pointwise to a limit function, h, under what conditions will $P_n h_n \to Ph$ for every sequence $\{P_n\}$ converging weakly to P?

Consider first the special case where P concentrates at a single point x. Choose P_n concentrated at x_n, for any sequence of real numbers $\{x_n\}$ converging to x. In this case the requirement reduces to $h_n(x_n) \to h(x)$. This shows that something stronger than mere pointwise convergence of h_n to

h is needed. (Notice that if $h_n = h$ for every n, the requirement is equivalent to continuity of h at x.)

In general it will suffice if $h_n(x_n) \to h(x)$ for every x in a set of P measure one and every sequence $\{x_n\}$ converging to such an x. Represent $\{P_n\}$ by an almost surely convergent sequence $\{X_n\}$ of random variables. Then $h_n(X_n) \to h(X)$ almost surely. By dominated convergence, $\mathbb{P}h_n(X_n) \to \mathbb{P}h(X)$, which implies that $P_n h_n \to Ph$. □

33 Example. Let G be an open subset of $\mathbb{R}$. Suppose that $P_n \rightsquigarrow P$. The Representation Theorem can be used to show that $\liminf P_n(G) \geq P(G)$. (The proof of the Convergence Lemma also contains the result implicitly.)

Switch to almost surely convergent representations X_n and X. For an ω at which convergence holds, if $X(\omega)$ belongs to G then so does $X_n(\omega)$ for all n large enough. That is,

$$\liminf\{X_n \in G\} \geq \{X \in G\} \quad \text{almost surely.}$$

Apply Fatou's lemma.

$$\begin{aligned} \liminf P_n G &= \liminf \mathbb{P}\{X_n \in G\} \\ &\geq \mathbb{P} \liminf\{X_n \in G\} \\ &\geq \mathbb{P}\{X \in G\} \\ &= P(G). \end{aligned}$$

□

34 Example. Let $\{\phi_n\}$ be the characteristic functions corresponding to a weakly convergent sequence of probability measures $\{P_n\}$, and ϕ be the characteristic function of the limit distribution P. Take $\{X_n\}$ as the almost surely convergent representing sequence of random variables. For each fixed t, the sequence $\{\exp(itX_n)\}$ converges almost surely to $\exp(itX)$. Moreover, the convergence is uniform on each bounded interval I of t-values:

$$\begin{aligned} \sup_I |\exp(itX_n) - \exp(itX)| &\leq \sup_I |\exp(it(X_n - X)) - 1)| \\ &\to 0 \quad \text{almost surely,} \end{aligned}$$

by virtue of the continuity of the exponential function at zero. Because

$$\sup_I |\mathbb{P} \exp(itX_n) - \mathbb{P} \exp(itX)| \leq \mathbb{P} \sup_I |\exp(it(X_n - X)) - 1)|,$$

the sequence $\{\phi_n(t)\}$ converges uniformly to $\phi(t)$ over bounded intervals. □

The argument for the proof of the Representation Theorem made little explicit use of the limit distribution function F. The essentials were:

(i) existence of $\lim F_n(t)$ for each t in a dense subset T_0;

(ii) for each fixed u in $(0, 1)$ there exists a t in T_0 for which $\liminf F_n(t) > u$, and an s in T_0 for which $\limsup F_n(s) < u$.

The second property is equivalent to the existence, for each fixed $\varepsilon > 0$, of a constant K such that

$$\limsup P_n[-K, K]^c < \varepsilon. \tag{35}$$

Of course K depends on ε. A sequence $\{P_n\}$ with this property is said to be uniformly tight. The corresponding property for a sequence of random variables $\{X_n\}$—that for each $\varepsilon > 0$ there exists a K such that

$$\limsup \mathbb{P}\{|X_n| > K\} < \varepsilon$$

—is also called uniform tightness.

36 Theorem. *Each uniformly tight sequence of probability measures on the real line has a subsequence that converges weakly.*

PROOF. Use Cantor's diagonalization procedure to select a subsequence $\{P_{n'}\}$ for which $\{F_{n'}(t)\}$ converges for each rational t. The subsequence satisfies both of the conditions (i) and (ii) noted above. Construct an almost surely convergent sequence of random variables $\{X_{n'}\}$ such that $X_{n'}$ has distribution $P_{n'}$ and $X_{n'} \to X$ almost surely. The sequence $\{P_{n'}\}$ converges weakly to the distribution of X. □

By providing a method for constructing probability measures on $\mathscr{B}(\mathbb{R})$, this theorem allows specification of the limit distribution to be omitted from some important results. For example, the Continuity Theorem for characteristic functions, as proved (Theorem 29) in Section 5, can be improved upon slightly.

37 Continuity Theorem (General Form). *Let $\{X_n\}$ be a sequence of real random variables whose characteristic functions $\{\phi_n\}$ converge pointwise to some function ϕ. If ϕ is continuous at the origin then it must be the characteristic function of an X for which $X_n \rightsquigarrow X$.*

PROOF. Once we prove that ϕ is a characteristic function, Theorem 29 will do the rest. Let us show that $\{X_n\}$ is uniformly tight. Then, by Theorem 36 it will have at least one subsequence that converges in distribution. The limit of any such subsequence must have ϕ as its characteristic function.

The inequality (35), which defines uniform tightness, calls for a constraint to be placed on the tails of the distributions of the random variables $\{X_n\}$. Here we can use an inequality, valid for any random variable Z with characteristic function ρ, that relates the tail behavior of a distribution to the values of its characteristic function near the origin: there exists a positive constant α (approximately 6.308) such that

$$\mathbb{P}\{|Z| \geq h^{-1}\} \leq (\alpha/2h) \int_{-h}^{h} [1 - \rho(t)]\, dt \tag{38}$$

for every positive h. This follows (Fubini's theorem) from the equality

$$(2h)^{-1}\int_{-h}^{h}[1-\mathbb{P}\exp(iZt)]\,dt=\mathbb{P}[1-(\sin hZ)/hZ].$$

Interpret $(\sin 0)/0$ as 1. Because the integrand is non-negative everywhere and greater than $\alpha^{-1}=1-\sin 1$ on the set $\{|Z|\geq h^{-1}\}$, the expectation is greater than $\alpha^{-1}\mathbb{P}\{|Z|\geq h^{-1}\}$.

Apply (38) with $Z=X_n$ and $\rho=\phi_n$.

$$\limsup_n \mathbb{P}\{|X_n|\geq h^{-1}\}\leq\limsup_n(\alpha/2h)\int_{-h}^{h}[1-\phi_n(t)]\,dt$$
$$=(\alpha/2h)\int_{-h}^{h}[1-\phi(t)]\,dt$$

by dominated convergence. Thanks to the continuity of ϕ at the origin, the last expression can be brought close to

$$\alpha[1-\phi(0)]=\lim_n\alpha[1-\phi_n(0)]=0$$

by choosing h small enough. □

Notes

Everything in this chapter is classical.

The name Continuous Mapping Theorem seems to have taken hold in the literature, despite its evident inappropriateness. Billingsley (1968, Section 5) has more on the history of the theorem. Some authors refer to P-almost-sure-continuity as P-Riemann-integrability.

Convolution with a normal kernel as a means of smoothing is an ancient idea. Weierstrass (1885) used it to prove his famous approximation theorem. Liapounoff (1900) introduced an auxiliary normal variable as a device for smoothing a sum of independent variables. That enabled him to complete the argument of Glaisher (1872), and thereby prove with full rigor a central limit theorem. The Liapounoff condition did not appear in the 1900 paper; instead Liapounoff imposed a stronger requirement on the maximum third absolute moment. In his 1901 paper he improved this to a requirement on the sum of $(2+\delta)$th moments about the mean. (Problem 19 gives a minor variation on his 1901 result.)

Section 4 follows in part the lucid exposition of Lindeberg (1922), with a little help from Billingsley (1968, Section 7).

Uspensky (1937, Chapter XIV) attributed the Continuity Theorem to Liapounoff (1900); the key idea was implicit in Liapounoff's proof. Lévy (1922) proved it explicitly. Both Liapounoff and Lévy required uniform convergence on compacta. The dominated convergence trick in the text is often called Scheffé's lemma.

Problems

[1] Suppose X, X_1, $X_2, \ldots$ are real random variables such that $\mathbb{P}\{X_n \le x\} \to \mathbb{P}\{X \le x\}$ whenever $\mathbb{P}\{X = x\} = 0$. Prove that $X_n \rightsquigarrow X$. [For a fixed f in $\mathscr{C}(\mathbb{R})$ find points $\{\alpha_i\}$ with $-\infty < \alpha_1 < \alpha_2 < \cdots < \alpha_k < \infty$ such that $\mathbb{P}\{X = \alpha_i\} = 0$ and both of $\mathbb{P}\{X \le \alpha_1\}$ and $\mathbb{P}\{X > \alpha_k\}$ are small. Choose the $\{\alpha_i\}$ close enough together to ensure that f changes by very little over each interval $[\alpha_i, \alpha_{i+1}]$. Approximate f by a step function constant on each of those intervals.] Extend the argument to higher dimensions.

[2] Let $\{x_n\}$ be a sequence of points in $\mathbb{R}^k$, and $\{X_n\}$ be a sequence of degenerate random vectors with $\mathbb{P}\{X_n = x_n\} = 1$. Show that $\{X_n\}$ converges in distribution if and only if $\{x_n\}$ converges.

[3] Let P be a probability measure on the σ-field $\mathscr{B}(\mathbb{R}^k)$. Show that for each $\varepsilon > 0$ and each borel set B there exists a closed set F_ε contained in B and an open set G_ε containing B such that $P(G_\varepsilon - F_\varepsilon) < \varepsilon$. [Prove that the class of all such B forms a σ-field. It contains every closed set, because closed sets can be written as countable intersections of open sets.] Deduce that if $\mathbb{P}f(X) = \mathbb{P}f(Y)$ for every bounded uniformly continuous f then X and Y have the same distribution. [Represent open sets as the limits of increasing sequences of uniformly continuous functions.]

[4] A sequence of probability measures on $\mathscr{B}(\mathbb{R}^k)$ can converge weakly to at most one limit distribution: if $P_n \rightsquigarrow P$ and $P_n \rightsquigarrow Q$ then P and Q agree for all borel sets. [Use Problem 3.]

[5] If $X_n \rightsquigarrow X$ then $X_n + o_p(1) \rightsquigarrow X$. [Appendix A explains the $o_p(\cdot)$ notation. This result is sometimes called Slutsky's theorem.]

[6] If $\{X_n\}$ converges in probability to X then $X_n \rightsquigarrow X$. The converse is false.

[7] If $X_n \rightsquigarrow X$ and $\mathbb{P}\{X = c\} = 1$, for some constant c, then $\{X_n\}$ converges in probability to c.

[8] Give an example of a continuous function g and a sequence of random vectors with $X_n \rightsquigarrow X$ for which $\{\mathbb{P}g(X_n)\}$ does not converge to $\mathbb{P}g(X)$. [Don't forget the word *bounded* in the definition of convergence in distribution.]

[9] Give an example of a map H from $\mathbb{R}^k$ into itself and a sequence of random vectors in $\mathbb{R}^k$ with $X_n \rightsquigarrow X$ for which HX_n does not converge in distribution to HX.

[10] The set D of discontinuity points of a map H from $\mathbb{R}^k$ into $\mathbb{R}^s$ is borel measurable. [Start from the set $D_{m,n}$ of all those points x for which $|Hy - Hz| > m^{-1}$ for at least one pair of points with $|x - y| < n^{-1}$ and $|x - z| < n^{-1}$. Prove that $D_{m,n}$ is open. Topsøe.]

[11] A random $\mathbb{R}^k$-vector X and a random $\mathbb{R}^j$-vector Y defined on the same probability space can be combined into a single random $\mathbb{R}^{j+k}$-vector (X, Y). If $X_n \rightsquigarrow X$, as random vectors in $\mathbb{R}^k$, and $Y_n \rightsquigarrow Y$, as random vectors in $\mathbb{R}^j$, it need not follow that $(X_n, Y_n) \rightsquigarrow (X, Y)$. But if $\mathbb{P}\{X = c\} = 1$ for some constant c the result is true. [Consider expectations of bounded uniformly continuous functions on $\mathbb{R}^{j+k}$.] Use characteristic functions to prove that the result also holds if X is independent of Y and X_n is independent of Y_n, for each n.

[12] If X_n has a t-distribution on n degrees of freedom then $X_n \rightsquigarrow N(0, 1)$ as $n \to \infty$. [Use Problem 11.]

[13] If real random variables $\{X_n\}$ converge in distribution to X and $\{x_n\}$ is a sequence of real numbers converging to an x for which $\mathbb{P}\{X = x\} = 0$, prove that $\mathbb{P}\{X_n \geq x_n\} \to \mathbb{P}\{X \geq x\}$.

[14] If a sequence of random vectors $\{X_n\}$ converges in distribution then $X_n = O_p(1)$. [Appendix A explains the $O_p(\cdot)$ notation.]

[15] Let H be a measurable map from $\mathbb{R}^k$ into $\mathbb{R}^s$ that is differentiable at a point x_0. That is, there exists a linear map L from $\mathbb{R}^k$ into $\mathbb{R}^s$ such that

$$Hx = Hx_0 + L(x - x_0) + o(x - x_0) \quad \text{near } x_0.$$

If $n^{1/2}(X_n - x_0) \rightsquigarrow Z$, prove that $n^{1/2}(HX_n - Hx_0) \rightsquigarrow LZ$. [Some authors call this the delta method.]

[16] If $X_n \sim \text{Bin}(n, \theta)$ for some fixed θ in $(0, 1)$, find the limiting distribution of $n^{1/2}[\text{arsin}(X_n/n)^{1/2} - \text{arsin}\,\theta^{1/2}]$ as $n \to \infty$.

[17] For y real, define $H_n(y) = \exp(iy) - 1 - iy - \cdots - (iy)^n/n!$. Prove that $|H_n(y)| \leq |y|^{n+1}/(n + 1)!$. [Proceed inductively, using $i\int_0^t H_n(s)\,ds = H_{n+1}(t)$ for $t > 0$. Take complex conjugates for $t < 0$. Borrowed from Feller (1971).]

[18] Use the inequality in the previous problem to show that the characteristic function of $n^{-1/2}(\text{Poisson}(n) - n)$ converges pointwise to $\exp(-t^2/2)$. [This is one way to find the characteristic function of the $N(0, 1)$ distribution.]

[19] For each n let Z_n be a sum of independent random variables $\xi_{n1}, \ldots, \xi_{nk(n)}$ with zero means and variances that sum to one. If for some $\delta > 0$,

$$\sum_{j=1}^{k(n)} \mathbb{P}|\xi_{nj}|^{2+\delta} \to 0 \quad \text{as} \quad n \to \infty$$

then $Z_n \rightsquigarrow N(0, 1)$. [Apply the Lindeberg Central Limit Theorem. Liapounoff.]

[20] Suppose a random vector X has a characteristic function ϕ that is integrable with respect to lebesgue measure on $\mathbb{R}^k$. Prove that the distribution of X has a bounded, uniformly continuous density

$$g(z) = (2\pi)^{-k} \int \exp(-iw \cdot z)\phi(w)\,dw.$$

[Choose a continuous f vanishing outside a bounded region of $\mathbb{R}^k$. Start from the expression for $\mathbb{P}f(X + \sigma Y)$ derived in Section 5:

$$\iint (2\pi\sigma)^{-k} f(z) \exp(iz \cdot y/\sigma - \tfrac{1}{2}|y|^2)\phi(-y/\sigma)\,dy\,dz.$$

Make a change of variable $w = -y/\sigma$, then take the limit as $\sigma \to 0$ under the integral sign. Adapt Problem 3 to complete the proof.]

[21] In the proof of the Representation Theorem write Q for the quantile function corresponding to F. Show that $\{Q_n(u)\}$ converges to $Q(u)$ for each u that is a point of continuity for Q.

CHAPTER IV

Convergence in Distribution in Metric Spaces

... in which that theory from Chapter III depending only on the metric space properties of $\mathbb{R}^k$ is extended to general metric spaces. It is argued that the theory should consider not just borel-measurable random elements. A Continuous Mapping Theorem and an analogue of the almost sure Representation Theorem survive the generalization. A compactness condition—uniform tightness—is shown to guarantee existence of cluster points of sequences of probability measures.

IV.1. Measurability

We write a statistic as a functional on the sample paths of a stochastic process in order to break an analysis of the statistic into two parts: the study of continuity properties of the functional; the study of the stochastic process as a random element of a space of functions. The method has its greatest appeal when many different statistics can be written as functionals on the same process, or when the process has a form that suggests a simple approximation, as in the goodness-of-fit example from Chapter I. There we expressed various statistics as functionals on the empirical process U_n, which defines a random element of $D[0, 1]$. Doob's heuristic argument suggested that U_n should behave like a brownian bridge, in some distributional sense.

Formalization of the heuristic, the task we embark upon in this chapter, requires a notion of convergence in distribution for random elements of $D[0, 1]$. As for euclidean spaces, the definition will involve convergence of expectations of bounded, continuous functions of the processes. For this we need a notion of distance. Equip $D[0, 1]$ with its uniform metric, which assigns the maximum separation

$$\|x - y\| = \sup_t |x(t) - y(t)|$$

as the distance between x and y. We shall find it easiest to prove convergence in distribution of $\{U_n\}$ using this metric, even though it does create some minor measurability difficulties. Chapter VI will examine another metric, for which these difficulties disappear, at the cost of greater topological complexity.

An expectation $\mathbb{P}f(U_n)$ is well defined only when $f(U_n)$ is measurable. If U_n lives on a probability space $(\Omega, \mathscr{E}, \mathbb{P})$, we can arrange for measurability by equipping $D[0, 1]$ with a σ-field, $\mathscr{P}$ say, then checking $\mathscr{E}/\mathscr{P}$-measurability of U_n and $\mathscr{P}$-measurability of f. The borel σ-field will not be the best choice for $\mathscr{P}$. The definition of convergence in distribution for random elements of a general metric space anticipates this complication for $D[0, 1]$.

1 Definition. An $\mathscr{E}/\mathscr{A}$-measurable map X from a probability space $(\Omega, \mathscr{E}, \mathbb{P})$ into a set $\mathscr{X}$ with σ-field $\mathscr{A}$ is called a random element of $\mathscr{X}$.

If $\mathscr{X}$ is a metric space, the set of all bounded, continuous, $\mathscr{A}/\mathscr{B}(\mathbb{R})$-measurable, real-valued functions on $\mathscr{X}$ is denoted by $\mathscr{C}(\mathscr{X}; \mathscr{A})$.

A sequence $\{X_n\}$ of random elements of $\mathscr{X}$ converges in distribution to a random element X, written $X_n \rightsquigarrow X$, if $\mathbb{P}f(X_n) \to \mathbb{P}f(X)$ for each f in $\mathscr{C}(\mathscr{X}; \mathscr{A})$.

A sequence $\{P_n\}$ of probability measures on $\mathscr{A}$ converges weakly to P, written $P_n \rightsquigarrow P$, if $P_n f \to Pf$ for every f in $\mathscr{C}(\mathscr{X}; \mathscr{A})$. □

The borel σ-field $\mathscr{B}(\mathscr{X})$, the σ-field generated by the closed sets, will always contain $\mathscr{A}$. For those spaces where we need $\mathscr{A}$ strictly smaller than the borel σ-field, we will usually have it generated by the collection of all closed balls in $\mathscr{X}$. Also the trace of $\mathscr{A}$ on each separable subset of $\mathscr{X}$ will coincide with the trace of the borel σ-field on the same subset. Limit distributions will always be borel measures concentrating on separable, $\mathscr{A}$-measurable subsets of $\mathscr{X}$. We could build these properties into the definition of weak convergence, but it would neither save us any extra work, nor simplify the theory much.

2 Example. If $D[0, 1]$ is equipped with the borel σ-field $\mathscr{B}$ generated by the closed sets under the uniform metric, the empirical processes $\{U_n\}$ will not be random elements of $D[0, 1]$ in the sense of Definition 1. That is, U_n is not $\mathscr{E}/\mathscr{B}$-measurable.

Consider, for example, the situation for a sample of size one. (Problem 1 extends the argument to larger sample sizes.) For each subset A of $[0, 1]$ define

$$G_A = \{x \in D[0, 1]: x \text{ has a jump at some point of } A\}.$$

Each G_A is open because $|x(t) - x(t-)|$ depends continuously upon x, for fixed t. If U_1 were $\mathscr{E}/\mathscr{B}$-measurable, the set $\{U_1 \in G_A\} = \{\xi_1 \in A\}$ would belong to $\mathscr{E}$. A probability measure μ could be defined on the class of all subsets of $[0, 1]$ by setting $\mu(A) = \mathbb{P}\{\xi_1 \in A\}$. This μ would be an extension of the uniform distribution to all subsets of $[0, 1]$. Unfortunately, such an extension cannot coexist with the usual axioms of set theory (Oxtoby 1971, Section 5): if we wish to retain the axiom of choice, or accept the continuum hypothesis, we must give up borel measurability of U_1. The borel σ-field generated by the uniform metric on $D[0, 1]$ contains too many sets.

There is a simple alternative to the borel σ-field. For each fixed t, the map $U_n(\cdot, t)$ from Ω into $\mathbb{R}$ is a random variable. That is, if π_t denotes the coordinate projection map that takes a function x in $D[0, 1]$ onto its value at t, the composition $\pi_t \circ U_n$ is $\mathscr{E}/\mathscr{B}(\mathbb{R})$-measurable. Each U_n is measurable with respect to the σ-field $\mathscr{P}$ generated by the coordinate projection maps (Problem 2). Call $\mathscr{P}$ the projection σ-field. Problem 4 shows that $\mathscr{P}$ coincides with the σ-field generated by the closed balls. All interesting functionals on $D[0, 1]$ are $\mathscr{P}$-measurable. □

Too large a σ-field $\mathscr{A}$ makes it too difficult for a map into $\mathscr{X}$ to be a random element. We must also guard against too small an $\mathscr{A}$. Even though the metric on $\mathscr{X}$ has lost the right to have $\mathscr{A}$ equal to the borel σ-field, it can still demand some degree of compatibility before a fruitful weak convergence theory will result. If $\mathscr{C}(\mathscr{X}; \mathscr{A})$ contains too few functions, the approximation arguments underlying the Continuous Mapping Theorem will fail. Without that key theorem, weak convergence becomes a barren theory. An extreme example should give you some idea of the worst that might happen.

3 Example. Allow the real line to retain its usual euclidean metric, but change its σ-field to the one generated by the intervals of the form $[n, n + 1)$, with n ranging over the integers. Call this σ-field $\mathscr{R}$. Functions measurable with respect to $\mathscr{R}$ must stay constant over each of the generating intervals. For a continuous function, this imposes a harsh restriction; continuity at each integer forces an $\mathscr{R}$-measurable function to be constant over the whole real line. This completely degrades the weak convergence concept: every sequence of $\mathscr{R}$-measurable random elements converges in distribution. It bodes ill for a sensible Continuous Mapping Theorem.

Consider the map H from the disfigured real line into the real real line (equipped with its usual metric and σ-field) defined by $Hx = 1$ if $0 \le x < 3$ and $Hx = 0$ otherwise. It is a perfectly good $\mathscr{R}$-measurable map, continuous at the point 1. Apply it to random elements $\{X_n\}$ identically equal to 3, and X identically equal to 1. Even though $X_n \rightsquigarrow X$ in the sense of Definition 1, $\{HX_n\}$ does not converge in distribution to HX. □

IV.2. The Continuous Mapping Theorem

Suppose $X_n \rightsquigarrow X$, as $\mathscr{A}$-measurable random elements of a metric space $\mathscr{X}$, and let H be an $\mathscr{A}/\mathscr{A}'$-measurable map from $\mathscr{X}$ into another metric space $\mathscr{X}'$. If H is continuous at each point of an $\mathscr{A}$-measurable set C with $\mathbb{P}\{X \in C\} = 1$, does it follow that $HX_n \rightsquigarrow HX$? That is, does $\{\mathbb{P}f(HX_n)\}$ converge to $\mathbb{P}f(HX)$ for every f in $\mathscr{C}(\mathscr{X}'; \mathscr{A}')$?

We found an answer to the analogous question for random vectors in Section III.2 by reducing it to an application of the Convergence Lemma. The same approach works here. We need to prove $\mathbb{P}h(X_n) \to \mathbb{P}h(X)$ for every bounded, $\mathscr{A}$-measurable, real-valued h that is continuous at each point of C. Were $\mathscr{A}$ equal to the borel σ-field $\mathscr{B}(\mathscr{X})$, the proof would go through almost exactly as before, with only a few words difference. For borel-measurable random elements of metric spaces, the theory parallels the theory in Chapter III very closely, at least as far as the Continuous Mapping Theorem is concerned. Example 3 warns us that non-borel σ-fields require more careful handling.

With this in mind, let's rework the Convergence Lemma of Chapter III, paying more attention to measurability difficulties. To begin with we assume only that $\mathscr{A}$ is a sub-σ-field of $\mathscr{B}(\mathscr{X})$. Define

(4) $$\mathscr{F} = \{f \in \mathscr{C}(\mathscr{X}; \mathscr{A}): f \leq h\}.$$

Last time we constructed a countable subfamily of $\mathscr{F}$ whose pointwise supremum achieved the upper bound h at each point of C. Functions in the subfamily took the form

$$f_{m,r}(x) = r \wedge md(x, \{h \leq r\})$$

Continuity of $f_{m,r}$ suffices for borel measurability, but it needn't imply $\mathscr{A}$-measurability. We must find a substitute for these functions. This is possible if we impose a regularity condition, which ensures that the pointwise supremum of $\mathscr{F}$ equals h at each point of C. If C is separable (meaning that it has a countable, dense subset), we can then extract from $\mathscr{F}$ a countable subfamily having the same supremum as $\mathscr{F}$ at each point of C. The regularity condition will capture the key property enjoyed by $f_{m,r}$.

Without loss of generality suppose $h > 0$. Suppose also that h is continuous at a point x. Choose r with $0 < r < h(x)$. Look for an f in $\mathscr{F}$ with $f(x) \geq r$. Continuity provides a $\delta > 0$ such that $h(y) > r$ on the closed ball $B(x, \delta)$ centered at x. If we could find a g in $\mathscr{C}(\mathscr{X}; \mathscr{A})$ with $0 \leq g \leq B(x, \delta)$ and $g(x) = 1$, the function rg would meet our requirements. Notice the similarity to the topological notion of complete regularity (Simmons 1963, Section 27). If $\mathscr{A}$ happened to contain all the closed balls centered at x, a property enjoyed by the projection σ-field on $D[0, 1]$ (Problem 4), the function

(5) $$g(y) = [1 - \delta^{-1} d(x, y)]^+$$

would do, because $\{g \geq 1 - s\} = B(x, s\delta)$. For general $\mathscr{A}$ we must postulate existence of the appropriate g.

To maintain the parallel with euclidean spaces as closely as possible, strengthen the requirements on g to include uniform continuity. We lose only a scintilla of generality thereby; the special g of (5) still passes the test.

6 Definition. Call a point x in $\mathscr{X}$ completely regular (with respect to the metric d and the σ-field $\mathscr{A}$) if to each neighborhood V of x there exists a uniformly continuous, $\mathscr{A}$-measurable function g with $g(x) = 1$ and $g \leq V$. □

You might well object to yet another mathematical notion attaining the status of regularity; the world is already overloaded with instances of "regular" as a synonym for "amenable to our current theory." At least it has the virtue of reminding us of its topological counterpart. (A more sadistic author might have called it $T_{3\frac{1}{2}}$.) The terminology would not be wasted if we were to expand our weak convergence theory to cover borel measures on general topological spaces, for there topological complete regularity seems just the thing needed for a well-behaved theory.

7 Convergence Lemma. *Let h be a bounded, $\mathscr{A}$-measurable, real-valued function on $\mathscr{X}$. If h is continuous at each point of some separable, $\mathscr{A}$-measurable set C of completely regular points, then:*

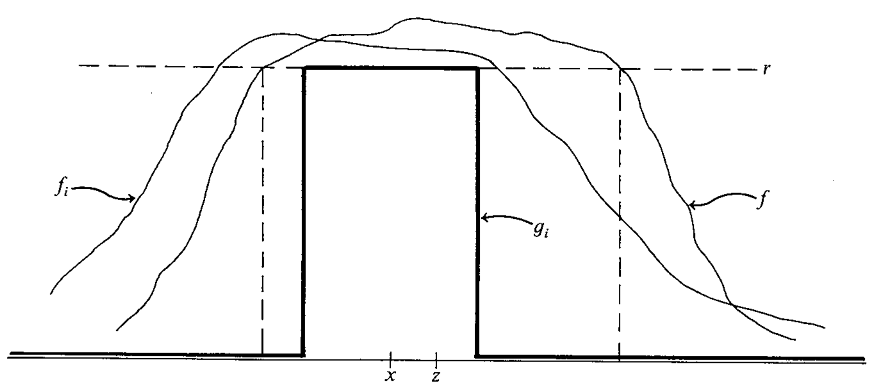
(i) *$X_n \rightsquigarrow X$ and $\mathbb{P}\{X \in C\} = 1$ imply $\mathbb{P}h(X_n) \to \mathbb{P}h(X)$;*
(ii) *$P_n \rightsquigarrow P$ and $PC = 1$ imply $P_n h \to Ph$.*

PROOF. As the arguments for both assertions are quite similar, let us prove (ii) only. Assume that $h > 0$ (add a constant to h if necessary). Define $\mathscr{F}$ as in (4), but with the continuity requirement strengthened to uniform continuity. At those completely regular points of $\mathscr{X}$ where h is continuous, the supremum of $\mathscr{F}$ equals h. This applies to points in C.

Separability of C will enable us to extract a suitable countable subfamily from $\mathscr{F}$. Argue as for the classical Lindelöf theorem (Simmons 1963, Section 18). Let C_0 be a countable, dense subset of C. Let $\{g_1, g_2, \ldots\}$ be the set of all those functions of the form rB, with r rational, B a closed ball of rational radius centered at a point of C_0, and $rB \le f$ for at least one f in $\mathscr{F}$. For each g_i choose one f satisfying the inequality $g_i \le f$. Denote it by f_i. This picks out the required countable subfamily:

(8) $$\sup_i f_i = \sup \mathscr{F} \quad \text{on } C.$$

To see this, consider any point z in C and any f in $\mathscr{F}$. For each rational number r such that $f(z) > r > 0$ choose a rational ε for which $f > r$ at all points within a distance 2ε of z. Let B be the closed ball of radius ε centered at a point x of C_0 for which $d(x, z) < \varepsilon$. The function rB lies completely below f; it must be one of the $\{g_i\}$. The corresponding f_i takes a value greater than r at z. Assertion (8) follows.



Complete the argument as for the Convergence Lemma of Section III.2. Assume without loss of generality that $f_i \uparrow h$ at points of C. Then

$$\begin{aligned} \liminf P_n h &\geq \liminf P_n f_i && \text{for each } i \\ &= P f_i && \text{because } P_n \rightsquigarrow P \\ &\to Ph && \text{as } i \to \infty, \quad \text{by monotone convergence.} \end{aligned}$$

Replace h by $-h +$ (a big constant) to get the companion inequality for the limsup. □

9 Corollary. *If* $\mathbb{P}f(X_n) \to \mathbb{P}f(X)$ *for each bounded, uniformly continuous,* $\mathscr{A}$*-measurable* f, *and if* X *concentrates on a separable set of completely regular points, then* $X_n \rightsquigarrow X$. □

The corollary flows directly from the decision to insist upon uniformly continuous separating functions in the definition of a completely regular point. As with its counterpart for euclidean spaces, it makes some weak convergence arguments just a little bit more straightforward than the corresponding arguments with continuous functions.

10 Example. Let $\mathscr{X}$ be a space equipped with a σ-field $\mathscr{A}$ and metric d, and $\mathscr{Y}$ be a space equipped with a σ-field $\mathscr{B}$ and metric e. Equip $\mathscr{X} \otimes \mathscr{Y}$ with its product σ-field and the metric σ defined by

$$\sigma[(x, y), (x', y')] = \max[d(x, x'), e(y, y')].$$

Suppose $X_n \rightsquigarrow X$, as random elements of $\mathscr{X}$. If $\mathbb{P}_X$ concentrates on a separable set of completely regular points, and $Y_n \to y_0$ in probability for some fixed completely regular point y_0 in $\mathscr{Y}$, then $(X_n, Y_n) \rightsquigarrow (X, y_0)$, as random elements of the product space $\mathscr{X} \otimes \mathscr{Y}$.

Of course the assertion only makes sense if X_n and Y_n are defined on the same probability space. Given that prerequisite, measurability with respect to the product σ-field presents no problem, because

$$(X_n, Y_n)^{-1}(A \otimes B) = (X_n^{-1}A) \cap (Y_n^{-1}B),$$

and similarly for (X, y_0).

Write C for the separable set on which $\mathbb{P}_X$ concentrates. Then $\mathbb{P}_{X,y_0}$ concentrates on the product set $C \otimes \{y_0\}$, which is separable. Each point of this set is completely regular: if $f(c) = 1$ and $f = 0$ outside the ball of d-radius ε, and $g(y_0) = 1$ and $g = 0$ outside a ball of e-radius ε, then the product $f(x)g(y)$ equals 1 at (c, y_0) and vanishes outside a ball of σ-radius ε. The product is uniformly continuous if both f and g are bounded and uniformly continuous; it is $\mathscr{A} \otimes \mathscr{B}$-measurable if f is $\mathscr{A}$-measurable and g is $\mathscr{B}$-measurable.

By virtue of Corollary 9, to prove $(X_n, Y_n) \rightsquigarrow (X, y_0)$ we have only to check that $\mathbb{P}h(X_n, Y_n) \to \mathbb{P}h(X, y_0)$ for each bounded, uniformly continuous, $\mathscr{A} \otimes \mathscr{B}$-measurable, real function h on $\mathscr{X} \otimes \mathscr{Y}$. Given $\varepsilon > 0$ choose

$\delta > 0$ so that $|h(x, y) - h(x', y')| < \varepsilon$ whenever $\sigma[(x, y), (x', y')] < \delta$. Write $k(\cdot)$ for the bounded, uniformly continuous, $\mathscr{A}$-measurable function $h(\cdot, y_0)$. Then

$$|\mathbb{P}h(X_n, Y_n) - \mathbb{P}h(X, y_0)| \leq \varepsilon + 2\|h\|\mathbb{P}^*\{e(Y_n, y_0) \geq \delta\} + |\mathbb{P}k(X_n) - \mathbb{P}k(X)|.$$

Convergence in probability of Y_n to y_0 makes the middle term converge to zero. (Notice the outer measure $\mathbb{P}^*$. By definition, $\mathbb{P}^*Z$ equals the infimum of $\mathbb{P}W$ over all $\mathscr{E}$-measurable real functions with $W \geq Z$. For most applications $e(\cdot, y_0)$ will be $\mathscr{A}$-measurable, in which case $\mathbb{P}^*$ can be replaced by $\mathbb{P}$.) The last term converges to zero because $X_n \rightsquigarrow X$. □

11 Example (Convergence in Distribution via Uniform Approximation). Let $X, X_1, X_2, \ldots$ be random elements of $\mathscr{X}$ with $\mathbb{P}_X$ concentrated on a separable set of completely regular points. Suppose, for each $\varepsilon > 0$ and $\delta > 0$, there exist approximating random elements $AX, AX_1, AX_2, \ldots$ such that:

(i) $\mathbb{P}^*\{d(X, AX) \geq \delta\} < \varepsilon$;
(ii) $\limsup \mathbb{P}^*\{d(X_n, AX_n) \geq \delta\} < \varepsilon$;
(iii) $AX_n \rightsquigarrow AX$.

Then $X_n \rightsquigarrow X$. Notice again the use of outer measure to guard against non-measurability.

We have already met a special case of this result in Lemma III.11, where $AX_n = X_n + \sigma Y$. In applications to stochastic processes, the approximations are typically constructed from the values of the processes at a fixed, finite set of index points. For such approximations, classical weak convergence methods can handle (iii). The assumptions (i) and (ii) place restrictions on the irregularity of the sample paths. Chapter V will take up this idea.

The convergence $X_n \rightsquigarrow X$ follows from convergence of expectations for every bounded, uniformly continuous, $\mathscr{A}$-measurable f. If $|f(x) - f(y)| < \varepsilon$ whenever $d(x, y) < \delta$ then $|\mathbb{P}f(X_n) - \mathbb{P}f(X)|$ is less than

$$\mathbb{P}|f(X_n) - f(AX_n)| + |\mathbb{P}f(AX_n) - \mathbb{P}f(AX)| + \mathbb{P}|f(AX) - f(X)|.$$

The convergence (iii) takes care of the middle term. Handle the first term by splitting it into the contributions from $\{d(X_n, AX_n) \geq \delta\}$ and its complement; and similarly for the last term. □

The Convergence Lemma has one other important corollary, the result that tells us how to transfer convergence in distribution of random elements of $\mathscr{X}$ to convergence in distribution of selected functionals of those random elements. For substantial applications turn to Chapter V.

12 Continuous Mapping Theorem. *Let H be an $\mathscr{A}/\mathscr{A}'$-measurable map from $\mathscr{X}$ into another metric space $\mathscr{X}'$. If H is continuous at each point of some separable, $\mathscr{A}$-measurable set C of completely regular points, then $X_n \rightsquigarrow X$ and $\mathbb{P}\{X \in C\} = 1$ together imply $HX_n \rightsquigarrow HX$.* □

IV.3. Representation by Almost Surely Convergent Sequences

In Section III.6 we used the quantile transformation to construct almost surely convergent sequences of random variables representing weakly convergent sequences of probability measures. That method will not work for probabilities on more general spaces; it even breaks down for $\mathbb{R}^2$. But the representation result itself still holds.

13 Representation Theorem. *Let $\{P_n\}$ be a sequence of probability measures on a metric space. If $P_n \rightsquigarrow P$ and P concentrates on a separable set of completely regular points, then there exist random elements $\{X_n\}$ and X with distributions $\{P_n\}$ and P such that $X_n \to X$ almost surely.* □

The new construction makes repeated use of a lemma that can be applied to any two probability measures P and Q that are close in a weak convergence sense. Roughly speaking, the idea is to cut up the metric space $\mathscr{X}$ into pieces $B_0, B_1, \ldots, B_k$ for which $PB_i \approx QB_i$ for each i, so that the set B_0 has small P measure and each of the other B_i's has small diameter. We use these sets to construct a random element Y of $\mathscr{X}$, starting from an X with distribution P. If X lands in B_i choose Y in B_i according to the conditional distribution $Q(\cdot \mid B_i)$. For $i \geq 1$ this forces Y to lie close to X, because B_i doesn't contain any pairs of points too far apart. The random element Y has approximately the distribution Q:

$$
\begin{aligned}
\text{(14)} \qquad \mathbb{P}\{Y \in A\} &= \sum_{i=0}^{k} \mathbb{P}\{Y \in A \mid X \in B_i\}\mathbb{P}\{X \in B_i\} \\
&= \sum_{i=0}^{k} Q(A \mid B_i)P(B_i) \\
&\approx \sum_{i=0}^{k} Q(A \mid B_i)Q(B_i) \\
&= Q(A).
\end{aligned}
$$

A slight refinement of the construction will turn the approximation into an equality. When applied with $Q = P_n$ and partitions growing finer with n, it will generate the sequence $\{X_n\}$ promised by the Representation Theorem.

15 Lemma. *For each $\varepsilon > 0$ and each P concentrating on a separable set of completely regular points, the space $\mathscr{X}$ can be partitioned into finitely many disjoint, $\mathscr{A}$-measurable sets $B_0, B_1, \ldots, B_k$ such that:*

(i) *the boundary of each B_i has zero P measure (a P-continuity set);*
(ii) $P(B_0) < \varepsilon$;
(iii) *diameter*$(B_i) < 2\varepsilon$ *for* $i = 1, 2, \ldots, k$.

PROOF. Call the separable set C. To each x in C there exists a uniformly continuous, $\mathscr{A}$-measurable f with $f(x) = 1$ and $f = 0$ for points a distance greater than ε from x. The open sets of the form $\{f > \alpha\}$, for $0 < \alpha < 1$, are all $\mathscr{A}$-measurable and of diameter less than 2ε. At each point on the boundary of $\{f > \alpha\}$, the continuous function f takes the value α. Because $P\{f = \alpha\}$ can be non-zero for at most countably many different values of α, there must exist at least one α for which the probability equals zero. Choose and fix such an α, then write $G(x)$ for the corresponding set $\{f > \alpha\}$. It has diameter less than 2ε and is a P-continuity set.

The union of the family of open sets $\{G(x): x \in C\}$ contains the separable set C. Extract a countable subfamily $\{G(x_i): i = 1, 2, \ldots\}$ containing C. (Every open cover of a separable subset of a metric space has a countable subcover: Problem 5.) Because

$$P\left[\bigcup_{i=1}^{k} G(x_i)\right] \uparrow P\left[\bigcup_{i=1}^{\infty} G(x_i)\right] \geq P(C) = 1,$$

there exists a k such that

$$P\left[\bigcup_{i=1}^{k} G(x_i)\right] > 1 - \varepsilon.$$

Define $B_i = G(x_i)\backslash[G(x_1) \cup \cdots \cup G(x_{i-1})]$ for $i = 1, \ldots, k$ and $B_0 = [G(x_1) \cup \cdots \cup G(x_k)]^c$, a process known to the uncouth as disjointification. The boundary of B_i is covered by the union of the boundaries of the P-continuity sets $G(x_1), \ldots, G(x_k)$. Each B_i lies completely inside the corresponding $G(x_i)$, a set of diameter less than 2ε if $i \geq 1$. □

PROOF OF THEOREM 13. Holding ε fixed for the moment, carry out the construction detailed in the proof of the lemma, generating P-continuity sets $B_0, B_1, \ldots, B_k$ as described.

The indicator function of B_i is almost surely continuous $[P]$ because it has discontinuities only at the boundary of B_i. So by the Convergence Lemma $P_n(B_i) \to P(B_i)$. When n is large enough, say $n \geq n(\varepsilon)$,

$$(16) \qquad P_n(B_i) \geq (1 - \varepsilon)P(B_i) \quad \text{for} \quad i = 0, 1, \ldots, k.$$

Write n_m for $n(2^{-m})$. Without loss of generality suppose $1 = n_1 < n_2 < \cdots$. For $n_m \leq n < n_{m+1}$, construct X_n using the $\{B_i\}$ partition corresponding to $\varepsilon_m = 2^{-m}$. Notice that B_i now depends on n through the value of m.

Let ξ be a random variable that has a Uniform(0, 1) distribution independent of X. If $\xi \leq 1 - \varepsilon_m$ and X lands in B_i, choose X_n according to the conditional distribution $P_n(\cdot|B_i)$. So far no B_i has received more than its quota of P_n measure, because of (16). The extra probability will be distributed over the space $\mathscr{X}$ to bring X_n up to its desired distribution P_n. If $\xi > 1 - \varepsilon_m$ choose X_n according to the distribution μ_n determined by

$$P_n(A) = \mu_n(A)\mathbb{P}\{\xi > 1 - \varepsilon_m\} + \sum_{i=0}^{k} P_n(A|B_i)(1 - \varepsilon_m)P(B_i).$$

That is,

$$\mu_n(A) = \varepsilon_m^{-1} \sum_{i=0}^{k} P_n(A \mid B_i)[P_n(B_i) - (1 - \varepsilon_m)P(B_i)].$$

By (16), the right-hand side is non-negative. And clearly $\mu_n \mathscr{X} = 1$.

Except on the set $\Omega_m = \{X \in B_0 \text{ or } \xi > 1 - \varepsilon_m\}$, which has measure at most $2\varepsilon_m$, the random elements X and X_n lie within $2\varepsilon_m$ of each other. On the complement of the set $\{\Omega_m \text{ infinitely often}\}$, the sequence $\{X_n\}$ converges to X. By the Borel–Cantelli lemma $\mathbb{P}\{\Omega_m \text{ infinitely often}\} = 0$. □

The applications of Theorem 13 follow the same pattern as in Section III.6. Problems of weak convergence transform into problems of almost sure convergence, to which the standard tools (monotone convergence, dominated convergence, and so on) can be applied.

17 Example. Most of the proof of the Convergence Lemma did not use the full force of almost sure continuity for the function h. To get the inequality for the liminf we only needed lower-semicontinuity of h at points of C. (Remember that semicontinuity imposes only half the constraint of continuity: only a lower bound is set on the oscillations of h in a neighborhood of a point. Problem 9 will refresh your memory on semicontinuity.) The Representation Theorem gives a quick proof of the same result.

If g is bounded below, lower-semicontinuous, and $\mathscr{A}$-measurable (automatic if $\mathscr{A}$ equals the borel σ-field), then liminf $P_n g \geq Pg$ whenever $P_n \rightsquigarrow P$ with P concentrated on a separable set of completely regular points. To prove it, switch to almost surely convergent representations. Lower-semicontinuity at $X(\omega)$ plus almost sure convergence of the representing sequence imply

$$\liminf g(X_n(\omega)) \geq g(X(\omega)) \quad \text{almost surely.}$$

Take expectations.

$$\begin{aligned} \liminf P_n g &= \liminf \mathbb{P} g(X_n) \\ &\geq \mathbb{P} g(X) \quad \text{by Fatou's lemma} \\ &= Pg. \end{aligned}$$

A similar inequality holds for upper-semicontinuous, $\mathscr{A}$-measurable functions that are bounded above. As a special case,

$$\limsup P_n F \leq PF \tag{18}$$

for each closed, $\mathscr{A}$-measurable set F. If inequality (18) holds for all such F then necessarily $P_n \rightsquigarrow P$ (Problem 12). □

19 Example. Let $\mathscr{G}$ be a uniformly bounded class of $\mathscr{A}$-measurable, real functions on $\mathscr{X}$. Suppose that $P_n \rightsquigarrow P$, with P concentrated on a separable set of completely regular points. Suppose also that $\mathscr{G}$ is equicontinuous at almost all points $[P]$ of $\mathscr{X}$. That is, for almost all x and each $\varepsilon > 0$ there exists a $\delta > 0$, depending on x but not on g, such that $|g(y) - g(x)| < \varepsilon$ whenever $d(x, y) < \delta$, for every g in $\mathscr{G}$. Then

$$\sup_{\mathscr{G}} |P_n g - Pg| \to 0. \tag{20}$$

This result underlies the success of most of the functions that have been constructed in the literature to metrize the topology of weak convergence.

To prove (20), represent the probability measures by almost surely convergent random elements $\{X_n\}$, then deduce from equicontinuity that

$$\sup_{\mathscr{G}} |g(X_n) - g(X)| \to 0 \quad \text{almost surely.} \tag{21}$$

It would be tempting to appeal to dominated convergence to get

$$\sup_{\mathscr{G}} |\mathbb{P}g(X_n) - \mathbb{P}g(X)| \leq \mathbb{P} \sup_{\mathscr{G}} |g(X_n) - g(X)| \to 0,$$

but that would assume measurability of the supremum in (21). Instead, note that (20) could fail only if, for some $\varepsilon > 0$, there were functions $\{g_n\}$ in $\mathscr{G}$ for which $|P_n g_n - P g_n| \geq \varepsilon$ infinitely often. Apply the dominated convergence argument to the countable family $\mathscr{G}_0 = \{g_1, g_2, \ldots\}$ to reach a contradiction. □

22 Example (The Bounded-Lipschitz Metric for Weak Convergence). Suppose that $\mathscr{A}$ contains all the closed balls, as in the case of $D[0, 1]$ under its uniform metric. The function $f(\cdot) = r[1 - md(\cdot, z)]^+$, which serves to separate z from points outside a small neighborhood of z, has the strong uniformity property

$$|f(x) - f(y)| \leq mr\, d(x, y).$$

A function satisfying such a condition, with mr replaced possibly by a different constant, is called a Lipschitz function. For the proof of the Convergence Lemma, $P_n f \to Pf$ for each bounded, $\mathscr{A}$-measurable Lipschitz function would have sufficed; convergence for bounded Lipschitz functions implies weak convergence. From Example 19 we draw a sharper conclusion.

Define $\mathscr{L}$ to be the set of all $\mathscr{A}$-measurable Lipschitz functions for which $|f(x) - f(y)| \leq d(x, y)$ and $\sup_x |f(x)| \leq 1$. The class $\mathscr{L}$ is equicontinuous at each point of $\mathscr{X}$. Every bounded Lipschitz function can be expressed as a multiple of a function in $\mathscr{L}$.

Define the distance between two probability measures on $\mathscr{A}$ by

$$\lambda(P, Q) = \sup\{|Pf - Qf| : f \in \mathscr{L}\}.$$

You can check that λ has all the properties required of a metric. If P concentrates on a separable set and $P_n \rightsquigarrow P$, the distance $\lambda(P_n, P)$ converges to zero, in obedience to the uniformity result of Example 19. Conversely, the convergence of $\lambda(P_n, P)$ to zero would ensure that $P_n f \to Pf$ for each bounded Lipschitz function f, which, as noted above, implies weak convergence. □

23 Example (The Prohorov Metric for Weak Convergence). Suppose $\mathscr{X}$ is a separable metric space equipped with its borel σ-field. For each $\delta > 0$ and each borel subset A of $\mathscr{X}$ define

$$A^\delta = \{x \in \mathscr{X}: d(x, A) < \delta\}.$$

(Visualize the open set A^δ as A wearing a halo of thickness δ.) Define the Prohorov distance between two borel probability measures as

$$\rho(P, Q) = \inf\{\delta > 0: PA^\delta + \delta \geq QA \quad \text{for every } A\}.$$

This distance has great appeal for robustniks, who interpret the delta halo as a way of constraining small migrations of Q mass and the added delta as insurance against a small proportion of gross changes. To us it will be just another metric for weak convergence.

It is not obvious that ρ is symmetric, one of the properties required of a metric. We need to show that $QA^\delta + \delta \geq PA$ for every A, whenever $\rho(P, Q) < \delta$. Set B equal to the complement of A^δ. We know that $QB \leq PB^\delta + \delta$. Subtract both sides from 1, after replacing B^δ by the complement of A, a larger set. (No point of A can be less than δ from a point in B.) We have symmetry.

If $\rho(P, Q) = 0$ then certainly $PF^\delta + \delta \geq QF$ for every closed F and every $\delta > 0$. Hold F fixed but let δ tend to zero through a sequence of values. The sequence $\{F^\delta\}$ shrinks to F, giving $PF \geq QF$ in the limit. Interchange the roles of P and Q then repeat the argument to deduce that P and Q agree on all closed sets, and hence (Problem 11) on all borel sets.

For the triangle inequality, suppose that $\rho(P, Q) < \delta$ and $\rho(Q, R) < \eta$. Temporarily set $B = A^\eta$. Then

$$RA \leq QA^\eta + \eta = QB + \eta \leq PB^\delta + \eta + \delta.$$

Check that $A^{\delta+\eta}$ contains B^δ. Deduce that $\rho(R, P) \leq \eta + \delta$.

Next, show that weak convergence implies convergence in the ρ metric. It suffices to deduce that $\rho(P_n, P) \leq \delta$ eventually if $P_n \rightsquigarrow P$. For each borel set A define

$$f_A(x) = [1 - \delta^{-1} d(x, A)]^+.$$

Notice that $A^\delta \geq f_A \geq A$. Also, because

$$|f_A(x) - f_A(y)| \leq \delta^{-1}|d(x, A) - d(y, A)| \leq \delta^{-1} d(x, y),$$

the class of all such f_A functions is equicontinuous. By Example 19,

$$\sup_A |P_n f_A - P f_A| \to 0.$$

Call this supremum Δ_n. Then

$$PA^\delta \geq Pf_A \geq P_n f_A - \Delta_n \geq P_n A - \Delta_n$$

for every A. Wait until $\Delta_n \leq \delta$ to be able to assert that $\rho(P, P_n) \leq \delta$.

Finally, if $\rho(P_n, P) \to 0$ then, for fixed closed F,

$$\limsup P_n F \leq PF^\delta + \delta$$

for every $\delta > 0$. Let δ decrease to zero then deduce from Problem 12 that $P_n \rightsquigarrow P$. Convergence in the ρ metric is equivalent to weak convergence. □

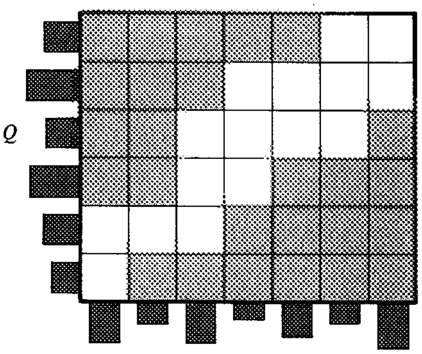
IV.4. Coupling

The Representation Theorems of Sections III.6 and IV.3 both depended upon methods for coupling distributions P_n and P. That is, we needed to construct random elements X_n and X, on the same probability space, such that X_n had distribution P_n and X had distribution P. Closeness of P_n and P, in a weak convergence sense, allowed us to choose X_n and X close in a stronger, almost sure sense. This section will examine coupling in more detail.

A coupling of probability measures P and Q, on a space $\mathscr{X}$, can be realized as a measure M on the product space $\mathscr{X} \otimes \mathscr{X}$, with X and Y defined by the coordinate projections. The product measure $P \otimes Q$ is a coupling, albeit a not very informative one. More useful are those couplings for which M concentrates near the diagonal. For example, in the Representation Theorem we put as much mass as possible on the set $\{(x, y): d(x, y) \leq \varepsilon\}$.

Roughly speaking, one can construct such couplings in two steps. First treat the desired property—that as much mass as possible be allocated to a particular region D in the product space—as a strict requirement. Imagine building up M slowly by drawing off mass from the P marginal measure and relocating it within D, subject to a matching constraint: to put an amount δ near (x, y) one must deplete the P supply near x by δ and the Q supply near y by δ. When as much mass as possible has been shifted into D by this method, forget about the constraint imposed by D. In the second step, complete the transfer of mass from P into the product space subject only to the matching constraint. The final M will have the correct marginals, P and Q.

A precise formulation of the coupling algorithm just sketched is easiest when both P and Q concentrate on a finite set of points. The first step can be represented by a picture that looks like a crossword puzzle. Label the points on which Q concentrates as $1, \ldots, r$; let these correspond to rows of a two-way array of cells. Similarly, let $1, \ldots, c$ label both the points on which P concentrates and the columns of the two-way array. The stack beside row i represents the mass Q puts on point i, and the stack under column j represents the mass P puts on j. The unshaded cells correspond to D. The aim is to place as much mass as possible in the unshaded cells without violating the constraint that the total mass in a row or column should not exceed the amount originally in the marginal stacks.



This formulation makes sense even if the marginal supplies don't both correspond to measures with total mass one. In general we could allow any non-negative masses $R(i)$ and $C(j)$ in the supply stacks for row i and column j. We would seek a non-negative allocation $M(i,j)$ of as much mass as possible into the unshaded cells, subject to

$$\sum_i M(i,j) \le C(j) \quad \text{and} \quad \sum_j M(i,j) \le R(i)$$

for each i and j. A continuous analogue of the classical marriage lemma (a sort of fractional polygamy) will give the necessary and sufficient conditions for existence of an M that turns the inequalities for the columns into equalities.

Treat C and R as measures. Write $C(J)$ for the sum of supply masses in a set of columns J. Denote by D_J the set of rows i for which cell (i,j) belongs to D for at least one column j in J. It is easy to see that M can have column marginal C only if $R(D_J) \ge C(J)$ for every J, because the rows in D_J contain all the D-cells in the columns of J. Sufficiency is a little trickier.

24 Allocation Lemma. *If $R(D_J) \ge C(J)$ for every set of columns J, then there exists an allocation $M(i,j)$ into the cells of D such that*

$$\sum_i M(i,j) = C(j) \quad \textit{and} \quad \sum_j M(i,j) \le R(i)$$

for every i and j.

Proof. Use induction on the number of columns. The result is trivial for $c = 1$. Suppose it is true for every number of columns strictly less than c.

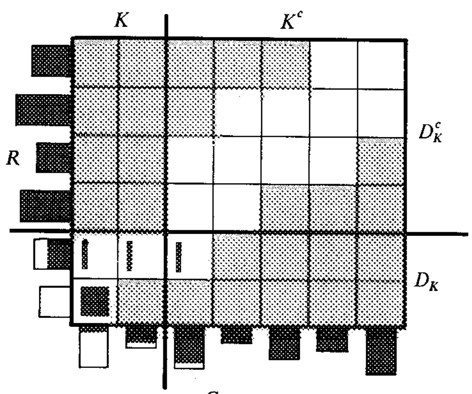
Construct M by transferring mass from the column margins into D. Shift mass at a constant rate into each of the D-cells in row r. For any mass shifted from $C(j)$ into (r, j) discard an equal amount from $R(r)$. If $R(r)$ becomes exhausted, move on to row $r - 1$, and so on. Stop when either:

(i) some $C(j)$ is exhausted; or
(ii) one of the constraints $R(D_J) \geq C(J)$ would be violated by continuation of the current method of allocation.

Here R and C are used as variable measures that decrease as mass is drawn off; the supply stacks diminish as the allocation proceeds. Notice that the mass transferred at each step can be specified as the largest solution to a system of linear inequalities.

If the allocation halts because of (i), the problem is transformed into an allocation for $c - 1$ columns. The inductive hypothesis can be invoked to complete the allocation.

If allocation halts because of (ii), then there must now exist some K for which $R(D_K) = C(K)$. Continued allocation would have caused $R(D_K) < C(K)$. The matching-constraint prevents K from containing every column: the total column supply always decreases at the same rate as the total row supply. Write K^c for the non-empty set of columns not in K.



If the marginal demands of the columns in K are to be met, the entire remaining supply $R(D_K)$ must be devoted to those columns. With this requirement the problem splits into two subproblems: rows in D_K may match only mass drawn off from the columns in K; from the rows D_K^c not in D_K, match mass from the columns in K^c. Both subproblems satisfy the initial assumptions of the lemma. For subsets of K this follows because allocation halted before $R(D_J) < C(J)$ for any J. For subsets of K^c, it follows from

$$\begin{aligned} R(D_J \cap D_K^c) &= R(D_{J \cup K}) - R(D_K) \\ &\geq C(J \cup K) - C(K) \\ &= C(J). \end{aligned}$$

Invoke the inductive hypothesis for both subproblems to complete the proof of the lemma. □

25 Corollary. *If R and C have the same total mass and $R(D_J) \geq C(J)$ for every J, then the allocation measure M has marginal measures R and C.* □

The Allocation Lemma applies directly only to discrete distributions supported by finite sets. For distributions not of that type a preliminary discretization, as in the proof of the Representation Theorem, is needed.

26 Example. Let P and Q be borel probability measures on a separable metric space. The Prohorov distance $\rho(P, Q)$ determines how closely P and Q can be coupled, in the sense that $\rho(P, Q)$ equals the infimum of those values of ε such that

$$\mathbb{P}\{d(X, Y) \geq \varepsilon\} \leq \varepsilon, \tag{27}$$

with X having distribution P and Y having distribution Q. We can use the Allocation Lemma to help prove this.

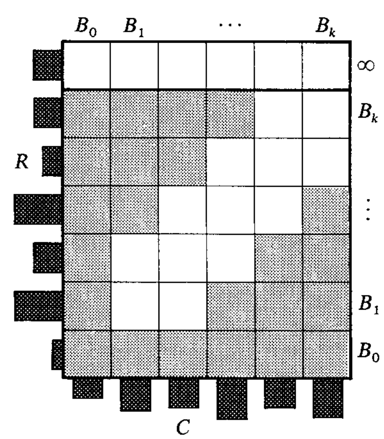
Half of the argument is easy. From (27) deduce, for every A,

$$\begin{aligned} QA &= \mathbb{P}\{Y \in A\} \\ &\leq \mathbb{P}\{X \in A^\varepsilon\} + \mathbb{P}\{d(X, Y) \geq \varepsilon\} \\ &\leq PA^\varepsilon + \varepsilon, \end{aligned}$$

whence $\rho(P, Q) \leq \varepsilon$.

For the other half of the argument suppose $\rho(P, Q) < \varepsilon$. Construct X and Y by means of a two-stage coupling. Apply the method of Lemma 15 twice to partition the underlying space into sets $B_0, B_1, \ldots, B_k$ with both $QB_0 < \delta$ and $PB_0 < \delta$, and diameter$(B_i) < \delta$ for $i = 1, \ldots, k$. Choose δ as a quantity much smaller than ε; it will eventually be forced down to zero while ε stays fixed. The requirement that each B_i be a Q- or P-continuity set is irrelevant to our present purpose.

Set $R(i)$ equal to QB_i and $C(j)$ equal to PB_j. Into the region D allow only those cells (i, j), for $1 \leq i \leq k$ and $1 \leq j \leq k$, whose corresponding B_i and B_j contain a pair of points, one in B_i and one in B_j, a distance $\leq \varepsilon$ apart. Augment the double array by one more row, call it ∞, whose row stack contains mass $\varepsilon + 2\delta$. Include $(\infty, 0), \ldots, (\infty, k)$ in the region D.



The hypotheses of the Allocation Lemma are satisfied. For any collection of columns J,

$$\begin{aligned}
C(J) &\leq PB_0 + P\Big(\bigcup_{J\backslash\{0\}} B_j\Big) \\
&< \delta + Q\Big(\bigcup_{J\backslash\{0\}} B_j\Big)^{\varepsilon} + \varepsilon \\
&\leq \delta + Q\Big(\bigcup_{D_J\backslash\{\infty\}} B_i\Big) + QB_0 + \varepsilon \quad \text{by definition of } D \\
&< \delta + R(D_J\backslash\{\infty\}) + \delta + \varepsilon \\
&= R(D_J).
\end{aligned}$$

Distribute all the mass from the column stacks into D, as in the Allocation Lemma. The ∞ row acts as a temporary repository for the small amount of mass that cannot legally be shifted into the desired small-diameter cells. Return the mass in this row to the column stacks, leaving at least $1 - \varepsilon - 2\delta$ of the original C mass in the desired cells.

Strip away the ∞ row. Allocate the remaining mass in the column stacks after expanding D to include all cells (i, j), for $0 \leq i \leq k$ and $0 \leq j \leq k$.

So far we have only decided the allocation of masses $M(i, j)$ between the cells. Within the cells distribute according to the product measures

$$M(i, j) \quad Q(\cdot \mid B_i) \otimes P(\cdot \mid B_j).$$

The resulting M on $\mathscr{X} \otimes \mathscr{X}$ has marginal measures P and Q. For example, within B_0 the column marginal is

$$\sum_i M(i, 0) Q(B_i \mid B_i) P(\cdot \mid B_0) = P(B_0) P(\cdot \mid B_0) = P(\cdot B_0).$$

The M measure concentrates at least $1 - \varepsilon - 2\delta$ of its mass within the original D, a cluster of cells each of diameter less than δ in both row and column directions. For a point (x, y) lying in a cell (i, j) of this cluster, there exists points z_i and z_j with

$$d(x, z_i) < \delta, \qquad d(z_i, z_j) \leq \varepsilon, \qquad d(z_j, y) < \delta,$$

which gives $d(x, y) < \varepsilon + 2\delta$. Put another way, if X and Y denote the coordinate projections then

$$\mathbb{P}\{d(X, Y) \geq \varepsilon + 2\delta\} \leq \varepsilon + 2\delta.$$

As δ can be chosen arbitrarily small, and ε can be chosen as close to $\rho(P, Q)$ as we please, we have the desired result.

Problem 17 gives a condition under which the bound $\rho(P, Q)$ can be achieved by a coupling of P and Q. □

IV.5. Weakly Convergent Subsequences

A reader not interested in existence theorems could skip this section, which presents a method for constructing measures on metric spaces. The results will be used in Section V.3 to prove existence of the brownian bridge. The method will be generalized in Chapter VII.

We saw in Section III.6 how to modify the quantile-transformation construction of the one-dimensional Representation Theorem to turn it into an existence theorem, a method for constructing a probability measure as the distribution of the almost sure limit of a sequence of random variables. We had to impose a uniform tightness constraint to stop the sequence from drifting off to infinity. The analogous result for probabilities on metric spaces plays a much more important role than in euclidean spaces, because existence theorems of any sort are so much harder to come by in abstract spaces. Again the key to the construction is a uniform tightness property, which ensures that sequences that ought to converge really do converge. The setting is still that of a metric space $\mathscr{X}$ equipped with a sub-σ-field $\mathscr{A}$ of its borel σ-field.

28 Definitions. Call a probability measure P on $\mathscr{A}$ *tight* if for every $\varepsilon > 0$ there exists a compact set $K(\varepsilon)$ of completely regular points such that $PK(\varepsilon) > 1 - \varepsilon$.

Call a sequence $\{P_n\}$ of probability measures on $\mathscr{A}$ *uniformly tight* if for every $\varepsilon > 0$ there exists a compact set $K(\varepsilon)$ of completely regular points such that $\liminf P_n G > 1 - \varepsilon$ for every open, $\mathscr{A}$-measurable G containing $K(\varepsilon)$. □

Problem 7 justifies the implicit assumption of $\mathscr{A}$-measurability for the $K(\varepsilon)$ in the definition of tightness; every compact set of completely regular points can be written as a countable intersection of open, $\mathscr{A}$-measurable sets.

If G is replaced by $K(\varepsilon)$, the uniform tightness condition becomes a slightly tidier, but stronger, condition. It is, however, more natural to retain the open G. If $P_n \rightsquigarrow P$ and P is tight then, by virtue of the results proved in Example 17, the liminf condition for open G is satisfied; it might not be satisfied if G were replaced by $K(\varepsilon)$. More importantly, one does not need the stronger condition to get weakly convergent subsequences, as will be shown in the next theorem.

For the proof of the theorem we shall make use of a property of compact sets:

> If $\{x_n\}$ is a Cauchy sequence in a metric space, and if $d(x_n, K) \to 0$ for some fixed compact set K, then $\{x_n\}$ converges to a point of K.

This follows easily from one of a set of alternative characterizations of compactness in metric spaces. As we shall be making free use of these characterizations in later chapters, a short digression on the topic will not go amiss.

To prove the assertion we have only to choose, according to the definition of $d(x_n, K)$, points $\{y_n\}$ in K for which $d(x_n, y_n) \to 0$. From $\{y_n\}$ we can extract a subsequence converging to a point y in K. For if no subsequence of $\{y_n\}$ converged to a point of K, then around each x in K we could put an open neighborhood G_x that excluded y_n for all large enough values of n. This would imply that $\{y_n\}$ is eventually outside the union of the finite collection of G_x sets covering the compact K, a contradiction. The corresponding subsequence of $\{x_n\}$ also converges to y. The Cauchy property forces $\{x_n\}$ to follow the subsequence in converging to y.

A set with the property that every sequence has a convergent subsequence (with limit point in the set) is said to be sequentially compact. Every compact set is sequentially compact. This leads to another characterization of compactness:

> A sequentially compact set is complete (every Cauchy sequence converges to a point of the set) and totally bounded (for every positive ε, the set can be covered by a finite union of closed balls of radius less than ε).

For clearly a Cauchy sequence in a sequentially compact K must converge to the same limit as the convergent subsequence. And if K were not totally bounded, there would be some positive ε for which no finite collection of balls of radius ε could cover K. We could extract a sequence $\{x_n\}$ in K with x_{n+1} at least ε away from each of $x_1, \ldots, x_n$ for every n. No subsequence of $\{x_n\}$ could converge, in defiance of sequential compactness.

For us the last link in the chain of characterizations will be the most important:

> A complete, totally bounded subset of a metric space is compact.

Suppose, to the contrary, that $\{G_i\}$ is an open cover of a totally bounded set K for which no finite union of $\{G_i\}$ sets covers K. We can cover K by a finite union of closed balls of radius $\frac{1}{2}$, though. There must be at least one such ball, B_1 say, for which $K \cap B_1$ has no finite $\{G_i\}$ subcover. Cover $K \cap B_1$ by finitely many closed balls of radius $\frac{1}{4}$. For at least one of these balls, B_2 say, $K \cap B_1 \cap B_2$ has no finite $\{G_i\}$ subcover. Continuing in this way we discover a sequence of closed balls $\{B_n\}$ of radii $\{2^{-n}\}$ for which $K \cap B_1 \cap \cdots \cap B_n$ has no finite $\{G_i\}$ cover. Choose a point x_n from this (necessarily non-empty) intersection. The sequence $\{x_n\}$ is Cauchy. If K were also complete, $\{x_n\}$ would converge to some x in K. Certainly x would belong to some G_i, which would necessarily contain B_n for n large enough. A single G_i is about as finite a subcover as one could wish for. Completeness would indeed force $\{G_i\}$ to have a finite subcover for K. End of digression.

29 Compactness Theorem. *Every uniformly tight sequence of probability measures contains a subsequence that converges weakly to a tight borel measure.*

Proof. Write $\{P_n\}$ for the uniformly tight sequence, and K_k for the compact set $K(\varepsilon_k)$, for a fixed sequence $\{\varepsilon_k\}$ that converges to zero. We may assume that $\{K_k\}$ is an increasing sequence of sets.

The proof will use a coupling to represent a subsequence of $\{P_n\}$ by an almost surely convergent sequence of random elements. The limit of these random elements will concentrate on the union of the compact K_k sets; it will induce the tight borel measure on $\mathscr{X}$ to which the subsequence $\{P_n\}$ will converge weakly.

Complete regularity of each point in K_k allows us to cover K_k by a collection of open $\mathscr{A}$-measurable sets, each of diameter less than ε_k. Invoke compactness to extract a finite subcover, $\{U_{ki}: 1 \leq i \leq i_k\}$. Define $\mathscr{A}_m$ to be the finite subfield of $\mathscr{A}$ generated by the open sets U_{ki} for $1 \leq k \leq m$ and $1 \leq i \leq i_k$.

The union of the fields $\{\mathscr{A}_m\}$ is a countable subfield $\mathscr{A}_\infty$ of $\mathscr{A}$. Apply Cantor's diagonalization argument to extract a subsequence of $\{P_n\}$ along which $\lim P_n A$ exists for each A in $\mathscr{A}_\infty$. Write λA for this limit. It is a finitely additive measure on the field $\mathscr{A}_\infty$. Avoid the mess of double-subscripting by assuming, with no loss of generality, that the subsequence is $\{P_n\}$ itself.

If $\{P_n\}$ were weakly convergent to a measure P we would be able to deduce that $P(\text{interior of } A) \leq \lambda A \leq P(\text{closure of } A)$ for each A in $\mathscr{A}_\infty$. If we could assume further that P put zero mass on the boundary of each such A, we would know the P measure of enough sets to allow almost surely convergent representing sequences to be constructed as in the Representation Theorem. Unfortunately there is no reason to expect P to cooperate in this way. Instead, we must turn to λ as a surrogate for the unknown, but sought after, probability measure P.

Since λ need not be countably additive, it would be wicked of us to presume the existence of a random element of $\mathscr{X}$ having distribution λ. We must take a more devious approach.

We can build a passable imitation of $\mathscr{A}_\infty$ on the unit interval. Partition $(0, 1)$ into as many intervals as there are atoms of $\mathscr{A}_1$, making the lebesgue measure of each interval $\bar{A}$ equal to the λ measure of the corresponding A in $\mathscr{A}_1$. These intervals generate a finite field $\bar{\mathscr{A}}_1$ on $(0, 1)$. Partition each atom $\bar{A}$ in $\bar{\mathscr{A}}_1$ into as many subintervals as there are atoms of $\mathscr{A}_2$ in A, matching up lebesgue and λ measures as before. The subintervals together generate a second field $\bar{\mathscr{A}}_2$ on $(0, 1)$, finer than $\bar{\mathscr{A}}_1$. Continuing in similar fashion, we set up an increasing sequence of fields $\{\bar{\mathscr{A}}_k\}$ on $(0, 1)$ that fit together in the same way as the fields $\{\mathscr{A}_k\}$ on $\mathscr{X}$. The union of the $\bar{\mathscr{A}}_k$'s is a countable subfield $\bar{\mathscr{A}}_\infty$ of $(0, 1)$. There is a bijection $\bar{A} \leftrightarrow A$ between $\bar{\mathscr{A}}_\infty$ and $\mathscr{A}_\infty$ that preserves inclusion, maps $\bar{\mathscr{A}}_k$ onto $\mathscr{A}_k$, and preserves measure, in the sense that the lebesgue measure of $\bar{A}$ equals λA. The construction ensures that, if η has a Uniform(0, 1) distribution, $\mathbb{P}\{\eta \in \bar{A}\} = \lambda A$ for every A in $\mathscr{A}_\infty$. The random variable η chooses between the sets in $\bar{\mathscr{A}}_k$ in much the same way as a random element X with distribution P would choose between the sets in $\mathscr{A}_k$.

By definition of λ, there exists an $n(k)$ such that

(30) $\quad P_n A \geq (1 - \varepsilon_k)\lambda A \quad$ for every A in $\mathscr{A}_k$ whenever $\quad n \geq n(k)$.

Lighten the notation by assuming that $n(k) = k$. (If you suspect these notational tricks for avoiding an orgy of subsequencing, feel free to rewrite the argument using, by now, triple subscripting.) As in the proof of the Representation Theorem, this allows us to construct a random element X_n, with distribution P_n, by means of an auxiliary random variable ξ that has a Uniform(0, 1) distribution independent of η:

> For each atom A of $\mathscr{A}_n$, if η falls in the corresponding $\bar{A}$ of $\bar{\mathscr{A}}_n$ and $\xi \leq 1 - \varepsilon_n$, distribute X_n on A according to the conditional distribution $P_n(\cdot \mid A)$. If $\xi > 1 - \varepsilon_n$ distribute X_n with whatever conditional distribution is necessary to bring its overall distribution up to P_n.

We have coupled each P_n with lebesgue measure on the unit square.

To emphasize that X_n depends on η, ξ, and the randomization necessary to generate observations on $P_n(\cdot \mid A)$, write it as $X_n(\omega, \eta, \xi)$. Notice that the same η and ξ figure in the construction of every X_n.

It will suffice for us to prove that $\{X_n(\omega, \eta, \xi)\}$ converges to a point $X(\omega, \eta, \xi)$ of K_k for every ω and every pair (η, ξ) lying in a region of probability at least $(1 - \varepsilon_k)^2$, a result stronger than mere almost sure convergence to a point in the union of the compact sets $\{K_k\}$. Problem 16 provides the extra details needed to deduce borel measurability of X.

For each m greater than k, let G_{mk} be the smallest open, $\mathscr{A}_m$-measurable set containing K_k. Uniform tightness tells us that

$$\lambda G_{mk} = \liminf P_n G_{mk} > 1 - \varepsilon_k,$$

which implies $\mathbb{P}\{\eta \in \bar{G}_{mk}\} > 1 - \varepsilon_k$. Define $\bar{G}_k$ as the intersection of the decreasing sequence of sets $\{\bar{G}_{mk}\}$ for $m = k, k + 1, \ldots$. The overbar here is slightly misleading, because $\bar{G}_k$ need not belong to $\bar{\mathscr{A}}_\infty$. But it is a borel subset of (0, 1). Countable additivity of lebesgue measure allows us to deduce that $\mathbb{P}\{\eta \in \bar{G}_k\} \geq 1 - \varepsilon_k$. Notice how we have gotten around lack of countable additivity for λ, by pulling the construction back into a more familiar measure space.

Whenever η falls in $\bar{G}_k$ and $\xi \leq 1 - \varepsilon_k$, which occurs with probability at least $(1 - \varepsilon_k)^2$, the random elements $X_k, X_{k+1}, \ldots$ crowd together into a shrinking neighborhood of a point of K_k. There exists a decreasing sequence $\{A_m\}$ with:

(i) A_m is an atom of $\mathscr{A}_m$;
(ii) A_m is contained in G_{mk};
(iii) $X_m(\omega, \eta, \xi)$ lies in A_m.

Properties (i) and (iii) are consequences of the method of construction for X_m; property (ii) holds because $\bar{G}_k$ is a subset of $\bar{G}_{mk}$. The set G_{mk}, being the smallest open, $\mathscr{A}_m$-measurable set containing K_k, must be contained within the union of those U_{mi} that intersect K_k. The atom A_m must lie wholly within one such U_{mi}, a set of diameter less than ε_m. So whenever η falls in $\overline{G}_k$ and $\xi \leq 1 - \varepsilon_k$, the sequence $\{X_m\}$ satisfies:

(i) $d(X_m(\omega, \eta, \xi), X_n(\omega, \eta, \xi)) \leq \varepsilon_m$ for $k \leq m \leq n$;
(ii) $d(X_m(\omega, \eta, \xi), K_k) \leq \varepsilon_m$ for $k \leq m$.

As explained at the start of the digression, this forces convergence to a point $X(\omega, \eta, \xi)$ of K_k. □

Notes

Any reader uncomfortable with the metric space ideas used in this chapter might consult Simmons (1963, especially Chapters 2 and 5).

The advantages of equipping a metric space with a σ-field different from the borel σ-field were first exploited by Dudley (1966a, 1967a), who developed a weak convergence theory for measures living on the σ-field generated by the closed balls. The measurability problem for empirical processes (Example 2) was noted by Chibisov (1965); he opted for the Skorohod metric. Pyke and Shorack (1968) suggested another way out: $X_n \rightsquigarrow X$ should mean $\mathbb{P}f(X_n) \to \mathbb{P}f(X)$ for all those bounded, continuous f that make $f(X_n)$ and $f(X)$ measurable. They noted the equivalence of this definition to the definition based on the Skorohod metric, for random elements of $D[0, 1]$ converging to a process with continuous sample paths.

Separability has a curious role in the theory. With it, the closed balls generate the borel σ-field (Problem 6); but this can also hold without separability (Talagrand 1978). Borel measures usually have separable support (Dudley 1967a, 1976, Lecture 5).

Alexandroff (1940, 1941, 1943) laid the foundation for a theory of weak convergence on abstract spaces, not necessarily topological. Prohorov (1956) reset the theory in complete, separable metric space, where most probabilistic and statistical applications can flourish. He and LeCam (1957) proved different versions of the Compactness Theorem, whose form (but not the proof) I have borrowed from Dudley (1966a). Weak convergence of baire measures on general topological spaces was thoroughly investigated by Varadarajan (1965). Topsøe (1970) put together a weak convergence theory for borel measures; he used the liminf property for semicontinuous functions (Example 17) to define weak convergence. These two authors made clear the need for added regularity conditions on the limit measure and separation properties on the topology. One particularly nice combination—a completely regular topology and a τ-additive limit measure—corresponds closely to my assumption that limit measures concentrate on separable sets of completely regular points.

The best references to the weak convergence theory for borel measures on metric spaces remain Billingsley (1968, 1971) and Parthasarathy (1967).

Dudley's (1976) lecture notes offer an excellent condensed exposition of both the mathematical theory and the statistical applications.

Example 11 is usually attributed to Wichura (1971), although Hájek (1965) used a similar approximation idea to prove convergence for random elements of $C[0, 1]$.

Skorohod (1956) hit upon the idea of representing sequences that converge in distribution by sequences that converge almost surely, for the case of random elements of complete, separable metric spaces. The proof in Section 3 is adapted from Dudley (1968). He paid more attention to some of the points glossed over in my proof—for example, he showed how to construct a probability space supporting all the $\{X_n\}$. Here, and in Section 5, one needs the existence theorem for product measures on infinite-product spaces. Pyke (1969, 1970) has been a most persuasive advocate of this method for proving theorems about weak convergence. Many of the applications now belong to the folklore.

The uniformity result of Example 19 comes from Ranga Rao (1962); Billingsley and Topsøe (1967) and Topsøe (1970) perfected the idea. Not surprisingly, the original proofs of this type of result made direct use of the dissection technique of Lemma 15. Prohorov (1956) defined the Prohorov metric; Dudley (1966b) defined the bounded Lipschitz metric.

Strassen (1965) invoked convexity arguments to establish the coupling characterization of the Prohorov metric (Example 26). My proof comes essentially from Dudley (1968), via Dudley (1976, Lecture 18), who introduced the idea of building a coupling between discrete measures by application of the marriage lemma. The Allocation Lemma can also be proved by the max-flow–min-cut theorem (an elementary result from graph theory; for a proof see Bollobás (1979)). The conditions of my Lemma ensure that the minimum capacity of a cut will correspond to the total column mass. Appendix B of Jacobs (1978) contains an exposition of this approach, following Hansel and Troallic (1978). Major (1978) has described more refined forms of coupling.

Problems

[1] Suppose the empirical process U_2 were measurable with respect to the borel σ-field on $D[0, 1]$ generated by the uniform metric. For each subset A of $(1, 2)$ define J_A as the open set of functions in $D[0, 1]$ with jumps at some pair of distinct points t_1 and t_2 in $[0, 1]$ with $t_1 + t_2$ in A. Define a non-atomic measure on the class of all subsets of $(1, 2)$ by setting $\gamma(A) = \mathbb{P}\{U_2 \in J_A\}$. This contradicts the continuum hypothesis (Oxtoby 1971, Section 5). Manufacture from γ an extension of the uniform distribution to all subsets of $(1, 2)$ if you would like to offend the axiom of choice as well. Extend the argument to larger sample sizes.

[2] Write $\mathscr{A}$ for the σ-field on a set $\mathscr{X}$ generated by a family $\{f_i\}$ of real-valued functions on $\mathscr{X}$. That is, $\mathscr{A}$ is the smallest σ-field containing $f_i^{-1}B$ for each i and each borel set B. Prove that a map X from $(\Omega, \mathscr{E})$ into $\mathscr{X}$ is $\mathscr{E}/\mathscr{A}$-measurable if and only if the composition $f_i \circ X$ is $\mathscr{E}/\mathscr{B}(\mathbb{R})$-measurable for each i.

[3] Every function in $D[0, 1]$ is bounded: $|x(t_n)| \to \infty$ as $n \to \infty$ would violate either the right continuity or the existence of the left limit at some cluster point of the sequence $\{t_n\}$.

[4] Write $\mathscr{P}$ for the projection σ-field on $D[0, 1]$ and $\mathscr{B}_0$ for the σ-field generated by the closed balls of the uniform metric. Write π_t for the projection map that takes an x in $D[0, 1]$ onto its value $x(t)$.

(a) Prove that each π_t is $\mathscr{B}_0$-measurable. [Express $\{x: \pi_t x > \alpha\}$ as a countable union of closed balls $B(x_n, n)$, where x_n equals α plus $(n + n^{-1})$ times the indicator function of $[t, t + n^{-1})$.] Deduce that $\mathscr{B}_0$ contains $\mathscr{P}$.

(b) Prove that the σ-field $\mathscr{P}$ contains each closed ball $B(x, r)$. [Express the ball as an intersection of sets $\{z: |\pi_t x - \pi_t z| \le r\}$, with t rational.] Deduce that $\mathscr{P}$ contains $\mathscr{B}_0$.

[5] Let $\{G_i\}$ be a family of open sets whose union covers a separable subset C of a metric space. Adapt the argument of Lemma 7 to prove that C is contained in the union of some countable subfamily of the $\{G_i\}$. [This is Lindelöf's theorem.]

[6] Every separable, open subset of a metric space can be written as a countable union of closed balls. [Rational radii, centered at points of the countable dense set.] The closed balls generate the borel σ-field on a separable metric space.

[7] Every closed, separable set of completely regular points belongs to $\mathscr{A}$. [Cover it with open, $\mathscr{A}$-measurable sets of small diameter. Use Lindelöf's theorem to extract a countable subcover. The union of these sets belongs to $\mathscr{A}$. Represent the closed set as a countable intersection of such unions.]

[8] Let C_0 be the countable subset of $C[0, 1]$ consisting of all piecewise linear functions with corners at only a finite set of rational pairs (t_i, r_i). Argue from uniform continuity to prove that $C[0, 1]$ equals the closure of C_0. Deduce that $C[0, 1]$ is a projection-measurable subset of $D[0, 1]$.

[9] A function h is said to be lower-semicontinuous at a point x if, for each $M < h(x)$, h is greater than M in some neighborhood of x. To say h is lower-semicontinuous means that it is lower-semicontinuous at every point. Show that the upper envelope of any set of continuous functions is lower-semicontinuous. Adapt the construction of Lemma 7 to prove that every lower-semicontinuous function that is bounded below can be represented on a separable set of completely regular points as the pointwise limit of an increasing sequence of continuous functions. How would one define upper-semicontinuity? Which sets should have upper-semicontinuous indicator functions? What does a combination of both semicontinuities imply?

[10] If $X_n \rightsquigarrow X$ as random elements of a metric space $\mathscr{X}$ and $d(X_n, Y_n) \to 0$ in probability, then $Y_n \rightsquigarrow X$, provided that $\mathbb{P}_X$ concentrates on a separable set of completely regular points. [Convergence in probability means $\mathbb{P}^*\{d(X_n, Y_n) > \varepsilon\} \to 0$ for each $\varepsilon > 0$.]

[11] Let P be a borel measure on a metric space. For every borel set B there exists an open G_ε containing B and a closed F_ε contained in B with $P(G_\varepsilon \backslash F_\varepsilon) < \varepsilon$. [The class of all sets with this property forms a σ-field. Each closed set has the property because it can be written as a countable intersection of open sets.] Deduce that P is uniquely determined by the values it gives to closed sets. Extend the result to measures defined on the σ-field generated by the closed balls.

[12] Suppose limsup $P_n F \leq PF$ for each closed, $\mathscr{A}$-measurable set F. Prove that $P_n \rightsquigarrow P$ by applying the inequalities

$$k^{-1}\sum_{i=1}^{\infty}\{f \geq i/k\} \quad \leq \quad f \quad \leq \quad k^{-1} + k^{-1}\sum_{i=1}^{\infty}\{f \geq i/k\}$$

for each non-negative f in $\mathscr{C}(\mathscr{X}; \mathscr{A})$. [The summands are identically zero for all i large enough. Apply the same argument to $-f +$ (a big constant).]

[13] If $P_n B \to PB$ for each $\mathscr{A}$-measurable set B whose boundary has zero P measure then $P_n \rightsquigarrow P$. [Replace the levels i/k of the previous problem by levels t_i for which $P\{f = t_i\} = 0$.]

[14] The functions in $\mathscr{C}(\mathscr{X}; \mathscr{A})$ generate a sub-σ-field $\mathscr{B}_c$ of $\mathscr{A}$. A map X from $(\Omega, \mathscr{E})$ into $\mathscr{X}$ is $\mathscr{E}/\mathscr{B}_c$-measurable if and only if $f(X)$ is $\mathscr{E}/\mathscr{B}(\mathbb{R})$-measurable for each f in $\mathscr{C}(\mathscr{X}; \mathscr{A})$.

[15] (Continued). The trace of $\mathscr{B}_c$ on any separable set S of completely regular points of $\mathscr{X}$ coincides with the borel σ-field on S. [Sets of the form $\{f > 0\} \cap S$, with f in $\mathscr{C}(\mathscr{X}; \mathscr{A})$, form a basis for the relative topology on S. Every relatively open subset of S is a countable union of such sets, by Lindelöf's theorem.]

[16] (Continued). Let $\{X_m\}$ be a sequence of $\mathscr{E}/\mathscr{A}$-measurable random elements of $\mathscr{X}$ that converges pointwise to a map X. Prove that X is $\mathscr{E}/\mathscr{B}_c$-measurable. If X takes values only in a fixed separable set of completely regular points, then it is $\mathscr{E}/\mathscr{B}(\mathscr{X})$-measurable.

[17] Let P and Q be tight probability measures on the borel σ-field of a separable metric space $\mathscr{X}$. There exists a coupling for which $\mathbb{P}\{d(X, Y) \geq \Delta\} \leq \Delta$, where $\Delta = \rho(P, Q)$, the Prohorov distance between P and Q. [From Example 26, there exist measures M_n on $\mathscr{X} \otimes \mathscr{X}$, with marginals P and Q, for which

$$M_n\{(x, y): d(x, y) \geq \Delta + n^{-1}\} \leq \Delta + n^{-1}.$$

The sequence $\{M_n\}$ is uniformly tight. The limit of a weakly convergent subsequence defines the required coupling. Is separability of $\mathscr{X}$ really needed?]

[18] Let $\mathscr{X}$ be a compact metric space, and $\mathscr{C}(\mathscr{X})$ be the vector space of all bounded, continuous, real functions on $\mathscr{X}$. Let T be a non-negative linear functional on $\mathscr{C}(\mathscr{X})$ with $T1 = 1$. These steps show that $Tf = Pf$ for some borel probability measure P:

(a) Given $\gamma > 0$ find functions $g_1, \ldots, g_k$ in $\mathscr{C}^+(\mathscr{X})$ with diameter$\{g_i > 0\} < 2\gamma$ and $g_1 + \cdots + g_k = 1$. [Find f_x in $\mathscr{C}^+(\mathscr{X})$ with $f_x(x) > 0$ and $f_x(y) = 0$ for $d(y, x) > \gamma$. Cover K by finitely many open sets $\{f_x > 0\}$. Standardize the corresponding functions to sum to one everywhere.]

(b) Choose x_i with $g_i(x_i) > 0$. Define P_γ as the discrete probability measure that puts mass Tg_i at x_i, for each i. If h belongs to $\mathscr{C}(\mathscr{X})$, show that $|P_\gamma h - Th| \to 0$ as $\gamma \to 0$.

(c) Extract a subsequence of $\{P_\gamma\}$ that converges weakly. Show that the limit measure P has the desired property.

CHAPTER V

The Uniform Metric on Spaces of Cadlag Functions

... in which random elements of metric spaces of cadlag functions—stochastic processes whose sample paths have at worst simple jump discontinuities—are treated. Necessary and sufficient conditions for convergence in distribution are found then specialized to prove limit theorems for empirical processes and processes with independent increments. The two most commonly occurring limit processes, brownian motion and brownian bridge, are studied.

V.1. Approximation of Stochastic Processes

The theory developed in Chapter IV justifies its existence by what it has to say about the limiting distributions of functionals defined on sequences of stochastic processes. Processes $\{X_n(t): t \in T\}$ are identified with random elements of some space $\mathscr{X}$ of functions on T, a space large enough to contain the sample paths of every X_n; the functionals are maps defined on $\mathscr{X}$ to which the Continuous Mapping Theorem can be applied. In this chapter we shall specialize the general theory to the particular function space $D[0, 1]$, under its uniform metric. It will turn out that most applications, especially those that come up with brownian bridges and brownian motions as limit processes, require no fancier setting than this.

Recall that for a function defined on a subset T of the extended real line $[-\infty, \infty]$ to deserve the title *cadlag* it must be right continuous ("continue à droite") and have left limits ("limites à gauche") at each point of T. Many of the well-studied stochastic processes—processes with independent increments, or the markov property, or some form of martingale structure—have versions with cadlag sample paths. For $T = [0, 1]$, these processes can be analyzed as random elements of $D[0, 1]$, the set of all real-valued cadlag functions on $[0, 1]$. Notice that right continuity at 1 and existence of the left limit at 0 become empty requirements for this space. The analysis for other spaces of cadlag functions on compact intervals (such as $D[-\infty, \infty]$, the natural space in which to study empirical distribution functions over the real line), and the extensions to non-compact index sets (Section 5), will call for only slight variations on the methods developed for $D[0, 1]$.

For each finite subset S of $[0, 1]$ write π_S for the projection map from $D[0, 1]$ into $\mathbb{R}^S$ that takes an x onto its vector of values $\{x(t): t \in S\}$. Abbreviate $\pi_{\{t\}}$ to π_t. These projections generate the projection σ-field, $\mathscr{P}$. Questions of measurability in this chapter will always refer to this σ-field. A stochastic process X on $(\Omega, \mathscr{E}, \mathbb{P})$ with sample paths in $D[0, 1]$, such as an empirical process, is $\mathscr{E}/\mathscr{P}$-measurable provided $\pi_t \circ X$ is $\mathscr{E}/\mathscr{B}(\mathbb{R})$-measurable for each fixed t (Problem IV.2).

Probability measures on $\mathscr{P}$ are uniquely determined by the values they give to the generating sets $\{\pi_S^{-1}B\}$, with S a finite subset of $[0, 1]$ and B a borel subset of $\mathbb{R}^S$. Equivalently, the distribution of a random element of $D[0, 1]$ is uniquely specified by giving the distributions of all its finite-dimensional projections.

As every cadlag function on $[0, 1]$ is bounded (Problem IV.3), the uniform distance

$$d(x, y) = \|x - y\| = \sup\{|x(t) - y(t)| : 0 \leq t \leq 1\}$$

defines a metric on $D[0, 1]$. No other metric will be used for $D[0, 1]$ in this chapter. The closed balls for d generate $\mathscr{P}$ but not the larger borel σ-field (Problem IV.4). Every point in $D[0, 1]$ is completely regular—see Definition IV.6 and the discussion preceding it.

The difficulty with the σ-field, which can be blamed on the lack of a countable, dense subset of functions in $D[0, 1]$, has dissuaded many authors from working with the uniform metric. Chapter IV showed that the difficulty can be surmounted, at least when limit distributions concentrate on separable subsets of $D[0, 1]$. As compensation for persisting with the uniform metric, we shall find its topological properties much easier to understand and manipulate than those of its main competitor, the Skorohod metric, which will be discussed in Chapter VI. That will make life more pleasant for us in Section 3 when we come to apply the Compactness Theorem.

The limit processes for the applications in this chapter will always concentrate in a separable subset of $D[0, 1]$, usually $C[0, 1]$, the set of all continuous, real functions on $[0, 1]$. As a closed (uniform convergence preserves continuity), separable, $\mathscr{P}$-measurable subset of $D[0, 1]$ (Problems IV.8), $C[0, 1]$ has several attractive properties. For example, it inherits completeness from $D[0, 1]$, and its projection σ-field coincides with its borel σ-field for the uniform metric (Problem IV.6).

How do we establish convergence in distribution of a sequence $\{X_n\}$ of random elements of $D[0, 1]$ to a limit process X? Certainly we need the finite-dimensional projections $\{\pi_S X_n\}$ to converge in distribution, as random vectors in $\mathbb{R}^S$, to the finite-dimensional projection $\pi_S X$, for each finite subset S of $[0, 1]$. Continuity of π_S and the Continuous Mapping Theorem make that a necessary condition. The methods of Chapter III usually help here. But that alone could hardly suffice. Continuous functionals such as $Mx = \|x\|$, a typical non-trivial example, depend on x through more than its values at a fixed, finite S. Indeed, direct examination of that very functional gives the clue to the extra condition needed.

Intuitively, it should be possible to approximate MX_n by taking the maximum of $|X_n(s)|$ over a large, finely spaced, finite subset S of $[0, 1]$. If S expands up to a countable, dense subset of $[0, 1]$ containing 1, then

$$M_S X_n = \max_S |X_n(s)| \to MX_n$$

for every sample path of X_n. The cadlag property assures us of this. (Notice the special treatment demanded by 1.) Given any $\delta > 0$ and $\varepsilon > 0$, a large enough S could be found to ensure that

$$\text{(1)} \qquad \mathbb{P}\{|M_S X_n - MX_n| \geq \delta\} < \varepsilon.$$

At first sight this seems to have solved the problem. Because $M_S X_n$ depends continuously on $\pi_S X_n$, it converges in distribution to $M_S X$ as $n \to \infty$. For f bounded and uniformly continuous, $\mathbb{P}f(M_S X_n) \to \mathbb{P}f(M_S X)$. From (1) with δ chosen appropriately,

$$|\mathbb{P}f(M_S X_n) - \mathbb{P}f(MX_n)| \leq \varepsilon + 2\|f\|\varepsilon.$$

A similar inequality holds for X. Taken together, these seem to add up to convergence in distribution of MX_n to MX. But the argument is flawed.

The problem lies with the choice of S in (1). If the sample paths of X_n jumped about more and more wildly as n increased, S would have to get bigger with n. That would undermine the convergence of $\{M_S X_n\}$ to $M_S X$, since finite-dimensional convergence says nothing about $\{\pi_S X_n\}$ for S varying with n. We need the same S to work for each n.

A similar argument can be invoked for any other continuous functional H on $D[0, 1]$. We approximate HX_n by a continuous function of $\pi_S X_n$ for some large, finite set S not depending on n. We construct the approximation by applying H to an element $A_S X_n$ in $D[0, 1]$. For simplicity, suppose S has been rearranged into increasing order and augmented by the points 0 and 1, if necessary, to form a grid $0 = s_0 < s_1 < \cdots < s_k = 1$. For x in $D[0, 1]$ define the approximating path $A_S x$ by

$$\text{(2)} \qquad (A_S x)(t) = x(s_i) \quad \text{for} \quad s_i \leq t < s_{i-1},$$

and $(A_S x)(1) = x(1)$. Notice that $A_S x$ depends on x only through $\pi_S x$.

In order that $H(A_S x)$ be close to Hx, it suffices that $\|A_S x - x\|$ be small. Consider, for example, a uniformly continuous H. There exists a $\delta > 0$ such that $|H(A_S x) - Hx| < \varepsilon$ whenever $\|A_S x - x\| < \delta$; the random variable HX_n lies within ε of $H(A_S X_n)$ with probability no less than $\mathbb{P}\{\|A_S X_n - X_n\| < \delta\}$. Again, the same grid would have to work for every X_n—or at least for every X_n with n large enough—to allow finite-dimensional convergence to imply convergence in distribution of $\{A_S X_n\}$ to $A_S X$. (The example of the empirical process U_n, whose sample paths have jumps of size $n^{-1/2}$, shows that uniform approximation of X_n by $A_S X_n$ can only be required for large values of n, anyway.) If such a map A_S does exist, the argument sketched above for the supremum functional M will carry over to the functional H, proving that $HX_n \rightsquigarrow HX$. (The argument might seem familiar—it is a special case of Example IV.11).

For the sake of brevity, from now on shorten "finite-dimensional distributions" to fidis and "finite-dimensional projections" to fidi projections.

3 Theorem. *Let $X, X_1, X_2, \ldots$ be random elements of $D[0, 1]$ (under its uniform metric and projection σ-field). Suppose $\mathbb{P}\{X \in C\} = 1$ for some separable subset C of $D[0, 1]$. The necessary and sufficient conditions for $\{X_n\}$ to converge in distribution to X are:*

(i) *the fidis of X_n converge to the fidis of X; that is, $\pi_S X_n \rightsquigarrow \pi_S X$ for each finite subset S of $[0, 1]$;*
(ii) *to each $\varepsilon > 0$ and $\delta > 0$ there corresponds a grid $0 = t_0 < t_1 < \cdots < t_m = 1$ such that*

$$\text{(4)} \qquad \limsup \mathbb{P}\left\{\max_i \sup_{J_i} |X_n(t) - X_n(t_i)| > \delta\right\} < \varepsilon,$$

where J_i denotes the interval $[t_i, t_{i+1})$, for $i = 0, 1, \ldots, m - 1$.

Proof. Suppose that $X_n \rightsquigarrow X$. The projection map π_S is both projection-measurable (by definition) and continuous. Condition (i) follows by the Continuous Mapping Theorem.

To simplify the proof of (ii) suppose that the separable subset C equals $C[0, 1]$. Continuity of the sample paths makes the choice of the grid in (ii) easier. (Problem 4 outlines the extra arguments needed for the general case.) Let $\{s_0, s_1, \ldots\}$ be a countable, dense subset of $[0, 1]$. To avoid trivialities, assume that $s_0 = 0$ and $s_1 = 1$. Write A_k for the interpolation map constructed as in (2) from the grid obtained by rearranging $s_0, \ldots, s_k$ into increasing order. For fixed x in $C[0, 1]$, the distance $\|A_k x - x\|$ converges to zero as k increases, by virtue of the uniform continuity of x. Applied to the sample paths of X, this shows that $\|A_k X - X\| \to 0$ almost surely. Convergence in probability would be enough to assure the existence of some k for which

$$\text{(5)} \qquad \mathbb{P}\{\|A_k X - X\| \geq \delta\} < \varepsilon.$$

Choose and fix such a k. Because $\|A_k x - x\|$ varies continuously with x, the set

$$F = \{x \in D[0, 1]: \|A_k x - x\| \geq \delta\}$$

is closed. By Example IV.17 (inequality IV.18 to be precise),

$$\limsup \mathbb{P}\{X_n \in F\} \leq \mathbb{P}\{X \in F\}.$$

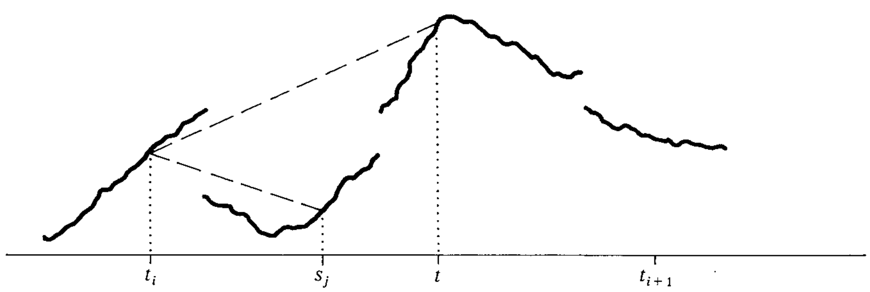
The left-hand side bounds the limsup in (4). We could require "$\geq \delta$" rather than "$> \delta$" in (4), but that would create a minor complication in the next lemma.

Now let us show that (i) and (ii) imply $X_n \rightsquigarrow X$. Retain the assumption that X has continuous sample paths (see Problem 4 for the general case), so that (5) still holds. Choose any bounded, uniformly continuous, projection-measurable, real function f on $D[0, 1]$. Given $\varepsilon > 0$ find $\delta > 0$ such that $|f(x) - f(y)| < \varepsilon$ whenever $\|x - y\| \le \delta$. Write A_T for the approximation map constructed from the grid in (ii) corresponding to this δ and ε. Condition (4) becomes

$$\limsup \mathbb{P}\{\|A_T X_n - X_n\| > \delta\} < \varepsilon.$$

With no loss of generality we may assume that the A_k of (5) equals A_T, because if we combine the two underlying grids we have at worst to replace δ by 2δ to preserve the two approximations. For example, if $t_i \le s_j \le t < t_{i+1}$ then, whenever $\|A_T X_n - X_n\| < \delta$,

$$|X_n(t) - X_n(s_j)| \le |X_n(t) - X_n(t_i)| + |X_n(s_j) - X_n(t_i)| < 2\delta.$$



The argument now follows the lines sketched out before. Write the composition $f \circ A_T$ as $g \circ \pi_T$ where g is a bounded, continuous function on $D[0, 1]$—a fancy way of saying that $f(A_T x)$ depends on x continuously through the values that x takes at the grid points.

$$\begin{aligned}
&|\mathbb{P}f(X_n) - \mathbb{P}f(X)| \\
&\quad \le \mathbb{P}|f(X_n) - f(A_T X_n)| + |\mathbb{P}f(A_T X_n) - \mathbb{P}f(A_T X)| \\
&\qquad + \mathbb{P}|f(A_T X) - f(X)| \\
&\quad \le \varepsilon + 2\|f\|\mathbb{P}\{\|X_n - A_T X_n\| > \delta\} + |\mathbb{P}g(\pi_T X_n) - \mathbb{P}g(\pi_T X)| \\
&\qquad + \varepsilon + 2\|f\|\mathbb{P}\{\|X - A_T X\| > \delta\}.
\end{aligned}$$

The middle term in the last line converges to zero as $n \to \infty$, because of the fidi convergence $\pi_T X_n \rightsquigarrow \pi_T X$. □

We now know the price we pay for wanting to make probability statements about functionals that depend on the whole sample path of a stochastic process: with high probability we need to rule out nasty behavior between the grid points. For uniform approximation of sample paths, a large value of $|X_n(t) - X_n(t_i)|$ would be nasty; inequality (4) rules it out. So how do we control the left-hand side of (4)? It involves the probability of a union of m events, which we may bound by the sum of the probabilities of those events,

$$\sum_{i=0}^{m-1} \mathbb{P}\left\{\sup_{J_i} |X_n(t) - X_n(t_i)| > \delta\right\}. \tag{6}$$

Then we can concentrate on what happens in an interval J_i between adjacent grid points. For many stochastic processes, good behavior of the increment $X_n(t_{i+1}) - X_n(t_i)$ forces good behavior for the whole segment of sample path over J_i.

7 Lemma. *Let $\{Z(t): 0 \le t \le b\}$ be a process with cadlag sample paths taking the value zero at $t = 0$. Suppose $Z(t)$ is $\mathscr{E}_t$-measurable, for some increasing family of σ-fields $\{\mathscr{E}_t: 0 \le t \le b\}$. If at each point of $\{|Z(t)| > \delta\}$,*

$$\mathbb{P}\{|Z(b) - Z(t)| \le \tfrac{1}{2}|Z(t)| \,|\, \mathscr{E}_t\} \ge \beta,$$

where β is a positive number depending only on δ, then

$$\mathbb{P}\left\{\sup_{0 \le t \le b} |Z(t)| > \delta\right\} \le \beta^{-1}\mathbb{P}\{|Z(b)| > \tfrac{1}{2}\delta\}.$$

Proof. Let S be a finite subset of $[0, b]$ containing the point b. By virtue of right continuity, as S expands up to a countable, dense subset of $[0, b]$,

$$\max_S |Z(t)| \to \sup_{[0,b]} |Z(t)| \quad \text{for every sample path of } Z.$$

Notice the minor complication here if "$> \delta$" were replaced by "$\ge \delta$". It is good enough to establish the inequality with the supremum taken over S.

Define τ as the first point of S for which $|Z(t)| > \delta$ if there is such a t, otherwise set $\tau = \infty$. The event $\{\tau = t\}$ belongs to $\mathscr{E}_t$. For distinct points t and t' in S the events $\{\tau = t\}$ and $\{\tau = t'\}$ are disjoint. This justifies a first-passage decomposition.

$$\begin{aligned}
\mathbb{P}\left\{\max_S |Z(t)| > \delta\right\} &= \sum_S \mathbb{P}\{\tau = t\} \\
&\le \beta^{-1} \sum_S \mathbb{P}[\{\tau = t\}\mathbb{P}\{|Z(b) - Z(t)| \le \tfrac{1}{2}|Z(t)| \,|\, \mathscr{E}_t\}] \\
&= \beta^{-1} \sum_S \mathbb{P}\{\tau = t, |Z(b) - Z(t)| \le \tfrac{1}{2}|Z(t)|\} \\
&\le \beta^{-1} \sum_S \mathbb{P}\{\tau = t, |Z(b)| > \tfrac{1}{2}\delta\} \\
&\le \beta^{-1}\mathbb{P}\{|Z(b)| > \tfrac{1}{2}\delta\}.
\end{aligned}$$

□

For the lemma to help us we need to know something about the conditional distribution of $Z(b) - Z(t)$ given what has happened up to time t. When do we have such information? Here are some possibilities. The distribution of the increment might not depend on the past at all—the process might have independent increments. In that case the requirement imposed by the lemma reduces to something involving the marginal distributions of the increments. Or the distribution of $Z(b) - Z(t)$ might depend on the past only through the value of $Z(t)$—the markov property. In that case the requirement turns into a condition on the transition probabilities. Or $Z(b) - Z(t)$ might merely have zero conditional expectation—the martingale property. And then we have all those clever stopping time and maximal inequality tricks to play with.

In each of these cases it turns out to be quite easy to find sufficient conditions for strengthening fidi convergence to the full convergence in distribution of the processes as random elements of $D[0, 1]$. Sections 4 and 2 illustrate the first two possibilities; Chapter VIII will play tricks with martingales. In each case the limit process will be either a brownian bridge, a brownian motion, or a close relative.

8 Definition. A brownian motion is a process with continuous sample paths and

(i) $B(0) = 0$;
(ii) for $0 \leq t_1 < \cdots < t_k$ the increments $B(t_2) - B(t_1), \ldots, B(t_k) - B(t_{k-1})$ are mutually independent; each $B(t_i) - B(t_{i-1})$ has a $N(0, t_i - t_{i-1})$ distribution.

A brownian bridge, or tied-down brownian motion, is a process on $[0, 1]$ with continuous sample paths and

(i) $U(0) = U(1) = 0$;
(ii) for each finite subset S of $[0, 1]$ the random vector $\pi_S U$ has a multivariate normal distribution with zero means and covariances given by $\mathbb{P}U(s)U(t) = s(1 - t)$ for $s \leq t$. □

The name tied-down brownian motion comes from one method for constructing a random element distributed like U. Apply to B the continuous transformation that takes the function $x(t)$ onto the function $x(t) - tx(1)$. That is, tie down the loose end of B at $t = 1$ to force upon it the constraint (i) placed on U. The tied-down process has gaussian fidi projections whose means and covariances agree with those for the brownian bridge; the processes generate the same distribution on the projection σ-field.

V.2. Empirical Processes

From a sequence $\{\xi_i\}$ of independent random variables, each having a Uniform(0, 1) distribution, construct the empirical process

$$U_n(t) = n^{-1/2} \sum_{i=1}^{n} [\{\xi_i \leq t\} - t] \quad \text{for} \quad 0 \leq t \leq 1.$$

For fixed finite S, the projection $\pi_S U_n$ is a normed sum of independent, identically distributed random vectors. By the Multivariate Central Limit Theorem (III.30) it converges to a zero-mean multivariate normal distribution with the same covariance structure as the brownian bridge; that is, the fidis of U_n converge to the fidis of U. (No accident of course—that's why U got the covariances it did.) By design, the first condition of Theorem 3 is satisfied. A markov property of U_n will strengthen this to convergence in distribution.

9 Empirical Central Limit Theorem (Uniform Case). *The empirical processes $\{U_n\}$ constructed by independent sampling from Uniform*(0, 1) *converge in distribution, as random elements of* $D[0, 1]$, *to the brownian bridge.*

PROOF. We need to find a grid $0 = t_0 < t_1 < \cdots < t_m = 1$ satisfying (4), for fixed ε and δ. Do this by making the sum in (6), with X_n replaced by U_n, small. Take the $\{t_i\}$ equally spaced. By reasons of symmetry, or stationarity if you prefer, the sum reduces to $m\mathbb{P}\{\sup_{0\le t\le b} |U_n(t)| > \delta\}$, where $b = m^{-1}$. We seek an m that makes it less than ε.

Take $\mathscr{E}_t$ as the σ-field generated by $U_n(s)$ for $0 \le s \le t$. It tells us how many of the observations $\xi_1, \ldots, \xi_n$ have landed in $[0, t]$, and where they lie.

Suppose we know that $[0, t]$ contains exactly k of the observations. Given this information, the other $n - k$ observations distribute themselves uniformly in the interval $(t, 1]$. Formally, on $\{U_n(t) = n^{-1/2}(k - nt)\}$ the conditional distribution of $U_n(b) - U_n(t)$ given $\mathscr{E}_t$ is

$$n^{-1/2}[\text{Bin}(n - k, \theta) - n(b - t)],$$

where $\theta = (b - t)/(1 - t)$. Notice the markov property: the conditional distribution depends only on $U_n(t)$. Apply Tchebychev's inequality on the set where $|U_n(t)| > \delta$, that is, where $|k - nt| \ge n^{1/2}\delta$.

$$\begin{aligned}
&\mathbb{P}\{|U_n(b) - U_n(t)| > \tfrac{1}{2}|U_n(t)|\,|\mathscr{E}_t\} \\
&\quad = \mathbb{P}\{|\text{Bin}(n - k, \theta) - n(b - t)| > \tfrac{1}{2}|k - nt|\} \\
&\quad \le 4[(n - k)\theta(1 - \theta) + [(n - k)\theta - n(b - t)]^2]/(k - nt)^2 \\
&\quad \le 4n\theta/(k - nt)^2 + 4\theta^2 \\
&\quad \le [4b/(1 - b)]/\delta^2 + 4b^2/(1 - b)^2 \\
&\quad \le \tfrac{1}{2} \quad \text{for small enough values of } b.
\end{aligned}$$

Notice that the argument would break down if we replaced $\frac{1}{2}|U_n(t)|$ by $\frac{1}{2}\delta$ in the first line. Apply Lemma 7 with $\beta = \frac{1}{2}$ and m large enough.

$$m\mathbb{P}\left\{\sup_{0\le t\le b} |U_n(t)| > \delta\right\} \le 2m\mathbb{P}\{|U_n(b)| > \tfrac{1}{2}\delta\}.$$

With m fixed, let $n \to \infty$. Invoke fidi convergence to make the right-hand side of the inequality converge to

$$\begin{aligned}
2m\mathbb{P}\{|N(0, b - b^2)| > \tfrac{1}{2}\delta\} &\le 2m[(b - b^2)^2(\tfrac{1}{2}\delta)^{-4}]\mathbb{P}|N(0, 1)|^4 \\
&\le 32m^{-1}\delta^{-4}\mathbb{P}|N(0, 1)|^4,
\end{aligned}$$

which is less than ε for m large enough. □

The restriction to samples from the uniform distribution was unnecessary. A similar result holds for the empirical process

$$E_n(r) = n^{1/2}[F_n(r) - F(r)] = n^{-1/2} \sum_{i=1}^{n} [\{\eta_i \le r\} - F(r)],$$

constructed by independent sampling from any distribution function F on the real line. Think of E_n as a random element of $D[-\infty, \infty]$, the space of cadlag functions on $[-\infty, \infty]$. Right continuity at $-\infty$ can be achieved by setting $E_n(-\infty)$ equal to zero; the left limit at $+\infty$ also equals zero, the natural value for $E_n(+\infty)$.

It would cause us no great hardship to carry all the theory for $D[0, 1]$ over to $D[-\infty, \infty]$ then track through the proof of the general Empirical Central Limit Theorem. The uniform metric is well-defined, because of the boundedness (Problem 6) of functions in $D[-\infty, \infty]$; the projection σ-field plays the same role as in $D[0, 1]$. This time the limit gaussian process E would have sample paths in $D[-\infty, \infty]$ and multivariate normal fidis with zero means and covariance kernel

$$\mathbb{P}E(r)E(s) = F(r) - F(r)F(s) \quad \text{for} \quad r \le s. \tag{10}$$

The E process need not have continuous sample paths; it jumps where F jumps.

Some small complications would arise with the choice of grid points for the general empirical process, but these could be overcome easily (Problem 7). Everything else would go through in much the same way as for the uniform distribution. There is, however, a much simpler way to prove the general Empirical Central Limit Theorem: use the quantile transformation (Section III.6) to represent E_n as a continuous image of U_n.

11 Empirical Central Limit Theorem. *The empirical processes $\{E_n\}$ constructed by independent sampling from a distribution function F converge in distribution, as random elements of $D[-\infty, \infty]$, to the gaussian process E with covariance kernel given by* (10).

PROOF. Define a map H from $D[0, 1]$ into $D[-\infty, \infty]$ by setting $(Hx)(r) = x(F(r))$. It is measurable (both spaces are equipped with their projection σ-fields) and uniformly continuous:

$$\|Hx - Hy\| = \sup_r |x(F(r)) - y(F(r))| \le \|x - y\|.$$

By the uniform case of the Empirical Central Limit Theorem and the Continuous Mapping Theorem, $HU_n \rightsquigarrow HU$. Remember from Section III.6 that the quantile function has the property: $\xi_i \le F(r)$ if and only if $Q(\xi_i) \le r$.

$$\begin{aligned}(HU_n)(r) &= n^{-1/2} \sum_{i=1}^{n} [\{\xi_i \le F(r)\} - F(r)] \\ &= n^{-1/2} \sum_{i=1}^{n} [\{Q(\xi_i) \le r\} - F(r)].\end{aligned}$$

The random variables $\{Q(\xi_i)\}$ form an independent sample from the distribution F; the last sum has the same distribution as E_n. You can complete the proof by checking that HU satisfies the defining requirements for E. (This provides one method for proving the existence of E.) □

Even though the quantile transformation worked like a charm here, you should beware of applying it unnecessarily to force processes that want to live in $D[-\infty, \infty]$ into migrating to $D[0, 1]$. Gratuitous rescaling of the time axis, perhaps with the aim of reducing a problem to a case involving only uniformly distributed random variables, can complicate an otherwise straightforward argument.

12 Example. The Central Limit Theorem for the sample median, proved by direct methods in Section III.4, can be deduced from the Empirical Central Limit Theorem. As before, assume that the sampling distribution F has a continuous, positive density f in a neighborhood of its median m.

To finesse the problems caused by the jumps in the empirical distribution function F_n, regard any random variable m_n for which

$$F_n(m_n) = \tfrac{1}{2} + o_p(n^{-1/2})$$

as a sample median. From the condition on f,

$$F(t) = \tfrac{1}{2} + (t - m)[f(m) + o(1)].$$

Because m_n must converge to m (Example II.1), we deduce that

$$F(m_n) = \tfrac{1}{2} + (m_n - m)[f(m) + o_p(1)].$$

Combine the expressions for $F_n(m_n)$ and $F(m_n)$ to get

$$E_n(m_n) = -n^{1/2}(m_n - m)[f(m) + o_p(1)] + o_p(1),$$

which rearranges to

$$n^{1/2}(m_n - m) = [-E_n(m_n) - o_p(1)]/[f(m) + o_p(1)]. \tag{13}$$

Think of the right-hand side as a function of the four random elements E_n, m_n, $o_p(1)$, and $o_p(1)$. (Perhaps we should distinguish between the two $o_p(1)$ symbols—they stand for different variables.) Better still, think of it as a function of the random element $(E_n, m_n, o_p(1), o_p(1))$ of $D[-\infty, \infty] \otimes \mathbb{R}^3$. Since $(m_n, o_p(1), o_p(1)) \to (m, 0, 0)$ in probability and $E_n \rightsquigarrow E$, and since Problem 8 shows that E concentrates on a separable subset of $D[-\infty, \infty]$, Example IV.10 lets us deduce that

$$(E_n, m_n, o_p(1), o_p(1)) \rightsquigarrow (E, m, 0, 0). \tag{14}$$

The right-hand side of (13) can be constructed from the left-hand side of (14) through application of the map H from $D[-\infty, \infty] \otimes \mathbb{R}^3$ into $\mathbb{R}$ defined by

$$H(x, \alpha, \beta, \gamma) = [-x(\alpha) - \beta]/[f(m) + \gamma].$$

H is continuous at $(x, m, 0, 0)$ for every x that has m as a point of continuity. Almost every sample path of E has this property because, as shown by the representation of E given in the proof of Theorem 11, each such sample path can be written in the form $u \circ F$, where u belongs to $C[0, 1]$. (Don't forget that F is continuous at m.) Call out the Continuous Mapping Theorem.

$$\begin{aligned} n^{1/2}(m_n - m) &= H(E_n, m_n, o_p(1), o_p(1)) \\ &\rightsquigarrow H(E, m, 0, 0) \\ &= -E(m)/f(m), \end{aligned}$$

which has the desired $N(0, \frac{1}{4}f(m)^{-2})$ distribution. □

15 Example. What happens to the limit distribution of Kolmogorov's goodness-of-fit statistic when parameters are estimated? That is, if F_n is obtained by independent sampling on the distribution function $F(\cdot, \theta_0)$, with θ_0 the true value of an unknown parameter θ, what is the asymptotic behavior of the statistic

$$D_n = n^{1/2} \sup_t |F_n(t) - F(t, \theta_n)|$$

constructed using an estimator θ_n for θ?

Suppose F were uniformly differentiable, in the strong sense (Problem 9) that

$$\|F(\cdot, \theta) - F(\cdot, \theta_0) - (\theta - \theta_0)\Delta(\cdot)\| = o(\theta - \theta_0), \quad \text{near } \theta_0,$$

for some fixed function $\Delta(\cdot)$ in $D[-\infty, \infty]$. Then, provided θ_n were one of those nice estimators that converge in towards θ_0 at the $O_p(n^{-1/2})$ rate, D_n could be written as

$$\begin{aligned} D_n &= n^{1/2}\|F_n(\cdot) - F(\cdot, \theta_0) - (\theta_n - \theta_0)\Delta(\cdot)\| + o_p(n^{1/2}(\theta_n - \theta_0)) \\ &= \|E_n - n^{1/2}(\theta_n - \theta_0)\Delta\| + o_p(1). \end{aligned}$$

We know how E_n behaves for large n—like the gaussian process E—but what will be the effect of adding on the extra term $n^{1/2}(\theta_n - \theta_0)\Delta$? That depends upon the joint distribution of E_n and θ_n.

According to the statistical folklore, good estimators can often be coerced into the form

$$\theta_n = \theta_0 + n^{-1} \sum_{i=1}^{n} L(\eta_i) + o_p(n^{-1/2})$$

for some function L satisfying $\mathbb{P}L(\eta_i) = 0$ and $\mathbb{P}L(\eta_i)^2 = \sigma^2 < \infty$. Let's assume that our estimator has such a representation. That makes it much easier to analyze the random elements $Z_n = (E_n, n^{1/2}(\theta_n - \theta_0))$ of $D[-\infty, \infty] \otimes \mathbb{R}$. Look at the fidis. Write $(E_n(t_1), \ldots, E_n(t_k), n^{1/2}(\theta_n - \theta_0))$ as

$$n^{-1/2} \sum_{i=1}^{n} (\{\eta_i \le t_1\} - F(t_1, \theta_0), \ldots, \{\eta_i \le t_k\} - F(t_k, \theta_0), L(\eta_i)) + o_p(1).$$

The Multivariate Central Limit Theorem gives convergence in distribution to a zero mean multivariate normal random vector $(E(t_1), \ldots, E(t_k), \gamma)$, where the covariances amongst the first k components follow (10) and $\mathbb{P}E(t)\gamma = \mathbb{P}\{\eta_i \leq t\}L(\eta_i)$. That suggests a limit process (E, γ) for Z_n, with E and γ having a sort of gaussian distribution in $D[-\infty, \infty] \otimes \mathbb{R}$.

We actually already have all the tools needed to formalize the limit result for Z_n. All that maximal inequality and finite-dimensional approximation stuff goes through as for the Empirical Central Limit Theorems; most of the brain work has been carried out, and only very messy details remain. But that would be terribly inelegant. Why not wait for the neater proof in Section VII.5, and just accept the fidi proof as sufficient evidence in the meantime? Elegance will return when the functions $\{\eta_i \leq t\}$ and $L(\eta_i)$ are accorded equal status.

Check continuity and measurability for the map from $D[-\infty, \infty] \otimes \mathbb{R}$ into $\mathbb{R}$ that sends (x, r) onto $\|x(\cdot) - r\Delta(\cdot)\|$, then bring out the Continuous Mapping Theorem, yet again, to deduce that $D_n \rightsquigarrow D = \|E(\cdot) - \gamma\Delta(\cdot)\|$. The parameter estimation propagates through to the limit to add a random drift term $\gamma\Delta(\cdot)$ onto the gaussian process $E(\cdot)$. If only we could calculate the distribution of D then we would have solved the problem completely. Problem 10 looks at a special case to show what difficulties this can present.

Try reworking the example with all the processes in $D[-\infty, \infty]$ rescaled using $F(\cdot, \theta_0)$ if you want to discover some of the perils of too automatic a recourse to the quantile transformation. □

V.3. Existence of Brownian Bridge and Brownian Motion

This section proves the existence of gaussian processes satisfying the requirements of Definition 8. The proof uses the Compactness Theorem from Section IV.5. This is neither the simplest nor the fastest method known to mankind; but it will generalize readily into an existence proof for more exotic gaussian processes. Any reader who found Section IV.5 too tedious could safely skip onto the next section, with perhaps just a quick look at Problem 11.

A good way to construct a process is to set up a sequence of approximations that should converge to the process, if it exists. One then needs to show that the approximating sequence, or one of its subsequences, actually does converge to something. One hopes this something has all the required properties of the desired process.

We have not yet considered any sequence of processes that should converge to a brownian motion, but in the uniform empirical processes $\{U_n\}$ we do have a sequence converging to a close relative, the brownian bridge. Let us extract a brownian bridge as the limit of a convergent subsequence.

16 Theorem. *There exists a brownian bridge.*

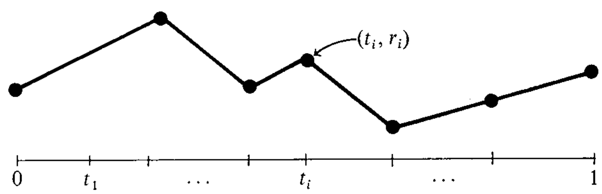
Proof. Check the uniform tightness condition in the Compactness Theorem of Section IV.5. We need to find a compact subset K of $D[0, 1]$ for which

$$\liminf \mathbb{P}\{U_n \in G\} > 1 - \varepsilon, \tag{17}$$

whenever G is an open, projection-measurable set containing K. To force the brownian bridge to live in $C[0, 1]$, we shall also want K to contain only continuous functions.

Our K will be represented as the intersection of a sequence of closed sets $\{D_k\}$, with D_k a finite union of closed balls of radius $2/k$ (any sequence of radii decreasing to zero would suffice). As the digression after the start of Section IV.5 pointed out, such a K is compact: it inherits completeness from $D[0, 1]$ and, by the choice of $\{D_k\}$, it is certainly totally bounded.

The closed balls making up each D_k center themselves on piecewise linear, continuous functions obtained by linear interpolation between a finite set of vertices $(0, r_0), \ldots, (t_i, r_i), \ldots, (1, r_m)$:



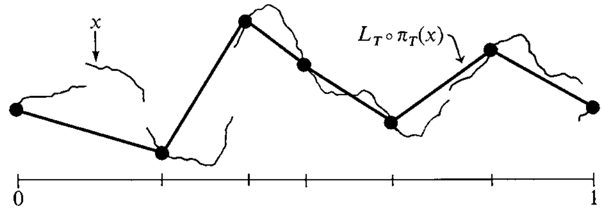
With T standing for the grid $0 = t_0 < \cdots < t_m = 1$, write L_T for the map from $\mathbb{R}^T$ to $C[0, 1]$ that takes the vector $\mathbf{r} = (r_0, \ldots, r_m)$ onto this interpolated function. The map satisfies a Lipschitz condition,

$$\|L_T(\mathbf{r}) - L_T(\mathbf{s})\| \leq \max_i |r_i - s_i|. \tag{18}$$

The composition $L_T \circ \pi_T$ defines a map from $D[0, 1]$ into $C[0, 1]$, a piecewise linear, continuous approximation map for which

$$\|L_T \circ \pi_T(x) - x\| \leq 2 \max_i \sup_{J_i} |x(t) - x(t_i)|,$$

where $J_i = [t_i, t_{i+1})$. The right-hand side should look familiar. It is the functional of x, call it $h_T(x)$, that lurks behind the convergence criterion of Theorem 3.



For given $\delta > 0$ and $\varepsilon > 0$, the proof of Theorem 9 came up with a grid T for which

$$\limsup \mathbb{P}\{h_T(U_n) > \delta\} < \varepsilon.$$

You should convince yourself that it was unnecessary to assume the existence of the brownian bridge process to get this inequality. Reinterpret it as a statement of how well U_n can be approximated by linear interpolation: given ε and δ there exists a grid T for which

$$\limsup \mathbb{P}\{\|L_T \circ \pi_T(U_n) - U_n\| > \delta\} < \varepsilon.$$

Why are there no measurability difficulties here?

Now come back to the task of specifying the closed balls for the D_k sets. For each k choose a grid $T(k)$ such that

$$\liminf \mathbb{P}\{\|L_{T(k)} \circ \pi_{T(k)}(U_n) - U_n\| \leq 1/k\} > 1 - \varepsilon/2^{k+1}.$$

Because the random vectors $\{\pi_{T(k)}(U_n)\}$ converge in distribution, there exists a compact subset H_k of $\mathbb{R}^{T(k)}$ such that

$$\liminf \mathbb{P}\{\pi_{T(k)} U_n \in H_k\} > 1 - \varepsilon/2^{k+1}$$

hence

$$\liminf \mathbb{P}\{L_{T(k)} \circ \pi_{T(k)}(U_n) \in L_{T(k)}(H_k)\} > 1 - \varepsilon/2^{k+1}.$$

As a continuous image of a compact set, $L_{T(k)}(H_k)$ is compact, and so can be covered by a finite collection of closed balls with radius $1/k$. Set D_k equal to the union of the finite collection of closed balls with the same centers, but radius $2/k$.

The larger radius allows for a $1/k$ distance between U_n and its linearly interpolated approximation $L_{T(k)} \circ \pi_{T(k)}(U_n)$:

$$\liminf \mathbb{P}\{U_n \in D_k\} > 1 - \varepsilon/2^k \quad \text{for every } k.$$

The $\varepsilon/2^k$ was contrived so that

$$\liminf \mathbb{P}\{U_n \in D_1 \cap \cdots \cap D_k\} > 1 - \varepsilon \quad \text{for every } k.$$

Of course we cannot just let k tend to infinity now and hope to replace the finite intersections by their limit K. We really do need that open set G in the definition of uniform tightness.

Inequality (17) will follow from what we have proved about the $\{D_k\}$ if we show that $G \supseteq D_1 \cap \cdots \cap D_k$ for some k. If there were no such k, we could choose an x_k from each closed set $F_k = G^c \cap D_1 \cap \cdots \cap D_k$. As shown in the next paragraph, we could then apply Cantor's diagonalization argument to extract a Cauchy subsequence from $\{x_k\}$. The limit of the Cauchy subsequence (remember that $D[0, 1]$ is complete) would belong to F_k for every k, an impossibility because

$$\bigcap_k F_k = G^c \cap K = \varnothing.$$

The existence of the desired k would follow.

Here is how we would construct the Cauchy subsequence. D_1 is a finite union of closed balls of radius 2; some subsequence of $\{x_k\}$ would lie completely within one of these, B_1 say. $B_1 \cap D_2$ would be contained in a finite union of closed balls of radius 1; some sub-subsequence would lie completely within $B_1 \cap B_2$, with B_2 a closed ball of radius 1. $B_1 \cap B_2 \cap D_3$ would be contained in a finite union of closed balls of radius $\frac{2}{3}$; a sub-sub-subsequence would lie within $B_1 \cap B_2 \cap B_3$, with B_3 a closed ball of radius $\frac{2}{3}$. And so on. The desired Cauchy sequence would take the first element of the first subsequence, the second element of the sub-subsequence, the third element of the sub-sub-subsequence, and so on.

Every function in K is continuous, a uniform limit of piecewise linear, continuous functions from the sets $L_{T(k)}(H_k)$.

By the Compactness Theorem, some subsequence of the uniformly tight $\{U_n\}$ converges in distribution to a probability measure concentrating on a countable union of compact subsets of $C[0, 1]$. Its fidis, the limits of the corresponding empirical process fidis, identify it as the distribution of the sought-after brownian bridge. □

There are two ways of constructing brownian motion from the brownian bridge U. One can either set

$$B_1(t) = U(t) + tZ \quad \text{for} \quad 0 \le t \le 1,$$

where Z has a $N(0, 1)$ distribution independent of U, or one can rescale by setting

$$B_2(t) = (1 + t)U\left(\frac{t}{1+t}\right) \quad \text{for} \quad 0 \le t < \infty.$$

In both cases one gets a gaussian process with continuous sample paths. Direct calculation of means and covariances, which uniquely determine multivariate normal distributions, shows that B_1 has the right fidis to identify it as a brownian motion on $[0, 1]$ and B_2 has the right fidis for brownian motion on $[0, \infty)$.

V.4. Processes with Independent Increments

A stochastic process Z indexed by an interval of the real line is said to have independent increments if $Z(t_0)$, $Z(t_1) - Z(t_0), \ldots, Z(t_k) - Z(t_{k-1})$ are mutually independent whenever $t_0 < t_1 < \cdots < t_k$. For random elements $\{X_n\}$ of $D[0, 1]$ with independent increments satisfying a mild regularity condition, the criterion for convergence in distribution given in Theorem 3 reduces to a particularly simple form when the limit process has continuous sample paths. Essentially we have only to check convergence for each increment $X_n(t) - X_n(s)$. Just one thing can spoil the nice clean characterization. Because the increments would be unchanged by the addition of an arbitrary constant to X_n, they surely cannot determine uniquely its finite-dimensional distributions; $X_n(0)$ must be specified.

19 Theorem. *Let $X, X_1, X_2, \ldots$ be random elements of $D[0, 1]$, each with independent increments. Suppose X has continuous sample paths. Then $X_n \rightsquigarrow X$ if and only if:*

(i) $X_n(0) \rightsquigarrow X(0)$;
(ii) *for each pair $s < t$, the increment $X_n(t) - X_n(s)$ converges in distribution to $X(t) - X(s)$;*
(iii) *given $\delta > 0$ there exist $\alpha > 0$ and $\beta > 0$ and an integer n_0 such that $\mathbb{P}\{|X_n(t) - X_n(s)| \le \frac{1}{2}\delta\} \ge \beta$ whenever $|t - s| < \alpha$ and $n \ge n_0$.*

PROOF. Forgive the mysterious factor of $\frac{1}{2}$ in (iii); it's there to make the notation fit better with Lemma 7.

Suppose $X_n \rightsquigarrow X$. Conditions (i) and (ii) follow from convergence of the fidis. Condition (iii) is a simple consequence of the Continuous Mapping Theorem applied to the continuous map from $D[0, 1]$ into $\mathbb{R}$ defined by

$$H_\alpha x = \sup\{|x(t) - x(s)|: |t - s| < \alpha\}.$$

(Take the supremum over rational pairs to check measurability.) At each continuous x the supremum converges to zero as $\alpha \to 0$, because continuity on a compact interval implies uniform continuity. As this applies to every sample path of X, there must exist an $\alpha > 0$ and $\beta > 0$ such that $\mathbb{P}\{H_\alpha X < \frac{1}{2}\delta\} = 2\beta$. Because $H_\alpha X_n \rightsquigarrow H_\alpha X$ and because $(-\infty, \frac{1}{2}\delta)$ is open,

$$\liminf_n \mathbb{P}\{H_\alpha X_n < \tfrac{1}{2}\delta\} \ge 2\beta$$

by authority of Example IV.17. For all n large enough,

$$\mathbb{P}\{|X_n(t) - X_n(s)| \le \tfrac{1}{2}\delta\} \ge \mathbb{P}\{H_\alpha X_n < \tfrac{1}{2}\delta\} > \beta,$$

whenever $|t - s| < \alpha$.

Now let us show that the two requirements listed in Theorem 3 follow from conditions (i), (ii), and (iii). Choose $0 = s_0 < s_1 < \cdots < s_k$. The joint characteristic function of the random vector

$$(X_n(s_0), X_n(s_1) - X_n(s_0), \ldots, X_n(s_k) - X_n(s_{k-1}))$$

factorizes into a product of one-dimensional characteristic functions, because X_n has independent increments. By (i) and (ii), the product converges to the corresponding product for X, the joint characteristic function of the random vector

$$(X(s_0), X(s_1) - X(s_0), \ldots, X(s_k) - X(s_{k-1})).$$

With a continuous (linear, even) transformation, recover the desired fidi convergence.

For the second requirement of Theorem 3 try a grid $0 = t_0 < \cdots < t_m = 1$ for which $\max_i(t_{i+1} - t_i) < \gamma < \alpha$, with α as given by (iii). The value of γ

will be specified at the end of the proof. Fix for the moment a value of n greater than n_0. Write $\mathscr{E}_t$ for the σ-field generated by $\{X_n(s): s \leq t\}$. Suppose $t_i \leq t < t_{i+1}$. Then, on the set $\{|X_n(t) - X_n(t_i)| > \delta\}$,

$$\begin{aligned}\mathbb{P}\{|X_n(t_{i+1}) - X_n(t)| &\leq \tfrac{1}{2}|X_n(t) - X_n(t_i)|\,|\mathscr{E}_t\} \\ &\geq \mathbb{P}\{|X_n(t_{i+1}) - X_n(t)| \leq \tfrac{1}{2}\delta\} \\ &\geq \beta.\end{aligned}$$

Apply Lemma 7 to the process $X_n(\cdot) - X_n(t_i)$ on each interval $J_i = [t_i, t_{i+1})$.

$$\begin{aligned}&\limsup \mathbb{P}\left\{\max_i \sup_{J_i} |X_n(t) - X_n(t_i)| > \delta\right\} \\ &\quad\leq \limsup \sum_{i=0}^{m-1} \mathbb{P}\left\{\sup_{J_i} |X_n(t) - X_n(t_i)| > \delta\right\} \\ &\quad\leq \beta^{-1} \sum_{i=0}^{m-1} \limsup \mathbb{P}\{|X_n(t_{i+1}) - X_n(t_i)| > \tfrac{1}{2}\delta\}\end{aligned}$$

which is less than

$$\beta^{-1} \sum_{i=0}^{m-1} \mathbb{P}\{|X(t_{i+1}) - X(t_i)| \geq \tfrac{1}{2}\delta\}$$

because $(X_n(t_{i+1}), X_n(t_i)) \rightsquigarrow (X(t_{i+1}), X(t_i))$ and $\{(u, v) \in \mathbb{R}^2: |u - v| \geq \frac{1}{2}\delta\}$ is closed. Write $E_0, \ldots, E_{m-1}$ for the independent events appearing in the last summation. From the inequality $\exp(-\mathbb{P}E_i) \geq 1 - \mathbb{P}E_i$ deduce that

$$\begin{aligned}\sum_{i=0}^{m-1} \mathbb{P}E_i &\leq -\log \mathbb{P} \bigcap_i E_i^c \\ &= -\log\left(1 - \mathbb{P} \bigcup_i E_i\right) \\ &\leq -\log(1 - \mathbb{P}\{H_\gamma X \geq \tfrac{1}{2}\delta\}) \quad \text{because } \max_i (t_{i+1} - t_i) < \gamma.\end{aligned}$$

Argue as before that $H_\gamma X \downarrow 0$ for each sample path, as $\gamma \downarrow 0$, to see that γ could have been chosen to make the logarithmic expression less than $\beta\varepsilon$. □

Notice how we used the sample path continuity of the limit process X. The argument would break down if X were allowed to grow by jumps. For example, if X were a poisson process, with $X(t_i) - X(t_{i-1})$ distributed Poisson$(\lambda(t_i - t_{i-1}))$, the sum would not decrease as γ tended to zero:

$$\begin{aligned}\sum_{i=0}^{m-1} \mathbb{P}\{|X(t_{i+1}) - X(t_i)| \geq \tfrac{1}{2}\delta\} &= \sum_{i=0}^{m-1} [\lambda(t_{i+1} - t_i) + o(t_{i+1} - t_i)] \\ &= \lambda + o(1).\end{aligned}$$

The uniform metric is inappropriate for such an X. A better metric, which can tolerate a jumpy limit process, will be introduced in Chapter VI.

Theorem 19 really only says something about convergence in distribution to gaussian limit processes. If $X(0)$ has a normal distribution, the combination of independent increments and continuous sample paths forces X to be a rescaled and recentered brownian motion (the gaussian process described in Definition 8). An approximation argument shows why. Because $H_\gamma(X) \to 0$ almost surely as $\gamma \to 0$, there exists a sequence $\{\varepsilon_n\}$ with $\varepsilon_n \downarrow 0$ and $\mathbb{P}\{H_{2/n}(X) > \varepsilon_n\} \to 0$. Set $\xi_{ni} = X_n(i/n) - X_n((i-1)/n)$. Define step-function approximations to X by

$$X_n(t) = X(0) + \sum_{i=1}^{n} \{i/n \le t\}\xi_{ni}\{|\xi_{ni}| \le \varepsilon_n\}.$$

Whenever $H_{2/n}(X) \le \varepsilon_n$, the X_n process lies uniformly within ε_n of X. Thus $\|X_n - X\| \to 0$ in probability.

Fix t. Write σ_n^2 for the variance of the sum appearing in the definition of $X_n(t)$, and a_n for its expectation. If $\sigma_n \to 0$, then $X_n(t) = X(0) + a_n + o_p(1)$ from which we get $X(t) - X(0) =$ constant, a degenerate sort of normality. If, on the other hand, $\{\sigma_n\}$ does not tend to zero, then, along some subsequence, $(X_n(t) - X(0) - a_n)/\sigma_n \rightsquigarrow N(0, 1)$ by the Lindeberg Central Limit Theorem. (Necessarily $\varepsilon_n/\sigma_n \to 0$ along any subsequence for which σ_n is bounded away from zero.) Convergence of types (Breiman 1968, Section 8.8) allows nothing but normality for the distribution of $X(t) - X(0)$. A similar argument works for any other increment of X.

Even with the gloss of apparent generality rubbed off the limit distribution, Theorem 19 still has enough content to handle some non-classical problems about sums of independent random variables.

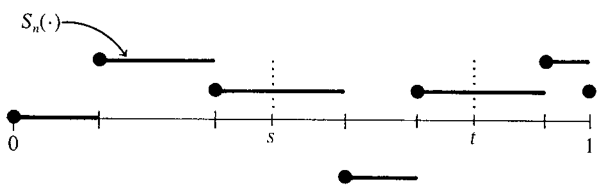
20 Example. Let $\{\xi_{ni}: i = 1, \ldots, k(n); n = 1, 2, \ldots\}$ be a triangular array of random variables satisfying the conditions of the Lindeberg Central Limit Theorem. That is, they are independent within each row, they have zero means, they have variances $\{\sigma_{ni}^2\}$ that sum to one within each row, and

$$\sum_i \mathbb{P}\xi_{ni}^2\{|\xi_{ni}| \ge \varepsilon\} \to 0 \tag{21}$$

for each fixed $\varepsilon > 0$. Set $S_{ni} = \xi_{n1} + \cdots + \xi_{ni}$. What is the asymptotic behavior of a random variable such as $\max_i S_{ni}$? We may write it as a functional of the partial-sum process $S_n(\cdot)$, the random element of $D[0, 1]$ defined by

$$S_n(t) = S_{ni} \quad \text{for} \quad \operatorname{var} S_{ni} \le t < \operatorname{var} S_{n,i+1}.$$

The curious choice for the location of the jumps of $S_n(\cdot)$ has the virtue that $\operatorname{var} S_n(t) \to t$ as $n \to \infty$, for each fixed t, because $\max_i \sigma_{ni}^2 \to 0$. The increments of $S_n(\cdot)$ over disjoint subintervals of $[0, 1]$ are sums over disjoint groups of the $\{\xi_{ni}\}$; the process has independent increments. If we joined up the vertices, to produce an interpolated partial-sum process with sample paths in $C[0, 1]$, we would destroy this property.



For fixed s and t, with s less than t, the increment $S_n(t) - S_n(s)$ is also a sum of the elements in a triangular array of independent random variables. Because

$$\text{var}[S_n(t) - S_n(s)] \to t - s$$

and because (21) also holds for summation over any subset of the $\{\xi_{ni}\}$, the Lindeberg Central Limit Theorem makes $S_n(t) - S_n(s)$ converge to a $N(0, t - s)$ distribution—one of the conditions we need to make $S_n(\cdot)$ converge in distribution to a brownian motion.

What about condition (iii) of Theorem 19? Tchebychev's inequality is good enough.

$$\begin{aligned} \mathbb{P}\{|S_n(t) - S_n(s)| \le \tfrac{1}{2}\delta\} &\ge 1 - \text{var}[S_n(t) - S_n(s)]/(\tfrac{1}{2}\delta)^2 \\ &\to 1 - (t - s)/(\tfrac{1}{2}\delta)^2, \end{aligned}$$

uniformly in t and s. You can figure out what α should be from this.

Only our ingenuity in thinking up functionals that are continuous at the sample paths of brownian motion can curtail our supply of limit theorems for the partial sums. One example:

$$\max_i S_{ni} = \sup_t S_n(t) \rightsquigarrow \sup_t B(t).$$

Calculation of the limit distribution in closed form awaits us in Section 6, where this and several other functionals of brownian motion and brownian bridge will be examined. □

V.5. Infinite Time Scales

Both the spaces $D[0, 1]$ and $D[-\infty, \infty]$ have compact intervals of the extended real line as their index sets; the theories for convergence in distribution of random elements of these spaces differ only superficially. For spaces with non-compact index sets some extra complications arise. As a typical example, consider $D[0, \infty)$, the set of all real-valued cadlag functions on $[0, \infty)$.

A function x in $D[0, \infty)$ must be right-continuous at each point of $[0, \infty)$, with a finite left limit existing at each point of $(0, \infty)$. Because the limit point $+\infty$ does not belong to the index set, no constraint is placed on the behavior of $x(t)$ as $t \to \infty$; it could diverge or oscillate about in any fashion. Such a function need not be bounded. The uniform distance between two functions in $D[0, \infty)$ might be infinite.

Even for bounded random elements of $D[0, \infty)$, convergence in distribution in the sense of the uniform metric may impose far stronger requirements on their tail behavior (that is, on what happens as $t \to \infty$) than we can hope to verify. Sometimes the best we can try for is control over large compact subintervals of $[0, \infty)$. That corresponds to convergence in the sense of the metric for uniform convergence on compacta.

22 Definition. A sequence of functions $\{x_n\}$ in $D[0, \infty)$ converges uniformly on compacta to a function x if $\sup_{t \le k} |x_n(t) - x(t)| \to 0$ as $n \to \infty$ for each fixed k. Equivalently, $d(x_n, x) \to 0$, where

$$d(x_n, x) = \sum_{k=1}^{\infty} 2^{-k} \min[1, d_k(x_n, x)]$$

$$d_k(x_n, x) = \sup_{t \le k} |x_n(t) - x(t)|. \qquad \square$$

In all that follows, $D[0, \infty)$ will be equipped with the metric d and the projection σ-field. With that combination, each x in $D[0, \infty)$ is completely regular (Problem 12).

For each k define a truncation map L_k from $D[0, \infty)$ into $D[0, k]$; set $L_k x$ equal to the restriction of x to the interval $[0, k]$. By construction, $\{x_n\}$ converges to x if and only if $\{L_k x_n\}$ converges uniformly to $L_k x$ for each k. Convergence in distribution has a similar characterization.

23 Theorem. *Let $X, X_1, X_2, \ldots$ be random elements of $D[0, \infty)$, with $\mathbb{P}\{X \in C\}$ for some separable set C. Then $X_n \rightsquigarrow X$ if and only if $L_k X_n \rightsquigarrow L_k X$, as random elements of $D[0, k]$, for each fixed k.*

Proof. Necessity of the condition follows from the Continuous Mapping Theorem, because each L_k is both continuous and measurable.

For sufficiency, define a continuous, measurable map H_k from $D[0, k]$ into $D[0, \infty)$ by $(H_k z)(t) = z(t \wedge k)$. It comes as close as possible to defining an inverse to the map L_k: the function $H_k \circ L_k x$ equals x on $[0, k]$. Deduce that $d(x, H_k \circ L_k x) \le 2^{-k}$ for every x in $D[0, \infty)$.

Suppose $L_k X_n \rightsquigarrow L_k X$. Because H_k is continuous, $H_k \circ L_k X_n \rightsquigarrow H_k \circ L_k X$. Pick k large enough to make $2^{-k} < \varepsilon$. Then both $\mathbb{P}\{d(X_n, H_k \circ L_k X_n) > \varepsilon\}$ and $\mathbb{P}\{d(X, H_k \circ L_k X) > \varepsilon\}$ equal zero. Example IV.11, with $H_k \circ L_k$ playing the role of the approximation map, does the rest. $\square$

Typically C equals $C[0, \infty)$, the set of all continuous functions in $D[0, \infty)$. As in the case of compact time intervals, $C[0, \infty)$ sits inside $D[0, \infty)$ as a closed, separable, measurable subset (Problem 13). The borel σ-field on $C[0, \infty)$, generated by the open sets for the d metric, coincides with the projection σ-field. The most important of the limit processes living in $C[0, \infty)$ is brownian motion on $[0, \infty)$.

24 Example. Let S_n denote the nth partial sum of a sequence $\xi_1, \xi_2, \ldots$ of independent, identically distributed random variables for which $\mathbb{P}\xi_i = 0$ and $\mathbb{P}\xi_i^2 = 1$. What is the asymptotic behavior of the hitting time

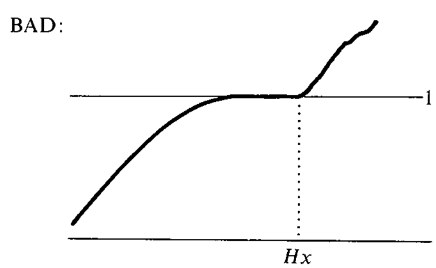
$$\tau_n = \inf\{j: n^{-1/2}S_j > 1\}$$

as n tends to infinity? Define $Hx = \inf\{t \geq 0: x(t) > 1\}$, with the usual convention that the empty set has infimum $+\infty$. Right continuity of functions in $D[0, \infty)$ allows us to take the infimum over rational t values to verify measurability of H. Express $n^{-1}\tau_n$ as the functional HX_n of the process

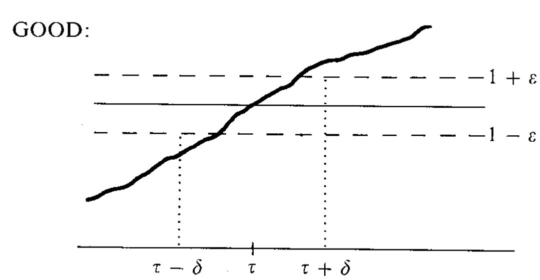
$$X_n(t) = n^{-1/2}S_j, \qquad j/n \leq t < (j+1)/n.$$

Use Theorem 23 to check that $X_n \rightsquigarrow B$, a brownian motion on $[0, \infty)$. For fixed k, the truncated process $L_k X_n$ has the same form as the partial-sum process of Example 20, but stretched out and rescaled to fit the interval $[0, k]$ instead of $[0, 1]$. The modifications have little effect on the arguments adduced there to prove convergence to brownian motion; exactly the same idea works in $D[0, k]$.

If we are to use the Continuous Mapping Theorem to prove $HX_n \rightsquigarrow HB$ we will need H continuous at almost all sample paths of B. By itself, continuity of a sample path x will not suffice for continuity of H at x. You can construct sequences of functions $\{x_n\}$ that converge uniformly to a bad continuous x without having Hx_n converging to Hx.



The functional H has better success with a good path like this:



Here $Hx = \tau$. If $x(t) \leq 1 - \varepsilon$ for $t \leq \tau - \delta$ (a continuous function achieves its maximum on a compact interval) and $x(\tau + \delta) \geq 1 + \varepsilon$, then any y in $D[0, \infty)$ with $d_{\tau+\delta}(x, y) < \varepsilon$ must satisfy $\tau - \delta \leq Hy \leq \tau + \delta$. That brownian motion has only good sample paths, almost surely, can best be shown by arguments based on a strong markov property, a topic to be taken up in the next section. The distribution of the functional HB of brownian motion will also be derived in that section. For the moment, we must content ourselves with knowing that $n^{-1}\tau_n$ has a limiting distribution that we could calculate if we were better acquainted with the limit process of the sequence $\{X_n\}$. □

V.6. Functionals of Brownian Motion and Brownian Bridge

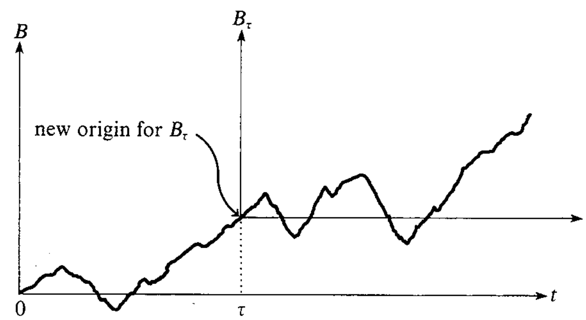
From the definition of brownian motion on $[0, \infty)$ it is easy to deduce (Problem 14), for each fixed τ, that the shifted process

$$B_\tau(t) = B(\tau + t) - B(\tau) \quad \text{for} \quad 0 \leq t < \infty$$

is a new brownian motion, independent of the σ-field $\mathscr{E}_\tau$ generated by the random variables $\{B(s): 0 \leq s \leq \tau\}$. Equivalently,

$$\mathbb{P}f(B_\tau)A = \mathbb{P}f(B)\mathbb{P}A \tag{25}$$

for every A in $\mathscr{E}_\tau$ and every bounded measurable f on $C[0, \infty)$.



The assertion is also valid for a wide class of random τ values. The precisely formulated generalization, known as the strong markov property, underlies most of the clever tricks one can perform with brownian motion. It will get us the distributions of a few interesting functionals.

The random variable τ will need to be a stopping time, that is, a random variable, taking values in $[0, \infty]$, for which $\{\tau \leq t\}$ belongs to $\mathscr{E}_t$ for each t.

Interpret $\mathscr{E}_\infty$ as the smallest σ-field containing every $\mathscr{E}_t$—we learn everything about the brownian motion if we watch it forever. Define

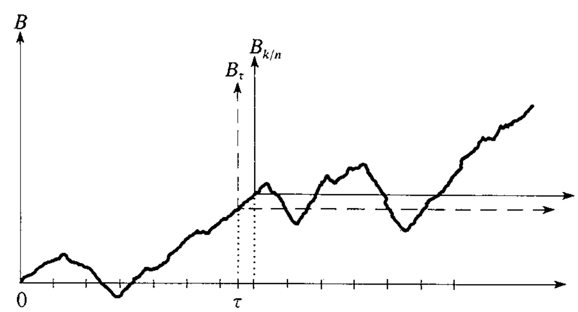
$$B_\tau(\omega, t) = [B(\omega, t + \tau(\omega)) - B(\omega, \tau(\omega))]\{\tau(\omega) < \infty\}.$$

Problem 17 shows that B_τ is projection-measurable, as a map from Ω into $C[0, \infty)$.

The stopping time τ determines a σ-field $\mathscr{E}_\tau$, which captures the intuitive idea of an event being observable before time τ. By definition, an event A belongs to $\mathscr{E}_\tau$ if and only if $A\{\tau \le t\}$ belongs to $\mathscr{E}_t$ for every t. The strong markov property asserts that B_τ is a brownian motion independent of $\mathscr{E}_\tau$ on $\{\tau < \infty\}$. Equivalently, (25) holds for every $\mathscr{E}_\tau$-measurable A contained in $\{\tau < \infty\}$ and every bounded, measurable f on $C[0, \infty)$. To prove the assertion it suffices that we check the equality for each bounded, uniformly continuous f. (Apply Problem 15 to the distributions induced on $C[0, \infty)$ by B, under $\mathbb{P}$, and B_τ, under $\mathbb{P}(\cdot \mid A)$.) With such an f, continuity of sample paths implies

$$f(B_\tau)A = \lim \sum_{k=1}^{\infty} f(B_{k/n})A\{(k-1)/n \le \tau < k/n\}.$$

Of course only one term in the sum will be non-zero for each fixed n.

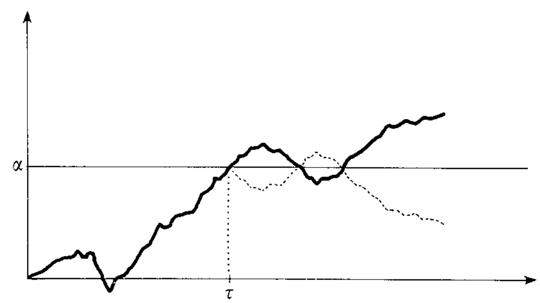


The event $A\{(k-1)/n \le \tau < k/n\}$ belongs to $\mathscr{E}_{k/n}$; the independence asserted in (25) can be invoked for each k/n value.

$$\begin{aligned}
\mathbb{P}f(B_\tau)A &= \lim \sum_{k=1}^{\infty} \mathbb{P}f(B_{k/n})A\{(k-1)/n \le \tau < k/n\} \\
&= \lim \sum_{k=1}^{\infty} \mathbb{P}f(B)\,\mathbb{P}A\{(k-1)/n \le \tau < k/n\} \quad \text{by (25)} \\
&= \mathbb{P}f(B)\mathbb{P}A \quad \text{because } \tau < \infty \text{ on } A.
\end{aligned}$$

The strong markov property is established.

26 Example. The classical reflection principle for brownian motion involves a hidden appeal to the strong markov property. It enables us to find the distribution of the stopping time $\tau = \inf\{t: B(t) = \alpha\}$ for fixed $\alpha > 0$.



The reflection principle uses the symmetry of brownian motion—the processes B_τ and $-B_\tau$ have the same distribution—to argue that B should be just as likely to hit α and end up with $B(t) > \alpha$ as it is to hit α and end up with $B(t) < \alpha$. The probability that B hits α before time t should be twice the probability that $B(t) > \alpha$. A formal proof works backwards.

$$\begin{aligned}\mathbb{P}\{B(t) > \alpha\} &= \mathbb{P}\{B(t) > \alpha, \tau < t\} \\ &= \mathbb{P}\{B(\tau) + B_\tau(t-\tau) > \alpha, \tau < t\} \\ &= \mathbb{P}\{B_\tau(t-\tau) > 0, \tau < t\} \quad \text{because } B(\tau) = \alpha \text{ if } \tau < \infty \\ &= \mathbb{P}[\{\tau < t\}\mathbb{P}\{B_\tau(t-\tau) > 0 \mid \mathscr{E}_\tau\}].\end{aligned}$$

Because τ is $\mathscr{E}_\tau$-measurable it can be treated as a constant inside the conditional probability. (A slightly more formal justification would invoke Fubini's theorem.) On $\{\tau < t\}$, symmetry of the normal distribution allows us to replace the conditional probability by $\frac{1}{2}$, and then deduce

$$\mathbb{P}\{B(t) > \alpha\} = \tfrac{1}{2}\mathbb{P}\{\tau < t\}.$$

An explicit expression for $\mathbb{P}\{\tau < t\}$ can be found by using the $N(0, t)$ distribution for $B(t)$.

For our later purposes it will be more important that we know the Laplace transform $L_\alpha(\lambda) = \mathbb{P}\exp(-\lambda\tau)$ for $\lambda \geq 0$. Two applications of Fubini's theorem, then a differentiation, will lead to a simple differential equation for L_α.

$$\begin{aligned}L_\alpha(\lambda) &= \mathbb{P}\int\{\tau < t < \infty\}\lambda e^{-\lambda t}\,dt \\ &= \int_0^\infty \lambda e^{-\lambda t}\mathbb{P}\{\tau < t\}\,dt \\ &= \int_0^\infty \lambda e^{-\lambda t} 2\mathbb{P}\{N(0, t) > \alpha\}\,dt\end{aligned}$$

$$= \mathbb{P} \int \lambda e^{-\lambda t}\{t > \alpha^2/N(0,1)^2\}\, dt$$
$$= \mathbb{P} \exp(-\lambda\alpha^2/N(0,1)^2)$$
$$= (2/\pi)^{1/2} \int_0^\infty \exp(-\lambda\alpha^2 z^{-2} - \tfrac{1}{2}z^2)\, dz.$$

Differentiate with respect to λ, then change variables by setting $y = (2\lambda)^{1/2}\alpha z^{-1}$.

$$L'_\alpha(\lambda) = (2/\pi)^{1/2} \int_0^\infty -\alpha^2 z^{-2} \exp(-\lambda\alpha^2 z^{-2} - \tfrac{1}{2}z^2)\, dz$$
$$= -\alpha(2\lambda)^{-1/2}(2/\pi)^{1/2} \int_0^\infty \exp(-\tfrac{1}{2}y^2 - \lambda\alpha^2 y^{-2})\, dy$$
$$= -\alpha(2\lambda)^{-1/2} L_\alpha(\lambda).$$

The only solution with $L_\alpha(0) = 1$ is $L_\alpha(\lambda) = \exp(-\alpha(2\lambda)^{1/2})$. □

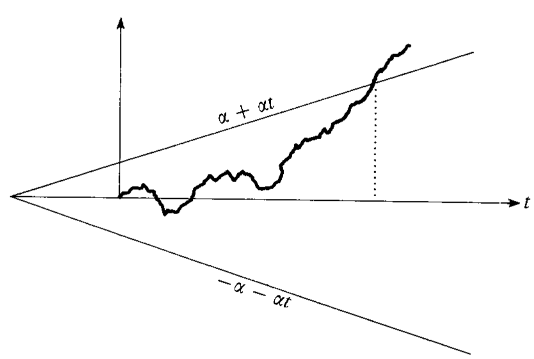
Doob (1949), in the non-heuristic part of his paper, was able to parlay the result from Example 26 into an expression for the distribution function of $\|U\|$, which his heuristic argument suggested as the limit for Kolmogorov's goodness-of-fit statistic. His approach provides a case study in the application of the strong markov property.

First one transforms the brownian bridge U by a scale change:

$$B(t) = (1+t)U\left(\frac{t}{1+t}\right) \quad \text{for} \quad 0 \le t < \infty.$$

The means and covariances of this gaussian random element of $C[0,\infty)$ identify it as a brownian motion (the same trick as in Section 3). Because $U(1) = 0$,

$$\mathbb{P}\{\|U\| \ge \alpha\} = \mathbb{P}\left\{\sup\left|U\left(\frac{t}{1+t}\right)\right| \ge \alpha : 0 \le t < \infty\right\}$$
$$= \mathbb{P}\{|B(t)| = \alpha(1+t) \text{ for at least one } t\}$$
$$= \mathbb{P}\{B \text{ hits } \alpha + \alpha t \text{ or } B \text{ hits } -\alpha - \alpha t\}.$$

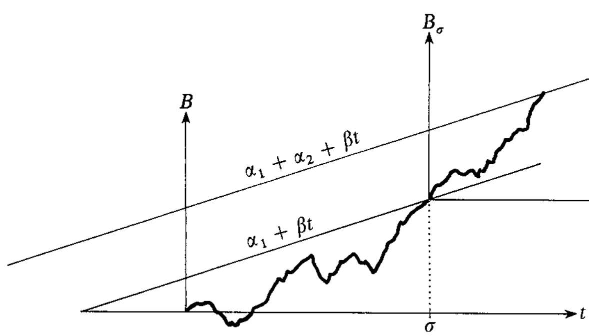


To find the probability that B ever leaves the region bounded by the two sloping lines $\pm(\alpha + \alpha t)$, one first solves the simpler problem where there is only one sloping line. For a fixed $\beta > 0$ (and not just $\beta = \alpha$) set

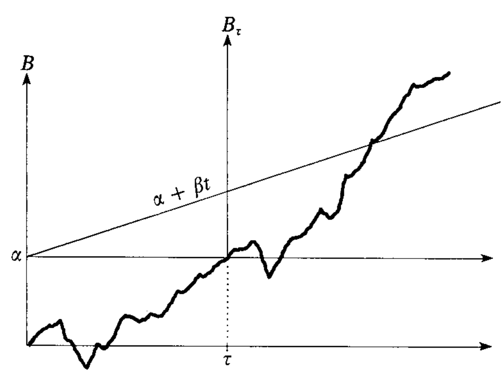
$$\phi(\alpha) = \mathbb{P}\{B \text{ hits } \alpha + \beta t\},$$

the dependence on β being suppressed for the while. If B is ever to hit the line $\alpha_1 + \alpha_2 + \beta t$, for $\alpha_1 > 0$ and $\alpha_2 > 0$, it must first hit the line $\alpha_1 + \beta t$. Call the stopping time at which this happens σ. Invoke the strong markov property.

$$\begin{aligned}\phi(\alpha_1 + \alpha_2) &= \mathbb{P}\{B \text{ hits } \alpha_1 + \alpha_2 + \beta t\}\\ &= \mathbb{P}\{\sigma < \infty, B_\sigma \text{ hits } \alpha_2 + \beta t\}\\ &= \mathbb{P}\{\sigma < \infty\}\mathbb{P}\{B \text{ hits } \alpha_2 + \beta t\}\\ &= \phi(\alpha_1)\phi(\alpha_2).\end{aligned}$$



The equation $\phi(\alpha_1 + \alpha_2) = \phi(\alpha_1)\phi(\alpha_2)$ has non-negative, decreasing solutions with $\phi(0) = 1$ only of the form $\phi(\alpha) = \exp(-c\alpha)$ for $\alpha > 0$. The positive constant c remains to be determined.



Another strong-markov argument, working with the stopping time τ at which B first hits level α, gives the value of the constant.

$$\begin{aligned}\exp(-c\alpha) &= \mathbb{P}\{B \text{ hits } \alpha + \beta t\}\\ &= \mathbb{P}\{\tau < \infty, B_\tau \text{ hits } \beta\tau + \beta t\}\\ &= \mathbb{P}[\{\tau < \infty\}\mathbb{P}\{B_\tau \text{ hits } \beta\tau + \beta t \mid \mathscr{E}_\tau\}]\\ &= \mathbb{P}\{\tau < \infty\}\phi(\beta\tau)\\ &= \mathbb{P}\exp(-c\beta\tau)\\ &= L_\alpha(c\beta) \quad \text{in the notation of Example 26}\\ &= \exp(-\alpha(2c\beta)^{1/2}).\end{aligned}$$

Solve for c.

(27) $$\mathbb{P}\{B \text{ hits } \alpha + \beta t\} = \exp(-2\alpha\beta) \quad \text{for} \quad \alpha > 0, \beta > 0.$$

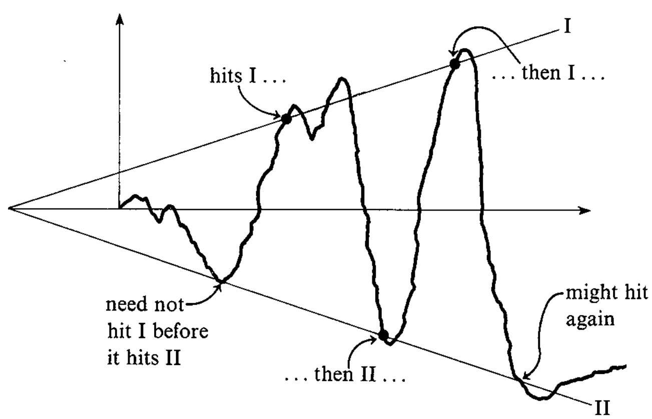
Notice that the hitting probability depends on α and β only through their product. The probability of B hitting the line $k\alpha + k^{-1}\beta t$ is the same for each $k > 0$.

The two barrier problem is slightly harder. For brevity write, temporarily, I for the line $\alpha + \beta t$ and II for the line $-\alpha - \beta t$. Then

$$\mathbb{P}\{|B(t)| = \alpha + \beta t \text{ for at least one } t\} = \mathbb{P}\{B \text{ hits I or } B \text{ hits II}\}.$$

The two events $\{B \text{ hits I}\}$ and $\{B \text{ hits II}\}$ are not disjoint. An inclusion–exclusion argument is needed. The infinite sum

$$\begin{aligned}&\{\text{hit I}\} - \{\text{hit I then II}\} + \{\text{hit I then II then I}\} - \cdots\\ &\quad + \{\text{hit II}\} - \{\text{hit II then I}\} + \{\text{hit II then I then II}\} - \cdots\end{aligned}$$



takes the value 0 if B never hits a barrier, it takes the value 1 if B makes a finite (positive) number of alternating hits, and it is undefined if B makes infinitely many alternating hits. If we prove that

(28) $$\mathbb{P}\{B \text{ makes infinitely many alternating hits}\} = 0$$

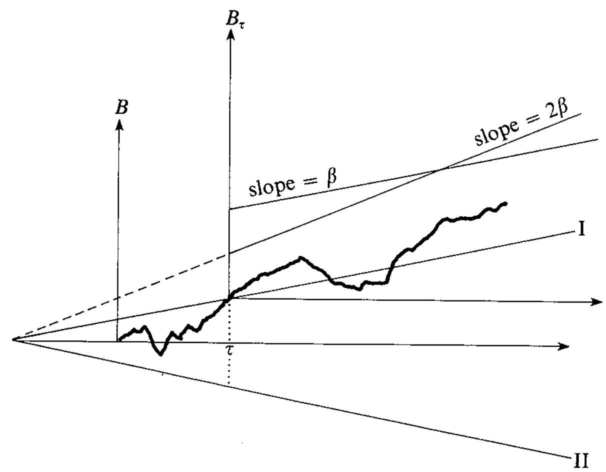
then dominated convergence will justify

(29) $$\begin{aligned}&\mathbb{P}\{B \text{ hits I or } B \text{ hits II}\}\\ &= \mathbb{P}\{\text{hit I}\} - \mathbb{P}\{\text{hit I then II}\} + \mathbb{P}\{\text{hit I then II then I}\} - \cdots\\ &\quad + \mathbb{P}\{\text{hit II}\} - \mathbb{P}\{\text{hit II then I}\} + \mathbb{P}\{\text{hit II then I then II}\} - \cdots.\end{aligned}$$

Our success with the calculation for one barrier will extend, in strongly markov fashion, to each of the individual terms in this sum of probabilities.

To avoid notational indigestion let us chew on just one of the terms, $\mathbb{P}\{\text{hit I then II}\}$, say. Define the stopping time τ (different from the last τ) as the first t at which B hits I. From the perspective of B_τ, barrier II is defined by the line of slope $-\beta$ and intercept $-2(\alpha + \beta\tau)$. (We can treat τ as if it were a constant if we condition on $\mathscr{E}_\tau$.) By symmetry, B_τ has the same probability of hitting the line $3(\alpha + \beta\tau) + \beta(t - \tau)$, which it sees as having slope β and intercept $2(\alpha + \beta\tau)$. Formula (27) allows B_τ to double the slope and halve the intercept of the line it is trying to hit. That is,

$$\begin{aligned}\mathbb{P}\{B_\tau \text{ hits II} \mid \mathscr{E}_\tau\} &= \mathbb{P}\{B_\tau \text{ hits } 3(\alpha + \beta\tau) + \beta(t - \tau) \mid \mathscr{E}_\tau\} \\ &= \mathbb{P}\{B_\tau \text{ hits } 2(\alpha + \beta\tau) + 2\beta(t - \tau) \mid \mathscr{E}_\tau\} \\ &= \mathbb{P}\{B_\tau \text{ hits } 2\alpha + 2\beta t \mid \mathscr{E}_\tau\}.\end{aligned}$$



Notice that the new line sits above $\alpha + \beta t$ for $t \geq 0$. Integrate out over the event $\{\tau < \infty\}$.

$$\begin{aligned}\mathbb{P}\{B \text{ hits I then II}\} &= \mathbb{P}\{\tau < \infty, B_\tau \text{ hits II}\} \\ &= \mathbb{P}\{\tau < \infty, B_\tau \text{ hits } 2\alpha + 2\beta t\} \\ &= \mathbb{P}\{B \text{ hits } 2\alpha + 2\beta t\} \\ &= \exp[-2(2\alpha)(2\beta)] \quad \text{from (27).}\end{aligned}$$

Similar reasoning succeeds with higher numbers of alternating hits. For example, if σ denotes the time of the first hit on II after a hit on I, then

$$\begin{aligned}\mathbb{P}\{B \text{ hits I then II then I}\} &= \mathbb{P}\{\tau < \infty, \sigma < \infty, B_\sigma \text{ hits I}\} \\ &= \mathbb{P}\{\tau < \infty, \sigma < \infty, B_\sigma \text{ hits } -2\alpha - 2\beta t\} \\ &= \mathbb{P}\{\tau < \infty, B_\tau \text{ hits } -2\alpha - 2\beta t\} \\ &= \mathbb{P}\{\tau < \infty, B_\tau \text{ hits } 3\alpha + 3\beta t\} \\ &= \mathbb{P}\{B \text{ hits } 3\alpha + 3\beta t\} \\ &= \exp[-2(3\alpha)(3\beta)].\end{aligned}$$

You will observe the significance of the point $-\alpha/\beta$ on the t axis if you draw a picture for the argument.

In general, for a string of m alternating hits on barriers I and II, the probability is $\exp(-2m^2\alpha\beta)$. Let m tend to infinity to prove (28). Formula (29) with $\alpha = \beta$ provides the solution to the problem with which we began:

$$\mathbb{P}\{\|U_n\| \geq \alpha\} \to \mathbb{P}\{\|U\| \geq \alpha\} = 2 \sum_{m=1}^{\infty} (-1)^{m+1} \exp(-2m^2\alpha^2).$$

Read Doob's paper.

Notes

Most authors avoid the uniform metric once they recognize the measurability problems it creates with the empirical process. Billingsley (1968, Section 18) made the case against it clear. Dudley (1966a, 1967a) had already pointed the way around the problem, but his solution has been mostly overlooked in the literature. Dudley (1966a) had even proved a multivariate version of the Empirical Central Limit Theorem. He used a markov property of the empirical process.

The idea of interpreting conditions like (4) as a means for constructing simple approximations to stochastic processes appears explicitly in Wichura (1971). Hájek (1965) had already applied the idea to characterize weak convergence in $C[0, 1]$. Skorohod (1956) based his study of weak convergence in $D[0, 1]$ under its various Skorohod metrics on the same idea. Unfortunately this simple approach seems to have lost out to the uniform tightness approach (exposited by Billingsley (1968), for example), possibly because the approximation method appears to demand a separate proof for existence of the limit process as a random element of a space such as $C[0, 1]$. Actually, (4) plus fidi convergence imply uniform tightness; the argument in Section 3 is easily generalized. Lemma 7, specialized to processes with independent increments and with $\frac{1}{2}\delta$ replacing $\frac{1}{2}|Z(t)|$ in the hypothesis, is sometimes attributed to Skorohod (1957), although it clearly has a strong similarity to Lévy's symmetrization inequality. Notice that the $\frac{1}{2}|Z(t)|$ does gain us something—one benefit was noted during the proof of the Empirical Central Limit Theorem, in Section 2. Gihman and Skorohod (1974, VI.5) made more systematic use of a form of the lemma in a study of weak convergence for markov processes. Stute (1982a) obtained delicate oscillation results for empirical processes by exploiting their markovian properties.

The Empirical Central Limit Theorem usually goes by the name of Donsker's theorem; Donsker (1952) proved it (using the uniform metric) in justifying Doob's (1949) heuristic approach to the Kolmogorov and Smirnov theorems. Donsker (1952) used a poissonization trick in his proof; he got it from Chung (1949), who got it from Kolmogorov (1933). Kac (1949) knew that a poissonized process was easier to analyze, because of its independent increments. Breiman (1968, Section 13.6) used another version of the trick—representation of uniform order statistics as rescaled points of a poisson process—in relating the empirical process to a partial-sum process. The quantile-transformation trick belongs to the folklore.

The interest aroused when Durbin (1973a) applied weak convergence methods to get limit distributions for statistics analogous to those of Kolmogorov and Smirnov, but with estimated parameters, died down when the intractable limit processes asserted themselves. The papers in the conference on empirical processes (Gaenssler and Révész 1976), the lecture notes by Durbin (1973b), and Pollard (1980), offer several perspectives on minimum distance methods. Parr (1982) has compiled a bibliography.

The existence proof in Section 3 adapts an argument of Dudley (1966a, Proposition 2; 1978, Lemma 1.3).

Erdős and Kac (1946, 1947) proved that the limit distributions for certain functions of partial sums of independent random variables depend on the summands only through their first two moments. Their paper led Donsker (1951) to formulate the results as limit theorems for functionals on a partial-sum process. Donsker proved convergence of the process to brownian motion. Skorohod (1957) studied processes with independent increments in great detail; a nice account of some of his results has appeared in Gihman and Skorohod (1969, Chapter IX).

Infinite time scales with various topologies (uniform or Skorohod) of convergence on compacta often cause minor confusion. The simple theory for $C[0, \infty)$ (Whitt 1970) proves remarkably unwilling to carry over to $D[0, \infty)$, at least for Skorohod metrics (Lindvall 1973, Stone 1963). The obvious nature of the main theorem (Theorem 23) for uniform metrics makes it hard to appreciate why Skorohod metrics should be any more difficult. More about this in Chapter VI.

Billingsley (1968, Section 11) has derived distributions for functionals of brownian motion by taking limits for functionals of simple random walks (partial-sum processes). Stopping-time arguments for brownian motion give other first-exit distributions (Breiman 1968, Section 13.7). The results can also be derived by elegant martingale methods (Loève 1978, Complements to Section 42).

Problems

[1] $D[0, 1]$ is complete under its uniform metric. [If $\{x_n\}$ is a uniform Cauchy sequence prove pointwise, and then uniform, convergence. Show that uniform convergence preserves the cadlag property.]

[2] Let C be a separable subset of $D[0, 1]$ under its uniform metric. There exists a countable subset T_0 of $[0, 1]$ such that functions in C can have discontinuities only at points of T_0. [A function in $D[0, 1]$ has only finitely many jumps $> \varepsilon$, otherwise the cadlag property would fail at a cluster point of those jumps. Take the union over rational ε. If x is a uniform limit of $\{x_n\}$ then it can jump only where one of the $\{x_n\}$ jumps.]

[3] Calculate the covariances of the tied-down process $B(t) - tB(1)$ to show that its fidis agree with those of the brownian bridge. [Means and covariances determine a multivariate normal distribution.]

[4] Extend the proof of Theorem 3 to limit processes with jumps. [If X has paths in a separable set C, and if the grid points for the approximation maps A_k pick up every jump point of C as $k \to \infty$, then $\|A_k X - X\| \to 0$ almost surely.]

[5] Would the argument in the proof of Theorem 9 for bounding $U_n(t)$ over $[0, b]$ work for a different interval, say $[1 - b, 1]$? This would be necessary if U_n were not stationary; direct analysis of the non-uniform empirical process would require such an extension. [Try a different definition for $\mathscr{E}_t$.]

[6] Every function in $D[-\infty, \infty]$ is bounded. [If $|x(t_n)| \geq n$ then x would not have the cadlag property at a cluster point of $\{t_n\}$.]

[7] Give a direct proof of the Empirical Central Limit Theorem for sampling from a general F. [Make sure the jump points of F appear in the sequence of grids from which fidi approximations are calculated.]

[8] The image of a separable metric space under a continuous map is separable. The empirical process E concentrates on a separable subset of $D[-\infty, \infty]$. [The image of a dense subset is dense.]

[9] Suppose the distribution function $F(x, \theta)$ has a bounded partial derivative $\Delta(x, \theta)$ with respect to θ. If $\Delta(\cdot, \cdot)$ is uniformly continuous (or even just $\Delta(x, \cdot)$ equicontinuous), then

$$\sup_x |F(x, \theta) - F(x, \theta_0) - (\theta - \theta_0)\Delta(x, \theta)| = o(\theta - \theta_0).$$

[10] Find the limiting distribution of Kolmogorov's goodness-of-fit statistic for sampling from the $N(\theta, 1)$ distribution, when θ is estimated by the sample mean. Show that this limit does not depend on the true value θ_0; express it as a functional of the gaussian process with covariance function $\Phi(s)[1 - \Phi(t)] - \phi(s)\phi(t)$, where Φ and ϕ denote the distribution function and density function of the $N(0, 1)$ distribution.

[11] Here is another construction for the brownian bridge.

(a) Temporarily suppose there exists a brownian bridge U. Define, recursively, new processes $Y_0, Y_1, \ldots$ and $Z_0, Z_1, \ldots$ by setting $Y_0 = U$, $Y_{n+1} = Y_n - Z_n$, and

$$Z_n(t) = \sum_{j=1}^{2^n} h_{nj}(t) Y_n((2j - 1)/2^{n+1}),$$

where h_{nj} is the function whose graph is an isosceles triangle of height one sitting on the base $[(j - 1)/2^n, j/2^n]$. Show that Y_{n+1} and Z_n are independent. Deduce that the $\{Z_n\}$ are mutually independent. [Calculate covariances. The process Y_n would be obtained from the brownian bridge by tying it down at the points $\{(2j - 1)/2^n\}$; it is made up of 2^n independent, rescaled brownian bridges sitting side by side. The covariance calculations all reduce to the same thing.]

(b) Show that the process $Z_0 + \cdots + Z_n$ interpolates linearly between the vertices $(j/2^{n+1}, U(j/2^{n+1}))$, for $j = 0, \ldots, 2^{n+1}$.

(c) Now run the argument the other way. Construct processes $\{X_n\}$ with the same distributions as the $\{Z_n\}$, then recover the brownian bridge as a limit of sums of the $\{X_n\}$. Let $\{\xi_{nj}: j = 1, \ldots, 2^n;\ n = 0, 1, \ldots\}$ be independent random variables, with ξ_{nj} distributed $N(0, 1/2^{n+2})$. Define

$$X_n(t) = \sum_{j=1}^{2^n} h_{nj}(t)\xi_{nj}.$$

At the points $\{j/2^{n+1}: j = 0, \ldots, 2^{n+1}\}$ the sum $X_0 + \cdots + X_n$ has the right fidis for a brownian bridge. [Only the fidis of U, which we know are well defined, were needed to calculate the distributional properties of $Z_0 + \cdots + Z_n$.]

(d) Show that $\mathbb{P}\{\|X_n\| \geq \varepsilon_n\} \leq 2^{n+1}\exp(-2^{n+1}\varepsilon_n^2)$. [Apply the exponential inequality from Appendix B for normal tails: $\|X_n\|$ is a maximum of 2^n independent $|N(0, 1/2^{n+2})|$ random variables.]

(e) By choosing $\varepsilon_n = (2n/2^{n+1})^{1/2}$ and then applying the Borel–Cantelli lemma, show that $\sum_{n=0}^{\infty} X_n(t)$ converges uniformly in t, almost surely; it defines a process X with continuous sample paths, almost surely.

(f) At dyadic rational values for t, the series for X contains only finitely many non-zero terms. The fidi projections of X at a dense subset of $[0, 1]$ have the distributions of a brownian bridge.

The process X, with a negligible set of sample paths discarded, is a brownian bridge.

[12] Every point of $D[0, \infty)$ is completely regular. [The distance $d_k(x, y)$ can be expressed as a supremum involving only rational time points. For each x and k, the function $d_k(x, \cdot)$ is projection measurable. Use $[1 - md_k(x, \cdot)]^+$ as a separating function.]

[13] $C[0, \infty)$ is a closed, separable subset of $D[0, \infty)$. [It is the closure of a countable collection of piecewise linear, continuous functions, each constant over an interval of the form $[t, \infty)$.]

[14] For each fixed τ and each finite subset S of $[0, \infty)$, the random vector $\pi_S B_\tau$ is independent of $\mathscr{E}_\tau$, because brownian motion has independent increments. Deduce independence of B_τ and $\mathscr{E}_\tau$. [The projection maps $\{\pi_S\}$ generate the projection σ-field on $C[0, \infty)$.]

[15] Let P and Q be probability measures on the σ-field $\mathscr{B}_0$ generated by the closed balls of a metric space. If $Pf = Qf$ for every bounded, uniformly continuous, $\mathscr{B}_0$-measurable f then $P = Q$. [Every closed ball is a pointwise decreasing limit of a sequence of such functions. The same is true for the intersection of any finite collection of closed balls. These sets form a generating class for $\mathscr{B}_0$ that is closed under the formation of finite intersections.]

[16] The borel σ-field on the product $\mathscr{X} \otimes \mathscr{Y}$ of two separable metric spaces coincides with the product σ-field $\mathscr{B}(\mathscr{X}) \otimes \mathscr{B}(\mathscr{Y})$. [Every open set in the product space is a countable union of sets of the form $G_x \otimes G_y$, with both G_x and G_y open. Compare with Problem IV.5.]

[17] Show that the shifted process B_τ is projection measurable for each random time τ. [Prove measurability of $B(\omega, t + \tau(\omega))$ by writing it as a composition of two

measurable maps: $\omega \mapsto (\tau(\omega), B(\omega, \cdot))$ from Ω into $[0, \infty] \otimes C[0, \infty)$ and $(s, x) \mapsto \lim_n x((t + s) \wedge n)[1 \wedge (n - s)^+]$ from $[0, \infty] \otimes C[0, \infty)$ into $\mathbb{R}$.

[18] In the sense of Example 24, almost all brownian motion sample paths are good. If τ denotes the first time that B hits level 1, prove that

$$\mathbb{P}\{B_\tau(s) \leq 0 \text{ for } 0 \leq s \leq \delta\} = 0.$$

Let δ tend to zero through a sequence to show that bad paths belong to a set of probability zero. [You could try letting the level α in Example 26 sink to zero.]

CHAPTER VI

The Skorohod Metric on $D[0, \infty)$

... in which an alternative to the metric of uniform convergence on compacta is studied. With the new metric the limit processes need not confine their jumps to a countable set of time points. Amongst the convergence criteria developed is an elegant condition based on random increments, due to Aldous. The chapter might be regarded as an extended appendix to Chapter V.

VI.1. Properties of the Metric

The uniform metric on $D[0, 1]$ is the best choice for applications where the limit distribution concentrates on $C[0, 1]$, or on some other separable subset of $D[0, 1]$. It is well suited for convergence to brownian motion, brownian bridge, and the gaussian processes that appear as limits in the Empirical Central Limit Theorem. But it excludes, for example, poisson processes and other non-gaussian processes with independent increments, whose jumps are not constrained to lie in a fixed, countable subset of $[0, 1]$. To analyze such processes, Skorohod (1956) introduced four new metrics, all weaker than the uniform metric. Of these, the J_1 metric has since become the most popular. (Too popular in my opinion—too often it is dragged into problems for which the uniform metric would suffice.) But Skorohod's J_1 metric on $D[0, 1]$ will not be the main concern of this chapter. Instead we shall investigate a sort of J_1 convergence on compacta for $D[0, \infty)$, the space where the interesting applications live.

With the results from Section V.5 in mind, and even without seeing the J_1 metric defined, you might suspect that convergence $X_n \rightsquigarrow X$ of random elements of $D[0, \infty)$ should reduce to convergence of their restrictions to each finite interval $[0, T]$, in the sense of the J_1 metric on $D[0, T]$. This is almost true. We need to avoid those values of T at which X has positive probability of jumping. The difficulty arises because projection maps are not automatically continuous for J_1 metrics. Both the points 0 and T require special treatment from the J_1 metric on $D[0, T]$, whereas only 0 has an *a priori* right to special treatment in $D[0, \infty)$. That tiny distinction makes it slightly more convenient to study $D[0, \infty)$ directly than to deduce all its properties from those of $D[0, T]$. As we shall not be concerned with

Skorohod's J_2, M_1, and M_2 metrics, let us drop the J_1 designation from now on.

1 Definition. For each finite T and each pair of functions x and y in $D[0, \infty)$ define the distance $d_T(x, y)$ as the infimum of all those values of δ for which there exist grids $0 = t_0 < t_1 < \cdots < t_k$, with $t_k \geq T$, and $0 = s_0 < s_1 < \cdots < s_k$, with $s_k \geq T$, such that $|t_i - s_i| \leq \delta$ for $i = 0, \ldots, k$, and

$$|x(t) - y(s)| \leq \delta \quad \text{if} \quad t_i \leq t < t_{i+1} \quad \text{and} \quad s_i \leq s < s_{i+1},$$

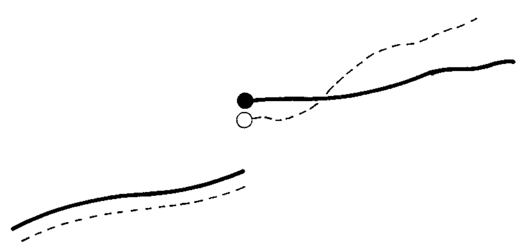
for $i = 0, \ldots, k - 1$. The weighted sum

$$d(x, y) = \sum_{k=1}^{\infty} 2^{-k} \min[1, d_k(x, y)]$$

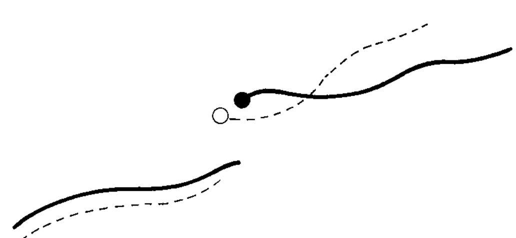
defines the Skorohod metric on $D[0, \infty)$. □

The proof that d is a metric contains no surprises (Problem 1). The requirement $t_k \geq T$ prevents discontinuities near T from overinflating the distance $d_T(x, y)$: if, say, y jumps just before T then it can be matched by a similar jump in x just a little beyond T. By allowing d_T to depend on more than the segment of path over $[0, T]$, we avoid a difficulty that other authors have encountered in defining a metric for J_1 convergence on compacta.

The main difference between convergence in the uniform sense and convergence in the J_1 sense appears at the discontinuity points of a limit function. If $\{x_n\}$ converges uniformly on compacta to x and if x has a jump



at t_0, then each x_n, for n large, must have a jump of almost the same magnitude precisely at t_0. Skorohod's metric still forces each x_n to have a jump of almost the same magnitude, but not precisely at t_0.



Straight from the definition we deduce that convergence of x_n to x under the d metric is equivalent to $d_T(x_n, x) \to 0$ for each finite T. Another equivalent form is (Problem 2):

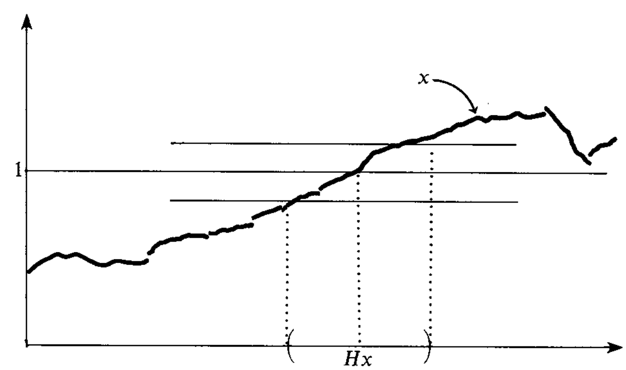
> $d(x_n, x) \to 0$ if and only if there exist continuous, strictly increasing maps $\{\lambda_n\}$ from $[0, \infty)$ onto itself such that, uniformly over compact sets of t values, $\lambda_n(t) - t \to 0$ and $x(\lambda_n(t)) - x_n(t) \to 0$.

Sometimes continuity of a functional on $D[0, \infty)$ can be verified more easily when convergence is invoked in this form.

2 Example. Consider once again the first passage time functional from Example V.24: for x in $D[0, \infty)$,

$$Hx = \inf\{t \geq 0 : x(t) > 1\}$$

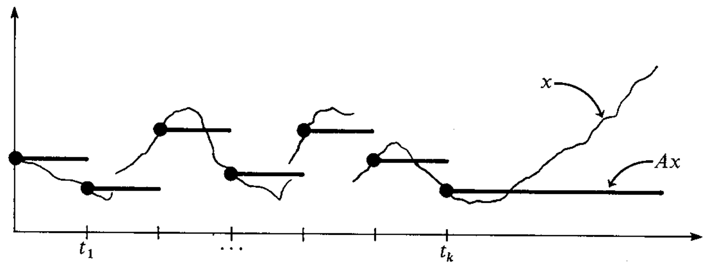
with the infimum over an empty set defined to be $+\infty$. The functional is continuous, in the sense of the Skorohod metric, at every good path x.



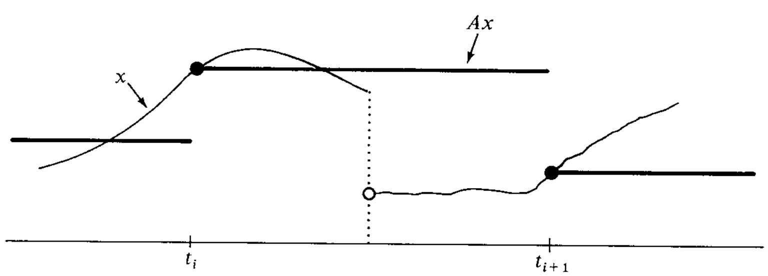
If $d(x_n, x) \to 0$ then the functions $\{y_n\}$ constructed from $\{x_n\}$ and $\{\lambda_n\}$ by setting $y_n(t) = x_n(\lambda_n^{-1}(t))$ converge uniformly to x over compact sets of t values. The argument from Example V.24 gives $Hy_n \to Hx$. Complete the proof of continuity of H by checking that $Hy_n = \lambda_n(Hx_n)$. □

As in Chapter V, we shall find a criterion for weak convergence in $D[0, \infty)$ by considering the approximations to processes built from the values they take at a fixed, finite grid of points $0 = t_0 < t_1 < \cdots < t_k$. Define the interpolation map A from $D[0, \infty)$ into $D[0, \infty)$ by

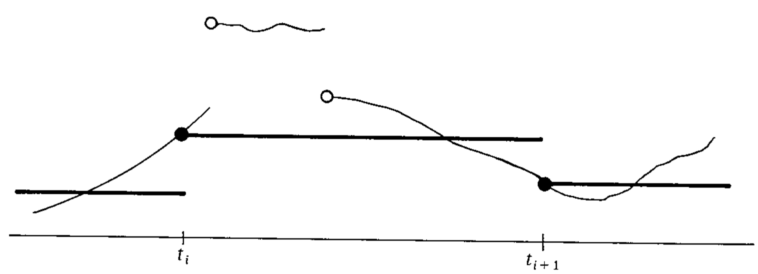
$$(Ax)(t) = x(t_k)\{t_k \leq t < \infty\} + \sum_{i=0}^{k-1} x(t_i)\{t_i \leq t < t_{i+1}\}.$$



All the jumps of Ax occur at grid points. For a fine grid, x and Ax will be close if x does not vary much between the grid points, except possibly for a single jump in each interval. In that case values of x before the jump are close to the value at the left grid point; values of x after the jump are close to the value at the right grid point:



Multiple jumps, though, could be missed completely:



Bad behavior of x in $[t_k, \infty)$ affects $d(x, Ax)$ only through those d_T with $T \geq t_k$. Good behavior of x in $[0, T]$ may be quantified by a modulus function.

3 Definition. For each x in $D[0, \infty)$ and each finite interval $[a, b]$, the modulus function $\Delta(x, [a, b])$ equals the infimum of those positive ε for which there exists a point s in $(a, b]$ such that $|x(t) - x(a)| < \varepsilon$ if $a \leq t < s$ and $|x(t) - x(b)| < \varepsilon$ if $s \leq t \leq b$. □

4 Lemma. *Let Ax be the approximation to x constructed from the values taken at grid points $0 = t_0 < \cdots < t_k = T$, with $t_i - t_{i-1} \le \delta$ for each i. Then*

$$d_T(x, Ax) \le \delta \vee \max_{i \le k} \Delta(x, [t_{i-1}, t_i]).$$

Proof. Write Δ for the maximum $\Delta(x, [t_{i-1}, t_i])$ value. Choose $\varepsilon > 0$. The definition of the modulus function gives points $\{s_i\}$ with

$$0 = s_0 = t_0 < s_1 \le t_1 < s_2 \le \cdots < s_k \le t_k$$

and

$$|x(s) - x(t_i)| < \Delta + \varepsilon \quad \text{if} \quad s_i \le s < s_{i+1} \quad \text{for} \quad i = 0, \ldots, k-1,$$

$$|x(s) - x(t_k)| < \Delta + \varepsilon \quad \text{if} \quad s_k \le s \le t_k.$$

Possibly s_k will be strictly less than T. Add on one more point s_{k+1} with $t_k < s_{k+1} < t_k + \delta$ and

$$|x(s) - x(t_k)| < \Delta + \varepsilon \quad \text{for} \quad t_k \le s \le s_{k+1}.$$

Set $t_{k+1} = s_{k+1}$. The extra points make $\{s_i\}$ and $\{t_i\}$ suitable grids for the calculation of $d_T(x, Ax)$. Since $(Ax)(t) = x(t_i)$ throughout $[t_i, t_{i+1})$,

$$|(Ax)(t) - x(s)| < \Delta + \varepsilon \quad \text{for} \quad t_i \le t < t_{i+1} \quad \text{and} \quad s_i \le s < s_{i+1}.$$

Deduce that $d_T(x, Ax) \le \delta \vee (\Delta + \varepsilon)$, then let ε tend to zero. □

The lemma shows that the error committed in the approximation of x by Ax depends upon the variability of x, as measured by the modulus function. If we sufficiently refine the grid used to construct Ax, the approximation improves; for a fixed x, we can make the error as small as we please by choosing the grid fine enough.

5 Lemma. *For each fixed x in $D[0, \infty)$, each fixed T, and each $\varepsilon > 0$, there exists a $\delta > 0$ such that*

$$\max_i \Delta(x, [t_{i-1}, t_i]) < \varepsilon$$

for every grid $0 = t_0 < \cdots < t_k = T$ with $t_i - t_{i-1} \le \delta$ for each i.

Proof. If $\Delta(x, [\alpha, \alpha']) > \varepsilon$, we get points $\alpha < \tau < \beta \le \alpha'$ with

$$|x(\tau) - x(\alpha)| \ge \tfrac{1}{2}\varepsilon \quad \text{and} \quad |x(\tau) - x(\beta)| \ge \tfrac{1}{2}\varepsilon$$

by setting

$$\tau = \inf\{t > \alpha : |x(t) - x(\alpha)| \ge \tfrac{1}{2}\varepsilon\},$$

$$\beta = \inf\{t > \tau : |x(t) - x(\tau)| \ge \tfrac{1}{2}\varepsilon\}.$$

So if the assertion were false there would exist sequences of points $\alpha_n < \tau_n < \beta_n \leq T$ for which $\beta_n - \alpha_n \to 0$ and

$$|x(\tau_n) - x(\alpha_n)| \geq \tfrac{1}{2}\varepsilon \quad \text{and} \quad |x(\tau_n) - x(\beta_n)| \geq \tfrac{1}{2}\varepsilon.$$

We could extract subsequences along which $\alpha_n \to \sigma, \tau_n \to \sigma, \beta_n \to \sigma$, for some σ in $[0, T]$. This would violate the cadlag property at σ: there must exist a $\delta > 0$ for which

$$|x(s) - x(\sigma-)| < \tfrac{1}{2}\varepsilon \quad \text{for} \quad \sigma - \delta \leq s < \sigma,$$

$$|x(s) - x(\sigma)| < \tfrac{1}{2}\varepsilon \quad \text{for} \quad \sigma \leq s < \sigma + \delta,$$

but eventually both members of at least one of the pairs (α_n, τ_n) or (τ_n, β_n) would have to lie on the same side of σ. □

Our main use for Lemmas 4 and 5 will be the construction of finite-dimensional approximating processes to characterize convergence in distribution for random elements of $D[0, \infty)$. They also help us dispose of measurability complications.

6 Theorem. *Under its Skorohod metric $D[0, \infty)$ is separable. The borel σ-field coincides with the projection σ-field.*

Proof. Write D_0 for the countable class of functions that take constant, rational values on each of the intervals $[0, 1/N), \ldots, [(N^2 - 1)/N, N)$, $[N, \infty)$ for $N = 1, 2, \ldots$. Let us show that the class D_0 is dense in $D[0, \infty)$.

Write $S(N)$ for the set of grid points $\{j/N : j = 0, \ldots, N^2\}$. As usual, $\pi_{S(N)}$ denotes the map that projects an x in $D[0, \infty)$ onto its vector of values at the points of $S(N)$. Define h_N as the continuous map from $\mathbb{R}^{S(N)}$ into $D[0, \infty)$ that takes a vector $\mathbf{r} = (r_0, \ldots, r_{N^2})$ onto the step function with constant value r_j on the jth grid interval. If the components of $\mathbf{r}$ are rational, $h_N(\mathbf{r})$ belongs to D_0.

Given x and an $\varepsilon > 0$, choose N with 2^{-N} less than ε and N^{-1} less than the δ of Lemma 5. We may assume that $\delta < \varepsilon$. Construct the approximation $A_N x$ to x using the grid points $S(N)$. That is, set $A_N = h_N \circ \pi_{S(N)}$. From Lemma 4, $d_N(x, A_N x) \leq \varepsilon$. Hence $d(x, A_N x) < 2\varepsilon$. Find rational numbers r_j with $|r_j - x(j/N)| < \varepsilon$. Then $h_N(\mathbf{r})$ belongs to D_0 and $d(x, h_N(\mathbf{r})) < 3\varepsilon$.

For the moment write $\mathscr{B}$ for the borel σ-field and $\mathscr{P}$ for the projection σ-field on $D[0, \infty)$. To prove that $\mathscr{P} \subseteq \mathscr{B}$, observe that $x(t_0) = \lim H_n(x)$ for fixed t_0, where

$$H_n(x) = \sup_t x(t)s_n(t),$$

$$s_n(t) = (1 - n|t - t_0 - 2/n|)^+.$$

For fixed n, the functional H_n is continuous (Problem 3). As a pointwise limit of continuous functions on $D[0, \infty)$, the projection π_{t_0} must be $\mathscr{B}$-measurable. The projections generate $\mathscr{P}$.

To prove that $\mathscr{B} \subseteq \mathscr{P}$, it is enough to establish that each continuous, real function f on $D[0, \infty)$ is $\mathscr{P}/\mathscr{B}(\mathbb{R})$-measurable. (Every closed set in a metric space can be represented as $f^{-1}\{0\}$ for some continuous f.)

We know that $d(x, h_N \circ \pi_{S(N)}(x)) \to 0$ as $N \to \infty$. Thus $f \circ h_N \circ \pi_{S(N)}(x) \to f(x)$ for each x, by continuity of f. The map $f \circ h_N$ is continuous from $\mathbb{R}^{S(N)}$ into $\mathbb{R}$, and hence $\mathscr{B}(\mathbb{R}^{S(N)})/\mathscr{B}(\mathbb{R})$-measurable. The map $\pi_{S(N)}$ is, by definition, $\mathscr{P}/\mathscr{B}(\mathbb{R}^{S(N)})$-measurable. Thus their composition $f \circ h_N \circ \pi_{S(N)}$ must be $\mathscr{P}/\mathscr{B}(\mathbb{R})$-measurable. As a pointwise limit of such functions, f must also be $\mathscr{P}/\mathscr{B}(\mathbb{R})$-measurable. □

Needless to say, from now on we shall always equip $D[0, \infty)$ with its projection σ-field, alias the borel σ-field for the Skorohod metric. Every point of $D[0, \infty)$ is completely regular under this σ-field.

7 Example. The asymptotic theory for maxima of independent random variables bears some similarity to the theory for sums of independent random variables. The role played by the normal is taken over by the extreme-value distributions, whose distribution functions are of the form $\exp(-G(x))$ for $G(x)$ equal to one of e^{-x}, or $x^{-\alpha}\{x \geq 0\}$, or $(-x)^{-\alpha}\{x \leq 0\}$, with α a positive parameter. If the maximum M_n of n independent observations from a distribution function F can be standardized to converge in distribution, then the limit must be one of these: for constants a_n (positive) and b_n,

$$(8) \qquad \mathbb{P}\{M_n \leq a_n x + b_n\} = F^n(a_n x + b_n) \to \exp(-G(x)).$$

This convergence implies a much stronger result for the joint asymptotic behavior of the maxima at different sample sizes, a result analogous to the convergence of the partial-sum process to brownian motion (Example V.20).

Define the maxima process $Y_n(\cdot)$ as the random element of $D[0, \infty)$ with

$$Y_n(t) = (M_j - b_n)/a_n \quad \text{for} \quad j/n \leq t < (j+1)/n.$$

The assumption (8) gives convergence for $\{Y_n(1)\}$. Using only the facts about the Skorohod metric that we have so far accumulated, we can strengthen this to convergence in distribution of the $\{Y_n\}$ process.

The method of proof depends upon a representation of Y_n as a continuous transformation of a poisson process. To minimize extraneous detail, assume $G(x) = x^{-\alpha}\{x \geq 0\}$. Trivial modifications of the argument would cover the other two cases.

Define a sequence of measures $\{H_n\}$ on $(0, \infty)$ by means of their distribution functions:

$$H_1(0, x] = \exp(-G(x)),$$

$$H_n(0, x] = n \exp(-G(x)/n) - (n-1)\exp(-G(x)/(n-1)).$$

Calculate their density functions if you doubt that these are well-defined measures. On $S = [0, \infty) \otimes (0, \infty)$ generate independent poisson processes

$\{\pi_n\}$ with intensity measures $\{\lambda \otimes H_n\}$, where λ denotes lebesgue measure on $[0, \infty)$. The sum $\sigma_n = \pi_1 + \cdots + \pi_n$ is also a poisson process. As n tends to infinity, σ_n increases to a poisson process σ on S with intensity measure

$$\sum_{i=1}^{\infty} \lambda \otimes H_i = \lambda \otimes \lim_n \sum_{i=1}^{n} H_i = \lambda \otimes \gamma,$$

the measure γ being determined on $(0, \infty)$ by

$$\gamma(x, \infty) = \lim_n [n - n \exp(-G(x)/n)] = G(x).$$

Label the points of σ_n as (η_{ni}, h_{ni}), where $\eta_{n1} < \eta_{n2} < \cdots$. The $\{\eta_{ni}\}$ form a poisson process on $[0, \infty)$, with intensity $n\lambda$, independent of the $\{h_{ni}\}$; the gaps between adjacent η_{ni} have independent exponential distributions with mean n^{-1}. The $\{h_{ni}\}$ are independent observations on the distribution function $\exp(-G/n)$. When subjected to a slight vertical perturbation they will become standardized observations on F.

Let Q be the quantile transformation corresponding to the distribution function F (Section III.6). Define

$$T_n(y) = [Q(\exp(-G(y)/n)) - b_n]/a_n.$$

For large n, this transformation hardly disturbs y:

$$\begin{aligned} T_n(y) &= \inf\{(z - b_n)/a_n : F(z) \geq \exp(-G(y)/n)\} \\ &= \inf\{x : F^n(a_n x + b) \geq \exp(-G(y))\} \\ &\to \inf\{x : \exp(-G(x)) \geq \exp(-G(y))\} \quad \text{by (8)} \\ &= y. \end{aligned}$$

The transformed variables $\{a_n T_n(h_{ni}) + b_n\}$ form a sequence of independent observations on F, because h_{ni} has distribution function $\exp(-G/n)$:

$$\begin{aligned} \mathbb{P}\{a_n T_n(h_{ni}) + b_n \leq x\} &= \mathbb{P}\{Q(\exp(-G(h_{ni})/n)) \leq x\} \\ &= \mathbb{P}\{\text{Uniform}(0, 1) \leq F(x)\} \\ &= F(x). \end{aligned}$$

The T_n has the desired effect, in the vertical direction, on the points of σ_n.

Define a random element Z_n of $D[0, \infty)$ by setting

$$Z_n(t) = \sup\{T_n(h_{ni}) : \eta_{ni} \leq t\}.$$

If Z_n had its jumps at $n^{-1}, 2n^{-1}, \ldots$ instead of at $\eta_{n1}, \eta_{n2}, \ldots$ it would be a probabilistic copy of Y_n. Remedy the defect by applying to the time axis the random, piecewise linear transformation γ_n that sends 0 onto 0 and j/n onto η_{nj}, for $j = 1, 2, \ldots$. The processes $Y_n(\cdot)$ and $Z_n \circ \gamma_n(\cdot)$ have the same distribution as random elements of $D[0, \infty)$.

By the weak law of large numbers, $\gamma_n(t) \to t$ in probability uniformly on compact intervals (Problem 4). Thus $d(Z_n, Z_n \circ \gamma_n) \to 0$ in probability. The random elements $\{Z_n\}$ themselves converge almost surely to the random element

$$Z(t) = \sup\{h_i : \eta_i \le t\},$$

where $(\eta_1, h_1), (\eta_2, h_2), \ldots$ denote the points of the poisson process σ arranged in order of increasing time coordinate. Deduce that $\{Z_n\}$, and hence $\{Y_n\}$, converges in distribution to Z. □

VI.2. Convergence in Distribution

In Section V.1 we found a necessary and sufficient condition for convergence in distribution of random elements $\{X_n\}$ of $D[0, 1]$, under its uniform metric, to a limit process X concentrating on a separable subset. The separability allowed X to have discontinuities only at fixed locations. For the proof we constructed an approximation AX_n to each X_n based on the values it took at a fixed finite grid on $[0, 1]$. The conditions we imposed ensured that, with high probability, the AX_n process was uniformly close to X_n.

A similar method of proof will apply for convergence in distribution of random elements of $D[0, \infty)$ under d, its Skorohod metric. The constraint on the limit process will disappear, because $D[0, \infty)$ itself is separable under d. Each approximation AX_n will, with high probability, be close to its X_n in the sense of d distance. But one extra complication will arise because the fidi projections are not automatically continuous.

If x belongs to $D[0, \infty)$ and $x(\tau) \ne x(\tau-)$, the projection map π_τ is not continuous at x. For example, if $x_n(t) = x(nt/(n+1))$ then $d(x_n, x) \to 0$ but $\pi_\tau x_n \to x(\tau-) \ne \pi_\tau x$. For τ a continuity point of x, however, π_τ is continuous at x (Problem 5). Necessarily, π_0 is continuous at every x, because every increasing λ that maps $[0, \infty)$ onto itself must set $\lambda(0)$ equal to 0. For a random element X of $D[0, \infty)$, the projection π_τ will be continuous at all sample paths except those that have a jump at τ.

9 Lemma. *For each random element X of $D[0, \infty)$ there exists a subset Γ_X of $[0, \infty)$ such that $[0, \infty)\backslash\Gamma_X$ is countable and $\mathbb{P}\{X(t) = X(t-)\} = 1$ for t in Γ_X. The projection π_t is $\mathbb{P}_X$ almost surely continuous at each t in Γ_X.*

Proof. It is enough to show that if $\varepsilon > 0$ then

$$\mathbb{P}\{|X(t) - X(t-)| \ge \varepsilon\} \ge \varepsilon$$

for at most finitely many t values in each bounded interval $[0, T]$. Write $J(t)$ for $\{|X(t) - X(t-)| \ge \varepsilon\}$, If $\mathbb{P}J(t_n) \ge \varepsilon$ for an infinite sequence $\{t_n\}$ of distinct points in $[0, T]$ then $\mathbb{P}\{J(t_n) \text{ infinitely often}\} \ge \varepsilon$. There would exist an ω belonging to infinitely many of the $J(t_n)$ sets. At some cluster point t in

$[0, T]$ the inequality $|X(\omega, t_n) - X(\omega, t_n-)| \geq \varepsilon$ would hold for infinitely many distinct t_n values in every neighborhood of t. This would violate the cadlag property of $X(\omega, \cdot)$ at t. □

Because the projection π_0 is continuous at every x in $D[0, \infty)$, let us also admit 0 as a point of Γ_X, even though $X(0-)$ is not defined.

10 Theorem. *Let $X, X_1, X_2, \ldots$ be random elements of $D[0, \infty)$. Necessary and sufficient conditions for $X_n \rightsquigarrow X$, in the sense of the Skorohod metric, are:*

(i) *the fidis of X_n corresponding to finite subsets of Γ_X converge to the fidis of X;*
(ii) *for each $\varepsilon > 0$, each $\eta > 0$, and each finite T in Γ_X, there exists a grid $0 = t_0 < \cdots < t_K = T$ of points from Γ_X with*

$$\limsup \mathbb{P}\left\{\max_i \Delta(X_n, [t_{i-1}, t_i]) > \eta\right\} < \varepsilon.$$

Proof of Necessity. Appeal to the Representation Theorem (IV.13) for a new sequence $\{\tilde{X}_n\}$, with the same distributions as $\{X_n\}$, and an $\tilde{X}$ with the same distribution as X, for which

$$d(\tilde{X}_n(\omega, \cdot), \tilde{X}(\omega, \cdot)) \to 0 \quad \text{for almost all } \omega. \tag{11}$$

For t in Γ_X,

$$\tilde{X}_n(\omega, t) \to \tilde{X}(\omega, t) \quad \text{at almost every } \omega.$$

Fidi convergence for $\{X_n\}$ at points of Γ_X follows.

Find a grid $0 = t_0 < \cdots < t_k = T$ of points from Γ_X with

$$\mathbb{P}\left\{\max_i \Delta(X, [t_{i-1}, t_i]) \geq \eta\right\} < \varepsilon. \tag{12}$$

Such a grid exists by virtue of Lemma 5: as a sequence of grids is refined down to a countable, dense subset of $[0, T]$, the maximum value of Δ over the grid intervals converges to zero at each sample path of X. Write $\Delta_T(x)$ as an abbreviation for $\max_i \Delta(x, [t_{i-1}, t_i])$.

Consider an ω at which the convergence (11) holds. Write $x_n(\cdot)$ for $\tilde{X}_n(\omega, \cdot)$, and $x(\cdot)$ for $\tilde{X}(\omega, \cdot)$. If we show $\limsup \Delta_T(x_n) \leq 2\Delta_T(x)$, then it will follow that

$$\begin{aligned} \limsup \mathbb{P}\{\Delta_T(X_n) \geq 2\eta\} &= \limsup \mathbb{P}\{\Delta_T(\tilde{X}_n) \geq 2\eta\} \\ &\leq \mathbb{P}\{\limsup \Delta_T(\tilde{X}_n) \geq 2\eta\} \\ &\leq \mathbb{P}\{\Delta_T(\tilde{X}) \geq \eta\} \\ &< \varepsilon \quad \text{by (12)} \end{aligned}$$

as required by hypothesis (ii).

Choose $\delta > 0$. By definition of $\Delta_T(x)$, there exists points $\{s_i\}$ with $t_{i-1} < s_i \le t_i$ and

$$|x(t) - x(t_{i-1})| < \Delta_T(x) + \delta \quad \text{for} \quad t_{i-1} \le t < s_i,$$
$$|x(t) - x(t_i)| \quad < \Delta_T(x) + \delta \quad \text{for} \quad s_i \le t \le t_i.$$

Continuity of x at both t_{i-1} and t_i allows us to assume that the strict inequality $s_i < t_i$ holds, and also that the range of validity for the first inequality is $t_{i-1} - \delta' < t < s_i$, and for the second $s_i \le t < t_i + \delta'$, for some $\delta' > 0$. Because $d(x_n, x) \to 0$, there exists a sequence of continuous, increasing maps $\{\lambda_n\}$ from $[0, \infty)$ onto $[0, \infty)$ such that

$$\lambda_n(t) - t \to 0 \quad \text{and} \quad x(\lambda_n(t)) - x_n(t) \to 0$$

uniformly on compacta. When n is large enough, $t_{i-1} < \lambda_n^{-1}(s_i) < t_i$. Use $s_{i,n} = \lambda_n^{-1}(s_i)$ as the split point in $[t_{i-1}, t_i]$ for bounding $\Delta_T(x_n)$. If $t_{i-1} \le t < s_{i,n}$, and n is large enough,

$$\begin{aligned}|x_n(t) - x_n(t_{i-1})| &\le |x(\lambda_n(t)) - x(\lambda_n(t_{i-1}))| + \delta \\ &\le |x(\lambda_n(t)) - x(t_{i-1})| + |x(t_{i-1}) - x(\lambda_n(t_{i-1}))| + \delta \\ &< 2\Delta_T(x) + 3\delta,\end{aligned}$$

because, eventually, $t_{i-1} - \delta' < \lambda_n(t_{i-1}) \le \lambda_n(t) < \lambda_n(s_{i,n}) = s_i$. A similar argument applies to t in $[s_{i,n}, t_i]$.

Proof of Sufficiency. Let f be a bounded, uniformly continuous, real function on $D[0, \infty)$. We need to show that $\mathbb{P}f(X_n) \to \mathbb{P}f(X)$. Given $\varepsilon > 0$, find $\eta > 0$ so that $|f(x) - f(y)| < \varepsilon$ whenever $d(x, y) \le 2\eta$. Choose from Γ_X a T large enough to ensure $d(x, y) \le d_T(x, y) + \eta$ for every pair x, y.

Let $\Delta_T(\cdot)$ have the same meaning as in the proof of necessity. According to hypothesis (ii) there exists a grid on $[0, T]$ for which

$$\mathbb{P}\{\Delta_T(X_n) > \eta\} < \varepsilon \quad \text{if} \quad n \ge n_0.$$

Also we may assume that the grid points are less than η apart and, by the same reasoning as for (12), that $\mathbb{P}\{\Delta_T(X) > \eta\} < \varepsilon$. Lemma 4 shows that the approximations constructed from this grid are, with high probability, close to the sample paths of the processes:

$$\mathbb{P}\{d_T(X_n, AX_n) > \eta\} < \varepsilon \quad \text{if} \quad n \ge n_0,$$
$$\mathbb{P}\{d_T(X, AX) > \eta\} < \varepsilon.$$

Complete the proof in the usual way. Write A for the interpolation map constructed from the values at the grid points.

$$\begin{aligned}&|\mathbb{P}f(X_n) - \mathbb{P}f(X)| \\ &\quad \le \mathbb{P}|f(X_n) - f(AX_n)| + |\mathbb{P}f(AX_n) - \mathbb{P}f(AX)| \\ &\qquad + \mathbb{P}|f(AX) - f(X)| \\ &\quad \le \varepsilon + 2\|f\|\mathbb{P}\{d(X_n, AX_n) > 2\eta\} + |\mathbb{P}f(AX_n) - \mathbb{P}f(AX)| \\ &\qquad + \varepsilon + 2\|f\|\mathbb{P}\{d(X, AX) > 2\eta\} \\ &\quad \le \varepsilon + 2\|f\|\varepsilon + \varepsilon + \varepsilon + 2\|f\|\varepsilon \quad \text{eventually,}\end{aligned}$$

because $AX_n \rightsquigarrow AX$ and $d(x, Ax) \le d_T(x, Ax) + \eta$ for every x. □

Roughly speaking, the modulus $\Delta(X_n, [t_{i-1}, t_i])$ will be small if X_n has at worst one large jump in the interval $[t_{i-1}, t_i]$. To prove convergence in distribution, we need some way of stopping large jumps from piling up in a small interval. If X_n has a jump at τ_n, which in general will be a random time, then we need

$$\mathbb{P}\{|X_n(\tau_n + t) - X_n(\tau_n)| \text{ small, for all small } t\} \approx 1,$$

and the approximation should hold uniformly in n. A maximal inequality seems required. Contrast this vague, formidable task with an elegant sufficient condition due to Aldous (1978): $X_n \rightsquigarrow X$ if the fidis converge and, for each fixed T,

$$X_n(\rho_n + \delta_n) - X_n(\rho_n) \to 0 \quad \text{in probability,} \tag{13}$$

whenever $\{\delta_n\}$ is a sequence of positive numbers converging to zero and $\{\rho_n\}$ is a sequence of stopping times taking values in $[0, T]$. (The stopping time property means that the event $\{\rho_n \le t\}$ should belong to the σ-field generated by the random variables $X_n(s)$, for $0 \le s \le t$.)

An equivalent form of (13) is: for each T, each $\eta > 0$, and each $\varepsilon > 0$, there exists a $\delta > 0$ and an n_0 such that

$$\mathbb{P}\{|X_n(\rho_n + \delta') - X_n(\rho_n)| \ge \eta\} < \varepsilon \quad \text{for} \quad n \ge n_0, \tag{14}$$

whenever ρ_n is a stopping time for X_n that takes values in $[0, T]$ and δ' is a real number with $0 \le \delta' \le \delta$. The proof of Aldous's result is built up by repeated application of inequality (14).

15 Lemma. *Let Z be a random element of $D[0, \infty)$ for which*

$$\mathbb{P}\{|Z(\rho + \delta') - Z(\rho)| \ge \eta\} < \varepsilon$$

for each real δ' in $[0, \delta]$ and each stopping time ρ taking values in $[0, T]$. If σ and τ are stopping times for which $\sigma \le \tau$ and $|Z(\tau) - Z(\sigma)| \ge 2\eta$ on $\{\tau < \infty\}$, then $\mathbb{P}\{\tau \le T \wedge (\sigma + \frac{1}{2}\delta)\} < 4\varepsilon$.

Proof. Integrate both sides of the inequality for ρ with respect to lebesgue measure on $[0, \delta]$, interchange the order of integration, then make a change of variable in the inner integral.

$$\varepsilon\delta > \mathbb{P} \int \{|Z(s) - Z(\rho)| \ge \eta\}\{\rho \le s \le \rho + \delta\}\, ds.$$

Apply this inequality twice with ρ equal to the stopping times $\sigma_T = T \wedge \sigma$ and $\tau_T = T \wedge \tau$, then add.

$$\begin{aligned} 2\varepsilon\delta > \mathbb{P} \int & \{|Z(s) - Z(\sigma_T)| \ge \eta\}\{\sigma_T \le s \le \sigma_T + \delta\} \\ & + \{|Z(s) - Z(\tau_T)| \ge \eta\}\{\tau_T \le s \le \tau_T + \delta\}\, ds \\ \ge \mathbb{P} \int & [\{Z(s) - Z(\sigma_T)| \ge \eta\} + \{|Z(s) - Z(\tau_T)| \ge \eta\}] \\ & \times \{\tau_T \le s \le \sigma_T + \delta\}\, ds. \end{aligned}$$

On the set $\{\tau \leq T\}$, the sum of the two indicators is at least 1, because at least one of the inequalities

$$|Z(s) - Z(\sigma_T)| \geq \eta, \qquad |Z(s) - Z(\tau_T)| \geq \eta$$

must hold if $\sigma_T = \sigma$, $\tau_T = \tau$, and $|Z(\sigma) - Z(\tau)| \geq 2\eta$. Deduce that

$$\begin{aligned} 2\varepsilon\delta &> \mathbb{P} \int \{\tau \leq T\}\{\tau \leq s \leq \sigma + \delta\}\, ds \\ &\geq \mathbb{P} \int \{\tau \leq T \wedge (\sigma + \tfrac{1}{2}\delta)\}\{\sigma + \tfrac{1}{2}\delta \leq s \leq \sigma + \delta\}\, ds \\ &\geq \tfrac{1}{2}\delta\mathbb{P}\{\tau \leq T \wedge (\sigma + \tfrac{1}{2}\delta)\}. \end{aligned}$$

□

16 Theorem. *Let $X, X_1, X_2, \ldots$ be random elements of $D[0, \infty)$ for which:*

(i) *the fidis of X_n corresponding to finite subsets of Γ_X converge to the fidis of X;*
(ii) *Aldous's condition* (13) *holds.*

Then $X_n \rightsquigarrow X$ in the sense of the Skorohod metric.

Proof. Verify condition (ii) of Theorem 10. Set down a grid $0 = t_0 < \cdots < t_k = T$ of points from Γ_X with the maximum grid interval shorter than $\frac{1}{2}\alpha$, for a value of α that will be specified soon.

Fix an n, then define stopping times for X_n by

$$\begin{aligned} \tau_0 &= 0, \\ \tau_{j+1} &= \inf\{t > \tau_j : |X_n(t) - X_n(\tau_j)| \geq 2\eta\} \end{aligned}$$

with the usual convention that the infimum of the empty set equals $+\infty$. We should perhaps add an extra subscript n to each τ_j.

If $(t_{i-1}, t_i]$ contains at most one of the $\{\tau_j\}$ then we must have $\Delta(X_n, [t_{i-1}, t_i]) \leq 4\eta$. For if $\tau_{j-1} \leq t_{i-1} < \tau_j \leq t_i < \tau_{j+1}$ then

$$\begin{aligned} |X_n(t) - X_n(\tau_{j-1})| &< 2\eta \quad \text{if} \quad t_{i-1} \leq t < \tau_j, \\ |X_n(t) - X_n(\tau_j)| &< 2\eta \quad \text{if} \quad \tau_j \leq t \leq t_i \end{aligned}$$

and hence

$$\begin{aligned} |X_n(t) - X_n(t_{i-1})| &< 4\eta \quad \text{if} \quad t_{i-1} \leq t < \tau_j, \\ |X_n(t) - X_n(t_i)| &< 4\eta \quad \text{if} \quad \tau_j \leq t \leq t_i. \end{aligned}$$

Apply the same reasoning to each grid interval.

$$\begin{aligned} &\left\{\max_i \Delta(X_n, [t_{i-1}, t_i]) > 4\eta\right\} \\ &\qquad \leq \{\text{some } (t_{i-1}, t_i] \text{ contains at least two of the } \{\tau_j\}\} \\ &\qquad \leq \{\text{some pair } \tau_{j-1}, \tau_j \text{ has } \tau_j \leq T \wedge (\tau_{j-1} + \tfrac{1}{2}\alpha)\}. \end{aligned}$$

Fix an integer m, whose value will be specified soon. Bound the last indicator function by

$$\sum_{j=1}^{2m} \{\tau_j \leq T \wedge (\tau_{j-1} + \tfrac{1}{2}\alpha)\} + m^{-1} \sum_{j=1}^{2m} \{\tau_j \leq T \wedge (\tau_{j-1} + T/m)\}.$$

The reasoning here is: either $\tau_{2m} \geq T$, in which case the pair τ_{j-1}, τ_j would be detected by the first sum; or $\tau_{2m} < T$, in which case at least m terms of the second sum must equal one, for otherwise $[0, T]$ would have to contain $(m + 1)$ disjoint intervals (τ_{j-1}, τ_j) of length greater than T/m. Take expectations.

$$\text{(17)} \qquad \mathbb{P}\left\{\max_i \Delta(X_n, [t_{i-1}, t_i]) > 4\eta\right\} \leq 2m \max_{j \leq 2m} \mathbb{P}\{\tau_j \leq T \wedge (\tau_{j-1} + \tfrac{1}{2}\alpha)\} + 2 \max_{j \leq 2m} \mathbb{P}\{\tau_j \leq T \wedge (\tau_{j-1} + T/m)\}.$$

Now we choose m and α.

Invoke Lemma 15 for $Z = X_n$ with $\sigma = \tau_{j-1}$, $\tau = \tau_j$ and the δ provided by (14). For $n \geq n_0$,

$$\mathbb{P}\{\tau_j \leq T \wedge (\tau_{j-1} + \tfrac{1}{2}\delta)\} < 4\varepsilon.$$

Choose m so that $T/m < \frac{1}{2}\delta$; hold it fixed. The second term in the bound (17) is then less than the 8ε if $n \geq n_0$. From (14) find the α for which

$$\mathbb{P}\{|X_n(\rho_n + \delta') - X_n(\rho_n)| \geq \eta\} < \varepsilon/m \quad \text{for} \quad n \geq n_1$$

if $0 \leq \delta' \leq \alpha$ and if ρ_n is a stopping time for X_n with $0 \leq \rho_n \leq T$. Invoke Lemma 15 again, but replace δ by α and ε by ε/m. For $n \geq n_1$,

$$\mathbb{P}\{\tau_j \leq T \wedge (\tau_{j-1} + \tfrac{1}{2}\alpha)\} < 4\varepsilon/m.$$

The first term in the bound (17) is less than 8ε if $n \geq n_1$. In summary:

$$\mathbb{P}\left\{\max_i \Delta(X_n, [t_{i-1}, t_i]) > 4\eta\right\} < 16\varepsilon \quad \text{for} \quad n \geq \max(n_0, n_1),$$

which completes the proof. □

18 Example. For processes with independent increments, the criterion for convergence in distribution is particularly simple. Under the uniform metric for $D[0, 1]$, Theorem V.19 showed that fidi convergence plus an equicontinuity condition on the increments suffices if the limit process X has continuous sample paths. With a minor variation we get Aldous's condition; essentially, we can drop the constraint on the sample paths of X if we reinterpret the result as convergence in the Skorohod sense.

Suppose that for each $\varepsilon > 0$, $\eta > 0$, and $T < \infty$, there exists a $\delta > 0$ and an n_0 such that $\mathbb{P}\{|X_n(t + \delta') - X_n(t)| \geq \eta\} < \varepsilon$ whenever $0 \leq t \leq T$, $0 \leq \delta' \leq \delta$, and $n \geq n_0$. If ρ_n is a stopping time that takes only finitely many different values, and these values all lie in $[0, T]$, then for $n \geq n_0$,

$$\begin{aligned} &\mathbb{P}\{|X_n(\rho_n + \delta') - X_n(\rho_n)| \geq \eta\} \\ &\quad = \sum_t \mathbb{P}[\{\rho_n = t\}\mathbb{P}\{|X_n(t + \delta') - X_n(t)| \geq \eta \,|\, \rho_n = t\}] \\ &\quad < \varepsilon. \end{aligned}$$

Every stopping time can be approximated arbitrarily closely from above by a stopping time that takes only finitely many different values (round up to the next value on a finely spaced grid); the cadlag property of the sample paths carries the inequality over to the stopping times covered by Aldous's condition. □

Problem VIII.8 will give a more interesting application of Theorem 16 to convergence of martingales.

Notes

Much of this chapter draws ideas from Billingsley (1968, Chapter 4) and Gihman and Skorohod (1974, Sections III.4 and VI.5).

Skorohod (1956) defined on $D[0, 1]$ several metrics, which allowed different sorts of behavior of a convergent sequence near a discontinuity of the limit function. Billingsley (1968) introduced a variation on Skorohod's J_1 metric, thereby making $D[0, 1]$ complete. This is not so important if we consider only convergence to a known limit process, but it does greatly simplify the theory if existence of the limit process must be proved by compactness arguments. The metric of Definition 1 is modeled on the metric of Kolmogorov (1956) for $D[0, 1]$.

Whitt (1980), and Lindvall (1973) have shown how difficult it is to write down a metric for J_1 convergence on compacta, as defined by Stone (1963).

The modulus function is a fixed-grid analogue of Skorohod's (1956) Δ_{J_1}, or the Δ_c of Gihman and Skorohod (1974, page 423). Billingsley (1968, Section 14) has defined other moduli for $D[0, 1]$. If existence of the limit process is not assumed in Theorem 10, the conditions will not guarantee its existence; the conditions of the theorem do not translate directly into a characterization of uniform tightness.

Theorem 6 borrows from Parthasarathy (1967, Section VII.6). The elegant argument for inclusion of the borel σ-field in the projection σ-field was attributed by Straf (1969, page 67) to Wichura.

Example 7 is based on the method of Resnick (1975). The idea of embedding the maxima process into a two-dimensional poisson process comes from Pickands (1971).

The convergence criterion of Theorem 10 corresponds to Theorem VI.5.2 of Gihman and Skorohod (1974).

Aldous (1978) proved a result slightly different from Theorem 16; he gave a sufficient condition for uniform tightness in $D[0, 1]$. I found Kurtz's (1981, Chapter 2) rearrangement of the proof helpful.

Skorohod (1957) studied convergence of processes with independent increments. His paper contains many fascinating sample-path arguments.

PROBLEMS

[1] Prove that the d of Definition 1 is a metric on $D[0, \infty)$. [To show that $d(x, y) = 0$ implies $x = y$, fix a t, then choose T larger than t. For each τ with $t < \tau < T$, deduce from the definition of d_T the existence of a sequence $\{s_n\}$ with $s_n \to \tau$ and $y(s_n) \to x(\tau)$. Deduce that either $x(\tau) = y(\tau)$ or $x(\tau) = y(\tau -)$. Choose a sequence $\{\tau_i\}$ strictly decreasing to t. Right continuity of y at t gives both $y(\tau_i) \to y(t)$ and $y(\tau_i -) \to y(t)$; right continuity of x gives $x(\tau_i) \to x(t)$.]

[2] Show that $d(x_n, x) \to 0$ if and only if there exist continuous, strictly increasing maps $\{\lambda_n\}$ from $[0, \infty)$ onto itself such that, uniformly on compact sets of t values,

$$\lambda_n(t) - t \to 0 \quad \text{and} \quad x(\lambda_n(t)) - x_n(t) \to 0.$$

[Construct λ_n as a piecewise linear map that takes gridpoints for x_n onto gridpoints for x, for pairs of grids chosen according to the definition of $d_T(x_n, x)$ with T depending on n.]

[3] Prove continuity of the functional H_n that appeared in the proof of Theorem 6. [If $d(x, x_i) \to 0$, choose continuous, increasing $\{\lambda_i\}$ with $x(\lambda_i(t)) - x_i(t) \to 0$ and $\lambda_i(t) - t \to 0$ uniformly on compact t sets. For i large enough, bound $|H_n(x) - H_n(x_i)|$ by

$$\sup_{t \le c} |x(t)| \sup_{t} |s_n(\lambda_i(t)) - s_n(t)| + \sup_{t \le c} |x(\lambda_i(t)) - x_i(t)|$$

for some constant c.]

[4] Let $\{\gamma_n\}$ be a sequence of random, increasing maps from $[0, \infty)$ onto itself such that $\gamma_n(t) \to t$ in probability, for each fixed t. Show that

$$\sup_{0 \le t \le T} |\gamma_n(t) - t| \to 0 \quad \text{in probability}$$

for each fixed T. [If $|\gamma_n(s) - s| < \varepsilon$ and $|\gamma_n(t) - t| < \varepsilon$ then $|\gamma_n(u) - u| < \varepsilon + |s - t|$ for u between s and t.]

[5] Suppose $d(x_n, x) \to 0$ and that x is continuous at τ. Show that $x_n(\tau) \to x(\tau)$. [Use $x(\lambda_n(\tau)) - x_n(\tau) \to 0$ and $\lambda_n(\tau) \to \tau$.]

[6] If $d(x_n, x) \to 0$ and x belongs to $C[0, \infty)$ then x_n converges to x uniformly on compacta.

[7] If $X_n \rightsquigarrow X$ in the Skorohod sense, and if X has sample paths in $C[0, \infty)$, then $X_n \rightsquigarrow X$ in the sense of the metric for uniform convergence on compacta. [Switch to versions that converge almost surely in the Skorohod sense.]

CHAPTER VII
Central Limit Theorems

... in which the chaining method for proving maximal inequalities for the increments of stochastic processes is established. Applications include construction of gaussian processes with continuous sample paths, central limit theorems for empirical measures, and justification of a stochastic equicontinuity assumption that is needed to prove central limit theorems for statistics defined by minimization of a stochastic process.

VII.1. Stochastic Equicontinuity

Much asymptotic theory boils down to careful application of Taylor's theorem. To bound remainder terms we impose regularity conditions, which add rigor to informal approximation arguments, but usually at the cost of increased technical detail. For some asymptotics problems, especially those concerned with central limit theorems for statistics defined by maximization or minimization of a random process, many of the technicalities can be drawn off into a single stochastic equicontinuity condition. This section shows how. Empirical process methods for establishing stochastic equicontinuity will be developed later in the chapter.

Maximum likelihood estimation is the prime example of a method that defines a statistic by maximization of a random criterion function. Independent observations $\xi_1, \ldots, \xi_n$ are drawn from a distribution P, which is assumed to be a member of a parametric family defined by density functions $\{p(\cdot, \theta)\}$. For simplicity take θ real-valued. The true, but unknown, θ_0 can be estimated by the value θ_n that maximizes

$$G_n(\theta) = n^{-1} \sum_{i=1}^{n} \log p(\xi_i, \theta).$$

Let us recall how one proves asymptotic normality for θ_n, assuming it is consistent for θ_0.

Write $g_0(\cdot, \theta)$ for $\log p(\cdot, \theta)$, and $g_1(\cdot, \theta), g_2(\cdot, \theta), g_3(\cdot, \theta)$, for the first three partial derivatives with respect to θ, whose existence we impose as a regularity condition. Using Taylor's theorem, expand $g_0(\cdot, \theta)$ into

$$g_0(\cdot, \theta_0) + (\theta - \theta_0)g_1(\cdot, \theta_0) + \tfrac{1}{2}(\theta - \theta_0)^2 g_2(\cdot, \theta_0) + \tfrac{1}{6}(\theta - \theta_0)^3 g_3(\cdot, \theta^*)$$

with θ^* between θ_0 and θ. Integrate with respect to the empirical measure P_n.

$$G_n(\theta) = G_n(\theta_0) + (\theta - \theta_0)P_n g_1 + \tfrac{1}{2}(\theta - \theta_0)^2 P_n g_2 + R_n(\theta).$$

If we impose, as an extra regularity condition, the domination

$$|g_3(\cdot, \theta)| \le H(\cdot) \quad \text{for all } \theta,$$

then the remainder term will satisfy

$$|R_n(\theta)| \le \tfrac{1}{6}|\theta - \theta_0|^3 P_n |g_3(\cdot, \theta^*)| \le \tfrac{1}{6}|\theta - \theta_0|^3 P_n H.$$

Assume $PH < \infty$ and $P|g_2| < \infty$. Then, by the strong law of large numbers, for each sequence of shrinking neighborhoods of θ_0 we can absorb the remainder term into the quadratic, leaving

(1) $$G_n(\theta) = G_n(\theta_0) + (\theta - \theta_0)P_n g_1 + \tfrac{1}{2}(\theta - \theta_0)^2(Pg_2 + o_p(1)) \quad \text{near } \theta_0.$$

The $o_p(1)$ stands for a sequence of random functions of θ that are bounded uniformly on the shrinking neighborhoods of θ_0 by random variables of order $o_p(1)$. Provided $Pg_2 < 0$, such a bound on the error of approximation will lead to the usual central limit theorem for $\{n^{1/2}(\theta_n - \theta_0)\}$. As a more general result will be proved soon, let us not pursue that part of the argument further. Instead, reconsider the regularity conditions.

The third partial derivative of $g_0(\cdot, \theta)$ was needed only to bound the remainder term in the Taylor expansion. The second partial derivative enters (1) only through its integrated value Pg_2. But the first partial derivative plays a critical role; its value at each ξ_i comes into the linear term. That suggests we might relax the assumptions about existence of the higher derivatives and still get (1). We can. In place of Pg_2 we shall require a second derivative for $Pg_0(\cdot, \theta)$; and for the remainder term we shall invoke stochastic equicontinuity.

In its abstract form stochastic equicontinuity refers to a sequence of stochastic processes $\{Z_n(t)\colon t \in T\}$ whose shared index set T comes equipped with a semimetric $d(\cdot, \cdot)$. (In case you have forgotten, a semimetric has all the properties of a metric except that $d(s, t) = 0$ need not imply that s equals t.) We shall later need it in that generality.

2 Definition. Call $\{Z_n\}$ stochastically equicontinuous at t_0 if for each $\eta > 0$ and $\varepsilon > 0$ there exists a neighborhood U of t_0 for which

$$\limsup \mathbb{P}\left\{\sup_U |Z_n(t) - Z_n(t_0)| > \eta\right\} < \varepsilon. \qquad \square$$

There might be measure theoretic difficulties related to taking a supremum over an uncountable set of t values. We shall ignore them as far as possible during the course of this chapter. A more careful treatment of measurability details appears in Appendix C.

Because stochastic equicontinuity bounds Z_n uniformly over the neighborhood U, it also applies to any randomly chosen point in the neighborhood.

If $\{\tau_n\}$ is a sequence of random elements of T that converges in probability to t_0, then

(3) $$Z_n(\tau_n) - Z_n(t_0) \to 0 \quad \text{in probability,}$$

because, with probability tending to one, τ_n will belong to each U. When we come to check for stochastic equicontinuity the form in Definition 2 will be the one we use; the form in (3) will be easier to apply, especially when behavior of a particular $\{\tau_n\}$ sequence is under investigation.

The maximum likelihood method generalizes to other maximization problems, where $\{\log p(\cdot, \theta)\}$ is replaced by other families of functions. For future reference it will be more convenient if we pose them as minimization problems.

Suppose $\mathscr{F} = \{f(\cdot, t): t \in T\}$, with T a subset of $\mathbb{R}^k$, is a collection of real, P-integrable functions on the set S where P lives. Denote by P_n the empirical measure formed from n independent observations on P, and define the empirical process E_n as the signed measure $n^{1/2}(P_n - P)$. Define

$$F(t) = Pf(\cdot, t),$$
$$F_n(t) = P_n f(\cdot, t).$$

We shall prove a central limit theorem for sequences $\{\tau_n\}$ that come close enough to minimizing the $\{F_n(\cdot)\}$.

Suppose $f(\cdot, t)$ has a linear approximation near the t_0 at which $F(\cdot)$ takes on its minimum value:

(4) $$f(\cdot, t) = f(\cdot, t_0) + (t - t_0)'\Delta(\cdot) + |t - t_0| r(\cdot, t).$$

For completeness set $r(\cdot, t_0) = 0$. The $\Delta(\cdot)$ is a vector of k real functions on S. Of course, if the approximation is to be of any use to us, the remainder function $r(\cdot, t)$ must in some sense be small near t_0. If we want a central limit theorem for $\{\tau_n\}$, stochastic equicontinuity of $\{E_n r(\cdot, t)\}$ at t_0 is the appropriate sense.

Usually $r(\cdot, t)$ will also tend to zero in the $\mathscr{L}^2(P)$ sense: $P|r(\cdot, t)|^2 \to 0$ as $t \to t_0$. That is, $f(\cdot, t)$ will be differentiable in quadratic mean. In that case, we may work directly with the $\mathscr{L}^2(P)$ seminorm ρ_P on the set $\mathscr{R}$ of all remainder functions $\{r(\cdot, t)\}$. Stochastic equicontinuity of $\{E_n r(\cdot, t)\}$ would then follow from: for each $\varepsilon > 0$ and $\eta > 0$ there exists in $\mathscr{R}$ a neighborhood V of 0 such that

$$\limsup \mathbb{P}\left\{\sup_V |E_n r| > \eta\right\} < \varepsilon.$$

The neighborhood V would take the form $\{r \in \mathscr{R}: \rho_P(r) \le \delta\}$ for some $\delta > 0$. This would be convenient for empirical process calculations. Differentiability in quadratic mean would also imply that $P\Delta = 0$. For if $P\Delta$ were non-zero the integrated form of (4),

$$Pf(\cdot, t) = Pf(\cdot, t_0) + (t - t_0)'P\Delta + o(t - t_0) \quad \text{near } t_0,$$

would contradict existence of even a local minimum at t_0.

5 Theorem. *Suppose $\{\tau_n\}$ is a sequence of random vectors converging in probability to the value t_0 at which $F(\cdot)$ has its minimum. Define $r(\cdot, t)$ and the vector of functions $\Delta(\cdot)$ by* (4). *If*

(i) *t_0 is an interior point of the parameter set T;*
(ii) *$F(\cdot)$ has a non-singular second derivative matrix V at t_0;*
(iii) $F_n(\tau_n) = o_p(n^{-1}) + \inf_t F_n(t)$;
(iv) *the components of $\Delta(\cdot)$ all belong to $\mathscr{L}^2(P)$;*
(v) *the sequence $\{E_n r(\cdot, t)\}$ is stochastically equicontinuous at t_0;*

then $n^{1/2}(\tau_n - t_0) \rightsquigarrow N(0, V^{-1}[P(\Delta\Delta') - (P\Delta)(P\Delta)']V^{-1})$.

Proof. Reparametrize to make t_0 equal to zero and V equal to the identity matrix. Then (ii) implies

$$F(t) = F(0) + \tfrac{1}{2}|t|^2 + o(|t|^2) \quad \text{near } 0.$$

Separate the stochastic and deterministic contributions to the function $F_n(t)$ by writing P_n as the sum $P + n^{-1/2}E_n$. Write $Z_n(t)$ for $E_n r(\cdot, t)$. Stochastic equicontinuity implies $Z_n(\tau_n) = o_p(1)$. For values of t near zero,

$$\begin{aligned} (6) \quad F_n(t) - F_n(0) &= P[f(\cdot, t) - f(\cdot, 0)] + n^{-1/2}E_n[f(\cdot, t) - f(\cdot, 0)] \\ &= \tfrac{1}{2}|t|^2 + o(|t|^2) + n^{-1/2}t'E_n\Delta + n^{-1/2}|t|Z_n(t). \end{aligned}$$

Invoke (iii). Because $F_n(\tau_n)$ comes within $o_p(n^{-1})$ of the infimum, which is smaller than $F_n(0)$,

$$\begin{aligned} o_p(n^{-1}) &\geq F_n(\tau_n) - F_n(0) \\ &= \tfrac{1}{2}|\tau_n|^2 + o_p(|\tau_n|^2) + n^{-1/2}\tau_n' E_n\Delta + o_p(n^{-1/2}|\tau_n|). \end{aligned}$$

The random vector $E_n\Delta$ has an asymptotic $N(0, P(\Delta\Delta') - (P\Delta)(P\Delta)')$ distribution; it is of order $O_p(1)$. Consequently, by the Cauchy–Schwarz inequality, $\tau_n' E_n\Delta \geq -|\tau_n|O_p(1)$. Tidy up the last inequality.

$$\begin{aligned} o_p(n^{-1}) &\geq [\tfrac{1}{2} - o_p(1)]|\tau_n|^2 - n^{-1/2}|\tau_n|O_p(1) - o_p(n^{-1/2}|\tau_n|) \\ &= [\tfrac{1}{2} - o_p(1)][|\tau_n| - O_p(n^{-1/2})]^2 - O_p(n^{-1}). \end{aligned}$$

It follows that the squared term is at most $O_p(n^{-1})$, and hence $\tau_n = O_p(n^{-1/2})$. (Look at Appendix A if you want to see the argument written without the $o_p(\cdot)$ and $O_p(\cdot)$ symbols.) Representation (6) for $t = \tau_n$ now simplifies:

$$\begin{aligned} F_n(\tau_n) &= F_n(0) + \tfrac{1}{2}|\tau_n|^2 + n^{-1/2}\tau_n' E_n\Delta + o_p(n^{-1}) \\ &= F_n(0) + \tfrac{1}{2}|\tau_n + n^{-1/2}E_n\Delta|^2 - \tfrac{1}{2}n^{-1}|E_n\Delta|^2 + o_p(n^{-1}). \end{aligned}$$

The same simplification would apply to any other sequence of t values of order $O_p(n^{-1/2})$. In particular,

$$F_n(-n^{-1/2}E_n\Delta) = F_n(0) - \tfrac{1}{2}n^{-1}|E_n\Delta|^2 + o_p(n^{-1}).$$

Notice the surreptitious appeal to (i). We need $n^{-1/2}E_n\Delta$ to be a point of T before stochastic equicontinuity applies; with probability tending to one as $n \to \infty$, it is.

Now invoke (iii) again, comparing the values of F_n at τ_n and $-n^{-1/2}E_n\Delta$ to get

$$\tfrac{1}{2}|\tau_n + n^{-1/2}E_n\Delta|^2 = o_p(n^{-1}),$$

whence $n^{1/2}\tau_n = -E_n\Delta + o_p(1)$. When transformed back to the old parametrization, this gives

$$\begin{aligned} n^{1/2}V^{1/2}(\tau_n - t_0) &= -V^{-1/2}E_n\Delta + o_p(1) \\ &\rightsquigarrow V^{-1/2}N(0, P(\Delta\Delta') - (P\Delta)(P\Delta)'). \qquad \square \end{aligned}$$

Examples 18 and 19 in Section 4 will apply the theorem just proved. But before we can get to the applications we must acquire the means for verifying the stochastic equicontinuity condition.

VII.2. Chaining

Chaining is a technique for proving maximal inequalities for stochastic processes, the sorts of things required if we want to check the stochastic equicontinuity condition defined in Section 1. It applies to any process $\{Z(t): t \in T\}$ whose index set is equipped with a semimetric $d(\cdot,\cdot)$ that controls the increments:

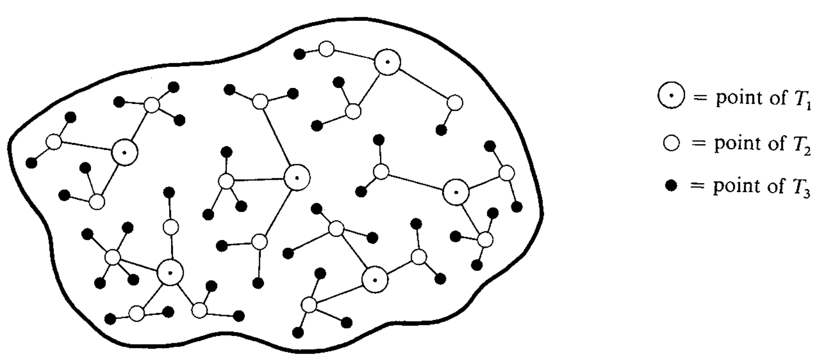
$$\mathbb{P}\{|Z(s) - Z(t)| > \eta\} \le \Delta(\eta, d(s,t)) \quad \text{for} \quad \eta > 0.$$

It works best when $\Delta(\cdot,\cdot)$ takes the form

$$\Delta(\eta, \delta) = 2\exp(-\tfrac{1}{2}\eta^2/D^2\delta^2),$$

with D a positive constant. Under some assumptions about covering numbers for T, the chaining technique will lead to an economical bound on the tail probabilities for a supremum of $|Z(s) - Z(t)|$ over pairs (s, t).

The idea behind chaining, and the reason for its name, is easiest to understand when T is finite. Suppose $T_1, T_2, \ldots, T_{k+1} = T$ are subsets with the property that each t lies within δ_i of at least one point in T_i. Imagine each point of T_{i+1} linked to its nearest neighbor in T_i, for $i = 1, \ldots, k$. From every t stretches a chain with links $t = t_{k+1}, t_k, \ldots, t_1$ joining it to a point in T_1.



The value of the process at t equals its value at t_1 plus a sum of increments across the links joining t to t_1. The error involved in approximating $Z(t)$ by $Z(t_1)$ is bounded, uniformly in t, by

$$\sum_{i=1}^{k} \max |Z(t_{i+1}) - Z(t_i)|.$$

If T_i contains N_i points, the maximum in the ith summand runs over N_{i+1} different increments, each across a link of length at most δ_i. The probability of the summand exceeding η_i is bounded by a sum of N_{i+1} terms, each less than $\Delta(\eta_i, \delta_i)$.

$$\text{(7)} \qquad \mathbb{P}\left\{\max_t |Z(t) - Z(t_1)| > \eta_1 + \cdots + \eta_k\right\} \leq \sum_{i=1}^{k} N_{i+1}\Delta(\eta_i, \delta_i).$$

This inequality is useful if we can choose η_i, δ_i, and T_i to make both the right-hand side and the sum of the $\{\eta_i\}$ small. In that case the maximum of $|Z(s) - Z(t)|$ over all pairs in T is, with high probability, close to the maximum for pairs taken from the smaller class T_1.

When $\Delta(\eta, \delta) = 2\exp(-\frac{1}{2}\eta^2/D^2\delta^2)$, a good combination seems to be: $\{\delta_i\}$ decreasing geometrically and $\{\eta_i\}$ chosen so that $N_{i+1}\Delta(\eta_i, \delta_i) = 2\delta_i$, that is,

$$\eta_i = D\delta_i[2\log(N_{i+1}/\delta_i)]^{1/2}.$$

With these choices the right-hand side of (7) is bounded by the tail of the geometric series $\sum_i \delta_i$, and the sum of the $\{\eta_i\}$ on the left-hand side can be approximated by an integral that reflects the rate at which N_i increases as δ_i decreases.

8 Definition. The covering number $N(\delta)$, or $N(\delta, d, T)$ if there is any risk of ambiguity, is the size of the smallest δ-net for T. That is, $N(\delta)$ equals the smallest m for which there exist points $t_1, \ldots, t_m$ with $\min_i d(t, t_i) \leq \delta$ for every t in T. The associated covering integral is

$$J(\delta) = J(\delta, d, T) = \int_0^\delta [2\log(N(u)^2/u)]^{1/2}\, du \quad \text{for} \quad 0 < \delta \leq 1. \qquad \square$$

The $N(u)^2$, in place of $N(u)$, will allow us to bound maxima over more than just the nearest-neighbor links from T_{i+1} to T_i.

If we interpret P as standing for the $\mathscr{L}^1(P)$ or $\mathscr{L}^2(P)$ semimetrics on $\mathscr{F}$, the notation $N_1(\delta, P, \mathscr{F})$ and $N_2(\delta, P, \mathscr{F})$ used in Chapter II almost agrees with Definition 8. Here we implicitly restrict $t_1, \ldots, t_m$ to be points of T. In Chapter II the approximating functions were allowed to lie outside $\mathscr{F}$. They could have been restricted to lie in $\mathscr{F}$ without seriously affecting any of the results.

The proof of our main result, the Chaining Lemma, will be slightly more complicated than indicated above. To achieve the most precise inequality, we replace η_i by a function of the link lengths. And we eliminate a few pesky details by being fastidious in the construction of the approximating sets T_i. But apart from that, the idea behind the proof is the same.

As you read through the argument please notice that it would also work if $N(\cdot)$ were replaced throughout by any upper bound and, of course, $J(\cdot)$ were increased accordingly. This trivial observation will turn out to be most important for applications; we seldom know the covering numbers exactly, but we often have upper bounds for them.

9 Chaining Lemma. *Let $\{Z(t): t \in T\}$ be a stochastic process whose index set has a finite covering integral $J(\cdot)$. Suppose there exists a constant D such that, for all s and t,*

(10) $\quad \mathbb{P}\{|Z(s) - Z(t)| > \eta d(s, t)\} \leq 2\exp(-\tfrac{1}{2}\eta^2/D^2) \quad$ *for* $\quad \eta > 0.$

Then there exists a countable dense subset T^* of T such that, for $0 < \varepsilon < 1$,

$$\mathbb{P}\{|Z(s) - Z(t)| > 26DJ(d(s,t)) \text{ for some } s, t \text{ in } T^* \text{ with } d(s,t) \leq \varepsilon\} \leq 2\varepsilon$$

We can replace T^* by T if Z has continuous sample paths.

Proof. Write $H(u)$ for $[2\log(N(u)^2/u)]^{1/2}$. It increases as u decreases. Set $\delta_i = \varepsilon/2^i$ for $i = 1, 2, \ldots$. Construct $2\delta_i$-nets T_i in a special way, to ensure that $T_1 \subseteq T_2 \subseteq \cdots$. (The extra 2 has little effect on the chaining argument.)

Start with any point t_1. If possible choose a t_2 with $d(t_2, t_1) > 2\delta_1$; then a t_3 with $d(t_3, t_1) > 2\delta_1$ and $d(t_3, t_2) > 2\delta_1$; and so on. After some t_m, with m no greater than $N(\delta_1)$, the process must stop: if $m > N(\delta_1)$ then some pair t_i, t_j would have to fall into one of the $N(\delta_1)$ closed balls of radius δ_1 that cover T. Take T_1 as the set $\{t_1, \ldots, t_m\}$. Every t in T lies within $2\delta_1$ of at least one point in T_1.

Next choose t_{m+1}, if possible, with $d(t_{m+1}, t_i) > 2\delta_2$ for $i \leq m$; then t_{m+2} with $d(t_{m+2}, t_i) > 2\delta_2$ for $i \leq m+1$; and so on. When that process stops we have built T_1 up to T_2, a $2\delta_2$-net of at most $N(\delta_2)$ points.

The sets $T_3, T_4, \ldots$ are constructed in similar fashion. Define T^* to be the union of all the $\{T_i\}$.

For the chaining argument sketched earlier (for finite T) we bounded the increment of Z across each link joining a point of T_{i+1} to its nearest neighbor in T_i. This time T_{i+1} contains T_i; all the links run between points of T_{i+1}. With only an insignificant increase in the probability bound we can increase the collection of links to cover all pairs in T_{i+1}, provided we replace the suggested η_i by a quantity depending on the length of the link. Set

$$A_i = \{|Z(s) - Z(t)| > Dd(s,t)H(\delta_i) \text{ for some } s, t \text{ in } T_i\}.$$

It is a union of at most $N(\delta_i)^2$ events, each of whose probabilities can be bounded using (10).

$$\mathbb{P}A_i \leq 2N(\delta_i)^2 \exp[-\tfrac{1}{2}H(\delta_i)^2] = 2\delta_i.$$

The union of all the $\{A_i\}$, call it A, has probability at most 2ε.

Consider any pair (s, t) in T^* for which $d(s, t) \le \varepsilon$. Find the n for which $\delta_n < d(s, t) \le 2\delta_n$. Because the $\{T_i\}$ expand as i increases, both s and t belong to some T_{m+1} with $m > n$. With a chain $s = s_{m+1}, s_m, \ldots, s_n$ link s to an s_n in T_n, choosing each s_i to be the closest point of T_i to s_{i+1}, thereby ensuring that $d(s_{i+1}, s_i) \le 2\delta_i$. Define a chain $\{t_i\}$ for t similarly. Break $Z(s) - Z(t)$ into $Z(s_n) - Z(t_n)$ plus sums of increments across the links of the two chains; $|Z(s) - Z(t)|$ is no greater than

$$|Z(s_n) - Z(t_n)| + \sum_{i=n}^{m} [|Z(s_{i+1}) - Z(s_i)| + |Z(t_{i+1}) - Z(t_i)|].$$

Both s_{i+1} and s_i belong to T_{i+1}. On A_{i+1}^c,

$$|Z(s_{i+1}) - Z(s_i)| \le Dd(s_{i+1}, s_i)H(\delta_{i+1}) \le 2D\delta_i H(\delta_{i+1}).$$

On A^c, these bounds, together with their companions for (s_n, t_n) and (t_{i+1}, t_i), allow $|Z(s) - Z(t)|$ to be at most

$$Dd(s_n, t_n)H(\delta_n) + 2\sum_{i=n}^{m} 2D\delta_i H(\delta_{i+1}).$$

The distance $d(s_n, t_n)$ is at most

$$d(s, t) + \sum_{i=n}^{m} d(s_{i+1}, s_i) + \sum_{i=n}^{m} d(t_{i+1}, t_i) \le 2\delta_n + 2\sum_{i=n}^{m} 2\delta_i \le 10\delta_n.$$

Also $\delta_i = 4(\delta_{i+1} - \delta_{i+2})$. Thus, on A^c,

$$\begin{aligned}
|Z(s) - Z(t)| &\le 10D\delta_n H(\delta_n) + 4D\sum_{i=n}^{m} 4(\delta_{i+1} - \delta_{i+2})H(\delta_{i+1}) \\
&\le 10D\delta_n H(\delta_n) + 16D\sum_{i=n}^{m} \int \{\delta_{i+2} < u \le \delta_{i+1}\}H(u)\,du \\
&\le 10D\delta_n H(\delta_n) + 16D\,J(\delta_{n+1}) \\
&\le 26D\,J(d(s, t)).
\end{aligned}$$

If Z has continuous sample paths, the inequality with T^* replaced by T is the limiting case of the inequalities for T^* with ε replaced by $\varepsilon + n^{-1}$. □

Often we will apply the inequality from the Chaining Lemma in the weaker form:

$$\mathbb{P}\{|Z(s) - Z(t)| > 26D\,J(\varepsilon) \text{ for some } s, t \text{ in } T^* \text{ with } d(s, t) \le \varepsilon\} \le 2\varepsilon.$$

A direct derivation of the weaker inequality would be slightly simpler than the proof of the lemma. But there are applications where the stronger result is needed.

11 Example. Brownian motion on $[0, 1]$, you will recall, is a stochastic process $\{B(\cdot, t): 0 \le t \le 1\}$ with continuous sample paths, independent increments, $B(\cdot, 0) = 0$, and $B(t) - B(s)$ distributed $N(0, t-s)$ for $t \ge s$. If we measure distances between points of $[0, 1]$ in a strange way, the Chaining Lemma will give a so-called modulus of continuity for the sample paths of B.

The normal distribution has tails that decrease exponentially fast: from Appendix B,

$$\mathbb{P}\{|B(t) - B(s)| > \eta\} \le 2\exp(-\tfrac{1}{2}\eta^2/|t-s|).$$

Define a new metric on $[0, 1]$ by setting $d(s, t) = |s - t|^{1/2}$. Then B satisfies inequality (10) with $D = 1$. The covering number $N(\delta, d, [0, 1])$ is smaller than $2\delta^{-2}$, which gives the bound

$$\begin{aligned} J(\delta) &\le \int_0^\delta [2\log 4 + 10\log(1/u)]^{1/2}\,du \\ &\le (2\log 4)^{1/2}\delta + \sqrt{10}[\log(1/\delta)]^{-1/2}\int_0^\delta \log(1/u)\,du \\ &\le 4\delta[\log(1/\delta)]^{1/2} \quad \text{for } \delta \text{ small enough.} \end{aligned}$$

From the Chaining Lemma,

$$\mathbb{P}\{|B(s) - B(t)| > 26J(d(s, t)) \text{ for some pair with } d(s, t) \le \delta\} \le 2\delta.$$

The event appearing on the left-hand side gets smaller as δ decreases. Let $\delta \downarrow 0$. Conclude that for almost all ω,

$$|B(\omega, s) - B(\omega, t)| \le 74|(s-t)\log|s-t||^{1/2}$$

for $|s - t|^{1/2} < \delta(\omega)$. Except for the unimportant factor of 74, this is the best modulus possible (McKean 1969, Section 1.6). □

VII.3. Gaussian Processes

In Section 5 we shall generalize the Empirical Central Limit Theorem of Chapter V to empirical processes indexed by classes of functions. The limit processes will be analogues of the brownian bridge, gaussian processes with sample paths continuous in an appropriate sense. Even though existence of the limits will be guaranteed by the method of proof, it is no waste of effort if we devote a few pages here to a direct construction, which makes non-trivial application of the Chaining Lemma. The direct argument tells us more about the sample path properties of the gaussian processes.

We start with analogues of brownian motion. The argument will extend an idea already touched on in Example 11.

Look at brownian motion in a different way. Regard it as a stochastic process indexed by the class of indicator functions

$$\mathscr{F} = \{[0, t] : 0 \le t \le 1\}.$$

The covariance $\mathbb{P}[B(\cdot, f)B(\cdot, g)]$ can then be written as $P(fg)$, where $P =$ Uniform[0, 1]. The process maps the subset $\mathscr{F}$ of $\mathscr{L}^2(P)$ into the space $\mathscr{L}^2(\mathbb{P})$ in such a way that inner products are preserved. From this perspective it becomes more natural to characterize the sample path property as continuity with respect to the $\mathscr{L}^2(P)$ seminorm ρ_P on $\mathscr{F}$. Notice that

$$\rho_P(|[0, s] - [0, t]|) = (P|[0, s] - [0, t]|^2)^{1/2} = |s - t|^{1/2}.$$

It is no accident that we used the same distance function in Example 11.

The new notion of sample path continuity also makes sense for stochastic processes indexed by subclasses of other $\mathscr{L}^2(P)$ spaces, for probability measures different from Uniform[0, 1].

12 Definition. Let $\mathscr{F}$ be a class of measurable functions on a set S with a σ-field supporting a probability measure P. Suppose $\mathscr{F}$ is contained in $\mathscr{L}^2(P)$. A P-motion is a stochastic process $\{B_P(\cdot, f): f \in \mathscr{F}\}$ indexed by $\mathscr{F}$ for which:

(i) B_P has joint normal finite-dimensional distributions with zero means and covariance $\mathbb{P}[B_P(\cdot, f)B_P(\cdot, g)] = P(fg)$;
(ii) each sample path $B_P(\omega, \cdot)$ is bounded and uniformly continuous with respect to the $\mathscr{L}^2(P)$ seminorm $\rho_P(\cdot)$ on $\mathscr{F}$.

The name does not quite fit unless one reads "Uniform[0, 1]" as "brownian," but it is easy to remember. The uniform continuity and boundedness that crept into the definition come automatically for brownian motion on the compact interval [0, 1]. In general $\mathscr{F}$ need not be a compact subset of $\mathscr{L}^2(P)$, although it must be totally bounded if it is to index a P-motion (Problem 3); uniformly continuous functions on a totally bounded $\mathscr{F}$ must be bounded.

We seek conditions on P and $\mathscr{F}$ for existence of the P-motion. The Chaining Lemma will give us much more: a bound on the increments of the process in terms of the covering integral

$$J(\delta) = J(\delta, \rho_P, \mathscr{F}) = \int_0^\delta [2 \log(N(u, \rho_P, \mathscr{F})^2/u)]^{1/2}\, du.$$

Finiteness of $J(\cdot)$ will guarantee existence of B_P.

13 Theorem. *Let $\mathscr{F}$ be a subset of $\mathscr{L}^2(P)$ with a finite covering integral, $J(\cdot)$, under the $\mathscr{L}^2(P)$ seminorm $\rho_P(\cdot)$. There exists a P-motion, B_P, indexed by $\mathscr{F}$, for which*

$$|B_P(\omega, f) - B_P(\omega, g)| \le 26 J(\rho_P(f - g)) \quad \textit{if} \quad \rho_P(f - g) < \delta(\omega),$$

with $\delta(\omega)$ finite for every ω.

PROOF. Construct the process first on a countable dense subset $\mathscr{F}_0 = \{f_j\}$ of $\mathscr{F}$. Such a subset exists because $\mathscr{F}$ has a finite δ-net for each $\delta > 0$ (otherwise J could not be finite). Apply the Gram–Schmidt procedure to $\mathscr{F}_0$, generating an orthonormal sequence of functions $\{u_j\}$. Each f in $\mathscr{F}_0$ is a finite linear combination $\sum_j \langle u_j, f\rangle u_j$ because $\{u_1, \ldots, u_n\}$ spans the same subspace as $\{f_1, \ldots, f_n\}$. Here, temporarily, $\langle u, f\rangle$ denotes the inner product in $\mathscr{L}^2(P)$: $\langle u, f\rangle = P(uf)$. Choose a probability space $(\Omega, \mathscr{E}, \mathbb{P})$ supporting a sequence $\{U_j\}$ of independent $N(0, 1)$ random variables. For each f in $\mathscr{F}_0$ and ω in Ω define

$$Z(\omega, f) = \sum_j \langle u_j, f\rangle U_j(\omega).$$

The sum converges for every ω, because only finitely many of the coefficients $\langle u_j, f\rangle$ are non-zero. The finite-dimensional distributions of Z are joint normal with zero means and the desired covariances:

$$\begin{aligned}\mathbb{P}[Z(\cdot, f)Z(\cdot, g)] &= \sum_{i,j} \langle u_i, f\rangle\langle u_j, g\rangle \mathbb{P}(U_i U_j)\\ &= \sum_i \langle u_i, f\rangle\langle u_i, g\rangle\\ &= \langle f, g\rangle\end{aligned}$$

as required for a P-motion.

The $\mathscr{L}^2(P)$ seminorm is tailor-made for the chaining argument. Because $\mathbb{P}\{|N(0, 1)| \geq x\} \leq 2\exp(-\frac{1}{2}x^2)$ for $x \geq 0$ (Appendix B),

$$\begin{aligned}\mathbb{P}\{|Z(f) - Z(g)| \geq \eta\} &\leq 2\exp(-\tfrac{1}{2}\eta^2/\mathbb{P}[Z(f) - Z(g)]^2)\\ &= 2\exp(-\tfrac{1}{2}\eta^2/\rho_P(f - g)^2).\end{aligned}$$

Apply the Chaining Lemma to the process Z on $\mathscr{F}_0$. Because $\mathscr{F}_0$ itself is countable we may as well assume the countable dense subset promised by the Lemma coincides with $\mathscr{F}_0$. Let $G(\delta)$ denote the set of ω for which

$$|Z(f) - Z(g)| > 26J(\rho_P(f - g)) \quad \text{for some pair with} \quad \rho_P(f - g) \leq \delta.$$

Then $\mathbb{P}G(\delta) \leq 2\delta$ for every $\delta > 0$. As δ decreases, $G(\delta)$ contracts to a negligible set $G(0)$. For each ω not in $G(0)$,

$$|Z(\omega, f) - Z(\omega, g)| \leq 26J(\rho_P(f - g)) \quad \text{if} \quad \rho_P(f - g) < \delta(\omega).$$

Reduce Ω to $\Omega\backslash G(0)$. Then each sample path $Z(\omega, \cdot)$ is uniformly continuous. Extend it from the dense $\mathscr{F}_0$ up to a uniformly continuous function on the whole of $\mathscr{F}$. The extension preserves the bound on the increments, because both J and ρ_P are continuous. Complete the proof by checking that the resulting process has the finite dimensional distributions of a P-motion. □

For brownian motion, continuity of sample paths in the usual sense coincides with continuity in the ρ_P sense, with $P = \text{Uniform}[0, 1]$. The P-motion processes for different P measures on $[0, 1]$ (or on $\mathbb{R}$, or on $\mathbb{R}^k$)

do not necessarily have the same property. If P has an atom of mass α at a point t_0, the sample paths of the B_P indexed by intervals $\{[0, t]\}$ will all have a jump at t_0. The size of the jump will be $N(0, \alpha)$ distributed independently of all increments that don't involve a pair of intervals bracketing t_0. All sample paths are cadlag in the usual sense.

We encountered similar behavior in the gaussian limit processes for the Empirical Central Limit Theorem (V.11) on the real line. We represented the limit as $U(F(\cdot))$, with U a brownian bridge and F the distribution function for the sampling measure P. We can also manufacture the limit process directly from the P-motion, in much the same way that we get a brownian bridge from brownian motion. Denote by 1 the function taking the constant value one. Then the process obtained from B_P by setting

$$E_P(\cdot, f) = B_P(\cdot, f) - (Pf)B_P(\cdot, 1),$$

is a gaussian process analogous to the brownian bridge.

14 Definition. Call a stochastic process E_P indexed by a subclass $\mathscr{F}$ of $\mathscr{L}^2(P)$ a P-bridge over $\mathscr{F}$ if

(i) E_P has joint normal finite-dimensional distributions with zero means and covariance $\mathbb{P}[E_P(\cdot, f)E_P(\cdot, g)] = P(fg) - (Pf)(Pg)$;
(ii) each sample path $E_P(\omega, \cdot)$ is bounded and uniformly continuous with respect to the $\mathscr{L}^2(P)$ seminorm on $\mathscr{F}$. □

The P-bridge will return in Section 5 as the limit in a central limit theorem for empirical processes indexed by a class of functions.

VII.4. Random Covering Numbers

The two methods developed in Chapter II, for proving uniform strong laws of large numbers, can be adapted to the task of proving the maximal inequalities that lurk behind the stochastic equicontinuity conditions introduced in Section 1. The second method, the one based on symmetrization of the empirical measure, lends itself more readily to the new purpose because it is the easier to upgrade by means of a chaining argument. We have the tools for controlling the rate at which covering numbers grow; we have a clean exponential bound for the conditional distribution of the increments of the symmetrized process. The introduction of chaining into the first method is complicated by a messier exponential bound. Section 6 will tackle that problem.

Recall that the symmetrization method relates $P_n - P$ to the random signed measure P_n° that puts mass $\pm n^{-1}$ at each of $\xi_1, \ldots, \xi_n$, the signs being allocated independently plus or minus, each with probability $\frac{1}{2}$. For central limit theorem calculations it is neater to work with the symmetrized empirical process $E_n^\circ = n^{1/2}P_n^\circ$. Hoeffding's Inequality (Appendix B) gives the clean exponential bound for E_n° conditional on everything but the random signs. For each fixed function f,

$$\begin{aligned}\mathbb{P}\{|E_n^\circ f| > \eta \,|\, \boldsymbol{\xi}\} &= \mathbb{P}\left\{\left|\sum_{i=1}^n \pm f(\xi_i)\right| > \eta n^{1/2} \,|\, \boldsymbol{\xi}\right\} \\ &\le 2\exp\left[-2(\eta n^{1/2})^2 \Big/ \sum_{i=1}^n 4f(\xi_i)^2\right] \\ &= 2\exp[-\tfrac{1}{2}\eta^2/P_n f^2].\end{aligned}$$

That is, if distances between functions are measured using the $\mathscr{L}^2(P_n)$ seminorm then tail probabilities of E_n° under $\mathbb{P}(\cdot|\boldsymbol{\xi})$ satisfy the exponential bound required by the Chaining Lemma, with $D = 1$. For the purposes of the chaining argument, E_n° will behave very much like the gaussian process B_P of Section 3, except that the bound involves the random covering number calculated using the $\mathscr{L}^2(P_n)$ seminorm. Write

$$J_2(\delta, P_n, \mathscr{F}) = \int_0^\delta [2\log(N_2(u, P_n, \mathscr{F})^2/u)]^{1/2}\,du$$

for the corresponding covering integral.

Stochastic equicontinuity of the empirical processes $\{E_n\}$ at a function f_0 in $\mathscr{F}$ means roughly that, with high probability and for all n large enough, $|E_n f - E_n f_0|$ should be uniformly small for all f close enough to f_0. Here closeness should be measured by the $\mathscr{L}^2(P)$ seminorm ρ_P. With the Chaining Lemma in hand we can just as easily check for what seems a stronger property —but if you look carefully you'll see that it's equivalent to stochastic equicontinuity for a larger class of functions. Of course we need $\mathscr{F}$ to be permissible (Appendix C).

15 Equicontinuity Lemma. *Let $\mathscr{F}$ be a permissible class of functions with envelope F in $\mathscr{L}^2(P)$. Suppose the random covering numbers satisfy the uniformity condition: for each $\eta > 0$ and $\varepsilon > 0$ there exists a $\gamma > 0$ such that*

$$\limsup_n \mathbb{P}\{J_2(\gamma, P_n, \mathscr{F}) > \eta\} < \varepsilon. \tag{16}$$

Then there exists a $\delta > 0$ for which

$$\limsup \mathbb{P}\left\{\sup_{[\delta]} |E_n(f - g)| > \eta\right\} < \varepsilon,$$

where $[\delta] = \{(f, g): f, g \in \mathscr{F}$ and $\rho_P(f - g) \le \delta\}$.

Proof. The idea will be: replace E_n by the symmetrized process E_n°; replace $[\delta]$ by a random analogue,

$$\langle 2\delta\rangle = \{(f, g): f, g \in \mathscr{F} \text{ and } (P_n(f - g)^2)^{1/2} < 2\delta\};$$

then apply the Chaining Lemma for the conditional distributions $\mathbb{P}(\cdot|\boldsymbol{\xi})$.

For fixed f and g in $[\delta]$ we have $\operatorname{var}(E_n(f-g)) = P(f-g)^2 \leq \delta^2$. Argue as in the FIRST and SECOND SYMMETRIZATION steps of Section II.3: when δ is small enough,

$$\mathbb{P}\left\{\sup_{[\delta]} |E_n(f-g)| > \eta\right\} \leq 4\mathbb{P}\left\{\sup_{[\delta]} |E_n^\circ(f-g)| > \tfrac{1}{4}\eta\right\}.$$

That gets rid of E_n.

If with probability tending to one the class $\langle 2\delta \rangle$ contains $[\delta]$, we will waste only a tiny bit of probability in replacing $[\delta]$ by $\langle 2\delta \rangle$:

$$\mathbb{P}\left\{\sup_{[\delta]} |E_n^\circ(f-g)| > \tfrac{1}{4}\eta\right\} \leq \mathbb{P}\left\{\sup_{\langle 2\delta \rangle} |E_n^\circ(f-g)| > \tfrac{1}{4}\eta\right\} + \mathbb{P}\{[\delta] \not\subseteq \langle 2\delta \rangle\}.$$

It would suffice if we showed $\sup_{\mathscr{F}_2} |P_n h - Ph| \to 0$ almost surely, where $\mathscr{F}_2 = \{(f-g)^2 : f, g \in \mathscr{F}\}$. This follows from Theorem II.24 because the condition (16) implies

$$\text{(17)} \qquad \log N_1(\delta, P_n, \mathscr{F}_2) = o_p(n) \quad \text{for each} \quad \delta > 0.$$

Problems 5 and 6 provides the details behind (17). That gets rid of $[\delta]$.

The reason we needed to replace $[\delta]$ by $\langle 2\delta \rangle$ becomes evident when we condition on ξ. Write $\rho_n(\cdot)$ for the $\mathscr{L}^2(P_n)$ seminorm. We have no direct control over $\rho_n(f-g)$ for functions in $[\delta]$; but for $\langle 2\delta \rangle$, whose members are determined as soon as ξ is specified, $\rho_n(f-g) < 2\delta$. Apply the Chaining Lemma.

$$\mathbb{P}\{|E_n^\circ(f-g)| > 26J_2(2\delta, P_n, \mathscr{F}) \text{ for some } (f, g) \text{ in } \langle 2\delta \rangle^* | \xi\} \leq 4\delta.$$

The countable dense subclass $\langle 2\delta \rangle^*$ can be replaced by $\langle 2\delta \rangle$ itself, because E_n° is a continuous function on $\mathscr{F}$ for each fixed ξ:

$$|E_n^\circ(f-g)| \leq n^{1/2} P_n |f-g| \leq n^{1/2} \rho_n(f-g).$$

Integrate out over ξ, then choose δ so that both $\mathbb{P}\{26J_2(2\delta, P_n, \mathscr{F}) > \tfrac{1}{4}\eta\}$ and 4δ are small. □

Now that we have the maximal inequalities for empirical processes, we can take up again the central limit theorems for statistics defined by minimization of a random process, the topic we left hanging at the end of Section 1.

Recall that we need the processes $\{E_n r(\cdot, t)\}$, which is indexed by the class $\mathscr{R} = \{r(\cdot, t) : t \in T\}$ of remainder functions, stochastically equicontinuous at t_0. If $f(\cdot, t)$ is differentiable in quadratic mean at t_0, it will suffice if we find a neighborhood $V = \{r \in \mathscr{R} : \rho_P(r) \leq \delta\}$ for which

$$\limsup \mathbb{P}\left\{\sup_V |E_n r| > \eta\right\} < \varepsilon.$$

Notice that $V \subseteq \{r_1 - r_2 : \rho_P(r_1 - r_2) \leq \delta\}$, because $r(\cdot, t_0) = 0$ by definition. Thus we may check for stochastic equicontinuity by showing:

(i) The class $\mathscr{R}$ has an envelope belonging to $\mathscr{L}^2(P)$.

(ii) $f(\cdot, t)$ is differentiable in quadratic mean at t_0. From (i), this follows by dominated convergence if $r(\cdot, t) \to 0$ almost surely $[P]$ as $t \to t_0$.
(iii) Condition (16) is satisfied for $\mathscr{F} = \mathscr{R}$.

These three conditions place constraints on the class $\{f(\cdot, t)\}$.

18 Example. The spatial median of a bivariate distribution P is the value of θ that minimizes $M(\theta) = P|x - \theta|$. Estimate it by the θ_n that minimizes $M_n(\theta) = P_n|x - \theta|$. Example II.26 gave conditions for consistency of such an estimator. Those conditions apply when P equals the symmetric normal $N(0, I_2)$, a pleasant distribution to work with because explicit values can be calculated for all the quantities connected with the asymptotics for $\{\theta_n\}$. For this P, convexity and symmetry force $M(\cdot)$ to have its unique minimum at zero, so θ_n converges almost surely to zero. Theorem 5 will produce the central limit theorem,

$$n^{1/2}\theta_n \rightsquigarrow N(0, (4/\pi)I_2),$$

after we check its non-obvious conditions (ii), (iv), and (v).

Change variables to reexpress $M(\theta)$ in a form that makes it easier to find derivatives.

$$M(\theta) = (2\pi)^{-1} \int |x| \exp(-\tfrac{1}{2}|x + \theta|^2)\, dx.$$

Differentiate under the integral sign.

$$M'(0) = 0, \quad \text{of course,}$$

$$M''(0) = (2\pi)^{-1} \int |x|(xx' - I_2) \exp(-\tfrac{1}{2}|x|^2)\, dx.$$

A random vector X with a $N(0, I_2)$ distribution has the factorization $X = RU$ where $R^2 = |X|^2$ has a χ_2^2-distribution independent of the random unit vector $U = X/|X|$, which is uniformly distributed around the rim of the unit circle.

$$\begin{aligned} V = M''(0) &= \mathbb{P}(R^3UU' - RI_2) \\ &= \mathbb{P}R^3\,\mathbb{P}UU' - (\mathbb{P}R)I_2 \\ &= (\pi/8)^{1/2}I_2. \end{aligned}$$

Condition (ii) wasn't so hard to check.

To figure out the $\Delta(x)$ that should appear in the linear approximation

$$|x - \theta| = |x| + \theta'\Delta(x) + |\theta| r(x, \theta),$$

carry out the usual pointwise differentiation. That gives $\Delta(x) = x/|x|$ for $x \neq 0$. Set $\Delta(0) = 0$, for completeness. The components of $\Delta(\cdot)$ all belong to $\mathscr{L}^2(P)$. Indeed, $P\Delta\Delta' = \mathbb{P}UU' = \tfrac{1}{2}I_2$. That's condition (iv) taken care of.

Now comes the hard part—or at least it would be hard if we hadn't already proved the Equicontinuity Lemma. Start by checking that the class $\mathscr{R}$ of remainder functions $r(\cdot, \theta)$ has an envelope in $\mathscr{L}^2(P)$. For $\theta \neq 0$,

$$\begin{aligned} |r(x, \theta)| &= \big| |x - \theta| - |x| - \theta'\Delta(x) \big| / |\theta| \\ &\leq |\theta|^{-1}(|x - \theta|^2 - |x|^2)/(|x - \theta| + |x|) + 1 \\ &\leq (2|x| + |\theta|)/(|x - \theta| + |x|) + 1 \\ &\leq 4. \end{aligned}$$

It follows that $|\cdot - \theta|$ is differentiable in quadratic mean at $\theta = 0$. We have only to verify condition (16) of the Equicontinuity Lemma to complete the proof of stochastic equicontinuity.

Each $r(\cdot, \theta)$, for $\theta \neq 0$, can be broken into a difference of two bounded functions:

$$\begin{aligned} r_1(\cdot, \theta) &= \theta'\Delta(\cdot)/|\theta|, \\ r_2(\cdot, \theta) &= (|x - \theta| - |x|)/|\theta|. \end{aligned}$$

Write $\mathscr{R}_1$ and $\mathscr{R}_2$ for the corresponding classes of functions.

The linear space spanned by $\mathscr{R}_1$ has finite dimension; the graphs have polynomial discrimination, by Lemma II.28; the covering numbers $N_2(u, P_n, \mathscr{R}_1)$ are bounded by a polynomial Au^{-W} in u^{-1}, with A and W not depending on P_n (Lemma II.36).

The graphs of functions in $\mathscr{R}_2$ also have polynomial discrimination, because $\{(x, t): |x - \theta| - |x| \geq |\theta| t\}$ can be written as

$$\{-2\theta'x + |\theta|^2 \geq 2|\theta||x|t + |\theta|^2 t^2\} \cap \{|x| + |\theta|t \geq 0\} \cup \{|x| + |\theta|t < 0\}.$$

This is built up from sets of the form $\{g \geq 0\}$ with g in the finite-dimensional vector space of functions

$$g_{\alpha, \beta, \gamma, \delta, \varepsilon, \zeta}(x, t) = \alpha'x + \beta|x| + \gamma|x|t + \delta t + \varepsilon t^2 + \zeta.$$

The covering numbers for $\mathscr{R}_2$ are also uniformly bounded by a polynomial in u^{-1}.

These two polynomial bounds combine (Problem II.18) to give a similar uniform bound for the covering numbers of $\mathscr{R}$, which amply suffices for the Equicontinuity Lemma: for each $\eta > 0$ there exists a γ such that $J_2(\gamma, P_n, \mathscr{R}) \leq \eta$ for every P_n. The conditions of Theorem 5 are all satisfied; the central limit theorem for $\{\theta_n\}$ is established. □

19 Example. Independent observations are sampled from a distribution P on the real line. The optimal 2-means cluster centers a_n, b_n minimize $W(a, b, P_n) = P_n f_{a,b}$, where $f_{a,b}(x) = |x - a|^2 \wedge |x - b|^2$. In Examples II.4 and II.29 we found conditions under which a_n, b_n converge almost surely to the centers a^*, b^* that minimize $W(a, b, P) = Pf_{a,b}$. Theorem 5 refines the result to a central limit theorem.

Keep the calculations simple by taking P as the Uniform[0, 1] distribution. The argument could be extended to other P distributions, higher dimensions, and more clusters, at the cost of more cumbersome notation and the imposition of a few extra regularity conditions.

The parameter set consists of all pairs (a, b) with $0 \le a \le b \le 1$. For the Uniform[0, 1] distribution direct calculation gives explicitly the values a^*, b^* that minimize $W(a, b, P)$.

$$\begin{aligned} W(a, b, P) &= \int \{0 \le x < \tfrac{1}{2}(a + b)\} |x - a|^2 \\ &\qquad + \{\tfrac{1}{2}(a + b) \le x \le 1\} |x - b|^2 \, dx \\ &= \tfrac{1}{3}a^3 + \tfrac{1}{3}(1 - b)^3 + \tfrac{1}{12}(b - a)^3. \end{aligned}$$

Minimizing values: $a^* = \frac{1}{4}$, $b^* = \frac{3}{4}$, as you might expect. Near these optimal centers,

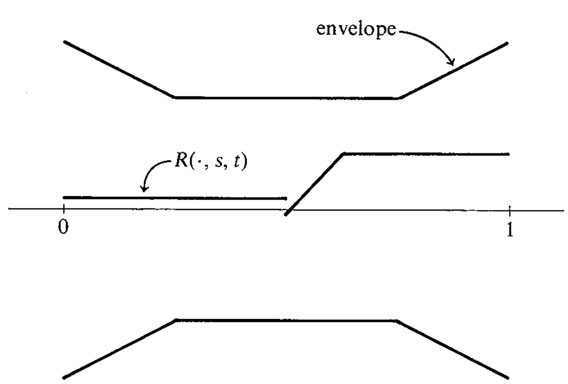
$$\begin{aligned} W(a, b, P) = \tfrac{1}{48} + \tfrac{3}{8}(a - \tfrac{1}{4})^2 - \tfrac{1}{4}(a - \tfrac{1}{4})(b - \tfrac{3}{4}) + \tfrac{3}{8}(b - \tfrac{3}{4})^2 \\ + \text{cubic terms} \end{aligned}$$

The function $f_{a,b}(x)$ has partial derivatives with respect to a and b except when $x = \frac{1}{2}(a + b)$. That suggests for $\Delta(x)$ the two components

$$\begin{aligned} \Delta_a(x) &= -2(x - \tfrac{1}{4})\{0 \le x < \tfrac{1}{2}\}, \\ \Delta_b(x) &= -2(x - \tfrac{3}{4})\{\tfrac{1}{2} \le x \le 1\}. \end{aligned}$$

Both functions belong to $\mathscr{L}^2(P)$. The remainder function is defined by subtraction of the linear approximation from $f_{a,b}$. Simplify the notation by writing $s = a - \frac{1}{4}$, $t = b - \frac{3}{4}$; change $f_{a,b}$ to $g_{s,t}$ and $r(\cdot, a, b)$ to $R(\cdot, s, t)$.

$$\begin{aligned} (|s| + |t|)R(x, s, t) &= g_{s,t}(x) - g_{0,0}(x) + 2s(x - \tfrac{1}{4})\{0 \le x < \tfrac{1}{2}\} \\ &\qquad + 2t(x - \tfrac{3}{4})\{\tfrac{1}{2} \le x \le 1\} \\ &= \text{a piecewise linear function of } x. \end{aligned}$$



The remainder functions are bounded by a fixed envelope in $\mathscr{L}^2(P)$.

$$\begin{aligned}|R(x, s, t)| &\le [|(x - \tfrac{1}{4} - s)^2 - (x - \tfrac{1}{4})^2| + |(x - \tfrac{3}{4} - t)^2 - (x - \tfrac{3}{4})^2| \\ &\quad + 2|s||x - \tfrac{1}{4}| + 2|t||x - \tfrac{3}{4}|]/(|s| + |t|) \\ &\le 4|x - \tfrac{1}{4}| + 4|x - \tfrac{3}{4}| + 2.\end{aligned}$$

Deduce differentiability in quadratic mean of $f_{a,b}$ at the optimal centers.

The graphs of the piecewise linear functions in $\mathscr{R}$ have only polynomial discrimination, and they have an envelope in $\mathscr{L}^2(P)$. Lemma II.36 gives a uniform bound on the covering numbers that ensures $J_2(\gamma, P_n, \mathscr{R}) \le \eta$ for every P_n if γ is chosen small enough. The Equicontinuity Lemma applies; the processes $\{E_n r(\cdot, a, b)\}$ are stochastically equicontinuous at $(\frac{1}{4}, \frac{3}{4})$; the optimal centers obey a central limit theorem

$$(n^{1/2}(a_n - \tfrac{1}{4}), n^{1/2}(b_n - \tfrac{3}{4})) \rightsquigarrow N(0, V^{-1}P(\Delta\Delta')V^{-1}),$$

where

$$V = \begin{bmatrix} \frac{3}{4} & -\frac{1}{4} \\ -\frac{1}{4} & \frac{3}{4} \end{bmatrix}, \qquad P(\Delta\Delta') = \begin{bmatrix} \frac{1}{24} & 0 \\ 0 & \frac{1}{24} \end{bmatrix}. \qquad \square$$

VII.5. Empirical Central Limit Theorems

As random elements of $D[0, 1]$, the uniform empirical processes $\{U_n\}$ converge in distribution to a brownian bridge. More generally, the empirical processes $\{E_n\}$ for observations from an arbitrary distribution on the real line converge in distribution, as random elements of $D[-\infty, \infty]$, to a gaussian process obtained by stretching out the brownian bridge. Both results treat the empirical measure as a process indexed by intervals of the real line. In this section we shall generalize these results to empirical measures indexed by classes of functions.

Convergence in distribution, as we have defined it, deals with random elements of metric spaces. Once we leave the safety of intervals on the real line it becomes quite a problem to decide what metric space of functions empirical process sample paths should belong to. Without the natural ordering of the intervals, it is difficult to find a completely satisfactory substitute for the cadlag property; without the simplification of cadlag sample paths, empirical processes run straight into the measure-theoretic complications we have so carefully been avoiding. Appendix C describes one way of overcoming these complications. A class of functions satisfying the regularity conditions described there is said to be permissible. Most classes that arise from specific applications are permissible.

Let $\mathscr{F}$ be a pointwise bounded, permissible class of functions for which $\sup_{\mathscr{F}} |Pf| < \infty$. The empirical processes $\{E_n\}$ define bounded functions on $\mathscr{F}$; their sample paths belong to the space $\mathscr{X}$ of all bounded, real functions on $\mathscr{F}$. To avoid some of the confusion that might be caused by the hierarchy of functions on spaces of functions on spaces of functions, call members of $\mathscr{X}$ functionals. Equip $\mathscr{X}$ with the metric generated by the uniform norm, $\|x\| = \sup_{\mathscr{F}} |x(f)|$. Be careful not to confuse the norm $\|\cdot\|$ on $\mathscr{X}$ with the $\mathscr{L}^2(P)$ seminorm $\rho_P(\cdot)$ on $\mathscr{F}$.

The choice of σ-field for $\mathscr{X}$ is tied up with the measurability problems handled in Appendix C. We need it small enough to make E_n a measurable random element of $\mathscr{X}$, but large enough to support a rich supply of measurable, continuous functions. The limit distributions must concentrate on sets of completely regular points (Section IV.2). That suggests that the σ-field should at least contain the balls centered at the functionals that are uniformly continuous for the ρ_P seminorm.

20 Definition. Write $C(\mathscr{F}, P)$ for the set of all functionals $x(\cdot)$ in $\mathscr{X}$ that are uniformly continuous with respect to the $\mathscr{L}^2(P)$ seminorm on $\mathscr{F}$. That is, to each $\varepsilon > 0$ there should exist a $\delta > 0$ for which $|x(f) - x(g)| < \varepsilon$ whenever $\rho_P(f - g) < \delta$. Define $\mathscr{B}^P$ as the smallest σ-field on $\mathscr{X}$ that: (i) contains all the closed balls with centers in $C(\mathscr{F}, P)$; (ii) makes all the finite-dimensional projections measurable. □

Notice that $C(\mathscr{F}, P)$ is complete, because it is a closed subset of the complete metric space $(\mathscr{X}, \|\cdot\|)$. Notice also that $\mathscr{B}^P$ depends on the sampling distribution P. Each E_n is a $\mathscr{B}^P$-measurable random element of $\mathscr{X}$ under mild regularity conditions (Appendix C).

The finite-dimensional projections of $\{E_n\}$ (the fidis) converge in distribution to the fidis of E_P, the P-bridge process over $\mathscr{F}$ (Definition 14). Of course some doubts arise over the existence of E_P; getting a version with sample paths in $C(\mathscr{F}, P)$ is no simple matter, as we saw in Section 3. Happily, the questions of existence and convergence are both taken care of by a single property of the empirical processes, uniform tightness.

Recall from Section IV.5 that uniform tightness for $\{E_n\}$ requires existence of a compact set K_ε of completely regular points in $\mathscr{X}$ such that

$$\liminf \mathbb{P}\{E_n \in G\} > 1 - \varepsilon$$

for every open, $\mathscr{B}^P$-measurable set G containing K_ε. From uniform tightness we would get a subsequence of $\{E_n\}$ that converged in distribution to a tight borel measure on $\mathscr{X}$. If $C(\mathscr{F}, P)$ contained each K_ε, the limit would concentrate in $C(\mathscr{F}, P)$. Its fidis would identify it (Problem 8) as the P-bridge over $\mathscr{F}$.

Uniform tightness of $\{E_n\}$ would also imply convergence of the whole sequence to E_P. For if $\{\mathbb{P}h(E_n)\}$ did not converge to $\mathbb{P}h(E_P)$ for some bounded, continuous, $\mathscr{B}^P$-measurable h on $\mathscr{X}$ then

$$|\mathbb{P}h(E_n) - \mathbb{P}h(E_P)| > \varepsilon \quad \text{infinitely often}$$

for some $\varepsilon > 0$. The subsequence along which the inequality held would also be uniformly tight; it would have a sub-subsequence converging to a process whose fidis still identified it as a P-bridge. That would give a contradiction: along the sub-subsequence, $\{\mathbb{P}h(E_n)\}$ would converge to $\mathbb{P}h(E_P)$ without ever getting closer than ε.

21 Theorem. *Let $\mathscr{F}$ be a pointwise bounded, totally bounded, permissible subset of $\mathscr{L}^2(P)$. If for each $\eta > 0$ and $\varepsilon > 0$ there exists a $\delta > 0$ for which*

$$\limsup \mathbb{P}\left\{\sup_{[\delta]} |E_n(f - g)| > \eta\right\} < \varepsilon, \tag{22}$$

where $[\delta] = \{(f, g): f, g \in \mathscr{F} \text{ and } \rho_P(f - g) < \delta\}$, then $E_n \rightsquigarrow E_P$ as random elements of $\mathscr{X}$. The limit P-bridge process E_P is a tight, gaussian random element of $\mathscr{X}$ whose sample paths all belong to $C(\mathscr{F}, P)$.

Proof. Check the uniform tightness. Given $\varepsilon > 0$ find a compact subset K of $C(\mathscr{F}, P)$ with $\liminf \mathbb{P}\{E_n \in G\} > 1 - \varepsilon$ for every open, $\mathscr{B}^P$-measurable G containing K. Construct K as an intersection of sets $D_1, D_2, \ldots$, where D_k is a finite union of closed balls of radius k^{-1} centered at points of $C(\mathscr{F}, P)$. Every functional in D_k will lie uniformly within k^{-1} of a member of $C(\mathscr{F}, P)$; every functional in K will therefore belong to $C(\mathscr{F}, P)$, being a uniform limit of functionals in $C(\mathscr{F}, P)$. The proof follows closely the ideas used in Theorem V.16 to prove existence of the brownian bridge. Only the continuous interpolation between values taken at a finite grid requires modification.

Fix $\eta > 0$ and $\varepsilon > 0$ for the moment, and choose δ according to (22). Invoke the total boundedness assumption on $\mathscr{F}$ to find a finite subclass $\mathscr{F}_\delta = \{f_1, \ldots, f_m\}$ of $\mathscr{F}$ such that each f in $\mathscr{F}$ has an f^* in $\mathscr{F}_\delta$ for which $\rho_P(f - f^*) < \frac{1}{2}\delta$.

We need to find a finite collection of closed balls in $\mathscr{X}$, each with a specified small radius and centered on a functional in $C(\mathscr{F}, P)$, such that E_n lies with specified large probability in the union of the balls. Construct the centers for the balls by a continuous interpolation between the values taken on at each f_i in $\mathscr{F}_\delta$ by realizations of an E_n.

For each f_i, the sequence of random variables $\{E_n(f_i)\}$ converges in distribution. There exists a constant C for which

$$\limsup \mathbb{P}\left\{\max_i |E_n(f_i)| > C\right\} < \varepsilon.$$

Define

$$\Omega_n = \left\{\omega: \sup_{[\delta]} |E_n(\omega, f - g)| \leq \eta \text{ and } \max_i |E_n(\omega, f_i)| \leq C\right\}.$$

By the choice of δ and C we ensure that $\liminf \mathbb{P}\Omega_n > 1 - 2\varepsilon$. Write S_n for the bounded set of all points in $\mathbb{R}^m$ with coordinates $E_n(\omega, f_i)$ for an ω in Ω_n, and S for the union of the $\{S_n\}$. If $\mathbf{r}$ belongs to S then $|r_i| \leq C$ for each coordinate and $|r_i - r_j| \leq 2\eta$ whenever there exists an f in $\mathscr{F}$ with $\rho_P(f - f_i) < \delta$ and $\rho_P(f - f_j) < \delta$.

Construct weight functions $\Delta_i(\cdot)$ on $\mathscr{F}$ by

$$v_i(f) = [1 - \rho_P(f - f_i)/\delta]^+,$$

$$\Delta_i(f) = v_i(f)/[v_1(f) + \cdots + v_m(f)].$$

Each $v_i(\cdot)$ is a uniformly continuous function (under the ρ_P seminorm) that vanishes outside a ball of radius δ about f_i. For every f there is at least one f_i, its f^*, for which $v_i(f) > \frac{1}{2}$; the denominator in the defining quotient for Δ_i is never less than $\frac{1}{2}$. The $\Delta_i(\cdot)$ are non-negative, uniformly continuous functions that sum to one everywhere in $\mathscr{F}$.

For $\mathbf{r}$ in S define an interpolation function by $x(f, \mathbf{r}) = \sum_{i=1}^m \Delta_i(f) r_i$. Each $x(\cdot, \mathbf{r})$ belongs to $C(\mathscr{F}, P)$. If $\rho_P(f - f_j) < \delta$, all the r_i values corresponding to non-zero $\{\Delta_i(f)\}$ satisfy $|r_i - r_j| \leq 2\eta$. As a convex combination of these values, $x(f, \mathbf{r})$ must also lie within 2η of r_j. An $E_n(\omega, \cdot)$ corresponding to an ω in Ω_n has a similar property:

$$|E_n(\omega, f) - E_n(\omega, f_j)| \leq \eta \quad \text{when} \quad \rho_P(f - f_j) < \delta.$$

Thus, if ω belongs to Ω_n and $|E_n(\omega, f_j) - r_j| \leq \eta$ for every j then

$$\sup_{\mathscr{F}} |E_n(\omega, f) - x(f, \mathbf{r})| \leq 4\eta.$$

Choose from the bounded set S a finite subset $\{\mathbf{r}(1), \ldots, \mathbf{r}(p)\}$ for which

$$\min_k \max_j |r_j - r_j(k)| \leq \eta \quad \text{for every } \mathbf{r} \text{ in } S.$$

Abbreviate $x(\cdot, \mathbf{r}(k))$ to $x_k(\cdot)$, for $k = 1, \ldots, p$. Then from what we have just proved

$$\min_k \sup_{\mathscr{F}} |E_n(\omega, f) - x_k(f)| \leq 4\eta,$$

whenever ω belongs to Ω_n. If we set D equal to the union of the balls $B(x_1, 4\eta), \ldots, B(x_p, 4\eta)$, then

$$\liminf \mathbb{P}\{E_n \in D\} > 1 - 2\varepsilon.$$

Repeat the argument with η replaced by $(4k)^{-1}$ and ε replaced by $\varepsilon/2^{k+1}$, for $k = 1, 2, \ldots$, to get each of the D_k sets promised at the start of the proof: D_k is a finite union of closed balls of radius k^{-1} and

$$\liminf \mathbb{P}\{E_n \in D_k\} > 1 - \varepsilon/2^k.$$

The remainder of the proof follows Theorem V.16 almost exactly.

The intersection of the sets $D_1, D_2, \ldots$ is a closed and totally bounded subset of the complete metric space $C(\mathscr{F}, P)$; it defines the sought-after compact K. The open G contains some finite intersection $D_1 \cap \cdots \cap D_k$. If not, there would exist a sequence $Y = \{y_k\}$ with y_k in $G^c \cap D_1 \cap \cdots \cap D_k$ for each k. Some subsequence Y' of Y would lie within one of the balls making up D_1; some subsequence Y'' of Y' would lie within one of the balls making up D_2; and so on. The sequence constructed by taking the first member of Y', the second member of Y'', and so on, would be Cauchy; it would converge to a point y ($\mathscr{X}$ is complete) belonging to all the closed sets $\{G^c \cap D_1 \cap \cdots \cap D_k\}$. This would contradict

$$\bigcap_{k=1}^{\infty} G^c \cap D_1 \cap \cdots \cap D_k = G^c \cap K = \varnothing.$$

Complete the uniform tightness proof by noting that

$$\liminf \mathbb{P}\{E_n \in G\} \geq \liminf \mathbb{P}\{E_n \in D_1 \cap \cdots \cap D_k\} > 1 - \varepsilon$$

if G contains $D_1 \cap \cdots \cap D_k$. □

Condition (22) points the way towards mass production of empirical central limit theorems. The Chaining Lemma makes it easy. For example, from the Equicontinuity Lemma of Section 4, we get conditions on the random covering numbers under which $\{E_n\}$ converges in distribution. The next section will describe other sufficient conditions.

23 Example. We left unfinished back in Example V.15 a limit problem for goodness-of-fit statistics with estimated parameters. The empirical processes were indexed by intervals of the real line; the estimators took the form

$$\theta_n = \theta_0 + n^{-1} \sum_{i=1}^{n} L(\xi_i) + o_p(n^{-1/2})$$

for an L with $PL = 0$, $PL^2 < \infty$. We wanted to find the limiting distribution of

$$D_n = \sup_t |E_n(-\infty, t] - n^{1/2}(\theta_n - \theta_0)\Delta(t)| + o_p(1)$$

the $\Delta(\cdot)$ being a fixed cadlag function on $[-\infty, \infty]$.

Set $\mathscr{F}$ equal to $\{L\} \cup \{(-\infty, t]: -\infty < t < \infty\}$. Express D_n in terms of a function on the corresponding $\mathscr{X}$. Define

$$H(x) = \sup_t |x((-\infty, t]) - x(L)\Delta(t)|.$$

Guard against measurability evils by restricting the supremum to rational t values: it makes no difference to E_n, Δ, or the limiting P-bridge. Clearly $H(\cdot)$ is a continuous function on $\mathscr{X}$.

You can check condition (22) by means of the Equicontinuity Lemma. The intervals have only polynomial discrimination; the inclusion of the single extra function L has a barely perceptible effect on the covering numbers. Deduce that $D_n = H(E_n) + o_p(1) \rightsquigarrow H(E_P)$. □

VII.6. Restricted Chaining

In this section the method of the Chaining Lemma is modified to develop another approach to empirical central limit theorems. The arguments for three representative examples are sketched. You might want to skip over the section at the first reading.

The chaining arguments in Section 2 assumed that the increments of the stochastic process had exponentially decreasing tail probabilities,

$$\text{(24)} \qquad \mathbb{P}\{|Z(s) - Z(t)| > \eta\} \le 2\exp(-\tfrac{1}{2}\eta^2/D^2\delta^2) \quad \text{if} \quad d(s, t) \le \delta.$$

The inequality held for every $\eta > 0$ and $\delta > 0$. We shall carry the argument further to cover processes, such as the empirical process, for which the inequality holds only in a restricted region $\mathscr{P}$ of (η, δ) pairs.

Suppose f is a bounded function, $|f| \le C$. Let δ^2 be an upper bound for the variance $\sigma^2(f) = Pf^2 - (Pf)^2$. Bennett's Inequality (Appendix B) gives

$$\begin{aligned}\text{(25)} \quad \mathbb{P}\{|E_n f| > \eta\} &= \mathbb{P}\left\{\left|\sum_{i=1}^{n} f(\xi_i) - Pf\right| > \eta n^{1/2}\right\} \\ &\le 2\exp[-\tfrac{1}{2}(\eta^2/\delta^2)B(2C\eta/(n^{1/2}\delta^2))] \\ &\le 2\exp(-\tfrac{1}{2}\lambda\eta^2/\delta^2) \quad \text{if} \quad \delta^2/\eta \ge 2C/(n^{1/2}B^{-1}(\lambda))\end{aligned}$$

for any fixed λ between 0 and 1, because $B(\cdot)$ is a continuous, decreasing function, with $B(0) = 1$.

The restricted range complicates the task of proving maximal inequalities for the stochastic process $\{Z(t)\colon t \in T\}$. We can chain as in Section 2 as long as the (η_i, δ_i) pairs remain within $\mathscr{P}$, but eventually the chain will hit the boundary of $\mathscr{P}$, when the links are getting down to lengths less than some tiny α, say. That leaves the problem of how to bound increments of Z across little links from points in T to their nearest neighbors in an α-net for T.

Remember the abbreviations $N(\delta)$, for the covering number $N(\delta, d, T)$, and $J(\delta)$, for the covering integral

$$J(\delta, d, T) = \int_0^\delta [2\log(N(u)^2/u)]^{1/2}\, du.$$

The chaining argument will work for maximal deviations down to about $J(\alpha)$. That explains the constraint $J(\alpha) \le \gamma/12D$ in the next theorem. The other constraints on α and γ are cosmetic.

26 Theorem. *Let $\{Z(t)\colon t \in T\}$ be a stochastic process that satisfies the exponential inequality* (24) *for every $\eta > 0$ and $\delta > 0$ with $\delta \ge \alpha\eta^{1/2}$, for some constant α. Suppose T has a finite covering integral $J(\cdot)$. Let $T(\alpha)$ be an α-net (containing $N(\alpha)$ points) for T; let t_α be the closest point in $T(\alpha)$ to t; and let*

$[\delta]$ *denote the set of pairs* (s, t) *with* $d(s, t) \le \delta$. *Given* $\varepsilon > 0$ *and* $\gamma > 0$ *there exists a* $\delta > 0$, *depending on* ε, γ, *and* $J(\cdot)$, *for which*

$$\mathbb{P}\left\{\sup_{[\delta]} |Z(s) - Z(t)| > 5\gamma\right\} \le 2\varepsilon + \mathbb{P}\left\{\sup_{T} |Z(t) - Z(t_\alpha)| > \gamma\right\}$$

provided $\alpha \le \frac{1}{3}\varepsilon$ *and* $\gamma \le 144$ *and* $J(\alpha) \le \min\{\gamma/12D, 3/D\}$.

PROOF. The argument is similar to the one used for the Chaining Lemma. Write $H(u)$ for $[2\log(N(u)^2/u)]^{1/2}$, as before. Choose the largest δ for which $\delta \le \frac{1}{3}\varepsilon$ and $J(\delta) \le \gamma/12D$. The assumptions about α ensure $\delta \ge \alpha$. Find the integer k for which $\delta < 3^k\alpha \le 3\delta$ then define

$$\delta_i = 3^{k-i}\alpha \quad \text{and} \quad \eta_i = D\delta_i H(\delta_{i+1}) \quad \text{for} \quad i = 0, \ldots, k.$$

Notice that $\delta_1 \le \delta < \delta_0$ and $\delta_k = \alpha$. Also

$$\begin{aligned}
\eta_0 + \cdots + \eta_{k-1} &= \sum_{i=0}^{k-1} \tfrac{9}{2} D(\delta_{i+1} - \delta_{i+2}) H(\delta_{i+1}) \\
&\le \tfrac{9}{2} D \sum_{i=0}^{\infty} \int \{\delta_{i+2} \le u < \delta_{i+1}\} H(u)\, du \\
&\le \tfrac{9}{2} D J(\delta_1) \\
&< \gamma \quad \text{because } J(\delta_1) \le J(\delta) \le \gamma/12D.
\end{aligned}$$

Choose δ_i-nets T_i containing $N(\delta_i)$ points, making sure that $T_k = T(\alpha)$. Link each t to a t_0 in T_0 through a chain of points,

$$t = t_{k+1}, \qquad t_\alpha = t_k, \qquad t_{k-1}, \ldots, t_0,$$

with t_i being the closest point of T_i to t_{i+1}. By this construction, $d(t_{i+1}, t_i) \le \delta_i$.

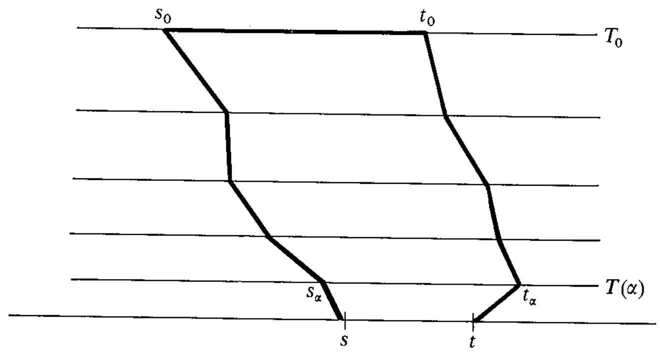
The smallest value of the ratios $\{\delta_i^2/\eta_i\}$, for $i = 0, \ldots, k-1$, occurs at $i = k-1$; all the ratios are greater than

$$3\alpha/DH(\alpha) \ge 3\alpha^2/DJ(\alpha) \ge \alpha^2.$$

The (η_i, δ_i) pairs all belong to the region in which the exponential inequality (24) holds. Apply the inequality for increments across links of the chains.

$$\begin{aligned}
\mathbb{P}\left\{\max_{T(\alpha)} |Z(t_\alpha) - Z(t_0)| > \gamma\right\} &\le \sum_{i=0}^{k-1} \mathbb{P}\left\{\max_{T_{i+1}} |Z(t_{i+1}) - Z(t_i)| > \eta_i\right\} \\
&\le \sum_{i=0}^{k-1} N(\delta_{i+1}) 2 \exp(-\tfrac{1}{2}\eta_i^2/D^2\delta_i^2) \\
&\le \sum_{i=0}^{\infty} 2N(\delta_{i+1}) \exp[-\log(N(\delta_{i+1})^2/\delta_{i+1})] \\
&\le \sum_{i=0}^{\infty} 2\delta_{i+1}
\end{aligned}$$

Notice how one of the $N(\delta_{i+1})$ factors was wasted; both factors will be needed later. The last series sums to less than ε because $\delta_1 \le \frac{1}{3}\varepsilon$ and the $\{\delta_i\}$ decrease geometrically.



Join each (s, t) pair in $[\delta]$ by two chains leading up to T_0 plus a link between s_0 and t_0.

$$\sup_{[\delta]} |Z(s) - Z(t)| \leq 2 \sup_{T} |Z(t) - Z(t_\alpha)| + 2 \max_{T(\alpha)} |Z(t_\alpha) - Z(t_0)|$$
$$+ \sup_{[\delta]} |Z(s_0) - Z(t_0)|.$$

Partition the 5γ correspondingly.

$$\mathbb{P}\left\{\sup_{[\delta]} |Z(s) - Z(t)| > 5\gamma\right\} \leq \mathbb{P}\left\{\sup_{T} |Z(t) - Z(t_\alpha)| > \gamma\right\} + \varepsilon$$
$$+ \mathbb{P}\left\{\sup_{[\delta]} |Z(s_0) - Z(t_0)| > \gamma\right\}.$$

The distance between the s_0 and t_0 of each pair appearing in the last term is less than

$$\begin{aligned} &d(s_0, s_\alpha) + d(s_\alpha, s) + d(s, t) + d(t, t_\alpha) + d(t_\alpha, t_0) \\ &\leq \sum_{i=0}^{k-1} \delta_i + \alpha + \delta + \alpha + \sum_{i=0}^{k-1} \delta_i \\ &\leq 3\delta_0 + 2\alpha + \delta \\ &\leq 12\delta. \end{aligned}$$

There are at most $N(\delta_0)^2$ such pairs. The exponential inequality holds for each pair, because $12\delta/(\alpha\gamma^{1/2}) \geq 12\gamma^{-1/2} \geq 1$.

$$\begin{aligned} &\mathbb{P}\left\{\sup_{[\delta]} |Z(s_0) - Z(t_0)| > \gamma\right\} \\ &\leq N(\delta_0)^2 2 \exp(-\tfrac{1}{2}\gamma^2/144D^2\delta^2) &&\text{because } d(s_0, t_0) \leq 12\delta \\ &\leq 2\delta \exp[\log(N(\delta)^2/\delta) - \tfrac{1}{2}\gamma^2/144D^2\delta^2] &&\text{because } N(\delta_0) \leq N(\delta) \\ &\leq 2\delta \exp[\tfrac{1}{2}\delta^{-2}(H(\delta)^2\delta^2 - \gamma^2/144D^2)] \\ &\leq 2\delta \quad \text{because} \quad H(\delta)\delta \leq J(\delta) \leq \gamma/12D \\ &< \varepsilon. \end{aligned}$$

□

Theorem 21 states a sufficient condition for empirical processes indexed by a pointwise bounded, totally bounded, permissible subclass $\mathscr{F}$ of $\mathscr{L}^2(P)$ to converge in distribution to a P-bridge: given $\eta > 0$ there exists a $\delta > 0$ for which

$$\limsup \mathbb{P}\left\{\sup_{[\delta]} |E_n(f - g)| > \eta\right\} < \varepsilon.$$

For a permissible class of bounded functions, say $0 \le f \le 1$, any condition implying finiteness of $N_1(\cdot, P, \mathscr{F})$ or $N_2(\cdot, P, \mathscr{F})$ will take care of the total boundedness. Finiteness of a covering integral will allow us to apply Theorem 26, leaving only a supremum over the class $\mathscr{H} = \{f - f_\alpha : f \in \mathscr{F}\}$ of little links. It will then suffice to prove $\sup_{\mathscr{H}} |E_n h| = o_p(1)$ to get the empirical central limit theorem. Notice that α, and hence $\mathscr{H}$, will depend on n. The next three examples sketch typical methods for handling $\mathscr{H}$.

27 Example. Equip $\mathscr{F}$ with the semimetric $d(f, g) = (P|f - g|)^{1/2}$. (This is the $\mathscr{L}^2(P)$ seminorm applied to the function $|f - g|^{1/2}$.) The square root ensures that the variance $\sigma^2(f - g)$ is less than $d(f, g)^2$. If we take $\lambda = \frac{1}{4}$, the exponential bound (25) becomes, for $d(f, g) \le \delta$,

$$\mathbb{P}\{|E_n(f - g)| > \eta\} \le 2\exp(-\tfrac{1}{8}\eta^2/\delta^2) \quad \text{if} \quad \delta^2/\eta \ge 2/(n^{1/2}B^{-1}(\tfrac{1}{4})).$$

That is, $D = 2$ and $\alpha = (2/B^{-1}(\frac{1}{4}))^{1/2} n^{-1/4}$ for Theorem 26.

The covering numbers for $d(\cdot, \cdot)$ are closely related to the $\mathscr{L}^1(P)$ covering numbers: in terms of the covering integral,

$$(28) \qquad J(\delta, d, \mathscr{F}) = \int_0^\delta [2\log(N_1(u^2, P, \mathscr{F})^2/u)]^{1/2}\, du \quad \text{for} \quad 0 < \delta < 1.$$

If J is finite, Theorem 26 can chain down to leave a class $\mathscr{H}$ of little links with $|h| \le 1$ and $P|h| \le \alpha^2$. If we add to this the condition

$$(29) \qquad \log N_1(cn^{-1/2}, P_n, \mathscr{H}) = o_p(n^{1/2}) \quad \text{for each } c > 0,$$

the empirical central limit theorem will hold.

The methods of Section II.6 work for the class $\mathscr{H}_{1/2} = \{|h|^{1/2} : h \in \mathscr{H}\}$. Notice that

$$N_2(\delta, P_n, \mathscr{H}_{1/2}) \le N_1(\delta^2, P_n, \mathscr{H})$$

because $P_n(|h_1|^{1/2} - |h_2|^{1/2})^2 \le P_n|h_1 - h_2|$. From Lemma II.33,

$$(30) \qquad \mathbb{P}\left\{\sup_{\mathscr{H}} (P_n|h|)^{1/2} > 8\alpha\right\} \le 4\mathbb{P}[N_1(\alpha^2, P_n, \mathscr{H})\exp(-n\alpha^2) \wedge 1]$$
$$= 4\mathbb{P}[\exp(\log N_1(\alpha^2, P_n, \mathscr{H}) - n\alpha^2) \wedge 1]$$
$$\to 0 \quad \text{by (29)}.$$

Symmetrize. For n large enough,

$$\mathbb{P}\left\{\sup_{\mathscr{H}} |E_n h| > 4\gamma\right\} \leq 4\mathbb{P}\left\{\sup_{\mathscr{H}} |E_n^\circ h| > \gamma\right\}.$$

Condition on ξ. Cover $\mathscr{H}$ by $M = N_1(\frac{1}{2}\gamma n^{-1/2}, P_n, \mathscr{H})$ balls, for the $\mathscr{L}^1(P_n)$ seminorm, with centers $g_1, \ldots, g_M$ in $\mathscr{H}$. Then as in Section II.6,

$$\mathbb{P}\left\{\sup_{\mathscr{H}} |E_n^\circ h| > \gamma|\xi\right\} \leq M \max_j \mathbb{P}\{|E_n^\circ g_j| > \tfrac{1}{2}\gamma|\xi\}.$$

On the set of ξ where $\sup_{\mathscr{H}} P_n|h| \leq 64\alpha^2$, Hoeffding's Inequality bounds the right-hand side by

$$2\exp[\log M - \tfrac{1}{2}(\tfrac{1}{2}\gamma)^2/(64\alpha^2)]$$

which is of order $o_p(1)$ because (29) says $\log M = o_p(n^{1/2})$. The central limit theorem follows. □

31 Example. The direct approximation method of Section II.2 gave uniform strong laws of large numbers. With a suitable bound on the number of functions needed for the approximations, we get central limit theorems.

Define a direct covering number $\Delta(\delta, P, \mathscr{H})$ as the smallest M for which there exist functions $g_1, \ldots, g_M$ such that, for every h in $\mathscr{H}$,

$$|h| \leq g_i \quad \text{and} \quad Pg_i \leq \delta + P|h| \quad \text{for some } i.$$

We may assume $0 \leq g_i \leq 1$. If

(32) $$\log \Delta(cn^{-1/2}, P, \mathscr{H}) = o(n^{1/2}) \quad \text{for each } c > 0,$$

and if the covering integral (28) from the previous example is finite, then the empirical central limit theorem holds.

Given $\gamma > 0$, choose λ in the exponential inequality (25) so that $2/B^{-1}(\lambda) = \gamma$. The dependence of λ on γ does not vitiate the chaining argument in Theorem 26; it does ensure that functions in $\mathscr{H}$ satisfy

$$P|h| \leq \alpha^2 = \gamma n^{-1/2}.$$

Find $g_1, \ldots, g_M$ according to the definition of $\Delta(\gamma n^{-1/2}, P, \mathscr{H})$. Because $Pg_i \leq 2\gamma n^{-1/2}$ for each i, the contributions of the means to E_n are small.

$$\begin{aligned}
\mathbb{P}\left\{\sup_{\mathscr{H}} |E_n h| > 4\gamma\right\} &\leq \mathbb{P}\left\{\sup_{\mathscr{H}} n^{1/2}P_n|h| > 3\gamma\right\} && \text{because } n^{1/2}P|h| \leq \gamma \\
&\leq \mathbb{P}\left\{\max_i n^{1/2}P_n g_i > 3\gamma\right\} && \text{because } |h| \leq g_i \text{ for some } i \\
&\leq M \max_i \mathbb{P}\{E_n g_i > \gamma\} && \text{because } n^{1/2}Pg_i \leq 2\gamma \\
&\leq M \max_i 2\exp[-\tfrac{1}{2}(\gamma^2/Pg_i)B(2\gamma/(n^{1/2}Pg_i))] && \text{from (25)} \\
&\leq 2\exp[\log M - \tfrac{1}{4}\gamma n^{1/2}B(1)] \\
&= o(1) \quad \text{by (32)}.
\end{aligned}$$

□

33 Example. In the previous two examples, the method of chaining left links of small $\mathscr{L}^1(P)$ seminorm at the end of the chain; $\mathscr{L}^1$ approximation methods took care of $\mathscr{H}$. If we chain instead with $\mathscr{L}^2(P)$ covering numbers, we need $\mathscr{L}^2$ approximation methods for $\mathscr{H}$.

Set $d(\cdot, \cdot)$ equal to the $\mathscr{L}^2(P)$ semimetric. Because $\sigma^2(f - g) \le d(f, g)^2$, the chaining down to $\mathscr{H}$ requires $J_2(1, P, \mathscr{F})$ finite. At the end

$$Ph^2 \le \alpha^2 = (2/B^{-1}(\tfrac{1}{4}))n^{-1/2}.$$

Invoke Lemma II.33.

$$\mathbb{P}\left\{\sup_{\mathscr{H}}(P_n h^2)^{1/2} \le 8\alpha\right\} \to 1 \quad \text{as} \quad n \to \infty$$

if the random covering numbers satisfy

$$\log N_2(cn^{-1/4}, P_n, \mathscr{H}) = o_p(n^{1/2}) \quad \text{for each } c > 0.$$

This would follow from

(34) $$J_2(cn^{-1/4}, P_n, \mathscr{H}) = o_p(1) \quad \text{for each } c > 0,$$

because

$$\begin{aligned} o_p(n^{1/4}) &= (cn^{-1/4})^{-1} J_2(cn^{-1/4}, P_n, \mathscr{H}) \\ &\ge [2\log(N_2(cn^{-1/4}, P_n, \mathscr{H})^2 n^{1/4}/c)]^{1/2}. \end{aligned}$$

Symmetrize. For all n large enough,

$$\mathbb{P}\left\{\sup_{\mathscr{H}} |E_n h| > 4\gamma\right\} \le 4\mathbb{P}\left\{\sup_{\mathscr{H}} |E_n^\circ h| > \gamma\right\}.$$

Now we are back to the sort of problem we were solving in Section 4. Condition on ξ. On the set of those ξ for which $\sup_{\mathscr{H}}(P_n h^2)^{1/2} \le 8\alpha$, chain using the Hoeffding Inequality to bound the tail probabilities. Apply the Chaining Lemma for $\mathbb{P}(\cdot \mid \xi)$, the $\mathscr{L}^2(P_n)$ seminorm, and $\varepsilon = 8\alpha$.

$$\mathbb{P}\left\{\sup_{\mathscr{H}} |E_n^\circ h| > 26 J_2(8\alpha, P_n, \mathscr{H}) \,\middle|\, \xi\right\} \le 16\alpha \quad \text{if} \quad \sup_{\mathscr{H}}(P_n h^2)^{1/2} \le 8\alpha.$$

Condition (34) and finiteness of $J_2(1, P, \mathscr{F})$ are sufficient for the empirical central limit theorem to hold. □

Notes

Theorem 5 draws on ideas from Chernoff (1954), but substitutes stochastic equicontinuity where he placed domination conditions on third-order partial derivatives. The theorem also holds if t_0 is just a local minimum for $F(\cdot)$, or if τ_n is a minimum for $F_n(\cdot)$ over a large enough neighborhood of t_0. Huber (1967, Lemma 3) made explicit the role of stochastic equicontinuity in a proof of the central limit theorem for an M-estimator.

The chaining argument abstracts the idea behind construction of processes on a dyadic rational skeleton. It appears to have entered weak convergence theory through the work of Kolmogorov and the Soviet School; it is closely related to the arguments for construction of measures in function spaces (Gihman and Skorohod 1974, Sections III.4, III.5). The Chaining Lemma is based on an arrangement by Le Cam (1983) of an argument of Dudley (1967a, 1973, 1978). Le Cam's approach avoids the complications introduced into Dudley's proof by the nuisance possibility that covering numbers $N(\delta)$ might not increase rapidly enough as δ decreases to zero. Alexander (1984a, 1984b) has refined Dudley's form of the chaining argument to prove the most precise maximal inequalities for general empirical processes to be found in the literature.

Theorem 13 is based on Theorem 2.1 of Dudley (1973), but with his modulus function increased slightly to take advantage of Le Cam's (1983) cleaner bound for the error term. The extra $(\delta \log(1/\delta))^{1/2}$ does not change the order of magnitude of the modulus for most processes.

The argument in Section 4 is based on Pollard (1982c), except for the substitution of convergence in probability (condition (16)) for uniform convergence. Kolchinsky (1982) developed a similar technique to prove a similar central limit theorem for bounded classes of functions. He imposed finiteness of $J_2(\cdot, P, \mathscr{F})$ plus a growth condition on $N_1(\cdot, P_n, \mathscr{F})$ to get results closer to those of my Example 27. Giné and Zinn (1984) have found a necessary and sufficient random entropy condition for the empirical central limit theorem.

Brown (1983) sketched the large-sample theory for the spatial median. He referred to Brown and Kildea (1979) and the appendix he wrote for Maritz (1981) for rigorous proofs, which depend on a form of stochastic equicontinuity.

The central limit theorem for k-means was proved by Pollard (1982b, 1982d) for a fixed number of clusters in euclidean space. The one-dimensional result was proved by Hartigan (1978), using a different method.

Dudley (1978, 1981a, 1981b, 1984) has developed the application of metric entropy (covering numbers) to empirical process theory. These papers extended his earlier work on entropy and sample path properties of gaussian processes (1967b, 1973), and on the multidimensional empirical distribution function (1966a).

Dudley (1966a, 1978) introduced most of the ideas needed to prove central limit theorems for empirical processes indexed by sets. He extended these ideas to classes of functions in (1981a, 1981b). His lecture notes (1984) provide the best available overview of empirical process theory, as of this writing. The proof of my Theorem 21 was inspired by Chapter 4 of those lecture notes, which reworked ideas from Dudley and Philipp (1983). If $Pf = 0$ for each f in $\mathscr{F}$, a standardization that can be imposed without affecting E_n or E_P, the conditions of Theorem 21 are also necessary for the empirical central limit theorem.

The first central limit theorems for empirical processes indexed by classes of sets were proved by the direct approximation method. Bolthausen (1978) worked with the class of compact, convex subsets of the unit square in $\mathbb{R}^2$. He applied an entropy bound due to Dudley (1974). Révész (1976) indexed the processes by classes of sets with smooth boundaries. Earlier work of Sun was, unfortunately, not published until quite recently (Pyke and Sun 1982). Dudley's (1978) Theorem 5.1 imposed a condition on the "metric entropy with inclusion" that corresponds to finiteness of a covering integral. Strassen and Dudley (1969) proved a central limit theorem for empirical processes indexed by classes of smooth functions. They deduced the result from their central limit theorem for sums of independent random elements of spaces of continuous functions. All these theorems depend on existence of good bounds for the rate of growth of entropy functions (covering numbers). For more about this see Dudley (1984, Sections 6 and 7) and Gaenssler (1984).

Theorem 26 resets an argument of Le Cam (1983). Such an approximation theorem has been implicit in the work of Dudley. Giné and Zinn (1984) have pointed out the benefits of stripping off the $\mathscr{L}^2(P)$ chaining argument, to expose more clearly the problem of how to handle the little links left at the end of the chain. They have also stressed the strong parallels between empirical processes and gaussian processes. The examples in Section 6 follow the lead of Giné and Zinn: Example 27 is based on their adaptation of Le Cam's (1983) square-root trick; Example 31 is based on their improvement of Dudley's (1978) "metric entropy with inclusion" method; Example 33 is based on their Theorem 5.5.

Problems

[1] Prove that the stochastic equicontinuity concept of Definition 2 follows from: $Z_n(\tau_n) - Z_n(t_0) \to 0$ in probability for every sequence $\{\tau_n\}$ that converges in probability to t_0. [Suppose the defining property fails for some $\eta > 0$ and $\varepsilon > 0$. For a sequence of neighborhoods $\{U_k\}$ that shrink to t_0 find positive integers $n(1) < n(2) < \cdots$ with

$$\mathbb{P}\left\{\sup_{U_k} |Z_{n(k)}(t) - Z_{n(k)}(t_0)| > \eta\right\} > \tfrac{1}{2}\varepsilon$$

for every k. Choose random elements $\{\tau_n\}$ of T such that, for $n(k) \le n < n(k+1)$,

$$|Z_n(\omega, \tau_n(\omega)) - Z_n(\omega, t_0)| \ge \tfrac{1}{2} \sup_{U_k} |Z_n(\omega, t) - Z_n(\omega, t_0)|$$

and $\tau_n(\omega)$ belongs to U_k. Appendix C covers measurability of τ_n.]

[2] Let $\{f(\cdot, t): t \in T\}$ be a collection of $\mathbb{R}^k$-valued functions indexed by a subset of $\mathbb{R}^k$. Suppose $P|f(\cdot, t)|^2 < \infty$ for each t. Set $F(t) = Pf(\cdot, t)$ and $F_n(t) = P_n f(\cdot, t)$. Let $\{\tau_n\}$ be a sequence converging in probability to a value t_0 at which $F(t_0) = 0$. If
(a) $F(\cdot)$ has a non-singular derivative matrix D at t_0;
(b) $F_n(\tau_n) = o_p(n^{-1/2})$;

(c) $\{E_n f(\cdot, t)\}$ is stochastically equicontinuous at t_0;
then $n^{1/2}(\tau_n - t_0) \rightsquigarrow N(0, D^{-1}P[f(\cdot, t_0)f(\cdot, t_0)']D^{-1})$. [Compare with Huber (1967).]

[3] For a class $\mathscr{F}$ to index a P-motion it must be totally bounded under the $\mathscr{L}^2(P)$ seminorm ρ_P. [First show $\mathscr{F}$ is bounded: otherwise $|B_P(f_n)| \to \infty$ in probability for some $\{f_n\}$, violating boundedness of P-motion sample paths. Total boundedness will then follow from: for each $\varepsilon > 0$, every f lies within ε of some linear combination of a fixed, finite subclass of $\mathscr{F}$. If for some ε no such finite subclass exists, find $\{f_n\}$ such that

$$f_{n+1} = g_{n+1} + \sum_{j=1}^{n} a_{nj} f_j,$$

where $\rho_P(g_{n+1}) \ge \varepsilon$ and g_{n+1} is orthogonal to $f_1, \ldots, f_n$. Fix an M. Show that there exists a $\delta > 0$, depending on M and ε, for which

$$\mathbb{P}\{B_P(f_{n+1}) \ge M \,|\, B_P(f_1), \ldots, B_P(f_n)\} \ge \delta.$$

Deduce that $\mathbb{P}\{\sup_n B_P(f_n) \ge M\} = 1$ for every M, which contradicts boundedness of the sample paths. Notice that continuity of the sample paths does not enter the argument. Dudley (1967b).]

[4] If $\sup_{\mathscr{F}} |Pf|$ is finite then $\mathscr{F}$ must be totally bounded under the $\mathscr{L}^2(P)$ seminorm ρ_P if it supports a P-bridge. [Choose Z with a $N(0, 1)$ distribution independent of E_P. The process $B(f) = E_P(f) + ZPf$ is a P-motion with bounded sample paths. Invoke Problem 3. The condition on the means is needed—consider the $\mathscr{F}$ consisting of all constant functions. The P-bridge is unaffected by addition of arbitrary constants to functions in $\mathscr{F}$; it depends only on the projection of $\mathscr{F}$ onto the subspace of $\mathscr{L}^2(P)$ orthogonal to the constants.]

[5] Let $\mathscr{H}_1$ be a class of functions with an envelope H in $\mathscr{L}^2(P)$. Set $\mathscr{H}_2 = \{h^2 : h \in \mathscr{H}_1\}$. Show that

$$N_1(4\varepsilon(QH^2)^{1/2}, Q, \mathscr{H}_2) \le N_2(2\varepsilon, Q, \mathscr{H}_1).$$

[By the Cauchy–Schwarz inequality,

$$Q|h_1^2 - h_2^2| \le Q(2H|h_1 - h_2|) \le 2(QH^2)^{1/2}(Q|h_1 - h_2|^2)^{1/2}$$

if both $|h_1| \le H$ and $|h_2| \le H$.]

[6] Let $\mathscr{F}$ be a permissible class of functions with envelope F. Suppose

$$J_2(\delta, P_n, \mathscr{F}) = o_p(n^{1/2}) \quad \text{for each} \quad \delta > 0.$$

[Condition (16) of the Equicontinuity Lemma implies that $J_2(\delta, P_n, \mathscr{F}) = O_p(1)$ for each $\delta > 0$.] Show that $\mathscr{H}_2 = \{(f - g)^2 : f - g \in \mathscr{F}\}$ satisfies the sufficient condition (Theorem II.24) for the uniform strong law of large numbers:

$$\log N_1(\varepsilon, P_n, \mathscr{H}_2) = o_p(n), \quad \text{for each } \varepsilon > 0.$$

[Set $H = 2F$ and $\mathscr{H}_1 = \{f - g : f, g \in \mathscr{F}\}$. Show that, for $1 > \varepsilon > 0$,

$$N_2(2\varepsilon, P_n, \mathscr{H}_1) \le N_2(\varepsilon, P_n, \mathscr{F})^2 \le \varepsilon \exp(\tfrac{1}{2} J_2(\varepsilon, P_n, \mathscr{F})^2/\varepsilon^2).$$

Deduce from this inequality, Problem 5, and the strong law of large numbers for $\{P_nH^2\}$ that, if $1 > \varepsilon > 0$,

$$\begin{aligned}
&\mathbb{P}\{\log N_1(4\varepsilon(2PH^2)^{1/2}, P_n, \mathscr{H}_2) > n\eta\} \\
&\quad \le \mathbb{P}\{\log N_1(4\varepsilon(P_nH^2)^{1/2}, P_n, \mathscr{H}_2) > n\eta\} + \mathbb{P}\{P_nH^2 > 2PH^2\} \\
&\quad \le \mathbb{P}\{\log N_2(2\varepsilon, P_n, \mathscr{H}_1) > n\eta\} + \mathbb{P}\{P_nH^2 > 2PH^2\} \\
&\quad \le \mathbb{P}\{\tfrac{1}{2}J_2(\varepsilon, P_n, \mathscr{F})^2/\varepsilon^2 > n\eta\} + \mathbb{P}\{P_nH^2 > 2PH^2\} \\
&\quad \to 0.
\end{aligned}$$

A weaker result was proved by Pollard (1982c).]

[7] If $\mathscr{F}$ is totally bounded under the $\mathscr{L}^2(P)$ seminorm, then the space $C(\mathscr{F}, P)$ of bounded, uniformly continuous, real functions on $\mathscr{F}$ is separable. [Suppose $|x(f) - x(g)| < \varepsilon$ whenever $\rho_P(f - g) \le 2\delta$. Choose $\{f_1, \ldots, f_m\}$ as a maximal set with $\rho_P(f_i - f_j) \ge \frac{1}{2}\delta$. Use the weighting functions $\Delta_i(\cdot)$ from the proof of Theorem 21 to interpolate between rational approximations to the $\{x(f_i)\}$.]

[8] Suppose $\mathscr{F}$ is totally bounded under the $\mathscr{L}^2(P)$ seminorm. If two probability measures λ and μ on the σ-field $\mathscr{B}^P$ have the same fidis, and if both concentrate on $C(\mathscr{F}, P)$, then they must agree everywhere on $\mathscr{B}^P$. [Show that λ and μ agree for all finite intersections of fidi sets and closed balls with centers in $C(\mathscr{F}, P)$. For example, consider a closed ball $B(x, r)$ with x in $C(\mathscr{F}, P)$. Let $\{f_1, f_2, \ldots\}$ be a countable, dense subset of $C(\mathscr{F}, P)$. Define

$$B_n = \{z \in C(\mathscr{F}, P) : |z(f_i) - x(f_i)| \le r \text{ for } 1 \le i \le n\}.$$

Show that $\mu B(x, r) \le \mu B_n = \lambda B_n \to \lambda B(x, r)$ as $n \to \infty$. Extend the result to finite collections of closed balls and fidi sets, then apply a generating-class argument.]

[9] The property that the graphs have only polynomial discrimination is not preserved by the operation of summing two classes of functions. That is, both $\mathscr{F}$ and $\mathscr{G}$ can have the property without the class $\mathscr{S} = \{f + g : f \in \mathscr{F}, g \in \mathscr{G}\}$ having it. Let $\mathscr{D} = \{D_1, D_2, \ldots\}$ be the set of indicator functions of all finite sets of rational numbers in $[0, 1]$. Let $\mathscr{F} = \{2n + D_n : n = 1, 2, \ldots\}$ and $\mathscr{G} = \{-2n : n = 1, 2, \ldots\}$. The graphs from neither class can shatter two-point sets, but $\mathscr{S}$ can shatter arbitrarily large finite sets of rationals in $[0, 1]$. [The roundabout reasoning used to bound the covering numbers in Example 18 may not be completely unnecessary.]

CHAPTER VIII
Martingales

... in which martingale central limit theorems in discrete and continuous time are proved. An extended non-trivial application to Kaplan–Meier estimators—estimation of distribution functions from censored data—is sketched.

VIII.1. A Central Limit Theorem for Martingale-Difference Arrays

Martingale theory must surely be the most successful of all the attempts to extend the classical theory for sums of independent random variables to cover dependent variables. Many of the classical limit theorems have martingale analogues that rival them for elegance and far exceed them in diversity of application. We shall explore two of these martingale theorems in this chapter.

One main change in technique will become apparent. Where proofs for independent summands use truncation to protect against occasional abnormally large increments—the sort of thing implicit in something like the Lindeberg condition—martingale proofs can resort to stopping time arguments. Optional stopping preserves the conditional expectation connection within a martingale sequence as long as the decision to stop is based only on past behavior of the sequence. The prohibition against peering into the future to anticipate abnormally large increments imposes a characteristic feature on martingale theorems. One needs conditions that protect against the behavior of the worst single increment, because the decision to stop can be taken only after that increment has had its effect.

In central limit theorems, independence allows one to factorize expectations of products and separate out the contribution of a particular increment from contributions of past and future increments. The martingale property allows a weaker factorization, for expectations conditional on the past alone. Conditional variances take over the role played by variances of independent summands. But apart from that, the arguments for martingales share the same inspiration as the proof of the Lindeberg Central Limit Theorem in Section III.4.

We shall prove asymptotic normality for row sums of martingale-difference arrays. That is, for each n we have random variables $\xi_{n1}, \ldots, \xi_{nn}$ (we avoid some messy notation, and lose no generality, by assuming exactly n variables in the nth row) and σ-fields $\mathscr{E}_{n0} \subseteq \cdots \subseteq \mathscr{E}_{nn}$ for which

(a) $\mathbb{P}(\xi_{nj} | \mathscr{E}_{n,j-1}) = 0$ for $j = 1, \ldots, n$;
(b) ξ_{nj} is $\mathscr{E}_{nj}$-measurable.

Define conditional variances

$$v_{nj} = \mathbb{P}(\xi_{nj}^2 | \mathscr{E}_{n,j-1}) \quad \text{for} \quad j = 1, \ldots, n.$$

Notice that v_{nj} is an $\mathscr{E}_{n,j-1}$-measurable random variable. Convergence of sums of conditional variances will be the only connection tying together the variables in different rows of the array.

1 Theorem. *Let $\{\xi_{nj}\}$ be a martingale-difference array. If as $n \to \infty$,*

(i) *$\sum_j v_{nj} \to \sigma^2$ in probability, with σ^2 a positive constant;*
(ii) *for every $\varepsilon > 0$, the sum $\sum_j \mathbb{P}(\xi_{nj}^2\{|\xi_{nj}| > \varepsilon\} | \mathscr{E}_{n,j-1})$ converges in probability to zero (a Lindeberg condition);*

then $\xi_{n1} + \cdots + \xi_{nn} \rightsquigarrow N(0, \sigma^2)$.

Proof. Without loss of generality set $\sigma^2 = 1$. Let us check pointwise convergence of characteristic functions:

$$\mathbb{P} \exp[it(\xi_{n1} + \cdots + \xi_{nn})] \to \exp(-\tfrac{1}{2}t^2).$$

We really will need some of the special multiplicative properties of the complex exponential function, and not just its smoothness by way of three bounded, continuous derivatives as in Section III.4. The randomness of the conditional variances fouls up the argument based on successive substitution of matching normal increments, which worked for independent summands.

At the risk of notational abuse (more will follow) abbreviate the conditional expectation $\mathbb{P}(\cdot | \mathscr{E}_{nj})$ to $\mathbb{P}_j(\cdot)$. Write $R(x)$ for the remainder term $e^{ix} - 1 - ix$ and S_{nj} for the partial sum $\xi_{n1} + \cdots + \xi_{nj}$. Define r_{nj} as the conditional expectation $\mathbb{P}_{j-1}R(t\xi_{nj})$. When the $\{\xi_{nj}\}$ are small, in a sense to be made precise by condition (ii), we shall get $r_{nj} \approx -\frac{1}{2}v_{nj}$.

The proof will work by successive conditioning. We pin down the effect of individual increments by evaluating

$$\mathbb{P} \exp(itS_{nn}) = \mathbb{P}(\mathbb{P}_0(\mathbb{P}_1(\mathbb{P}_2(\cdots(\mathbb{P}_{n-1} \exp(itS_{nn})\cdots))))),$$

a layer at a time. Start with the innermost conditional expectation. Factorize out the part depending on $\mathscr{E}_{n,j-1}$ then expand the remaining $\exp(it\xi_{nn})$.

$$\exp(itS_{n,n-1})\mathbb{P}_{n-1}[1 + it\xi_{nn} + R(t\xi_{nn})] = \exp(itS_{n,n-1})(1 + r_{nn}).$$

The $1 + r_{nn}$ factor foils the attempt to work the same idea with $\mathbb{P}_{n-2}$; it won't cooperate by slipping outside the conditional expectation, leaving $\exp(itS_{n,n-1})$ to enjoy the same treatment from $\mathbb{P}_{n-2}$ that $\exp(itS_{nn})$ received from $\mathbb{P}_{n-1}$. We could clear away the obstacle by dividing out the offending factor.

$$\begin{aligned}\mathbb{P}_{n-2}\mathbb{P}_{n-1}[(1 + r_{nn})^{-1}\exp(itS_{nn})] &= \mathbb{P}_{n-2}\exp(itS_{n,n-1})\\ &= \exp(itS_{n,n-2})(1 + r_{n,n-1}).\end{aligned}$$

We could get rid of the $(1 + r_{n,n-1})$ in a similar fashion.

If you pursue this idea through each layer of conditioning, you will see the sense in starting from

$$\prod_{j=1}^{n}(1 + r_{nj})^{-1}\exp(itS_{nn}).$$

The remote possibility that one r_{nj} might get close to -1 could cause minor difficulty when we come to bound the product term. To avoid this, replace $(1 + r_{nj})^{-1}$ by $(1 - r_{nj})$. When $r_{nj} \approx 0$ the change has little effect. Define

$$T_{nk} = \prod_{j=1}^{k}(1 - r_{nj}) \quad \text{and} \quad Z_{nk} = T_{nk}\exp(itS_{nk}).$$

If we could show $\mathbb{P}|T_{nn} - \exp(\frac{1}{2}t^2)| \to 0$ and $\mathbb{P}Z_{nn} \to 1$, then it would follow that

$$\begin{aligned}&|\mathbb{P}\exp(itS_{nn}) - \exp(-\tfrac{1}{2}t^2)|\\ &\quad \le \exp(-\tfrac{1}{2}t^2)[\mathbb{P}|\exp(itS_{nn} + \tfrac{1}{2}t^2) - Z_{nn}| + |\mathbb{P}Z_{nn} - 1|]\\ &\quad \to 0.\end{aligned}$$

We would get the desired results for $\{T_{nn}\}$ and $\{Z_{nn}\}$ from:

(a) $\sum_j r_{nj} \to -\frac{1}{2}t^2$ in probability;
(b) $\sum_j |r_{nj}| \le t^2$;
(c) $\max_j |r_{nj}| \to 0$ in probability.

The second of these requirements need not be satisfied but, without loss of generality, we may act as if it were.

We replace ξ_{nj} by $\xi_{nj}\{j \le \sigma_n\}$, where $\sigma_n = \max\{k: \sum_{j=1}^{k} v_{nj} \le 2\}$. Interpret σ_n as zero if $v_{n1} > 2$. Because v_{nj} is $\mathscr{E}_{n,j-1}$-measurable, the event $\{j \le \sigma_n\}$ is $\mathscr{E}_{n,j-1}$-measurable; σ_n is a stopping time. The new variables are martingale differences. The new row sums have the same asymptotic behavior as the original row sums:

$$\mathbb{P}\left\{\sum_{j=1}^{n}\xi_{nj} \ne \sum_{j=1}^{n}\xi_{nj}\{j \le \sigma_n\}\right\} \le \mathbb{P}\{n > \sigma_n\} = \mathbb{P}\left\{\sum_{j=1}^{n} v_{nj} > 2\right\} \to 0.$$

The quantities $w_{nj} = \mathbb{P}_{j-1}R(t\xi_{nj}\{j \le \sigma_n\})$ satisfy analogues of the requirements (a), (b), and (c). The argument depends on the inequalities (Problem 1) for real x,

$$|R(x)| \le \tfrac{1}{2}|x|^2;$$
$$|R(x) + \tfrac{1}{2}x^2| \le \min\{|x|^2, \tfrac{1}{6}|x|^3\}.$$

For the analogue of (c): if $\delta > 0$, no $|w_{nj}|$ exceeds

$$\max_j \mathbb{P}_{j-1}\tfrac{1}{2}t^2\xi_{nj}^2\{j \le \sigma_n\} \le \tfrac{1}{2}t^2\left[\delta^2 + \sum_j \mathbb{P}_{j-1}\xi_{nj}^2\{|\xi_{nj}| > \delta\}\right].$$

Set δ small, then invoke the Lindeberg condition. For the analogue of (b):

$$\begin{aligned}\sum_{j=1}^n |w_{nj}| &\le \tfrac{1}{2}t^2 \sum_{j=1}^n \mathbb{P}_{j-1}\xi_{nj}^2\{j \le \sigma_n\}\\ &= \tfrac{1}{2}t^2 \sum_{j=1}^n \{j \le \sigma_n\}v_{nj} \quad \text{because } \{j \le \sigma_n\} \text{ is } \mathscr{E}_{n,j-1}\text{-measurable}\\ &\le t^2 \quad \text{by definition of } \sigma_n.\end{aligned}$$

For the analogue of (a), fix $\delta > 0$:

$$\begin{aligned}&\left|\sum_{j=1}^n w_{nj} + \tfrac{1}{2}t^2 \sum_{j=1}^n v_{nj}\right|\\ &\le \sum_j \mathbb{P}_{j-1}|R(t\xi_{nj}\{j \le \sigma_n\}) + \tfrac{1}{2}t^2\xi_{nj}^2\{j \le \sigma_n\} + \tfrac{1}{2}t^2\xi_{nj}^2\{j > \sigma_n\}|\\ &\le \sum_j \mathbb{P}_{j-1}[\tfrac{1}{6}|t\xi_{nj}\{j \le \sigma_n\}|^3\{|\xi_{nj}| \le \delta\}\\ &\qquad + t^2\xi_{nj}^2\{j \le \sigma_n\}\{|\xi_{nj}| > \delta\} + \tfrac{1}{2}t^2\xi_{nj}^2\{j > \sigma_n\}]\\ &\le \sum_j \tfrac{1}{6}\delta|t|^3 v_{nj} + t^2\sum_j \mathbb{P}_{j-1}\xi_{nj}^2\{|\xi_{nj}| > \delta\} + \tfrac{1}{2}t^2\{n > \sigma_n\}\sum_j v_{nj}.\end{aligned}$$

The first sum can be made small with high probability, by an appropriate choice of δ; the last two terms converge in probability to zero.

We could carry σ_n along throughout the rest of the argument, but that would clutter up the notation. Instead, let us assume that (a), (b), and (c) hold. While we are at it, let us drop the n subscript for variables in the nth row of the array; all calculations take place within the nth row. Our task is to prove that

$$\text{(2)} \qquad \mathbb{P}|T_n - \exp(\tfrac{1}{2}t^2)| \to 0 \qquad \text{and} \qquad \mathbb{P}Z_n \to 1,$$

where $T_k = \prod_{j=1}^k(1 - r_j)$ and $Z_k = T_k\exp(itS_k)$.

The path leading from (a), (b), and (c) to the first assertion in (2) is well-worn (Chung 1968, Section 7.1). For complex θ,

$$|\log(1 - \theta) + \theta| \le |\theta|^2 \quad \text{if} \quad |\theta| \le \tfrac{1}{2}.$$

Apply the inequality to each r_j. When $\max_j|r_j| \le \frac{1}{2}$,

$$\begin{aligned}\left|\sum_j \log(1 - r_j) + \sum_j r_j\right| &\le \sum_j |r_j|^2\\ &\le t^2 \max_j|r_j| \quad \text{by (b).}\end{aligned}$$

It follows from (a), (c), and continuity of the exponential function that $\{T_n\}$ converges in probability to $\exp(\frac{1}{2}t^2)$. Each $|T_n|$ is bounded, because of (b):

$$(3) \qquad |T_n| \leq \prod_j (1 + |r_j|) \leq \exp\left(\sum_j |r_j|\right) \leq \exp(t^2).$$

Boundedness plus convergence in probability imply convergence in L^1.

For the proof of the second assertion in (2), bound the errors that accrue during the calculation of conditional expectations layer-by-layer.

$$\begin{aligned} \mathbb{P}_{j-1}Z_j &= \mathbb{P}_{j-1}[T_j \exp(itS_{j-1} + it\xi_j)] \\ &= T_j \exp(itS_{j-1})\mathbb{P}_{j-1}\exp(it\xi_j) \\ &= (1 - r_j)Z_{j-1}[1 + \mathbb{P}_{j-1}(it\xi_j) + \mathbb{P}_{j-1}R(t\xi_j)] \\ &= (1 - r_j)Z_{j-1}(1 + r_j) \\ &= Z_{j-1} - r_j^2 Z_{j-1}. \end{aligned}$$

Thus $|\mathbb{P}Z_j - \mathbb{P}Z_{j-1}| \leq \mathbb{P}|r_j^2 Z_{j-1}| \leq \exp(t^2)\mathbb{P}|r_j^2|$, because inequality (3) implies $|Z_{j-1}| = |T_{j-1}| \leq \exp(t^2)$. Sum over j.

$$|\mathbb{P}Z_n - 1| \leq \exp(t^2)\mathbb{P}\sum_{j=1}^{n} |r_j|^2 \to 0 \quad \text{as} \quad n \to \infty,$$

because $\sum_j |r_j|^2 \leq \min\{t^4, t^2 \max_j |r_j|\}$. □

4 Example. A sequence of real random variables $X_0, X_1, \ldots$ is called an autoregression of order one if $X_n = \theta_0 X_{n-1} + u_n$ for some fixed θ_0. The innovations $\{u_n\}$ are assumed independent and identically distributed with zero mean and finite variance σ^2. The initial value X_0 is independent of the $\{u_n\}$. The least-squares estimator θ_n minimizes

$$\sum_{j=1}^{n} (X_j - \theta X_{j-1})^2.$$

Solving for θ_n and standardizing, we get

$$(5) \qquad n^{1/2}(\theta_n - \theta_0) = \left[n^{-1/2}\sum_{j=1}^{n} u_j X_{j-1}\right] \Big/ \left[n^{-1}\sum_{j=1}^{n} X_{j-1}^2\right].$$

With the help of Theorem 1 we can prove

$$n^{1/2}(\theta_n - \theta_0) \rightsquigarrow N(0, 1 - \theta_0^2)$$

provided $|\theta_0| < 1$ and $\mathbb{P}X_0^2 < \infty$. This will follow from convergence results for the denominator and numerator in (5):

$$n^{-1}\sum_{j=1}^{n} X_{j-1}^2 \to \sigma^2/(1 - \theta_0^2) \quad \text{in probability,}$$

$$\sum_{j=1}^{n} n^{-1/2} u_j X_{j-1} \rightsquigarrow N(0, \sigma^4/(1 - \theta_0^2)).$$

Start with the denominator.

Square both sides of the defining equation of the autoregression, then sum over j.

$$\sum_{j=1}^{n} X_j^2 = \sum_{j=1}^{n} u_j^2 + 2\theta_0 \sum_{j=1}^{n} u_j X_{j-1} + \theta_0^2 \sum_{j=1}^{n} X_{j-1}^2.$$

Rearrange then divide through by n.

$$(6) \qquad (1-\theta_0^2)n^{-1} \sum_{j=1}^{n} X_{j-1}^2 = n^{-1} \sum_{j=1}^{n} u_j^2 + 2\theta_0 n^{-1} \sum_{j=1}^{n} u_j X_{j-1} + n^{-1}(X_0^2 - X_n^2).$$

On the right-hand side, the first term converges almost surely to σ^2, by the strong law of large numbers. The third term converges in L^1 to zero, because repeated application of the equality

$$\mathbb{P}X_n^2 = \mathbb{P}u_n^2 + 2\theta_0 \mathbb{P}u_n \mathbb{P}X_{n-1} + \theta_0^2 \mathbb{P}X_{n-1}^2 = \sigma^2 + \theta_0^2 \mathbb{P}X_{n-1}^2,$$

yields

$$\mathbb{P}X_n^2 = \sigma^2(1 + \theta_0^2 + \cdots + \theta_0^{2n-2}) + \theta_0^{2n}\mathbb{P}X_0^2 \to \sigma^2/(1-\theta_0^2).$$

The n^{-1} brings the limit down to zero. The middle term on the right-hand side of (6) converges in L^2 to zero:

$$\begin{aligned} n^{-2}\mathbb{P}&\left(\sum_{j=1}^{n} u_j X_{j-1}\right)^2 \\ &= n^{-2} \sum_{j=1}^{n} \mathbb{P}u_j^2 X_{j-1}^2 \quad \text{independence kills the cross-product terms} \\ &= n^{-2} \sum_{j=1}^{n} \sigma^2 \mathbb{P}X_{j-1}^2 \\ &= O(n^{-1}) \quad \text{because } \{\mathbb{P}X_j^2\} \text{ is convergent.} \end{aligned}$$

So much for the denominator.

Write $\mathscr{E}_n$ for the σ-field generated by $\{X_0, u_1, \ldots, u_n\}$. Abbreviate $\mathbb{P}(\cdot \mid \mathscr{E}_j)$ to $\mathbb{P}_j(\cdot)$. The variables $\{u_j X_{j-1}\}$ are martingale differences for $\{\mathscr{E}_j\}$. Apply Theorem 1 to the sum in the numerator of (5). For condition (i):

$$v_{nj} = \mathbb{P}_{j-1}(n^{-1}u_j^2 X_{j-1}^2) = n^{-1}\sigma^2 X_{j-1}^2,$$

$$\sum_{j=1}^{n} v_{nj} = \sigma^2 n^{-1} \sum_{j=1}^{n} X_{j-1}^2 \to \sigma^4/(1-\theta_0^2) \quad \text{in probability.}$$

The Lindeberg condition demands a more delicate argument if higher-moment constraints on the innovations are to be avoided.

$$\begin{aligned} n^{-1} &\sum_{j=1}^{n} \mathbb{P}_{j-1} u_j^2 X_{j-1}^2 \{|u_j X_{j-1}| > \varepsilon n^{1/2}\} \\ &\le n^{-1} \sum_{j=1}^{n} \mathbb{P}_{j-1} u_j^2 X_{j-1}^2 [\{u_j^2 > \varepsilon n^{1/2}\} + \{X_{j-1}^2 > \varepsilon n^{1/2}\}] \\ &= n^{-1} \sum_{j=1}^{n} X_{j-1}^2 \mathbb{P}u_1^2 \{u_1^2 > \varepsilon n^{1/2}\} + n^{-1} \sum_{j=1}^{n} \sigma^2 X_{j-1}^2 \{X_{j-1}^2 > \varepsilon n^{1/2}\}. \end{aligned}$$

The first sum converges to zero in probability, because u_1^2 is integrable. The second sum converges in L^1 to zero because the sequence $\{X_n^2\}$ is uniformly integrable (Problem 2). □

VIII.2. Continuous Time Martingales

A stochastic process $\{Z(t): 0 \le t < \infty\}$ is said to be a martingale with respect to an increasing family of σ-fields $\{\mathscr{E}_t: 0 \le t < \infty\}$ if $Z(t)$ is adapted to the σ-fields (that is, $Z(t)$ is $\mathscr{E}_t$-measurable) and $\mathbb{P}(Z(s) \mid \mathscr{E}_t) = Z(t)$ whenever $s > t$. After some fiddling around with sets of measure zero it can usually be arranged that such a process has cadlag sample paths (Dellacherie and Meyer 1982, Section VI.1), in which case it may be studied as a random element of $D[0, \infty)$.

Call Z an L^2-martingale if it has cadlag sample paths and $\mathbb{P}Z(t)^2 < \infty$ for each t. The behavior of an L^2-martingale is largely determined by the conditional variances of its increments. The conditional expectation of $[Z(t+\delta) - Z(t)]^2$ given $\mathscr{E}_t$ plays a role similar to that of the conditional variance v_{nj} in Section 1. The most economical way to explain this uses some deeper results from the Strasbourg theory of stochastic processes. We could avoid the appeal to the deeper theory by building its special consequences into the martingale calculations for each particular application. That would always work for martingales that evolve by discrete jumps; the calculations would be similar to those in Section 1. The theory would be more self-contained, but it would disguise the unifying concept of the conditional variance process.

According to the Doob–Meyer decomposition (Theorem VII.12 of Dellacherie and Meyer (1982), applied to the supermartingale $-Z^2$), for each L^2-martingale Z, the process Z^2 has a unique representation as a sum $V + M$ of a martingale M and an increasing, predictable, conditional variance process V with $V(0) = 0$. Both M and V have cadlag sample paths. (Strictly speaking, for this decomposition we need the σ-fields $\{\mathscr{E}_t\}$ to satisfy the "usual conditions": $\mathscr{E}_0$ should contain all $\mathbb{P}$-negligible sets and each $\mathscr{E}_t$ should equal the intersection of the σ-fields $\mathscr{E}_s$ for $s > t$.) The adjective "predictable" has the technical meaning that $V(\omega, t)$ is measurable with respect to the σ-field on $\Omega \otimes [0, \infty)$ generated by the class of all adapted, left-continuous processes. So V behaves something like a process with left-continuous sample paths; its paths can be predicted a tiny instant into the future. We will need the predictability property only in Lemma 11.

If the martingale Z changes only by jumps $\xi_1, \xi_2, \ldots$ occurring at fixed times $t_1 < t_2 < \cdots$, and if $\mathscr{E}_t = \mathscr{E}_{t_k}$ for $t_k \le t < t_{k+1}$, then V is just a sum of conditional variances:

$$V(t) = \mathbb{P}(\xi_1^2 \mid \mathscr{E}_0) + \mathbb{P}(\xi_2^2 \mid \mathscr{E}_{t_1}) + \cdots + \mathbb{P}(\xi_k^2 \mid \mathscr{E}_{t_{k-1}}) \tag{7}$$

whenever $t_k \leq t < t_{k+1}$. You can check directly that $Z^2 - V$ is a martingale and that there exists a sequence of left-continuous, adapted processes converging pointwise on $\Omega \otimes [0, \infty)$ to V. The value of V at t_k corresponds to what we would have written as $v_1 + \cdots + v_k$ in Section 1. You might take this as your guiding example for the rest of the section if you wish to avoid all appeals to the Strasbourg theory.

The process V carries information about the conditional variances of the increments of Z. If $s > t$,

$$\begin{aligned} (8) \qquad & \mathbb{P}([Z(s) - Z(t)]^2 \mid \mathscr{E}_t) \\ &= \mathbb{P}(Z(s)^2 \mid \mathscr{E}_t) - Z(t)^2 - 2Z(t)\mathbb{P}(Z(s) - Z(t) \mid \mathscr{E}_t) \\ &= \mathbb{P}(V(s) \mid \mathscr{E}_t) + \mathbb{P}(M(s) \mid \mathscr{E}_t) - V(t) - M(t) \\ &= \mathbb{P}(V(s) - V(t) \mid \mathscr{E}_t). \end{aligned}$$

For s very close to t, the predictability of V makes $V(s) - V(t)$ almost $\mathscr{E}_t$-measurable; the last conditional expectation almost equals $V(s) - V(t)$, in a sense that will receive a more rigorous meaning later.

By means of a simple Tchebychev inequality we get from V a bound on the size of the increments of Z in precisely the form required by the stopping time argument of Lemma V.7. The maximal inequality provided by that lemma lies at the heart of any proof for convergence in distribution in spaces of cadlag functions.

9 Lemma. *Let $\{Z(t): 0 \leq t \leq b\}$ be an L^2-martingale with conditional variance process V. If, for every t,*

$$\mathbb{P}(V(b) - V(t) \mid \mathscr{E}_t) \leq \delta^2/12 \quad \textit{almost surely,}$$

then $\mathbb{P}\left\{\sup_{t \leq b} |Z(t) - Z(0)| > \delta\right\} \leq 3\mathbb{P}\{|Z(b) - Z(0)| > \tfrac{1}{2}\delta\}.$

Proof. With no loss of generality we may assume $Z(0) = 0$. Write $\mathbb{P}_t(\cdot)$ for expectation conditional on $\mathscr{E}_t$. Lemma V.7 invites us to check that

$$\mathbb{P}_t\{|Z(b) - Z(t)| \leq \tfrac{1}{2}|Z(t)|\} \geq \tfrac{1}{3} \quad \text{on} \quad \{|Z(t)| > \delta\}.$$

We shall do this by bounding the conditional probability from below by

$$(10) \qquad \tfrac{2}{3} - 4|Z(t)|^{-2}\mathbb{P}_t|Z(b) - Z(t)|^2,$$

which is greater than $\frac{2}{3} - 4\delta^{-2}(\delta^2/12)$ on the set $\{|Z(t)| > \delta\}$. Start from the inequality

$$\begin{aligned} |Z(b)| &\leq |Z(b) - Z(t)| + |Z(t)| \\ &\leq \tfrac{3}{2}|Z(t)|\{|Z(b) - Z(t)| \leq \tfrac{1}{2}|Z(t)|\} \\ &\quad + 3|Z(b) - Z(t)|\{|Z(b) - Z(t)| > \tfrac{1}{2}|Z(t)|\}. \end{aligned}$$

Keeping in mind that the absolute value of a martingale is a submartingale, take conditional expectations given $\mathscr{E}_t$.

$$\begin{aligned}|Z(t)| &\le \mathbb{P}_t|Z(b)| \\ &\le \tfrac{3}{2}|Z(t)|\mathbb{P}_t\{|Z(b) - Z(t)| \le \tfrac{1}{2}|Z(t)|\} \\ &\quad + 3\mathbb{P}_t|Z(b) - Z(t)|\{|Z(b) - Z(t)| > \tfrac{1}{2}|Z(t)|\} \\ &\le \tfrac{3}{2}|Z(t)|\mathbb{P}_t\{|Z(b) - Z(t)| \le \tfrac{1}{2}|Z(t)|\} \\ &\quad + 6|Z(t)|^{-1}\mathbb{P}_t|Z(b) - Z(t)|^2.\end{aligned}$$

On $\{|Z(t)| > \delta\}$ we may divide through by $\tfrac{3}{2}|Z(t)|$ to get the bound (10). □

Martingales with cadlag sample paths define random elements of the space $D[0, \infty)$ under its projection σ-field. As we shall be concerned only with convergence in distribution to limit processes having continuous sample paths, we equip $D[0, \infty)$ with its metric for uniform convergence on compacta (Section V.5).

The main theorem will give conditions for a sequence of L^2-martingales to converge in distribution to a stretched-out brownian motion B_H, the process constructed from brownian motion by applying a continuous, strictly increasing transformation $H(\cdot)$ to the time scale: $B_H(t) = B(H(t))$. This new gaussian process is an L^2-martingale with conditional variance process $H(t)$, provided $H(0) = 0$.

One hypothesis of the theorem will be pointwise convergence of the conditional variance processes to H. It is analogous to the hypothesis in Theorem 1 that the sum of conditional variances converges to one. As in the proof of that theorem, we shall use the hypothesis to introduce stopping times for the martingales. For fullest generality we must draw upon the Strasbourg theory for the properties of predictable stopping times; for the special case of martingales with conditional variance processes as in (7), the argument from Theorem 1 would suffice.

11 Lemma. *Let $\{X_n\}$ be a sequence of L^2-martingales whose conditional variance processes converge pointwise to a fixed, continuous, increasing function: $V_n(t) \to H(t)$ in probability, for each fixed t. Then there exists a sequence of stopping times $\{\sigma_n\}$ and constants $\{\varepsilon_n\}$, with $\sigma_n \to \infty$ in probability and $\varepsilon_n \downarrow 0$, such that* $\sup_t |V_n(t \wedge \sigma_n) - H(t \wedge \sigma_n)| \le \varepsilon_n$ *almost surely.*

Proof. First fix an $\varepsilon > 0$, and define a stopping time

$$\tau_n = \inf\{t > 0 : |V_n(t) - H(t)| > \varepsilon\}.$$

The hypothesis implies that $\tau_n \to \infty$ in probability: if $0 = s_0 < s_1 < \cdots < s_k$ are chosen so that $H(s_i) - H(s_{i-1}) < \tfrac{1}{2}\varepsilon$ for each i, then (by monotonicity of V_n and H)

$$\mathbb{P}\{\tau_n < s_k\} \le \mathbb{P}\{|V_n(s_i) - H(s_i)| > \tfrac{1}{2}\varepsilon \text{ for some } i\} \to 0.$$

From the definition of τ_n,

(12) $$|V_n(t) - H(t)| \le \varepsilon \quad \text{for} \quad t < \tau_n.$$

Only the time $t = \tau_n$ might cause trouble; V_n might have a jump there. That is where predictability comes to the rescue.

There exists an increasing sequence of stopping times $\{\tau_{nj}\}$ with $\tau_{nj} < \tau_n$ and $\tau_{nj} \uparrow \tau_n$ almost surely (Dellacherie and Meyer 1978, IV.69, IV.77), because predictability allows us to peer just a little distance into the future for V_n. Replace τ_n by $\tau_{n,j(n)}$ for some $j(n)$ such that $\tau_{n,j(n)} \to \infty$ in probability. This lops off the troublesome point $t = \tau_n$ in (12); the inequality holds for every t in the range $[0, \tau_{n,j(n)}]$.

If the argument works for fixed ε it must also work for a sequence $\{\varepsilon_n\}$ decreasing slowly enough. Write $\sigma_n(\varepsilon)$ for the stopping time $\tau_{n,j(n)}$ just identified. There exist integers $n(1) < n(2) < \cdots$ such that

$$\mathbb{P}\{\sigma_n(k^{-1}) \ge k\} < k^{-1} \quad \text{for} \quad n \ge n(k).$$

Set $\varepsilon_n = k^{-1}$ and $\sigma_n = \sigma_n(k^{-1})$ when $n(k) \le n < n(k+1)$. □

Stopping time tricks like the one used in this proof pop up all over the place in martingale limit theory. Sample path properties that hold with probability tending to one can often be made to hold with probability one by enforcing an appropriate stopping rule.

Something must be added to the convergence of conditional variance processes. Otherwise we could specialize the result to processes with independent increments, obtaining as a by-product the Central Limit Theorem for sums of independent random variables without having to impose anything like a Lindeberg condition. The extra something takes the form of a constraint on the maximum jump in the sample path. Define the jump functional J_T on $D[0, \infty)$ by:

$$J_T(x) = \max\{|x(s) - x(s-)|: 0 \le s \le T\}.$$

It is both continuous and measurable (Problem 4). If $X_n \rightsquigarrow B_H$ then certainly $J_T(X_n) \rightsquigarrow J_T(B_H) = 0$; convergence in probability of $\{J_T(X_n)\}$ to zero is a necessary condition. The theorem assumes just a little bit more to get a sufficient condition.

13 Theorem. *Let $\{X_n\}$ be a sequence of L^2-martingales with conditional variance processes $\{V_n\}$. Let H be a continuous, increasing function on $[0, \infty)$ with $H(0) = 0$. Sufficient conditions for convergence in distribution of $\{X_n\}$, as random elements of $D[0, \infty)$, to the stretched-out brownian motion B_H are:*

(i) $X_n(0) \to 0$ *in probability;*
(ii) $V_n(t) \to H(t)$ *in probability, for each fixed t;*
(iii) $\mathbb{P}J_k(X_n)^2 \to 0$ *for each fixed k, as $n \to \infty$.*

PROOF. By virtue of Theorem V.23, we have only to prove convergence in distribution for the truncations of the processes to each compact interval $[0, T]$. The argument for the typical case $T = 1$ will suffice. According to Theorem V.3 we shall need to establish

(a) *Fidi Convergence*: the fidis of the $\{X_n\}$ converge to the fidis of B_H
(b) *The Grid Condition*: to each $\varepsilon > 0$ and $\delta > 0$ there corresponds a grid $0 = t_0 < t_1 < \cdots < t_m = 1$ such that

$$\limsup_n \mathbb{P}\left\{\max_j \sup_{[t_j, t_{j+1})} |X_n(t) - X_n(t_j)| > \delta\right\} < \varepsilon$$

Theorem 1 will take care of (a); Lemma 9, applied to the X_n processes stopped appropriately, will take care of (b).

By Lemma 11 there exist stopping times $\{\sigma_n\}$ and constants $\{\varepsilon_n\}$, with $\sigma_n \to \infty$ in probability and $\varepsilon_n \downarrow 0$, such that $|V_n(t \wedge \sigma_n) - H(t \wedge \sigma_n)| \leq \varepsilon_n$ almost surely. We may assume that X_n has at most one jump of size greater than ε_n up to time σ_n. Formally, we would replace $\{\varepsilon_n\}$ by a more slowly decreasing sequence such that

$$\mathbb{P}\{J_{k(n)}(X_n) > \varepsilon_n\} \to 0 \quad \text{as} \quad n \to \infty$$

for some slowly diverging sequence $\{k(n)\}$. Such sequences exist because of (iii). Then we would replace σ_n by

$$\inf\{t \leq \sigma_n : |X_n(t) - X_n(t-)| > \varepsilon_n\}.$$

These modifications would not disturb the other properties of $\{\sigma_n\}$ and $\{\varepsilon_n\}$.

The stopped martingale $X_n^*(t) = X_n(t \wedge \sigma_n)$ has conditional variance process $V_n^*(\cdot) = V_n(\cdot \wedge \sigma_n)$. It enjoys (i), (ii), and (iii) in strengthened form:

(i)′ $X_n^*(0) = X_n(0) \to 0$ in probability;
(ii)′ $|V_n^*(t) - H(t \wedge \sigma_n)| \leq \varepsilon_n$ for all t;
(iii)′ X_n^* has at most one jump of size $> \varepsilon_n$ and $\mathbb{P}J_1(X_n^*)^2 \to 0$.

These will make it easy to prove that $X_n^* \rightsquigarrow B_H$. The required convergence of the truncation of X_n to $[0, 1]$ will then follow because

$$\mathbb{P}\{X_n(t) = X_n^*(t) \text{ for } 0 \leq t \leq 1\} \geq \mathbb{P}\{\sigma_n \geq 1\} \to 1.$$

Simplify the notation by dropping the star.

Fidi Convergence

Let us prove only that $X_n(1) \rightsquigarrow N(0, H(1))$. Problem 6 extends the result to higher-dimensional fidis. Because of (i)′, we can do this by breaking $X_n(1) - X_n(0)$ into a sum of martingale differences that satisfy the conditions of Theorem 1.

Focus for the moment on a fixed X_n by setting $Z(t) = X_n(t) - X_n(0)$ and writing V instead of V_n for its conditional variance process. Break $Z(1)$ into a sum of increments $Z(\tau_j) - Z(\tau_{j-1})$, with $0 = \tau_0 < \tau_1 < \cdots$ a sequence of stopping times defined inductively by

$$\tau_{j+1} = \inf\{t > \tau_j : |Z(t) - Z(\tau_j)| \geq \varepsilon_n\} \wedge 1 \wedge (\tau_j + \delta_n).$$

Choose $\{\delta_n\}$ so that $|H(t + \delta_n) - H(t)| \leq \varepsilon_n$ for every t in $[0, 1]$. Denote expectations given $\mathscr{E}_{\tau_j}$ by $\mathbb{P}_j(\cdot)$; write $\Delta_j f$ for the increment $f(\tau_j) - f(\tau_{j-1})$ of any function f between successive stopping times. Then

$$Z(1) = \sum_{j=1}^{\infty} \Delta_j Z.$$

Along any particular sample path of Z all except finitely many of these increments equal zero; there is no problem with convergence of the sum.

Check the conditions of Theorem 1 for the martingale differences $\{\Delta_j Z\}$. Strictly speaking the theorem applies only to triangular arrays with a finite number of variables in each row, but there may be infinitely many $\Delta_j Z$ increments. We would need to apply the theorem to a finite sum of $\Delta_j Z$, for $1 \leq j \leq j(n)$, with $j(n)$ chosen so that both

$$\sum_{j(n)}^{\infty} \Delta_j(Z) \quad \text{and} \quad \sum_{j(n)}^{\infty} \mathbb{P}_{j-1}(\Delta_j Z)$$

converge to zero in probability.

By property (8) of the conditional variance process,

$$\sum_j \mathbb{P}_{j-1}(\Delta_j Z)^2 = \sum_j \mathbb{P}_{j-1}(\Delta_j V).$$

Informally speaking, predictability of V almost makes $\Delta_j V$ measurable with respect to $\mathscr{E}_{\tau_{j-1}}$; the last sum almost equals $\sum_j \Delta_j V$, which we know will converge in probability to $H(1)$ as $n \to \infty$. Formally, $\sum_j (\Delta_j V - \mathbb{P}_{j-1}\Delta_j V)$ is a sum of martingale differences with zero mean and variance less than

$$\begin{aligned}
&\sum_j \mathbb{P}(\Delta_j V)^2 \quad \text{because the cross-product terms vanish} \\
&\leq \sum_j \mathbb{P}[(2\varepsilon_n + \Delta_j H(\cdot \wedge \sigma_n))\Delta_j V] \quad \text{by (ii)}' \\
&\leq 3\varepsilon_n \mathbb{P}\left(\sum_j \Delta_j V\right) \quad \text{by the choice of } \{\delta_n\} \\
&\leq 3\varepsilon_n(\varepsilon_n + H(1)) \\
&\to 0 \quad \text{as} \quad n \to \infty.
\end{aligned}$$

That takes care of condition (i) of Theorem 1.

For the Lindeberg condition it suffices to check the stronger L^1 convergence.

$$\begin{aligned}
\mathbb{P}\sum_j \mathbb{P}_{j-1}(\Delta_j Z)^2\{|\Delta_j Z| > \varepsilon\} &= \mathbb{P}\sum_j (\Delta_j Z)^2\{|\Delta_j Z| > \varepsilon\} \\
&\leq \mathbb{P}(\varepsilon_n + J_1(X_n))^2 \quad \text{if } \varepsilon > 2\varepsilon_n
\end{aligned}$$

The inequality follows, by the definition of τ_j, from:

$$\begin{aligned}|Z(\tau_j) - Z(\tau_{j-1})| &\le |Z(\tau_j) - Z(\tau_j -)| + |Z(\tau_j -) - Z(\tau_{j-1})| \\ &\le |\text{jump at } \tau_j| + \varepsilon_n\end{aligned}$$

At most one increment $\Delta_j Z$ can exceed $2\varepsilon_n$ in absolute value, and that happens only if Z has its one jump greater than ε_n at τ_j. An appeal to the second part of (iii)′ completes the proof of the fidi convergence.

The Grid Condition

Choose $0 = t_0 < \cdots < t_m = 1$ so that $H(t_{j+1}) - H(t_j) \le \delta^2/24$ for each j. For n large enough to make $2\varepsilon_n \le \delta^2/24$, the strengthened condition (ii)′ implies

(14) $$\mathbb{P}(V_n(t_{j+1}) - V_n(t) \mid \mathscr{E}_t) \le \delta^2/12 \quad \text{almost surely}$$

if $t_j \le t < t_{j+1}$. Invoke Lemma 9.

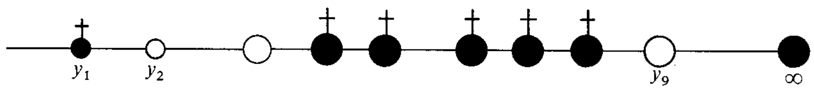
(15) $$\begin{aligned}&\limsup_n \mathbb{P}\left\{\max_j \sup_{[t_j, t_{j+1})} |X_n(t) - X_n(t_j)| > \delta\right\} \\ &\le \sum_{j=0}^{m-1} \limsup_n 3\mathbb{P}\{|X_n(t_{j+1}) - X_n(t_j)| \ge \tfrac{1}{2}\delta\} \\ &\le \sum_{j=0}^{m-1} 3\mathbb{P}\{|B_H(t_{j+1}) - B_H(t_j)| \ge \tfrac{1}{2}\delta\} \quad \text{by fidi convergence.} \\ &= \sum_{j=0}^{m-1} 3\mathbb{P}\{|N(0, H(t_{j+1}) - H(t_j))| \ge \tfrac{1}{2}\delta\} \\ &\le 48\delta^{-4} \sum_{j=0}^{m-1} [H(t_{j+1}) - H(t_j)]^2 \mathbb{P}|N(0,1)|^4 \\ &\le 48\delta^{-4} \max_j [H(t_{j+1}) - H(t_j)] H(1) \mathbb{P}|N(0,1)|^4\end{aligned}$$

which is less than ε if the grid points are close enough together. □

VIII.3. Estimation from Censored Data

The empirical distribution function based on an independent sample $\xi_1, \ldots, \xi_n$ from P is a natural estimator for the distribution function of P. How should one modify it when the observations are subject to censorship? One possibility, the Kaplan–Meier estimator, can be analyzed by the martingale methods from the previous section. What follows is a heuristic account. The notes to the chapter will point you towards more rigorous treatments, which draw on results from the theory of stochastic integration.

Consider the simplest model for censorship. Independent variables $c_1, \dots, c_n$, drawn from a censoring distribution C, cut short the natural lifetime ξ_i; we observe the value $y_i = \xi_i \wedge c_i$. Each y_i has distribution Q, where $Q(s, \infty) = P(s, \infty)C(s, \infty)$. In addition, we can tell whether the y_i represents natural death, $\xi_i < c_i$, or a case of death by censorship, $\xi_i \geq c_i$.



To construct the Kaplan–Meier empirical measure K_n, start from the usual empirical measure Q_n for the observations $y_1, \dots, y_n$. Working from the left-hand end, distribute successively all the mass from each censored point equally amongst all the $\{y_i\}$ lying to its right. In the situation pictured, y_1 keeps its mass $\frac{1}{9}$; then y_2 surrenders mass $\frac{1}{7} \times \frac{1}{9}$ to each of its seven successors $y_3, \dots, y_9$; then y_3 surrenders $\frac{1}{6} \times (\frac{1}{9} + \frac{1}{7} \times \frac{1}{9})$ to each of $y_4, \dots, y_9$; and so on. At the last point, y_9, there are no more $\{y_i\}$ to inherit its mass, so dump it all down on a fictitious super-survivor out at $+\infty$. If y_9 had not been censored, it would have kept all its inherited mass. In any case, K_n will distribute its total mass of one amongst the naturally deceased $\{y_i\}$, with maybe a little bit on $+\infty$. Notice that $K_n[0, t] \leq Q_n[0, t]$ for each t.

Make the analysis as simple as possible by assuming that both P and C are continuous distributions living on $[0, \infty)$, with neither concentrated on a finite interval. Write $\mathscr{E}_t$ for the σ-field corresponding to everything we learn up to time t about which ξ_i have died or been censored. Write $\mathbb{P}_t(\cdot)$ for $\mathbb{P}(\cdot \mid \mathscr{E}_t)$.

Calculate (to first-order terms) the conditional distribution of the increment $\Delta K_n = K_n(t, t + h]$ given $\mathscr{E}_t$, for tiny positive h. From $\mathscr{E}_t$ we learn the value $Q_n[0, t]$. Define m to be $nQ_n[0, t]$. The remaining $n - m$ observations in (t, ∞) are generated by choosing each ξ_i from the conditional distribution $P(\cdot \mid \xi > t)$, then censoring it by a c_i chosen from $C(\cdot \mid c > t)$. Write ΔP for $P(t, t + h]$ and ΔK_n for $K_n(t, t + h]$. To first order, each of the $n - m$ observations has conditional probability $\Delta P / P(t, \infty)$ of registering a natural death during the interval $(t, t + h]$. A single such observation would receive a fraction $(n - m)^{-1}$ of the K_n measure for $(t, \infty]$. Thus, to first order,

$$\mathbb{P}_t \Delta K_n = (n - m)^{-1} K_n(t, \infty](n - m)\Delta P / P(t, \infty).$$

This suggests that

$$\mathbb{P}_t \Delta K_n / K_n(t, \infty] = \Delta P / P(t, \infty) + \text{lower order terms},$$

which would lead us to believe that $\log K_n(t, \infty] - \log P(t, \infty)$ is a continuous-time martingale for each n.

An attempt to add rigor to the first-order analysis would reveal a few illegal divisions by zero. There is a positive probability that either $n - m$ or $K_n(t, \infty]$ could equal zero. A suitable stopping time can save us from embarrassment. Let $\{\alpha_n\}$ be a sequence of positive numbers converging to zero. Make sure that $n\alpha_n$ is an integer. Define ρ_n as the first t for which $Q_n(t, \infty)$ equals α_n. Then certainly

$$K_n(t \wedge \rho_n, \infty] \geq Q_n(t \wedge \rho_n, \infty) \geq \alpha_n > 0 \quad \text{for all } t.$$

The first-order analysis could be made rigorous enough to show that the process

$$X_n(t) = \log K_n(t \wedge \rho_n, \infty] - \log P(t \wedge \rho_n, \infty)$$

is a continuous-time martingale for each n.

On the set $\{\rho_n > t\}$, the increment in V_n, the conditional variance process of X_n, would be

$$\mathbb{P}_t(\Delta X_n)^2 = (n - m)^{-2}(n - m)\Delta P/P(t, \infty) + \text{smaller order terms}$$

On $\{\rho_n \leq t\}$ the increment would be zero. Recover V_n as a limit of sums of conditional variances (see Problem 5).

$$V_n(t) = n^{-1} \int \{0 \leq s \leq t \wedge \rho_n\} Q_n(s, \infty)^{-1} P(s, \infty)^{-1} P(ds).$$

By definition of ρ_n,

$$V_n(t) \leq (n\alpha_n)^{-1} \int_0^t P(s, \infty)^{-1} P(ds).$$

Thus $\{V_n\}$ converges in probability to zero uniformly over compact sets.

Apply Lemma 9. For b fixed and n large enough,

$$\begin{aligned} \mathbb{P}\left\{\sup_{t \leq b} |X_n(t)| > \delta\right\} &\leq 3\mathbb{P}\{|X_n(b)| > \tfrac{1}{2}\delta\} \\ &\leq 12\delta^{-2}\, \mathbb{P}X_n(b)^2 \\ &= 12\delta^{-2}\, \mathbb{P}V_n(b) \\ &\to 0. \end{aligned}$$

That is, for each fixed b,

$$\sup_{t \leq b} |\log K_n(t \wedge \rho_n, \infty] - \log P(t \wedge \rho_n, \infty)| \to 0 \text{ in probability.}$$

Because both K_n and P are probability measures, and because $\rho_n \to \infty$ in probability, it follows that

(16) $$\sup_t |K_n[0, t] - P[0, t]| \to 0 \quad \text{in probability.}$$

The Kaplan–Meier measure estimates P consistently, in the sense of uniform convergence of distribution functions.

Now consider the normalized martingale $n^{1/2}X_n(t)$. It has conditional variance process nV_n, which converges in probability to the continuous function

$$H(t) = \int \{0 \leq s \leq t\} Q(s, \infty)^{-1} P(s, \infty)^{-1} P(ds).$$

That suggests $\{n^{1/2}X_n\}$ converges in distribution to a brownian motion stretched out by H. Theorem 13 proves it.

We already have the first two requirements of the theorem holding. What about the maximum jump of $n^{1/2}X_n$ on an interval $[0, t]$? It can only occur at one of the $\{y_i\}$ that died naturally. In the worst case, where all except the largest y_i in $[0, t]$ are censored, the measure K_n puts mass $(nQ_n(t \wedge \rho_n, \infty))^{-1}$ at that largest y_i. The corresponding jump J in $n^{1/2}X_n$ cannot exceed $n^{1/2}(n\alpha_n)^{-1}\alpha_n^{-1}$, which converges to zero provided $\alpha_n \gg n^{-1/4}$. That is more than enough to force $\mathbb{P}J^2 \to 0$ as $n \to \infty$. Thus $n^{1/2}X_n \rightsquigarrow B_H$.

The effect of the stopping wears off as $n \to \infty$, because $\rho_n \to \infty$ in probability. So we get from the convergence result for $\{X_n\}$ that

$$n^{1/2} \log[K_n(t, \infty]/P(t, \infty)] \rightsquigarrow B_H$$

as random elements of $D[0, \infty)$. Rewrite the left-hand side as

$$\begin{aligned} &n^{1/2} \log[1 + (P[0, t] - K_n[0, t])/P(t, \infty)] \\ &\quad = - n^{1/2}(K_n[0, t] - P[0, t])(1 + o_p(1))/P(t, \infty) \end{aligned}$$

with the $o_p(1)$ converging in probability to zero uniformly over compact intervals, by virtue of (16). Deduce that

$$n^{1/2}(K_n[0, t] - P[0, t])/P(t, \infty) \rightsquigarrow B_H,$$

or, equivalently,

$$n^{1/2}(K_n[0, t] - P[0, t]) \rightsquigarrow P(t, \infty)B_H.$$

The covariances of the gaussian process $P(\cdot, \infty)B_H(\cdot)$ depend on both P and C.

Notes

Martingales have been one of the hottest topics on the weak convergence market for some years now, partly because of a boom in stochastic integral stocks. For discrete-time martingales key papers are Brown (1971), McLeish (1974), and Aldous (1978). Both McLeish and Aldous obtained results sharper than those in Sections 1 and 2. Hall and Heyde (1980, Section 6.4) gave a much fancier version of Example 4. The literature on continuous time martingales has grown rapidly; the Shiryayev (1981) survey, Métivier (1982), the two-volume monograph by Liptser and Shiryayev (1977, 1978), and Gihman and Skorohod (1979), are accessible. Méyer (1966), and Dellacherie and Meyer (1978, 1982), make harder reading. Dellacherie (1972) contains much about subtle σ-fields.

What I called the conditional variance process usually gets written as $\langle Z, Z\rangle$. The more widely definable quadratic variation process $[Z, Z]$ has taken over the role I gave to $\langle Z, Z\rangle$ in Section 2. In the discrete-time theory, $[Z, Z]$ would be the sum of squared increments; $\langle Z, Z\rangle$ would be the sum of conditional variances.

Martingales and stochastic integrals simplify the theory for estimation of distribution functions from censored data. Gill (1980) covered this in some detail; Jacobsen (1982) explained the essential prerequisites from the Strasbourg theory of processes. Brémaud (1981) gave applications to point processes. The martingale proof of the central limit theorem for Kaplan–Meier estimators makes an interesting comparison with a more traditional weak convergence approach of Breslow and Crowley (1974).

Problems

[1] For y real, define $H_n(y) = e^{iy} - 1 - iy - \cdots - (iy)^n/n!$. Prove that $|H_n(y)| \leq |y|^{n+1}/(n+1)!$. [Proceed inductively, using $i\int_0^t H_n(s)\,ds = H_{n+1}(t)$ for $t > 0$.] Deduce that $|e^{iy} - 1 - iy + \frac{1}{2}y^2| \leq \min\{\frac{1}{2}y^2 + \frac{1}{2}y^2, \frac{1}{6}|y|^3\}$.

[2] Prove uniform integrability of the squared autoregressive process $\{X_j^2\}$ from Example 4. Break u_j into a sum $v_j + w_j$, where $v_j = u_j\{|u_j| \leq M\} - \mathbb{P}u_j\{|u_j| \leq M\}$ for some large truncation level M. Define

$$Y_n = v_n + \theta_0 v_{n-1} + \cdots + \theta_0^{n-1} v_1,$$

$$Z_n = w_n + \theta_0 w_{n-1} + \cdots + \theta_0^{n-1} w_1.$$

Show that $X_n = Y_n + Z_n + \theta_0^n X_0$. The sequence $\{Y_n^2\}$ is uniformly bounded. Because $\mathbb{P}Z_n^2 = \mathbb{P}w_n^2 + \cdots + \theta_0^{2n-2}\mathbb{P}w_1^2$, the M can be chosen to make $\mathbb{P}Z_n^2 < \varepsilon$ for every n. [A similar argument would work for $\{u_j\}$ a sequence of martingale differences. Solution provided by Kai Fun Yu.]

[3] For the uniform empirical process U_n prove that $\{U_n(t)/(1-t) : 0 \leq t < 1\}$ is a martingale with respect to an appropriate family of σ-fields. Bound $\mathbb{P}\{\sup_{t\leq b}|U_n(t)| > \delta\}$ for small enough b, then deduce the Empirical Central Limit Theorem. Generalize to E_n. [Over $[a, b]$ take $\mathscr{E}_t$ as the σ-field determined by the locations of sample points in $[a, t]$, for $a \leq t \leq b$.]

[4] The jump functional is projection measurable; $J_T(x)$ is the limit of $\max_i |x(t_i) - x(t_{i-1})|$ as the grid $0 = t_0 < \cdots < t_m = T$ is refined down to a countable dense subset of $[0, T]$.

[5] Let Z be an L^2 martingale on $[0, 1]$ with conditional variance process V. If V has continuous sample paths, adapt the arguments in the *Fidi Convergence* part of the proof of Theorem 13 to show that

$$\sum_i \mathbb{P}([Z(t_i) - Z(t_{i-1})]^2 | \mathscr{E}_{t_{i-1}}) \to V(1) \quad \text{in probability},$$

as the maximum spacing in the grid $0 = t_0 < \cdots < t_m = 1$ tends to zero. [The intuitive interpretation of V as a limit of a sum of conditional variances of increments is close to the mark.]

[6] Suppose that martingales $\{X_n\}$ satisfy the conditions of Theorem 13. Prove that their fidis converge to the fidis of B_H. [If $0 \leq t_1 < \cdots < t_k \leq 1$, then $Y_n(t) = \sum_{j=1}^k \alpha_j X_n(t \wedge t_j)$ is a martingale with $Y_n(1) = \sum_j \alpha_j X_n(t_j)$.]

[7] Using the notation from Section 3, show that

$$Z_n(t) = K_n[0, t \wedge \rho_n] - \int \{0 \le s \le t \wedge \rho_n\} K_n(s \wedge \rho_n, \infty] P(s, \infty)^{-1} P(ds)$$

is a martingale. Deduce from the consistency result for K_n that $\{n^{1/2} Z_n\}$ converges in distribution to a stretched-out gaussian process. Reprove the central limit theorem for the Kaplan–Meier estimator.

[8] Let $\{X_n\}$ be a sequence of L^2-martingales with conditional variance processes $\{V_n\}$. Suppose the fidis converge to the fidis of a stretched out brownian motion B_H, with H continuous. If $V_n(t) \to H(t)$ in probability for each fixed t, then $X_n \rightsquigarrow B_H$. [If ρ_n is a stopping time taking only finitely many different values, then

$$\mathbb{P}[X_n(\rho_n + \delta_n) - X_n(\rho_n)]^2 = \mathbb{P}[V_n(\rho_n + \delta_n) - V_n(\rho_n)].$$

Check Aldous's condition (VI.13) for suitably stopped processes. Use Lemma 11.]

APPENDIX A

Stochastic-Order Symbols

Mann and Wald (1943) defined the symbols $O_p(\cdot)$ and $o_p(\cdot)$, the stochastic analogues of $O(\cdot)$ and $o(\cdot)$, as a way of avoiding many of the messy details that bedevil asymptotic calculations in probability theory.

Let $\{X_n\}$ and $\{Y_n\}$ be sequences of random vectors. Write $X_n = O_p(Y_n)$ to mean: for each $\varepsilon > 0$ there exists a real number M such that $\mathbb{P}\{|X_n| \geq M|Y_n|\} < \varepsilon$ if n is large enough. Write $X_n = o_p(Y_n)$ to mean: $\mathbb{P}\{|X_n| \geq \varepsilon|Y_n|\} \to 0$ for each $\varepsilon > 0$. Of course X_n and Y_n must be defined on the same probability space.

Typically $\{Y_n\}$ will be non-random. For example: $X_n = o_p(1)$ is another way of writing $X_n \to 0$ in probability; a sequence of order $O_p(1)$ is said to be stochastically bounded; statistical estimators commonly come within $O_p(n^{-1/2})$ of the parameters they estimate.

With the $O_p(\cdot)$, $o_p(\cdot)$ notation, Problem III.15 (the delta method) has a quick solution. Differentiability of the map H means

$$H(x) = H(x_0) + L(x - x_0) + o(x - x_0) \quad \text{near } x_0,$$

where L is a fixed $s \times k$ matrix. If $n^{1/2}(X_n - x_0) \rightsquigarrow Z$ then

$$\begin{aligned} H(X_n) &= H(x_0) + L(X_n - x_0) + o_p(X_n - x_0) \\ &= H(x_0) + L(X_n - x_0) + o_p(n^{-1/2}). \end{aligned}$$

Thus

$$n^{1/2}(H(X_n) - H(x_0)) = Ln^{1/2}(X_n - x_0) + o_p(1) \rightsquigarrow LZ.$$

A few hidden assertions here need justification. For example:

$$\begin{aligned} &n^{1/2}(X_n - x_0) \rightsquigarrow Z \quad \text{implies} \quad X_n - x_0 = O_p(n^{-1/2}); \\ &o(O_p(n^{-1/2})) = o_p(n^{-1/2}); \\ &\text{if} \quad Z_n \rightsquigarrow Z \quad \text{then} \quad LZ_n + o_p(1) \rightsquigarrow LZ. \end{aligned}$$

These conceal the asymptotic details of the proof.

For comparison, here is part of the proof of Theorem VII.5 written without the benefit of the notation. Remember that

$$\begin{aligned} f(\cdot, t) &= f(\cdot, 0) + t'\Delta(\cdot) + |t| r(\cdot, t), \\ F_n(t) &= P_n f(\cdot, t), \\ F(t) &= Pf(\cdot, t) = F(0) + \tfrac{1}{2}|t|^2 + h(t), \end{aligned}$$

where for each $\varepsilon > 0$ there exists a $\delta > 0$ such that

$$|h(t)| \le \varepsilon |t|^2 \quad \text{if} \quad |t| \le \delta.$$

By assumption,

(1) $$F_n(\tau_n) = X_n + \inf_t F_n(t),$$

where $nX_n \to 0$ in probability as $n \to \infty$. Also we know that

$$n^{1/2}(F_n(\tau_n) - F(\tau_n)) = E_n f(\cdot, 0) + \tau_n' E_n \Delta + |\tau_n| Y_n$$

or

(2) $$F_n(\tau_n) = F(\tau_n) + n^{-1/2} E_n f(\cdot, 0) + n^{-1/2} \tau_n' E_n \Delta + n^{-1/2} |\tau_n| Y_n,$$

where $Y_n = E_n r(\cdot, \tau_n) \to 0$ in probability.

From (1) and (2),

$$\begin{aligned} X_n &\ge F_n(\tau_n) - F_n(0) \\ &= F(\tau_n) + n^{-1/2}\tau_n' E_n \Delta + n^{-1/2}|\tau_n| Y_n - F(0) \\ &= \tfrac{1}{2}|\tau_n|^2 + \tfrac{1}{2}h(\tau_n) + n^{-1/2}\tau_n' E_n \Delta + n^{-1/2}|\tau_n| Y_n. \end{aligned}$$

Take absolute values. Because $X_n \ge 0$,

$$\begin{aligned} nX_n &\ge \tfrac{1}{2}n|\tau_n|^2 - \tfrac{1}{2}n|h(\tau_n)| - n^{1/2}|\tau_n||E_n\Delta| - n^{1/2}|\tau_n||Y_n| \\ &\ge \tfrac{1}{2}n(1-\varepsilon)|\tau_n|^2 - n^{1/2}|\tau_n|(|E_n\Delta| + |Y_n|) \quad \text{if} \quad |\tau_n| \le \delta \\ &\ge \tfrac{1}{2}(1-\varepsilon)(n^{1/2}|\tau_n| - W_n)^2 - \tfrac{1}{2}(1-\varepsilon) W_n^2, \end{aligned}$$

where $W_n = (1-\varepsilon)^{-1}(|E_n\Delta| + |Y_n|)$. For fixed M deduce that

$$\begin{aligned} \mathbb{P}\{(n^{1/2}|\tau_n| - W_n)^2 > 2M\} \le \mathbb{P}\{|\tau_n| \ge \delta\} &+ \mathbb{P}\{2(1-\varepsilon)^{-1} nX_n > M\} \\ &+ \mathbb{P}\{W_n^2 > M\}. \end{aligned}$$

The first and second terms converge to zero. If M is large enough, the third term is eventually less than any fixed α because ... well, because W_n is of order $O_p(1)$. The mess gets too great if we try to avoid the stochastic-order symbols any longer.

If you need further convincing read the beautiful expository article by Chernoff (1956).

APPENDIX B

Exponential Inequalities

The central limit theorem leads one to expect sums of independent random variables to behave as if they were normally distributed; tail probabilities for standardized sums can be approximated by normal tail probabilities. For the limit theorems proved in this book, we need upper bounds rather than approximations. A few such bounds are collected together here.

The tails of the normal distribution decay rapidly. For $\eta > 0$,

$$\left(\frac{1}{\eta} - \frac{1}{\eta^3}\right)\frac{\exp(-\frac{1}{2}\eta^2)}{\sqrt{2\pi}} < \mathbb{P}\{N(0, 1) \geq \eta\} < \frac{1}{\eta}\frac{\exp(-\frac{1}{2}\eta^2)}{\sqrt{2\pi}}$$

The important factor is the $\exp(-\frac{1}{2}\eta^2)$. We need a similar upper bound for the tail probabilities of a sum of independent random variables $Y_1, \ldots, Y_n$. Set $S = Y_1 + \cdots + Y_n$. For each $t > 0$,

$$(1) \qquad \mathbb{P}\{S \geq \eta\} \leq \exp(-\eta t)\mathbb{P}\exp(tS) = \exp(-\eta t)\prod_{i=1}^{n}\mathbb{P}\exp(tY_i).$$

The trick will be to find a t that makes the last product small. For the normal distribution it is easy to find the best t directly:

$$\mathbb{P}\{N(0, 1) \geq \eta\} \leq \inf_{t>0}\exp(\tfrac{1}{2}t^2 - \eta t) = \exp(-\tfrac{1}{2}\eta^2).$$

For other distributions we have to work harder. We must maneuver the moment generating function of Y_i into a tractable form that gives us some clue about which value of t to choose.

2 Hoeffding's Inequality. *Let $Y_1, Y_2, \ldots, Y_n$ be independent random variables with zero means and bounded ranges: $a_i \leq Y_i \leq b_i$. For each $\eta > 0$,*

$$\mathbb{P}\{Y_1 + \cdots + Y_n \geq \eta\} \leq \exp\left[-2\eta^2 \Big/ \sum_{i=1}^{n}(b_i - a_i)^2\right].$$

PROOF. Use convexity to bound the moment generating function of Y_i. Drop the subscript i temporarily.

$$e^{tY} \le e^{ta}(b - Y)/(b - a) + e^{tb}(Y - a)/(b - a).$$

Take expectations, remembering that $\mathbb{P}Y = 0$.

$$\mathbb{P}e^{tY} \le e^{ta}b/(b - a) - e^{tb}a/(b - a).$$

Set $\alpha = 1 - \beta = -a/(b - a)$ and $u = t(b - a)$. Note: $\alpha > 0$ because $a < 0 < b$.

$$\log \mathbb{P}e^{tY} \le \log(\beta e^{-\alpha u} + \alpha e^{\beta u}) = -\alpha u + \log(\beta + \alpha e^u).$$

Write $L(u)$ for this function of u. Differentiate twice.

$$L'(u) = -\alpha + \alpha e^u/(\beta + \alpha e^u) = -\alpha + \alpha/(\alpha + \beta e^{-u}),$$

$$L''(u) = \alpha\beta e^{-u}/(\alpha + \beta e^{-u})^2 = [\alpha/(\alpha + \beta e^{-u})][\beta e^{-u}/(\alpha + \beta e^{-u})] \le \tfrac{1}{4}.$$

The inequality is a special case of: $x(1 - x) \le \frac{1}{4}$ for $0 \le x \le 1$. Expand by Taylor's theorem.

$$L(u) = L(0) + uL'(0) + \tfrac{1}{2}y^2L''(u^*) \le \tfrac{1}{2}u^2\tfrac{1}{4} = \tfrac{1}{8}t^2(b - a)^2.$$

Apply the inequality to each Y_i, then use (1)

$$\mathbb{P}\{Y_1 + \cdots + Y_n \ge \eta\} \le \exp[-\eta t + \tfrac{1}{8}t^2 \sum_{i=1}^{n} (b_i - a_i)^2].$$

Set $t = 4\eta/\sum_i (b_i - a_i)^2$ to minimize the quadratic. □

3 Corollary. *Apply the same argument to $\{-Y_i\}$ then combine with the inequality for $\{Y_i\}$ to get a two-sided bound under the same conditions*:

$$\mathbb{P}\{|Y_1 + \cdots + Y_n| \ge \eta\} \le 2\exp\left[-2\eta^2 \Big/ \sum_{i=1}^{n} (b_i - a_i)^2\right]. \quad \square$$

In one special case the proof can be shortened slightly. If Y_i takes only two values, $\pm a_i$, each with probability $\frac{1}{2}$, then

$$\mathbb{P}e^{tY_i} = \tfrac{1}{2}[e^{ta_i} + e^{-ta_i}] = \sum_{k=0}^{\infty} (a_i t)^{2k}/(2k)! \le \exp(\tfrac{1}{2}a_i^2t^2).$$

The rest of the proof is the same as before. We only need the Hoeffding Inequality for this special case.

4 Bennett's Inequality. *Let $Y_1, \ldots, Y_n$ be independent random variables with zero means and bounded ranges: $|Y_i| \le M$. Write σ_i^2 for the variance of Y_i. Suppose $V \ge \sigma_1^2 + \cdots + \sigma_n^2$. Then for each $\eta > 0$,*

$$\mathbb{P}\{|Y_1 + \cdots + Y_n| > \eta\} \le 2\exp[-\tfrac{1}{2}\eta^2V^{-1}B(M\eta V^{-1})],$$

where $B(\lambda) = 2\lambda^{-2}[(1 + \lambda)\log(1 + \lambda) - \lambda]$ for $\lambda > 0$.

PROOF. It suffices to establish the corresponding one-sided inequality. The two-sided inequality will then follow by combining it with the companion inequality for $\{-Y_i\}$.

Bound the moment generating function of Y_i. Drop the subscript i temporarily.

$$\begin{aligned}\mathbb{P}e^{tY} &= 1 + t\mathbb{P}Y + \sum_{k=2}^{\infty}(t^k/k!)\mathbb{P}(Y^2Y^{k-2}) \\ &\leq 1 + \sum_{k=2}^{\infty}(t^k/k!)\sigma^2M^{k-2} \\ &= 1 + \sigma^2 g(t) \quad \text{where } g(t) = (e^{tM} - 1 - tM)/M^2 \\ &\leq \exp[\sigma^2 g(t)].\end{aligned}$$

From (1) deduce $\mathbb{P}\{S \geq \eta\} \leq \exp[Vg(t) - \eta t]$. Differentiate to find the minimizing value, $t = M^{-1}\log(1 + M\eta V^{-1})$, which is positive. □

The function $B(\cdot)$ is well-behaved: continuous, decreasing, and $B(0+) = 1$. When λ is large, $B(\lambda) \approx 2\lambda^{-1}\log\lambda$ in the sense that the ratio tends to one as $\lambda \to \infty$; the Bennett Inequality does not give a true exponential bound for η large compared to V/M. For smaller η it comes very close to the bound for normal tail probabilities.

Problem 2 shows that $B(\lambda) \geq (1 + \frac{1}{3}\lambda)^{-1}$ for all $\lambda > 0$. If we replace $B(\cdot)$ by this lower bound we get

$$\mathbb{P}\{|S| \geq \eta\} \leq 2\exp[-\tfrac{1}{2}\eta^2/(V + \tfrac{1}{3}M\eta)],$$

which is known as Bernstein's inequality.

NOTES

Feller (1968, Chapter VII) analyzed the tail probabilities of binomial and normal distributions—sharp results obtained by elementary methods. Bennett (1962) and Hoeffding (1963) derived and compared a number of inequalities on tail probabilities for sums of independent random variables. Dudley (1984) noted the simpler derivation of Hoeffding's Inequality when Y_i takes only values $\pm a_i$. Bernstein's inequality apparently dates from the 1920's; it appeared as Problem X.14 in Uspensky's (1937) book.

PROBLEMS

[1] For independent $N(0, 1)$-distributed random variables $Y_1, Y_2, \ldots$, show that $(\max_{j \leq n} Y_j)/(2\log n)^{1/2}$ converges in probability to one. [Write M_n for the maximum. Show

$$\mathbb{P}\{M_n \leq (2\eta\log n)^{1/2}\} = [1 - \mathbb{P}\{N(0,1) > (2\eta\log n)^{1/2}\}]^n,$$

then use the exponential inequalities for normal tails.]

[2] For the function $B(\cdot)$ appearing in Bennett's Inequality prove that

$$(1 + \tfrac{1}{3}\lambda)B(\lambda) \geq 1 \quad \text{for all} \quad \lambda > 0.$$

[Apply l'Hôpital's rule twice to reduce the left-hand side to

$$(1 + \tfrac{1}{3}\lambda^*)(1 + \lambda^*)^{-1} + \tfrac{2}{3}\log(1 + \lambda^*)$$

for some λ^* less than λ. Then use $\log(1 + \lambda^*) \geq \lambda^*/(1 + \lambda^*)$.]

[3] If a random variable Y has zero mean, finite variance σ^2, and is bounded above by a constant M, then for $t > 0$,

$$\mathbb{P}e^{tY} \leq (\sigma^2 e^{tM} + M^2 e^{-t\sigma^2/M})/(\sigma^2 + M^2).$$

[Subject to the constraints $\mathbb{P}Y = 0$ and $\mathbb{P}Y^2 \leq \sigma^2$, the value of $\mathbb{P}e^{tY}$ is maximized when $\mathbb{P}_Y$ concentrates on the two values M and $-\sigma^2/M$.] To prove this let $\phi(y)$ be the quadratic

$$e^{-t\sigma^2/M}C^{-1}(M - y)[1 + (C^{-1} + t)(y + \sigma^2/M)] + e^{tM}C^{-2}(y + \sigma^2/M)^2,$$

where $C = M + \sigma^2/M$. Check that the coefficient of y^2 is strictly positive and that ϕ satisfies

$$\phi(M) = e^{tM}, \qquad \phi(-\sigma^2/M) = e^{-t\sigma^2/M}, \qquad \phi'(-\sigma^2/M) = te^{-t\sigma^2/M}.$$

Show that $\phi(y) \geq e^{ty}$ for $y \leq M$, with equality at $y = M$ and $y = -\sigma^2/M$. [The function $h(y) = e^{-ty}\phi(y)$ has a local minimum of 1 at $-\sigma^2/M$. Also $h(M) = 1$. If $h(y^*)$ were equal to 1 for some y^* in the interval $(-\sigma^2/M, y^*)$, the quadratic $e^{ty}h'(y)$ would have three real roots: one at $-\sigma^2/M$, one in the interval $(-\sigma^2/M, y^*)$, and one in the interval (y^*, M).] The distribution $\mathbb{P}_Y$ concentrated at M and $-\sigma^2/M$ achieves equality in $\mathbb{P}e^{tY} \leq \mathbb{P}\phi(Y)$. Bennett (1962, page 42).]

[4] For the one-sided form of Bennett's Inequality one needs only zero means and $Y_i \leq M$ for each i. Reexpress the inequality from the previous problem as

$$\mathbb{P}e^{tY} \leq \exp[tM + \log f(1 + \sigma^2/M^2)],$$

where $f(y) = 1 - y^{-1} + y^{-1}e^{-tMy}$. Prove that $d^2/dy^2 \log f(y)$ equals

$$-2y^{-3}e^{-tMy}[e^{tMy} - 1 - tMy - \tfrac{1}{2}(tMy)^2]/f(y) - [f'(y)/f(y)]^2$$

which is less than zero for $y \geq 1$. Deduce from a Taylor expansion to quadratic terms that

$$\log f(y) \leq \log f(1) + (y - 1)f'(1)/f(1) \quad \text{for} \quad y \geq 1,$$

whence $\mathbb{P}e^{ty} \leq \exp[\sigma^2(e^{tM} - 1 - tM)/M^2]$. Complete the argument as before. [Hoeffding (1963, page 24).]

APPENDIX C
Measurability

The defining properties of σ-fields ensure that the usual countable operations in probability theory—countable unions and intersections, pointwise limits of sequences, and the like—cause no measurability difficulties. In Chapters II and VII, however, we needed to take suprema over uncountable families of measurable functions. The possibility of a non-measurable supremum was brushed aside by an assurance that a regularity condition, dubbed permissibility, would take care of everything. This appendix will supply the missing details.

The discussion will take as axiomatic certain properties of analytic sets. A complete treatment may be found in Sections III.1 to III.20, III.27 to III.33, and III.44 to 45, of Dellacherie and Meyer (1978). Square brackets containing the initials DM followed by a number will point to the section of that book where you can find the justification for any unproved assertions.

Suppose M is a set equipped with a σ-field $\mathscr{M}$. The analytic ($\mathscr{M}$-analytic in DM terminology) subsets of M form a class slightly larger than $\mathscr{M}$. Denote it by $\mathscr{A}(M)$. If $\mathscr{M}$ is complete for some probability measure μ, (that is, $\mathscr{M}$ contains all the sets of zero μ measure) then $\mathscr{A}(M) = \mathscr{M}$ [DM 33]. For example, the analytic subsets generated by $\mathscr{B}[0, 1]$ contain that σ-field properly, but the σ-field of lebesgue measurable subsets of [0, 1] coincides with its analytic sets. You see from this example that we should be writing $\mathscr{A}(\mathscr{M})$ rather than $\mathscr{A}(M)$. The ambiguity is not serious when M is equipped with only one σ-field. For metric spaces, we will always choose $\mathscr{M}$ to be the borel σ-field; for product spaces, it will always be the product σ-field.

We considered empirical processes indexed by a class of functions. Formally, $\xi_1, \xi_2, \ldots$ were measurable maps from a probability space $(\Omega, \mathscr{E}, \mathbb{P})$ into a set S equipped with a σ-field $\mathscr{S}$. A class $\mathscr{F}$ of $\mathscr{S}/\mathscr{B}(\mathbb{R})$-measurable, real-valued functions on S was given. The empirical measure P_n attached to each f in $\mathscr{F}$ the real number

$$P_n f = n^{-1} \sum_{i=1}^{n} f(\xi_i(\omega)).$$

We were assuming measurability of functions of ω such as $\sup_{\mathscr{F}} |P_n f - Pf|$. Let us now consider more carefully the dependence on ω, which we emphasize by writing $P_n(\omega, \cdot)$ instead of $P_n(\cdot)$.

Permissible Classes

Suppose that the class $\mathscr{F}$ is indexed by a parameter t that ranges over some set T. That is, $\mathscr{F} = \{f(\cdot, t): t \in T\}$. We lose no generality by assuming $\mathscr{F}$ so indexed; T could be $\mathscr{F}$ itself. When more convenient, write f_t instead of $f(\cdot, t)$.

Assume T is a separable metric space. The metric on T will be important only insofar as it determines the borel σ-field $\mathscr{B}(T)$ on T.

1 Definition. Call the class $\mathscr{F}$ permissible if it can be indexed by a T in such a way that

(i) the function $f(\cdot, \cdot)$ is $\mathscr{S} \otimes \mathscr{B}(T)$-measurable as a function from $S \otimes T$ into the real line;
(ii) T is an analytic subset of a compact metric space $\overline{T}$ (from which it inherits its metric and borel σ-field). □

Some authors call a T satisfying (ii) a Souslin measurable space [DM 16]. The usual sorts of class parametrized by borel subsets [DM 12] of an euclidean space are permissible. (Take $\overline{T}$ as the one-point compactification.) So are fancier classes such as all indicator functions of compact, convex subsets of euclidean space (Problem 2).

Assume from now on that $\mathscr{F}$ is permissible and that $(\Omega, \mathscr{E}, \mathbb{P})$ is complete.

Here are the properties of analytic sets that make the definition of permissibility a good one for empirical process applications. For every measurable space $(M, \mathscr{M})$,

(a) $\mathscr{A}(M \otimes T)$ contains the product σ-field $\mathscr{M} \otimes \mathscr{B}(T)$;
(b) for each H in $\mathscr{A}(M \otimes T)$, and in particular for each $\mathscr{M} \otimes \mathscr{B}(T)$-measurable set, the projection $\pi_M H$ of H onto M is in $\mathscr{A}(M)$ [DM 13, DM 9: the set H is also in $\mathscr{A}(M \otimes \overline{T})$, because T is analytic];
(c) for each A in $\mathscr{A}(M)$ and each $\mathscr{E}/\mathscr{M}$-measurable map η from $(\Omega, \mathscr{E}, \mathbb{P})$ into M, the set $\{\eta \in A\}$ is an analytic subset of Ω [DM 11]; hence $\{\eta \in \mathscr{A}\}$ belongs to $\mathscr{E}$, because $(\Omega, \mathscr{E}, \mathbb{P})$ is complete.

From these properties we shall be able to deduce measurability for functions defined by certain uncountable operations.

Measurable Suprema

Suppose $g(\cdot, \cdot)$ is an $\mathscr{M} \otimes \mathscr{B}(T)$-measurable real function on $M \otimes T$. Set $G(m) = \sup_t g(m, t)$. Then by (a), $\mathscr{A}(M \otimes T)$ contains the set

$$H_\alpha = \{(m, t): g(m, t) > \alpha\}.$$

The projection of H_α onto M is an analytic set, by (b). It consists of all those m for which $G(m) > \alpha$. Thus $\{G > \alpha\}$ belongs to $\mathscr{A}(M)$ for each real α. If η is a measurable map from a complete probability space $(\Omega, \mathscr{E}, \mathbb{P})$ into M then, by (c), the set $\{\omega: G(\eta(\omega)) > \alpha\}$ is $\mathscr{E}$-measurable. That is, $\sup_t g(\eta(\omega), t)$ is an $\mathscr{E}$-measurable function of ω.

If $\mathscr{F}$ is permissible and if $P|f_t| < \infty$ for each t, requirement (i) of Definition 1 plus Fubini's theorem make Pf_t a measurable function of t. Apply the argument given above, with $M = S^n$, $\mathscr{M} =$ the product σ-field $\mathscr{S}^n$, $\eta =$ the vector $(\xi_1, \ldots, \xi_n)$, and

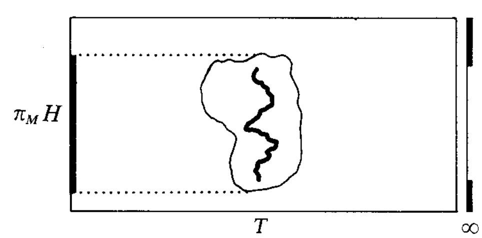
$$g(\mathbf{s}, t) = \left| n^{-1} \sum_{i=1}^{n} [f(s_i, t) - Pf_t] \right|$$

to deduce that $\sup_{\mathscr{F}} |P_n(\omega, f) - Pf|$ is a measurable function of ω.

Measurable Cross-Sections

The Symmetrization Lemma II.8 made use of another property of analytic sets. We had a stochastic process $\{Z_t: t \in T\}$; we assumed existence of a random τ for which, almost surely, $|Z_\tau| > \varepsilon$ whenever $\sup_t |Z_t| > \varepsilon$. For this we need a cross-section theorem [DM44–45]. Under requirement (ii) of Definition 1, and for any complete probability space $(M, \mathscr{M}, \mu)$,

(d) if H belongs to $\mathscr{A}(M \otimes T)$ there exists a measurable map h from M into $T \cup \{\infty\}$ (where ∞ is an ideal point added to T) such that: $(m, h(m))$ belongs to H whenever $h(m) \neq \infty$; and $h(m) \neq \infty$ for μ almost all m in the projection $\pi_M H$. Call h a measurable cross-section for H.



Write $Z(\omega, t)$ to emphasize the role of Z as an $\mathscr{E} \otimes \mathscr{B}(T)$-measurable function on $\Omega \otimes T$. Let $\{\varepsilon_j\}$ be a strictly decreasing sequence of real numbers converging to ε, with $\varepsilon_1 = \infty$. Set

$$A_j = \left\{\omega: \varepsilon_{j+1} < \sup_t |Z(\omega, t)| \leq \varepsilon_j\right\};$$

$$B_j = \{(\omega, t): \varepsilon_{j+1} < |Z(\omega, t)| \leq \varepsilon_j\}.$$

The sets $\{A_j\}$ all belong to $\mathscr{E}$. The sets $\{B_j\}$ all belong to $\mathscr{E} \otimes \mathscr{B}(T)$, and hence are analytic. Let t_0 be any fixed element of T. Choose a measurable cross-section τ_j for each B_j. Set τ equal to τ_j on A_j, and equal to t_0 outside the union of the $\{A_j\}$. Redefine $\tau(\omega)$ to be t_0 whenever (d) would set it equal to ∞. For almost all ω, if $\sup_t |Z(\omega, t)| > \varepsilon$ then $(\omega, \tau(\omega))$ belongs to B_j for some j; that is,

$$\varepsilon_{j+1} < Z(\omega, \tau(\omega)) \leq \varepsilon_j \quad \text{for some } j,$$

as required.

A formal proof of the Symmetrization Lemma II.8 is now possible. Require Z and Z' to be defined on a product space $\Omega \otimes \Omega'$ equipped with product measure $\mathbb{P} \otimes \mathbb{P}'$,

$$Z(\omega, \omega', t) = Z(\omega, t),$$

$$Z'(\omega, \omega', t) = Z'(\omega', t).$$

The τ constructed above need depend only on ω. For almost all ω,

$$\mathbb{P}'\{\omega' : |Z'(\omega', \tau(\omega))| \leq \alpha\} \geq \beta.$$

The rest of the proof goes through as before, with Fubini's theorem formalizing the conditioning argument.

Shattered Sets

Theorem II.21 placed a condition on the behavior of $V_n(\xi_1, \ldots, \xi_n)$, the smallest integer k such that $\mathscr{D}$ shatters no collection of k points from $\{\xi_1, \ldots, \xi_n\}$. Assume that the indicator functions of sets in $\mathscr{D}$ form a permissible class $\mathscr{F}$. Then V_n is measurable.

For example, here is how to prove that $\{V_n \leq 2\}$ belongs to $\mathscr{E}$. Define $g(s_1, s_2, t_1, t_2, t_3, t_4)$ as

$$\begin{aligned} &f(s_1, t_1)f(s_2, t_1) + f(s_1, t_2)[1 - f(s_2, t_2)] \\ &\quad + [1 - f(s_1, t_3)]f(s_2, t_3) + [1 - f(s_1, t_4)][1 - f(s_2, t_4)]. \end{aligned}$$

Clearly g is $\mathscr{S}^2 \otimes \mathscr{B}(T^4)$-measurable. The function

$$G(\omega) = \max_{i,j} \sup_{\mathbf{t}} g(\xi_i(\omega), \xi_j(\omega), \mathbf{t})$$

is $\mathscr{E}$-measurable. The set $\{V_n \leq 2\}$ equals $\{G < 4\}$.

Covering Numbers

Functions of the empirical measure that enter into arguments depending on Fubini's theorem demand that we take measurability seriously. But for other functions, such as the random covering numbers appearing in Sections

II.5, II.6, VI.4, and VI.6, we do not really need all the machinery of analytic sets. For example, we could just as well interpret a condition like

$$\log N_1(\delta, P_n, \mathscr{F}) = o_p(n)$$

in terms of outer measure: for each $\varepsilon > 0$,

$$\mathbb{P}^*\{\log N_1(\delta, P_n, \mathscr{F}) > n\varepsilon\} \to 0.$$

The proofs go through almost exactly as before. Equivalently, we could interpret the conditions on the covering numbers to mean $\mathbb{P}\{Z_n > n\varepsilon\} \to 0$ for some measurable random variable Z_n greater than $\log N_1(\delta, P_n, \mathscr{F})$.

For permissible classes, another solution would be to replace covering numbers by packing numbers: define $M_1(\delta, P_n, \mathscr{F})$ as the smallest m for which there exist functions $f_1, \ldots, f_m$ in $\mathscr{F}$ with $P_n|f_j - f_k| > \delta$ for $j \neq k$. This is a measurable function of ω (and even jointly measurable in ω and δ); the set

$$\{\omega: M_1(\delta, P_n(\omega, \cdot), \mathscr{F}) \geq m\}$$

equals the projection on Ω of the $\mathscr{E} \otimes \mathscr{B}(T^m)$-measurable set

$$\left\{(\omega, \mathbf{t}): \min_{j \neq k} n^{-1} \sum_{i=1}^{n} |f(\xi_i(\omega), t_j) - f(\xi_i(\omega), t_k)| > \delta\right\}.$$

The packing numbers are closely related to covering numbers:

$$M_1(2\delta, P_n, \mathscr{F}) \leq N_1(\delta, P_n, \mathscr{F}) \leq M_1(\delta, P_n, \mathscr{F}).$$

Theorems stated in terms of random covering numbers have equivalent versions for random packing numbers.

I cannot prove measurability for covering numbers of permissible classes. The set of ω where $N_1(\delta, P_n, \mathscr{F})$ strictly exceeds m is the complement of a projection of a complement of an analytic set, which apparently need not be measurable.

The Function Space $\mathscr{X}$

For the results in Chapter VII we required the class $\mathscr{F}$ to be pointwise bounded and have $\sup_{\mathscr{F}}|Pf|$ finite. The empirical process,

$$E_n(\omega, f) = n^{-1/2} \sum_{i=1}^{n} [f(\xi_i(\omega)) - Pf]$$

had bounded sample paths; the functions $E_n(\omega, \cdot)$ all belonged to the set $\mathscr{X}$ of bounded, real functions on $\mathscr{F}$. To avoid confusion, we called members of $\mathscr{X}$ functionals. We equipped $\mathscr{X}$ with the uniform norm, $\|x\| = \sup_{\mathscr{F}}|x(f)|$.

The limit processes had sample paths in the set $C(\mathscr{F}, P)$ of functionals that were uniformly continuous with respect to the $\mathscr{L}^2(P)$ seminorm ρ_P on $\mathscr{F}$. The σ-field $\mathscr{B}^P$ was smallest for which

(i) all the closed balls (for $\|\cdot\|$) belonged to $\mathscr{B}^P$;
(ii) all the finite-dimensional projections were $\mathscr{B}^P$-measurable.

We assumed in Chapter VII that E_n is $\mathscr{E}/\mathscr{B}^P$-measurable. That is true for a permissible $\mathscr{F}$ that is separable under the ρ_P seminorm.

Because each $E_n(\cdot, f)$ is a real random variable, the finite-dimensional projections create no difficulty for the $\mathscr{E}/\mathscr{B}^P$-measurability. The properties of analytic sets are needed to prove that $\{\omega: \|E_n(\omega, \cdot) - x(\cdot)\| \le r\}$ belongs to $\mathscr{E}$ whenever $x(\cdot)$ belongs to $C(\mathscr{F}, P)$. Introduce the index set T as described in Definition 1. Problem 1 shows that $x(f_t)$ is $\mathscr{B}(T)$-measurable. Equip S^n with its product σ-field $\mathscr{S}^n$. The function

$$g(\mathbf{s}, t) = \left| n^{-1/2} \sum_{i=1}^{n} [f(s_i, t) - Pf_t] - x(f_t) \right|$$

is $\mathscr{S}^n \otimes \mathscr{B}(T)$-measurable. The argument in MEASURABLE SUPREMA establishes $\mathscr{E}$-measurability of $\sup_t g(\xi(\omega), t)$, which equals $\|E_n(\omega, \cdot) - x(\cdot)\|$.

NOTES

Dudley (1978) introduced a condition, which he termed $P\in$-Suslin, as a way of handling the measurability problems for empirical processes indexed by classes of sets. Since then he has refined the definition several times. He has called the latest version (Dudley 1984) of the condition "image admissible Suslin" (via a parameter space); it is almost the same as permissibility.

Le Cam (1983) has also imposed a Suslin type of condition.

Dudley and Philipp (1983) have systematically replaced measurability assumptions by conditions framed using measurable cover functions (the idea touched on in COVERING NUMBERS—an unfortunate clash of terminology—with the Z_n random variables).

PROBLEMS

[1] Suppose $\mathscr{F}$ has a countable, dense subset $\{g_j\}$ under the $\mathscr{L}^2(P)$ seminorm ρ_P. Suppose also that $\mathscr{F}$ satisfies condition (i) of Definition 1. If $x(\cdot)$ is a bounded real functional on $\mathscr{F}$ that is ρ_P continuous, show that $x(f_t)$ is a measurable function of t. [Assume $x \ge 0$. Represent $x(f_t)$ as

$$\limsup_{n\to\infty} \sup_j x(g_j)\{r_j(t) < n^{-1}\},$$

where $r_j(t) = \rho_P(f_t - g_j)$. Use Fubini's theorem to prove measurability of r_j.]

[2] The class $\mathscr{C}$ of all non-empty, compact, convex subsets of the unit square $[0, 1]^2$ is permissible. [Equip $\mathscr{C}$ with the metric

$$d(C_1, C_2) = \inf\{\varepsilon > 0: C_1 \subseteq C_2^\varepsilon \text{ and } C_2 \subseteq C_1^\varepsilon\},$$

where C^ε denotes the set of x less than ε away from at least one point of C. Under this metric, $\mathscr{C}$ is compact (Eggleston 1969, Section 4.2). The set of (x, C) with x in C is a closed subset of $[0, 1]^2 \otimes \mathscr{C}$. Take $\mathscr{C}$ as its own indexing set.]

[3] The class of sets $\{f > 0\}$, for f running through a permissible class, is itself permissible.

References

Aldous, D. (1978). Stopping times and tightness. *Annals of Probability* **6**:335–340.

Alexander, K. S. (1984a). Probability inequalities for empirical processes and a law of the iterated logarithm. *Annals of Probability*. (Based on his 1982 Ph.D. dissertation, MIT.) To appear.

Alexander, K. S. (1984b). Rates of growth for weighted empirical processes. (The Proceedings of the Neyman–Kiefer Memorial Conference, Berkeley, 1983. Belmont, CA: Wadsworth.) To appear.

Alexandroff, A. D. (1940). Additive set functions in abstract spaces. *Mat. Sbornik* **50**(NS 8):307–342. (Chapter 1.)

Alexandroff, A. D. (1941). Additive set functions in abstract spaces. *Mat. Sbornik* **51**(NS 9):563–621. (Chapters 2 and 3.)

Alexandroff, A. D. (1943). Additive set functions in abstract spaces. *Mat. Sbornik* **55**(NS 13):169–234. (Chapters 4 and 5.)

Araujo, A. and Giné, E. (1980). *The Central Limit Theorem for Real and Banach Valued Random Variables*. New York: Wiley.

Ash, R. B. (1972). *Real Analysis and Probability*. New York: Academic Press.

Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association* **57**:33–45

Bertrand-Retali, M. (1974). Convergence uniforme d'un estimateur d'une densité de probabilité dans R^s. *Comptes Rendus de l'Académie des Sciences, Paris* **278**:451–453.

Bertrand-Retali, M. (1978). Convergence uniforme d'un estimateur de la densité par la méthode du noyau. *Revue Roumaine de Mathématiques Pures et Appliqués* **23**:361–385.

Billingsley, P. (1968). *Convergence of Probability Measures*. New York: Wiley.

Billingsley, P. (1971). *Weak Convergence of Measures: Applications in Probability*. Philadelphia: Society for Industrial and Applied Mathematics. (Regional Conference Series in Applied Mathematics # 5.)

Billingsley, P. and Topsøe, F. (1967). Uniformity in weak convergence. *Zeitschrift für Wahrscheinlichkeitstheorie und Verw. Geb.* **7**:1–16.

Bollobás, B. (1979). *Graph Theory: An Introductory Course*. New York: Springer-Verlag.

Bolthausen, E. (1978). Weak convergence of an empirical process indexed by the closed convex subsets of I^2. *Zeitschrift für Wahrscheinlichkeitstheorie und Verw. Geb.* **43**:173–181.

Breiman, L. (1968). *Probability.* Reading, MA: Addison-Wesley.

Brémaud, P. (1981). *Point Processes and Queues: Martingale Dynamics.* New York: Springer-Verlag.

Breslow, N. and Crowley, J. (1974). A large sample study of the life table and product limit estimates under random censorship. *Annals of Statistics* **2**:437–453.

Brown, B. M. (1971). Martingale central limit theorems. *Annals of Mathematical Statistics* **42**:59–66.

Brown, B. M. (1983). Statistical uses of the spatial median. *Journal of the Royal Statistical Society, Series B* **45**:25–30.

Brown, B. M. and Kildea, D. G. (1979). Outlier-detection tests and robust estimators based on signs of residuals. *Communications in Statistics—Theory and Methods* **A8**:257–269.

Chernoff, H. (1954). On the distribution of the likelihood ratio. *Annals of Mathematical Statistics* **25**:573–578.

Chernoff, H. (1956). Large sample theory: parametric case. *Annals of Mathematical Statistics* **27**:1–22.

Chibisov, D. M. (1965). An investigation of the asymptotic power of tests of fit. *Theory of Probability and Its Applications* **10**:421–437.

Chung, K. L. (1949). An estimate concerning the Kolmogorov limit distribution. *Transactions of the American Mathematical Society* **67**:36–50.

Chung, K. L. (1968). *A Course in Probability Theory.* New York: Harcourt, Brace and World.

DeHardt, J. (1971). Generalizations of the Glivenko–Cantelli theorem. *Annals of Mathematical Statistics* **42**:2050–2055.

Dellacherie, C. (1972). *Capacités et Processus Stochastiques.* Heidelberg–Berlin: Springer-Verlag.

Dellacherie, C. and Meyer, P-A. (1978). *Probabilities and Potential, Part A.* Amsterdam: North-Holland. Translation of the 1975 French second edition of the first four chapters from Meyer (1966).)

Dellacherie, C. and Meyer, P-A. (1982). *Probabilities and Potential, Part B.* Amsterdam: North-Holland.

Denby, L. and Martin, R. D. (1979). Robust estimation of the first-order autoregressive parameter. *Journal of the American Statistical Association* **74**:140–146.

Donsker, M. D. (1951). An invariance principle for certain probability limit theorems. *Memoirs of the American Mathematical Society* **6**:1–12.

Donsker, M. D. (1952). Justification and extension of Doob's heuristic approach to the Kolmogorov–Smirnov theorems. *Annals of Mathematical Statistics* **23**: 277–281.

Doob, J. L. (1949). Heuristic approach to the Kolmogorov–Smirnov theorems. *Annals of Mathematical Statistics* **20**:393–403.

Doob, J. L. (1953). *Stochastic Processes.* New York: Wiley.

Dudley, R. M. (1966a). Weak convergence of probabilities on nonseparable metric spaces and empirical measures on euclidean spaces. *Illinois Journal of Mathematics* **10**:109–126.

Dudley, R. M. (1966b). Convergence of Baire measures. *Studia Mathematica* **27**:251–268.

Dudley, R. M. (1967a). Measures on non-separable metric spaces. *Illinois Journal of Mathematics* **11**:449–453.

Dudley, R. M. (1967b). The sizes of compact subsets of Hilbert space and continuity of gaussian processes. *Journal of Functional Analysis* **1**:290–330.

Dudley, R. M. (1968). Distances of probability measures and random variables. *Annals of Mathematical Statistics* **39**:1563–1572.

Dudley, R. M. (1973). Sample functions of the gaussian process. *Annals of Probability* **1**:66–103.

Dudley, R. M. (1974). Metric entropy of some classes of sets with differentiable boundary. *Journal of Approximation Theory* **10**:227–236.

Dudley, R. M. (1976). *Probabilities and Metrics: Convergence of laws on metric spaces, with a view to statistical testing.* Aarhus: Aarhus University. (Mathematics Institute Lecture Note Series #45.)

Dudley, R. M. (1978). Central limit theorems for empirical measures. *Annals of Probability* **6**:899–929. (Correction, *ibid.* **7** (1979):909–911.)

Dudley, R. M. (1981a). Donsker classes of functions. In Csörgö, M., Dawson, D. A., Rao, J. N. K., and Salek, A. K. Md. E. (Editors), *Statistics and Related Topics*, pages 341–352. Amsterdam: North-Holland.

Dudley, R. M. (1981b). Vapnik–Cervonenkis Donsker classes of functions. In *Aspects Statistiques et Aspects Physiques des Processus Gaussiens.* CNRS, Paris, St. Flour 1980.

Dudley, R. M. (1984). *A Course on Empirical Processes.* Springer Lecture Notes in Mathematics. (Lectures given at Ecole d'Eté de Probabilités de St. Flour, 1982.) To appear.

Dudley, R. M. and Philipp, W. (1983). Invariance principles for sums of Banach space valued random elements and empirical processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verw. Geb.* **62**:509–552.

Durbin, J. (1973a). Weak convergence of the sample distribution function when parameters are estimated. *Annals of Statistics* **1**:279–290.

Durbin, J. (1973b). *Distribution Theory for Tests Based on the Sample Distribution Function.* Philadelphia: Society for Industrial and Applied Mathematics. (Regional Conference Series in Applied Mathematics #9.)

Durst, M. and Dudley, R. M. (1981). Empirical processes, Vapnik–Cervonenkis classes and Poisson processes. *Probability and Mathematical Statistics (Wrocław)* **1**:109–115.

Eggleston, H. G. (1969). *Convexity.* Cambridge: Cambridge University Press.

Erdös, P. and Kac, M. (1946). On certain limit theorems in the theory of probability. *Bulletin of the American Mathematical Society* **52**:292–302.

Erdös, P. and Kac, M. (1947). On the number of positive sums of independent random variables. *Bulletin of the American Mathematical Society* **53**:1011–1020.

Feller, W. (1968). *An Introduction to Probability Theory and Its Applications, Vol. I* (Third edition). New York: Wiley.

Feller, W. (1971). *An Introduction to Probability Theory and Its Applications, Vol. II* (Second edition). New York: Wiley.

Gaenssler, P. (1984). Empirical Processes. Institute of Mathematical Statistics Lecture Notes—Monograph Series **3**.

Gaenssler, P. and Révész, P. (Editors) (1976). *Empirical Distributions and Processes.* Springer Lecture Notes in Mathematics **566**.

Gaenssler, P. and Stute, W. (1979). Empirical processes: a survey of results for independent and identically distributed random variables. *Annals of Probability* **7**:193–243.

Gihman, I. I. and Skorohod, A. V. (1969). *Introduction to the Theory of Random Processes.* Philadelphia: Saunders. (Earlier version of the three volumes 1974, 1975, 1979.)

Gihman, I. I. and Skorohod, A. V. (1974). *The Theory of Stochastic Processes, I.* New York: Springer-Verlag.

Gihman, I. I. and Skorohod, A. V. (1975). *The Theory of Stochastic Processes, II.* New York: Springer-Verlag.

Gihman, I. I. and Skorohod, A. V. (1979). *The Theory of Stochastic Processes, III.* New York: Springer-Verlag.

Gill, R. D. (1980). *Censoring and Stochastic Integrals.* Mathematical Centre Tract **124**. Amsterdam: Mathematisch Centrum.

Giné, E. and Zinn, J. (1984). On the central limit theorem for empirical processes. *Annals. of Probability*. To appear.

Glaisher, J. W. L. (1872). On the law of facility of errors of observations, and on the method of least squares. *Memoirs of the Royal Astronomical Society* **39**:75–124.

Gnedenko, B. V. (1968). *The Theory of Probability* (Fourth edition). New York: Chelsea.

Hájek, J. (1965). Extension of the Kolmogorov–Smirnov test to regression alternatives. In J. Neyman and L. M. Le Cam (Editors), *Bernoulli, Bayes, Laplace: Anniversary Volume*, pages 45–60. New York: Springer-Verlag.

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Applications*. New York: Academic Press.

Halmos, P. R. (1969). *Measure Theory*. New York: Van Nostrand.

Hansel, G. and Troallic, J. P. (1978). Mesures marginales et théorème de Ford–Fulkerson. *Zeitschrift für Wahrscheinlichkeitstheorie und Verw. Geb.* **43**:245–251.

Hartigan, J. A. (1975). *Clustering Algorithms*. New York: Wiley.

Hartigan, J. A. (1978). Asymptotic distributions for clustering criteria. *Annals of Statistics* **6**:117–131.

Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* **58**:13–30.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Fifth Berkeley Symposium on Mathematical Statistics and Probability*, pages 221–233. Berkeley, CA: University of California.

Jacobs, K. (1978). *Measure and Integral*. New York: Academic Press.

Jacobsen, M. (1982). *Statistical Analysis of Counting Processes*. Springer Lecture Notes in Statistics **12**.

Kac, M. (1949). On deviations between theoretical and empirical distributions. *Proceedings of the National Academy of Science USA* **35**:252–257.

Kelley, J. L. (1955). *General Topology*. New York: Van Nostrand.

Kolchinsky, V. I. (1982). On the central limit theorem for empirical measures. *Theory of Probability and Mathematical Statistics (Kiev)* **24**:71–82.

Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Instituto Ital. degli Attuari* **4**:83–91.

Kolmogorov, A. N. (1956). On Skorohod convergence. *Theory of Probability and Its Applications* **1**:213–222.

Kurtz. T. G. (1981). *Approximation of Population Processes*. Philadelphia: Society for Industrial and Applied Mathematics. (Regional Conference Series in Applied Mathematics #36.)

Le Cam, L. (1957). Convergence in distribution of stochastic processes. *University of California Publications in Statistics* **2**:207–236.

Le Cam, L. (1983). A remark on empirical measures. In Bickel, P., Doksum, K., and Hodges, J. (Editors), *Festschrift for E. L. Lehmann*, pages 305–327. Belmont, CA: Wadsworth.

Lévy, P. (1922). Sur la determination des lois de probabilité par leurs fonctions caractéristiques. *Comptes Rendus de l'Académie des Sciences, Paris* **175**:854–856.

Liapounoff, A. M. (1900). Sur une proposition de la théorie des probabilités. *Bulletin de l'Académie impériale des Sciences de St. Pétersbourg* **13**:359–386.

Liapounoff, A. M. (1901). Nouvelle forme du theoreme sur la limite de probabilités. *Mémoires de l'Académie impériale des Sciences de St. Pétersbourg* **12**:Number 5.

Lindeberg, J. W. (1922). Eine neue Herleitung des Exponentialgesetzes in der Wahrscheinlichkeitsrechnung. *Mathematische Zeitschrift* **15**:211–225.

Lindvall, T. (1973). Weak convergence of probability measures and random functions in the function space $D[0, \infty)$. *Journal of Applied Probability* **10**:109–121.

Liptser, R. S. and Shiryayev, A. N. (1977). *Statistics of Random Processes, I: General Theory*. New York: Springer-Verlag.

Liptser, R. S. and Shiryayev, A. N. (1978). *Statistics of Random Processes, II: Applications.* New York: Springer-Verlag.

Loève, M. (1978). *Probability Theory, II.* (Fourth Edition) New York: Springer-Verlag.

Major, P. (1978). On the invariance principle for sums of independent identically distributed random variables. *Journal of Multivariate Analysis* **8**:487–517.

Mann, H. B. and Wald, A. (1943). On stochastic limit and order relationships. *Annals of Mathematical Statistics* **14**:217–226.

Maritz, J. S. (1981). *Distribution-Free Statistical Methods.* London: Chapman and Hall.

McKean, H. P., Jr. (1969). *Stochastic Integrals.* New York: Academic Press.

McLeish, D. L. (1974). Dependent central limit theorems and invariance principles. *Annals of Probability* **2**:620–628.

Métivier, M. (1982). *Semimartingales: A Course on Stochastic Processes.* Berlin: De-Gruyter.

Meyer, P-A. (1966). *Probability and Potential.* (First edition). Boston: Blaisdell.

Neveu, J. (1975). *Discrete-Parameter Martingales.* Amsterdam: North-Holland.

Neyman, J. (1937). 'Smooth' test for goodness of fit. *Skandinavisk Aktuarietidskrift* **20**: 149–199. (Reprinted in "A Selection of Early Statistical Papers of J. Neyman," Cambridge: Cambridge University Press, 1967.)

Oxtoby, J. C. (1971). *Measure and Category.* New York: Springer-Verlag.

Parr, W. C. (1982). Minimum distance estimation: a bibliography. *Communications in Statistics, A* **11**:1511–1518.

Parthasarathy, K. R. (1967). *Probability Measures on Metric Spaces.* New York: Academic Press.

Pickands, J. (1971). The two-dimensional Poisson process and extremal processes. *Journal of Applied Probability* **8**:745–756.

Pollard, D. (1980). The minimum distance method of testing. *Metrika* **27**:43–70.

Pollard, D. (1981a). Limit theorems for empirical processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verw. Geb.* **57**:181–195.

Pollard, D. (1981b). Strong consistency of k-means clustering. *Annals of Statistics* **9**:135–140.

Pollard, D. (1982a). Beyond the heuristic approach to Kolmogorov–Smirnov theorems. In Gani J. and Hannan E. J. (Editors), *Essays in Statistical Science. Journal of Applied Probability*, Special Volume 19A, pages 359–365. (Festschrift for P. A. P. Moran.)

Pollard, D. (1982b). Quantization and the method of k-means. *IEEE Transactions on Information Theory* **IT-28**: 199–205.

Pollard, D. (1982c). A central limit theorem for empirical processes. *Journal of the Australian Mathematical Society (Series A)* **33**:235–248.

Pollard, D. (1982d). A central limit theorem for k-means clustering. *Annals of Probability* **10**:919–926.

Prohorov, Yu. V. (1956). Convergence of random processes and limit theorems in probability. *Theory of Probability and Its Applications* **1**:157–214.

Pyke, R. (1969). Applications of almost surely convergent constructions of weakly convergent processes. *Springer Lecture Notes in Mathematics* **89**: 187–200.

Pyke, R. (1970). Asymptotic results for rank statistics. In M. L. Puri (Editor), *Nonparametric Techniques in Statistical Inference*, pages 21–37. Cambridge: Cambridge University Press.

Pyke, R. and Shorack, G. (1968). Weak convergence of a two-sample empirical process and a new approach to Chernoff–Savage theorems. *Annals of Mathematical Statistics* **39**:755–771.

Pyke, R. and Sun, T. G. (1982). Weak convergence of empirical processes. (University of Washington Statistics Department Technical Report # 19.)

Ranga Rao, R. (1962). Relations between weak and uniform convergence of measures with applications. *Annals of Mathematical Statistics* **33**:659–680.

Resnick, S. I. (1975). Weak convergence to extremal processes. *Annals of Probability* **3**:951–960.

Révész, P. (1976). Three theorems of multivariate empirical process. *Springer Lecture Notes in Mathematics* **566**:106–126.

Sauer, N. (1972). On the density of families of sets. *Journal of Combinatorial Theory (A)* **13**:145–147.

Shelah, S. (1971). Stability, the f.c.p., and superstability; model theoretic properties of formulas in first-order theory. *Annals of Mathematical Logic* **3**:271–362.

Shelah, S. (1972). A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics* **41**:247–261.

Shiryayev, A. N. (1981). Martingales: recent developments, results and applications. *International Statistical Review* **49** : 199–233.

Silverman, B. W. (1978). Weak and strong uniform consistency of the kernel estimate of a density and its derivatives. *Annals of Statistics* **6**:177–184.

Simmons, G. F. (1963). *Introduction to Topology and Modern Analysis.* New York: McGraw-Hill.

Skorohod, A. V. (1956). Limit theorems for stochastic processes. *Theory of Probability and Its Applications* **1**:261–290.

Skorohod, A. V. (1957). Limit theorems for stochastic processes with independent increments. *Theory of Probability and Its Applications* **2**:138–171.

Steele, J. M. (1975). *Combinatorial Entropy and Uniform Limit Laws.* Ph.D. thesis, Stanford. (Reproduced by University Microfilms, Ann Arbor.)

Steele, J. M. (1978). Empirical discrepancies and subadditive processes. *Annals of Probability* **6**:118–127.

Stone, C. (1963). Weak convergence of stochastic processes defined on semi-infinite time intervals. *Proceedings of the American Mathematical Society* **14**:694–696.

Straf, M. (1969). *A general Skorohod space and its applications to the weak convergence of stochastic processes with several parameters.* Ph.D. thesis, University of Chicago.

Strassen, V. (1965). The existence of probability measures with given marginals. *Annals of Mathematical Statistics* **36**:423–439.

Strassen, V. and Dudley, R. M. (1969). The central limit theorem and ε-entropy. *Springer Lecture Notes in Mathematics* **89**:224–231.

Stute, W. (1982a). The oscillation behavior of empirical processes. *Annals of Probability* **10**:86–107.

Stute, W. (1982b). A law of the logarithm for kernel density estimators. *Annals of Probability* **10**:414–422.

Talagrand, M. (1978). Les boules peuvent-elles engendrer la tribu borélienne d'un espace métrisable non séparable? *Séminaire Choquet, Paris* 17e annee: C5.

Tong, Y. L. (1980). *Probability Inequalities in Multivariate Distributions.* New York: Academic Press.

Topsøe, F. (1970). *Topology and Measure.* Springer Lecture Notes in Mathematics **133**.

Uspensky, J. V. (1937). *Introduction to Mathematical Probability.* New York: McGraw-Hill.

Vapnik, V. N. and Cervonenkis, A. Ya. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and Its Applications* **16**:264–280.

Vapnik, V. N. and Cervonenkis, A. Ya. (1981). Necessary and sufficient conditions for the uniform convergence of means to their expectations. *Theory of Probability and Its Applications* **26**:532–553.

Varadarajan, V. S. (1965). Measures on topological spaces. *American Mathematical Society Translations* **48**:161–228. (Translation from *Mat. Sbornik* **55** (97) (1961):35–100.)

Weierstrass, K. (1885). Über die analytische Darstellbarkeit sogenannter willkürlicher Funktionen reeller Argumente. *Sitzungsberichte der Preussischen Akademie der Wissenschaften*: 633–640, 789–906. (Also in Collected Works, Vol. III, pp. 1–37, Berlin: Mayer u. Müller 1903.)

Whitt, W. (1970). Weak convergence of probability measures on the function space $C[0, \infty)$. *Annals of Mathematical Statistics* **41**: 939–944.

Whitt, W. (1980). Some useful functions for functional limit theorems. *Mathematics of Operation Research* **5**: 67–85. (First circulated as preprint in 1970.)

Wichura, M. (1971). A note on the weak convergence of stochastic processes. *Annals of Mathematical Statistics* **42**: 1769–1772.

Author Index

Aldous, D. 133, 137, 185, 187
Alexander, K. S. 38, 166
Alexandroff, A. D. 85
Amato, B. R. viii
Araujo, A. 37
Ash, R. B. 39

Barry, D. G. viii
Bennett, G. 193–194
Bertrand-Retali, M. 38, 41
Billingsley, P. 36, 61, 85–86, 117–118, 136
Bollobás, B. 86
Bolthausen, E. 167
Boyce, J. S. viii
Breiman, L. 4, 22, 106, 118
Brémaud, P. 186
Breslow, N. 186
Brown, B. M. 166, 185

Cervonenkis, A. Ya. 37
Chernoff, H. 165, 190
Chibisov, D. M. 85
Chung, K. L. 117, 173
Crowley, J. 186

DeHardt, J. 36
Dellacherie, C. 176, 179, 185, 195–197
Denby, L. 37
Donsker, M. D. 117–118
Doob, J. L. 3, 4, 64, 113, 117
Dudley, R. M. vii, 37–39, 85–86, 117–118, 166–168, 193, 200
Durbin, J. 118
Durst, M. 39

Eggleston, H. G. 200
Erdös, P. 118

Feller, W. 53, 63, 193

Gaenssler, P. vii, 36–37, 40, 118, 167
Gihman, I. I. 4, 117–118, 136–137, 166, 185
Gill, R. D. 186
Giné, E. 37, 40, 166–167
Glaisher, J. W. L. 61
Gnedenko, B. V. 37

Hájek, J. 86, 117
Hall, P. 185
Halmos, P. R. 39
Hansel, G. 86
Hartigan, J. A. viii, 36, 166
Heyde, C. C. 185
Hoeffding, W. 193–194
Huber, P. J. 38, 165, 168

Jacobs, K. 86
Jacobsen, M. 186

Kac, M. 117–118
Kelley, J. L. 39
Kildea, D. G. 166
Kolchinsky, V. I. 38, 166
Kolmogorov, A. N. 99, 117, 136, 166
Kurtz, T. G. 137

Le Cam, L. 37–38, 85, 166–167, 200
Lévy, P. 61
Liapounoff, A. M. 61, 63
Lindeberg, J. W. 61
Lindvall, T. 118, 136
Liptser, R. S. 185
Loève, M. 118

Major, P. 86
Mann, H. B. 189
Maritz, J. S. 166
Martin, R. D. 37
McKean, H. P. 146
McLeish, D. L. 185
Métivier, M. 185
Meyer, P-A. 176, 179, 185, 195–197

Neveu, J. 22
Neyman, J. 47
Nolan, D. A. viii

Oxtoby, J. C. 65, 86

Parr, W. 118
Parthasarathy, K. R. 85, 136
Philipp, W. 166, 200
Pickands, J. 136
Pollard, D. 36–39, 118, 166, 169
Prohorov, Yu. V. 85–86
Pyke, R. vii, 85–86, 167

Ranga Rao, R. 86
Resnick, S. I. 136
Révész, P. 118, 167

Sauer, N. 37
Shelah, S. 37
Shiryayev, A. N. 185
Shorack, G. vii–viii, 85
Silverman, B. W. 38
Simmons, G. F. 67–68, 85
Skorohod, A. V. 4, 86, 117–118, 122, 136–137, 166, 185
Steele, J. M. 37
Stone, C. 118, 136
Straf, M. 136
Strassen, V. 86, 167
Stute, W. vii, 36–38, 117
Sun, T. G. 167

Talagrand, M. 85
Tong, Y. L. 41
Topsøe, F. 36–37, 62, 85–86
Troallic, J. P. 86

Uspensky, J. V. 61, 193

Vapnik, V. N. 37
Varadarajan, V. S. 85

Wald, A. 189
Weierstrass, K. 61
Wellner, J. viii
Whitt, W. 118, 136
Wichura, M. 86, 117, 136

Yu, K. F. 186

Zinn, J. 37, 38, 40, 166–167

Subject Index

abuse, of notation 44, 171
adapted 176
Aldous's condition 133
Allocation Lemma 77
almost-sure representation. *See* Representation Theorem
Approximation Lemma 27
autoregression 12, 174
axiom of choice 65, 86

B_H. *See* brownian motion, stretched out
B_P. *See* P-motion
$\mathscr{B}^P$ 156
Bennett's Inequality 160, 192, 194
Bernstein's inequality 193
bias, of density estimator 35, 42
binomial coefficient 19
bounded variation 42
brownian bridge 64
  defined 3, 95
  existence 101, 119
brownian motion
  construction from brownian bridge 103
  defined 95
  modulus of continuity for 146
  stretched-out 178
  tied-down 3

$C(\mathscr{F}, P)$ 156
$C[0, \infty)$ 108
$C[0, 1]$ 1, 90
cadlag 3, 89, 176
Cauchy sequence 81
censored data, estimation from 182
Central Limit Theorem
  Empirical, for distribution functions 97
  Empirical, uniform case 96
  Liapounoff 51
  Lindeberg 52
  Multivariate 57
central limit theorem
  autoregression 174
  empirical process 157, 163–165
  martingale-difference array 171
  minimization functional 141
Chaining Lemma 144
chaining 37
  restricted 160–163
  for stochastic processes 142–145
characteristic function 54
  Continuity Theorem 55, 60
  inversion formula 63
  uniform convergence on compacta 59
  uniquely determines distribution 55
chi-square, Pearson's 46
clustering. *See* k-means
combinatorial method 13
compactness
  in metric spaces 81–82
  sequential 82
Compactness Theorem 82

completely regular
  point of metric space 67
  topological space 68
conditional variance
  of increment 171
  process 176
continuity set 71
Continuity Theorem, for characteristic functions 55, 60
Continuous Mapping Theorem
  for euclidean space 46
  failure of 66
  for metric spaces 70
consistency
  for autoregression 12
  of k-means 9
  of optimal solutions 12
convergence in distribution
  Aldous's condition for, in $D[0, \infty)$ 134
  in $D[0, \infty)$, for Skorohod metric 131
  in $D[0, \infty)$, for uniform convergence on compacta 108
  of empirical process 97, 157
  in euclidean space 43
  of L^2 martingale 179
  in metric spaces 65
  process with independent increments 104, 135
  via semicontinuity 73
  via uniform approximation 70
  of uniform empirical process 96
Convergence Lemma
  for euclidean space 44–46
  for metric spaces 68
convex sets, shattering 17, 23
convolution 35, 49, 54
coupling 76–80
  in Compactness Theorem 84
covering integral
  for metric space 143
  as modulus of continuity 147
covering numbers
  direct 164
  inequalities between 34
  $\mathscr{L}^1$, for classes of functions 25
  $\mathscr{L}^2$, for classes of functions 31
  measurability of 198
  for metric spaces 143
  for polynomial classes 27
  random 150
criterion function
  minimization of 28
  random 12
cross-section, measurable 197
crossword puzzle 76

$D[0, \infty)$
  definition of 107
  metric of uniform convergence on compacta 108
  Skorohod metric 123
$D[0, 1]$ 3, 89–90
$\Delta_n(\xi)$ 21
delta method 63, 189
density
  convolution 55
  estimation 35
differentiability, in quadratic mean 140, 151
direct approximation 8, 164
discrimination, linear; polynomial; quadratic 17
distribution function, pointwise convergence of 43, 46, 53
Donsker's theorem. *See* Central Limit Theorem, Empirical
Doob–Meyer decomposition 176

$\frac{8}{5}$-argument 24
E_n. *See* empirical process
E_P. *See* P-bridge
empirical measure
  bivariate 12
  defined 6
empirical process
  central limit theorem for 96–97
  defined 2, 95, 97, 140
  measurability of 65, 199
envelope, for classes of functions 24, 151
equicontinuity
  of class of functions 74
  stochastic 139
Equicontinuity Lemma 150
ergodicity 9, 12
EXPONENTIAL BOUNDS 16
exponential bound. *See* tail probability
extreme-value distribution 128

fidi projection 92. *See also* projection maps
fidis (finite-dimensional distributions) 3, 90, 92
function space $\mathscr{X}$ 155, 199

functional
  first passage time 109, 124
  in function space 156
  jump 179
  terminology 2, 7
functions, infinitely differentiable 49

gaussian process
  characterization 106
  indexed by functions 146
Glivenko–Cantelli theorem
  classical 6–7, 13
  generalized, for classes of functions 25
  generalized, for classes of sets 18, 22
goodness-of-fit statistic 2
  with estimated parameters 99, 159
  Kolmogorov's, limiting distribution of 113
  Neyman's 47
graph, of real-valued function 27
grid, approximations constructed from 91, 124

heuristic argument, Doob's 4, 64
Hewitt–Savage, zero-one law of 22
Hoeffding's Inequality 16, 26, 31, 150, 164, 191

independent increments
  convergence in distribution 104, 135
  definition 103
innovations 174
INTEGRATION 16, 21
interpolation
  continuous 101
  in $D[0, \infty)$ 124
  in $D[0, 1]$ 92
  in a function space 158

$J(\delta)$ xiv. *See also* covering integral

Kaplan–Meier estimator 182
kernel
  smoothing 35
  translates of 42
k-means
  central limit theorem 153
  consistency 9, 30
  method of 9

Laplace transform, for hitting time 112
Liapounoff condition 51
Lindeberg condition 52, 106
  for martingale-difference arrays 171
Lindelöf's theorem 68, 72, 87
Lipschitz condition 45, 74, 101
little links 163
location, center of 28

marginal stacks, with coupling 77
markov property for empirical process 96
marriage lemma 77
martingale
  continuous time 176
  difference array 171
  L^2 176
  L^2, convergence to gaussian process 179
  reversed 39
  reversed, empirical measure as 22
  stopped 180
MAXIMAL INEQUALITY 15
maximal inequality
  for cadlag processes 94
  by chaining 144
  for L^2-martingales 177
maximum likelihood 138
measurability, for random elements of metric spaces 64–66
median
  central limit theorem for 53, 98
  consistency 7
  spatial, central limit theorem for 152
  spatial, consistency of 28
M-estimator
  generalized 12
  stochastic equicontinuity 165
metric
  bounded-Lipschitz 74
  entropy 166
  Prohorov 75, 79, 88
  Skorohod, on $D[0, \infty)$ 123–124
minimization of random criterion function 140
modulus function for $D[0, \infty)$ 125
multinomial distribution 46

$N(\delta)$ xiv. *See also* covering number
$\|\cdot\|$. *See* norm
norm 14, 24, 156

normal distribution
characteristic function of 63
multivariate 29
tail probabilities 191

optimal centers 10
outer measure 70

$\mathscr{P}$. *See* projection σ-field
P_n. *See* empirical measure
P_n° 15, 26, 149
partial sum process 106, 109
P-bridge 149
permissible
classes of functions 24
classes of sets 17
definition 196
π_S. *See* projection maps
P-motion 147
polynomial
bound for number of sets picked out 19
class 17, 27
discriminating 17
discrimination 17, 20
orthogonal 47
poisson process 105, 129
predictable 176
product σ-field 69
projection, measurability of 196
projection maps 66, 90
for $D[0, \infty)$ 130
projection σ-field 66, 90
on $D[0, \infty)$ 127
on $D[0, 1]$ 87
Prohorov's theorem. *See* Compactness Theorem

quadratic variation process 185
quantile transformation 57, 97, 129

random element of metric space 1, 65
random vector
characteristic functions 56
convergence in distribution of 43, 56
perturbation of 48
rates of convergence 30
reflection principle, for brownian motion 112
remainder term, in Taylor expansion 50, 139
Representation Theorem
with random elements 71
with random variables 58
ρ_P xiv
robustnik 75

sample path 1
Scheffé's lemma 61
semicontinuity 73, 87
seminorm xiv
separability
of $C[0, 1]$ 87
of $D[0, \infty)$, under Skorohod metric 127
and σ-fields 85
for subsets of metric spaces 67
universal 38
shatter 18, 21, 198
sigma field. *See also* $\mathscr{B}^P$, $\mathscr{P}$
generated by balls 87
sign variables 15
signed measure. *See* P_n°
Slutsky's theorem 62
Souslin 196
square-root trick 32, 37
Stirling's approximation 21
stochastic
equicontinuity 139–140
order symbol 141, 189
process 1
stopping time 110, 172
Strasbourg theory 176
strong markov property, for brownian motion 111
submartingale 178
reversed 22, 25
subsets
hidden 19
picked out 18
substitution, of increments 49
symmetrization 32, 149
inequality 14–15
SYMMETRIZATION, FIRST 14
Symmetrization Lemma 14, 198
SYMMETRIZATION, SECOND 15
symmetry, and empirical measures 21

tail probability
characteristic function bound 60
exponential bound 160, 191

tied-down brownian motion. *See* brownian bridge
tight measure
  on metric space 81
  on real line 60
total boundedness
  of metric space 82
  and P-motion 168
triangular array 51, 106
truncation 25
2-means. *See* k-means

U_n. *See* empirical process, uniform
uniform convergence of density estimator 36
uniform integrability 176
uniform strong law of large numbers 7
  for classes of functions 25, 168
  for classes of sets 18, 22
  for convex sets 22
uniform tightness
  for general empirical processes 156
  for uniform empirical processes 102
usual conditions 176

VC class 37
vector space, finite-dimensional 20, 30, 38

weak convergence
  in euclidean space 44
  in metric spaces 65
weight functions 158

$\mathscr{X}$. *See* function space