

# **Springer Series in Statistics**

Aad W. van der Vaart  
Jon A. Wellner

# **Weak Convergence and Empirical Processes**

With Applications  
to Statistics



**Springer**

# **Springer Series in Statistics**

*Advisors:*

P. Bickel, P. Diggle, S. Fienberg, K. Krickeberg,  
I. Olkin, N. Wermuth, S. Zeger

Springer Science+Business Media, LLC

# **Springer Series in Statistics**

---

- Andersen/Borgan/Gill/Keiding*: Statistical Models Based on Counting Processes.
- Andrews/Herzberg*: Data: A Collection of Problems from Many Fields for the Student and Research Worker.
- Anscombe*: Computing in Statistical Science through APL.
- Berger*: Statistical Decision Theory and Bayesian Analysis, 2nd edition.
- Bolfarine/Zacks*: Prediction Theory for Finite Populations.
- Brémaud*: Point Processes and Queues: Martingale Dynamics.
- Brockwell/Davis*: Time Series: Theory and Methods, 2nd edition.
- Daley/Vere-Jones*: An Introduction to the Theory of Point Processes.
- Dzhaparidze*: Parameter Estimation and Hypothesis Testing in Spectral Analysis of Stationary Time Series.
- Fahrmeir/Tutz*: Multivariate Statistical Modelling Based on Generalized Linear Models.
- Farrell*: Multivariate Calculation.
- Federer*: Statistical Design and Analysis for Intercropping Experiments.
- Fienberg/Hoaglin/Kruskal/Tanur (Eds.)*: A Statistical Model: Frederick Mosteller's Contributions to Statistics, Science and Public Policy.
- Fisher/Sen*: The Collected Works of Wassily Hoeffding.
- Good*: Permutation Tests: A Practical Guide to Resampling Methods for Testing Hypotheses.
- Goodman/Kruskal*: Measures of Association for Cross Classifications.
- Grandell*: Aspects of Risk Theory.
- Hall*: The Bootstrap and Edgeworth Expansion.
- Härdle*: Smoothing Techniques: With Implementation in S.
- Hartigan*: Bayes Theory.
- Heyer*: Theory of Statistical Experiments.
- Jolliffe*: Principal Component Analysis.
- Kolen/Brennan*: Test Equating: Methods and Practices.
- Kotz/Johnson (Eds.)*: Breakthroughs in Statistics Volume I.
- Kotz/Johnson (Eds.)*: Breakthroughs in Statistics Volume II.
- Kres*: Statistical Tables for Multivariate Analysis.
- Le Cam*: Asymptotic Methods in Statistical Decision Theory.
- Le Cam/Yang*: Asymptotics in Statistics: Some Basic Concepts.
- Longford*: Models for Uncertainty in Educational Testing.
- Manoukian*: Modern Concepts and Theorems of Mathematical Statistics.
- Miller, Jr.*: Simultaneous Statistical Inference, 2nd edition.
- Mosteller/Wallace*: Applied Bayesian and Classical Inference: The Case of *The Federalist Papers*.
- Pollard*: Convergence of Stochastic Processes.
- Pratt/Gibbons*: Concepts of Nonparametric Theory.

(continued after index)

Aad W. van der Vaart Jon A. Wellner

# Weak Convergence and Empirical Processes

With Applications to Statistics



Springer

Aad W. van der Vaart  
Department of Mathematics  
and Computer Science  
Free University  
De Boelelaan 1081a  
1081 HV Amsterdam  
The Netherlands  
aad@cs.vu.nl

Jon A. Wellner  
University of Washington  
Statistics  
Box 354322  
Seattle, WA 98195-4322  
jaw@stat.washington.edu

With one illustration.

Library of Congress Cataloging-in-Publication Data  
Vaart, A.W. van der.  
Weak convergence and empirical processes / by Aad van der Vaart  
and Jon A. Wellner.  
p. cm. — (Springer series in statistics)  
Includes bibliographical references and indexes.  
ISBN 978-1-4757-2547-6 ISBN 978-1-4757-2545-2 (eBook)  
DOI 10.1007/978-1-4757-2545-2  
1. Stochastic processes. 2. Convergence. 3. Distribution  
(Probability theory) 4. Sampling (Statistics) I. Wellner, Jon A.,  
1945— . II. Title. III. Series.  
QA274.V33 1996  
519.2—dc20 95-49099

Printed on acid-free paper.

© 1996 by Springer Science+Business Media New York  
Originally published by Springer-Verlag New York, Inc. in 1996  
Softcover reprint of the hardcover 1st edition 1996

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

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

Production managed by Frank Ganz; manufacturing supervised by Jacqui Ashri.  
Photocomposed pages prepared from the author's TeX files.

9 8 7 6 5 4 3 2 1

ISBN 978-1-4757-2547-6

SPIN 10522999

*To Maryse*

*To Cynthia*

# Preface

This book tries to do three things. The first goal is to give an exposition of certain modes of stochastic convergence, in particular convergence in distribution. The classical theory of this subject was developed mostly in the 1950s and is well summarized in Billingsley (1968). During the last 15 years, the need for a more general theory allowing random elements that are not Borel measurable has become well established, particularly in developing the theory of empirical processes. Part 1 of the book, *Stochastic Convergence*, gives an exposition of such a theory following the ideas of J. Hoffmann-Jørgensen and R. M. Dudley.

A second goal is to use the weak convergence theory background developed in Part 1 to present an account of major components of the modern theory of empirical processes indexed by classes of sets and functions. The weak convergence theory developed in Part 1 is important for this, simply because the empirical processes studied in Part 2, *Empirical Processes*, are naturally viewed as taking values in nonseparable Banach spaces, even in the most elementary cases, and are typically *not* Borel measurable. Much of the theory presented in Part 2 has previously been scattered in the journal literature and has, as a result, been accessible only to a relatively small number of specialists. In view of the importance of this theory for statistics, we hope that the presentation given here will make this theory more accessible to statisticians as well as to probabilists interested in statistical applications.

Our third goal is to illustrate the usefulness of modern weak convergence theory and modern empirical process theory for statistics by a wide

variety of applications. On the one hand, as is also clear through the work of David Pollard, the theory of empirical processes provides a collection of extremely powerful tools for proving many of the main limit theorems of asymptotic statistics. On the other hand, the empirical distribution indexed by a collection of sets or functions is an object of independent statistical interest; for instance, as a measure of goodness of fit. The topics included in Part 3 of the book, *Statistical Applications*, range from rates of convergence in semiparametric estimation, to the functional delta-method, bootstrap, permutation empirical processes, and the convolution theorem. We have not aimed at giving an exhaustive coverage of the statistical background or related literature in presenting these applications. The choices made reflect our personal interests and research efforts over the past few years, so the reader should understand that many equally interesting applications and further research directions are not covered here. For instance, we expect significant progress in semiparametric theory through the use of empirical processes in the next few years. Wellner (1992) reviews various applications of empirical process methods through 1992.

This project began with a joint effort to develop some results in the Hoffmann-Jørgensen theory (namely Prohorov's Theorem 1.3.9) needed to prove the convolution and asymptotic minimax theorem for estimators with values in nonseparable Banach spaces (see Chapter 3.11). That effort was successful, but it resulted in a manuscript that was too awkward for publication in a journal. The generality offered by the new weak convergence theory and the many applications of general empirical process theory in statistics led us to explore the area further. The result, several years later, is this book.

Along the way we have learned much from the main contributors to empirical process theory, from colleagues, and from friends. In particular, we owe thanks to Lucien Birgé, R. M. Dudley, Peter Gaensler, Sara van de Geer, Richard Gill, Evarist Giné, Piet Groeneboom, Lucien Le Cam, Michel Ledoux, Pascal Massart, Susan Murphy, David Pollard, Michel Talagrand, and Joel Zinn, for advice, corrections, discussions, clarifications, inspiration, preprints, and help.

The authors presented preliminary versions of various parts of the manuscript in courses at the Free University, Amsterdam, and the University of Washington to patient and sharp-eyed groups of students, including Jian Huang, Marianne Jonker, Brad McNeney, Jinko Graham, Jens Praestgaard, and Anne Sheehy. We owe them as well as Franz Strobl, Klaus Ziegler, and Lutz Duembgen thanks for picking up errors and points needing clarification.

The authors have been partially supported in their research efforts over the period of work on this book (1988–1995) by grants from the National Science Foundation, NIAID, NWO, NATO, CNRS, Stieltjes Institute, and Free University Amsterdam (which funded several visits of the second author to that fair city). Our thanks also go to our editor at Springer, Martin

Gilchrist, and to the staff at Springer-Verlag for their efficient job in turning the  $\text{\TeX}$  manuscript into a published book.

Amsterdam and Seattle  
August 1995

# Reading Guide

This book consists of three parts and an appendix. Part 1 is an exposition of (mostly) three modes of convergence of stochastic variables with values in metric spaces: convergence in distribution, convergence in probability, and almost sure convergence. A new aspect of this part, as compared to existing literature, is the fact that the stochastic variables are not assumed measurable with respect to the Borel  $\sigma$ -field. Part 1 is also useful in collecting a large number of results that have thus far not been available in book form in sufficient generality. In particular this concerns the systematic treatment of convergence in distribution in spaces of bounded functions equipped with the supremum metric. Most subjects treated in Part 1 are used in the two later parts of the book, but Chapters 1.2, 1.3, 1.5, 1.9, and 1.10 probably form the core of this part, and should be read in this order. Part 1 begins with a thorough introduction.

Part 2 is mostly concerned with the empirical measure and empirical process of a sample of observations, indexed by a class of functions. These are the maps  $f \mapsto \sum_{i=1}^n f(X_i)$  and  $f \mapsto n^{-1/2} \sum_{i=1}^n (f(X_i) - Pf)$  whose domain is a class  $\mathcal{F}$  of measurable functions. The main results in this part are contained in Chapters 2.4 and 2.5 and concern the uniform law of large numbers (Glivenko-Cantelli theorem) and uniform central limit theorem (Donsker theorem), respectively. Chapters 2.6, 2.7, and 2.10 contain many examples of classes that satisfy the conditions of the theorems in these chapters. Chapter 2.1 is an introduction and Chapters 2.2 and 2.3 are necessary preparation for the main results of Part 2. Again, many of the results from the other chapters reappear later in the book but need not be studied in sequential order.

Part 3 consists of 11 chapters, which in principle can be read independently. This part shows the wide range of applications of the results obtained in the earlier parts in statistics, ranging from parametric and nonparametric estimation to the bootstrap, the functional delta-method, and Kolmogorov-Smirnov statistics. Every chapter assumes familiarity of the basic notation of Parts 1 and 2, but no section requires knowledge of more than a few sections of Part 2. Chapter 3.1 gives an overview of this part.

The material presented in the three parts is self-contained to a reasonable extent. The appendix covers a number of auxiliary subjects that are used to develop some of the material in the three main Parts. Some results are presented with proof and some without, to serve as an easy reference.

Most of the chapters contain a number of “problems and complements” at the end. A number of these are real, textbook-style problems, but a lot of this material is meant as a supplement to the main text. Some problems present technical details, while other problems concern additional results of interest. We should warn the reader that the density of errors in these problem sections is probably higher than in the main text. Many problems and complements have not been double-checked.

# Contents

<b>Preface</b> . . . . .	vii
<b>Reading Guide</b> . . . . .	xi

<b>1. Stochastic Convergence</b> . . . . .	1
1.1. Introduction . . . . .	2
1.2. Outer Integrals and Measurable Majorants . . . . .	6
1.3. Weak Convergence . . . . .	16
1.4. Product Spaces . . . . .	29
1.5. Spaces of Bounded Functions . . . . .	34
1.6. Spaces of Locally Bounded Functions . . . . .	43
1.7. The Ball Sigma-Field and Measurability of Suprema . . . . .	45
1.8. Hilbert Spaces . . . . .	49
1.9. Convergence: Almost Surely and in Probability . . . . .	52
1.10. Convergence: Weak, Almost Uniform, and in Probability . . . . .	57
1.11. Refinements . . . . .	67
1.12. Uniformity and Metrization . . . . .	71
<i>Notes</i> . . . . .	75

<b>2. Empirical Processes . . . . .</b>	<b>79</b>
2.1. Introduction . . . . .	80
2.1.1. Overview of Chapters 2.3–2.14 . . . . .	83
2.1.2. Asymptotic Equicontinuity . . . . .	89
2.1.3. Maximal Inequalities . . . . .	90
*2.1.4. The Central Limit Theorem in Banach Spaces . . . . .	91
2.2. Maximal Inequalities and Covering Numbers . . . . .	95
2.2.1. Sub-Gaussian Inequalities . . . . .	100
2.2.2. Bernstein’s Inequality . . . . .	102
*2.2.3. Tightness Under an Increment Bound . . . . .	104
2.3. Symmetrization and Measurability . . . . .	107
2.3.1. Symmetrization . . . . .	107
*2.3.2. More Symmetrization . . . . .	111
*2.3.3. Separable Versions . . . . .	115
2.4. Glivenko-Cantelli Theorems . . . . .	122
2.5. Donsker Theorems . . . . .	127
2.5.1. Uniform Entropy . . . . .	127
2.5.2. Bracketing . . . . .	129
2.6. Uniform Entropy Numbers . . . . .	134
2.6.1. VC-Classes of Sets . . . . .	134
2.6.2. VC-Classes of Functions . . . . .	140
2.6.3. Convex Hulls and VC-Hull Classes . . . . .	142
2.6.4. VC-Major Classes . . . . .	145
2.6.5. Examples and Permanence Properties . . . . .	146
2.7. Bracketing Numbers . . . . .	154
2.7.1. Smooth Functions and Sets . . . . .	154
2.7.2. Monotone Functions . . . . .	159
2.7.3. Closed Convex Sets and Convex Functions . . . . .	162
2.7.4. Classes That Are Lipschitz in a Parameter . . . . .	164
2.8. Uniformity in the Underlying Distribution . . . . .	166
2.8.1. Glivenko-Cantelli Theorems . . . . .	166
2.8.2. Donsker Theorems . . . . .	168
2.8.3. Central Limit Theorem Under Sequences . . . . .	173
2.9. Multiplier Central Limit Theorems . . . . .	176
2.10. Permanence of the Donsker Property . . . . .	190
2.10.1. Closures and Convex Hulls . . . . .	190
2.10.2. Lipschitz Transformations . . . . .	192
2.10.3. Permanence of the Uniform Entropy Bound . . . . .	198
2.10.4. Partitions of the Sample Space . . . . .	200
2.11. The Central Limit Theorem for Processes . . . . .	205
2.11.1. Random Entropy . . . . .	205
2.11.2. Bracketing . . . . .	210
2.11.3. Classes of Functions Changing with $n$ . . . . .	220
2.12. Partial-Sum Processes . . . . .	225

2.12.1. The Sequential Empirical Process . . . . .	225
2.12.2. Partial-Sum Processes on Lattices . . . . .	228
2.13. Other Donsker Classes . . . . .	232
2.13.1. Sequences . . . . .	232
2.13.2. Elliptical Classes . . . . .	233
2.13.3. Classes of Sets . . . . .	236
2.14. Tail Bounds . . . . .	238
2.14.1. Finite Entropy Integrals . . . . .	238
2.14.2. Uniformly Bounded Classes . . . . .	245
2.14.3. Deviations from the Mean . . . . .	254
2.14.4. Proof of Theorem 2.14.13 . . . . .	257
<i>Notes</i> . . . . .	269
<b>3. Statistical Applications</b> . . . . .	277
3.1. Introduction . . . . .	278
3.2. M-Estimators . . . . .	284
3.2.1. The Argmax Theorem . . . . .	285
3.2.2. Rate of Convergence . . . . .	289
3.2.3. Examples . . . . .	294
3.2.4. Linearization . . . . .	300
3.3. Z-Estimators . . . . .	309
3.4. Rates of Convergence . . . . .	321
3.4.1. Maximum Likelihood . . . . .	326
3.4.2. Concave Parametrizations . . . . .	330
3.4.3. Least Squares Regression . . . . .	331
3.4.4. Least-Absolute-Deviation Regression . . . . .	336
3.5. Random Sample Size, Poissonization and Kac Processes . . . . .	339
3.5.1. Random Sample Size . . . . .	339
3.5.2. Poissonization . . . . .	341
3.6. The Bootstrap . . . . .	345
3.6.1. The Empirical Bootstrap . . . . .	345
3.6.2. The Exchangeable Bootstrap . . . . .	353
3.7. The Two-Sample Problem . . . . .	360
3.7.1. Permutation Empirical Processes . . . . .	362
3.7.2. Two-Sample Bootstrap . . . . .	365
3.8. Independence Empirical Processes . . . . .	367
3.9. The Delta-Method . . . . .	372
3.9.1. Main Result . . . . .	372
3.9.2. Gaussian Limits . . . . .	376
3.9.3. The Delta-Method for the Bootstrap . . . . .	377
3.9.4. Examples of the Delta-Method . . . . .	381
3.10. Contiguity . . . . .	401
3.10.1. The Empirical Process . . . . .	406

3.10.2. Change-Point Alternatives . . . . .	408
3.11. Convolution and Minimax Theorems . . . . .	412
3.11.1. Efficiency of the Empirical Distribution . . . . .	420
<i>Notes</i> . . . . .	423
<b>A. Appendix</b> . . . . .	429
A.1. Inequalities . . . . .	430
A.2. Gaussian Processes . . . . .	437
A.2.1. Inequalities and Gaussian Comparison . . . . .	437
A.2.2. Exponential Bounds . . . . .	442
A.2.3. Majorizing Measures . . . . .	445
A.2.4. Further Results . . . . .	447
A.3. Rademacher Processes . . . . .	449
A.4. Isoperimetric Inequalities for Product Measures . . . . .	451
A.5. Some Limit Theorems . . . . .	456
A.6. More Inequalities . . . . .	459
A.6.1. Binomial Random Variables . . . . .	459
A.6.2. Multinomial Random Vectors . . . . .	462
A.6.3. Rademacher Sums . . . . .	463
<i>Notes</i> . . . . .	465
<b>References</b> . . . . .	467
<b>Author Index</b> . . . . .	487
<b>Subject Index</b> . . . . .	493
<b>List of Symbols</b> . . . . .	506

PART 1

# Stochastic Convergence

# 1.1

## Introduction

The first goal in this book is to give an exposition of the modern weak convergence theory suitable for the study of empirical processes.

Let  $(\mathbb{D}, d)$  be a metric space, and let  $\{P_n\}$  and  $P$  be Borel probability measures on  $(\mathbb{D}, \mathcal{D})$ , where  $\mathcal{D}$  is the Borel  $\sigma$ -field on  $\mathbb{D}$ , the smallest  $\sigma$ -field containing all the open sets. Then the sequence  $P_n$  converges weakly to  $P$ , which we write as  $P_n \rightsquigarrow P$ , if and only if

$$(1.1.1) \quad \int_{\mathbb{D}} f \, dP_n \rightarrow \int_{\mathbb{D}} f \, dP, \quad \text{for all } f \in C_b(\mathbb{D}).$$

Here  $C_b(\mathbb{D})$  denotes the set of all bounded, continuous, real functions on  $\mathbb{D}$ . Equivalently, if  $X_n$  and  $X$  are  $\mathbb{D}$ -valued random variables with distributions  $P_n$  and  $P$  respectively, then  $X_n \rightsquigarrow X$  if and only if

$$(1.1.2) \quad \mathbf{E}f(X_n) \rightarrow \mathbf{E}f(X), \quad \text{for all } f \in C_b(\mathbb{D}).$$

These definitions yield the classical theory of weak convergence as treated in Billingsley (1968) and which has proved very useful in probability theory and statistics. The basic elements of this theory include the portmanteau theorem, continuous mapping theorems, Prohorov's theorem, tools for establishing tightness and uniform tightness, and weak convergence results for product spaces.

The classical theory requires that  $P_n$  is defined, for each  $n$ , on the Borel  $\sigma$ -field  $\mathcal{D}$ , or, equivalently, that  $X_n$  is a Borel measurable map for each  $n$ . If  $(\Omega_n, \mathcal{A}_n, P_n)$  are the underlying probability spaces on which the maps  $X_n$  are defined, this means that  $X_n^{-1}(D) \in \mathcal{A}_n$  for every Borel set

$D$ . This required measurability usually holds when  $\mathbb{D}$  is a separable metric space such as  $\mathbb{R}^k$  or  $C[0, 1]$  with the supremum metric and sometimes even when  $\mathbb{D}$  is nonseparable (for example, in the case of partial-sum processes).

However, this apparently modest requirement can and does easily fail when the metric space  $\mathbb{D}$  is not separable. For example, this occurs when  $\mathbb{D}$  is the Skorohod space  $D[0, 1]$  of all right-continuous functions on  $[0, 1]$  with left limits is endowed with the metric induced by the supremum norm or for the space  $\ell^\infty(\mathcal{F})$  of all bounded functions from a set  $\mathcal{F}$  to  $\mathbb{R}$  equipped with the supremum norm  $\|z\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} |z(f)|$ .

The classical example of this difficulty comes from empirical process theory. Suppose that for each  $n$  the random variables  $\xi_1, \dots, \xi_n$  are the independent, uniformly distributed variables, defined as the coordinate projections on the product probability space  $([0, 1], \mathcal{B}, \lambda)^n$ , where  $\lambda$  denotes Lebesgue measure on  $[0, 1]$  and  $\mathcal{B}$  the Borel  $\sigma$ -field. The empirical distribution function  $\mathbb{F}_n$  is the random function

$$\mathbb{F}_n(t) = \frac{1}{n} \sum_{i=1}^n 1_{[0,t]}(\xi_i), \quad 0 \leq t \leq 1;$$

and the uniform empirical process  $X_n$  is

$$X_n(t) = \sqrt{n}(\mathbb{F}_n(t) - t), \quad 0 \leq t \leq 1.$$

Both  $\mathbb{F}_n$  and  $X_n$  can be viewed as maps from  $[0, 1]^n$  into  $D[0, 1]$ , but neither  $\mathbb{F}_n$  nor  $X_n$  is a Borel measurable map if  $D[0, 1]$  is endowed with the supremum norm. It turns out that the Borel  $\sigma$ -field  $\mathcal{D}$  is so large in this case that the inclusion  $X_n^{-1}(\mathcal{D}) \subset \mathcal{B}^n$  fails to hold (cf. Problem 1.7.3). This failure was pointed out by Chibisov (1965) and is nicely explained by Billingsley (1968); see Billingsley's Chapter 18, pages 150–153. Thus, the basic definitions (1.1.1) and (1.1.2) of weak convergence cannot be used for  $X_n$  viewed as a random function with values in  $(D[0, 1], \|\cdot\|_\infty)$ , even though this is a very natural and useful space in which to view the processes  $X_n$ , and even though the natural limiting process  $X$ , a Brownian bridge process on  $[0, 1]$ , can be taken to be a Borel measurable map (which takes its values in the separable subspace  $C[0, 1]$ ).

Through the late 1960s, several different approaches were suggested to deal with this difficulty. Skorokhod (1956) and Billingsley (1968) endowed  $D[0, 1]$  with the Skorokhod metric (or a modification) under which  $D[0, 1]$  is separable (and complete) and the classical theory could be applied without difficulties. Dudley (1966, 1967a) developed an alternative weak convergence theory based on the smaller ball  $\sigma$ -field generated by the  $\|\cdot\|_\infty$ -open balls in  $D[0, 1]$ . Pyke and Shorack (1968) proposed yet another definition of weak convergence requiring convergence of the integrals only for those functions  $f \in C_b(\mathbb{D})$  for which  $f(X_n)$  is a measurable map from  $(\Omega_n, \mathcal{A}_n)$  into  $\mathbb{R}$ . While Dudley's proposition successfully handles the uniform empirical process and many simple extensions, it cannot handle the general empirical process and certain aspects of large sample theory in statistics.

The key idea put forward more recently by J. Hoffmann-Jørgensen is to drop the requirement of Borel measurability of each  $X_n$ , meanwhile upholding the requirement (1.1.2), where the expectations are now to be interpreted as *outer expectations* and the  $X_n$  may be arbitrary (possibly nonmeasurable) maps. Provided the limiting variable  $X$  is Borel measurable, a fruitful theory of weak convergence is still possible. This chapter is a systematic exposition of this theory, and it also investigates the relationships with other modes of stochastic convergence such as convergence in probability and almost sure convergence, which can also be extended to nonmeasurable maps.

The first job is to define outer integrals and the basic techniques related to these integrals. If  $T$  is an arbitrary (not necessarily measurable) map from a probability space  $(\Omega, \mathcal{A}, P)$  to the extended real line  $\bar{\mathbb{R}}$ , then the *outer integral* of  $T$  with respect to  $P$  is defined as

$$E^*T = \inf \left\{ EU : U \geq T, U : \Omega \mapsto \bar{\mathbb{R}} \text{ measurable and } EU \text{ exists} \right\}.$$

It is useful to know that the infimum is achieved in a strong sense: if  $E^*T < \infty$ , then  $E^*T = ET^*$ , where  $T^*$  is a “smallest measurable function above  $T$ .”

In terms of outer integrals, weak convergence of arbitrary, possibly nonmeasurable, maps  $X_n$  from underlying probability spaces  $(\Omega_n, \mathcal{A}_n, P_n)$  to a metric space  $\mathbb{D}$  is defined as follows. The sequence  $X_n$  converges weakly to a Borel measurable map  $X$ , and we write  $X_n \rightsquigarrow X$ , if

$$(1.1.3) \quad E^*f(X_n) \rightarrow Ef(X), \quad \text{for every } f \in C_b(\mathbb{D}).$$

This part develops the weak convergence theory connected with this definition, more or less in parallel to the classical theory, including a portmanteau theorem, continuous mapping theorems, Prohorov’s theorem, tightness and basic tools for establishing tightness, and weak convergence on product spaces.

Because of several important applications of nets of processes in statistics, we have formulated the theory in terms of *nets*  $\{X_\alpha\}_{\alpha \in A}$  with  $A$  a directed set rather than just for sequences. Recall that a *directed set*  $A$  is a set equipped with a partial order  $\leq$  with the further property that every pair  $\alpha_1, \alpha_2$  has a “successor”  $\alpha_3$  (satisfying  $\alpha_3 \geq \alpha_1$  and  $\alpha_3 \geq \alpha_2$ ). For a sequence, the directed set is the set of natural numbers with the usual ordering, and all results for nets are valid for sequences in particular.

Once the basic elements of the theory are available, Part 1 develops further connections and relationships. Because spaces of uniformly bounded functions are important for empirical process theory, Chapter 1.5 develops conditions for weak convergence in these spaces. For example, Theorem 1.5.4 asserts that asymptotic tightness together with weak convergence of the finite-dimensional distributions suffices for weak convergence. This

resembles the classical theorem for convergence in  $C[0, 1]$ , as in Billingsley (1968), but the present formulation is considerably more general and flexible.

The relationship between the present theory and measurability with respect to the ball  $\sigma$ -field is explored in Chapter 1.7. Chapter 1.8 develops the theory for the special case of a Hilbert space.

Just as it is useful to develop a theory of weak convergence for nonmeasurable maps, it is useful to extend the notions of “convergence in probability” and “convergence almost surely.” The appropriate analogues are developed and investigated in Chapter 1.9, including the corresponding continuous mapping theorems. Chapter 1.10 gives relationships and connections between these stronger convergence concepts and weak convergence as defined in (1.1.3). In particular, this section contains an analogue of the Skorokhod-Dudley-Wichura almost sure representation theorem for nonmeasurable maps.

The main result of Chapter 1.11 is the “extended continuous mapping theorem,” which applies to a sequence of functions  $g_n$  rather than to a fixed continuous function  $g$ . Chapter 1.12 explores uniformity of the convergence (1.1.3) in  $f$  for appropriate subsets  $\mathcal{F}$  of  $C_b(\mathbb{D})$  and uses this to metrize weak convergence under the additional condition that the limit  $X$  is separable.

Much of the theory of stochastic convergence for nonmeasurable maps follows the same lines as the classical theory. Some of the key differences between the present theory and the classical theory that emerge in the course of Part 1 are as follows:

- (i) The notion of (uniform) tightness of sequences needs modification.
- (ii) The “direct half” of Prohorov’s theorem (asymptotic tightness implies relative compactness) must be modified by the addition of the requirement of *asymptotic measurability*.
- (iii) Almost sure convergence or even convergence everywhere is meaningless without some (asymptotic) measurability.
- (iv) There is no general version of Fubini’s theorem (but see Lemmas 1.2.6 and 1.2.7).

In spite of these differences and slight difficulties resulting from sacrificing measurability of the converging random quantities, the weak convergence theory outlined in Part 1 parallels the classical theory to a remarkable degree and will prove valuable in dealing with empirical processes, which are studied in Part 2, and asymptotic statistical theory, which is treated in Part 3.

# 1.2

## Outer Integrals and Measurable Majorants

Let  $(\Omega, \mathcal{A}, P)$  be an arbitrary probability space and  $T: \Omega \mapsto \bar{\mathbb{R}}$  an arbitrary map. The *outer integral* of  $T$  with respect to  $P$  is defined as

$$E^*T = \inf \left\{ EU: U \geq T, U: \Omega \mapsto \bar{\mathbb{R}} \text{ measurable and } EU \text{ exists} \right\}.$$

Here, as usual,  $EU$  is understood to exist if at least one of  $EU^+$  or  $EU^-$  is finite. The *outer probability* of an arbitrary subset  $B$  of  $\Omega$  is

$$P^*(B) = \inf \left\{ P(A): A \supseteq B, A \in \mathcal{A} \right\}.$$

Note that the functions  $U$  in the definition of outer integral are allowed to take the value  $\infty$ , so that the infimum is never empty.

*Inner integral* and *inner probability* can be defined in a similar fashion – their definition should be obvious. Equivalently, they can be defined by  $E_*T = -E^*(-T)$  and  $P_*(B) = 1 - P^*(\Omega - B)$ , respectively.

A very useful fact is that the infima in the definitions of the outer integral and probability are always achieved. This even happens for an essentially minimal  $U$  and  $A$  as in the definitions (provided  $E^*T < \infty$ ), which will be denoted  $T^*$  and  $B^*$ . For outer integrals, this is contained in the following lemma; for outer probabilities, a proof is deferred to Lemma 1.2.3.

**1.2.1 Lemma (Measurable cover function).** *For any map  $T: \Omega \mapsto \bar{\mathbb{R}}$ , there exists a measurable function  $T^*: \Omega \mapsto \bar{\mathbb{R}}$  with*

- (i)  $T^* \geq T$ ;
- (ii)  $T^* \leq U$  a.s., for every measurable  $U: \Omega \mapsto \bar{\mathbb{R}}$  with  $U \geq T$  a.s.

For any  $T^*$  satisfying these requirements, it holds that  $E^*T = ET^*$ , provided  $ET^*$  exists. The latter is certainly true if  $E^*T < \infty$ .

**Proof.** Choose a measurable sequence  $U_m \geq T$  with  $E \arctan U_m \downarrow E^* \arctan T$ , and set

$$T^*(\omega) = \lim_{m \rightarrow \infty} \inf_{1 \leq k \leq m} U_k(\omega).$$

This defines a measurable function  $T^*$  taking values in the extended real line, with  $T^* \geq T$ , and by monotone convergence  $E \arctan T^* = E^* \arctan T$ . Every measurable  $U \geq T$  satisfies  $\arctan U \wedge T^* \geq \arctan T$ , so that  $E \arctan U \wedge T^* \geq E^* \arctan T = E \arctan T^*$ . But the integrand on the left side is trivially pointwise smaller than the integrand on the right side. Since they have the same expectation, they must be equal:  $\arctan U \wedge T^* = \arctan T^*$  a.s. This implies  $T^* \leq U$  a.s.

If  $ET^*$  exists, then it is larger than  $E^*T$  by (i) and smaller by (ii). Hence  $ET^* = E^*T$ . If  $E^*T < \infty$ , then there exists a measurable  $U \geq T$  with  $EU^+ < \infty$ . Then  $E(T^*)^+ \leq EU^+$  and  $ET^*$  exists. ■

The function  $T^*$  is called a *minimal measurable majorant* of  $T$ , or also a measurable cover or envelope function. It is unique only up to  $P$  null sets but, somewhat abusing notation, we write  $T^*$  for any member of its equivalence class of a.s. equal functions. Its expectation does not always exist, so some care is needed when applying the identity  $E^*T = ET^*$  (Problem 1.2.2). A *maximal measurable minorant* is defined by  $T_* = -(-T)^*$  and satisfies the obvious relations. Some of the properties of these functions are collected in the following lemma. It is tedious, but useful. The same is true for the other results in this section; perhaps it is better to skip them at first reading.

**1.2.2 Lemma.** The following statements are true a.s. for arbitrary maps  $S, T: \Omega \mapsto \bar{\mathbb{R}}$ , provided the statement is well-defined:

- (i)  $(S + T)^* \leq S^* + T^*$ , with equality if  $S$  is measurable;
  - (ii)  $(S - T)^* \geq S^* - T^*$ ;
  - (iii)  $|S^* - T^*| \leq |S - T|^*$ ;
  - (iv) If  $S$  is measurable, then  $(ST)^* = S1_{S>0}T^* + S1_{S<0}T_*$ ;
  - (v)  $(ST)^* \leq S^*T^*1_{S^*>0, T^*>0} + S^*T_*1_{S^*<0, T_*>0} + S_*T^*1_{S_*>0, T^*<0} + S_*T_*1_{S_*<0, T_*<0}$ ;
  - (vi)  $(1_{T>c})^* = 1_{T^*>c}$  for any  $c \in \mathbb{R}$ ;
  - (vii)  $|T|^* = T^* \vee (-T)^* = T^* \vee (-T_*) = |T^*| \vee |T_*|$ ;
  - (viii)  $(S \vee T)^* = S^* \vee T^*$ ;
  - (ix)  $(S \wedge T)^* \leq S^* \wedge T^*$  with equality if  $S$  is measurable.
- Moreover,  $P^*(T > c) = P(T^* > c)$  for any  $c \in \mathbb{R}$ .

**Proof.** Every inequality in this proof means inequality a.s. The inequality in (i) is trivial. If  $S$  is measurable and  $U \geq S + T$  and is measurable,

then  $U - S \geq T$  and is measurable, so  $U - S \geq T^*$ , which shows that  $(S + T)^* = S + T^*$ . Statement (ii) is a rearrangement and relabeling of (i). Next, (iii) follows from  $S^* - T^* \leq (S - T)^* \leq |S - T|^*$ .

In (iv) it is trivial that the left side is smaller than the right side. Suppose  $U \geq ST$  and is measurable. Then the following string of implications holds:

$$\begin{aligned} U1_{S>0} &\geq ST1_{S>0}, \\ U/S1_{S>0} &\geq T1_{S>0}, \\ U/S1_{S>0} + T^*1_{S\leq 0} &\geq T, \\ U/S1_{S>0} + T^*1_{S\leq 0} &\geq T^*, \\ U/S1_{S>0} &\geq T^*1_{S>0}, \\ U1_{S>0} &\geq ST^*1_{S>0}. \end{aligned}$$

Furthermore, since the starting inequality can also be written  $U \geq (-S)(-T)$ , we also have

$$U1_{-S>0} \geq (-S)(-T)^*1_{-S>0}.$$

Adding the last two displayed inequalities to the trivial one  $U1_{S=0} \geq 0$  yields  $U \geq S1_{S>0}T^* + S1_{S<0}T_*$ . This concludes the proof of (iv).

To obtain (v), first write  $ST \leq S^*T1_{T>0} + S_*T1_{T<0}$ . Next use (iv) repeatedly, together with a number of simple inequalities, such as  $(T1_{T>0})^* \leq T^*1_{T^*>0}$ .

In (vi) it is trivial that  $(1_{T>c})^* \leq 1_{T^*>c}$ . Suppose  $U \geq 1_{T>c}$  and is measurable. Then  $S = T^*1_{U\geq 1} + (T^*\wedge c)1_{U<1} \geq T$  and is measurable. Thus  $S \geq T^*$ , whence  $T^* \leq c$  if  $U < 1$ . So  $1_{T^*>c} \leq U$ .

It is immediate in (vii) that the functions are nondecreasing from left to right. Furthermore,  $|T^*| \leq |T|^*$  follows from (iii) applied with  $S = 0$ , while  $|T_*| = |(-T)^*| \leq |-T|^*$  by a second application of (iii).

In (viii) it is clear that  $S^* \vee T^*$  is a measurable majorant of  $S \vee T$ . This gives one-half of the equality. The other half follows from this: if  $U \geq S$  and  $U \geq T$  and is measurable, then  $U$  is greater than or equal to the measurable majorants of  $S$  and  $T$ , and hence  $U \geq S^* \vee T^*$ .

The first part of (ix) is trivial. To obtain the second part, first note that  $(S \wedge T)^* \leq S \wedge T^*$ . Next, if  $U \geq S \wedge T$  and is measurable, then  $U1_{U<S} \geq T1_{U<S}$ . Hence  $U1_{U<S} \geq (T1_{U<S})^* = T^*1_{U<S}$ , so that  $U \geq T^* \wedge S$ .

To obtain the last assertion of the theorem, write  $P(T^* > c) \geq P^*(T > c) \geq E^*1_{T>c} = E1_{T^*>c} = P(T^* > c)$ , where the inequalities are immediate consequences of the definitions and the equality follows from (vi). ■

The following lemma shows that outer probabilities are special cases of outer integrals. Furthermore, just as for outer integrals, the supremum in their definition is achieved, by measurable sets that we denote by  $B^*$ .

**1.2.3 Lemma.** For any subset  $B$  of  $\Omega$ ,

- (i)  $P^*(B) = E^*1_B$ ;  $P_*(B) = E(1_B)_*$ ;
- (ii) there exists a measurable set  $B^* \supset B$  with  $P(B^*) = P^*(B)$ ; for any such  $B^*$ , it holds that  $1_{B^*} = (1_B)^*$ ;
- (iii)  $(1_B)^* + (1_{\Omega-B})_* = 1$ .

**Proof.** From the definitions it is immediate that  $P^*(B) \geq E^*1_B$ . Next with  $A = \{1_B^* \geq 1\}$ , one has  $E^*1_B = E1_B^* \geq P(A) \geq P^*(B)$ , where the inequalities are direct consequences of the definitions. Combination yields that all the inequalities are in fact equalities. This shows the first part of (i) and (ii). The second part of (i) follows from  $P_*(B) = 1 - P^*(\Omega - B) = 1 - E(1 - 1_B)^* = 1 - E(1 - (1_B)_*)$ . The second part of (ii) follows from the trivial inequality  $1_{B^*} \geq (1_B)^*$  if  $B^* \supset B$  and  $E1_{B^*} = P(B^*) = E(1_B)^*$ . To obtain (iii), write  $(1_{\Omega-B})_* = (1 - 1_B)_* = 1 - (1_B)^*$ . ■

Though we didn't let this show up in the notation, the minimal cover function  $T^*$  depends on the underlying probability measure  $P$ . Suppose we have a set  $\mathcal{P}$  of probability measures. Is it possible to find one function that is a minimal measurable cover function for every  $P \in \mathcal{P}$  at the same time? If  $\mathcal{P}$  is dominated (by a  $\sigma$ -finite measure), the answer is affirmative.

**1.2.4 Lemma.** Let  $\mathcal{P}$  be a dominated set of probability measures on  $(\Omega, \mathcal{A})$ . Then there exists a measurable function  $T^*: \Omega \mapsto \bar{\mathbb{R}}$  with

- (i)  $T^* \geq T$ ;
- (ii)  $T^* \leq U$   $P$ -a.s., for every measurable  $U: \Omega \mapsto \bar{\mathbb{R}}$  with  $U \geq T$   $P$ -a.s., for every  $P \in \mathcal{P}$ .

**Proof.** Let  $P_0$  be a probability measure that dominates  $\mathcal{P}$ . Take  $T^*$  equal to the minimal measurable cover function that satisfies (i) and (ii) for  $\mathcal{P} = \{P_0\}$ . Given an arbitrary  $P \ll P_0$ , there is a decomposition  $\Omega = \Omega^a + \Omega^\perp$  of  $\Omega$  into disjoint, measurable sets with  $P(\Omega^\perp) = 0$  and  $P_0$  absolutely continuous with respect to  $P$  on  $\Omega^a$ . (Take densities  $p$  and  $p_0$ , and set  $\Omega^a = \{p > 0, p_0 > 0\}$  and  $\Omega^\perp = \{p = 0, \text{ or } p_0 = 0\}$ .) If  $U \geq T$ ,  $P$ -a.s., and is measurable, then  $U1_{\Omega^a} \geq T1_{\Omega^a}$ ,  $P_0$ -a.s., and is measurable. Hence  $U1_{\Omega^a} \geq (T1_{\Omega^a})^{*P_0} = T^*1_{\Omega^a}$ ,  $P_0$ -a.s. Since  $\mathcal{P}$  is dominated by  $P_0$ , this then also holds  $P$ -a.s., but then also  $U \geq T^*$ ,  $P$ -a.s., because  $P(\Omega^\perp) = 0$ . ■

Consider what happens to a minimal measurable cover if a map  $T: \Omega \mapsto \mathbb{R}$  is composed with a measurable map  $\phi: \tilde{\Omega} \mapsto \Omega$  defined on some probability space to form

$$T \circ \phi: (\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{P}) \xrightarrow{\phi} (\Omega, \mathcal{A}, \tilde{P} \circ \phi^{-1}) \xrightarrow{T} \mathbb{R}.$$

Let  $T^*$  be the minimal measurable cover of  $T$  for  $\tilde{P} \circ \phi^{-1}$ . Since  $T^* \circ \phi \geq T \circ \phi$

and is measurable, trivially  $(T \circ \phi)^* \leq T^* \circ \phi$ . The map  $\phi$  is called *perfect*<sup>†</sup> if  $(T \circ \phi)^* = T^* \circ \phi$ , for every bounded  $T: \Omega \mapsto \mathbb{R}$ . It is exactly the property that ensures that

$$E^*T \circ \phi = \int^* T d\tilde{P} \circ \phi^{-1}, \quad \text{for every bounded } T: \Omega \mapsto \mathbb{R}.$$

In particular,  $P^*(\phi \in A) = (\tilde{P} \circ \phi^{-1})^*(A)$  for every set  $A \subset \Omega$ .

Perfect maps do exist. An example that will be encountered frequently results from the following. Suppose a map  $T$  is defined on a product probability space  $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \times \mathcal{A}_2, P_1 \times P_2)$ , but really depends only on the first coordinate of  $\omega = (\omega_1, \omega_2)$ . Then  $T^*$  (for  $P_1 \times P_2$ ) can be computed (for  $P_1$ ) after just ignoring  $\Omega_2$  and thinking of  $T$  as a map on  $\Omega_1$ . More formally, suppose  $T = T_1 \circ \pi_1$ , where  $\pi_1$  is the projection on the first coordinate. Then the next lemma shows that  $T^* = T_1^* \circ \pi_1$ .

**1.2.5 Lemma.** *A coordinate projection on a product probability space (with product measure) is perfect.*

**Proof.** Let  $\pi_1: (\Omega_1 \times \Omega_2, \mathcal{A}_1 \times \mathcal{A}_2, P_1 \times P_2) \mapsto \Omega_1$  be the projection on the first coordinate and  $T: \Omega_1 \mapsto \mathbb{R}$  be bounded, but arbitrary otherwise. Let  $T^*$  be the least measurable cover of  $T$  for  $(P_1 \times P_2) \circ \pi_1^{-1} = P_1$ . Trivially  $(T \circ \pi_1)^* \leq T^* \circ \pi_1$ . Suppose  $U \geq T \circ \pi_1$ ,  $P_1 \times P_2$ -a.s. and is measurable (where  $U: \Omega_1 \times \Omega_2 \mapsto \mathbb{R}$ ). Then, by Fubini's theorem, for  $P_2$ -almost all  $\omega_2$ :  $U(\omega_1, \omega_2) \geq T(\omega_1)$  for  $P_1$ -almost all  $\omega_1$ . Since for  $\omega_2$  fixed,  $U$  is a measurable function of  $\omega_1$ , for  $P_2$ -almost all  $\omega_2$ :  $U(\omega_1, \omega_2) \geq T^*(\omega_1)$  for  $P_1$ -almost all  $\omega_1$ . By Fubini's theorem, the jointly measurable set  $\{(\omega_1, \omega_2): U < T^* \circ \pi_1\}$  is  $P_1 \times P_2$ -null. ■

Now we must consider Fubini's theorem. Unfortunately, measurability plays an important role in this result, and there is no general Fubini's theorem for outer expectations. This is the main reason why certain arguments work only under measurability assumptions.

Fubini's theorem is valid as a string of inequalities: repeated outer expectations are always less than joint outer integrals. This can only be formulated in a somewhat awkward notation. Let  $T$  be a real-valued map defined on a product space  $(\Omega_1 \times \Omega_2, \mathcal{A}_1 \times \mathcal{A}_2, P_1 \times P_2)$ . Then write  $E^*T$  for the outer expectation as before, and write  $E_1^*E_2^*T$  for the outer expectations

<sup>†</sup> There is some clash of terminology here. A probability space  $(\Omega, \mathcal{A}, P)$  is called *perfect* if for every Borel measurable map  $\phi: \Omega \mapsto \mathbb{R}$ , the collection of sets  $\{B: \phi^{-1}(B) \in \mathcal{A}\}$  is contained in the  $P \circ \phi^{-1}$ -completion of the Borel sets. (So the biggest  $\sigma$ -field on  $\mathbb{R}$  for which  $\phi$  is measurable is contained in the Borel  $\sigma$ -field up to differences in  $P \circ \phi^{-1}$ -null sets.) Dudley (1985) calls a  $\phi$  for which the condition is satisfied *quasi-perfect*, so that a probability space is perfect if every Borel measurable real function on it is quasi-perfect. Any perfect function is quasi-perfect, but not the other way around.

Note that perfectness of  $\phi$  depends on  $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{P})$  as well as on  $\phi$  and  $(\Omega, \mathcal{A})$ .

taken in turn: define for every  $\omega_1$ :

$$(E_2^*T)(\omega_1) = \inf \int U(\omega_2) dP_2(\omega_2),$$

where the infimum is taken over all measurable functions  $U: \Omega_2 \mapsto \bar{\mathbb{R}}$  with  $U(\omega_2) \geq T(\omega_1, \omega_2)$  for every  $\omega_2$  and such that  $\int U dP_2$  exists. Next  $E_1^*E_2^*T$  is the outer integral of the function  $E_2^*T: \Omega_1 \mapsto \bar{\mathbb{R}}$ . Repeated inner expectations are defined analogously.

**1.2.6 Lemma (Fubini's theorem).** *Let  $T$  be defined on a product probability space. Then  $E_*T \leq E_{1*}E_{2*}T \leq E_1^*E_2^*T \leq E^*T$ .*

**Proof.** For the inequality on the far right, it may be assumed that  $E^*T < \infty$ , so that  $E^*T = ET^*$  and  $(T^*)^+$  is integrable. Since  $T^*$  is jointly measurable for the product  $\sigma$ -field, the map  $\omega_2 \mapsto T^*(\omega_1, \omega_2)$  is a measurable majorant of  $\omega_2 \mapsto T(\omega_1, \omega_2)$  for  $P_1$ -almost all  $\omega_1$ . Hence  $\int T^*(\omega_1, \omega_2) dP_2(\omega_2) \geq (E_2^*T)(\omega_1)$  for almost every  $\omega_1$ . Here the left side is defined as the difference

$$\int (T^*)^+(\omega_1, \omega_2) dP_2(\omega_2) - \int (T^*)^-(\omega_1, \omega_2) dP_2(\omega_2),$$

in which the first term is finite for almost every  $\omega_1$  and the second is possibly infinite. By Fubini's theorem, both terms are measurable functions of  $\omega_1$ . It follows that the difference is a measurable majorant of the map  $\omega_1 \mapsto (E_2^*T)(\omega_1)$ , and its integral with respect to  $P_1$  is an upper bound for  $E_1^*E_2^*T$ , provided it exists. Since the first term of the difference is integrable, the integral of the difference exists and is the difference of the integrals. Next by Fubini's theorem, the double integrals in both terms can be replaced by a joint integral, which yields  $E(T^*)^+ - E(T^*)^- \geq E_1^*E_2^*T$ .

Finish the proof by considering  $-T$ . ■

If a function  $T$  on Euclidean space is continuous, then it is measurable and Fubini's theorem holds. In the case that  $T$  is continuous in only one argument but not necessarily measurable, it may still be possible to write the joint outer expectation as a repeated outer expectation (in one of the two possible orders). Consider a map  $T$  on a product probability space as before. Write  $T^{2*}$  for any map such that for every  $\omega_1$  the map  $\omega_2 \mapsto T^{2*}(\omega_1, \omega_2)$  is a measurable cover function for  $\omega_2 \mapsto T(\omega_1, \omega_2)$  (for  $P_2$ ).

**1.2.7 Lemma (Fubini's theorem).** *Let  $(\Omega_1, \mathcal{A}_1)$  be a separable metric space equipped with its Borel  $\sigma$ -field. Suppose the map  $T: \Omega_1 \times \Omega_2 \mapsto \mathbb{R}$  is defined on a product probability space and satisfies  $|T(\omega_1, \omega_2) - T(\omega'_1, \omega_2)| \leq d(\omega_1, \omega'_1) H(\omega_2)$ , for a function  $H$  with  $P_2^*H < \infty$  and for every sufficiently small  $d(\omega_1, \omega'_1)$ . Then  $E^*T = E_1E_2^*T$  whenever the left side is finite.*

**Proof.** For every  $n$ , partition  $\Omega_1 = \cup_{j=1}^{\infty} A_{n,j}$  into measurable sets of diameter at most  $1/n$ . Choose an arbitrary point  $a_{n,j}$  from each set  $A_{n,j}$  and define the discretization

$$T_n(\omega_1, \omega_2) = \sum_j 1_{A_{n,j}}(\omega_1) T(a_{n,j}, \omega_2).$$

Then

$$T_n^*(\omega_1, \omega_2) = \sum_j 1_{A_{n,j}}(\omega_1) T^{2*}(a_{n,j}, \omega_2).$$

Indeed, it is certainly true that  $T_n^*$  is jointly measurable and dominates  $T_n$ . If  $U \geq T_n$  and is jointly measurable, then  $U(\omega_1, \omega_2) \geq T(a_{n,j}, \omega_2)$  for every  $\omega_1 \in A_{n,j}$  and every  $\omega_2$ , whence  $U(\omega_1, \cdot) \geq T_n^{2*}(a_{n,j}, \cdot)$  almost surely for every  $\omega_1 \in A_{n,j}$ . In view of Fubini's theorem this shows that the jointly measurable set  $\{(\omega_1, \omega_2) : U(\omega_1, \omega_2) < T_n^*(\omega_1, \omega_2)\}$  is  $P_1 \times P_2$ -null.

Next  $|T_n^* - T^*| \leq |T_n - T|^* \leq 1/nH^*$ . Take the expectation on the left and right, and let  $n \rightarrow \infty$ , to conclude that

$$E^*T = \lim E^*T_n = \lim \sum_j P_1(A_{n,j}) E_2 T^{2*}(a_{n,j}, \cdot).$$

Since the map  $\omega_1 \mapsto E_2 T^{2*}(\omega_1, \cdot) = (E_2^* T)(\omega_1)$  is continuous, it follows that the right side of the preceding display is equal to  $E_1 E_2^* T$ . ■

The preceding lemma applies in particular to countable  $\Omega_1$  with discrete topology, in which case the continuity condition is automatic. In this case we also have that any versions  $T^{2*}(\omega_1, \cdot)$  (determined independently for different values of  $\omega_1$ ) automatically yield a measurable cover  $T^*(\omega_1, \omega_2) = T^{2*}(\omega_1, \omega_2)$  for  $T$ . This follows since  $T \leq T^{2*} \leq T^*$  always and for countable, discrete  $\Omega_1$ , any version  $T^{2*}$  is automatically jointly measurable.

We note that under the conditions of the preceding theorem, it need not be true that  $E^*T = E_2^* E_1 T$ .

## Problems and Complements

- (Essential infima)** Let  $(\Omega, \mathcal{A}, P)$  be a probability space and  $\mathcal{U}$  an arbitrary set of measurable functions  $U: \Omega \mapsto \bar{\mathbb{R}}$ . Show that there exists a measurable function  $U_*: \Omega \mapsto \bar{\mathbb{R}}$  with
  - $U_* \leq U$  for every  $U \in \mathcal{U}$ ;
  - $U_* \geq V$  a.s., for every measurable  $V$  with  $V \leq U$  a.s. for every  $U \in \mathcal{U}$ .
Such a function  $U_*$  is called the *essential infimum* of  $\mathcal{U}$ . Show that for a map  $T: \Omega \mapsto \bar{\mathbb{R}}$ , the minimal measurable cover  $T^*$  is the essential infimum of the set of all measurable  $U \geq T$ .

- 2. ( $E^*T$  is not always  $ET^*$ )** Let  $(\Omega, \mathcal{A}, P)$  be  $\mathbb{R}$  with the  $\sigma$ -field generated by  $[0, \infty)$  and the Borel sets in  $(-\infty, 0)$  and  $P$  equal to the Cauchy measure. The map  $T: \Omega \mapsto \mathbb{R}$  defined by  $T(\omega) = \omega$  has  $E^*T = \infty$ , but  $ET^*$  does not exist.
- 3. (Monotone convergence)** Let  $T_n, T: \Omega \mapsto \mathbb{R}$  be maps on a probability space with  $T_n \uparrow T$  pointwise on a set of probability one. Then  $T_n^* \uparrow T^*$  almost surely. If the maps are bounded from below, then  $E^*T_n \uparrow E^*T$ . For a decreasing sequence  $T_n$ , similar results are true for the lower starred versions, but not for the upper starred objects.

**[Hint:** Since  $T_n^* \leq T^*$  for every  $n$ , certainly  $\liminf T_n^* \leq \limsup T_n^* \leq T^*$ . Conversely,  $\liminf T_n^* \geq \liminf T_n = T$  and is a measurable function. If  $E^*T_n < \infty$  for every  $n$ , then  $E^*T_n = ET_n^* \uparrow ET^*$  by the monotone convergence theorem for measurable maps; consequently,  $ET^*$  exists and equals  $E^*T$ . If  $E^*T_n = \infty$  for some  $n$ , then it is infinite for every larger  $n$  and also  $E^*T = \infty$  from its definition.]

- 4. (Dominated convergence)** The “naive” dominated convergence theorem fails for outer integrals; in fact, on  $\Omega = [0, 1]$  with the uniform measure on the Borel sets, there exist maps  $T_n$  with  $T_n \downarrow 0$  everywhere and  $|T_n| \leq 1$  for every  $n$ , such that  $E^*T_n = 1$  for every  $n$ . However, if almost sure convergence is strengthened to  $|T_n - T|^* \xrightarrow{\text{as}} 0$  and  $|T_n| \leq S$  for every  $n$  and  $S$  with  $E^*S < \infty$ , then  $E^*T_n \rightarrow E^*T$ .

**[Hint:** If  $T_n \rightarrow T$  almost surely and  $|T_n| \leq S$  for every  $n$ , then  $|T_n - T|^* \leq 2S^*$ . If  $|T_n - T|^* \xrightarrow{\text{as}} 0$  and  $ES^* < \infty$ , then  $E|T_n - T|^* \rightarrow 0$ , which implies the result.]

- 5.** Let  $S, T: \Omega \mapsto \bar{\mathbb{R}}$  be arbitrary. Then  $|S^* - T_*| \leq |S - T|^* + (S^* - S_*) \wedge (T^* - T_*)$ . Also,  $|S - T|^* \leq (S^* - T_*) \vee (T^* - S_*) \leq |S - T|^* + (S^* - S_*) \wedge (T^* - T_*)$ .
- 6.** It can happen that  $(S \wedge T)^* < S^* \wedge T^*$  with probability 1.  
**[Hint:** There exist subsets of  $[0, 1]$  with  $\lambda_*(B) = 0$  and  $\lambda^*(B) = 1$ .]
- 7.** For arbitrary  $T: \Omega \mapsto \mathbb{R}$ , one has  $P(T^* \leq c) = P_*(T \leq c)$  for every  $c$ . It can happen that  $1 = P^*(T \leq c) > P(T^* \leq c) = 0$ .
- 8.** Let  $T: \Omega \mapsto \mathbb{R}$  be an arbitrary map defined on a probability space. If  $g: \mathbb{R} \mapsto \mathbb{R}$  is nondecreasing and left-continuous on an interval  $(a, b]$  that contains the range of both  $T$  and  $T^*$ , then  $g(T)^* = g(T^*)$ . If  $g$  is nonincreasing and right-continuous, then  $g(T)_* = g(T_*)$ . The continuity conditions can be replaced by the condition that  $g$  is one-to-one, is measurable, and has a measurable inverse.  
**[Hint:** That the left side is smaller than the right side is immediate from  $g(T^*) \geq g(T)$ . Suppose  $g(T^*) \geq U \geq g(T)$  almost surely for a measurable  $U$ . The “inverse”  $g^{-1}(u) = \sup\{x: g(x) \leq u\}$  satisfies  $g(x) \leq u$  if and only if  $x \leq g^{-1}(u)$ . Consequently,  $g^{-1}(U) \geq T$  almost surely. Since it is also measurable,  $g^{-1}(U) \geq T^*$ . This implies  $U \geq g(T^*)$  almost surely. The second statement follows from the first applied to the function  $x \mapsto -g(-x)$ .]

9. Let  $T_i: \Omega_i \mapsto \mathbb{R}$  be maps defined on probability spaces  $(\Omega_i, \mathcal{A}_i, P_i)$  and define  $(T_1, T_2)$  on  $\Omega_1 \times \Omega_2$  by  $(\omega_1, \omega_2) \mapsto (T_1(\omega_1), T_2(\omega_2))$ . Let  $g: \mathbb{R}^2 \mapsto \mathbb{R}$  be coordinatewise nondecreasing and left-continuous on an interval  $(a, \infty]$  that contains the range of  $(T_1, T_2)$ . Then  $g(T_1, T_2)^* = g(T_1^*, T_2^*)$ , where the first measurable cover is computed for  $P_1 \times P_2$  on the product  $\sigma$ -field  $\mathcal{A}_1 \times \mathcal{A}_2$  and the  $T_i^*$  on the right each for the corresponding  $P_i$ . In particular, for “independent”  $T_1$  and  $T_2$  as above:
- $(T_1 + T_2)^* = T_1^* + T_2^*$ ;
  - $(T_1 T_2)^* = T_1^* T_2^*$  if  $T_1 \geq 0$  and  $T_2 \geq 0$  everywhere.

[Hint: It is clear that the left side is smaller than the right side. If  $U \geq g(T_1, T_2)$  almost surely and is measurable, then, by Fubini’s theorem, for  $P_1$ -almost all  $\omega_1$ , one has  $U(\omega_1, \omega_2) \geq g(T_1(\omega_1), T_2(\omega_2))$  for  $P_2$ -almost all  $\omega_2$ . For every  $\omega_1$  for which this holds, it follows that  $U(\omega_1, \omega_2) \geq g(T_1(\omega_1), T_2^*(\omega_2))$  for  $P_2$ -almost all  $\omega_2$  by the previous exercise. Now reverse the roles of  $\omega_1$  and  $\omega_2$ .]

10. (*P-completion*) The *P-completion* of a probability space  $(\Omega, \mathcal{A}, P)$  is the triple  $(\Omega, \tilde{\mathcal{A}}, \tilde{P})$ , where  $\tilde{\mathcal{A}}$  consists of all sets  $A \cup N$  (and also  $A - N$ ) with  $A \in \mathcal{A}$  and  $N$  a subset of  $\Omega$  with  $P^*(N) = 0$ , and  $\tilde{P}(A \cup N) = P(A)$ . An equivalent description is that  $\tilde{\mathcal{A}}$  is the collection of all sets with equal inner and outer probabilities under  $P$ , and  $\tilde{P}$  maps each of these sets in this common value of  $P^*$  and  $P_*$ . A completion is a probability space, and for every measurable map  $\tilde{U}: (\Omega, \tilde{\mathcal{A}}) \mapsto \mathbb{R}$ , there is a measurable map  $U: (\Omega, \mathcal{A}) \mapsto \mathbb{R}$  with  $P^*(U \neq \tilde{U}) = 0$ .

[Hint: For the construction of  $U$ , first assume that  $\tilde{U} = \sum a_i 1_{\tilde{A}_i}$  with  $\tilde{A}_i \in \tilde{\mathcal{A}}$ . Take sets  $A_i \in \mathcal{A}$  with  $\tilde{A}_i = A_i \cup N_i$  and  $P^*(N_i) = 0$ . Then  $U = \sum a_i 1_{A_i}$  satisfies the requirements. Then a nonnegative  $\tilde{U}$  can be approximated pointwise from below by a sequence  $\tilde{U}_n$  of the form just considered. Construct  $U_n$  for every  $n$ , and set  $U = \liminf U_n$ . A general  $\tilde{U}$  can be split into its positive and negative parts.]

11. A minimal measurable cover  $T^*$  of a map  $T: (\Omega, \mathcal{A}, P) \mapsto \bar{\mathbb{R}}$  is also (a version of) a minimal measurable cover for  $T$  as a map on the *P-completion* of  $(\Omega, \mathcal{A}, P)$ .

[Hint: If  $\tilde{T}^{*c}$  is a minimal measurable cover for the completion, then there is measurable function  $T^{*c}: (\Omega, \mathcal{A}) \mapsto \mathbb{R}$  with  $T^{*c} = \tilde{T}^{*c}$ , almost surely. Check that  $T^{*c}$  is a version of  $T^*$ . So  $T^* = T^{*c} = \tilde{T}^{*c}$ , almost surely.]

12. Let  $(\Omega_i, \tilde{\mathcal{A}}_i, \tilde{P}_i)$  be the completions of probability spaces  $(\Omega_i, \mathcal{A}_i, P_i)$ , and let the probability space  $(\Omega_1 \times \Omega_2, \widetilde{\mathcal{A}_1 \times \mathcal{A}_2}, \widetilde{P_1 \times P_2})$  be the completion of their product. Then  $\widetilde{\mathcal{A}_1 \times \mathcal{A}_2} \subset \widetilde{\mathcal{A}_1 \times \mathcal{A}_2}$  and  $\widetilde{P_1 \times P_2}(\mathcal{A}_1 \times \Omega_2) = \widetilde{P}_1(\mathcal{A}_1)$ . Moreover, a map  $T: \Omega_1 \times \Omega_2 \mapsto \mathbb{R}$  of the form  $T(\omega_1, \omega_2) = T_1(\omega_1)$  is measurable for  $\widetilde{\mathcal{A}_1 \times \mathcal{A}_2}$  if and only if  $T_1$  is measurable for  $\widetilde{\mathcal{A}}_1$ . Finally,  $T^* = T_1^*$ , where the first is computed for  $P_1 \times P_2$  on  $\mathcal{A}_1 \times \mathcal{A}_2$  and the second for  $P_1$  on  $\mathcal{A}_1$ .

13. (**Outer measure**) A real function  $\mu^*$  defined on the collection of all subsets of a set  $\Omega$  is called an *outer measure* if  $\mu^*(\emptyset) = 0$ ,  $\mu^*(B_1) \leq \mu^*(B_2)$  if  $B_1 \subset B_2$  and  $\mu^*(\cup_{n=1}^{\infty} B_n) \leq \sum_{n=1}^{\infty} \mu^*(B_n)$ . A subset  $B$  of  $\Omega$  is said to be  $\mu^*$ -measurable if for every subset  $C$  one has  $\mu^*(C) = \mu^*(C \cap B) + \mu^*(C \cap \Omega - B)$ .

(According to a fundamental result of measure theory, the  $\mu^*$ -measurable sets form a  $\sigma$ -field and the restriction of  $\mu^*$  to this  $\sigma$ -field is a complete measure.) The outer probability  $P^*$  defined in the present section is an outer measure and the  $P^*$ -measurable sets are exactly the sets with  $P^*(B) = P_*(B)$ ; they are also the sets in the  $P$ -completion of  $\mathcal{A}$ .

[Hint: If  $B$  is  $P^*$ -measurable, then the outer measure rule applied to  $C = B^*$  yields  $P(B^*) = P(B^*) + P^*(B^* - B)$ .]

14. For any  $\varepsilon \in [0, 1]$ , there exists a set  $B \subset [0, 1]$  with  $\lambda_*(B) = 0$  and  $\lambda^*(B) = \varepsilon$ , where  $\lambda$  is Lebesgue measure. Hence there exist subsets  $A$  and  $B$  of  $[0, 1]$  with  $A \cap B = \emptyset$  and  $\lambda^*(A) = \lambda^*(B) = 1$ .

15. For any sets  $A$  and  $B$ ,

- (i)  $(A \cup B)^* = A^* \cup B^*$ ;     $(A \cap B)_* = A_* \cap B_*$ ;
- (ii)  $(A \cap B)^* \subset A^* \cap B^*$ ;     $(A \cup B)_* \supset A_* \cup B_*$ .

For sets  $A$  and  $B$  with  $A \cap B = \emptyset$ ,

- (iii)  $P_*(A) + P_*(B) \leq P_*(A \cup B) \leq P^*(A \cup B) \leq P^*(A) + P^*(B)$ .

The last two equalities in (iii) are always valid. The inclusions in (ii) cannot be replaced by equalities, in general, but they are equalities if one of the two sets is measurable.

16. (**Trace measure**) Let  $(\Omega, \mathcal{A}, P)$  be a probability space and  $B$  a subset of  $\Omega$ , possibly not measurable. The equality  $P_B(A \cap B) = P(A \cap B^*)$  defines a measure on the trace  $\sigma$ -field  $\mathcal{A} \cap B$  with the property  $P_B(C) = P^*(C)$  for every  $C \in \mathcal{A} \cap B$ . It is a probability measure if and only if  $P^*(B) = 1$ .

[Hint: If  $A_1$  and  $A_2$  are measurable sets with  $A_1 \cap B = A_2 \cap B$ , then their symmetric difference is measurable and disjoint with  $B$ , so it is also disjoint with (a version of)  $B^*$ . Thus  $P(A_1 \cap B^*) = P(A_2 \cap B^*)$  and  $P_B$  is well defined. It is clearly a measure. Next use that  $A \cap B^* = (A \cap B)^*$  for measurable  $A$ .]

17. Let  $(\Omega, \mathcal{A}, P)$  be a probability space and  $B \notin \mathcal{A}$ . For every constant  $p$  with  $P_*(B) \leq p \leq P^*(B)$  there exists an extension  $\tilde{P}$  of  $P$  to a  $\sigma$ -field that contains  $\mathcal{A}$  and  $B$  with the further property that  $\tilde{P}(B) = p$ .

[Hint: Take the  $\sigma$ -field equal to the collection of sets of the form  $(A \cup B) \cup (A \cap \Omega - B)$ . For  $p = P^*(B)$ , define  $\tilde{P}(A) = P_B(A \cap B) + P(A \cap (\Omega - B^*))$ , where  $P_B$  is the trace of  $P$  on  $\mathcal{A} \cap B$ . For  $p = P_*(B)$ , define  $\tilde{P}(A) = P(A \cap B_*) + P_{\Omega - B}(A \cap (\Omega - B))$ . For general  $p$ , take a linear combination.]

18. Let  $L$  be a finite Borel measure on a metric space  $\mathbb{D}$ , and let  $\mathbb{D}_0 \subset \mathbb{D}$  be an arbitrary subset. If  $G_n$  is a sequence of relatively open subsets of  $\mathbb{D}_0$  with  $\limsup G_n = \emptyset$ , then  $L^*(G_n) \rightarrow 0$ . “Relatively open” could just as well be “relatively Borel.”

[Hint: Consider the Borel measure  $L_{\mathbb{D}_0}(B) = L(B \cap \mathbb{D}_0^*)$  on  $\mathbb{D}_0$ .]

19. For  $i = 1, 2$ , let  $\phi_i: (\Omega_i, \mathcal{U}_i, P_i) \mapsto (\mathcal{X}, \mathcal{A})$  be measurable maps such that the induced measures  $P_1 \circ \phi_1^{-1} = P_2 \circ \phi_2^{-1}$  are the same (on  $\mathcal{A}$ ). Then it is not necessarily true that  $E^*T \circ \phi_1 = E^*T \circ \phi_2$  for every  $T: \mathcal{X} \mapsto [0, 1]$ .

[Hint: Take  $\mathcal{X} = \Omega_1 = \Omega_2$  and both maps  $\phi_i$  the identity. Take  $\mathcal{A} = \mathcal{U}_1$  and  $\mathcal{U}_2$  the smallest  $\sigma$ -field generated by  $\mathcal{U}_1$  and a nonmeasurable set  $B$ . Take  $T = 1_B$ .]

# 1.3

## Weak Convergence

In this section  $\mathbb{D}$  and  $\mathbb{E}$  are metric spaces with metrics  $d$  and  $e$ , respectively. The set of all continuous, bounded functions  $f: \mathbb{D} \mapsto \mathbb{R}$  is denoted  $C_b(\mathbb{D})$ .

The *Borel σ-field* on  $\mathbb{D}$  is the smallest  $\sigma$ -field containing the open sets. A function between two topological spaces is continuous if and only if the inverse image of every open set is open. Hence a continuous function is Borel measurable. This is always true; for a metric space  $\mathbb{D}$ , the case we consider throughout, there is also a converse.

**1.3.1 Lemma.** *The Borel σ-field on a metric space  $\mathbb{D}$  is the smallest σ-field making all elements of  $C_b(\mathbb{D})$  measurable (with respect to the Borel sets on  $\mathbb{R}$ ).*

**Proof.** A closed set  $F$  in  $\mathbb{D}$  is the null set  $\{x: f(x) = 0\}$  of the continuous, bounded function  $x \mapsto f(x) = d(x, F) \wedge 1$ . Hence it is contained in the  $\sigma$ -field generated by  $C_b(\mathbb{D})$ . Since the closed sets generate the Borel  $\sigma$ -field, all Borel sets must be contained in the  $\sigma$ -field generated by  $C_b(\mathbb{D})$ . The reverse inclusion was argued previously. ■

A finite Borel measure is simply a finite measure on the Borel sets.<sup>‡</sup> A Borel probability measure  $L$  is *tight* if for every  $\varepsilon > 0$  there exists a compact set  $K$  with  $L(K) \geq 1 - \varepsilon$ . A Borel measurable map  $X: \Omega \mapsto \mathbb{D}$  is called *tight* if its law  $\mathcal{L}(X) = P \circ X^{-1}$  is tight. This is equivalent

---

<sup>‡</sup> A general *Borel measure* is a measure  $\mu$  on the Borel sets with  $\mu(K) < \infty$  for every compact set  $K$ .

to there being a  $\sigma$ -compact set (a countable union of compacts) that has probability 1 under  $L$  or  $X$ . If there is a separable, measurable set with probability 1, then  $L$  or  $X$  is called *separable*. Since a  $\sigma$ -compact set in a metric space is separable, separability is slightly weaker than tightness. The following lemma shows that the two properties are the same if the metric space is complete. So the difference can be made to disappear by completing the space.<sup>b</sup> Actually, every separable Borel probability measure on a metric space is *pre-tight*: for every  $\varepsilon > 0$ , there exists a totally bounded, measurable set with probability at least  $1 - \varepsilon$ .<sup>#</sup> This concept is usually not considered, because it depends on the metric, while both separability and tightness depend on the topology only. Another topological property of a Borel probability measure is Polishness:  $L$  is *Polish* if it gives mass 1 to a Polish set.<sup>t</sup>

**1.3.2 Lemma.** *A Borel probability measure on a metric space is pre-tight if and only if it is separable. On a complete metric space separability, pre-tightness and tightness are equivalent. Any Polish Borel probability measure is tight.*

**Proof.** A pre-tight measure concentrates on a countable union of totally bounded sets; a totally bounded set is separable — so a pre-tight measure is certainly separable. Conversely, let  $\mathbb{D}_0$  be the closure of a separable set with probability 1. Take a sequence  $x_m$  that is dense in  $\mathbb{D}_0$ . For every  $\delta > 0$ , the balls of radius  $\delta$  around the  $x_m$  cover  $\mathbb{D}_0$ . Thus their union has mass 1; the union of some finitely many of them has mass at least  $1 - \varepsilon$ . Conclude that for every  $j$  there exist finitely many balls of radius  $1/j$  whose union  $G_j$  has mass at least  $1 - \varepsilon/2^j$ . The intersection  $\cap_{j=1}^{\infty} G_j$  is totally bounded and has mass at least  $1 - \varepsilon$ . This concludes the proof of the first assertion. The second is a consequence of the fact that a set is compact if and only if it is totally bounded and complete. ■

**1.3.3 Definition.** Let  $(\Omega_\alpha, \mathcal{A}_\alpha, P_\alpha)$  be a net of probability spaces and  $X_\alpha: \Omega_\alpha \rightarrow \mathbb{D}$  arbitrary maps. The net  $X_\alpha$  converges weakly to a Borel measure  $L$  if

$$\mathbf{E}^* f(X_\alpha) \rightarrow \int f dL, \quad \text{for every } f \in C_b(\mathbb{D}).$$

<sup>b</sup> Every metric space  $\mathbb{D}$  has a *completion*: a complete metric space of which a dense subset can be isometrically identified with  $\mathbb{D}$ .

<sup>#</sup> A subset of a semi-metric space is *totally bounded* if for every  $\varepsilon > 0$  it can be covered with finitely many balls of radius  $\varepsilon$ . This is equivalent to the completion of the space being compact. Also, a subset of a semimetric space is compact if and only if it is totally bounded and complete. Every totally bounded space is separable. For every semimetric space, there is a totally bounded semimetric that generates the same topology.

<sup>t</sup> A topological space is called *Polish* if it is separable and there exists a metric that generates the topology for which the space is complete. Any separable, complete metric space is Polish; so is any open subset of a Polish space. A metric space is Polish if and only if it is a  $G_\delta$ -subset of its completion (a countable intersection of open sets). Concrete examples of Polish spaces are  $\mathbb{R}$ ,  $(0,1)$ ,  $[0,1]$ , and  $\mathbb{R} - \mathbb{Q}$ .

This is denoted by  $X_\alpha \rightsquigarrow L$ . If  $X$  has a Borel law  $L$ , we also say that  $X_\alpha$  converges weakly to  $X$  and write  $X_\alpha \rightsquigarrow X$ . Throughout, it will be silently understood that the statements  $X_\alpha \rightsquigarrow L$  or  $X_\alpha \rightsquigarrow X$  include that  $L$  is a Borel measure and  $X$  Borel measurable, or in the latter case  $X$  can at least be chosen Borel measurable. The names “convergence in distribution”, “convergence in law,” and “weak star convergence” are sometimes used instead of “weak convergence.” In the present context, these terms are equivalent.

A closely related concept is weak convergence of Borel measures. A net  $L_\alpha$  of Borel measures on  $\mathbb{D}$  is said to converge weakly to  $L$  if

$$\int f dL_\alpha \rightarrow \int f dL, \quad \text{for every } f \in C_b(\mathbb{D}).$$

This is denoted  $L_\alpha \rightsquigarrow L$ . In the special case that every  $X_\alpha$  is Borel measurable, weak convergence is equivalent to weak convergence of their induced laws:  $X_\alpha \rightsquigarrow L$  if and only if  $P_\alpha \circ X_\alpha^{-1} \rightsquigarrow L$ . In general, such a reduction to induced laws is impossible. Instead, the definition must be based on outer expectations.

The measurable spaces  $(\Omega_\alpha, \mathcal{A}_\alpha, P_\alpha)$  are of crucial importance because they determine the outer expectations. They may be different for each  $\alpha$ , though in everything that is to follow it is not a loss of generality to take them all equal — the case of different spaces can be reduced to that of a single  $\Omega$  through suitable “canonical representations.”

In any case, we write  $P^*$  and  $E^*$  without an index  $\alpha$  to denote “general” probability and expectation, unless this would cause confusion. This index  $\alpha$ , of course, is understood to run through a directed set  $A$ . Some results are special to sequences; in this case  $\alpha$  will be changed to  $n$ , and it is silently understood that the directed set is formed by the natural numbers.

The portmanteau theorem gives equivalent ways of describing weak convergence. Characterization (vi) has intuitive meaning: weak convergence means convergence of probabilities of certain (but not all) sets. The other characterizations are mainly useful as technical tools.

**1.3.4 Theorem (Portmanteau).** *The following statements are equivalent:*

- (i)  $X_\alpha \rightsquigarrow L$ ;
- (ii)  $\liminf P_*(X_\alpha \in G) \geq L(G)$  for every open  $G$ ;
- (iii)  $\limsup P^*(X_\alpha \in F) \leq L(F)$  for every closed  $F$ ;
- (iv)  $\liminf E_* f(X_\alpha) \geq \int f dL$  for every lower semicontinuous  $f$  that is bounded below;<sup>†</sup>
- (v)  $\limsup E^* f(X_\alpha) \leq \int f dL$  for every upper semicontinuous  $f$  that is bounded above;

---

<sup>†</sup> A function  $f$  on  $\mathbb{D}$  is called upper semicontinuous if  $\{f \geq c\}$  is closed for every  $c$ , or equivalently  $\limsup_{y \rightarrow y_0} f(y) \leq f(y_0)$  for every  $y_0$ . A function  $f$  is lower semicontinuous if  $-f$  is upper semicontinuous.

- (vi)  $\lim P^*(X_\alpha \in B) = \lim P_*(X_\alpha \in B) = L(B)$  for every Borel set  $B$  with  $L(\delta B) = 0$ ,<sup>b</sup>
- (vii)  $\liminf E_* f(X_\alpha) \geq \int f dL$  for every bounded, Lipschitz continuous, nonnegative  $f$ .

**Proof.** The equivalence of (ii) and (iii) follows by taking complements; the equivalence of (iv) and (v) by replacing  $f$  by  $-f$ . The implication (i)  $\Rightarrow$  (vii) is trivial.

(vii)  $\Rightarrow$  (ii). Suppose (vii) holds. For every open  $G$ , there exists a sequence of Lipschitz continuous functions with  $0 \leq f_m \uparrow 1_G$ . For instance,  $f_m(x) = m d(x, \mathbb{D} - G) \wedge 1$ . For every fixed  $m$ ,  $\liminf P_*(X_\alpha \in G) \geq \liminf E_* f_m(X_\alpha) \geq \int f_m dL$ . Letting  $m \rightarrow \infty$  yields (ii).

(ii)  $\Rightarrow$  (iv). Let (ii) hold and  $f$  be lower semicontinuous with  $f \geq 0$ . Define the sequence  $f_m$  by  $f_m = \sum_{i=1}^{m^2} (1/m) 1_{G_i}$ , where  $G_i = \{x: f(x) > i/m\}$ ; this is  $f$  truncated to  $i/m$  if  $i/m < f(x) \leq (i+1)/m \leq m$  and truncated to  $m$  if  $f(x) > m$ . Thus  $f_m \leq f$  and  $|f_m - f|(x) \leq 1/m$  whenever  $f(x) \leq m$ . Fix  $m$ . Since every  $G_i$  is open

$$\liminf E_* f(X_\alpha) \geq \liminf E_* f_m(X_\alpha) \geq \sum_{i=1}^{m^2} \frac{1}{m} P(X \in G_i) = \int f_m dL.$$

Letting  $m \rightarrow \infty$  yields the assertion of (iv) for nonnegative, lower semicontinuous  $f$ . Add and subtract a constant to complete the proof.

Since a continuous function is both upper and lower semicontinuous, (v) implies (i). It remains to show the equivalence of (vi) and the others.

(ii)  $\Rightarrow$  (vi). If (ii) and (iii) hold, then

$$L(\text{int } B) \leq \liminf P_*(X_\alpha \in \text{int } B) \leq \limsup P^*(X_\alpha \in \bar{B}) \leq L(\bar{B}).$$

When  $L(\delta B) = 0$ , all inequalities in the previous display are equalities. This yields (vi).

(vi)  $\Rightarrow$  (iii). Suppose (vi) holds, and let  $F$  be closed. Write  $F^\varepsilon = \{x: d(x, F) < \varepsilon\}$ . The sets  $\delta F^\varepsilon$  are disjoint for different values of  $\varepsilon > 0$ , so that at most countably many of them can have nonzero  $L$ -measure. Choose a sequence  $\varepsilon_m \downarrow 0$  with  $L(\delta F^{\varepsilon_m}) = 0$  for every  $m$ . For fixed  $m$ ,

$$\limsup P^*(X_\alpha \in F) \leq \limsup P^*(X_\alpha \in \overline{F^{\varepsilon_m}}) = L(\overline{F^{\varepsilon_m}}).$$

Letting  $m \rightarrow \infty$  yields (iii). ■

---

<sup>b</sup> The set  $\delta B$  is the boundary of  $B$ , the closure minus the interior. A set  $B$  with  $L(\delta B) = 0$  is often called an *L-continuity set*.

**1.3.5 Example.** When  $\mathbb{D} = \mathbb{R}^k$ , the Borel  $\sigma$ -field is the usual  $\sigma$ -field generated by the cells  $(a, b]$  and every Borel measure is tight. Every Borel measure  $L$  is uniquely determined by its cumulative distribution function  $L(x) = L((-\infty, x])$ . In addition to the characterizations of the portmanteau theorem, weak convergence  $X_\alpha \rightsquigarrow L$  is equivalent to

- (viii)  $\lim P^*(X_\alpha \leq x) = \lim P_*(X_\alpha \leq x) = L(x)$  for all continuity points  $x$  of  $L$ ;

and, for measurable  $X_\alpha$ , also to

$$(ix) \lim E e^{it' X_\alpha} = \int e^{it' x} dL(x) \text{ for every } t \in \mathbb{R}^k.$$

The proofs are omitted. They can be based on Prohorov's theorem combined with the fact that a probability distribution on  $\mathbb{R}^k$  is uniquely determined by its cumulative distribution function or characteristic function.

It is immediate from the definition that  $X_\alpha \rightsquigarrow X$  implies  $g(X_\alpha) \rightsquigarrow g(X)$  for every continuous  $g$ . This is actually already true under only the condition that  $g$  is continuous almost surely under  $X$ . The continuous mapping theorem is not a very deep result, but it is what makes the concept of weak convergence successful.

**1.3.6 Theorem (Continuous mapping).** *Let  $g: \mathbb{D} \mapsto \mathbb{E}$  be continuous at every point of a set  $\mathbb{D}_0 \subset \mathbb{D}$ . If  $X_\alpha \rightsquigarrow X$  and  $X$  takes its values in  $\mathbb{D}_0$ , then  $g(X_\alpha) \rightsquigarrow g(X)$ .*

**Proof.** The set  $D_g$  of all points at which  $g$  is discontinuous can be written as  $D_g = \bigcup_{m=1}^{\infty} \bigcap_{k=1}^{\infty} G_k^m$ , where  $G_k^m$  is the set of all  $x$  for which there are  $y$  and  $z$  within the open ball of radius  $1/k$  around  $x$  with  $e(g(y), g(z)) > 1/m$ . Every  $G_k^m$  is open, so  $D_g$  is a Borel set. For every closed  $F$ , it holds that

$$\overline{g^{-1}(F)} \subset g^{-1}(F) \cup D_g.$$

Since  $g$  is continuous on the range of the limit variable,  $g(X)$  can be chosen Borel measurable. By the portmanteau theorem,  $\limsup P^*(g(X_\alpha) \in F) \leq \limsup P^*(X_\alpha \in g^{-1}(F)) \leq P(X \in g^{-1}(F))$ . Since the set of discontinuities has probability zero under  $X$ , the last expression equals  $P(g(X) \in F)$ . Finally, apply the portmanteau theorem again. ■

Next to the continuous mapping theorem, Prohorov's theorem is the most important theorem on weak convergence. To formulate the result, two new concepts are needed.

**1.3.7 Definition.** The net of maps  $X_\alpha$  is *asymptotically measurable* if and only if

$$E^* f(X_\alpha) - E_* f(X_\alpha) \rightarrow 0, \quad \text{for every } f \in C_b(\mathbb{D}).$$

The net  $X_\alpha$  is *asymptotically tight* if for every  $\varepsilon > 0$  there exists a compact set  $K$  such that

$$\liminf P_*(X_\alpha \in K^\delta) \geq 1 - \varepsilon, \quad \text{for every } \delta > 0.$$

Here  $K^\delta = \{y \in \mathbb{D}: d(y, K) < \delta\}$  is the “ $\delta$ -enlargement” around  $K$ .

The  $\delta$  in the definition of tightness may seem a bit overdone. It is not — asymptotic tightness as defined is essentially weaker than the same condition but with  $K$  instead of  $K^\delta$ . This is caused by a second difference with the classical concept of uniform tightness<sup>#</sup>: the (enlarged) compacts need to contain mass  $1 - \varepsilon$  only in the limit.

On the other hand, nothing is gained in simple cases: for Borel measurable maps in a Polish space, asymptotic tightness and uniform tightness are the same (Problem 1.3.9). It may also be noted that, although  $K^\delta$  is dependent on the metric, the property of asymptotic tightness depends on the topology only (Problem 1.3.6). One nice consequence of the present tightness concept is that weak convergence usually implies asymptotic measurability and tightness.

### 1.3.8 Lemma.

- (i) If  $X_\alpha \rightsquigarrow X$ , then  $X_\alpha$  is asymptotically measurable.
- (ii) If  $X_\alpha \rightsquigarrow X$ , then  $X_\alpha$  is asymptotically tight if and only if  $X$  is tight.

**Proof.** (i). This follows upon applying the definition of weak convergence to both  $f$  and  $-f$ .

(ii). Fix  $\varepsilon > 0$ . If  $X$  is tight, then there is a compact  $K$  with  $P(X \in K) > 1 - \varepsilon$ . By the portmanteau theorem,  $\liminf P_*(X_\alpha \in K^\delta) \geq P(X \in K^\delta)$ , which is larger than  $1 - \varepsilon$  for every  $\delta > 0$ . Conversely, if  $X_\alpha$  is tight, then there is a compact  $K$  with  $\liminf P_*(X_\alpha \in K^\delta) \geq 1 - \varepsilon$ . By the portmanteau theorem,  $P(X \in \overline{K^\delta}) \geq 1 - \varepsilon$ . Let  $\delta \downarrow 0$ . ■

The next version of Prohorov’s theorem may be considered a converse of the previous lemma. It comes in two parts, one for nets and one for sequences, neither one of which implies the other. The sequence case is the deepest of the two.

### 1.3.9 Theorem (Prohorov’s theorem).

- (i) If the net  $X_\alpha$  is asymptotically tight and asymptotically measurable, then it has a subnet  $X_{\alpha(\beta)}$  that converges in law to a tight Borel law.
- (ii) If the sequence  $X_n$  is asymptotically tight and asymptotically measurable, then it has a subsequence  $X_{n_j}$  that converges weakly to a tight Borel law.

---

<sup>#</sup> A collection of Borel measurable maps  $X_\alpha$  is called *uniformly tight* if, for every  $\varepsilon > 0$ , there is a compact  $K$  with  $P(X_\alpha \in K) \geq 1 - \varepsilon$  for every  $\alpha$ .

**Proof.** (i). Consider  $(E^* f(X_\alpha))_{f \in C_b(\mathbb{D})}$  as a net in the product space

$$\prod_{f \in C_b(\mathbb{D})} [-\|f\|_\infty, \|f\|_\infty].$$

By Tychonov's theorem, this space is compact in the product topology. Hence the given net has a converging subnet. This is equivalent to the existence of constants  $L(f) \in [-\|f\|_\infty, \|f\|_\infty]$  such that

$$E^* f(X_{\alpha(\beta)}) \rightarrow L(f), \quad \text{for every } f \in C_b(\mathbb{D}).$$

It suffices to show that the functional  $L: C_b(\mathbb{D}) \mapsto \mathbb{R}$  is representable by a Borel probability measure  $L$  in the sense that  $L(f) = \int f dL$  for every  $f \in C_b(\mathbb{D})$ .

Because of the asymptotic measurability, the numbers  $L(f)$  are also the limits of the corresponding inner expectations  $E_* f(X_\alpha)$ . Consequently,

$$\begin{aligned} L(f_1 + f_2) &\leq \lim \left( E^* f_1(X_{\alpha(\beta)}) + E^* f_2(X_{\alpha(\beta)}) \right) \\ &= L(f_1) + L(f_2) \\ &= \lim \left( E_* f_1(X_{\alpha(\beta)}) + E_* f_2(X_{\alpha(\beta)}) \right) \leq L(f_1 + f_2). \end{aligned}$$

Thus  $L: C_b(\mathbb{D}) \mapsto \mathbb{R}$  is additive. By a similar argument, it follows that  $L(\lambda f) = \lambda L(f)$  for every  $\lambda \in \mathbb{R}$ . So  $L$  is even linear. Trivially, it is positive: if  $f \geq 0$ , then  $L(f) \geq 0$ .

Finally,  $L$  also has another property of an integral: if  $f_m \downarrow 0$  pointwise, then  $L(f_m) \downarrow 0$ . Indeed, fix  $\varepsilon > 0$ . There is a compact  $K$  such that  $\liminf P_*(X_\alpha \in K^\delta) \geq 1 - \varepsilon$  for every  $\delta > 0$ . By Dini's theorem,  $f_m \downarrow 0$  uniformly on compacts. Hence for sufficiently large  $m$ , it holds that  $|f_m(x)| \leq \varepsilon$  for every  $x \in K$ . Fix an  $m$  for which this holds. By an easy argument using the compactness of  $K$ , there exists a  $\delta > 0$  such that  $|f_m(x)| \leq 2\varepsilon$  for every  $x \in K^\delta$ . One has that  $(1_{X_\alpha \in K^\delta})_* = 1_{A_\alpha}$  for a measurable set  $A_\alpha \subset \{X_\alpha \in K^\delta\}$ . Hence

$$L(f_m) = \lim E f_m(X_\alpha)^* ((1_{X_\alpha \in K^\delta})_* + (1_{X_\alpha \notin K^\delta})^*) \leq 2\varepsilon + \|f_m\|_\infty \varepsilon.$$

It has been shown that  $L$  has all the properties of an *abstract integral*. By the Daniell-Stone theorem<sup>†</sup>,  $L$  is representable by a Borel probability measure  $L$ . This concludes the proof of (i).

(ii). The proof of (ii) is the same as for (i) except in the first part, where for (ii) it has to be shown that there is a subsequence, rather than a subnet, along which outer expectations converge. This is achieved in the following manner.

For  $m \in \mathbb{N}$ , let  $K_m$  be a compact with  $\liminf P_*(X_n \in K_m^\delta) \geq 1 - 1/m$  for every  $\delta > 0$ . Since  $K_m$  is compact, the space  $C_b(K_m)$ , and hence also

---

<sup>†</sup> Bauer (1981), Theorem 7.1.4, or Dudley (1989), Theorem 4.5.2.

its unit ball  $\{f \in C_b(K_m) : |f(x)| \leq 1 \text{ for every } x \in K_m\}$ , is separable. According to Tietze's extension theorem<sup>†</sup>, every continuous function defined on a closed subset of a normal space with values in  $[-1, 1]$  can be extended to a continuous function on the whole space with values in  $[-1, 1]$ . Hence every  $f$  in the unit ball of  $C_b(K_m)$  can be extended to an  $f$  in the unit ball of  $C_b(\mathbb{D})$ . Combination of these facts yields that there exists a countable subset of the unit ball of  $C_b(\mathbb{D})$  of which the restrictions to  $K_m$  are dense in the unit ball of  $C_b(K_m)$ .

Pick such a countable set for every  $m$ , and let  $\mathcal{F}$  be the countably many functions obtained this way. For a fixed, bounded  $f$ , there clearly is a subsequence  $X_{n_j}$  such that  $E^* f(X_{n_j})$  converges to some number. By a diagonalization argument, obtain a subsequence such that

$$E^* f(X_{n_j}) \rightarrow L(f), \quad \text{for every } f \in \mathcal{F},$$

for numbers  $L(f) \in [-1, 1]$ .

Next let  $f \in C_b(\mathbb{D})$  take values in  $[-1, 1]$ , but be arbitrary otherwise. Fix  $\varepsilon > 0$  and  $m$ . There exists  $f_m \in \mathcal{F}$  with  $|f(x) - f_m(x)| \leq \varepsilon$ , for every  $x \in K_m$ . Then, as before, there exists  $\delta > 0$  such that  $|f(x) - f_m(x)| \leq 2\varepsilon$ , for every  $x \in K_m^\delta$ . Then

$$\begin{aligned} & |E^* f(X_n) - E^* f_m(X_n)| \\ & \leq E|f(X_n) - f_m(X_n)|^*(1_{X_n \in K_m^\delta})_* + 2P^*(X_n \notin K_m^\delta) \leq 2\varepsilon + 2/m, \end{aligned}$$

for sufficiently large  $n$ . Conclude that the sequence  $E^* f(X_{n_j})$  has the property that, for every  $\eta > 0$ , there is a converging subsequence of numbers that is eventually within distance  $\eta$ . This implies that  $E^* f(X_{n_j})$  itself converges to a limit.

Complete the proof as under (i). ■

Of course, under the conditions of Prohorov's theorem — that  $X_\alpha$  is asymptotically tight and measurable — the seemingly stronger conclusion is true that every subnet of  $X_\alpha$  has a further subnet that converges to a tight Borel law. A net with this property is called *relatively compact*. Relatively compact *sequences* are defined analogously, with subnets replaced by subsequences. The converse of this stronger form of Prohorov's theorem is false: a relatively compact net or sequence is not necessarily asymptotically tight. However, if, in addition, all limit points concentrate on a fixed Polish space, then a relatively compact net is asymptotically tight (Problem 1.12.4). Thus for Borel measures on Polish spaces the concepts “relatively compact,” “asymptotically tight,” and “uniformly tight” are all equivalent.

The next theorem is trivial but important. In many applications there are several possible spaces  $\mathbb{D}$  in which one could consider weak convergence.

---

<sup>†</sup> Jameson (1974), Theorem 12.4.

For instance, a net of positive real-valued maps can be considered as maps in  $\mathbb{R}^+$ , in  $\mathbb{R}$ , or even in  $\mathbb{R}^2$ . Less trivial examples arise with infinite-dimensional spaces. It wouldn't be nice if our choice of  $\mathbb{D}$  influenced the conclusion too much. Fortunately, as long as the topology is not changed, this is not the case.

**1.3.10 Theorem.** *Let  $\mathbb{D}_0 \subset \mathbb{D}$  be arbitrary, and let  $X$  and  $X_\alpha$  take their values in  $\mathbb{D}_0$ . Then  $X_\alpha \rightsquigarrow X$  as maps in  $\mathbb{D}_0$  if and only if  $X_\alpha \rightsquigarrow X$  as maps in  $\mathbb{D}$ . Here  $\mathbb{D}_0$  and  $\mathbb{D}$  are equipped with the same metric.*

**Proof.** Because a set  $G_0$  in  $\mathbb{D}_0$  is open if and only if it is of the form  $G \cap \mathbb{D}_0$  for an open set  $G$  in  $\mathbb{D}$ , this is an easy corollary of (ii) of the portmanteau theorem. ■

**1.3.11 Example (Weak convergence of discrete measures).** Let  $X$  and every  $X_\alpha$  be maps taking their values in a countable subset  $S \subset \mathbb{D}$  of isolated points. (Every  $s \in S$  has an open neighbourhood that contains no other points of  $S$ .) Then  $X_\alpha \rightsquigarrow X$  if and only if  $X$  is Borel measurable and  $P_*(X_\alpha = s) \rightarrow P(X = s)$  for every  $s \in S$ . In this case, both  $P^*(X_\alpha \in B)$  and  $P_*(X_\alpha \in B)$  converge to  $P(X \in B)$  for every Borel set  $B \subset \mathbb{D}$ .

Indeed, in view of the previous theorem, it may be assumed without loss of generality that  $S = \mathbb{D}$ . Then every set  $B \subset \mathbb{D}$  is open and closed at the same time and therefore has empty boundary. If  $X_\alpha \rightsquigarrow X$ , then outer and inner probabilities of every set converge by the portmanteau theorem. Conversely, suppose inner measures of one-point sets converge. Let  $G$  be arbitrary. For any finite subset  $K$  of  $G$ , one has  $P_*(X_\alpha \in K) \geq \sum_{s \in K} P_*(X_\alpha = s) \rightarrow P(X \in K)$ . Conclude that  $\liminf P_*(X_\alpha \in G) \geq P(X \in G)$  for every  $G$ , in particular open ones, so that  $X_\alpha \rightsquigarrow X$  by the portmanteau theorem.

The condition that the points of  $S$  are isolated cannot be omitted. It is crucial in the above that both  $X$  and every  $X_\alpha$  take their values in  $S$ .

As a consequence of the previous theorem, a net that converges weakly to a separable limit is asymptotically tight when seen as maps into the completion of  $\mathbb{D}$ . There are no known examples of nonseparable Borel measures. Thus, there is no great loss of generality in considering only those weakly convergent nets that are asymptotically tight. (It is actually known that, starting from the usual axioms of mathematics, including the axiom of choice, it is impossible to construct nonseparable Borel measures — the axiom that nonseparable Borel measures do not exist can be added to the Zermelo-Frankel system without creating inconsistencies. It is apparently unknown whether nonseparable Borel measures can consistently exist.)

A question that has been ignored so far is whether weak limits are unique — they are. A Borel measure is uniquely defined by the expectations it gives to the elements of  $C_b(\mathbb{D})$ . Tight Borel measures are already determined by the expectations of relatively small subclasses of  $C_b(\mathbb{D})$ .

**1.3.12 Lemma.** Let  $L_1$  and  $L_2$  be finite Borel measures on  $\mathbb{D}$ .

(i) If  $\int f dL_1 = \int f dL_2$  for every  $f \in C_b(\mathbb{D})$ , then  $L_1 = L_2$ .

Let  $L_1$  and  $L_2$  be tight Borel probability measures on  $\mathbb{D}$ .

(ii) If  $\int f dL_1 = \int f dL_2$  for every  $f$  in a vector lattice<sup>b</sup>  $\mathcal{F} \subset C_b(\mathbb{D})$  that contains the constant functions and separates points of  $\mathbb{D}$ , then  $L_1 = L_2$ .

**Proof.** (i). For every open  $G$ , there exists a sequence of continuous functions with  $0 \leq f_m \uparrow 1_G$  (compare the proof of the portmanteau theorem). By monotone convergence,  $L_1(G) = L_2(G)$  for every open  $G$ . Since  $L_1(\mathbb{D}) = L_2(\mathbb{D})$ , the collection of Borel sets for which  $L_1(B) = L_2(B)$  is a  $\sigma$ -field.

(ii). Fix  $\varepsilon > 0$ . Take a compact  $K$  such that  $L_1(K) \wedge L_2(K) \geq 1 - \varepsilon$ . According to a version of the Stone-Weierstrass theorem<sup>#</sup>, a vector lattice  $\mathcal{F} \subset C_b(K)$  that contains the constant functions and separates points of  $K$  is uniformly dense in  $C_b(K)$ . Given  $g \in C_b(\mathbb{D})$  with  $0 \leq g \leq 1$ , take  $f \in \mathcal{F}$  with  $|g(x) - f(x)| \leq \varepsilon$  for every  $x \in K$ . Then  $|\int g dL_1 - \int g dL_2| \leq |\int (f \wedge 1)^+ dL_1 - \int (f \wedge 1)^+ dL_2| + 4\varepsilon$ , which equals  $4\varepsilon$  because  $(f \wedge 1)^+ \in \mathcal{F}$ . Conclude that  $\int g dL_1 = \int g dL_2$ . By adding and multiplying with scalars, obtain the same result for every  $g \in C_b(\mathbb{D})$ . ■

Requiring only asymptotic measurability of a net  $X_\alpha$ , as opposed to Borel measurability of every  $X_\alpha$ , extends the applicability of weak convergence considerably. However, asymptotic measurability is hard to establish directly. The most promising method seems to be to prove measurability into some smaller  $\sigma$ -field plus some additional property. For asymptotically tight nets, the situation is nice — asymptotic tightness plus only a little bit of measurability gives asymptotic measurability. The next lemma is an abstract version of this result. Interesting special cases are discussed in the next sections.

**1.3.13 Lemma.** Let the net  $X_\alpha$  be asymptotically tight, and suppose  $E^*f(X_\alpha) - E_*f(X_\alpha) \rightarrow 0$  for every  $f$  in a subalgebra  $\mathcal{F}$  of  $C_b(\mathbb{D})$  that separates points of  $\mathbb{D}$ .<sup>†</sup> Then the net  $X_\alpha$  is asymptotically measurable.

**Proof.** Fix  $\varepsilon > 0$  and a compact  $K$  such that  $\limsup P^*(X_\alpha \notin K^\delta) \leq \varepsilon$  for every  $\delta > 0$ . Assume without loss of generality that  $\mathcal{F}$  contains the constant functions. By the Stone-Weierstrass theorem, the restrictions of the functions in  $\mathcal{F}$  to  $K$  are uniformly dense in  $C_b(K)$ . Hence given  $f \in$

<sup>b</sup> A vector lattice  $\mathcal{F} \subset C_b(\mathbb{D})$  is a vector space that is closed under taking positive parts: if  $f \in \mathcal{F}$ , then  $f^+ = f \vee 0 \in \mathcal{F}$ . Then automatically  $f \vee g \in \mathcal{F}$  and  $f \wedge g \in \mathcal{F}$  for every  $f, g \in \mathcal{F}$ . A set of functions on  $\mathbb{D}$  separates points of  $\mathbb{D}$  if, for every pair  $x \neq y \in \mathbb{D}$ , there is  $f \in \mathcal{F}$  with  $f(x) \neq f(y)$ .

<sup>#</sup> Jameson (1974), p. 263.

<sup>†</sup> An algebra  $\mathcal{F} \subset C_b(\mathbb{D})$  is a vector space that is closed under taking products: if  $f, g \in \mathcal{F}$ , then  $fg \in \mathcal{F}$ .

$C_b(\mathbb{D})$ , there exists  $g \in \mathcal{F}$  with  $|f(x) - g(x)| \leq \varepsilon/4$  for every  $x \in K$ . Using the compactness of  $K$ , it is easily seen that there is a  $\delta > 0$  such that  $|f(x) - g(x)| \leq \varepsilon/3$  for every  $x \in K^\delta$ . Let  $\{X_\alpha \in K^\delta\}_*$  be a measurable set contained in  $\{X_\alpha \in K^\delta\}$  and with the same inner measure. Then  $P(\Omega_\alpha - \{X_\alpha \in K^\delta\}_*) = P^*(X_\alpha \notin K^\delta)$  and, for large  $\alpha$ ,

$$\begin{aligned} P(|f(X_\alpha)^* - f(X_\alpha)_*| > \varepsilon) \\ &\leq P(|f(X_\alpha)^* - f(X_\alpha)_*| > \varepsilon \cap \{X_\alpha \in K^\delta\}_*) + 2\varepsilon \\ &\leq P(|g(X_\alpha)^* - g(X_\alpha)_*| > \varepsilon/3) + 2\varepsilon. \end{aligned}$$

Hence  $f(X_\alpha)^* - f(X_\alpha)_* \rightarrow 0$  in probability. By dominated convergence,  $E(f(X_\alpha)^* - f(X_\alpha)_*) \rightarrow 0$ . ■

## Problems and Complements

1. (**Regularity of Borel measures**) Every Borel probability measure  $L$  on a metric space is *outer regular*: for every Borel set  $B$ ,

$$L(B) = \sup_{\substack{F \subset B \\ F \text{ closed}}} L(F) = \inf_{\substack{G \supset B \\ G \text{ open}}} L(G).$$

A Borel probability measure  $L$  is called *inner regular* if, for every Borel set  $B$ ,

$$L(B) = \sup_{\substack{K \subset B \\ K \text{ compact}}} L(K).$$

A Borel probability measure on a metric space is inner regular if and only if it is tight. In particular, every Borel probability measure on a Polish space is inner regular. An inner regular measure is sometimes called a *Radon measure*. The first equation is also true with “ $F$  closed” replaced by “ $F$  totally bounded”. (More generally, every finite Borel measure on a Suslin space is regular.)

[**Hint:** The collection of all Borel sets  $B$  for which the first equation is valid can be seen to be a  $\sigma$ -field. It includes all open and closed sets.]

2. The metrics  $d(x, y) = |x - y|$  and  $e(x, y) = |\arctan x - \arctan y|$  generate the same topology on  $\mathbb{R}$ . However,  $\mathbb{R}$  is complete for  $d$ , but incomplete for  $e$ . The completion of  $\mathbb{R}$  under  $e$  is the extended real line  $[-\infty, \infty]$ . The real line is totally bounded for  $e$ , but not for  $d$ .
3. The *Baire*  $\sigma$ -field on a topological space  $\mathbb{D}$  (not necessarily metrizable) is the smallest  $\sigma$ -field for which all continuous (bounded) functions  $f: \mathbb{D} \mapsto \mathbb{R}$  are measurable. It is the smallest  $\sigma$ -field containing all open  $F_\sigma$ -sets and/or closed  $G_\delta$ -sets.

[**Hint:** Bauer (1981), page 206.]

**4. (Canonical representation)** Let  $X_\alpha: \Omega_\alpha \mapsto \mathbb{D}$  be an arbitrary collection of maps indexed by  $\alpha \in A$  and defined on probability spaces  $(\Omega_\alpha, \mathcal{A}_\alpha, P_\alpha)$ . Let  $(\Omega, \mathcal{A}, P)$  be the product of these probability spaces, and define  $\tilde{X}_\alpha: \Omega \mapsto \mathbb{D}$  by  $\tilde{X}_\alpha = X_\alpha \circ \pi_\alpha$ , where  $\pi_\alpha: \Omega \mapsto \Omega_\alpha$  is the projection on the  $\alpha$ -th coordinate. Then  $E^* f(\tilde{X}_\alpha) = E^* f(X_\alpha)$  for every bounded  $f: \mathbb{D} \mapsto \mathbb{R}$ . Consequently, whenever one has to do with maps  $X_\alpha$  and is interested only in their “laws,” it is no loss of generality to assume that the maps are defined on a single probability space.

**5.** Let  $d$  and  $e$  be metrics on the same set  $\mathbb{D}$  with the following property: if  $x_n \rightarrow x$  for  $e$  and a limit  $x \in \mathbb{D}_0$ , then  $x_n \rightarrow x$  for  $d$ ; for every sequence  $x_n$  and a given subset  $\mathbb{D}_0$  of  $\mathbb{D}$ . Then  $X_\alpha \rightsquigarrow X$  for  $e$ , where  $X$  takes its values in  $\mathbb{D}_0$ , implies the same for  $d$ .

[Hint: Apply the continuous mapping theorem to the identity map from  $(\mathbb{D}, e)$  to  $(\mathbb{D}, d)$ .]

**6.** A net  $X_\alpha$  is asymptotically tight if and only if, for every  $\varepsilon > 0$ , there exists a compact set  $K$  with  $\liminf P_*(X_\alpha \in G) \geq 1 - \varepsilon$  for every open  $G \supset K$ .

[Hint: For every open  $G \supset K$  where  $K$  is compact, there is a  $\delta > 0$  with  $G \supset K^\delta \supset K$ . If there were not such a  $\delta$ , then there would be a sequence  $x_n$  with  $d(x_n, K) \rightarrow 0$  and  $x_n \notin G$  for every  $n$ . By compactness of  $K$ , a subsequence would converge; the limit would be both in  $K$  and not in  $G$ .]

**7.** Let  $X_\alpha$  be maps in  $\mathbb{D}$  and  $g: \mathbb{D} \mapsto \mathbb{E}$  continuous.

- (i) If  $X_\alpha$  is asymptotically tight, then  $g(X_\alpha)$  is asymptotically tight.
- (ii) If  $X_\alpha$  is asymptotically measurable, then  $g(X_\alpha)$  is asymptotically measurable.

Is full continuity of  $g$  necessary?

[Hint: The function  $x \mapsto e(g(x), g(K))$  is continuous and identically zero on  $K$ . For a compact  $K$ , there exists for every  $\varepsilon > 0$  a  $\delta > 0$  with  $e(g(x), g(K)) < \varepsilon$  whenever  $d(x, K) < \delta$ .]

**8.** If  $X_\alpha \rightsquigarrow X$  and  $X$  is separable, then  $X_\alpha$  is asymptotically pre-tight: for every  $\varepsilon > 0$ , there exists a totally bounded measurable set  $K$  with  $\liminf P_*(X_\alpha \in K^\delta) \geq 1 - \varepsilon$  for every  $\delta > 0$ . Asymptotic pre-tightness cannot replace asymptotic tightness in Prohorov’s theorem: a net that is measurable and asymptotically pre-tight is not necessarily relatively compact.

[Hint: For a counterexample, let  $X_n$  be maps in the rationals  $\mathbb{D} = \mathbb{Q}$  with  $P(X_n = i/n) = 1/(n+1)$ , for  $i = 0, 1, \dots, n$ .]

**9. (Asymptotic and uniform tightness of a sequence)** According to the “classical” definition, a sequence of Borel measurable maps  $X_n$  is *uniformly tight* if, for every  $\varepsilon > 0$ , there exists a compact  $K$  with  $P(X_n \in K) \geq 1 - \varepsilon$  for every  $n$ . A sequence  $X_n$  is uniformly tight if and only if it is asymptotically tight and each element  $X_n$  is tight. In particular, for Borel measurable sequences in a Polish space, uniform tightness and asymptotic tightness are the same.

[Hint: Fix  $\varepsilon > 0$ . Take a compact  $K_0$  with  $\liminf P(X_n \in K_0^\delta) \geq 1 - \varepsilon$  for every  $\delta > 0$ . Choose  $n_1 < n_2 < \dots$  such that  $P(X_n \in K_0^{1/m}) \geq 1 - 2\varepsilon$

for  $n \geq n_m$ . For  $n_m < n \leq n_{m+1}$ , choose a compact  $K_n$  with  $P(X_n \in K_0^{1/m} - K_n) < \varepsilon$ . Now  $K = \cup_{n=0}^{\infty} K_n$  is compact and  $P(X_n \in K) \geq 1 - 3\varepsilon$  for every  $n$ .]

10. A relatively compact sequence of Borel measures on a metric space is not necessarily asymptotically tight.

[**Hint:** For every rational  $q$ , let  $L_q$  be the measure on  $\ell^\infty(\mathbb{Q})$  with  $L_q\{0\} = L_q\{z_q\} = 1/2$ , where 0 is the function that is identically zero and  $z_q$  is the function that is zero except at the point  $q$ , where it is 1. (So  $L_q$  is a two-point measure.) Take a sequence  $\{q_n : n = 1, 2, \dots\}$  of rational numbers that reaches every rational number infinitely often, e.g. 0, 1,  $-1, 0, 1/2, -1/2, 1, -1, 3/2, -3/2, 2, -2, 0, 1/3, \dots$ . Then the sequence  $L_{q_n}$  has all measures  $L_q$  as weak limit points, and it is relatively compact. If the sequence were asymptotically tight, then it would be uniformly tight (in the old sense), because every  $L_q$  is tight. But there is no compact  $K$  with  $L_q(K) > 3/4$  for every  $q$ . Indeed,  $K$  would have to contain both 0 and  $z_q$  for every  $q$ , which is impossible, since  $\|z_q - z_r\| = 1$  for every pair of distinct rationals  $q, r$ .]

11. Define  $X_n$  as  $Z_n(2 \log(n+1))^{-1/2}$  for i.i.d. standard normal variables  $Z_n$ . Then  $(X_1, X_2, \dots)$  takes its values with probability 1 in  $\ell_\infty$ , but it does not induce a tight Borel law on this space.

[**Hint:** By Anderson's lemma, every tight, Borel measurable centered Gaussian variable  $X$  satisfies  $P(\|X - x\| \leq \varepsilon) \leq P(\|X\| \leq \varepsilon)$  for every  $x$ . Conclude from this that  $P(\|X\| \leq \varepsilon)$  is positive for every  $\varepsilon > 0$ . For the given process, this probability is zero for  $\varepsilon < 1$ . (Instead of Anderson's lemma, one may use the inequality  $P(\|X - x\| \leq \varepsilon)^2 \leq P(\|X\| \leq \varepsilon\sqrt{2})$ , which can be established easily by rewriting the left side as  $P(\|X - x\| \leq \varepsilon, \|Y - x\| \leq \varepsilon)$  for an independent copy  $Y$  of  $X$ .)]

## 1.4

# Product Spaces

Let  $\mathbb{D}$  and  $\mathbb{E}$  be metric spaces with metrics  $d$  and  $e$ . Then the Cartesian product  $\mathbb{D} \times \mathbb{E}$  is a metric space for any of the metrics

$$\begin{aligned} c((x_1, y_1), (x_2, y_2)) &= d(x_1, x_2) \vee e(y_1, y_2), \\ c((x_1, y_1), (x_2, y_2)) &= \sqrt{d(x_1, x_2)^2 + e(y_1, y_2)^2}, \\ c((x_1, y_1), (x_2, y_2)) &= d(x_1, x_2) + e(y_1, y_2). \end{aligned}$$

These generate the same topology, the *product topology*.

On the product space  $\mathbb{D} \times \mathbb{E}$ , there are two natural  $\sigma$ -fields: the product of the Borel  $\sigma$ -fields and the Borel  $\sigma$ -field for the product topology. In general, these are not the same. The reason is that a product topology is built up from *arbitrary* unions of (open) cylinders  $G_1 \times G_2$ , whereas the product  $\sigma$ -field is generated through *countable* set-theoretic operations on these sets. A separable metric space has a countable base for its topology. Not surprisingly, for separable spaces the two  $\sigma$ -fields are the same.

**1.4.1 Lemma.** *If  $\mathbb{D}$  and  $\mathbb{E}$  are separable, then the product of the Borel  $\sigma$ -fields equals the Borel  $\sigma$ -field for the product topology on  $\mathbb{D} \times \mathbb{E}$ .*

**1.4.2 Lemma.** *A separable Borel probability measure  $L$  on  $\mathbb{D} \times \mathbb{E}$  is uniquely determined by the numbers  $\int f(x)g(y) dL(x, y)$ , where  $f$  and  $g$  range over the nonnegative Lipschitz functions in  $C_b(\mathbb{D})$  and  $C_b(\mathbb{E})$ , respectively.*

**Proofs.** Since  $\mathbb{D}$  and  $\mathbb{E}$  are separable (and metric), it is possible to find countable bases  $\mathcal{G}_1$  and  $\mathcal{G}_2$  for the open sets of their topologies. A set in

$\mathbb{D} \times \mathbb{E}$  is open if and only if it is the union of sets of the form  $G_1 \times G_2$  with  $G_1 \in \mathcal{G}_1$  and  $G_2 \in \mathcal{G}_2$ . Since there are only countably many sets of the latter type, a set is open in the product topology if and only if it is a countable union of sets of the form  $G_1 \times G_2$ , where  $G_1$  and  $G_2$  are open. Thus the Borel  $\sigma$ -field is generated by the sets  $G_1 \times G_2$ , where  $G_1$  and  $G_2$  are open. But so is the product of the Borel  $\sigma$ -fields.

For the proof of the second lemma, conclude first that the given integrals determine the probabilities  $L(G_1 \times G_2)$  for every pair of open  $G_1$  and  $G_2$ , because the indicator of  $G_1 \times G_2$  is the monotone limit from below of a sequence of functions of the form  $f_m(x) g_m(y)$ , with  $f_m$  and  $g_m$  Lipschitz continuous and  $0 \leq f_m \leq 1_{G_1}$  and  $0 \leq g_m \leq 1_{G_2}$ . The open product sets generate the trace of the Borel  $\sigma$ -field on any separable subset of  $\mathbb{D} \times \mathbb{E}$  by the first lemma. They also form an intersection stable collection of sets. So the trace of  $L$  on any separable subset is uniquely determined (by a monotone class theorem). ■

Given  $X_\alpha: \Omega_\alpha \mapsto \mathbb{D}$  and  $Y_\alpha: \Omega_\alpha \mapsto \mathbb{E}$ , one can form the joint variable  $(X_\alpha, Y_\alpha): \Omega_\alpha \mapsto \mathbb{D} \times \mathbb{E}$ . If both components are Borel measurable, then  $(X_\alpha, Y_\alpha)$  is measurable in the product of the Borel  $\sigma$ -fields, but not necessarily in the Borel  $\sigma$ -field of the product. This is an inconvenience, since we would want to use the second one in the weak convergence theory in the product space. The problem disappears if the metric spaces are separable. Furthermore, in the general case, asymptotic measurability is retained under tightness.

**1.4.3 Lemma.** *Nets  $X_\alpha: \Omega_\alpha \mapsto \mathbb{D}$  and  $Y_\alpha: \Omega_\alpha \mapsto \mathbb{E}$  are asymptotically tight if and only if the same is true for  $(X_\alpha, Y_\alpha): \Omega_\alpha \mapsto \mathbb{D} \times \mathbb{E}$ .*

**1.4.4 Lemma.** *Asymptotically tight nets  $X_\alpha: \Omega_\alpha \mapsto \mathbb{D}$  and  $Y_\alpha: \Omega_\alpha \mapsto \mathbb{E}$  are asymptotically measurable if and only if the same is true for  $(X_\alpha, Y_\alpha): \Omega_\alpha \mapsto \mathbb{D} \times \mathbb{E}$ .*

**Proofs.** A set of the form  $K_1 \times K_2$  is compact if and only if  $K_1$  and  $K_2$  are compact. Second, if  $K \subset \mathbb{D} \times \mathbb{E}$  is compact, then it is contained in  $\pi_1 K \times \pi_2 K$ , where  $\pi_i K$  are the projections on the first and second coordinates and are compact. Third, for the first of the three mentioned product metrics, it holds that  $(K_1 \times K_2)^\delta = K_1^\delta \times K_2^\delta$ . It is now easy to see that asymptotic tightness of both marginals is equivalent to asymptotic tightness of the joint maps.

Asymptotic measurability of the joint maps trivially implies asymptotic measurability of the marginals (Problem 1.3.7).

Finally, assume  $(X_\alpha, Y_\alpha)$  is asymptotically tight and  $X_\alpha$  and  $Y_\alpha$  are asymptotically measurable. Let  $f \in C_b(\mathbb{D})$  and  $g \in C_b(\mathbb{E})$  take values in

the interval  $[-1, 1]$ . Then

$$\begin{aligned} & (f(X_\alpha)g(X_\alpha))^* - (f(X_\alpha)g(X_\alpha))_* \\ & \leq (1 + f(X_\alpha))^*(1 + g(X_\alpha))^* - (1 + f(X_\alpha))_* - (1 + g(X_\alpha))_* + 1 \\ & \quad - (1 + f(X_\alpha))_*(1 + g(X_\alpha))_* + (1 + f(X_\alpha))^* + (1 + g(X_\alpha))^* - 1. \end{aligned}$$

This converges to zero in probability, because each of the four terms in the bottom line asymptotically cancels the corresponding term in the second line of the display. By similar reasoning it follows that  $h(X_\alpha, Y_\alpha)$  is asymptotically measurable for every  $h$  of the form  $h(x, y) = \sum_{i=1}^m f_i(x)g_i(y)$  with  $f_i \in C_b(\mathbb{D})$  and  $g_i \in C_b(\mathbb{E})$ . This set of functions  $h$  forms an algebra and separates points of  $\mathbb{D} \times \mathbb{E}$ . Thus  $(X_\alpha, Y_\alpha)$  is asymptotically measurable by Lemma 1.3.13. ■

A useful corollary is that for weak convergence of vectors  $(X_\alpha, Y_\alpha)$ , it suffices to consider expressions of the type  $E^*f(X_\alpha)g(Y_\alpha)$ .

**1.4.5 Corollary.** *Let  $(X_\alpha, Y_\alpha): \Omega_\alpha \mapsto \mathbb{D} \times \mathbb{E}$  be an arbitrary net with  $E^*f(X_\alpha)g(Y_\alpha) \rightarrow \int f(x)g(y) dL(x, y)$  for all bounded, nonnegative Lipschitz functions  $f: \mathbb{D} \mapsto \mathbb{R}$  and  $g: \mathbb{E} \mapsto \mathbb{R}$  and a separable Borel measure  $L$  on  $\mathbb{D} \times \mathbb{E}$ . Then  $(X_\alpha, Y_\alpha) \rightsquigarrow L$ .*

**Proof.** Since  $g$  and  $f$  can be taken equal to 1, it follows that the nets  $X_\alpha$  and  $Y_\alpha$  converge marginally in distribution. Thus the nets  $X_\alpha$  and  $Y_\alpha$  are asymptotically tight and asymptotically measurable in the completions of  $\mathbb{D}$  and  $\mathbb{E}$ , respectively. The joint maps  $(X_\alpha, Y_\alpha)$  are asymptotically tight in  $\mathbb{D} \times \mathbb{E}$ . By Prohorov's theorem, every subnet has a further weakly converging subnet. Every weak limit point gives the same expectation to the function  $(x, y) \mapsto f(x)g(y)$  as  $L$ . Thus  $L$  is the only limit point. ■

It is not hard to find an example of weakly convergent nets  $X_\alpha$  and  $Y_\alpha$  with tight limits for which the joint maps  $(X_\alpha, Y_\alpha)$  do not converge weakly. However, the previous result and Prohorov's theorem show that  $(X_\alpha, Y_\alpha)$  is at least relatively compact. Under special conditions there is only one limit point and the joint maps do converge weakly. The next two examples consider the case of independent coordinates and the case that one coordinate is asymptotically degenerate.

**1.4.6 Example.** Call  $X_\alpha$  and  $Y_\alpha$  *asymptotically independent* if

$$E^*f(X_\alpha)g(Y_\alpha) - E^*f(X_\alpha)E^*g(Y_\alpha) \rightarrow 0$$

for all bounded, nonnegative, Lipschitz functions  $f$  and  $g$  on  $\mathbb{D}$  and  $\mathbb{E}$ , respectively. If  $X_\alpha$  and  $Y_\alpha$  are asymptotically independent and  $X_\alpha \rightsquigarrow L_1$  and  $Y_\alpha \rightsquigarrow L_2$  for separable  $L_i$ , then  $(X_\alpha, Y_\alpha) \rightsquigarrow L_1 \otimes L_2$ .

To see this, assume without loss of generality that  $\mathbb{D}$  and  $\mathbb{E}$  are complete, so that the marginals of  $(X_\alpha, Y_\alpha)$  are asymptotically tight. By Prohorov's theorem the net of joint variables  $(X_\alpha, Y_\alpha)$  is relatively compact. Every limit point  $L$  must have  $\int f(x)g(y) dL(x, y) = \int f dL_1 \int g dL_2$ . These numbers uniquely identify the (one) limit point.

**1.4.7 Example (Slutsky's lemma).** If  $X_\alpha \rightsquigarrow X$  and  $Y_\alpha \rightsquigarrow c$  with  $X$  separable and  $c$  a constant, then  $(X_\alpha, Y_\alpha) \rightsquigarrow (X, c)$ .

Again assume without loss of generality that  $X$  is tight. Then the net  $(X_\alpha, Y_\alpha)$  is asymptotically tight and asymptotically measurable. The limit points have marginals  $X$  and  $c$  and so are uniquely determined as  $(X, c)$ . (Now that the result has been obtained, it is easy to see that this is actually a special case of the previous example: since  $X$  and  $c$  are independent,  $X_\alpha$  and  $Y_\alpha$  are asymptotically independent.)

If  $X_\alpha$  and  $Y_\alpha$  take their values in a fixed topological vector space, the previous result can be combined with the continuous mapping theorem to obtain *Slutsky's theorem*: if  $X_\alpha \rightsquigarrow X$  and  $Y_\alpha \rightsquigarrow c$  with  $X$  separable and  $c$  a constant, then  $X_\alpha + Y_\alpha \rightsquigarrow X + c$ . Similarly, under the same conditions,  $X_\alpha Y_\alpha \rightsquigarrow cX$  in the case  $Y_\alpha$  are scalars; and provided  $c > 0$ ,  $X_\alpha/Y_\alpha \rightsquigarrow X/c$  also.

The results of this section easily extend to products of finitely many metric spaces. They can also be extended to countable products, with slightly more effort. If  $\mathbb{D}_i$  is a metric space with metric  $d_i$  for every natural number  $i$ , then the Cartesian product  $\mathbb{D}_1 \times \mathbb{D}_2 \times \dots$  can be metrized by the metrics

$$\begin{aligned} d((x_1, x_2, \dots), (z_1, z_2, \dots)) &= \sup_i \frac{1}{i} (d_i(x_i, z_i) \wedge 1), \\ d((x_1, x_2, \dots), (z_1, z_2, \dots)) &= \sum_i 2^{-i} (d_i(x_i, z_i) \wedge 1). \end{aligned}$$

These generate the same topology; the product topology. The results of this section hold for  $\mathbb{D}_1 \times \mathbb{D}_2 \times \dots$ . In particular, coordinatewise asymptotic tightness is the same as joint asymptotic tightness, and given asymptotic tightness, coordinatewise asymptotic measurability is the same as joint asymptotic measurability.

For weak convergence, a countable product yields hardly anything new over finite products. A map  $X: \Omega \mapsto \mathbb{D}_1 \times \mathbb{D}_2 \times \dots$  is separable if and only if every of its coordinates is, and weak convergence to a separable limit is equivalent to weak convergence of every finite set of coordinates. Write  $X_i$  for the  $i$ th coordinate of  $X$ :  $X(\omega) = (X_1(\omega), X_2(\omega), \dots)$ .

**1.4.8 Theorem.** Let  $X_\alpha: \Omega_\alpha \mapsto \mathbb{D}_1 \times \mathbb{D}_2 \times \dots$  be an arbitrary net and  $X$  separable. Then  $X_\alpha \rightsquigarrow X$  if and only if  $(X_{\alpha,1}, \dots, X_{\alpha,m}) \rightsquigarrow (X_1, \dots, X_m)$ , for every  $m \in \mathbb{N}$ .

**Proof.** If  $X_\alpha \rightsquigarrow X$ , then all marginals converge by the continuous mapping theorem. For the converse, assume without loss of generality that every  $\mathbb{D}_i$  is complete. If  $X_{\alpha,i} \rightsquigarrow X_i$  with  $X_i$  separable, then  $X_{\alpha,i}$  is asymptotically tight and asymptotically measurable. By a straightforward argument, conclude that the vector  $X_\alpha$  is asymptotically tight in the product space. Moreover, by a similar argument as for Lemma 1.4.3, obtain that  $h(X_\alpha)$  is asymptotically measurable for every  $h$  in the linear span of the functions of the form  $f_1(x_1)f_2(x_2) \cdots f_m(x_m)$ , where  $f_i$  ranges over  $C_b(\mathbb{D}_i)$  and  $m \in \mathbb{N}$ . This set of  $h$  forms a subalgebra of  $C_b(\mathbb{D}_1 \times \mathbb{D}_2 \times \cdots)$  and separates points of the product. By Lemma 1.3.13, the net  $X_\alpha$  is asymptotically measurable. Apply Prohorov's theorem to see that it is relatively compact. Every limit point  $L$  has the same finite-dimensional marginal distributions as  $X$ , so that the numbers  $\int f_1(x_1) \cdots f_m(x_m) dL(x_1, x_2, \dots)$  are uniquely determined. These determine  $L$  completely. ■

# 1.5

## Spaces of Bounded Functions

Let  $T$  be an arbitrary set. The space  $\ell^\infty(T)$  is defined as the set of all uniformly bounded, real functions on  $T$ : all functions  $z: T \mapsto \mathbb{R}$  such that

$$\|z\|_T := \sup_{t \in T} |z(t)| < \infty.$$

It is a metric space with respect to the *uniform distance*  $d(z_1, z_2) = \|z_1 - z_2\|_T$ .

The space  $\ell^\infty(T)$ , or a suitable subspace of it, is a natural space for stochastic processes with bounded sample paths. A *stochastic process* is simply an indexed collection  $\{X(t): t \in T\}$  of random variables defined on the same probability space: every  $X(t): \Omega \mapsto \mathbb{R}$  is a measurable map. If every *sample path*  $t \mapsto X(t, \omega)$  is bounded, then a stochastic process yields a map  $X: \Omega \mapsto \ell^\infty(T)$ . Sometimes the sample paths have additional properties, such as measurability or continuity, and it may be fruitful to consider  $X$  as a map into a subspace of  $\ell^\infty(T)$ . If in either case the uniform metric is used, this does not make a difference for weak convergence of a net; but for measurability it can. Here is one example of this situation; more examples are discussed in the next section.

**1.5.1 Example (Continuous functions).** Let  $T$  be a compact semimetric space; for instance, a compact interval in the real line, or the extended real line  $[-\infty, \infty]$  with the metric  $\rho(s, t) = |\arctan s - \arctan t|$ . The set  $C(T)$  of all continuous functions  $z: T \mapsto \mathbb{R}$  is a separable, complete subspace of  $\ell^\infty(T)$ . The Borel  $\sigma$ -field of  $C(T)$  equals the  $\sigma$ -field generated by the coordinate projections  $z \mapsto z(t)$ , the *projection  $\sigma$ -field* (Problem 1.7.1).

Thus a map  $X: \Omega \mapsto C(T)$  is Borel measurable if and only if it is a stochastic process.

In most cases a map  $X: \Omega \mapsto \ell^\infty(T)$  is a stochastic process. The small amount of measurability this gives may already be enough for asymptotic measurability. The special role played by the *marginals*  $(X(t_1), \dots, X(t_k))$ , which are considered as maps into  $\mathbb{R}^k$ , is underlined by the following three results. Weak convergence in  $\ell^\infty(T)$  can be characterized as asymptotic tightness plus convergence of marginals.

**1.5.2 Lemma.** *Let  $X_\alpha: \Omega_\alpha \mapsto \ell^\infty(T)$  be asymptotically tight. Then it is asymptotically measurable if and only if  $X_\alpha(t)$  is asymptotically measurable for every  $t \in T$ .*

**1.5.3 Lemma.** *Let  $X$  and  $Y$  be tight Borel measurable maps into  $\ell^\infty(T)$ . Then  $X$  and  $Y$  are equal in Borel law if and only if all corresponding marginals of  $X$  and  $Y$  are equal in law.*

**1.5.4 Theorem.** *Let  $X_\alpha: \Omega_\alpha \mapsto \ell^\infty(T)$  be arbitrary. Then  $X_\alpha$  converges weakly to a tight limit if and only if  $X_\alpha$  is asymptotically tight and the marginals  $(X_\alpha(t_1), \dots, X_\alpha(t_k))$  converge weakly to a limit for every finite subset  $t_1, \dots, t_k$  of  $T$ . If  $X_\alpha$  is asymptotically tight and its marginals converge weakly to the marginals  $(X(t_1), \dots, X(t_k))$  of a stochastic process  $X$ , then there is a version of  $X$  with uniformly bounded sample paths and  $X_\alpha \rightsquigarrow X$ .*

**Proofs.** For the proof of both lemmas, consider the collection  $\mathcal{F}$  of all functions  $f: \ell^\infty(T) \mapsto \mathbb{R}$  of the form

$$f(z) = g(z(t_1), \dots, z(t_k)), \quad g \in C_b(\mathbb{R}^k), \quad t_1, \dots, t_k \in T, \quad k \in \mathbb{N}.$$

This forms an algebra and a vector lattice, contains the constant functions, and separates points of  $\ell^\infty(T)$ . Therefore, the lemmas are corollaries of Lemmas 1.3.13 and 1.3.12, respectively.

If  $X_\alpha$  is asymptotically tight and the marginals converge, then  $X_\alpha$  is asymptotically measurable by the first lemma. By Prohorov's theorem,  $X_\alpha$  is relatively compact. To prove weak convergence, it suffices to show that all limit points are the same. This follows from marginal convergence and the second lemma. ■

Marginal convergence can be established by any of the well-known methods for proving weak convergence on Euclidean space. Tightness can be given a more concrete form, either through finite approximation or (essentially) with the help of the Arzelà-Ascoli theorem. Finite approximation leads to the simpler of the two characterizations, but the second approach

is perhaps of more interest, because it connects tightness to (asymptotic, uniform, equi-) continuity of the sample paths  $t \mapsto X_\alpha(t)$ .

The idea of finite approximation is that for any  $\varepsilon > 0$  the index set  $T$  can be partitioned into finitely many subsets  $T_i$  such that (asymptotically) the variation of the sample paths  $t \mapsto X_\alpha(t)$  is less than  $\varepsilon$  on every one of the sets  $T_i$ . More precisely, it is assumed that for every  $\varepsilon, \eta > 0$ , there exists a partition  $T = \cup_{i=1}^k T_i$  such that

$$(1.5.5) \quad \limsup_{\alpha} P^* \left( \sup_i \sup_{s,t \in T_i} |X_\alpha(s) - X_\alpha(t)| > \varepsilon \right) < \eta.$$

Clearly, under this condition the asymptotic behavior of the process can be described within error margin  $\varepsilon, \eta$  by the behavior of the marginal  $(X_\alpha(t_1), \dots, X_\alpha(t_k))$  for arbitrary fixed points  $t_i \in T_i$ . If the process can thus be reduced to a finite set of coordinates for any  $\varepsilon, \eta > 0$  and the nets of marginal distributions are tight, then the net  $X_\alpha$  is asymptotically tight.

**1.5.6 Theorem.** *A net  $X_\alpha: \Omega_\alpha \mapsto \ell^\infty(T)$  is asymptotically tight if and only if  $X_\alpha(t)$  is asymptotically tight in  $\mathbb{R}$  for every  $t$  and, for all  $\varepsilon, \eta > 0$ , there exists a finite partition  $T = \cup_{i=1}^k T_i$  such that (1.5.5) holds.*

**Proof.** The necessity of the conditions follows easily from the next theorem. For instance, take the partition equal to (disjointified) balls of radius  $\delta$  for a semimetric on  $T$  as in the next theorem. We prove sufficiency.

For any partition, as in the condition of the theorem, the norm  $\|X_\alpha\|_T$  is bounded by  $\max_i |X_\alpha(t_i)| + \varepsilon$ , with inner probability at least  $1 - \eta$ , if  $t_i \in T_i$  for each  $i$ . Since a maximum of finitely many tight nets of real variables is tight, it follows that the net  $\|X_\alpha\|_T$  is asymptotically tight in  $\mathbb{R}$ .

Fix  $\zeta > 0$  and a sequence  $\varepsilon_m \downarrow 0$ . Take a constant  $M$  such that  $\limsup P^*(\|X_\alpha\|_T > M) < \zeta$ , and for each  $\varepsilon = \varepsilon_m$  and  $\eta = 2^{-m}\zeta$ , take a partition  $T = \cup_{i=1}^k T_i$  as in (1.5.5). For the moment  $m$  is fixed and we do not let it appear in the notation. Let  $z_1, \dots, z_p$  be the set of all functions in  $\ell^\infty(T)$  that are constant on each  $T_i$  and take on only the values  $0, \pm\varepsilon_m, \dots, \pm[M/\varepsilon_m]\varepsilon_m$ . (It is easy to express  $p$  in terms of the other constants, but it is only relevant that it is finite.) Let  $K_m$  be the union of the  $p$  closed balls of radius  $\varepsilon_m$  around the  $z_i$ . Then, by construction, the two conditions

$$\|X_\alpha\|_T \leq M \quad \text{and} \quad \sup_i \sup_{s,t \in T_i} |X_\alpha(s) - X_\alpha(t)| \leq \varepsilon_m$$

imply that  $X_\alpha \in K_m$ . This is true for each fixed  $m$ .

Let  $K = \cap_{m=1}^\infty K_m$ . Then  $K$  is closed and totally bounded (by construction of the  $K_m$  and because  $\varepsilon_m \downarrow 0$ ) and hence compact. Furthermore, for every  $\delta > 0$ , there is an  $m$  with  $K^\delta \supset \cap_{i=1}^m K_i$ . If not, then there would be a sequence  $z_m$  not in  $K^\delta$ , but with  $z_m \in \cap_{i=1}^m K_i$  for every  $m$ . This

would have a subsequence contained in one of the balls making up  $K_1$ , a further subsequence eventually contained in one of the balls making up  $K_2$ , and so on. The “diagonal” sequence, formed by taking the first of the first subsequence, the second of the second subsequence and so on, would eventually be contained in a ball of radius  $\varepsilon_m$  for every  $m$ ; hence Cauchy. Its limit would be in  $K$ , contradicting the fact that  $d(z_m, K) \geq \delta$  for every  $m$ .

Conclude that if  $X_\alpha$  is not in  $K^\delta$ , then it is not in  $\cap_{i=1}^m K_i$  for some fixed  $m$ . Then

$$\limsup P^*(X_\alpha \notin K^\delta) \leq \limsup P^*(X_\alpha \notin \bigcap_{i=1}^m K_i) \leq \zeta + \sum_{i=1}^m \zeta 2^{-m} < 2\zeta.$$

This concludes the proof of the theorem. ■

The second type of characterization of asymptotic tightness is deeper and relates the concept to asymptotic continuity of the sample paths. Suppose  $\rho$  is a semimetric on  $T$ . A net  $X_\alpha: \Omega_\alpha \mapsto \ell^\infty(T)$  is *asymptotically uniformly  $\rho$ -equicontinuous in probability* if for every  $\varepsilon, \eta > 0$  there exists a  $\delta > 0$  such that

$$\limsup_\alpha P^* \left( \sup_{\rho(s,t) < \delta} |X_\alpha(s) - X_\alpha(t)| > \varepsilon \right) < \eta.$$

**1.5.7 Theorem.** *A net  $X_\alpha: \Omega_\alpha \mapsto \ell^\infty(T)$  is asymptotically tight if and only if  $X_\alpha(t)$  is asymptotically tight in  $\mathbb{R}$  for every  $t$  and there exists a semimetric  $\rho$  on  $T$  such that  $(T, \rho)$  is totally bounded and  $X_\alpha$  is asymptotically uniformly  $\rho$ -equicontinuous in probability.*

**1.5.8 Addendum.** *If, moreover,  $X_\alpha \rightsquigarrow X$ , then almost all paths  $t \mapsto X(t, \omega)$  are uniformly  $\rho$ -continuous; and the semimetric  $\rho$  can without loss of generality be taken equal to any semimetric  $\rho$  for which this is true and  $(T, \rho)$  is totally bounded.*

**Proof.**  $\Leftarrow$ . The sufficiency follows from the previous theorem. First take  $\delta > 0$  sufficiently small so that the last displayed inequality is valid. Since  $T$  is totally bounded, it can be covered with finitely many balls of radius  $\delta$ . Construct a partition of  $T$  by disjointifying these balls.

$\Rightarrow$ . If  $X_\alpha$  is asymptotically tight, then  $g(X_\alpha)$  is asymptotically tight for every continuous map  $g$ ; in particular, for each coordinate projection.

Let  $K_1 \subset K_2 \subset \dots$  be compacts with  $\liminf P_*(X_\alpha \in K_m^\varepsilon) \geq 1 - 1/m$  for every  $\varepsilon > 0$ . For every fixed  $m$ , define a semimetric  $\rho_m$  on  $T$  by

$$\rho_m(s, t) = \sup_{z \in K_m} |z(s) - z(t)|, \quad s, t \in T.$$

Then  $(T, \rho_m)$  is totally bounded. Indeed, cover  $K_m$  by finitely many balls of (arbitrarily small) radius  $\eta$ , centered at  $z_1, \dots, z_k$ . Partition  $\mathbb{R}^k$  into cubes of

edge  $\eta$ , and for every cube pick at most one  $t \in T$  such that  $(z_1(t), \dots, z_k(t))$  is in the cube. Since  $z_1, \dots, z_k$  are uniformly bounded, this gives finitely many points  $t_1, \dots, t_p$ . Now the balls  $\{t: \rho_m(t, t_i) < 3\eta\}$  cover  $T$ :  $t$  is in the ball around  $t_i$  for which  $(z_1(t), \dots, z_k(t))$  and  $(z_1(t_i), \dots, z_k(t_i))$  fall in the same cube. This follows because  $\rho_m(t, t_i)$  can be bounded by  $2 \sup_{z \in K_m} \inf_i \|z - z_i\|_T + \sup_j |z_j(t_i) - z_j(t)|$ .

Next set

$$\rho(s, t) = \sum_{m=1}^{\infty} 2^{-m} (\rho_m(s, t) \wedge 1).$$

Fix  $\eta > 0$ . Take a natural number  $m$  with  $2^{-m} < \eta$ . Cover  $T$  with finitely many  $\rho_m$ -balls of radius  $\eta$ . Let  $t_1, \dots, t_p$  be their centers. Since  $\rho_1 \leq \rho_2 \leq \dots$ , there is for every  $t$  a  $t_i$  with  $\rho(t, t_i) \leq \sum_{k=1}^m 2^{-k} \rho_k(t, t_i) + 2^{-m} < 2\eta$ . Thus  $(T, \rho)$  is totally bounded for  $\rho$ , too.

It is clear from the definitions that  $|z(s) - z(t)| \leq \rho_m(s, t)$  for every  $z \in K_m$  and that  $\rho_m(s, t) \wedge 1 \leq 2^m \rho(s, t)$ . Also, if  $\|z_0 - z\|_T < \varepsilon$  for  $z \in K_m$ , then  $|z_0(s) - z_0(t)| < 2\varepsilon + |z(s) - z(t)|$  for any pair  $s, t$ . Deduce that

$$K_m^\varepsilon \subset \left\{ z: \sup_{\rho(s, t) < 2^{-m}\varepsilon} |z(s) - z(t)| \leq 3\varepsilon \right\}.$$

Thus for given  $\varepsilon$  and  $m$ , and for  $\delta < 2^{-m}\varepsilon$ ,

$$\liminf P_* \left( \sup_{\rho(s, t) < \delta} |X_\alpha(s) - X_\alpha(t)| \leq 3\varepsilon \right) \geq 1 - \frac{1}{m}.$$

Finally, we prove the addendum. If  $X_\alpha \rightsquigarrow X$ , then with notation as in the second part of the proof,  $P(X \in K_m) \geq 1 - 1/m$ ; hence  $X$  concentrates on  $\cup_{m=1}^{\infty} K_m$ . The elements of  $K_m$  are uniformly  $\rho_m$ -equicontinuous and hence also uniformly  $\rho$ -continuous. This yields the first statement.

The set of uniformly continuous functions on a totally bounded, semimetric space is complete and separable, so a map  $X$  that takes its values in this set is tight. Next if  $X_\alpha \rightsquigarrow X$  and  $X$  is tight, then  $X_\alpha$  is asymptotically tight and the compacts for asymptotical tightness can be chosen equal to the compacts for tightness of  $X$ . If  $X$  has uniformly continuous paths, then the latter compacts can be chosen within the space of uniformly continuous functions. Since a compact is totally bounded, every one of the compacts is necessarily uniformly equicontinuous. Combination of these facts proves the second statement of the addendum. ■

Asymptotic uniform  $\rho$ -equicontinuity is a fairly complicated concept. Nevertheless, it is what must be shown for weak convergence. For particular problems, reasonable methods are available; for instance, methods based on the Markov property or the chaining method. In Part 2 we develop such methods for empirical processes.

At a general level, a few things can be said about which semimetric to use to establish asymptotic equicontinuity of a net  $X_\alpha$ . Note that in

principle the index set  $T$  need not have a semimetric on it when it comes to us. Furthermore, if it does, this semimetric may not be the right one to use. The next few results of this section are somewhat technical, but they lead to the important result that for Gaussian limits one can always use a certain variance metric.

A possible limit  $X$  of a net  $X_\alpha$  can be identified from marginal convergence. Next, according to the addendum, we should look for a semimetric  $\rho$  that makes  $T$  totally bounded and the paths of  $X$  uniformly continuous. It can be shown that if any semimetric does this at all, the semimetric

$$\rho_0(s, t) = \mathbb{E} \arctan |X(s) - X(t)|$$

should do the job (Problem 1.5.2). However,  $\rho_0$  may not be the most convenient semimetric with which to work. Sometimes there is an obvious semimetric  $\rho$  to try. Alternatively, consider the family of semimetrics

$$\rho_p(s, t) = \left( \mathbb{E} |X(s) - X(t)|^p \right)^{1/(p \vee 1)} \quad (0 < p < \infty).$$

None of these will always qualify; in fact, the expectations need not even be finite. Furthermore, these semimetrics are special in that they make the process  $t \mapsto X(t)$  continuous in  $p$ th mean. It turns out that the latter is exactly what makes them work or not: if there is a semimetric  $\rho$  for which  $T$  is totally bounded, the paths of  $X$  are uniformly  $\rho$ -continuous, and also the process  $t \mapsto X(t)$  is uniformly  $\rho$ -continuous in  $p$ th mean, then and only then can  $\rho_p$  be used without loss of generality when showing tightness. Here uniform continuity in  $p$ th mean with respect to  $\rho$  means that  $\mathbb{E} |X(s_n) - X(t_n)|^p \rightarrow 0$  whenever  $\rho(s_n, t_n) \rightarrow 0$ .

**1.5.9 Lemma.** *Let  $X$  be a tight, Borel measurable map into  $\ell^\infty(T)$ . Then there is a semimetric on  $T$  for which almost all paths of  $X$  are uniformly continuous and  $T$  is totally bounded. Moreover, for any fixed  $p > 0$ , the following statements are equivalent:*

- (i) *for the semimetric  $\rho_p$ , the set  $T$  is totally bounded and almost all paths of  $X$  are uniformly  $\rho_p$ -continuous;*
- (ii) *for every semimetric  $\rho$  for which almost all paths of  $X$  are uniformly  $\rho$ -continuous, the map  $t \mapsto X(t)$  is uniformly  $\rho$ -continuous in  $p$ th mean;*
- (iii) *there is a semimetric  $\rho$  for which  $T$  is totally bounded, the map  $t \mapsto X(t)$  is uniformly  $\rho$ -continuous in  $p$ th mean, and almost all paths of  $X$  are uniformly  $\rho$ -continuous.*

In particular, if  $T$  is compact for a semimetric  $\rho$  such that the map  $t \mapsto \mathbb{E} |X(t)|^p$  and almost all sample paths are  $\rho$ -continuous, then (i) holds.

**Proof.** Since  $X \rightsquigarrow X$  and is asymptotically tight, the first statement is a special case of the addendum to the previous theorem.

(i)  $\Rightarrow$  (ii). Suppose that for sequences  $s_n$  and  $t_n$  in  $T$ ,  $\rho(s_n, t_n) \rightarrow 0$ , but there is an  $\varepsilon$  with  $\mathbb{E} |X(s_n) - X(t_n)|^p \geq \varepsilon > 0$  for every  $n$ . By

total boundedness of  $T$  under  $\rho_p$ , there exist subsequences  $s_{n'}$  and  $t_{n'}$  that converge with respect to  $\rho_p$  to limits  $s$  and  $t$  in the  $\rho_p$ -completion of  $T$ . These subsequences are Cauchy, which by definition of  $\rho_p$  is the same as  $X(s_{n'})$  and  $X(t_{n'})$  being Cauchy in the  $L_p$ -metric. Let  $X(s)$  and  $X(t)$  be their  $L_p$ -limits. Since almost all sample paths of  $X$  are uniformly  $\rho_p$ -continuous,  $X(s_{n'})$  and  $X(t_{n'})$  converge almost surely also; the almost sure limits must equal the  $L_p$ -limits  $X(s)$  and  $X(t)$ . Because almost all sample paths of  $X$  are also uniformly continuous with respect to  $\rho$  (!) and  $\rho(s_n, t_n) \rightarrow 0$  by assumption,  $X(s_n) - X(t_n) \rightarrow 0$  almost surely. Conclude that  $X(s) = X(t)$  almost surely, so that the two constructed subsequences have the same  $L_p$ -limits. It follows that  $E|X(s_{n'}) - X(t_{n'})|^p \rightarrow 0$ , contradicting the assumption.

(ii)  $\Rightarrow$  (iii). By the first statement, there always exists a semimetric for which  $T$  is totally bounded and  $X$  is almost surely uniformly continuous. If (ii) holds, then  $t \mapsto X(t)$  is uniformly continuous in  $p$ th mean for this semimetric. Hence we obtain a semimetric as in (iii).

(iii)  $\Rightarrow$  (i). Since the map  $t \mapsto X(t)$  is uniformly  $\rho$ -continuous in  $p$ th mean, there exists for every  $\varepsilon > 0$  a  $\delta > 0$  with  $\rho_p(s, t) < \varepsilon$  whenever  $\rho(s, t) < \delta$ . Thus the total boundedness of  $T$  for  $\rho$  implies the same for  $\rho_p$ . Now assume for simplicity that all paths of  $X$  are uniformly  $\rho$ -continuous. Every uniformly continuous function defined on a subset of a semimetric space has a unique extension to a continuous function on the closure of its domain. Therefore all sample paths of  $X$  can be extended to continuous functions on  $(\bar{T}, \rho)$ , where  $\bar{T}$  is the  $\rho$ -completion of  $T$ , which is compact.

If a path  $t \mapsto X(t, \omega)$  is not uniformly  $\rho_p$ -continuous, then there exist  $\varepsilon > 0$  and sequences  $s_n$  and  $t_n$  with  $\rho_p(s_n, t_n) \rightarrow 0$  and  $|X(s_n, \omega) - X(t_n, \omega)| \geq \varepsilon$  for every  $n$ . These have subsequences that converge with respect to  $\rho$  to limits  $s$  and  $t$  in  $\bar{T}$ . Consequently, first  $X(s_{n'}, \omega) - X(t_{n'}, \omega) \rightarrow X(s, \omega) - X(t, \omega)$ . Second, the subsequences converge to  $s$  and  $t$  under  $\rho_p$ , too, so  $\rho_p(s, t) = 0$ . Conclude that the path  $t \mapsto X(t, \omega)$  is uniformly  $\rho_p$ -continuous for every  $\omega$  for which there do not exist  $s, t \in \bar{T}$  with  $\rho_p(s, t) = 0$ , but  $X(s, \omega) \neq X(t, \omega)$ . Let  $N$  be the exceptional set of  $\omega$  for which there do exist such  $s, t$ . Take a countable,  $\rho$ -dense subset  $A$  of  $\{(s, t) \in \bar{T} \times \bar{T}: \rho_p(s, t) = 0\}$ . Since  $t \mapsto X(t, \omega)$  is  $\rho$ -continuous,  $N$  is also the set of all  $\omega$  such that there exist  $(s, t) \in A$  with  $X(s, \omega) \neq X(t, \omega)$ . From the definition of  $\rho_p$ , it is clear that for every fixed  $(s, t) \in A$ , it holds that  $X(s, \omega) = X(t, \omega)$  for almost every  $\omega$ . Conclude that  $N$  is a nullset. Hence, almost all paths of  $X$  are uniformly  $\rho_p$ -continuous.

For the last remark of the lemma, it suffices to note that  $X(t_n) \rightarrow X(t)$  almost surely and  $E|X(t_n)|^p \rightarrow E|X(t)|^p$  implies that  $X(t_n) \rightarrow X(t)$  in  $p$ th mean. ■

**1.5.10 Example.** A stochastic process  $X$  is called *Gaussian* if each of its finite-dimensional marginals  $(X(t_1), \dots, X(t_k))$  has a multivariate normal distribution on Euclidean space.

Let  $X$  be a Gaussian process with “intrinsic” semimetrics  $\rho_p$ , and let  $X_\alpha$  be a net of random elements with values in  $\ell^\infty(T)$ . Then there exists a version of  $X$  which is a tight Borel measurable map into  $\ell^\infty(T)$ , and  $X_\alpha$  converges weakly to  $X$  if and only if for some  $p$  (and then for all  $p$ ):

- the marginals of  $X_\alpha$  converge weakly to the corresponding marginals of  $X$ ;
- $X_\alpha$  is asymptotically equicontinuous in probability with respect to  $\rho_p$ ;
- $T$  is totally bounded for  $\rho_p$ .

The second moment metric  $\rho_2$  is often the easiest with which to work.

The sufficiency of the three conditions is immediate from the preceding theorems and is not special to Gaussian processes. In fact, for the sufficiency, the semimetric  $\rho_p$  can be replaced by any other semimetric. The extra information is that for Gaussian processes one can always use any of the intrinsic semimetrics. This necessity is not immediate from the previous results but can be derived as follows. In view of the addendum to Theorem 1.5.7, it is certainly enough to show that

A Gaussian process  $X$  in  $\ell^\infty(T)$  is tight if and only if  $(T, \rho_p)$  is totally bounded and almost all paths  $t \mapsto X(t, \omega)$  are uniformly  $\rho_p$ -continuous for some  $p$  (and then for all  $p$ ).

For any tight  $X$ , there is a semimetric  $\rho$  that makes  $T$  totally bounded and almost all paths of  $X$  uniformly continuous. The assertion would follow by (iii) of the previous lemma if the map  $t \mapsto X(t)$  is uniformly  $\rho$ -continuous in  $p$ th mean for a Gaussian process. This is automatically the case.

If a Gaussian process in  $\ell^\infty(T)$  has uniformly continuous paths with respect to a semimetric  $\rho$  that makes  $T$  totally bounded, then the map  $t \mapsto X(t)$  is uniformly  $\rho$ -continuous in  $p$ th mean.

Indeed, if  $\rho(s_n, t_n) \rightarrow 0$ , then  $X(s_n) - X(t_n) \rightarrow 0$  almost surely, hence weakly. Since all the random variables are normal, this can only happen if the mean and variance of  $X(s_n) - X(t_n)$  converge to zero. (Use characteristic functions.) For a normal variable, the  $p$ th absolute moment is a continuous function of the first two moments.

## Problems and Complements

1. (**Arzelà-Ascoli**) Let  $(T, \rho)$  be a totally bounded, semimetric space and  $K \subset \ell^\infty(T)$ . Suppose that for some  $\varepsilon, \delta > 0$  and every  $z \in K$ :  $|z(s) - z(t)| < \varepsilon$  whenever  $\rho(s, t) < \delta$ . Moreover, suppose  $\{z(t): z \in K\}$  is bounded for every  $t$ . Then
  - (i)  $K$  is uniformly bounded;
  - (ii)  $K$  can be covered with finitely many balls of radius  $2\varepsilon$  with centres in  $\text{UC}(T, \rho)$ .

Deduce the Arzelà-Ascoli theorem: a uniformly bounded, uniformly  $\rho$ -equicontinuous subset of  $\ell^\infty(T)$  is totally bounded.

[Hint: For (ii) choose a  $\delta/2$ -net  $t_1, \dots, t_p$  on  $T$ . Let  $d$  be the distance between the closest pair of  $t_i$ . Define functions

$$\begin{aligned}\beta_i(t) &= (1 - \delta^{-1} \rho(t, t_i))^+, \\ \alpha_i(t) &= 1 - (1 - d^{-1} \rho(t, t_i))^+, \\ w_i(t) &= \frac{\beta_i(t) \prod_{j \neq i} \alpha_j(t)}{\sum_i \beta_i(t) \prod_{j \neq i} \alpha_j(t)}.\end{aligned}$$

Then each  $w_i$  is defined and continuous on the  $\rho$ -completion of  $T$ , the  $w_i$  sum up to 1,  $w_i(t_i) = 1$ , and  $w_i$  can be nonzero only at  $t$  with  $\rho(t, t_i) < \delta$  that do not equal one of the other  $t_j$ . (The  $\alpha_i$  are just there to take care of the latter.) For any  $u \in \mathbb{R}^p$ , define a function  $z_u(t) = \sum_i w_i(t) u_i$ . This is uniformly continuous with  $z_u(t_i) = u_i$  for every  $i$ . Let  $S = \{0, \varepsilon, -\varepsilon, \dots, k\varepsilon, -k\varepsilon\}$  where  $k$  is chosen such that  $k\varepsilon$  is larger than a uniform bound on  $K$ . Then  $\{z_u : u \in S^p\}$  is a  $2\varepsilon$ -net over  $K$ : a given  $z$  is in the  $2\varepsilon$ -ball around  $z_u$  for which  $|u_i - z(t_i)| < \varepsilon$  for every  $i$ . This is true because, for every fixed  $t$ , the number  $z_u(t)$  is a convex combination of values  $u_i$  with  $w_i(t) \neq 0$ ; hence  $i$  with  $\rho(t, t_i) < \delta$ . For every such  $i$ :  $|z(t) - u_i| < |z(t) - z(t_i)| + \varepsilon < 2\varepsilon$ .]

2. Let  $(T, \rho)$  be a totally bounded, semimetric space and  $X$  a map into  $\ell^\infty(T)$  with uniformly  $\rho$ -continuous paths. Then  $T$  is totally bounded for the semimetric

$$\rho_0(s, t) = \mathbb{E} \arctan |X(s) - X(t)|.$$

Furthermore, almost all paths of  $X$  are uniformly continuous with respect to  $\rho_0$ .

3. Let  $X_\alpha : \Omega_\alpha \mapsto \ell^\infty(S)$  and  $Y_\alpha : \Omega_\alpha \mapsto \ell^\infty(T)$  be asymptotically tight nets such that

$$(X_\alpha(s_1), \dots, X_\alpha(s_k), Y_\alpha(t_1), \dots, Y_\alpha(t_l)) \rightsquigarrow (X(s_1), \dots, X(s_k), Y(t_1), \dots, Y(t_l)),$$

for stochastic processes  $X$  and  $Y$ . Then there exist versions of  $X$  and  $Y$  with bounded sample paths and  $(X_\alpha, Y_\alpha) \rightsquigarrow (X, Y)$  in  $\ell^\infty(S) \times \ell^\infty(T)$ .

[Hint: The net  $(X_\alpha, Y_\alpha)$  can be identified with a net in  $\ell^\infty(S \cup T)$ , where  $S \cup T$  is the formal union of  $S$  and  $T$ , counting every element of  $S$  different from every element of  $T$ . This identification is an isometry.]

# 1.6

## Spaces of Locally Bounded Functions

Let  $T_1 \subset T_2 \subset \dots$  be arbitrary sets and  $T = \cup_{i=1}^{\infty} T_i$ . The space  $\ell^{\infty}(T_1, T_2, \dots)$  is defined as the set of all functions  $z: T \mapsto \mathbb{R}$  that are uniformly bounded on every  $T_i$  (but not necessarily on  $T$ ). This is a complete metric space with respect to the metric

$$d(z_1, z_2) = \sum_{i=1}^{\infty} (\|z_1 - z_2\|_{T_i} \wedge 1) 2^{-i}.$$

A sequence converges in this metric if it converges uniformly on each  $T_i$ . In the case that  $T_i$  equals the interval  $[-i, i] \subset \mathbb{R}^d$ , the metric  $d$  induces the topology of uniform convergence on compacta.

The space  $\ell^{\infty}(T_1, T_2, \dots)$  is of interest in applications, but its weak convergence theory is uneventful. Weak convergence of a net is equivalent to convergence of each of the restrictions to  $T_i$ .

**1.6.1 Theorem.** *Let  $X_{\alpha}: \Omega_{\alpha} \mapsto \ell^{\infty}(T_1, T_2, \dots)$  be arbitrary maps. Then the net  $X_{\alpha}$  converges weakly to a tight limit if and only if every of the nets of restrictions  $X_{\alpha|T_i}: \Omega \mapsto \ell^{\infty}(T_i)$  converges weakly ( $i \in \mathbb{N}$ ) to a tight limit.*

**Proof.** The necessity of the weak convergence of all restrictions follows from the continuous mapping theorem. For the sufficiency part of the theorem, fix  $\varepsilon > 0$  and take for each  $i$  a compact  $K_i \subset \ell^{\infty}(T_i)$  such that

$$\limsup P^*(X_{\alpha|T_i} \notin K_i^{\delta}) < \frac{\varepsilon}{2^i}, \quad \text{for every } \delta > 0.$$

Define  $K$  as the set of all  $z:T \mapsto \mathbb{R}$  such that  $z|_{T_i - T_{i-1}}$  is contained in  $(K_i)|_{T_i - T_{i-1}}$  for every  $i$ . A diagonal argument shows that  $K$  is compact in  $\ell^\infty(T_1, T_2, \dots)$ . Furthermore, if  $z|_{T_i} \in K_i^\delta$  for every  $i$ , then  $z \in K^\delta$ . Conclude that

$$\limsup P^*(X_\alpha \notin K^\delta) < \varepsilon, \quad \text{for every } \delta > 0.$$

It follows that the net  $X_\alpha$  is asymptotically tight in  $\ell^\infty(T_1, T_2, \dots)$ .

Let  $\mathcal{F}$  be the set of all continuous functions  $f:\ell^\infty(T_1, T_2, \dots) \mapsto \mathbb{R}$  of the form  $f(z) = g(z(t_1), \dots, z(t_k))$  with  $g \in C_b(\mathbb{R}^k)$ ,  $t_1, \dots, t_k \in T$ , and  $k \in \mathbb{N}$ . Then  $\mathcal{F}$  is an algebra that separates points and contains the constant functions. Since the net  $f(X_\alpha)$  is asymptotically measurable for every  $f \in \mathcal{F}$ , Lemma 1.3.13 implies that  $X_\alpha$  is asymptotically measurable in  $\ell^\infty(T_1, T_2, \dots)$ .

Apply Prohorov's theorem to see that every subnet has a further weakly converging subnet with a tight limit. There is only one limit point, in view of the marginal convergence of the net and Lemma 1.3.12. ■

Given the preceding theorem, convergence in distribution in the space  $\ell^\infty(T_1, T_2, \dots)$  can be proved by showing weak convergence of all marginals  $(X_\alpha(t_1), \dots, X_\alpha(t_k))$  plus asymptotic equicontinuity on every  $T_i$ . For a given semimetric  $\rho$  on  $T$ , the latter takes the following form: for all  $\varepsilon > 0$ ,  $\eta > 0$ , and  $i$ , there exists  $\delta > 0$  such that

$$\limsup_\alpha P^*\left(\sup_{\substack{\rho(s,t) < \delta \\ s,t \in T_i}} |X_\alpha(s) - X_\alpha(t)| > \varepsilon\right) < \eta.$$

Under this condition, the limit process necessarily has uniformly  $\rho$ -continuous sample paths on each  $T_i$ . In particular, it has continuous sample paths on  $T$ . For a Gaussian limit, this is necessarily the case for the standard deviation metric.

## Problems and Complements

- (Convex processes)** Let  $X_\alpha$  be a net of stochastic processes indexed by a convex, open subset  $C$  of  $\mathbb{R}^k$  such that every sample path  $t \mapsto X_\alpha(t)$  is convex (on  $C$ ). If the net  $X_\alpha$  converges marginally in distribution to a limit, then it converges in distribution to a tight limit in the space  $\ell^\infty(K_1, K_2, \dots)$  for any sequence of compact sets  $K_1 \subset K_2 \subset \dots \subset C$ .

[Hint: For every compact  $K \subset C$  there exists an  $\varepsilon > 0$  such that  $K^\varepsilon \subset K^{2\varepsilon} \subset C$ . If  $x_\alpha:C \mapsto \mathbb{R}$  is a net of (deterministic) convex functions such that  $x_\alpha(t)$  is bounded for every  $t \in C$ , then the net is automatically uniformly bounded over  $K^\varepsilon$  (Problem 2.7.5). Conclude that the net  $\|X_\alpha\|_{K^\varepsilon}$  is asymptotically tight. A bounded, convex function  $x$  on  $K^\varepsilon$  is automatically Lipschitz on  $K$  with Lipschitz constant  $(2/\varepsilon)\|x\|_{K^\varepsilon}$  (Problem 2.7.4). Conclude that the net  $X_\alpha$  is asymptotically equicontinuous in  $\ell^\infty(K)$ .]

## 1.7

# The Ball Sigma-Field and Measurability of Suprema

The *ball  $\sigma$ -field* on  $\mathbb{D}$  is the smallest  $\sigma$ -field containing all the open (and/or closed) balls in  $\mathbb{D}$ . In general, this is smaller than the Borel  $\sigma$ -field, although the two  $\sigma$ -fields are equal for separable spaces (Problems 1.7.3 and 1.7.4). For some nonseparable spaces, it is even fairly common that maps are ball measurable even though they are not Borel measurable. Thus one may wonder about the possibility of a weak convergence theory for ball measurable maps. It turns out that the set of ball measurable  $f \in C_b(\mathbb{D})$  is rich enough to make this fruitful, but at the same time it is so rich that the theory is a special case of the theory that we have discussed so far.

However, establishing ball measurability is a good way to take care of the measurability part of weak convergence. Since maps of the form  $x \mapsto d(x, s)$  generate the ball  $\sigma$ -field, a map  $X$  in  $\mathbb{D}$  is ball measurable if and only if the map  $\omega \mapsto d(X(\omega), s)$  is measurable for every  $s \in \mathbb{D}$ . A more common situation is that the maps  $d(X(\omega), s)$  are measurable for a subset of  $s$ . This may yield enough measurability for weak convergence too.

**1.7.1 Lemma.** *Let  $X_\alpha: \Omega_\alpha \mapsto \mathbb{D}$  be asymptotically tight. Then it is asymptotically measurable if and only if  $f(X_\alpha)$  is asymptotically measurable for every ball measurable  $f \in C_b(\mathbb{D})$ .*

**1.7.2 Theorem.** *Let  $X$  be separable and Borel measurable. Then*

- (i)  $X_\alpha \rightsquigarrow X$  if and only if  $E^*f(X_\alpha) \rightarrow Ef(X)$  for all ball measurable  $f \in C_b(\mathbb{D})$ ;

- (ii)  $X_\alpha \rightsquigarrow X$  if and only if there exists a sequence  $s_i$  such that  $(d(X_\alpha, s_1), d(X_\alpha, s_2), \dots) \rightsquigarrow (d(X, s_1), d(X, s_2), \dots)$  in  $\mathbb{R}^\infty$  and with closure satisfying  $P(X \in \overline{\{s_1, s_2, \dots\}}) = 1$ .

**Proofs.** The set of all ball measurable  $f \in C_b(\mathbb{D})$  forms an algebra and separates points of  $\mathbb{D}$ . So the lemma is a special case of Lemma 1.3.13.

The “only if” parts of the theorem are trivial. Furthermore, the “if” part of (ii) is stronger than that of (i). It suffices to show that  $X_\alpha \rightsquigarrow X$  if the condition on the right side in (ii) holds.

Let  $\mathcal{F}$  be the collection of functions of the form

$$f(x) = (1 - p d(x, s_i))^+, \quad p \in \mathbb{Q}^+, \quad i \in \mathbb{N}.$$

Then for any open  $G$ ,

$$1_G(x) = \sup\{f(x) : 0 \leq f \leq 1_G, f \in \mathcal{F}\}, \quad \text{for every } x \in \overline{s_1, s_2, \dots}.$$

Place the countably many elements of  $\mathcal{F}$  with  $0 \leq f \leq 1_G$  in a sequence, and let  $f_m$  be the maximum of the first  $m$  functions. Then  $0 \leq f_m \leq 1_G$  and  $f_m \uparrow 1_G$  on a set of probability 1 under  $X$ . Hence first, for fixed  $m$ ,  $\liminf P_*(X_\alpha \in G) \geq \liminf E_* f_m(X_\alpha) = E f_m(X)$ . Next, letting  $m \rightarrow \infty$  completes the proof in view of the portmanteau theorem. ■

**1.7.3 Example (Cadlag functions).** The space  $D[a, b]$  is the set of all *cadlag functions* on an interval  $[a, b] \subset \bar{\mathbb{R}}$ : functions  $z: [a, b] \mapsto \mathbb{R}$  that are continuous from the right and have limits from the left everywhere. The ball  $\sigma$ -field for the uniform norm equals the  $\sigma$ -field generated by the coordinate projections (Problem 1.7.2). This means that a map  $X: \Omega \mapsto D[a, b]$  is ball measurable if and only if  $X(t): \Omega \mapsto \mathbb{R}$  is measurable for every  $t \in [a, b]$ , a rather weak requirement.

**1.7.4 Example (Pointwise separable processes).** The same idea applied to a more general situation leads to the following. Let  $\mathcal{H}$  be a uniformly bounded class of functions  $h: \mathcal{X} \mapsto \mathbb{R}$  defined on some measurable space  $(\mathcal{X}, \mathcal{B})$ , with the property that there is a countable subcollection  $\mathcal{H}_0 \subset \mathcal{H}$  such that every  $h \in \mathcal{H}$  is the pointwise limit of a sequence  $h_m$  in  $\mathcal{H}_0$ :  $h_m(x) \rightarrow h(x)$  for every  $x$ . Let  $D(\mathcal{H})$  be the subspace of  $\ell^\infty(\mathcal{H})$  of all  $z$  with the property

$$z(h_m) \rightarrow z(h), \quad \text{if } h_m \rightarrow h \text{ pointwise,}$$

for every sequence  $h_m$  in  $\mathcal{H}_0$ . Then the ball  $\sigma$ -field for the uniform norm is contained in the  $\sigma$ -field generated by the coordinate projections  $z \mapsto z(h)$ ,  $h \in \mathcal{H}$ . Thus every stochastic process  $X: \Omega \mapsto D(\mathcal{H})$  is ball measurable. Remember that  $X$  is called a stochastic process if every coordinate  $X(h)$  is a measurable map into  $\mathbb{R}$ .

A finite signed measure  $P$  on  $(\mathcal{X}, \mathcal{B})$  induces an element  $h \mapsto Ph = \int h dP$  of  $D(\mathcal{H})$ . So any stochastic process  $X$  for which every  $X(\omega)$  has the character of a signed measure can be viewed a ball measurable map into  $D(\mathcal{H})$ . An example is the empirical measure  $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$  of a sample  $X_1, \dots, X_n$  of random elements in  $(\mathcal{X}, \mathcal{B})$ ; another is the empirical process  $\sqrt{n}(\mathbb{P}_n - P)$ .

The condition on  $\mathcal{H}$  is met, for instance, by the set of all indicator functions of closed or open ellipsoids, rectangles, and half-spaces in a Euclidean space.

**1.7.5 Example (Measurable processes with Suslin index set).** Let  $T$  be a metric space equipped with its Borel  $\sigma$ -field. A map  $X: \Omega \mapsto \mathbb{R}^T$  is a *measurable stochastic process* if  $X: \Omega \times T \mapsto \mathbb{R}$  is jointly measurable for the product of  $\mathcal{A}$  and the Borel sets of  $T$ . This is a little stronger than measurability of every coordinate  $X(t)$ .

If  $T$  is a Borel subset of a Polish space and  $X: \Omega \mapsto \ell^\infty(T)$  is a measurable process, then  $d(X, z) = \|X - z\|_T$  is measurable for the  $P$ -completion of  $(\Omega, \mathcal{A})$  for every measurable  $z$ . Consequently, every measurable process  $X$  with bounded sample paths is a ball measurable map into  $\ell_m^\infty(T)$ , the subspace of all measurable  $z \in \ell^\infty(T)$ , at least if  $(\Omega, \mathcal{A}, P)$  is complete.

The claim follows from the theorem on measurable projections. The set  $\{\omega: \|X(\omega) - z\|_T > c\}$  is the projection on  $\Omega$  of the product measurable set  $\{(\omega, t): |X(t, \omega) - z(t)| > c\}$ . Projections are not always measurable, but under the stated conditions the present one is.<sup>†</sup>

At closer inspection the argument actually yields more. Suppose  $\mathbf{S}$  is a Suslin topological space<sup>‡</sup> and let  $\phi: \mathbf{S} \mapsto T$  be an arbitrary map. Then any  $X: \Omega \mapsto \ell^\infty(T)$  for which  $(s, \omega) \mapsto X(\phi(s), \omega)$  is jointly measurable has a measurable supremum  $\sup_{t \in \phi(\mathbf{S})} |X(t)|$ .

## Problems and Complements

- Let  $T$  be a compact semimetric space with dense subset  $S$ . The Borel  $\sigma$ -field on  $C(T)$  is the smallest  $\sigma$ -field making all projections  $z \mapsto z(s)$ ,  $s \in S$ , measurable.

[**Hint:** The closed ball around  $z_0$  of radius  $r$  equals  $\cup_{s \in S_0} \{z: |z(s) - z_0(s)| \leq r\}$ , where  $S_0$  is a countable dense subset of  $S$ . So every closed and consequently every open ball is projection measurable. For a separable metric space, the ball  $\sigma$ -field and Borel  $\sigma$ -field are the same.]

<sup>†</sup> If  $(\Omega, \mathcal{A})$  is a measurable space and  $Y$  a Polish space, then the projection on  $\Omega$  of any product measurable subset of  $\Omega \times Y$  (a set contained in the product  $\sigma$ -field of  $\mathcal{A}$  and the Borel sets of  $Y$ ) is *universally measurable*; that is, it is contained in the  $P$ -completion of  $\mathcal{A}$  for every probability measure  $P$  on  $(\Omega, \mathcal{A})$  (Cohn (1980), p. 281).

<sup>‡</sup> A (Hausdorff) topological space is called *Suslin* if it is the continuous image of a Polish space. A subset of a Polish space is called *analytic* if it is Suslin for the relative topology. Every Borel subset of a Polish space is analytic/Suslin. See Cohn (1980), page 292.

2. The ball  $\sigma$ -field on  $D[a, b]$  with respect to the uniform metric equals the  $\sigma$ -field generated by the coordinate projections.

[Hint: The set  $\{z: z(t) > c\}$  is the union of the balls of radius  $n$  around  $z_{n,c} = c + (n + n^{-1})1_{[t,t+n^{-1}]}$ .]

3. Equip  $D[0, 1]$  with the uniform metric, and define  $X: [0, 1] \mapsto \mathbb{D}$  by  $X(\omega) = 1_{[\omega, 1]}$  (the empirical process based on one observation  $\omega$ ). If  $[0, 1]$  is equipped with the Borel  $\sigma$ -field, then  $X$  is ball measurable, but not Borel measurable.

[Hint: Let  $B_s$  be the open ball of radius  $1/2$  in  $\mathbb{D}$  around the function  $1_{[s, 1]}$ . Since an open ball is open, so is the union  $G = \cup_{s \in S} B_s$ , for any  $S \subset [0, 1]$ . Now  $X(\omega) \in B_s$  if and only if  $\omega = s$ . Thus  $\{X \in G\} = S$ . The conclusion is that  $X$  is Borel measurable if and only if every subset of  $[0, 1]$  is a Borel set.]

4. (Properties of the ball  $\sigma$ -field)

- (i) For a separable, semimetric space, the ball  $\sigma$ -field equals the Borel  $\sigma$ -field.
- (ii) Every closed, separable set is ball measurable.
- (iii) If  $K$  is compact, then  $K^\delta$  is ball measurable.
- (iv) Every separable probability measure on the ball  $\sigma$ -field has a unique extension to the Borel  $\sigma$ -field.

[Hint: If the sequence  $s_i$  is dense in a closed  $F$ , then  $F^\delta = \cup_{i=1}^{\infty} B(s_i, \delta)$ , where  $B(s, \delta)$  is the ball of radius  $\delta$  around  $s$ . Since a compact set is closed and separable, (iii) follows. For (ii) note that  $F = \cap_{m=1}^{\infty} F^{1/m}$ ; (i) is a consequence of (ii). For (iv) let  $S$  be a closed, separable set with  $L(\mathbb{D}_0) = 1$ . For every Borel set  $B$ , the set  $B \cap \mathbb{D}_0$  is a Borel set contained in  $\mathbb{D}_0$ , hence is contained in the ball  $\sigma$ -field on  $\mathbb{D}_0$  by (i), but then also in the ball  $\sigma$ -field on  $\mathbb{D}$ . Define the extension by  $\tilde{L}(B) = L(B \cap \mathbb{D}_0)$ .]

5. (Completely regular points) Pollard (1984) calls a point  $x$  in a metric space  $\mathbb{D}$  equipped with a  $\sigma$ -field  $\mathcal{A}$  *completely regular* if, for every open neighborhood  $G$  of  $x$ , there is a measurable, uniformly continuous  $f: \mathbb{D} \mapsto \mathbb{R}$  with  $0 \leq f \leq 1_G$  and  $f(x) = 1$ . If  $X$  is an  $\mathcal{A}$ -measurable map that takes all its values in a separable set of completely regular points, then it is Borel measurable. If, moreover,  $X_\alpha$  are arbitrary maps with  $E^*f(X_\alpha) \rightarrow Ef(X)$  for every  $\mathcal{A}$ -measurable, uniformly continuous  $f$ , then  $X_\alpha \rightsquigarrow X$ . Since every point is completely regular for the ball  $\sigma$ -field, this extends the results obtained in the section about the ball  $\sigma$ -field.

[Hint: Pollard (1984), page 88, Exercises [14]–[16].]

## 1.8

# Hilbert Spaces

Let  $\mathbb{H}$  be a (real) Hilbert space with inner product  $\langle \cdot, \cdot \rangle$  and complete orthonormal system  $\{e_j : j \in J\}$ . Thus  $\langle e_i, e_j \rangle$  equals 0 or 1 if  $i \neq j$  or  $i = j$ , respectively, and every  $x \in \mathbb{H}$  can be written as

$$x = \sum_j \langle x, e_j \rangle e_j.$$

The sum contains at most countably many nonzero terms, because only countably many inner products  $\langle x, e_j \rangle$  are nonzero for every  $x$ . The series converges unconditionally and  $\|x\|^2 = \sum_j \langle x, e_j \rangle^2$ . The orthogonal projection of  $x$  into the linear span of a subset  $\{e_i : i \in I\}$  of the base equals  $\sum_{i \in I} \langle x, e_i \rangle e_i$ ; the square distance of  $x$  to this linear span equals  $\sum_{j \notin I} \langle x, e_j \rangle^2$ .

Call a net  $X_\alpha : \Omega_\alpha \mapsto \mathbb{H}$  *asymptotically finite-dimensional* if, for all  $\delta, \varepsilon > 0$ , there exists a finite subset  $\{e_i : i \in I\}$  of the orthonormal base such that

$$\limsup_\alpha P^* \left( \sum_{j \notin I} \langle X_\alpha, e_j \rangle^2 > \delta \right) < \varepsilon.$$

Compact sets in a Banach space can be characterized by being bounded and being contained in the  $\delta$ -shell of a finite-dimensional subspace, for every  $\delta$ . This leads to the following characterization of asymptotic tightness of a sequence of random maps.

**1.8.1 Lemma.** *A net of random maps  $X_\alpha : \Omega_\alpha \mapsto \mathbb{H}$  is asymptotically tight if and only if it is asymptotically finite-dimensional and the nets  $\langle X_\alpha, e_j \rangle$  are asymptotically tight for every  $j$ .*

**Proof.** The set  $K = \{\sum_{i \in I} a_i e_i : a \in \mathbb{R}^I, \max |a_i| \leq k\}$  is compact for every finite set  $I$  and constant  $k$ . For every fixed  $\delta$ ,

$$P^*(X_\alpha \notin K^\delta) \leq P^*\left(\max_{i \in I} |\langle X_\alpha, e_i \rangle| > k\right) + P^*\left(\sum_{j \notin I} \langle X_\alpha, e_j \rangle^2 \geq \delta^2\right).$$

If the net  $X_\alpha$  is asymptotically finite-dimensional, then the last probability can be made arbitrarily small by the choice of  $I$ . Next, the first probability on the right can be made arbitrarily small by the choice of  $k$ .

Conversely, suppose that the net  $X_\alpha$  is asymptotically tight. Then every net  $\langle X_\alpha, h \rangle$  is asymptotically tight by continuity of the inner product. Given  $\varepsilon > 0$ , let  $K$  be a compact set such that  $\limsup_\alpha P^*(X_\alpha \notin K^\delta) < \varepsilon$ , for every  $\delta > 0$ . Since  $K$  is compact, there exists a finite set  $I$  such that  $K \subset \text{lin}(e_i : i \in I)^\delta$ . Then every  $x \in K^\delta$  is at a distance of at most  $2\delta$  from  $\text{lin}(e_i : i \in I)$ ; the square of this distance equals  $\sum_{j \notin I} \langle x, e_j \rangle^2$ . Conclude that  $\sum_{j \notin I} \langle X_\alpha, e_j \rangle^2 \geq 4\delta^2$  implies that  $X_\alpha \notin K^\delta$ . Hence the net  $X_\alpha$  is asymptotically finite-dimensional. ■

**1.8.2 Lemma.** Let  $X_\alpha : \Omega_\alpha \mapsto \mathbb{H}$  be asymptotically tight. Then it is asymptotically measurable if and only if  $\langle X_\alpha, e_j \rangle$  is asymptotically measurable for every  $j$ .

**1.8.3 Lemma.** Tight Borel measurable random elements  $X$  and  $Y$  in  $\mathbb{H}$  are equal in distribution if and only if the random variables  $\langle X, h \rangle$  and  $\langle Y, h \rangle$  are equal in distribution for every  $h \in \mathbb{H}$ .

**1.8.4 Theorem.** A net of random maps  $X_\alpha : \Omega_\alpha \mapsto \mathbb{H}$  converges in distribution to a tight Borel measurable random variable  $X$  if and only if it is asymptotically finite-dimensional and the net  $\langle X_\alpha, h \rangle$  converges in distribution to  $\langle X, h \rangle$  for every  $h \in \mathbb{H}$ .

**Proofs.** For the lemmas, consider the collection  $\mathcal{F}$  of all functions  $f : \mathbb{H} \mapsto \mathbb{R}$  of the form

$$f(x) = g(\langle x, e_{j_1} \rangle, \dots, \langle x, e_{j_k} \rangle), \quad g \in C_b(\mathbb{R}^k), \quad j_1, \dots, j_k \in J, \quad k \in \mathbb{N}.$$

This collection forms an algebra and a vector lattice and separates points of  $\mathbb{H}$ . Therefore, the lemmas are consequences of Lemmas 1.3.13 and 1.3.12, respectively.

The theorem follows from combination of the lemmas with Prohorov's theorem. ■

**1.8.5 Example (Central limit theorem).** If  $X_1, X_2, \dots$  are i.i.d. Borel measurable random elements in a separable Hilbert space  $\mathbb{H}$  with mean zero (i.e.,  $E\langle X_1, h \rangle = 0$  for every  $h$ ), and  $E\|X_1\|^2 < \infty$ , then the sequence

$n^{-1/2} \sum_{i=1}^n X_i$  converges in distribution to a Gaussian variable  $G$ . The distribution of  $G$  is determined by the distribution of its marginals  $\langle G, h \rangle$ , which are  $N(0, E\langle X, h \rangle^2)$  distributed for every  $h \in \mathbb{H}$ .

This follows from the theorem. The sequence  $\langle n^{-1/2} \sum_{i=1}^n X_i, h \rangle$  converges in distribution to a normal distribution by the central limit theorem for real-valued random variables. Second, if  $e_1, e_2, \dots$  is an orthonormal base for  $\mathbb{H}$ , then

$$E \sum_{j>J} \left\langle n^{-1/2} \sum_{i=1}^n X_i, e_j \right\rangle^2 = E \sum_{j>J} \langle X_1, e_j \rangle^2 \rightarrow 0, \quad J \rightarrow \infty,$$

by dominated convergence, because the series on the right is bounded by  $\|X_1\|^2$ , by Bessel's inequality. Thus the sequence  $n^{-1/2} \sum_{i=1}^n X_i$  is asymptotically finite-dimensional.

**1.8.6 Example (Anderson-Darling statistic).** Let  $X_1, X_2, \dots$  be real-valued random variables with cumulative distribution function  $F$ . The random functions

$$Z_i(t) = \frac{1\{X_i \leq t\} - F(t)}{\sqrt{F(t)(1-F)(t)}}$$

are contained in  $L_2(\mathbb{R}, \mu)$ , and  $E\|Z_1\|_2^2 = \mu(\mathbb{R}) < \infty$  for every finite measure  $\mu$ . Conclude that the sequence  $n^{-1/2} \sum_{i=1}^n Z_i$  converges in distribution to a Gaussian variable  $G$  in  $L_2(\mathbb{R}, \mu)$ . One consequence is the weak convergence of the sequence of  $L_2(\mu)$ -norms

$$\int \frac{n(\mathbb{F}_n - F)^2(t)}{F(t)(1-F)(t)} d\mu(t) \rightsquigarrow \int G^2(t) d\mu(t),$$

where  $\mathbb{F}_n$  is the empirical distribution function of  $X_1, \dots, X_n$ . The choice  $\mu = F$  yields the Anderson-Darling statistic.

## Problems and Complements

- If  $K$  is compact, then for every  $\delta > 0$  there exists a finite set  $I$  with  $K \subset \text{lin}(e_i : i \in I)^\delta$ .

[**Hint:** If not then for every finite subset  $I$  there exists  $x_I \in K$  at distance at least  $\delta$  from  $\text{lin}(e_i : i \in I)$ . Direct the finite subsets by inclusion. The net has a converging subnet. Its limit point  $x$  has distance at least  $\delta$  to every finite-dimensional space  $\text{lin}(e_i : i \in I)$ . Thus  $\sum_{j \notin I} \langle x, e_j \rangle^2 \geq \delta^2$  for every finite set  $I$ .]

## 1.9

# Convergence: Almost Surely and in Probability

For nets of maps defined on a single, fixed probability space  $(\Omega, \mathcal{A}, P)$ , convergence almost surely and in probability are frequently used modes of stochastic convergence, stronger than weak convergence. In this section we consider their nonmeasurable extensions together with the concept of almost uniform convergence, which is equivalent to outer almost sure convergence for sequences, but stronger and more useful for general nets.

**1.9.1 Definition.** Let  $X_\alpha, X: \Omega \mapsto \mathbb{D}$  be arbitrary maps.

- (i)  $X_\alpha$  converges in outer probability to  $X$  if  $d(X_\alpha, X)^* \rightarrow 0$  in probability; this means that  $P(d(X_\alpha, X)^* > \varepsilon) = P^*(d(X_\alpha, X) > \varepsilon) \rightarrow 0$ , for every  $\varepsilon > 0$ , and is denoted by  $X_\alpha \xrightarrow{P^*} X$ .
- (ii)  $X_\alpha$  converges almost uniformly to  $X$  if, for every  $\varepsilon > 0$ , there exists a measurable set  $A$  with  $P(A) \geq 1 - \varepsilon$  and  $d(X_\alpha, X) \rightarrow 0$  uniformly on  $A$ ; this is denoted  $X_\alpha \xrightarrow{\text{au}} X$ .
- (iii)  $X_\alpha$  converges outer almost surely to  $X$  if  $d(X_\alpha, X)^* \rightarrow 0$  almost surely for some versions of  $d(X_\alpha, X)^*$ ; this is denoted  $X_\alpha \xrightarrow{\text{as}*} X$ .
- (iv)  $X_\alpha$  converges almost surely to  $X$  if  $P_*(\lim d(X_\alpha, X) = 0) = 1$ ; this is denoted  $X_\alpha \xrightarrow{\text{as}} X$ .

The first three concepts in this list are the most important. The fourth is tricky — it does not behave as one might expect, in general. In fact, for nonmeasurable sequences, even convergence *everywhere* (which is stronger than (iv)) does not imply any of the other three forms of convergence. This is one of those phenomena that remind us that measurability, though often

present, should not be taken too lightly. The counterexample given below isn't even very complicated, though contrived, perhaps.

For sequences the second and third modes of convergence are equivalent, but for general nets the third loses much of its value. First, there is the nuisance that for general nets outer almost sure convergence depends on the versions of the minimal measurable covers one uses. Second, even for measurable nets  $X_\alpha$ , outer almost sure convergence is rather weak. For instance, it does not imply convergence in probability. For these reasons we consider (iii) and (iv) for sequences only. The Problems and Complements section gives more details about the general implications between the four forms of convergence. In the following text, we use the notation  $X_\alpha$  for a general net and  $X_n$  for a sequence, without further mention.

For sequences things are partly as they should be. Outer almost sure convergence (iii) is stronger than convergence in outer probability. Equivalence of almost uniform and outer almost sure convergence, part (iii) of the following lemma, is known as *Egorov's theorem*.

**1.9.2 Lemma.** *Let  $X$  be Borel measurable. Then*

- (i)  $X_n \xrightarrow{\text{as*}} X$  implies  $X_n \xrightarrow{\text{P*}} X$ ;
- (ii)  $X_n \xrightarrow{\text{P*}} X$  if and only if every subsequence  $X_{n'}$  has a further subsequence  $X_{n''}$  with  $X_{n''} \xrightarrow{\text{as*}} X$ ;
- (iii)  $X_n \xrightarrow{\text{as*}} X$  if and only if  $X_n \xrightarrow{\text{au}} X$ .

**1.9.3 Lemma.** *Let  $X$  be Borel measurable. Then*

- (i)  $X_\alpha \xrightarrow{\text{au}} X$  if and only if  $\sup_{\beta \geq \alpha} d(X_\beta, X)^* \xrightarrow{\text{P}} 0$  if and only if  $\sup_{\beta \geq \alpha} d(X_\beta, X) \xrightarrow{\text{P*}} 0$ ;
- (ii)  $X_\alpha \xrightarrow{\text{au}} X$  implies  $X_\alpha \xrightarrow{\text{P*}} X$ .

**Proofs.** (iii). Suppose  $X_n$  converges outer almost surely to  $X$ . Fix  $\varepsilon > 0$ . Set  $A_n^k = \{\sup_{m \geq n} d(X_m, X)^* > 1/k\}$ . Then, for every fixed  $k$ , it holds that  $P(A_n^k) \downarrow 0$  as  $n \rightarrow \infty$ . Choose  $n_k$  with  $P(A_{n_k}^k) \leq \varepsilon/2^k$ , and set  $A = \Omega - \bigcup_{k=1}^{\infty} A_{n_k}^k$ . Then  $P(A) \geq 1 - \varepsilon$  and  $d(X_n, X)^* \leq 1/k$ , for  $n \geq n_k$  and  $\omega \in A$ . Thus  $X_n$  converges to  $X$  almost uniformly. Conversely, suppose  $X_n$  converges almost uniformly to  $X$ . Fix  $\varepsilon > 0$ , and let  $A$  be as in the definition of almost uniform convergence. Fix  $\eta > 0$ . Then  $d(X_n, X)^* 1_A = (d(X_n, X) 1_A)^* \leq \eta$  for sufficiently large  $n$ , since  $\eta$  is a measurable function and  $\eta \geq d(X_n, X) 1_A$  for sufficiently large  $n$ . Thus  $d(X_n, X)^* \rightarrow 0$  for almost all  $\omega \in A$ .

Next consider the second lemma first.

(i). It is easy to see that the first statement implies the second, which implies the third. For the converse, fix  $\varepsilon > 0$ . Take  $\alpha_k$  such that  $P^*(\sup_{\beta \geq \alpha_k} d(X_\beta, X) > 1/k) \leq \varepsilon/2^k$ . Call the set within brackets  $A_k$ , and set  $A = \Omega - \bigcup_{k=1}^{\infty} A_k^*$ . Then  $P(A) \geq 1 - \varepsilon$ , and for every  $\omega \in A$  and  $\alpha \geq \alpha_k$ , it holds that  $d(X_\alpha, X) \leq 1/k$ .

(ii). This is an immediate consequence of (i).

Finally, consider the first lemma again. Part (i) follows from a combination of the other statements. For (ii) suppose  $X_n \xrightarrow{P^*} X$ . Take  $n_1 < n_2 < \dots$  such that  $P(d(X_{n_j}, X)^* > 1/j) < 2^{-j}$ . Then  $P(d(X_{n_j}, X)^* > 1/j, \text{i.o.}) = 0$ , by the Borel-Cantelli lemma. Thus  $d(X_{n_j}, X) \leq 1/j$  eventually, for almost every  $\omega$ . The converse is trivial in view of (i). ■

**1.9.4 Counterexample.**  $X_n(\omega) \rightarrow 0$  for every  $\omega$  does not imply  $X_n \xrightarrow{\text{as}*} 0$ . Let  $(\Omega, \mathcal{A}, \lambda)$  be  $[0,1]$  with Borel sets and Lebesgue measure. There exists a decreasing sequence of sets  $B_n$ , with  $\cap_{n=1}^{\infty} B_n = \emptyset$ , but  $\lambda^*(B_n) = 1$  for every  $n$ . Thus  $1_{B_n}(\omega) \rightarrow 0$  for every  $\omega$ , but  $|1_{B_n} - 0|^* = 1_{B_n^*} = 1$  for every  $n$ , so that  $1_{B_n}$  certainly doesn't converge to zero almost uniformly.

One construction of the sequence  $B_n$  is as follows.<sup>#</sup> A set  $H \subset \mathbb{R}$  is called a *Hamel base* for  $\mathbb{R}$  over  $\mathbb{Q}$  if every element of  $\mathbb{R}$  can be expressed as a finite linear combination  $\sum_{i=1}^k q_i h_i$ , where  $q_i \in \mathbb{Q}$ ,  $h_i \in H$ , and  $k \in \mathbb{N}$ . Such a base exists by Zorn's lemma (which is equivalent to the axiom of choice). Any such Hamel base  $H$  is an uncountable set. Cut one such Hamel base  $H$  into countably many disjoint, nonempty subsets (this is possible by the axiom of choice), and let  $H_n$  be the union of the first  $n$  cuts. Thus  $H_n \uparrow$  and  $H = \cup_{n=1}^{\infty} H_n$ . Let  $C_n$  be the subspace of  $\mathbb{R}$  spanned by  $H_n$ : all finite linear combinations  $\sum_{i=1}^k q_i h_i$  with  $q_i \in \mathbb{Q}$  and  $h_i \in H_n$ . Then  $C_n \uparrow$  and  $\mathbb{R} = \cup_{n=1}^{\infty} C_n$ .

Let  $K$  be an arbitrary compact set. Then either  $\lambda(K) = 0$  or  $K - K$  contains an interval  $(-\delta, \delta)$  for some  $\delta > 0$ . Indeed, if  $\lambda(K) > 0$ , then for  $\delta > 0$  sufficiently small,  $\lambda(K^\delta) < 2\lambda(K)$ . If  $(-\delta, \delta)$  is not contained in  $K - K$ , then there is an  $|x| < \delta$  for which the sets  $K$  and  $K + x$  are disjoint. But this is impossible, since it would imply  $2\lambda(K) = \lambda(K \cup K + x) \leq \lambda(K^\delta)$ .

Now suppose that  $K$  is a compact set contained in  $C_n$ . Then  $K - K \subset C_n = C_n - C_n$ . Since  $C_n$  does not contain an interval, it can be inferred that  $\lambda(K) = 0$ . Together with the inner regularity of Lebesgue measure, this implies  $\lambda_*(C_n) = 0$ . Finally, set  $B_n = [0, 1] - C_n$ .

There is a continuous mapping theorem for convergence in outer probability and almost uniform convergence under exactly the same conditions as for weak convergence.

**1.9.5 Theorem (Continuous mapping).** *Let  $g: \mathbb{D} \mapsto \mathbb{E}$  be continuous at every point of a Borel set  $\mathbb{D}_0 \subset \mathbb{D}$ . Let  $X$  be Borel measurable with  $P(X \in \mathbb{D}_0) = 1$ . Then*

- (i)  $X_\alpha \xrightarrow{P^*} X$  implies that  $g(X_\alpha) \xrightarrow{P^*} g(X)$ ;
- (ii)  $X_\alpha \xrightarrow{\text{au}} X$  implies that  $g(X_\alpha) \xrightarrow{\text{au}} g(X)$ .

**Proof.** (i). Fix  $\varepsilon > 0$ . Let  $B_k$  be the set of all  $x$  for which there exist  $y$  and  $z$  within the open ball of radius  $1/k$  around  $x$  with  $e(g(y), g(z)) > \varepsilon$ . Then

---

<sup>#</sup> Cohn (1980).

$B_k$  is open and the sequence  $B_k$  is decreasing. Moreover,  $P(X \in B_k) \downarrow 0$ , since every point in  $\cap_{k=1}^{\infty} B_k$  is a point of discontinuity of  $g$ . Now, for every fixed  $k$ ,

$$\begin{aligned} P^*(e(g(X_\alpha), g(X)) > \varepsilon) &\leq P^*(X \in B_k \text{ or } d(X_\alpha, X) \geq 1/k) \\ &\rightarrow P^*(X \in B_k). \end{aligned}$$

Finally, let  $k \rightarrow \infty$ .

(ii). Add a supremum in the proof of (i) twice to obtain that  $\sup_{\beta \geq \alpha} d(X_\beta, X) \xrightarrow{P^*} 0$  implies that  $\sup_{\beta \geq \alpha} e(g(X_\beta), g(X)) \xrightarrow{P^*} 0$ . This is equivalent to the statement of (ii). ■

It was shown by counterexample that convergence almost surely does not imply convergence outer almost surely or in outer probability. The problem is a possible lack of measurability. Since convergence almost surely is so easy to work with, it is of interest to know how much measurability is sufficient to remove the difference. A trivial but useful result is that  $X_n \xrightarrow{\text{as}} X$  together with measurability of  $d(X_n, X)$  implies  $X_n \xrightarrow{\text{as*}} X$ . The next theorem gives an exact answer to the problem, although, admittedly, it may often not be more useful than the trivial sufficient condition just mentioned.

What is needed is some sort of asymptotic measurability of  $X_n$ . Asymptotic measurability as introduced before is slightly too weak. Call  $X_n$  *strongly asymptotically measurable* if

$$f(X_n)^* - f(X_n)_* \xrightarrow{\text{as}} 0, \quad \text{for every } f \in C_b(\mathbb{D}).$$

This is stronger than (ordinary) asymptotic measurability: the latter requires convergence in probability rather than almost sure convergence. Of course, both are implied by measurability for every  $n$ . It can be shown that  $X_n \xrightarrow{\text{as}} X$  together with asymptotic measurability implies  $X_n \xrightarrow{P^*} X$  (Problem 1.9.1). For the stronger conclusion that  $X_n \xrightarrow{\text{as*}} X$ , strong asymptotic measurability is necessary. Perhaps the best part of the next theorem is (iv), which implies that almost sure convergence plus ball measurability implies almost uniform convergence.

**1.9.6 Theorem.** Let  $X$  be Borel measurable and separable. Then the following statements are equivalent:

- (i)  $X_n \xrightarrow{\text{as*}} X$ ;
- (ii)  $X_n \xrightarrow{\text{as}} X$  and  $d(X_n, X)^* - d(X_n, X)_* \xrightarrow{\text{as}} 0$ ;
- (iii)  $X_n \xrightarrow{\text{as}} X$  and  $X_n$  is strongly asymptotically measurable;
- (iv)  $X_n \xrightarrow{\text{as}} X$  and  $d(X_n, s)$  is strongly asymptotically measurable for every  $s$  in a set  $S$  with  $P(X \in \bar{S}) = 1$ .

In particular, if  $X_n \xrightarrow{\text{as}} X$  and every  $X_n$  and  $X$  is ball measurable, then  $X_n \xrightarrow{\text{as*}} X$ .

**Proof.** The equivalence of (i) and (ii) is easy to see.

(i)  $\Rightarrow$  (iii)  $\Rightarrow$  (iv). If (i) holds, then trivially  $X_n \xrightarrow{\text{as}} X$ . Furthermore, by the continuous mapping theorem  $f(X_n) \xrightarrow{\text{as*}} f(X)$ , for every continuous  $f$ . Consequently

$$\begin{aligned} f(X_n)^* - f(X_n)_* &= (f(X_n) - f(X))^* - (f(X_n) - f(X))_* \\ &\leq 2|f(X_n) - f(X)|^* \xrightarrow{\text{as}} 0. \end{aligned}$$

Thus (iii) holds. Next, (iv) is immediate from the fact that the function  $x \mapsto f(d(x, s))$  is contained in  $C_b(\mathbb{D})$  for every bounded, continuous function  $f$  on  $\mathbb{R}$  and  $s \in \mathbb{D}$ .

(iv)  $\Rightarrow$  (i). Assume without loss of generality that  $S$  is countable and that  $\mathbb{D}$  is complete. Let  $f_1, f_2, \dots$  be the set of functions  $x \mapsto (1 - m d(x, s))^+$ , where  $m \in \mathbb{N}$  and  $s \in S$ . Define a semimetric on  $\mathbb{D}$  by

$$e(x, y) = \sup_{j \in \mathbb{N}} \frac{1}{j} |f_j(x) - f_j(y)|.$$

A sequence  $x_n$  in  $\mathbb{D}$  converges to a point  $x \in \overline{S}$  if and only if  $f_j(x_n) \rightarrow f_j(x)$  for every  $j$ ; equivalently, if and only if  $e(x_n, x) \rightarrow 0$ . Conclude that, for every compact  $K \subset \overline{S}$ , there is for every  $\varepsilon > 0$  a  $\delta > 0$  such that, for every  $x \in K$  and  $y \in \mathbb{D}$ ,  $e(y, x) < \delta$  implies  $d(y, x) < \varepsilon$ .

Fix  $j$ . If the first assertion of (iv) holds, then  $f_j(X_n) \xrightarrow{\text{as}} f_j(X)$ . Under the second assertion,

$$(f_j(X_n) - f_j(X))^* - (f_j(X_n) - f_j(X))_* = f_j(X_n)^* - f_j(X_n)_* \xrightarrow{\text{as*}} 0.$$

Combination yields that  $|f_j(X_n) - f_j(X)| \xrightarrow{\text{as*}} 0$ . By Egorov's theorem, there is a measurable set  $A_j$  with  $P(\Omega - A_j) < \varepsilon/2^j$  and  $f_j(X_n) - f_j(X) \rightarrow 0$  uniformly on  $A_j$ . The set  $A = \cap_{j=1}^{\infty} A_j$  has  $P(\Omega - A) < \varepsilon$  and  $e(X_n, X) \rightarrow 0$  uniformly on  $A$ .

Take a compact  $K$  with  $P(X \in K) \geq 1 - \varepsilon$ . Then  $B = A \cap \{X \in K\}$  has  $P(\Omega - B) < 2\varepsilon$  and  $d(X_n, X) \rightarrow 0$  uniformly on  $B$ . Thus  $X_n \xrightarrow{\text{as*}} X$ . ■

## Problems and Complements

- Let  $X$  be Borel measurable and separable. Then the following statements are equivalent:

- (i)  $X_n \xrightarrow{\text{as}} X$  and  $d(X_n, X)$  is asymptotically measurable;
- (ii)  $X_n \xrightarrow{\text{as}} X$  and  $X_n$  is asymptotically measurable;
- (iii)  $X_n \xrightarrow{\text{as}} X$  and  $d(X_n, s)$  is asymptotically measurable for all  $s$  in a set  $S$  with  $P(X \in \overline{S}) = 1$ .

Furthermore, the statements imply

- (iv)  $X_n \xrightarrow{\text{P*}} X$ .

It is possible to give this a fancier formulation by introducing the notion of *convergence in probability* (different from convergence in outer probability). Say that  $X_n \xrightarrow{\text{P}} X$  if every subsequence  $X_{n'}$  has a further subsequence with  $X_{n''} \xrightarrow{\text{as}} X$ . If, in (i) - (iii),  $X_n \xrightarrow{\text{as}} X$  is replaced by  $X_n \xrightarrow{\text{P}} X$ , then (i) - (iv) are equivalent.

## 1.10

# Convergence: Weak, Almost Uniform, and in Probability

Consider the relationships between the convergence concepts introduced in the previous section and weak convergence. First we shall be a bit formal and note that convergence in probability to a constant can be defined for maps with *different* domains  $(\Omega_\alpha, \mathcal{A}_\alpha, P_\alpha)$  too, so that it is not covered by Definition 1.9.1 in the preceding section.

**1.10.1 Definition.** Let  $X_\alpha: \Omega_\alpha \rightarrow \mathbb{D}$  be an arbitrary net of maps and  $c \in \mathbb{D}$ . Then  $X_\alpha$  converges in outer probability to  $c$  if  $P^*(d(X_\alpha, c) > \varepsilon) \rightarrow 0$ , for every  $\varepsilon > 0$ . This is denoted  $X_\alpha \xrightarrow{P^*} c$ .

In general, convergence in outer probability is stronger than weak convergence, though they are equivalent if the limit is constant.

**1.10.2 Lemma.** Let  $X_\alpha, Y_\alpha$  be arbitrary maps and  $X$  Borel measurable.

- (i) If  $X_\alpha \rightsquigarrow X$  and  $d(X_\alpha, Y_\alpha) \xrightarrow{P^*} 0$ , then  $Y_\alpha \rightsquigarrow X$ .
- (ii) If  $X_\alpha \xrightarrow{P^*} X$ , then  $X_\alpha \rightsquigarrow X$ .
- (iii)  $X_\alpha \xrightarrow{P^*} c$  if and only if  $X_\alpha \rightsquigarrow c$ .

**Proof.** (i). Let  $F$  be closed. Then, for every fixed  $\varepsilon > 0$ , it holds that  $\limsup P^*(Y_\alpha \in F) = \limsup P^*(Y_\alpha \in F \wedge d(X_\alpha, Y_\alpha)^* \leq \varepsilon) \leq \limsup P^*(X_\alpha \in \overline{F^\varepsilon}) \leq P(X \in \overline{F^\varepsilon})$ . Letting  $\varepsilon \downarrow 0$  completes the proof.

(ii). Clearly,  $X \rightsquigarrow X$  and  $d(X, X_\alpha) \xrightarrow{P^*} 0$ . Apply (i).

(iii). One direction follows from (ii). For the other, note that  $P^*(d(X_\alpha, c) > \varepsilon) = P^*(X_\alpha \notin B(c, \varepsilon))$ , where  $B(c, \varepsilon)$  is the ball of radius  $\varepsilon$

around  $c$ . By the portmanteau theorem, the limsup of this is smaller than or equal to  $P(c \notin B(c, \varepsilon)) = 0$ . ■

The second assertion of the previous lemma can certainly not be inverted: weakly convergent maps need not even be defined on the same probability space. However, according to the almost sure representation theorem, for every weakly convergent net there is an almost surely convergent net (defined on some probability space) that is the “same” as far as laws are concerned. The nonmeasurable version of this result sounds somewhat more complicated than the measurable version; so it is worth considering the case of measurable maps first. For every  $\alpha$ , let  $\mathcal{D}_\alpha$  be a  $\sigma$ -field on  $\mathbb{D}$ , not larger than the Borel  $\sigma$ -field. For convenience of notation, the limit variable will be written as  $X_\infty$  rather than  $X$  and a statement that is valid “for every  $\alpha$ ” will be understood to apply to  $\alpha = \infty$  too.

**1.10.3 Theorem (a.s. representations).** *Let  $X_\alpha: \Omega_\alpha \mapsto \mathbb{D}$  be  $\mathcal{D}_\alpha$ -measurable maps. If  $X_\alpha \rightsquigarrow X_\infty$  and  $X_\infty$  is separable, then there exist  $\mathcal{D}_\alpha$ -measurable maps  $\tilde{X}_\alpha: \tilde{\Omega} \mapsto \mathbb{D}$  defined on some probability space  $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{P})$  with*

- (i)  $\tilde{X}_\alpha \xrightarrow{\text{au}} \tilde{X}_\infty$ ;
- (ii)  $\tilde{X}_\alpha$  and  $X_\alpha$  are equal in law on  $\mathcal{D}_\alpha$  for every  $\alpha$ .

Usually this theorem will be applied with every  $\mathcal{D}_\alpha$  equal to the Borel or ball  $\sigma$ -field. In any case, the smaller the  $\mathcal{D}_\alpha$  are, the weaker the result. In the extreme case that the  $\mathcal{D}_\alpha$  are the trivial  $\sigma$ -fields, the theorem is still true, but it yields “representations”  $\tilde{X}_\alpha$  with no relationship to the original  $X_\alpha$  whatsoever. Thus it is worthwhile to pursue a stronger formulation for nonmeasurable maps. The problem is to generalize the statement that every pair  $\tilde{X}_\alpha$  and  $X_\alpha$  are equal in law. In the initial formulation, equality in law will be interpreted in the sense that

$$\mathbb{E}^* f(\tilde{X}_\alpha) = \mathbb{E}^* f(X_\alpha), \quad \text{for every bounded } f: \mathbb{D} \mapsto \mathbb{R}.$$

In particular,  $P^*(\tilde{X}_\alpha \in B) = P^*(X_\alpha \in B)$  for every set  $B$ , and the same for inner probabilities. If  $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{P})$  is complete (which may be assumed without loss of generality), this implies that the laws of  $\tilde{X}_\alpha$  and  $X_\alpha$  are the same on the maximal  $\sigma$ -field  $\{B \subset \mathbb{D}: X_\alpha^{-1}(B) \in \mathcal{A}_\alpha\}$  for which they are defined, that is, for which  $X_\alpha$  is measurable (Problem 1.2.10). However, in general, equality of the outer expectations says a lot more than just equality in law.

The following nonmeasurable representation theorem holds for sequences but not nets in general. Call a directed set *nontrivial* if it permits a net of strictly positive numbers  $\delta_\alpha$  with  $\delta_\alpha \rightarrow 0$ . Of course, the set of natural numbers with the usual ordering is nontrivial.<sup>†</sup>

---

<sup>†</sup> There do exist directed sets that are trivial; for each of them there is a net  $X_\alpha$  for which the theorem fails (Problem 1.10.7). On the other hand, for such a directed set  $X_\alpha \rightsquigarrow X_\infty$  with

**1.10.4 Theorem (a.s. representations).** Let  $X_\alpha: \Omega_\alpha \mapsto \mathbb{D}$  be an arbitrary net indexed by a nontrivial directed set, and let  $X_\infty$  be Borel measurable and separable. If  $X_\alpha \rightsquigarrow X_\infty$ , then there exists a probability space  $(\tilde{\Omega}, \tilde{\mathcal{A}}, \tilde{P})$  and maps  $\tilde{X}_\alpha: \tilde{\Omega} \mapsto \mathbb{D}$  with

- (i)  $\tilde{X}_\alpha \xrightarrow{\text{au}} \tilde{X}_\infty$ ;
- (ii)  $E^* f(\tilde{X}_\alpha) = E^* f(X_\alpha)$ , for every bounded  $f: \mathbb{D} \mapsto \mathbb{R}$  and every  $\alpha$ .

**1.10.5 Addendum.** In addition to (i) and (ii),  $\tilde{X}_\alpha$  can be chosen according to the following diagram:

$$\begin{array}{ccc} \Omega_\alpha & \xrightarrow{X_\alpha} & \mathbb{D} \\ \phi_\alpha \uparrow & \nearrow & \tilde{X}_\alpha = X_\alpha \circ \phi_\alpha \\ \tilde{\Omega} & & \end{array}$$

with the maps  $\phi_\alpha$  measurable and perfect, and  $P_\alpha = \tilde{P} \circ \phi_\alpha^{-1}$ .

The perfectness of  $\phi_\alpha$  asserted by the addendum improves part (ii) of the theorem. However, the main interest of the addendum is the implication that every  $\tilde{X}_\alpha$  can be forced to take no other values than the original  $X_\alpha$ . Thus  $\tilde{X}_\alpha$  and  $X_\alpha$  also share other properties, which may not be directly expressible in terms of their “laws.”

**1.10.6 Example.** Consider the empirical process  $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$  indexed by a set of functions  $\mathcal{H}$  on a measurable space  $(\mathcal{X}, \mathcal{B})$ . Under suitable conditions this converges weakly when seen as maps into  $\ell^\infty(\mathcal{H})$ . Of course, every realization  $\mathbb{G}_n(\omega)$  can also be seen as a signed measure on  $(\mathcal{X}, \mathcal{B})$ . According to the addendum, it is possible to construct the representations  $\tilde{\mathbb{G}}_n$  as maps into  $\ell^\infty(\mathcal{H})$  in such a way that the double interpretation — as a bounded function and a signed measure — is retained.

**Proofs.** If the directed set is trivial, then  $X_\alpha \rightsquigarrow X_\infty$  for a separable  $X$  can happen only if  $X_\alpha$  is Borel measurable (for the completion of the underlying probability space) and equal in law to  $X_\infty$  eventually (Problem 1.10.6). In that case, the measurable version of the theorem is trivial. For nontrivial directed sets, the nonmeasurable version implies the measurable version. For  $\tilde{X}_\alpha$  as in the addendum,

$$E^* f(\tilde{X}_\alpha) = \int^* f \circ X_\alpha \circ \phi_\alpha d\tilde{P} = \int^* f \circ X_\alpha d\tilde{P} \circ \phi_\alpha^{-1} = E^* f(X_\alpha),$$

for every bounded  $f$ , where perfectness of  $\phi_\alpha$  is used in the second equality. Thus it is enough to construct  $\tilde{X}_\alpha$  as in the addendum that also satisfy (i).

---

$X_\infty$  separable is only possible for  $X_\alpha$  eventually Borel measurable and equal in law to  $X_\infty$  (Problem 1.10.6). So only rather trivial cases have to be excluded from the present formulation of the almost sure representation theorem. The reason that these have to be excluded is that (ii) says things about the relation between  $\tilde{X}_\alpha$  and  $X_\alpha$  that go beyond their Borel laws.

The construction consists of five steps. Call a set  $B \subset \mathbb{D}$  a *continuity set* if  $P(X_\infty \in \delta B) = 0$ .

(i). For every  $\varepsilon > 0$ , there exists a partition of  $\mathbb{D}$  into finitely many disjoint continuity sets

$$B_0^{(\varepsilon)}, B_1^{(\varepsilon)}, \dots, B_{k_\varepsilon}^{(\varepsilon)},$$

with the properties  $P(X_\infty \in B_0^{(\varepsilon)}) < \varepsilon^2$  and  $\text{diam } B_i^{(\varepsilon)} < \varepsilon$  for  $i = 1, 2, \dots, k_\varepsilon$ .

Indeed, let the sequence  $s_1, s_2, \dots$  be dense in a set with probability 1 under  $X_\infty$ . Every open ball  $B(s_i, r)$  is a discontinuity set for at most countably many values of  $r$ . Take  $\varepsilon/3 < r_i < \varepsilon/2$  such that  $B(s_i, r_i)$  is a continuity set. For  $i \geq 1$  set  $B_i^{(\varepsilon)} = B(s_i, r_i) - \cup_{j < i} B(s_j, r_j)$ . Since the boundary of a finite intersection is contained in the union of the boundaries, every  $B_i^{(\varepsilon)}$  is a continuity set. Moreover, the union of all  $B_i^{(\varepsilon)}$  has probability 1 under  $X_\infty$ . Take  $k_\varepsilon$  sufficiently large, and set  $B_0^{(\varepsilon)} = \mathbb{D} - \cup_{i=1}^{k_\varepsilon} B_i^{(\varepsilon)}$ .

(ii). There exists a net  $\varepsilon_\alpha \rightarrow 0$  taking values  $1/m$  with  $m \in \mathbb{N}$  only, and

$$(1.10.7) \quad P_*(X_\alpha \in B_i^{(\varepsilon_\alpha)}) \geq (1 - \varepsilon_\alpha) P_\infty(X_\infty \in B_i^{(\varepsilon_\alpha)}), \quad i = 1, \dots, k_{\varepsilon_\alpha},$$

for all sufficiently large  $\alpha$ , say  $\alpha \geq \alpha_1$ .

Indeed, for every fixed  $\varepsilon > 0$  and every  $i$ , one has  $P_*(X_\alpha \in B_i^{(\varepsilon)}) \rightarrow P_\infty(X_\infty \in B_i^{(\varepsilon)})$ , by the portmanteau theorem. Let  $\alpha_m$  be such that, for  $\alpha \geq \alpha_m$  and  $i = 1, \dots, k_{1/m}$ ,

$$(1.10.8) \quad P_*(X_\alpha \in B_i^{(1/m)}) \geq (1 - \frac{1}{m}) P_\infty(X_\infty \in B_i^{(1/m)}).$$

Assume without loss of generality that  $\alpha_1 \leq \alpha_2 \leq \dots$ . Set

$$\eta_\alpha = \inf \left\{ \frac{1}{m} : \alpha \geq \alpha_m, m \in \mathbb{N} \right\},$$

where the infimum over the empty set is 2. For  $\alpha \geq \alpha_m$ , it holds that  $\eta_\alpha \leq 1/m$ ; hence  $\eta_\alpha \rightarrow 0$ . This net qualifies for  $\varepsilon_\alpha$  in case it is never zero. In general,  $\varepsilon_\alpha$  will be defined as  $\eta_\alpha$  kept away from zero. For a fixed  $\alpha \geq \alpha_1$ , either  $\eta_\alpha = 1/m$ , in which case  $\alpha \geq \alpha_m$ , or  $\eta_\alpha = 0$ , in which case  $\alpha \geq \alpha_m$  for every  $m$ . In the first case, (1.10.8) holds for  $1/m = \eta_\alpha$ ; in the second, (1.10.8) holds for every  $1/m$ ,  $m \in \mathbb{N}$ . Let  $\delta_\alpha$  be an arbitrary net with  $\delta_\alpha \rightarrow 0$  and taking values  $1/m$  with  $m \in \mathbb{N}$  only. Set  $\varepsilon_\alpha = \eta_\alpha$  if  $\eta_\alpha > 0$  and  $\varepsilon_\alpha = \delta_\alpha$  if  $\eta_\alpha = 0$ .

(iii). To simplify notation, assume that (1.5.5) holds for every  $\alpha$  and is positive (throw away the beginning of the net). For  $i = 1, \dots, k_{\varepsilon_\alpha}$ , let  $A_i^\alpha$  be a measurable set contained in  $\{X_\alpha \in B_i^{(\varepsilon_\alpha)}\}$  with the same (inner)

probability. Set  $A_0^\alpha = \Omega_\alpha - \cup_{i=1}^{k_{\varepsilon_\alpha}} A_i^\alpha$ . Define

$$\begin{aligned}\tilde{\Omega} &= \Omega_\infty \times \prod_\alpha \left[ \Omega_\alpha \times \prod_{i=0}^{k_{\varepsilon_\alpha}} A_i^\alpha \right] \times [0, 1], \\ \tilde{\mathcal{A}} &= \mathcal{A}_\infty \times \prod_\alpha \left[ \mathcal{A}_\alpha \times \prod_{i=0}^{k_{\varepsilon_\alpha}} \mathcal{A}_\alpha \cap A_i^\alpha \right] \times \mathcal{B}_o, \\ \tilde{P} &= P_\infty \times \prod_\alpha \left[ \mu_\alpha \times \prod_{i=0}^{k_{\varepsilon_\alpha}} P_\alpha(\cdot | A_i^\alpha) \right] \times \lambda.\end{aligned}$$

Here  $P_\alpha(\cdot | A_i^\alpha)$  is the conditional  $P_\alpha$ -measure given that  $A_i^\alpha$  and  $\mathcal{B}_o$  and  $\lambda$  are the Borel sets and Lebesgue measure on  $[0,1]$ , respectively. Furthermore,  $\mu_\alpha$  is the probability measure defined by

(1.10.9)

$$\mu_\alpha(A) = \varepsilon_\alpha^{-1} \sum_{i=0}^{k_{\varepsilon_\alpha}} P_\alpha(A | A_i^\alpha) \left[ P_\alpha(A_i^\alpha) - (1 - \varepsilon_\alpha) P_\infty(X_\infty \in B_i^{(\varepsilon_\alpha)}) \right].$$

Write an element  $\tilde{\omega}$  of  $\tilde{\Omega}$  as

$$\tilde{\omega} = (\omega_\infty, \dots, \omega_\alpha, \omega_{\alpha 0}, \omega_{\alpha 1}, \dots, \omega_{\alpha k_{\varepsilon_\alpha}}, \dots, \xi).$$

Define

$$\begin{aligned}\phi_\infty(\tilde{\omega}) &= \omega_\infty, \\ \phi_\alpha(\tilde{\omega}) &= \begin{cases} \omega_\alpha, & \text{if } \xi > 1 - \varepsilon_\alpha, \\ \omega_{\alpha i}, & \text{if } \xi \leq 1 - \varepsilon_\alpha \text{ and } X_\infty(\omega_\infty) \in B_i^{(\varepsilon_\alpha)}. \end{cases}\end{aligned}$$

Thus if  $X_\infty$  falls in  $B_i^{(\varepsilon_\alpha)}$ , “choose”  $\phi_\alpha$  with probability  $1 - \varepsilon_\alpha$  in  $A_i^\alpha$  (so that  $X_\alpha \circ \phi_\alpha \in B_i^{(\varepsilon_\alpha)}$ , close to  $X_\infty$  for  $i \geq 1$ ) according to the conditional distribution  $P_\alpha(\cdot | A_i^\alpha)$ . Next, the mass  $\varepsilon_\alpha$  is used to correct this “discretization” so as to ensure that  $\tilde{P} \circ \phi_\alpha^{-1} = P_\alpha$ . If  $\phi_\alpha$  is chosen with probability  $\varepsilon_\alpha$ , according to a law  $\mu_\alpha$ , then

$$\tilde{P}(\phi_\alpha \in A) = (1 - \varepsilon_\alpha) \sum_{i=0}^{k_{\varepsilon_\alpha}} \tilde{P}(\omega_{\alpha i} \in A \wedge X_\infty(\omega_\infty) \in B_i^{(\varepsilon_\alpha)}) + \varepsilon_\alpha \mu_\alpha(A).$$

Elementary algebra shows that to make this equal to  $P_\alpha(A)$ , the measure  $\mu_\alpha$  must be defined by (1.10.9). Because of (1.10.7), this is possible.

(iv). It is clear from the construction that  $d(X_\alpha, \tilde{X}_\infty) \leq \varepsilon_\alpha$ , for every  $\tilde{\omega}$ , with  $X_\infty(\omega_\infty) \notin B_0^{(\varepsilon_\alpha)}$  and  $\xi \leq 1 - \varepsilon_\alpha$ . Set  $A_k = \cup_{m \geq k} \{\tilde{\omega}: X_\infty(\omega_\infty) \in B_0^{(1/m)} \text{ or } \xi > 1 - 1/k\}$ . For every  $\varepsilon > 0$ , there exists a  $k$  such that  $\tilde{P}(A_k) \leq \varepsilon$ . For  $\tilde{\omega} \in \tilde{\Omega} - A_k$  and  $\alpha$  such that  $\varepsilon_\alpha \leq 1/k$ , it holds that  $d(\tilde{X}_\alpha, \tilde{X}_\infty) \leq \varepsilon_\alpha$ . Thus  $\tilde{X}_\alpha$  converges to  $\tilde{X}_\infty$  almost uniformly.

(v). Let  $T: \Omega_\alpha \mapsto \mathbb{R}$  be bounded, and let  $T^*$  be its minimal measurable cover for  $P_\alpha$ . Write

$$T \circ \phi_\alpha = 1_{\pi_\xi \leq 1-\varepsilon_\alpha} \sum_{i=0}^{k_{\varepsilon_\alpha}} T|_{A_i^\alpha} \circ \pi_{\alpha,i} 1_{X_\infty \circ \pi_\infty \in B_i^{(\varepsilon_\alpha)}} + 1_{\pi_\xi > 1-\varepsilon_\alpha} T \circ \pi_\alpha,$$

where  $\pi_\xi: \tilde{\Omega} \mapsto [0, 1]$ ,  $\pi_{\alpha,i}: \tilde{\Omega} \mapsto A_i^\alpha$ , and  $\pi_\alpha: \tilde{\Omega} \mapsto \Omega_\alpha$  are the coordinate projections. Then the minimal measurable cover of  $T \circ \phi_\alpha$  for  $\tilde{P}$  can be computed as

$$\begin{aligned} (T \circ \phi_\alpha)^* &= 1_{\pi_\xi \leq 1-\varepsilon_\alpha} \sum_{i=0}^{k_{\varepsilon_\alpha}} (T|_{A_i^\alpha} \circ \pi_{\alpha,i})^* 1_{X_\infty \circ \pi_\infty \in B_i^{(\varepsilon_\alpha)}} \\ &\quad + 1_{\pi_\xi > 1-\varepsilon_\alpha} (T \circ \pi_\alpha)^* \\ &= 1_{\pi_\xi \leq 1-\varepsilon_\alpha} \sum_{i=0}^{k_{\varepsilon_\alpha}} (T|_{A_i^\alpha})^{*P(\cdot | A_i^\alpha)} \circ \pi_{\alpha,i} 1_{X_\infty \circ \pi_\infty \in B_i^{(\varepsilon_\alpha)}} \\ &\quad + 1_{\pi_\xi > 1-\varepsilon_\alpha} T^{*\mu_\alpha} \circ \pi_\alpha, \end{aligned}$$

since coordinate projections are perfect. Because  $P_\alpha$  and  $P(\cdot | A_i^\alpha)$  or  $\mu_\alpha$  are equivalent on the spaces where they are defined, the measurable covers in the last formula can just as well be computed under  $P_\alpha$ , whence  $(T \circ \phi_\alpha)^* = T^* \circ \phi_\alpha$ . ■

A typical use of the almost sure representation theorem is to extend convergence theorems for expectations of almost surely convergent nets to weakly convergent nets.

**1.10.10 Example.** Let  $f, g: \mathbb{D} \mapsto \mathbb{R}$  be continuous and satisfy  $|f| \leq g$ . If  $X_n \rightsquigarrow X$  and  $E^*g(X_n) \rightarrow Eg(X) < \infty$ , then  $E^*f(X_n) \rightarrow Ef(X)$ .

To see this, let  $\tilde{X}_n \xrightarrow{\text{as*}} \tilde{X}$  be almost sure representations. By the continuous mapping theorem for outer almost sure convergence,  $f(\tilde{X}_n) \xrightarrow{\text{as*}} f(\tilde{X})$  and  $g(\tilde{X}_n) \xrightarrow{\text{as*}} g(\tilde{X})$ . The latter gives  $g(\tilde{X}_n)^* \xrightarrow{\text{as*}} g(\tilde{X})$ . Combined with convergence of expectations,  $Eg(\tilde{X}_n)^* \rightarrow Eg(\tilde{X})$ , it yields  $E|g(\tilde{X}_n)^* - g(\tilde{X})| \rightarrow 0$ , by a well-known convergence lemma. Thus the sequence  $g(\tilde{X}_n)^*$  is uniformly integrable. Since  $0 \leq |f(\tilde{X}_n)|^* \leq g(\tilde{X}_n)^*$ , the same is true for the sequence  $f(\tilde{X}_n)^*$ , whence  $Ef(\tilde{X}_n)^* \rightarrow Ef(\tilde{X})$ .

**1.10.11 Example.** A particular case of the previous example is the following. Let  $L_n$  be a sequence of Borel measures. For a continuous positive function  $g$ , consider the measures  $M_n(B) = \int_B g dL_n$ . If  $L_n \rightsquigarrow L_\infty$  and  $\int g dL_n \rightarrow \int g dL_\infty$ , then  $M_n \rightsquigarrow M_\infty$ . A closer look reveals that  $g$  need be continuous almost everywhere under the limit  $L_\infty$  only.

Certain results are easier to prove for measurable maps. The following proposition can often be used to turn a result for measurable maps into a general one.

**1.10.12 Proposition.** Let  $X_\alpha: \Omega_\alpha \mapsto \mathbb{D}$  be arbitrary. If  $X_\alpha \rightsquigarrow X$  and  $X$  is separable, then there exist Borel measurable  $Y_\alpha: \Omega_\alpha \mapsto \mathbb{D}$  with  $d(X_\alpha, Y_\alpha) \xrightarrow{P^*} 0$ .

**Proof.** As explained in the proof of the almost sure representation theorem, there exist for every  $\varepsilon > 0$  finitely many  $X$ -continuity sets  $B_0^{(\varepsilon)}, B_1^{(\varepsilon)}, \dots, B_{k_\varepsilon}^{(\varepsilon)}$  such that  $P(X \in B_0^{(\varepsilon)}) < \varepsilon$  and  $\text{diam } B_i^{(\varepsilon)} < \varepsilon$  for  $i = 1, 2, \dots, k_\varepsilon$ . Choose a point  $x_i^{(\varepsilon)}$  from every  $B_i^{(\varepsilon)}$ , and define

$$Y_\alpha^{(\varepsilon)} = \begin{cases} x_i^{(\varepsilon)}, & \text{on } \{X_\alpha \in B_i^{(\varepsilon)}\}_*, \quad i = 1, \dots, k_\varepsilon, \\ x_0^{(\varepsilon)}, & \text{on } \Omega_\alpha - \cup_{i=1}^{k_\varepsilon} \{X_\alpha \in B_i^{(\varepsilon)}\}_*. \end{cases}$$

Then each  $Y_\alpha^{(\varepsilon)}$  is measurable, and

$$\begin{aligned} P^*(d(Y_\alpha^{(\varepsilon)}, X_\alpha) > \varepsilon) &\leq \sum_{i=1}^{k_\varepsilon} P(\{X_\alpha \in B_i^{(\varepsilon)}\}^* - \{X_\alpha \in B_i^{(\varepsilon)}\}_*) \\ &\quad + P^*(X_\alpha \in B_0^{(\varepsilon)}) \rightarrow P(X \in B_0^{(\varepsilon)}) < \varepsilon. \end{aligned}$$

Assume for the moment that there exists a net  $\delta_\alpha \rightarrow 0$  taking values  $1/m$ ,  $m \in \mathbb{N}$  only. Let  $R(\alpha, \varepsilon)$  be the probability on the left side in the last displayed equation. For  $m \in \mathbb{N}$ , there exists  $\alpha_m$  such that, for all  $\alpha \geq \alpha_m$ ,

$$R\left(\alpha, \frac{1}{m}\right) < \frac{2}{m}.$$

Choose the  $\alpha_m$  increasing:  $\alpha_1 \leq \alpha_2 \leq \dots$ . Set

$$\eta_\alpha = \inf\left\{\frac{1}{m}: \alpha \geq \alpha_m\right\},$$

where the infimum over the empty set is 2. Then  $\eta_\alpha \rightarrow 0$ . Moreover, for  $\alpha \geq \alpha_1$ , either  $\eta_\alpha = 1/m$  for some  $m$ , in which case  $\alpha \geq \alpha_m$ , or  $\eta_\alpha = 0$ . In the first case,  $R(\alpha, 1/m) < 2/m$ ; while in the second case,  $R(\alpha, 1/m) < 2/m$ , for every  $m$ . Define  $\varepsilon_\alpha$  to be  $\eta_\alpha$  if  $\eta_\alpha > 0$  and  $\delta_\alpha$  if  $\eta_\alpha = 0$ . Then  $R(\alpha, \varepsilon_\alpha) < 2\varepsilon_\alpha$ , for every  $\alpha$ . Consequently,  $d(Y_\alpha^{(\varepsilon_\alpha)}, X_\alpha) \xrightarrow{P^*} 0$ .

Finally, if there does not exist a net  $\delta_\alpha$  as assumed, then necessarily  $X_\alpha$  is Borel measurable for the completion of  $(\Omega_\alpha, \mathcal{A}_\alpha, P_\alpha)$  and is equal in law to  $X$  for all sufficiently large  $\alpha$ . By a standard argument, there exists a Borel measurable map  $Y_\alpha: (\Omega_\alpha, \mathcal{A}_\alpha) \mapsto \mathbb{D}$  with  $P_*(X_\alpha = Y_\alpha) = 1$  for such  $\alpha$  (Problem 1.2.10). ■

## Problems and Complements

1. Let  $X_n$  and  $X$  be maps into  $\mathbb{R}$  with  $X$  Borel measurable.
  - (i)  $X_n \xrightarrow{\text{as*}} X$  if and only if  $X_n^* \xrightarrow{\text{as}} X$  and  $X_{n*} \xrightarrow{\text{as}} X$ .
  - (ii)  $X_n \rightsquigarrow X$  if and only if  $X_n^* \rightsquigarrow X$  and  $X_{n*} \rightsquigarrow X$ .

[Hint: For (i) use  $|X_n - X|^* = |X_n^* - X| \vee |X_{n*} - X|$ . For (ii) suppose  $X_n \rightsquigarrow X$ . Let  $\tilde{X}_n = X_n \circ \phi_n \xrightarrow{\text{as*}} \tilde{X}$  be almost sure representations. By (i)  $\tilde{X}_n^* \xrightarrow{\text{as}} \tilde{X}$ , whence  $\tilde{X}_n^* \rightsquigarrow \tilde{X}$ . For perfect  $\phi_n$ , one has  $\tilde{X}_n^* = X_n^* \circ \phi_n$ , and it follows that  $E_f(\tilde{X}_n^*) = E_f(X_n^*)$ , for every measurable  $f$ . So  $X_n^* \rightsquigarrow X$ . By considering the negatives of the functions, obtain  $X_{n*} \rightsquigarrow X$ . The converse follows from  $P(X_n^* \leq x) \leq P^*(X_n \leq x) \leq P(X_{n*} \leq x)$ , for every  $x$ .]
2. Let  $(\Omega_n, \mathcal{A}_n, P_n)$  be the product of  $n$  copies of a probability space  $(\Omega, \mathcal{A}, P)$ . Let  $X: \Omega \mapsto \mathbb{R}$  be a fixed map, let  $X_i = X$  for every  $i$ , and let  $S_n: \Omega_n \mapsto \mathbb{R}$  be defined by  $S_n(\omega_1, \dots, \omega_n) = \sum_{i=1}^n X_i(\omega_i)$ . Then the central limit theorem  $S_n/\sqrt{n} \rightsquigarrow N(0, 1)$  can hold only if  $X$  is measurable for the  $P$ -completion of  $\mathcal{A}$ . Similarly, the law of large numbers  $S_n/n \rightsquigarrow 0$  can hold only if  $X$  is  $P$ -completion measurable.
3. The following statements are *not* true:
  - (i)  $X_\alpha$  measurable and  $X_\alpha \xrightarrow{\text{as*}} 0$  imply  $X_\alpha \xrightarrow{P*} 0$  (for general nets);
  - (ii)  $X_\alpha \xrightarrow{\text{as*}} 0$  implies that  $X_\alpha$  is asymptotically measurable (for general nets);
  - (iii)  $X_n \xrightarrow{\text{as}} 0$  and  $X_n$  is asymptotically measurable imply  $X_n \xrightarrow{\text{as*}} X$ .

[Hint: Let  $A$  be the collection of all finite subsets of  $[0,1]$ , directed through inclusion:  $\alpha_1 \leq \alpha_2$  if and only if  $\alpha_1 \subset \alpha_2$ . Let  $P$  be Lebesgue measure on the Borel sets of  $[0,1]$ . For (i) define  $X_\alpha: [0,1] \mapsto [0,1]$  by  $X_\alpha = 1_{[0,1]-\alpha}$ . Then  $X_\alpha$  is Borel measurable, and  $X_\alpha(\omega) \rightarrow 0$  for every  $\omega$ . However,  $P(X_\alpha = 1) = 1$  for every  $\alpha$ , so  $X_\alpha$  does not converge to zero in probability; in fact,  $X_\alpha \xrightarrow{P*} 1$ . Note that the dominated convergence theorem fails too, since  $EX_\alpha = 1$  for every  $\alpha$ . For (ii) let  $B$  be a subset of  $[0,1]$  with  $P_*(B) < P^*(B)$ . Define  $X_\alpha = 1_{[0,1]-\alpha} 1_B$ . Then  $|X_\alpha|^* = 1_{[0,1]-\alpha}(1_B)^* \xrightarrow{\text{as*}} 0$ , while  $E(X_\alpha)^* - E(X_\alpha)_* = E(1_B)^* - E(1_B)_*$  does not converge to 0. Finally, for (iii), let  $I_n$  be the  $n$ th interval in the sequence  $[0,1], [0,1/2], [1/2,1], [0,1/4], \dots$ . For  $B_n$  as in Example 1.9.4, take  $X_n = 1_{I_n} 1_{B_n}$ . Then  $d(X_n, 0)^* = X_n^* = 1_{I_n}$  converges to zero in probability and weakly, but not almost surely.]
4. For  $\alpha$  running through a directed set, consider the statements
  - (i)  $X_\alpha \xrightarrow{\text{au}} X$ ;
  - (ii)  $\sup_{\beta \geq \alpha} d(X_\beta, X)^* \xrightarrow{P*} 0$ ;
  - (iii)  $X_\alpha \xrightarrow{\text{as*}} X$ ;
  - (iv)  $X_\alpha \xrightarrow{P*} X$ .

The following implications are always true:

(i)	$\longleftrightarrow$	(ii)
$\downarrow$	$\times$	$\downarrow$
(iii)		(iv)

There is a counterexample to every implication that is not indicated in the diagram.

5. There exist directed sets that cannot index a net of strictly positive numbers that converges to zero. Equivalently, there exist directed sets  $A$  such that, if  $\{\delta_\alpha : \alpha \in A\}$  is a net of real numbers with  $\delta_\alpha \rightarrow 0$ , then  $\delta_\alpha = 0$  eventually.

[Hint: A trivial example is obtained by taking  $A$  equal to a partially ordered set with a largest element; for instance,  $A = \{1, 2, \dots, n\}$  with usual order, or  $A = [0, 1]$  with reversed natural order. For a nontrivial example, take  $A$  equal to an uncountable, well-ordered set for which every section  $\{\alpha : \alpha < \alpha_0\}$  is countable. Such sets exist provided one assumes the axiom of choice (and the well-ordering theorem), though it is hard to give a description of one. (See Munkres (1975), page 66.) If  $\delta_\alpha$  is indexed by such a set and converges to zero without being zero eventually, then  $\varepsilon_\alpha = \sup_{\beta \geq \alpha} |\delta_\beta|$  is strictly positive and converges monotonically to zero. By the well-ordering, each set  $S_i = \{\alpha : \varepsilon_\alpha \leq 1/i\}$  has a smallest element  $\alpha_i$ . Since the net  $\varepsilon_\alpha$  is decreasing, it follows that  $S_i = \{\alpha : \alpha \geq \alpha_i\}$ . Since each  $\varepsilon_\alpha$  is positive, there is for each  $\alpha$  an  $i$  with  $\varepsilon_\alpha > 1/i$ , that is,  $\alpha < \alpha_i$ . Then  $A$  is covered by the sets  $\{\alpha < \alpha_i\}$ , but these are only countably many sets, each with countably many elements.]

6. Let  $X_\alpha$  be maps indexed by a directed set that is trivial defined on complete probability spaces.

- (i) If  $X_\alpha \rightsquigarrow X$  and  $X$  is separable, then  $X_\alpha$  is Borel measurable and equal in law to  $X$  eventually.  
(ii) If  $X_\alpha \xrightarrow{\text{au}} X$ , then  $P_*(X_\alpha = X) = 1$  eventually.

[Hint: For (i) use metrization of weak convergence to reduce  $X_\alpha \rightsquigarrow X$  to convergence of a set of numbers. For (ii) set  $\delta_\alpha = P^*(d(X_\alpha, X) > 0)$ . Then  $\delta_\alpha \rightarrow 0$ . Indeed, for every  $\varepsilon > 0$  there is a set  $B$  with  $\sup_{\omega \in B} d(X_\alpha, X) \rightarrow 0$  and  $P(B) > 1 - \varepsilon$ . Since the directed set is trivial, the supremum is zero for all sufficiently large  $\alpha$ , whence  $\delta_\alpha \leq \varepsilon$  eventually. Finally,  $\delta_\alpha \rightarrow 0$  and trivialness imply  $\delta_\alpha = 0$  eventually.]

7. Any directed set that is trivial indexes a weakly convergent net  $X_\alpha \rightsquigarrow X_\infty$  with values in  $\mathbb{R}$  for which there is no almost uniform representation: there is no net  $\tilde{X}_\alpha$  and a  $\tilde{X}_\infty$  with  $\tilde{X}_\alpha \xrightarrow{\text{au}} \tilde{X}_\infty$  and  $E^* f(X_\alpha) = E^* f(\tilde{X}_\alpha)$  for every  $\alpha$  and bounded  $f$ .

[Hint: Take a Borel measure on  $\mathbb{R}$  that can be extended in two different ways to a  $\sigma$ -field  $\mathbb{D}$  that is strictly larger than the Borel sets. (Every  $L$  for which there is a (nonmeasurable) set  $B$  with  $L^*(B) > L_*(B)$  does the job.) Let  $L_1$  and  $L_2$  be the extensions. Take  $(\Omega_\infty, \mathcal{A}_\infty, P_\infty) = (\mathbb{R}, \mathbb{D}, L_1)$  and  $(\Omega, \mathcal{A}_\alpha, P_\alpha) = (\mathbb{R}, \mathbb{D}, L_2)$  for every other  $\alpha$ ; let both  $X_\alpha$  and  $X_\infty$  be the identity map. The almost uniform representations would be  $\mathbb{D}$ -measurable maps into  $\mathbb{R}$ ;  $\tilde{X}_\infty$  would have law  $L_1$ , while the other  $\tilde{X}_\alpha$  would have law  $L_2$ . But according to Problem 1.10.6, also  $\tilde{X}_\infty = \tilde{X}_\alpha$  almost surely.]

8. Let  $X_n$  and  $X$  be Borel measurable elements in a metric space. For a fixed Borel set, define  $Y_n$  to be  $X_n$  if  $X \in B$  and to be  $X$  if  $X \notin B$ .

- (i) If  $X_n \xrightarrow{P} X$ , then  $Y_n \xrightarrow{P} X$ .  
(ii) If  $X_n \rightsquigarrow X$ , then not necessarily  $Y_n \rightsquigarrow X$ .

[**Hint:** For (ii) take  $X$  distributed as  $N(0, 1)$ ,  $X_n = -X$ , and  $B = [0, \infty)$ . Then  $Y_n$  is nonnegative with probability 1.]

# 1.11

## Refinements

The continuous mapping theorems for the three modes of stochastic convergence considered so far can be refined to cover maps  $g_n(X_n)$ , rather than  $g(X_n)$ , for a fixed  $g$ . Then the  $g_n$  should have a property that might be called *asymptotic equicontinuity* almost everywhere under the limit measure.

For simplicity, it will be assumed that the limit measure is separable, though this is not necessary for (iii) and can be replaced by other conditions for (i) and (ii) (Problem 1.11.1).

**1.11.1 Theorem (Extended continuous mapping).** Let  $\mathbb{D}_n \subset \mathbb{D}$  and  $g_n: \mathbb{D}_n \mapsto \mathbb{E}$  satisfy the following statements: if  $x_n \rightarrow x$  with  $x_n \in \mathbb{D}_n$  for every  $n$  and  $x \in \mathbb{D}_0$ , then  $g_n(x_n) \rightarrow g(x)$ , where  $\mathbb{D}_0 \subset \mathbb{D}$  and  $g: \mathbb{D}_0 \mapsto \mathbb{E}$ . Let  $X_n$  be maps with values in  $\mathbb{D}_n$ , let  $X$  be Borel measurable and separable, and take values in  $\mathbb{D}_0$ . Then

- (i)  $X_n \rightsquigarrow X$  implies that  $g_n(X_n) \rightsquigarrow g(X)$ ;
- (ii)  $X_n \xrightarrow{P_*} X$  implies that  $g_n(X_n) \xrightarrow{P_*} g(X)$ ;
- (iii)  $X_n \xrightarrow{\text{as*}} X$  implies that  $g_n(X_n) \xrightarrow{\text{as*}} g(X)$ .

**Proof.** Assume the weakest of the three assumptions: the one in (i) that  $X_n \rightsquigarrow X$ . Let  $\mathbb{D}_\infty$  be the set of all  $x$  for which there exists a sequence  $x_n$  with  $x_n \in \mathbb{D}_n$  and  $x_n \rightarrow x$ . First,  $P_*(X \in \mathbb{D}_\infty) = 1$ ; second, the restriction of  $g$  to  $\mathbb{D}_0 \cap \mathbb{D}_\infty$  is continuous; and third, if some subsequence satisfies  $x_{n'} \rightarrow x$  with  $x_{n'} \in \mathbb{D}_{n'}$  for every  $n'$  and  $x \in \mathbb{D}_0 \cap \mathbb{D}_\infty$ , then  $g_{n'}(x_{n'}) \rightarrow g(x)$ .

To see the first, invoke the almost sure representation theorem. If  $\tilde{X}_n \xrightarrow{\text{as*}} \tilde{X}$  are representing versions, then the range of  $\tilde{X}$  is contained in

$\mathbb{D}_\infty$  up to a null set, so  $P_*(X \in \mathbb{D}_\infty) = P_*(\tilde{X} \in \mathbb{D}_\infty) = 1$ . For the third, let the subsequence  $x_{n'}$  be given. Since its limit  $x$  is in  $\mathbb{D}_\infty$ , there is a sequence  $y_n \rightarrow x$  with  $y_n \in \mathbb{D}_n$  for every  $n$ . Fill out the subsequence  $x_{n'}$  to a whole sequence by putting  $x_n = y_n$  if  $n \notin \{n'\}$ . Then, by assumption,  $g_n(x_n) \rightarrow g(x)$  along the whole sequence, so also along the subsequence. To prove the second, let  $x_m \rightarrow x$  in  $\mathbb{D}_0 \cap \mathbb{D}_\infty$ . For every  $m$ , there is a sequence  $y_{m,n} \in \mathbb{D}_n$  with  $y_{m,n} \rightarrow x_m$  as  $n \rightarrow \infty$ . Since  $x_m \in \mathbb{D}_0$ , then also  $g_n(y_{m,n}) \rightarrow g(x_m)$ . For every  $m$ , take  $n_m$  such that both  $|y_{m,n_m} - x_m| < 1/m$  and  $|g_{n_m}(y_{m,n_m}) - g(x_m)| < 1/m$  and such that  $n_m$  is increasing with  $m$ . Then  $y_{m,n_m} \rightarrow x$ , so  $g_{n_m}(y_{m,n_m}) \rightarrow g(x)$ . This implies  $g(x_m) \rightarrow g(x)$ .

For simplicity of notation, now write  $\mathbb{D}_0$  for  $\mathbb{D}_0 \cap \mathbb{D}_\infty$ . The limit variable  $X$  may, without loss of generality, be assumed to take its values in  $\mathbb{D}_0$ . By the continuity property of  $g$ , the map  $g(X)$  is then Borel measurable.

(i). Let  $F$  be closed. Then

$$\bigcap_{n=1}^{\infty} \overline{\bigcup_{m=n}^{\infty} g_m^{-1}(F)} \subset g^{-1}(F) \cup (\mathbb{D} - \mathbb{D}_0).$$

Indeed, suppose  $x$  is in the left side. Then for every  $n$  there is  $m_n \geq n$  and  $x_{m_n} \in g_{m_n}^{-1}(F)$ , with  $d(x_{m_n}, x) < 1/n$ . Thus there exists  $x_{m'_n} \in \mathbb{D}_{m'_n}$ , with  $m'_n \rightarrow \infty$  and  $x_{m'_n} \rightarrow x$ . Then either  $g_{m'_n}(x_{m'_n}) \rightarrow g(x)$  or  $x \notin \mathbb{D}_0$ . Because  $F$  is closed, this implies  $g(x) \in F$  or  $x \notin \mathbb{D}_0$ .

Now, for every fixed  $k$ , by the portmanteau theorem,

$$\begin{aligned} \limsup P^*(g_n(X_n) \in F) &\leq \limsup P^*\left(X_n \in \overline{\bigcup_{m=k}^{\infty} g_m^{-1}(F)}\right) \\ &\leq P\left(X \in \overline{\bigcup_{m=k}^{\infty} g_m^{-1}(F)}\right). \end{aligned}$$

As  $k \rightarrow \infty$ , the last probability converges to  $P(X \in \bigcap_{k=1}^{\infty} \overline{\bigcup_{m=k}^{\infty} g_m^{-1}(F)})$ , which is smaller than or equal to  $P(g(X) \in F)$ .

(ii). Fix  $\varepsilon > 0$ . Choose  $\delta_n \downarrow 0$  with  $P^*(d(X_n, X) \geq \delta_n) \rightarrow 0$ . Let  $B_n$  be the set of all  $x$  for which there is  $y \in \mathbb{D}_n$ , with  $d(y, x) < \delta_n$  and  $e(g_n(y), g(x)) > \varepsilon$ . Suppose  $x \in B_n$  infinitely often. Then there is a sequence  $x_{n_m} \in \mathbb{D}_{n_m}$ , with  $x_{n_m} \rightarrow x$  and  $e(g_{n_m}(x_{n_m}), g(x)) > \varepsilon$ , for every  $m$ . So  $x \notin \mathbb{D}_0$ . Conclude that  $\limsup B_n \cap \mathbb{D}_0 = \emptyset$ . From the continuity of  $g$ , it is not hard to see that  $B_n \cap \mathbb{D}_0$  is relatively open in  $\mathbb{D}_0$  and hence relatively Borel. Consequently,  $P^*(X \in B_n) \rightarrow 0$  (Problem 1.2.18). Now

$$P^*(e(g_n(X_n), g(X)) > \varepsilon) \leq P^*(X \in B_n \text{ or } d(X_n, X) \geq \delta_n) \rightarrow 0.$$

(iii). It suffices to prove that  $\sup_{m \geq n} e(g_m(X_m), g(X)) \xrightarrow{P^*} 0$ . Choose  $\delta_n \downarrow 0$  such that  $P^*(\sup_{m \geq n} d(X_m, X) \geq \delta_n) \rightarrow 0$ . Define  $B_n$  as the set of all  $x$  such that there exists  $m \geq n$  and  $y \in \mathbb{D}_m$  with  $d(y, x) < \delta_n$  and  $e(g_m(y), g(x)) > \varepsilon$ . Finish the proof along the lines of the proof of (ii). ■

**1.11.2 Counterexample.** One might think that in part (iii) of the previous theorem it is enough that  $g_n(X_n(\omega)) \rightarrow g(X(\omega))$  for every  $\omega \in \Omega$ . This is not so. Take  $\Omega_n = [0, 1]$  with Borel sets and Lebesgue measure. Furthermore, set  $X_n(\omega) = X(\omega) = \omega$ , and let  $g_n(\omega) = 1_{B_n}(\omega)$  as in Example 1.9.4. Then  $g_n(X_n(\omega)) \rightarrow 0$  for every  $\omega$ , but  $|g_n(X_n)|^* = 1$  for every  $n$ , so that  $g_n(X_n)$  definitely does not converge to zero almost uniformly.

It is not difficult to find examples of weakly convergent nets  $X_\alpha \rightsquigarrow X$  and unbounded continuous functions  $f$  such that  $E^*f(X_\alpha)$  does not converge to  $Ef(X)$ . However, in many situations, such convergence for unbounded functions actually does hold true. Call a net of real-valued maps  $X_\alpha$  *asymptotically uniformly integrable* if

$$\lim_{M \rightarrow \infty} \limsup E^*|X_\alpha|\{|X_\alpha| > M\} = 0.$$

**1.11.3 Theorem.** Let  $f: \mathbb{D} \mapsto \mathbb{R}$  be continuous at every point in a set  $\mathbb{D}_0$ . Let  $X_\alpha \rightsquigarrow X$ , where  $X$  takes its values in  $\mathbb{D}_0$ .

- (i) If  $f(X_\alpha)$  is asymptotically uniformly integrable, then  $E^*f(X_\alpha) \rightarrow Ef(X)$ .
- (ii) If  $\limsup E^*|f(X_\alpha)| \leq E|f(X)| < \infty$ , then  $E^*f(X_\alpha) \rightarrow Ef(X)$ .

**Proof.** For  $M > 0$ , write  $f^M$  for  $f$  truncated to  $[-M, M]$ , that is,  $f^M(x) = (f(x) \vee (-M)) \wedge M$ . By the continuous mapping theorem,  $f^M(X_\alpha) \rightsquigarrow f^M(X)$  as maps into the interval  $[-M, M]$ , for every fixed  $M$ . Since the identity function is bounded and continuous on this interval, one has  $E^*f^M(X_\alpha) \rightarrow Ef^M(X)$ ; similarly for  $|f^M|$ .

First consider (ii). If  $y^M$  is an arbitrary real number  $y$  truncated to  $[-M, M]$ , then  $|y - y^M| = |y| - |y^M|$ . Consequently, for every  $M > 0$ ,

$$\begin{aligned} \limsup |E^*f(X_\alpha) - E^*f^M(X_\alpha)| &\leq \limsup E^*(|f(X_\alpha)| - |f^M(X_\alpha)|) \\ &\leq E|f(X)| - E|f^M(X)|. \end{aligned}$$

Fix  $M$  such that the right side is smaller than  $\varepsilon > 0$  and also  $|Ef^M(X) - Ef(X)| < \varepsilon$ . Then the net  $E^*f(X_\alpha)$  is eventually within distance  $\varepsilon$  of  $Ef^M(X)$ , and so within distance  $2\varepsilon$  of  $Ef(X)$ .

For (i) note first that  $|E^*|f|^M(X_\alpha) - E^*|f|(X_\alpha)|$  is bounded above by  $2E^*|f(X_\alpha)|\{|f(X_\alpha)| > M\}$ . Fix some  $\varepsilon > 0$ . For any  $M$  such that  $\limsup E^*|f(X_\alpha)|\{|f(X_\alpha)| > M\} < \varepsilon$ , the net  $E^*|f|(X_\alpha)$  is eventually within distance  $2\varepsilon$  of  $E|f|^M(X)$ . This happens for all sufficiently large  $M$ . By monotone convergence,  $E|f|^M(X) \rightarrow E|f|(X)$  as  $M \rightarrow \infty$ . So  $E^*|f|(X_\alpha) \rightarrow E|f|(X)$ . The asymptotic uniform integrability implies that the limit is finite. Finally, apply (ii) to obtain (i). ■

**1.11.4 Example.** If  $X_\alpha \rightsquigarrow X$  in  $\mathbb{R}$  and  $\limsup E^*|X_\alpha|^p < \infty$ , then  $E^*X_\alpha^q \rightarrow EX^q$  for every  $q < p$  for which the statement makes sense. It suffices to note that  $X_\alpha^q$  is asymptotically uniformly integrable, since  $\limsup E^*|X_\alpha|^q\{|X_\alpha| > M\} \leq M^{q-p} \limsup E^*|X_\alpha|^p < \infty$ .

**1.11.5 Example.** If  $X_\alpha \rightsquigarrow X$  in  $\mathbb{R}^k$  and every  $X_\alpha$  has a multivariate Gaussian distribution, then  $E\|X_\alpha\|^p \rightarrow E\|X\|^p$  for every  $p > 0$  and every norm on  $\mathbb{R}^k$ . The reason is that a Gaussian net can converge weakly only if the nets of means and covariances converge. Thus  $\limsup E^*\|X_\alpha\|^p < \infty$ , for every  $p > 0$ .

**1.11.6 Example.** If  $|X_\alpha| \leq Z_\alpha$  for every  $\alpha$  and  $Z_\alpha$  is asymptotically uniformly integrable, then  $X_\alpha$  is asymptotically uniformly integrable. In particular, the following dominated convergence theorem holds: *if  $X_\alpha \rightsquigarrow X$  in  $\mathbb{R}$  and  $|X_\alpha| \leq Z$  for every  $\alpha$  and some  $Z$  with  $E^*Z < \infty$ , then  $E^*X_\alpha \rightarrow EX$ .*

## Problems and Complements

1. Separability of  $X$  in the refined continuous mapping theorem 1.11.1 can be omitted from the conditions for (iii). If every  $\mathbb{D}_n$  contains  $\mathbb{D}_0$ , it is not necessary for (i) and (ii) either. Furthermore, suppose the condition on  $g_n: \mathbb{D}_n \mapsto \mathbb{E}$  is strengthened as follows: for every subsequence  $n'$ , if  $x_{n'} \rightarrow x$  with  $x_{n'} \in \mathbb{D}_{n'}$  and  $x \in \mathbb{D}_0$ , then  $g_{n'}(x_{n'}) \rightarrow g(x)$ . Then separability can be dropped for (ii) and replaced by the condition that  $g(X)$  is Borel measurable for (i).

[Hint: The proof as given uses separability only to ensure that the set  $\mathbb{D}_\infty$  has inner measure 1 under the limit. For (iii) this is always true. If  $\mathbb{D}_n$  contains  $\mathbb{D}_0$  for every  $n$ , then  $\mathbb{D}_\infty$  trivially contains  $\mathbb{D}_0$  and so always has inner measure 1. For (ii) the set  $\mathbb{D}_\infty$  can be replaced by the set of all  $x$  for which there is a subsequence  $x_{n'} \in \mathbb{D}_{n'}$  with  $x_{n'} \rightarrow x$ . Then the restriction of  $g$  to  $\mathbb{D}_\infty \cap \mathbb{D}_0$  is still continuous. Furthermore, since there is a subsequence such that  $X_{n'} \xrightarrow{\text{as*}} X$ , the inner measure of  $\mathbb{D}_\infty$  under  $X$  is 1 and the proof applies as it stands. Finally, the proof of (i) does not use separability, but only the conditions as given.]

## 1.12

# Uniformity and Metrization

In principle, weak convergence is the pointwise convergence of “operators”  $X_\alpha$  or  $L_\alpha$  on the space  $C_b(\mathbb{D})$ . However, there is automatically uniform convergence over certain subsets. These subsets can be fairly big: equicontinuity and boundedness suffice. On the other hand, there also exist small (countable) subsets such that pointwise convergence on such a subset is automatically uniform, and equivalent to pointwise convergence on the whole of  $C_b(\mathbb{D})$ , i.e. weak convergence. For separable  $\mathbb{D}$ , it is even possible to pick such a countable subset that works for every  $X_\alpha$  at the same time.

**1.12.1 Theorem.** *Let  $\mathcal{F} \subset C_b(\mathbb{D})$  be bounded and equicontinuous at every point  $x$  in a set  $\mathbb{D}_0$ . If  $X_\alpha \rightsquigarrow X$  where the limit  $X$  is separable and takes its values in  $\mathbb{D}_0$ , then  $E^*f(X_\alpha) \rightarrow Ef(X)$  uniformly in  $f \in \mathcal{F}$ .*

**Proof.** Add all Lipschitz functions with  $|f(x) - f(y)| \leq d(x, y)$  that are bounded by 1 to  $\mathcal{F}$ . Then  $\mathcal{F}$  is still bounded and equicontinuous on  $\mathbb{D}_0$ , and a new metric can be defined on  $\mathbb{D}$  through

$$e(x, y) = \sup_{f \in \mathcal{F}} |f(x) - f(y)|.$$

With respect to this metric, the class  $\mathcal{F}$  is uniformly equicontinuous on the whole of  $\mathbb{D}$ . Also, as a consequence of the original equicontinuity, if  $x_n \rightarrow x \in \mathbb{D}_0$  with respect to  $d$ , then  $x_n$  converges with respect to  $e$  to the same limit. Thus global continuity of  $f: \mathbb{D} \mapsto \mathbb{R}$  with respect to  $e$  implies continuity of  $f$  with respect to  $d$  at every  $x \in \mathbb{D}_0$ . If  $f$  is also bounded, then  $E^*f(X_\alpha) \rightarrow Ef(X)$  by the continuous mapping theorem. Thus  $X_\alpha \rightsquigarrow X$

also for  $e$ . Conclude that it is no loss of generality to assume that the original set  $\mathcal{F}$  is uniformly equicontinuous on the whole of  $\mathbb{D}$ .

Viewed as maps into the completion  $\bar{\mathbb{D}}$  of  $\mathbb{D}$ , the  $X_\alpha$  satisfy  $X_\alpha \rightsquigarrow X$  and form an asymptotically tight net. Fix  $\varepsilon > 0$ . Let  $K \subset \bar{\mathbb{D}}$  be a compact set with  $\liminf P_*(X_\alpha \in K^\delta) \geq 1 - \varepsilon$  for every  $\delta > 0$ . Since it is uniformly continuous, every element of  $\mathcal{F}$  can be extended to an element of  $C_b(\bar{\mathbb{D}})$ . The set of restrictions  $\mathcal{F}_K$  to  $K$  is totally bounded in  $C_b(K)$  by the Arzelà-Ascoli theorem. Take finitely many balls of radius  $\varepsilon > 0$  in  $C_b(K)$  that cover and intersect  $\mathcal{F}_K$ . Let  $f_1, \dots, f_m$  be the centers of the balls. By Tietze's theorem<sup>†</sup>, each  $f_i$  can be extended to an element of  $C_b(\bar{\mathbb{D}})$  of the same norm. Abusing notation, call the extensions  $f_1, \dots, f_m$  too. For sufficiently large  $\alpha$ , it is certainly true that  $|E^* f_i(X_\alpha) - Ef_i(X)| < \varepsilon$  for every  $i$ . For any  $f \in \mathcal{F}$ , there is an  $f_i$  and  $\delta > 0$  with  $|f(x) - f_i(x)| < \varepsilon$  for all  $x \in K^\delta$ . Then for sufficiently large  $\alpha$ , one has  $|E^* f(X_\alpha) - Ef(X)| < 3\varepsilon + 4\varepsilon M$  if  $M - \varepsilon$  is bigger than the uniform bound on  $\mathcal{F}$ . ■

**1.12.2 Theorem.** *For every separable subset  $\mathbb{D}_0 \subset \mathbb{D}$ , there is a countable, uniformly equicontinuous, and bounded collection  $\mathcal{F} \subset C_b(\mathbb{D})$  such that the following statements are equivalent for every Borel measurable  $X$  with  $P(X \in \mathbb{D}_0) = 1$ :*

- (i)  $X_\alpha \rightsquigarrow X$ ;
- (ii)  $E^* f(X_\alpha) \rightarrow Ef(X)$  for every  $f \in \mathcal{F}$ ;
- (iii)  $E^* f(X_\alpha) \rightarrow Ef(X)$  uniformly in  $f \in \mathcal{F}$ .

**1.12.3 Addendum.** *The collection  $\mathcal{F}$  can be taken equal to the set of functions of the form*

$$f(x) = q \left(1 - p_1 d(x, s_1)\right)^+ \vee \cdots \vee \left(1 - p_k d(x, s_k)\right)^+,$$

where  $k \in \mathbb{N}$ ;  $s_1, \dots, s_k$  range over a countable, dense subset of  $\mathbb{D}_0$ ;  $q \in \mathbb{Q}$ ; and each  $p_i \in \mathbb{Q}^+$ , with  $|q| \leq 1$  and  $|q|(p_1 \vee \cdots \vee p_k) \leq 1$ .

**Proof.** The collection  $\mathcal{F}$  as suggested in the addendum is uniformly bounded, and every  $f \in \mathcal{F}$  is Lipschitz continuous with Lipschitz constant smaller than 1. By Theorem 1.12.1, statement (i) implies (iii), which trivially implies (ii). Assume (ii), and let  $G$  be open. Let  $S$  be a countable, dense subset of  $\mathbb{D}_0$ . Then

$$1_G(y) = \sup \left\{ f(y) : 0 \leq f \leq 1_G, f(x) = (1 - p d(x, s))^+, p \in \mathbb{Q}^+, s \in S \right\},$$

for every  $y \in \mathbb{D}_0$ . Order the countably many functions in this supremum in a sequence, and let  $f_m$  be the maximum of the first  $m$  functions. Then  $0 \leq f_m \uparrow 1_G$  on  $\mathbb{D}_0$ . Now, for fixed  $m$ , it holds that  $\liminf P_*(X_\alpha \in G) \geq \liminf E_* f_m(X_\alpha) = Ef_m(X)$ , because every  $f_m$  is a multiple of a function in  $\mathcal{F}$ . Letting  $m \rightarrow \infty$  completes the proof. ■

---

<sup>†</sup> Jameson (1974), Theorem 12.4.

A consequence of the previous results is that weak convergence to separable limits is “metrizable.” Let  $\text{BL}_1$  be the set of all real functions on  $\mathbb{D}$  with a Lipschitz norm bounded by 1: for instance, all  $f$  with  $\|f\|_\infty \leq 1$  and  $|f(x) - f(y)| \leq d(x, y)$ , for every  $x, y$ . Then  $X_\alpha \rightsquigarrow X$ , where  $X$  is Borel measurable and separable if and only if

$$\sup_{f \in \text{BL}_1} |\mathbf{E}^* f(X_\alpha) - \mathbf{E} f(X)| \rightarrow 0.$$

In particular, weak convergence of separable Borel measures on a metric space is metrizable; for instance, by the metric

$$d_{\text{BL}}(L_1, L_2) = \sup_{f \in \text{BL}_1} \left| \int f \, dL_1 - \int f \, dL_2 \right|.$$

This is called the *bounded Lipschitz metric*.<sup>b</sup>

**1.12.4 Theorem.** Weak convergence of separable Borel probability measures on a metric space  $\mathbb{D}$  corresponds to a topology that is metrizable by the bounded Lipschitz metric. The set of all separable Borel probability measures is complete under this metric if and only if  $\mathbb{D}$  is complete. This set is separable for the weak topology if and only if  $\mathbb{D}$  is separable.

**Proof.** A Cauchy sequence is always totally bounded. A set of separable Borel measures on a complete metric space that is totally bounded for the bounded Lipschitz metric is relatively compact for the weak topology (Problem 1.12.1). This means that every sequence has a converging subsequence. A Cauchy sequence with a converging subsequence converges itself. ■

## Problems and Complements

1. **(Uniform tightness and weak compactness of a set of Borel measures)** Call a set  $\mathcal{L}$  of Borel probability measures on a metric space *uniformly tight* if, for every  $\varepsilon > 0$ , there exists a compact  $K$  with  $L(K) \geq 1 - \varepsilon$  for every  $L \in \mathcal{L}$ . The following statements are equivalent for a collection of separable Borel measures on a complete metric space:

- (i)  $\mathcal{L}$  is totally bounded for the bounded Lipschitz metric;
- (ii)  $\mathcal{L}$  is uniformly tight;
- (iii)  $\mathcal{L}$  is relatively compact.

In (iii) *relatively compact* may be taken to mean either that every net in  $\mathcal{L}$  has a weakly convergent subnet or that every sequence in  $\mathcal{L}$  has a weakly convergent subsequence. Thus  $\mathcal{L}$  is compact for the weak topology if and only it is uniformly tight and contains all its limit points.

<sup>b</sup> The class  $\text{BL}_1$  is the unit ball in the space  $\text{BL}(\mathbb{D})$  of bounded Lipschitz functions on  $\mathbb{D}$  if this is equipped with the norm  $\|f\|_{\text{BL}}$  equal to the maximum of  $\|f\|_\infty$  and  $\inf\{c: |f(x) - f(y)| \leq cd(x, y) \text{ for every } x, y\}$ . Taking a sum instead of a maximum in the definition of  $\|f\|_{\text{BL}}$  leads to a slightly different bounded Lipschitz metric on the set of measures.

[**Hint:** Prohorov's theorem implies that the second statement implies the third. Relative compactness implies total boundedness for any semimetric space. It suffices to show that (i) implies (ii). For any Borel set  $B$ , the function  $f(x) = (1 - \varepsilon^{-1}d(x, B))^+$  has Lipschitz constant  $\varepsilon^{-1}$  and is sandwiched between  $1_B$  and  $1_{B^\varepsilon}$ . Deduce that, for any probability measures  $L$  and  $L'$ ,  $L'(B^\varepsilon) \geq L(B) - \varepsilon^{-1}d_{BL}(L, L')$ . Fix  $\delta > 0$ . Take a finite set  $\mathcal{L}_0 \subset \mathcal{L}$  of measures with  $\mathcal{L} \subset \mathcal{L}_0^{\delta^2}$ . For every  $L_0 \in \mathcal{L}_0$ , take a compact with probability at least  $1 - \delta$ . Let  $F$  be the union of the compacts, and take a finite set  $G$  with  $F \subset G^\delta$ . Then, for every  $L \in \mathcal{L}$ ,  $L(G^{2\delta}) \geq L_0(G^\delta) - \delta^{-1}d_{BL}(L, L_0)$ , which is at least  $1 - 2\delta$  for some  $L_0$ . Take  $K$  equal to the closure of  $\cap_{m=1}^\infty G^{\varepsilon^{2^{-m}}}$ . Then  $K$  is totally bounded and complete, hence compact, with  $L(K) \geq 1 - \varepsilon$ .]

2. (**Completeness of the bounded Lipschitz metric**) The set of all separable Borel probability measures on a complete metric space is complete for the bounded Lipschitz metric.

[**Hint:** A Cauchy sequence is always totally bounded. According to Problem 1.12.1, a totally bounded set for the bounded Lipschitz metric is relatively compact. A Cauchy sequence with a converging subsequence converges itself.]

3. A collection of separable Borel measures  $\mathcal{L}$  on a complete metric space is compact for the weak topology if and only if it is uniformly tight and contains the limit of every converging sequence in  $\mathcal{L}$ .
4. (**A converse of Prohorov's theorem**) Let the net  $X_\alpha$  be relatively compact with all limit points concentrating on a fixed Polish subset  $\mathbb{D}_0 \subset \mathbb{D}$ . Then  $X_\alpha$  is asymptotically tight.

[**Hint:** Use “metrizability” of weak convergence to a separable limit to see that the set of all limit points is compact for the weak topology. (The proof is the same as for the proposition that the set of limit points of a relatively compact net in a metric space is compact.) Next use Problem 1.12.1.]

5. (**Lipschitz functions**) For a Lipschitz function  $f: \mathbb{D} \mapsto \mathbb{R}$ , define  $\|f\|_l = \sup\{|f(x) - f(y)|/d(x, y) : d(x, y) \neq 0\}$ . This is equal to the smallest  $c$  for which  $|f(x) - f(y)| \leq c d(x, y)$ , for every  $x, y$ . The expressions  $\|f\|_\infty + \|f\|_l$  and  $\|f\|_\infty \vee \|f\|_l$  define norms on the set of all Lipschitz functions on  $\mathbb{D}$ . They are equivalent and make the Lipschitz functions into a Banach space.

# 1

## Notes

**1.2.** Outer integrals and measurability are well known in probability theory. The minimal measurable cover is also the essential infimum of the collection of functions in the definition of  $E^*T$ ; see, e.g., Chow and Teicher (1978), page 190. Measurable cover or envelope functions were studied by Blumberg (1935) and then again independently by Eames and May (1967). Many parts of Lemma 1.2.2 are contained in May (1973), Dudley and Philipp (1983), Dudley (1984), and Dudley (1985). Perfect functions were introduced by Hoffmann-Jørgensen (1984, 1991).

**1.3.** Convergence in distribution of random variables in Euclidean spaces is an old concept. The study of weak convergence on abstract spaces started in the late 1940s and early 1950s with the study of the empirical process and the partial sum process by Doob (1949) and Donsker (1951, 1952). Prohorov (1956), Le Cam (1957), and Skorokhod (1956) gave a general theory for separable metric spaces. Billingsley (1968) and Parthasarathy (1967) have become standard references. The measurability problems were first ignored and then were solved through the introduction of “Skorokhod’s” topology on  $D[0, 1]$ . The more elegant solution to work with measurability in the ball  $\sigma$ -field was introduced by Dudley (1966). His stronger version of Prohorov’s theorem and characterization of tightness for nonseparable spaces was crucial for a more general development of weak convergence of the empirical process. Apparently, it went largely unnoticed until Gaenssler (1983) and Pollard (1984) showed the importance of the theory.

The more general approach to drop measurability altogether and work

with outer integrals, which we follow here, is due to Hoffmann-Jørgensen (1984), Chapter 7; see Hoffmann-Jørgensen (1991). Hoffmann-Jørgensen's approach was developed further in a series of papers by Andersen (1985) and Andersen and Dobrić (1987, 1988). We have restricted our treatment to the weak convergence of measures on metric spaces. For various aspects of weak convergence theory on general topological spaces (which is the framework frequently chosen by physicists and probabilists studying interacting particle systems), see Le Cam (1957), Varadarajan (1961), Smolyanov and Fomin (1976), and Mitoma (1983).

**1.4.** The results in this section are classical in the case of measurable random elements, but were apparently first proved in the nonmeasurable setting by Van der Vaart and Wellner (1989).

**1.5.** Theorem 1.5.7 was given by Andersen and Dobrić (1987); see their Theorem 2.12, page 167. They built on results of Dudley (1984, 1985) and Hoffmann-Jørgensen (1984). Lemma 1.5.9 originated with Dudley (1966, 1973). The first published proof is apparently that of Andersen and Dobrić (1987), Theorem 3.2, page 169.

**1.6.** For the function spaces  $C[0, \infty)$  and  $D[0, \infty)$ , results similar to those in this section were obtained by Stone (1963), Whitt (1970), Lindvall (1973), and Kim and Pollard (1990).

**1.7.** The failure of measurability of the uniform empirical process as a process in  $D[0, 1]$  with the uniform metric and Borel  $\sigma$ -field (Problem 1.7.3) was pointed out by Chibisov (1965); see Billingsley (1968), Section 18, pages 150–153, for a discussion. Dudley (1966, 1967a) initiated the study of weak convergence based on the ball  $\sigma$ -field. This theory was studied further by Wichura (1968), Gaenssler (1983), and Pollard (1984). Dudley (1978a, 1984) applied the theory of analytic sets to establish measurability properties in empirical process theory and introduced the concept of a class of functions that is “image-admissible-Suslin via a Suslin measurable space.” For a general introduction to the theory of analytic sets, including a proof of the projection theorem and references to the literature, see Cohn (1980), Chapter 8. For a thorough treatment, see also Hoffmann-Jørgensen (1970).

**1.8.** The basic result of this section goes back at least to Prohorov (1956); see his Theorem 1.13, page 171. For related results and further development, see Parthasarathy (1967), Chapter 6. For an application to tests of independence for directional data, see Jupp and Spurr (1985).

The characterization of tightness as approximate finite-dimensionality plus boundedness can be extended to general Banach spaces. See, for instance, Araujo and Giné (1980).

**1.9.** This section, including most of Lemma 1.9.3 and the Counterexample 1.9.4, is based on Dudley (1985). The continuous mapping theorem 1.9.5 extends the corresponding theorem in the classical theory. For a study of continuous functions for limit theorems in  $C[0, 1]$ ,  $D[0, 1]$ ,  $C[0, \infty)$ , and  $D[0, \infty)$ , see Whitt (1980). Theorem 1.9.6 is new.

**1.10.** A precursor of the “Skorokhod construction” theorem 1.10.3 is due to Hammersley (1952). Hammersley showed that if a sequence of random variables converges in distribution, then there is a construction of a sequence of random variables and a limit random variable with the same distributions on a common probability space for which convergence in probability holds. Constructions of almost surely convergent versions of processes that converge weakly apparently began with Skorokhod (1956) for processes with values in a complete separable metric space. Dudley (1968) removed completeness as a hypothesis, and Wichura (1970) proved a Skorokhod type result without separability. The first constructions were based on embedding in Brownian motion. The potential uses of such constructions in statistics were made clear by Pyke (1969, 1970). Billingsley (1971) gave a nice proof of Skorokhod’s theorem for complete and separable spaces. Theorem 1.10.4 extends these results within the framework of the Hoffmann-Jørgensen weak convergence theory, with almost sure convergence replaced by outer almost sure convergence. For sequences, the theorem is due to Dudley (1985), who shows that  $\tilde{\Omega}$  can be taken equal to the product of the  $\Omega_\alpha$  and one copy of  $[0, 1]$ , which is more economical than the space that is constructed here. Proposition 1.10.12 is due to Le Cam (1989). Exercise 1.10.2 is from Dudley (1984); see Theorem 3.3.1, page 25. Dudley and Philipp (1983) use a strong approximation approach to convergence of general empirical processes in order to avoid measurability difficulties.

**1.11.** The extended continuous mapping theorem 1.11.1 originated with Prohorov (1956) and continued with H. Rubin [see Anderson (1963) and Billingsley (1968), page 34, and Topsoe (1967a, 1967b)]. The current extension to nonmeasurable random elements in Theorem 1.11.1 is used by Wellner (1989) in connection with the delta-method for Hadamard differentiable functions.

**1.12.** The metrization of convergence of laws began with Lévy; he provided a metric for convergence of laws on the real line; see Gnedenko and Kolmogorov (1954), Chapter 2, for a treatment of the Lévy metric. Prohorov (1956) defined a generalization, now called the Prohorov metric, and showed that it metrized weak convergence of distributions on complete, separable metric spaces. See Dudley (1989), Section 11.3, for a summary of this and other metrics. Uniformity of weak convergence over classes of functions was investigated by Ranga Rao (1962) and Topsoe (1967a). Fortet and Mourier (1953) introduced the bounded Lipschitz metric  $d_{BL}$ , and Dudley (1966,

1968) studied it further. That weak convergence in the Hoffmann-Jørgensen theory to a Borel measurable, separable limit is equivalent to convergence of the bounded Lipschitz distance was proved independently by Dudley (1990) and Van der Vaart and Wellner (1989).

PART 2

# Empirical Processes

## 2.1

# Introduction

This part is concerned with convergence of a particular type of random map: the empirical process. The *empirical measure*  $\mathbb{P}_n$  of a sample of random elements  $X_1, \dots, X_n$  on a measurable space  $(\mathcal{X}, \mathcal{A})$  is the discrete random measure given by  $\mathbb{P}_n(C) = n^{-1} \#(1 \leq i \leq n : X_i \in C)$ . Alternatively (if points are measurable), it can be described as the random measure that puts mass  $1/n$  at each observation. We shall frequently write the empirical measure as the linear combination  $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_i}$  of the dirac measures at the observations.

Given a collection  $\mathcal{F}$  of measurable functions  $f: \mathcal{X} \rightarrow \mathbb{R}$ , the empirical measure induces a map from  $\mathcal{F}$  to  $\mathbb{R}$  given by

$$f \mapsto \mathbb{P}_n f.$$

Here, we use the abbreviation  $Qf = \int f dQ$  for a given measurable function  $f$  and signed measure  $Q$ . Let  $P$  be the common distribution of the  $X_i$ . The centered and scaled version of the given map is the  $\mathcal{F}$ -indexed *empirical process*  $\mathbb{G}_n$  given by

$$f \mapsto \mathbb{G}_n f = \sqrt{n}(\mathbb{P}_n - P)f = \frac{1}{\sqrt{n}} \sum_{i=1}^n (f(X_i) - Pf).$$

Frequently the signed measure  $\mathbb{G}_n = n^{-1/2} \sum_{i=1}^n (\delta_{X_i} - P)$  will be identified with the empirical process.

For a given function  $f$ , it follows from the law of large numbers and the central limit theorem that

$$\begin{aligned}\mathbb{P}_n f &\xrightarrow{\text{as}} Pf, \\ \mathbb{G}_n f &\rightsquigarrow N(0, P(f - Pf)^2),\end{aligned}$$

provided  $Pf$  exists and  $Pf^2 < \infty$ , respectively. This part is concerned with making these two statements uniform in  $f$  varying over a class  $\mathcal{F}$ .

Classical empirical process theory concerns the special cases when the sample space  $\mathcal{X}$  is the unit interval in  $[0, 1]$ , the real line  $\mathbb{R}$ , or  $\mathbb{R}^d$  and the indexing collection  $\mathcal{F}$  is taken to be the set of indicators of left half-lines in  $\mathbb{R}$  or lower-left orthants in  $\mathbb{R}^d$ . In this part we are concerned with empirical measures and processes indexed by these classes of functions  $\mathcal{F}$ , but also many others, including indicators of half-spaces, balls, ellipsoids, and other sets in  $\mathbb{R}^d$ , classes of smooth functions from  $\mathbb{R}^d$  to  $\mathbb{R}$ , and monotone functions.

With the notation  $\|Q\|_{\mathcal{F}} = \sup\{|Qf| : f \in \mathcal{F}\}$ , the uniform version of the law of large numbers becomes

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} \rightarrow 0,$$

where the convergence is in outer probability or is outer almost surely. A class  $\mathcal{F}$  for which this is true is called a *Glivenko-Cantelli class*, or also *P-Glivenko-Cantelli class* to bring out the dependence on the underlying measure  $P$ .

In order to discuss a uniform version of the central limit theorem, it is assumed that

$$\sup_{f \in \mathcal{F}} |f(x) - Pf| < \infty, \quad \text{for every } x.$$

Under this condition the empirical process  $\{\mathbb{G}_n f : f \in \mathcal{F}\}$  can be viewed as a map into  $\ell^\infty(\mathcal{F})$ . Consequently, it makes sense to investigate conditions under which

$$(2.1.1) \quad \mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P) \rightsquigarrow \mathbb{G}, \quad \text{in } \ell^\infty(\mathcal{F}),$$

where the limit  $\mathbb{G}$  is a tight Borel measurable element in  $\ell^\infty(\mathcal{F})$ . A class  $\mathcal{F}$  for which this is the true is called a *Donsker class*, or *P-Donsker class* to be more complete.

The nature of the limit process  $\mathbb{G}$  follows from consideration of its marginal distributions. The marginals  $\mathbb{G}_n f$  converge if and only if the functions  $f$  are square-integrable. In that case the multivariate central limit theorem yields that for any finite set  $f_1, \dots, f_k$  of functions

$$(\mathbb{G}_n f_1, \dots, \mathbb{G}_n f_k) \rightsquigarrow N_k(0, \Sigma),$$

where the  $k \times k$ -matrix  $\Sigma$  has  $(i, j)$ th element  $P(f_i - Pf_i)(f_j - Pf_j)$ . Since convergence in  $\ell^\infty(\mathcal{F})$  implies marginal convergence, it follows that the

limit process  $\{\mathbb{G}f: f \in \mathcal{F}\}$  must be a zero-mean Gaussian process with covariance function

$$(2.1.2) \quad \mathbb{E}\mathbb{G}f_1\mathbb{G}f_2 = P(f_1 - Pf_1)(f_2 - Pf_2) = Pf_1f_2 - Pf_1Pf_2.$$

According to Lemma 1.5.3, this and tightness completely determine the distribution of  $\mathbb{G}$  in  $\ell^\infty(\mathcal{F})$ . It is called the *P-Brownian bridge*.

An application of Slutsky's lemma shows that every Donsker class is a Glivenko-Cantelli class in probability. In fact, this is also true with "in probability" replaced by "almost surely." Conversely, not every Glivenko-Cantelli class is Donsker, but the collection of all Donsker classes contains many members of interest. Many classes are also *universally Donsker*, which is defined as *P-Donsker* for every probability measure  $P$  on the sample space.

**2.1.3 Example (Empirical distribution function).** Let  $X_1, \dots, X_n$  be i.i.d. random elements in  $\mathbb{R}^d$ , and let  $\mathcal{F}$  be the collection of all indicator functions of lower rectangles  $\{1\{(-\infty, t]\}: t \in \bar{\mathbb{R}}^d\}$ . The empirical measure indexed by  $\mathcal{F}$  can be identified with the empirical distribution function

$$t \mapsto \mathbb{P}_n 1\{(-\infty, t]\} = \frac{1}{n} \sum_{i=1}^n 1\{X_i \leq t\}.$$

In this case, it is natural to identify  $f = 1\{(-\infty, t]\}$  with  $t \in \bar{\mathbb{R}}^d$  and the space  $\ell^\infty(\mathcal{F})$  with the space  $\ell^\infty(\bar{\mathbb{R}}^d)$ . Of course, the sample paths of the empirical distribution function are all contained in a much smaller space (such as a  $D[-\infty, \infty]$ ), but as long as this space is equipped with the supremum metric, this is irrelevant for the central limit theorem.

It is known from classical results (for the line by Donsker) that the class of lower rectangles is Donsker for any underlying law  $P$  of  $X_1, \dots, X_n$ . These classical results are a simple consequence of the results of this part. Moreover, with no additional effort, analogous results are obtained for the empirical process indexed by the collection of closed balls, rectangles, half-spaces, and ellipsoids, as well as many collections of functions.

**2.1.4 Example (Empirical process indexed by sets).** Let  $\mathcal{C}$  be a collection of measurable sets in the sample space  $(\mathcal{X}, \mathcal{A})$ , and take the class of functions  $\mathcal{F}$  equal to the set of indicator functions of sets in  $\mathcal{C}$ . This leads to the empirical distribution indexed by sets

$$C \mapsto \mathbb{P}_n(C) = \frac{1}{n} \#(X_i \in C).$$

In this case, it is convenient to make the identification  $C \leftrightarrow 1\{C\}$ , both in notation and in terminology. Hence  $\mathcal{C}$  is called a Glivenko-Cantelli class if  $\|\mathbb{P}_n - P\|_c$  converges to zero in outer probability or outer almost surely, and  $\mathcal{C}$  is a Donsker class if  $\sqrt{n}(\mathbb{P}_n - P)$  converges weakly to a tight limit in  $\ell^\infty(\mathcal{C})$ .

An unfortunate technicality of the preceding definitions of Glivenko-Cantelli and Donsker classes is that they depend on the underlying probability space on which  $X_1, X_2, \dots$  are defined, because this determines the outer expectations. In this book, unless specified otherwise, this technicality is resolved by always assuming that  $X_1, X_2, \dots$  are defined *canonically*. Thus the underlying probability space is the product space  $(\mathcal{X}^\infty, \mathcal{B}^\infty, P^\infty)$ , and  $X_i$  is the projection onto the  $i$ th coordinate.<sup>†</sup>

### 2.1.1 Overview of Chapters 2.3–2.14

Whether a given class  $\mathcal{F}$  is a Glivenko-Cantelli or Donsker class depends on the size of the class. A finite class of square integrable functions is always Donsker, while at the other extreme the class of all square integrable, uniformly bounded functions is almost never Donsker. A relatively simple way to measure the size of a class  $\mathcal{F}$  is to use *entropy numbers*. The  $\varepsilon$ -entropy of  $\mathcal{F}$  is essentially the logarithm of the number of “balls” or “brackets” of size  $\varepsilon$  needed to cover  $\mathcal{F}$ . From this informal definition, it is already clear that the entropy numbers increase as  $\varepsilon$  decreases to zero. Simple sufficient conditions for a class to be Glivenko-Cantelli or Donsker can be given in terms of the rate of increase as  $\varepsilon$  tends to zero. The main results on Glivenko-Cantelli classes are given in Chapter 2.4, and the main results on Donsker classes are given in Chapter 2.5. These chapters are the core of Part 2.

For an informal description of the main results, we first define entropy with and without bracketing. Let  $(\mathcal{F}, \|\cdot\|)$  be a subset of a normed space of real functions  $f: \mathcal{X} \mapsto \mathbb{R}$  on some set. We are mostly thinking of  $L_r(Q)$ -spaces for probability measures  $Q$ .

**2.1.5 Definition (Covering numbers).** The *covering number*  $N(\varepsilon, \mathcal{F}, \|\cdot\|)$  is the minimal number of balls  $\{g: \|g - f\| < \varepsilon\}$  of radius  $\varepsilon$  needed to cover the set  $\mathcal{F}$ . The centers of the balls need not belong to  $\mathcal{F}$ , but they should have finite norms. The *entropy (without bracketing)* is the logarithm of the covering number.

**2.1.6 Definition (Bracketing numbers).** Given two functions  $l$  and  $u$ , the *bracket*  $[l, u]$  is the set of all functions  $f$  with  $l \leq f \leq u$ . An  $\varepsilon$ -bracket is a bracket  $[l, u]$  with  $\|u - l\| < \varepsilon$ . The *bracketing number*  $N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)$  is the minimum number of  $\varepsilon$ -brackets needed to cover  $\mathcal{F}$ . The *entropy with bracketing* is the logarithm of the bracketing number. In the definition of the bracketing number, the upper and lower bounds  $u$  and  $l$  of the brackets need not belong to  $\mathcal{F}$  themselves but are assumed to have finite norms.

The norms that will be of interest later, such as the  $L_r(Q)$ -norms, all possess the (Riesz) property: for a pair of functions, if  $|f| \leq |g|$ , then

---

<sup>†</sup> Since the empirical measure depends only on the first  $n$  coordinates, and the projection of  $\mathcal{X}^\infty$  onto  $\mathcal{X}^n$  is a perfect map, any outer expectation  $E^*h(\mathbb{P}_n)$  may be evaluated equivalently, assuming that  $X_1, \dots, X_n$  are defined on  $\mathcal{X}^n$  or on  $\mathcal{X}^\infty$ .

$\|f\| \leq \|g\|$ . For such norms, if  $f$  is in the  $2\varepsilon$ -bracket  $[l, u]$ , then it is in the ball of radius  $\varepsilon$  around  $(l + u)/2$ . Thus covering and bracketing numbers are related by

$$N(\varepsilon, \mathcal{F}, \|\cdot\|) \leq N_{[]} (2\varepsilon, \mathcal{F}, \|\cdot\|).$$

In general, there is no converse inequality, so that bracketing numbers may be bigger than covering numbers. (A notable exception is the uniform norm, for which the previous inequality is an identity.) However, sufficient conditions for the theorems of interest in terms of bracketing numbers can often be stated using only a single norm on  $\mathcal{F}$ , while sufficient conditions in terms of entropy numbers usually involve many norms. This is intuitively obvious since brackets control a function  $f(x)$  pointwise in its argument  $x$ , rather than in norm. As a consequence, sufficient conditions in terms of bracketing numbers or covering numbers are unrelated in general. Both are of interest.

We shall write  $N(\varepsilon, \mathcal{F}, L_r(Q))$  for covering numbers relative to the  $L_r(Q)$ -norm

$$\|f\|_{Q,r} = (\int |f|^r)^{1/r}.$$

This is similar for bracketing numbers. An *envelope function* of a class  $\mathcal{F}$  is any function  $x \mapsto F(x)$  such that  $|f(x)| \leq F(x)$ , for every  $x$  and  $f$ . The minimal envelope function is  $x \mapsto \sup_f |f(x)|$ . It will usually be assumed that this function is finite for every  $x$ . The *uniform entropy numbers* (relative to  $L_r$ ) are defined as

$$\sup_Q \log N(\varepsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q)),$$

where the supremum is over all probability measures  $Q$  on  $(\mathcal{X}, \mathcal{A})$ , with  $0 < QF^r < \infty$ .<sup>†</sup>

In Chapter 2.4, two Glivenko-Cantelli theorems are given, based on entropy with and without bracketing, respectively. The first theorem asserts that  $\mathcal{F}$  is  $P$ -Glivenko-Cantelli if

$$N_{[]} (\varepsilon, \mathcal{F}, L_1(P)) < \infty, \quad \text{for every } \varepsilon > 0.$$

The proof of this theorem is elementary and extends the usual proof of the classical Glivenko-Cantelli theorem for the empirical distribution function. The second theorem is more complicated but has a simple corollary involving the uniform covering numbers. A class  $\mathcal{F}$  is  $P$ -Glivenko-Cantelli if it is  $P$ -measurable with envelope  $F$  such that  $P^*F < \infty$  and satisfies

$$\sup_Q N(\varepsilon \|F\|_{Q,1}, \mathcal{F}, L_1(Q)) < \infty, \quad \text{for every } \varepsilon > 0.$$

The second theorem requires the assumption that  $\mathcal{F}$  is a “ $P$ -measurable class,” a concept that is defined in Definition 2.3.3. The presence of this

---

<sup>†</sup> Alternatively, the supremum could be taken over all discrete probability measures. This yields the same results.

measurability hypothesis is a consequence of the way uniform entropy is used to control suprema via “randomization by Rademacher random variables” (or “symmetrization”) together with the lack of a general version of Fubini’s theorem for outer integrals. This same type of additional measurability hypothesis also enters in the corresponding Donsker theorems with uniform entropy hypotheses. Without these hypotheses, the theorems are false, but on the other hand it requires some effort to construct a class for which the hypotheses fail. Thus, in applications, this type of measurability will hardly ever play a role.

The method of symmetrization is discussed in Chapter 2.3 together with measurability conditions. This section is a necessary preparation for the uniform entropy Glivenko-Cantelli and Donsker theorems in Chapters 2.4 and 2.5. The theorems that are proved with bracketing do not use symmetrization and do not need measurability hypotheses.

One of the two Donsker theorems proved in Chapter 2.5 uses bracketing entropy. A slightly simpler version of this theorem asserts that a class  $\mathcal{F}$  of functions is  $P$ -Donsker under an integrability condition on the  $L_2(P)$ -entropy with bracketing. A class  $\mathcal{F}$  of measurable functions is  $P$ -Donsker if

$$\int_0^\infty \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}, L_2(P))} d\varepsilon < \infty.$$

The other Donsker theorem in this chapter asserts that a class  $\mathcal{F}$  of functions is  $P$ -Donsker under a similar condition on uniform entropy numbers. If  $\mathcal{F}$  is class of functions such that

$$(2.1.7) \quad \int_0^\infty \sup_Q \sqrt{\log N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))} d\varepsilon < \infty,$$

then  $\mathcal{F}$  is  $P$ -Donsker for every probability measure  $P$  such that  $P^*F^2 < \infty$  for some envelope function  $F$  and for which certain measurability hypotheses are satisfied. The integral conditions both measure the rate at which the entropy grows as  $\varepsilon$  decreases to zero. Note that the entropies are zero for large  $\varepsilon$  (if  $\mathcal{F}$  is bounded), so that convergence of the integrals at  $\infty$  is automatic.

Once the main Glivenko-Cantelli theorems and Donsker theorems are in hand, the following question arises: how can we verify the hypotheses on uniform entropy numbers and bracketing numbers? Tools and techniques for this purpose are developed in Chapter 2.6 for uniform entropy numbers and in Chapter 2.7 for bracketing numbers.

One of the starting points for controlling uniform covering numbers is the notion of a *Vapnik-Červonenkis class of sets*, or simply *VC-class*. Say that a collection  $\mathcal{C}$  of subsets of the sample space  $\mathcal{X}$  *picks out* a certain subset of the finite set  $\{x_1, \dots, x_n\} \subset \mathcal{X}$  if it can be written as  $\{x_1, \dots, x_n\} \cap C$  for some  $C \in \mathcal{C}$ . The collection  $\mathcal{C}$  is said to *shatter*  $\{x_1, \dots, x_n\}$  if  $\mathcal{C}$  picks out each of its  $2^n$  subsets. The *VC-index*  $V(\mathcal{C})$  of  $\mathcal{C}$  is the smallest  $n$  for which no set of size  $n$  is shattered by  $\mathcal{C}$ . A collection  $\mathcal{C}$  of measurable

sets is called a *VC-class* if its index  $V(\mathcal{C})$  is finite. Thus, by definition, a VC-class of sets picks out strictly less than  $2^n$  subsets from any set of  $n \geq V(\mathcal{C})$  elements. The surprising fact is that such a class can necessarily pick out only a polynomial number  $O(n^{V(\mathcal{C})-1})$  of subsets, well below the  $2^n - 1$  that the definition appears to allow. This result is a consequence of a combinatorial result, known as Sauer's lemma, which states that the number  $\Delta_n(\mathcal{C}, x_1, \dots, x_n)$  of subsets picked out by a VC-class  $\mathcal{C}$  satisfies

$$\max_{x_1, \dots, x_n} \Delta_n(\mathcal{C}, x_1, \dots, x_n) \leq \sum_{j=0}^{V(\mathcal{C})-1} \binom{n}{j} \leq \left( \frac{ne}{V(\mathcal{C})-1} \right)^{V(\mathcal{C})-1}.$$

Some thought shows that the number of subsets picked out by a collection  $\mathcal{C}$  is closely related to the covering numbers of the class of indicator functions  $\{\mathbf{1}_C : C \in \mathcal{C}\}$  in  $L_1(Q)$  for discrete, empirical type measures  $Q$ . By a clever argument, Sauer's lemma can be used to bound the uniform covering (or entropy) numbers for this class. Theorem 2.6.4 asserts that there exists a universal constant  $K$  such that, for any VC-class  $\mathcal{C}$  of sets,

$$N(\varepsilon, \mathcal{C}, L_r(Q)) \leq KV(\mathcal{C})(4e)^{V(\mathcal{C})} \left( \frac{1}{\varepsilon} \right)^{r(V(\mathcal{C})-1)},$$

for any probability measure  $Q$  and  $r \geq 1$  and  $0 < \varepsilon < 1$ . Consequently, VC-classes are examples of *polynomial classes* in the sense that their covering numbers are bounded by a polynomial in  $1/\varepsilon$ . The upper bound shows that VC-classes satisfy the sufficient conditions for the Glivenko-Cantelli theorem and Donsker theorem discussed previously (with much to spare) provided they possess certain measurability properties.

It is possible to obtain similar results for classes of functions. A collection of real-valued, measurable functions  $\mathcal{F}$  on a sample space  $\mathcal{X}$  is called a *VC-subgraph class* (or simply a VC-class of functions) if the collection of all subgraphs of the functions in  $\mathcal{F}$  forms a VC-class of sets in  $\mathcal{X} \times \mathbb{R}$ . Just as for sets, the covering numbers of a VC-subgraph class  $\mathcal{F}$  grow polynomially in  $1/\varepsilon$ , so that suitably measurable VC-subgraph classes are Glivenko-Cantelli and Donsker under moment conditions on the envelope function.

Among the many examples of VC-classes are the collections of left half-lines and lower-left orthants, with which classical empirical process theory is concerned. Methods to construct a great variety of VC-classes are discussed in Section 2.6.5.

The application of the other main Glivenko-Cantelli and Donsker theorem requires estimates on the bracketing numbers of classes of functions. These are given in Chapter 2.7 for classes of smooth functions, sets with smooth boundaries, convex sets, monotone functions, and functions smoothly depending on a parameter. Coupled with the results of Chapter 2.4 and 2.5, these estimates yield many more interesting Donsker classes.

Given a basic collection of Glivenko-Cantelli and Donsker classes and estimates of their entropies, new classes with similar properties can be constructed in several ways. For instance, the *symmetric convex hull*  $\text{sconv } \mathcal{F}$  of a class  $\mathcal{F}$  is the collection of all functions of the form  $\sum_{i=1}^m \alpha_i f_i$ , with each  $f_i \in \mathcal{F}$  and  $\sum_{i=1}^m |\alpha_i| \leq 1$ . In Chapter 2.10 it is shown that the convex hull of a Donsker class is Donsker. In addition to this, Theorem 2.6.9 gives a useful bound for the entropy of a convex hull in the case that covering numbers for  $\mathcal{F}$  are polynomial: if

$$N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \leq C \left( \frac{1}{\varepsilon} \right)^V,$$

then the convex hull of  $\mathcal{F}$  satisfies

$$\log N(\varepsilon \|F\|_{Q,2}, \text{sconv } \mathcal{F}, L_2(Q)) \leq K \left( \frac{1}{\varepsilon} \right)^{2V/(V+2)}.$$

An important feature of this bound is that the power  $2V/(V+2)$  of  $1/\varepsilon$  is strictly less than 2 for any  $V$ , so that the convex hull of a polynomial class satisfies the uniform entropy condition (2.1.7). This bound is true in particular for *VC-hull classes*, which are defined as the sequential closures of the convex hulls of VC-classes.

Other operations preserving the Donsker property are considered in Chapter 2.10 and include the formation of classes of functions of the form  $\phi(f_1, \dots, f_k)$ , for  $f_1, \dots, f_k$  ranging over given Donsker classes  $\mathcal{F}_1, \dots, \mathcal{F}_k$  and a fixed Lipschitz function  $\phi$  on  $\mathbb{R}^k$ . For example, the classes of pairwise sums  $\mathcal{F} + \mathcal{G}$ , pairwise infima  $\mathcal{F} \wedge \mathcal{G}$ , and pairwise suprema  $\mathcal{F} \vee \mathcal{G}$ , as well as the union  $\mathcal{F} \cup \mathcal{G}$ , are Donsker classes if both  $\mathcal{F}$  and  $\mathcal{G}$  are Donsker. Similarly, pairwise products  $\mathcal{F}\mathcal{G}$  and quotients  $\mathcal{F}/\mathcal{G}$  are Donsker if  $\mathcal{F}$  and  $\mathcal{G}$  are Donsker and bounded above, or bounded away from, zero. The latter boundedness assumptions can be traded against stronger conditions on the entropies of the classes. Such “permanence properties” applied to the basic Donsker classes resulting from Chapters 2.6 and 2.7 provide a very effective method to verify the Donsker property in, for instance, statistical applications.

The remainder of Part 2 explores refinements, extensions, special classes  $\mathcal{F}$ , uniformity in the underlying distribution, multiplier processes, partial-sum processes, the non-i.i.d. situation, and moment and tail bounds for the supremum of the empirical process.

The uniformity of weak convergence of  $\mathbb{G}_n$  to  $\mathbb{G}$  with respect to the underlying probability distribution  $P$  generating the data is addressed in Chapter 2.8. Again, the main results use either uniform entropy or bracketing entropy.

The viewpoint in Chapter 2.9 on multiplier central limit theorems is that, for a Donsker class  $\mathcal{F}$ ,

$$\mathbb{G}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n (\delta_{X_i} - P) \equiv \frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i \rightsquigarrow \mathbb{G},$$

in  $\ell^\infty(\mathcal{F})$ . Then we pose the following question: for what sequences of i.i.d., real-valued random variables  $\xi_1, \dots, \xi_n$  independent from the original data does it follow that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i Z_i \rightsquigarrow \mathbb{G}?$$

In fact, these two displays turn out to be equivalent if the  $\xi_i$  have zero mean, have variance 1, and satisfy the  *$L_{2,1}$ -condition*

$$\|\xi_1\|_{2,1} = \int_0^\infty \sqrt{\mathbb{P}(|\xi| > t)} dt < \infty.$$

Chapter 2.9 also presents conditional versions of this multiplier central limit theorem. These are basic for part 3's development of limit theorems for the bootstrap empirical process.

Chapter 2.11 gives extensions of the Donsker theorem for sums of independent, but not identically distributed, processes  $\sum_{i=1}^n Z_{ni}$  indexed by an arbitrary collection  $\mathcal{F}$ . One example, which occurs naturally in statistical contexts and is covered by this situation, is the empirical process indexed by a collection  $\mathcal{F}$  of functions based on observations  $X_{n1}, \dots, X_{nn}$  that are independent but not identically distributed. Once again we present two main central limit theorems, based on entropy with and without bracketing. Even if specified to the i.i.d. situation, the theorems in Chapter 2.11 are more general than the theorems obtained before. For instance, we use a random entropy condition, rather than a uniform entropy condition, and also consider bracketing using majorizing measures.

Two types of *partial-sum processes* are studied in Chapter 2.12. The first type is the *sequential empirical process*

$$\mathbb{Z}_n(s, f) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor ns \rfloor} (f(X_i) - Pf) = \sqrt{\frac{\lfloor ns \rfloor}{n}} \mathbb{G}_{\lfloor ns \rfloor}(f),$$

which records the “history” of the empirical processes  $\mathbb{G}_n$  as sampling progresses. The second type of partial-sum processes studied are those generated by independent random variables located at the points of a lattice. These processes, which can be viewed as closely related to the multiplier processes studied in Chapter 2.9, since they are empirical processes with random masses (the multipliers) at (degenerately) random points, have some special features and deserve special attention.

Chapter 2.13 gives Donsker theorems for several special types of classes  $\mathcal{F}$ , namely, sequences and elliptical classes.

Part 2 concludes with a detailed study of moment bounds, tail bounds, and exponential bounds for the supremum  $\|\mathbb{G}_n\|_{\mathcal{F}}$  of the empirical process in Chapter 2.14. The moment bounds play an important role in Chapters 3.2 and 3.4 on the limiting theory of  $M$ -estimators.

### 2.1.2 Asymptotic Equicontinuity

A class of measurable functions is called *pre-Gaussian* if the (tight) limit process  $\mathbb{G}$  in the uniform central limit theorem (2.1.1) exists. By Kolmogorov's extension theorem, there always exists a zero-mean Gaussian process  $\{\mathbb{G}f: f \in \mathcal{F}\}$  with covariance function given by (2.1.2). A more precise description of pre-Gaussianity is that there exists a version of this Gaussian process that is a tight, Borel measurable map from some probability space into  $\ell^\infty(\mathcal{F})$ . Usually the name "Brownian bridge" introduced earlier refers to this tight process with its special sample path properties, rather than to the general stochastic process  $\mathbb{G}$ .

It is desirable to have a more concrete description of the tightness property of a Brownian bridge and hence of the notion of pre-Gaussianity. In Chapter 1.5 it was seen that tightness of a random map into  $\ell^\infty(\mathcal{F})$  is closely connected to continuity of its sample paths. Define a seminorm  $\rho_P$  by

$$\rho_P(f) = (P(f - Pf)^2)^{1/2}.$$

Then, by Example 1.5.10, a class  $\mathcal{F}$  is pre-Gaussian if and only if

- $\mathcal{F}$  is totally bounded for  $\rho_P$ ,
- there exists a version of  $\mathbb{G}$  with uniformly  $\rho_P$ -continuous sample paths  $f \mapsto \mathbb{G}(f)$ .

Actually, this example shows that instead of the centered  $L_2$ -norm  $\rho_P$ , any centered  $L_r$ -norm can be used as well. However, the  $L_2$ -norm appears to be the most convenient.<sup>b</sup>

While pre-Gaussianity of a class  $\mathcal{F}$  is necessary for the uniform central limit theorem, it is not sufficient. A Donsker class  $\mathcal{F}$  satisfies the stronger condition that the sequence  $\mathbb{G}_n$  is asymptotically tight. By Theorem 1.5.7, the latter entails replacing the condition that the sample paths of the limit process are continuous by the condition that the empirical process is asymptotically continuous: for every  $\varepsilon > 0$ ,

$$(2.1.8) \quad \lim_{\delta \downarrow 0} \limsup_{n \rightarrow \infty} P^* \left( \sup_{\rho_P(f-g) < \delta} |\mathbb{G}_n(f - g)| > \varepsilon \right) = 0.$$

Much of Part 2 is concerned with this equicontinuity condition. With the notation<sup>#</sup>  $\mathcal{F}_\delta = \{f - g: f, g \in \mathcal{F}, \rho_P(f - g) < \delta\}$ , it is equivalent to the following statement: for every decreasing sequence  $\delta_n \downarrow 0$ ,

$$\|\mathbb{G}_n\|_{\mathcal{F}_{\delta_n}} \xrightarrow{P^*} 0$$

<sup>b</sup> It may be noted that the seminorm  $\rho_P$  is the  $L_2(P)$ -norm of  $f$  centered at its expectation (in fact, the standard deviation of  $f(X_i)$ ). The centering is natural in view of the fact that the empirical process is invariant under the addition of constants to  $\mathcal{F}$ :  $\mathbb{G}_n f = \mathbb{G}_n(f - c)$  for every  $c$ . Under the condition that  $\sup_{f \in \mathcal{F}} |Pf| < \infty$ , the seminorm  $\rho_P$  can be replaced by the slightly simpler  $L_2(P)$ -seminorm without loss of generality (Problem 2.1.2).

<sup>#</sup> In most of this part, the notation  $\mathcal{F}_\delta$  means  $\mathcal{F}_\delta = \{f - g: f, g \in \mathcal{F}, \rho(f - g) < \delta\}$ , for  $\rho$  equal to either the  $L_2(P)$ -semimetric or  $\rho_P$ .

(Problem 2.1.5). In view of Theorem 1.5.7, a class  $\mathcal{F}$  is Donsker if and only if  $\mathcal{F}$  is totally bounded in  $\mathcal{L}_2(P)$  and satisfies the asymptotic equicontinuity condition as in the preceding displays.

### 2.1.3 Maximal Inequalities

It follows that both the law of large numbers and the central limit theorem are concerned with showing that a supremum of real-valued variables converges to zero. To a certain extent, the derivation of these results is done in parallel, as both need to make use of *maximal inequalities* that bound probabilities involving suprema of random variables. Chapter 2.2 develops such inequalities in an abstract setting, using Orlicz norms. If  $\psi$  is a non-decreasing, convex function with  $\psi(0) = 0$ , then the Orlicz norm  $\|X\|_\psi$  of a random variable  $X$  is defined by

$$\|X\|_\psi = \inf \left\{ C > 0 : \mathbb{E} \psi \left( \frac{|X|}{C} \right) \leq 1 \right\}.$$

The functions  $\psi$  of primary interest are  $\psi(x) = x^p$  and  $\psi(x) = \exp(x^p) - 1$  (for  $p \geq 1$  or adapted versions if  $0 < p < 1$ ). For  $\psi(x) = x^p$ , the Orlicz norm  $\|X\|_\psi$  is just the usual  $L_p$ -norm  $\|X\|_p$ . The key Lemma 2.2.2 provides a bound for the Orlicz norm of a finite maximum  $\max_{1 \leq i \leq m} X_i$  in terms of the Orlicz norms of the variables  $X_i$  and the inverse function  $\psi^{-1}(m)$ . Suppose that  $\limsup_{x,y \rightarrow \infty} \psi(x)\psi(y)/\psi(xy) < \infty$ . Then there is a constant  $K$  depending only on  $\psi$  so that, for arbitrary random variables  $X_1, \dots, X_m$ ,

$$\left\| \max_{1 \leq i \leq m} X_i \right\|_\psi \leq K \psi^{-1}(m) \max_{1 \leq i \leq m} \|X_i\|_\psi.$$

The interesting feature of this inequality is that a bound on  $\max_i \|X_i\|_\psi$  for a rapidly growing function  $\psi$ , such as  $\psi_2(x) = \exp(x^2) - 1$ , yields a bound on the rate of growth with  $m$  of the Orlicz norm of  $\max_i X_i$  governed by the slowly growing function  $\psi^{-1}(m)$ , such as  $\psi_2^{-1}(m) = \sqrt{\log(m+1)}$ . Since Gaussian random variables and sums of independent, bounded random variables enjoy finite  $\psi_2$ -Orlicz norms, the particular function  $\psi_2^{-1}(m)$  figures prominently in the development in later chapter.

Bounds on finite suprema can be extended to general maximal inequalities with the help of the *chaining method*, which Kolmogorov pioneered. The idea is to relate maxima over an infinite set to maxima of increments over an increasing sequence of finite sets. Here covering numbers are introduced to measure the size of the approximating finite sets. For a semimetric space  $(T, d)$ , the *covering number*  $N(\varepsilon, T, d)$  is defined as the minimal number of balls of radius  $\varepsilon$  needed to cover  $T$ . Given a separable stochastic process  $\{X_t : t \in T\}$  and a fixed function  $\psi$ , define a semimetric by  $d(s, t) = \|X_s - X_t\|_\psi$  for every  $s, t$ . Then, for any  $\eta, \delta > 0$  and a constant  $K$ ,

$$\left\| \sup_{d(s,t) \leq \delta} |X_s - X_t| \right\|_\psi \leq K \left[ \int_0^\eta \psi^{-1}(N(\varepsilon, d)) d\varepsilon + \delta \psi^{-1}(N^2(\eta, d)) \right].$$

Thus the  $\psi$ -norm of a supremum is bounded by an integral that involves the inverse  $\psi^{-1}$  and the covering numbers of the index set  $T$  with respect to the *intrinsic semimetric*  $d$ .

In particular, a process  $\{X_t: t \in T\}$  is said to be *sub-Gaussian* if the  $\psi_2$ -norms of the increments  $X_s - X_t$  are finite. This can be shown to be true if and only if, for some semimetric  $d$ ,

$$\mathbb{P}(|X_s - X_t| > x) \leq 2e^{-\frac{1}{2}x^2/d^2(s,t)}, \quad \text{for every } x > 0.$$

Then the preceding inequality can be simplified to

$$\mathbb{E} \sup_{d(s,t) \leq \delta} |X_s - X_t| \leq K \int_0^\delta \sqrt{\log N(\varepsilon, T, d)} \, d\varepsilon.$$

This inequality explains the appearance of the square root of the logarithm of the entropy in the sufficient integral conditions for a class to be Donsker, which were discussed in the preceding chapter.

#### \* 2.1.4 The Central Limit Theorem in Banach Spaces

The convergence in distribution in  $\ell^\infty(\mathcal{F})$  of the empirical process  $\mathbb{G}_n = n^{-1/2} \sum_{i=1}^n (\delta_{X_i} - P)$  can be considered a central limit theorem for the i.i.d. random elements  $\delta_{X_1} - P, \dots, \delta_{X_n} - P$  in the Banach space  $\ell^\infty(\mathcal{F})$ . In this section it is first shown that, conversely, any central limit theorem in a Banach space can be stated in terms of empirical processes. Next it is argued that this observation is only moderately useful.

Suppose  $Z_1, \dots, Z_n$  are i.i.d. random maps with values in a Banach space  $\mathbf{X}$ . Let  $\mathcal{F}$  be a subset of the dual space of  $\mathbf{X}$  such that  $\sup_f |f(x)| = \|x\|$  for every  $x$  and such that  $f(Z_i)$  is measurable for each  $f$ . A mean of  $Z_i$  is an element  $EZ_i$  of  $\mathbf{X}$  such that  $f(EZ_i) = Ef(Z_i)$ , for every  $f \in \mathcal{F}$ . Every given element  $x \in \mathbf{X}$  can be identified with a map

$$x^{**}: f \mapsto f(x),$$

from  $\mathcal{F}$  to  $\mathbb{R}$ . By assumption, this gives an isometric identification  $x \leftrightarrow x^{**}$  of  $\mathbf{X}$  with a subset of  $\ell^\infty(\mathcal{F})$ . Thus random maps  $Z_1, \dots, Z_n$  in  $\mathbf{X}$  satisfy the central limit theorem if and only if the random elements  $Z_1^{**}, \dots, Z_n^{**}$  satisfy the central limit theorem in  $\ell^\infty(\mathcal{F})$ . Since  $\sum_{i=1}^n (Z_i - EZ_i)^{**}(f) = \sum_{i=1}^n (f(Z_i) - Ef(Z_i))$ , this appears to be the case if and only if  $\mathcal{F}$  is a Donsker class of functions. However, for an accurate statement, it is necessary to describe the measurability structure more precisely.

\* This section may be skipped at first reading.

**2.1.9 Example.** In case of a separable Banach space it is usually assumed that  $Z_1, \dots, Z_n$  are Borel measurable maps. Then “i.i.d.” can be understood in the usual manner. If  $\mathcal{F}$  is taken equal to the unit ball of the dual space, then the equality  $\sup_f |f(x)| = \|x\|$  is valid by the Hahn-Banach theorem, and the  $\sigma$ -field generated by  $\mathcal{F}$  is precisely the Borel  $\sigma$ -field. It follows that the  $Z_i$  are Borel measurable if and only if every  $f(Z_i)$  is measurable.

The conclusion is that the sequence  $n^{-1/2} \sum_{i=1}^n (Z_i - EZ_i)$  converges in distribution if and only if the unit ball of the dual space is Donsker with respect to the common Borel law of the  $Z_i$ .

**2.1.10 Example.** Suppose  $Z_1, \dots, Z_n$  are i.i.d. stochastic processes with bounded sample paths, indexed by an arbitrary set  $\tilde{\mathcal{F}}$  (not necessarily a class of functions). The notion “i.i.d.” should include that the finite-dimensional marginals  $(Z_i(\tilde{f}_1), \dots, Z_i(\tilde{f}_k))$  are i.i.d. vectors. In general, this does not determine the “distribution” of the processes in  $\ell^\infty(\tilde{\mathcal{F}})$  adequately. Assume in addition that each  $Z_i$  is defined on a product probability space  $(\mathcal{X}^n, \mathcal{A}^n, P^n)$  as  $Z_i(x_1, \dots, x_n)\tilde{f} = Z(\tilde{f}, x_i)$ , for some fixed stochastic process  $\{Z(\tilde{f}, x) : \tilde{f} \in \tilde{\mathcal{F}}\}$ . Then  $\sum_{i=1}^n (Z_i(f) - EZ_i(\tilde{f})) = \sum_{i=1}^n (f(x_i) - Pf)$ , for the function  $f$  defined by  $f(x) = Z(\tilde{f}, x)$ . That  $Z$  is a stochastic process means precisely that each function  $x \mapsto f(x)$  is measurable. The processes  $Z_1, Z_2, \dots$  satisfy the central limit theorem in  $\ell^\infty(\tilde{\mathcal{F}})$  if and only if the class of functions  $\{f : \tilde{f} \in \tilde{\mathcal{F}}\}$  is  $P$ -Donsker.

Thus the central limit theorem in an arbitrary Banach space can be phrased as a result about empirical processes. Even though this conclusion is formally correct, the methods (traditionally) used for proving empirical central limit theorems yield unsatisfactory descriptions of the central limit theorem in special Banach spaces. For instance, the conditions for the central limit theorem in  $L_p$ -space are simple and well known. If  $(S, \Sigma, \mu)$  is a  $\sigma$ -finite measure space and  $Z_1, \dots, Z_n$  are i.i.d., zero-mean Borel measurable maps into  $L_p(S, \Sigma, \mu)$ , then the sequence  $n^{-1/2} \sum_{i=1}^n Z_i$  converges weakly if and only if  $P(\|Z_1\|_p > t) = o(t^{-2})$  as  $t \rightarrow \infty$  and

$$\int_S (EZ_1^2(s))^{p/2} d\mu(s) < \infty.$$

(In case  $p = 2$ , this can be simplified to the single requirement  $E\|Z_1\|_2^2 < \infty$ . For this case, the theorem is given in Chapter 1.8.) In terms of empirical processes, this becomes as follows.

**2.1.11 Proposition.** Let  $(S, \Sigma, \mu)$  be a  $\sigma$ -finite measure space, let  $1 \leq p < \infty$ , and let  $P$  be a Borel probability measure on  $L_p(S, \Sigma, \mu)$ . Then the unit ball of the dual space of  $L_p(S, \Sigma, \mu)$  is  $P$ -Donsker if and only if  $\int_S (\int z(\omega, s)^2 dP(\omega))^{p/2} d\mu(s) < \infty$  and  $P(\|z\|_p > t) = o(t^{-2})$  as  $t \rightarrow \infty$ .

This proposition is proved in texts on probability in Banach spaces.<sup>†</sup> The methods developed in this part do not yield this result. In fact, already the formulation of the proposition is unnecessarily complicated: it is awkward to push this central limit theorem into an empirical mold.

On the other hand, Part 2 gives many interesting results on the central limit theorem in Banach spaces of the type  $\ell^\infty(\mathcal{F})$ , where  $\mathcal{F}$  is a collection of measurable functions. For this large class of Banach spaces, sufficient conditions for the central limit theorem go beyond the Banach space structure, but they can be stated in terms of the structure of the function class  $\mathcal{F}$ .

## Problems and Complements

1. (**Total boundedness in  $L_2(P)$** ) If  $\mathcal{F}$  is totally bounded in  $L_2(P)$ , then it is totally bounded for the seminorm  $\rho_P$ . If  $\mathcal{F}$  is totally bounded for  $\rho_P$  and  $\|P\|_{\mathcal{F}} = \sup\{|Pf| : f \in \mathcal{F}\}$  is finite, then it is totally bounded in  $L_2(P)$ .

[Hint: Suppose  $f_1, \dots, f_m$  form an  $\varepsilon$ -net for  $\rho_P$ . If  $\sup\{|Pf| : f \in \mathcal{F}\}$  is finite, then the numbers  $Pf$  with  $f$  ranging over some  $\rho_P$ -ball of radius  $\varepsilon$  are contained in an interval. Choose for every  $f_i$  a finite collection  $f_{i,j}$  such that, for every  $f$  with  $\rho_P(f - f_i) < \varepsilon$ , there is an  $f_{i,j}$  with  $|Pf - Pf_{i,j}| < \varepsilon$ . Then the  $f_{i,j}$  form a  $\sqrt{5}\varepsilon$ -net for  $\mathcal{F}$  in  $L_2(P)$ .]

2. (**Using  $\|\cdot\|_{P,2}$  instead of  $\rho_P$** ) Suppose  $\sup\{|Pf| : f \in \mathcal{F}\}$  is finite. Then a class  $\mathcal{F}$  of measurable functions is  $P$ -Donsker if and only if  $\mathcal{F}$  is totally bounded in  $L_2(P)$  and the empirical process is asymptotically equicontinuous in probability for the  $L_2(P)$ -semimetric.

[Hint: Use the previous problem and the next problem.]

3. Suppose  $\rho$  is a semimetric on a class of measurable functions  $\mathcal{F}$  that is uniformly stronger than  $\rho_P$  in the sense that  $\rho_P(f - g) \leq \phi(\rho(f, g))$ , for a function  $\phi$  with  $\phi(\varepsilon) \rightarrow 0$  as  $\varepsilon \downarrow 0$ . Suppose  $\mathcal{F}$  is totally bounded under  $\rho$ . Then  $\mathcal{F}$  is  $P$ -Donsker if and only if the empirical process indexed by  $\mathcal{F}$  is asymptotically equicontinuous in probability with respect to  $\rho$ .

[Hint: Use the addendum to Theorem 1.5.7.]

4. (**Brownian motion**) If  $\mathcal{F}$  is  $P$ -pre-Gaussian and  $\|P\|_{\mathcal{F}} < \infty$ , then a  $P$ -Brownian bridge has uniformly continuous sample paths with respect to the  $L_2(P)$ -semimetric. The process  $Z(f) = G(f) + \xi P f$  for a standard normal variable  $\xi$  independent of  $G$  has a version that is a Borel measurable, tight map in  $\ell^\infty(\mathcal{F})$ . This process, which is Gaussian with mean zero and covariance function  $Pfg$ , is called a *Brownian motion*.

5. If  $a_n: [0, 1] \mapsto [0, \infty)$  is a sequence of nondecreasing functions, then there exists  $\delta_n \downarrow 0$  such that  $\limsup a_n(\delta_n) = \lim_{\delta \downarrow 0} \limsup_{n \rightarrow \infty} a_n(\delta)$ . Deduce that  $\lim_{\delta \downarrow 0} \limsup_{n \rightarrow \infty} a_n(\delta) = 0$  if and only if  $a_n(\delta_n) \rightarrow 0$  for every  $\delta_n \rightarrow 0$ .

---

<sup>†</sup> See Ledoux and Talagrand (1991), Theorem 10.10.

6. The packing numbers of a ball of radius  $R$  in  $\mathbb{R}^d$  satisfy

$$D(\varepsilon, B(0, R), \|\cdot\|) \leq \left(\frac{3R}{\varepsilon}\right)^d, \quad 0 < \varepsilon \leq R,$$

for the Euclidean norm  $\|\cdot\|$ .

[**Hint:** If  $x_1, \dots, x_m$  is an  $\varepsilon$ -separated subset in  $B(0, R)$ , then the balls of radii  $\varepsilon/2$  around the  $x_i$  are disjoint and contained in  $B(0, R + \varepsilon/2)$ . Comparison of the volume of their union and the volume of  $B(0, R + \varepsilon/2)$  shows that  $m(\varepsilon/2)^d \leq (R + \varepsilon/2)^d$ .]

## 2.2

# Maximal Inequalities and Covering Numbers

In this chapter we derive a class of maximal inequalities that can be used to establish the asymptotic equicontinuity of the empirical process. Since the inequalities have much wider applicability, we temporarily leave the empirical process framework.

Let  $\psi$  be a nondecreasing, convex function with  $\psi(0) = 0$  and  $X$  a random variable. Then the *Orlicz norm*  $\|X\|_\psi$  is defined as

$$\|X\|_\psi = \inf \left\{ C > 0 : \mathbb{E} \psi \left( \frac{|X|}{C} \right) \leq 1 \right\}.$$

(Here the infimum over the empty set is  $\infty$ .) Using Jensen's inequality, it is not difficult to check that this indeed defines a norm (on the set of random variables for which  $\|X\|_\psi$  is finite). The best-known examples of Orlicz norms are those corresponding to the functions  $x \mapsto x^p$  for  $p \geq 1$ : the corresponding Orlicz norm is simply the  $L_p$ -norm

$$\|X\|_p = (\mathbb{E}|X|^p)^{1/p}.$$

For our purposes, Orlicz norms of more interest are the ones given by  $\psi_p(x) = e^{x^p} - 1$  for  $p \geq 1$ , which give much more weight to the tails of  $X$ . The bound  $x^p \leq \psi_p(x)$  for all nonnegative  $x$  implies that  $\|X\|_p \leq \|X\|_{\psi_p}$  for each  $p$ . It is not true that the exponential Orlicz norms are all bigger than all  $L_p$ -norms. However, we have the inequalities

$$\begin{aligned} \|X\|_{\psi_p} &\leq \|X\|_{\psi_q} (\log 2)^{p/q}, & p \leq q, \\ \|X\|_p &\leq p! \|X\|_{\psi_1} \end{aligned}$$

(see Problem 2.2.5). Since for the present purposes fixed constants in inequalities are irrelevant, this means that a bound on an exponential Orlicz norm always gives a better result than a bound on an  $L_p$ -norm.

Any Orlicz norm can be used to obtain an estimate of the tail of a distribution. By Markov's inequality,

$$\mathrm{P}(|X| > x) \leq \mathrm{P}\left(\psi(|X|/\|X\|_\psi) \geq \psi(x/\|X\|_\psi)\right) \leq \frac{1}{\psi(x/\|X\|_\psi)}.$$

For  $\psi_p(x) = e^{x^p} - 1$ , this leads to tail estimates  $\exp(-Cx^p)$  for any random variable with a finite  $\psi_p$ -norm. Conversely, an exponential tail bound of this type shows that  $\|X\|_{\psi_p}$  is finite.

**2.2.1 Lemma.** *Let  $X$  be a random variable with  $\mathrm{P}(|X| > x) \leq Ke^{-Cx^p}$  for every  $x$ , for constants  $K$  and  $C$ , and for  $p \geq 1$ . Then its Orlicz norm satisfies  $\|X\|_{\psi_p} \leq ((1+K)/C)^{1/p}$ .*

**Proof.** By Fubini's theorem

$$\mathrm{E}(e^{D|X|^p} - 1) = \mathrm{E} \int_0^{|X|^p} De^{Ds} ds = \int_0^\infty P(|X| > s^{1/p}) De^{Ds} ds.$$

Now insert the inequality on the tails of  $|X|$  and obtain the explicit upper bound  $KD/(C-D)$ . This is less than or equal to 1 for  $D^{-1/p}$  greater than or equal to  $((1+K)/C)^{1/p}$ . ■

Next consider the  $\psi$ -norm of a maximum of finitely many random variables. Using the fact that  $\max |X_i|^p \leq \sum |X_i|^p$ , one easily obtains for the  $L_p$ -norms

$$\left\| \max_{1 \leq i \leq m} X_i \right\|_p = \left( \mathrm{E} \max_{1 \leq i \leq m} |X_i|^p \right)^{1/p} \leq m^{1/p} \max_{1 \leq i \leq m} \|X_i\|_p.$$

A similar inequality is valid for many Orlicz norms, in particular the exponential ones. Here, in the general case, the factor  $m^{1/p}$  becomes  $\psi^{-1}(m)$ , where  $\psi^{-1}$  is the inverse function of  $\psi$ .

**2.2.2 Lemma.** *Let  $\psi$  be a convex, nondecreasing, nonzero function with  $\psi(0) = 0$  and  $\limsup_{x,y \rightarrow \infty} \psi(x)\psi(y)/\psi(cx) < \infty$  for some constant  $c$ . Then, for any random variables  $X_1, \dots, X_m$ ,*

$$\left\| \max_{1 \leq i \leq m} X_i \right\|_\psi \leq K \psi^{-1}(m) \max_i \|X_i\|_\psi,$$

for a constant  $K$  depending only on  $\psi$ .

**Proof.** For simplicity of notation assume first that  $\psi(x)\psi(y) \leq \psi(cxy)$  for all  $x, y \geq 1$ . In that case  $\psi(x/y) \leq \psi(cx)/\psi(y)$  for all  $x \geq y \geq 1$ . Thus, for  $y \geq 1$  and any  $C$ ,

$$\begin{aligned} \max \psi\left(\frac{|X_i|}{Cy}\right) &\leq \max \left[ \frac{\psi(c|X_i|/C)}{\psi(y)} + \psi\left(\frac{|X_i|}{Cy}\right) \mathbf{1}\left\{\frac{|X_i|}{Cy} < 1\right\} \right] \\ &\leq \sum \frac{\psi(c|X_i|/C)}{\psi(y)} + \psi(1). \end{aligned}$$

Set  $C = c \max \|X_i\|_\psi$ , and take expectations to get

$$E\psi\left(\frac{\max |X_i|}{Cy}\right) \leq \frac{m}{\psi(y)} + \psi(1).$$

When  $\psi(1) \leq 1/2$ , this is less than or equal to 1 for  $y = \psi^{-1}(2m)$ , which is greater than 1 under the same condition. Thus

$$\|\max |X_i|\|_\psi \leq \psi^{-1}(2m) c \max \|X_i\|_\psi.$$

By the convexity of  $\psi$  and the fact that  $\psi(0) = 0$ , it follows that  $\psi^{-1}(2m) \leq 2\psi^{-1}(m)$ . The proof is complete for every special  $\psi$  that meets the conditions made previously.

For a general  $\psi$ , there are constants  $\sigma \leq 1$  and  $\tau > 0$  such that  $\phi(x) = \sigma\psi(\tau x)$  satisfies the conditions of the previous paragraph. Apply the inequality to  $\phi$ , and observe that  $\|X\|_\psi \leq \|X\|_\phi/(\sigma\tau) \leq \|X\|_\psi/\sigma$  (Problem 2.2.3). ■

For the present purposes, the value of the constant in the previous lemma is irrelevant. (For the  $L_p$ -norms, it can be taken equal to 1.) The important conclusion is that the inverse of the  $\psi$ -function determines the size of the  $\psi$ -norm of a maximum in comparison to the  $\psi$ -norms of the individual terms. The  $\psi$ -norm grows slowest for rapidly increasing  $\psi$ . For  $\psi(x) = e^{x^p} - 1$ , the growth is at most logarithmic, because

$$\psi_p^{-1}(m) = (\log(1+m))^{1/p}.$$

The previous lemma is useless in the case of a maximum over infinitely many variables. However, such a case can be handled via repeated application of the lemma via a method known as *chaining*. Every random variable in the supremum is written as a sum of “little links,” and the bound depends on the number and size of the little links needed. For a stochastic process  $\{X_t: t \in T\}$ , the number of links depends on the entropy of the index set for the semimetric

$$d(s, t) = \|X_s - X_t\|_\psi.$$

The general definition of “metric entropy” is as follows.

**2.2.3 Definition (Covering numbers).** Let  $(T, d)$  be an arbitrary semimetric space. Then the *covering number*  $N(\varepsilon, d)$  is the minimal number of balls of radius  $\varepsilon$  needed to cover  $T$ . Call a collection of points  $\varepsilon$ -separated if the distance between each pair of points is strictly larger than  $\varepsilon$ . The *packing number*  $D(\varepsilon, d)$  is the maximum number of  $\varepsilon$ -separated points in  $T$ . The corresponding *entropy numbers* are the logarithms of the covering and packing numbers, respectively.

For the present purposes, both covering and packing numbers can be used. In all arguments one can be replaced by the other through the inequalities

$$N(\varepsilon, d) \leq D(\varepsilon, d) \leq N\left(\frac{1}{2}\varepsilon, d\right).$$

Clearly, covering and packing numbers become bigger as  $\varepsilon \downarrow 0$ . By definition, the semimetric space  $T$  is totally bounded if and only if the covering and packing numbers are finite for every  $\varepsilon > 0$ . The upper bound in the following maximal inequality depends on the rate at which  $D(\varepsilon, d)$  grows as  $\varepsilon \downarrow 0$ , as measured through an integral criterion.

**2.2.4 Theorem.** Let  $\psi$  be a convex, nondecreasing, nonzero function with  $\psi(0) = 0$  and  $\limsup_{x,y \rightarrow \infty} \psi(x)\psi(y)/\psi(cxy) < \infty$ , for some constant  $c$ . Let  $\{X_t : t \in T\}$  be a separable stochastic process<sup>†</sup> with

$$\|X_s - X_t\|_\psi \leq C d(s, t), \quad \text{for every } s, t,$$

for some semimetric  $d$  on  $T$  and a constant  $C$ . Then, for any  $\eta, \delta > 0$ ,

$$\left\| \sup_{d(s,t) \leq \delta} |X_s - X_t| \right\|_\psi \leq K \left[ \int_0^\eta \psi^{-1}(D(\varepsilon, d)) d\varepsilon + \delta \psi^{-1}(D^2(\eta, d)) \right],$$

for a constant  $K$  depending on  $\psi$  and  $C$  only.

**2.2.5 Corollary.** The constant  $K$  can be chosen such that

$$\left\| \sup_{s,t} |X_s - X_t| \right\|_\psi \leq K \int_0^{\text{diam } T} \psi^{-1}(D(\varepsilon, d)) d\varepsilon,$$

where  $\text{diam } T$  is the diameter of  $T$ .

**Proof.** Assume without loss of generality that the packing numbers and the associated “covering integral” are finite. Construct nested sets  $T_0 \subset T_1 \subset T_2 \subset \dots \subset T$  such that every  $T_j$  is a maximal set of points such that  $d(s, t) > \eta 2^{-j}$  for every  $s, t \in T_j$  where “maximal” means that no point can be added without destroying the validity of the inequality. By

---

<sup>†</sup> Separable may be understood in the sense that  $\sup_{d(s,t) < \delta} |X_s - X_t|$  remains almost surely the same if the index set  $T$  is replaced by a suitable countable subset.

the definition of packing numbers, the number of points in  $T_j$  is less than or equal to  $D(\eta 2^{-j}, d)$ .

“Link” every point  $t_{j+1} \in T_{j+1}$  to a unique  $t_j \in T_j$  such that  $d(t_j, t_{j+1}) \leq \eta 2^{-j}$ . Thus obtain for every  $t_{k+1}$  a chain  $t_{k+1}, t_k, \dots, t_0$  that connects it to a point in  $T_0$ . For arbitrary points  $s_{k+1}, t_{k+1}$  in  $T_{k+1}$ , the difference in increments along their chains can be bounded by

$$\begin{aligned} |(X_{s_{k+1}} - X_{s_0}) - (X_{t_{k+1}} - X_{t_0})| &= \left| \sum_{j=0}^k (X_{s_{j+1}} - X_{s_j}) - \sum_{j=0}^k (X_{t_{j+1}} - X_{t_j}) \right| \\ &\leq 2 \sum_{j=0}^k \max |X_u - X_v|, \end{aligned}$$

where for fixed  $j$  the maximum is taken over all links  $(u, v)$  from  $T_{j+1}$  to  $T_j$ . Thus the  $j$ th maximum is taken over at most  $\#T_{j+1}$  links, with each link having a  $\psi$ -norm  $\|X_u - X_v\|_\psi$  bounded by  $C d(u, v) \leq C \eta 2^{-j}$ . It follows with the help of Lemma 2.2.2 that, for a constant depending only on  $\psi$  and  $C$ ,

$$\begin{aligned} \left\| \max_{s, t \in T_{k+1}} |(X_s - X_{s_0}) - (X_t - X_{t_0})| \right\|_\psi &\leq K \sum_{j=0}^k \psi^{-1}(D(\eta 2^{-j-1}, d)) \eta 2^{-j} \\ (2.2.6) \quad &\leq 4K \int_0^\eta \psi^{-1}(D(\varepsilon, d)) d\varepsilon. \end{aligned}$$

In this bound,  $s_0$  and  $t_0$  are the endpoints of the chains starting at  $s$  and  $t$ , respectively.

The maximum of the increments  $|X_{s_{k+1}} - X_{t_{k+1}}|$  can be bounded by the maximum on the left side of (2.2.6) plus the maximum of the discrepancies  $|X_{s_0} - X_{t_0}|$  at the end of the chains. The maximum of the latter discrepancies will be analyzed by a seemingly circular argument. For every pair of endpoints  $s_0, t_0$  of chains starting at two points in  $T_{k+1}$  within distance  $\delta$  of each other, choose exactly one pair  $s_{k+1}, t_{k+1}$  in  $T_{k+1}$ , with  $d(s_{k+1}, t_{k+1}) < \delta$ , whose chains end at  $s_0, t_0$ . By definition of  $T_0$ , this gives at most  $D^2(\eta, d)$  pairs. By the triangle inequality,

$$|X_{s_0} - X_{t_0}| \leq |(X_{s_0} - X_{s_{k+1}}) - (X_{t_0} - X_{t_{k+1}})| + |X_{s_{k+1}} - X_{t_{k+1}}|.$$

Take the maximum over all pairs of endpoints  $s_0, t_0$  as above. Then the corresponding maximum over the first term on the right in the last display is bounded by the maximum in the left side of (2.2.6). Its  $\psi$ -norm can be bounded by the right side of this equation. Combine this with (2.2.6) to find that

$$\left\| \max_{\substack{s, t \in T_{k+1} \\ d(s, t) < \delta}} |X_s - X_t| \right\|_\psi \leq 8K \int_0^\eta \psi^{-1}(D(\varepsilon, d)) d\varepsilon + \|\max |X_{s_{k+1}} - X_{t_{k+1}}|\|_\psi.$$

Here the maximum on the right is taken over the pairs  $s_{k+1}, t_{k+1}$  in  $T_{k+1}$  uniquely attached to the pairs  $s_0, t_0$  as above. Thus the maximum is over at most  $D^2(\eta, d)$  terms, each of whose  $\psi$ -norm is bounded by  $\delta$ . Its  $\psi$ -norm is bounded by  $K\psi^{-1}(D^2(\eta, d))\delta$ .

Thus the upper bound given by the theorem is a bound for the maximum of increments over  $T_{k+1}$ . Let  $k$  tend to infinity to conclude the proof.

The corollary follows immediately from the previous proof, after noting that, for  $\eta$  equal to the diameter of  $T$ , the set  $T_0$  consists of exactly one point. In that case  $s_0 = t_0$  for every pair  $s, t$ , and the increments at the end of the chains are zero. The corollary also follows from the theorem upon taking  $\eta = \delta = \text{diam } T$  and noting that  $D(\eta, d) = 1$ , so that the second term in the maximal inequality can also be written  $\delta\psi^{-1}(D(\eta, d))$ . Since the function  $\varepsilon \mapsto \psi^{-1}(D(\varepsilon, d))$  is decreasing, this term can be absorbed into the integral, perhaps at the cost of increasing the constant  $K$ . ■

Though the theorem gives a bound on the continuity modulus of the process, a bound on the maximum of the process will be needed. Of course, for any  $t_0$ ,

$$\left\| \sup_t |X_t| \right\|_\psi \leq \|X_{t_0}\|_\psi + \int_0^{\text{diam } T} \psi^{-1}(D(\varepsilon, d)) d\varepsilon.$$

Nevertheless, to state the maximal inequality in terms of the increments appears natural. The increment bound shows that the process  $X$  is continuous in  $\psi$ -norm, whenever the covering integral  $\int_0^\eta \psi^{-1}(D(\varepsilon, d)) d\varepsilon$  converges for some  $\eta > 0$ . (In that case, the right side in Theorem 2.2.4 can be made arbitrarily small by choosing first  $\eta$  small and next  $\delta$ .) It is a small step to deduce the continuity of almost all sample paths from this inequality, but this is not needed at this point (Problem 2.2.17).

### 2.2.1 Sub-Gaussian Inequalities

A standard normal variable has tails of the order  $x^{-1} \exp -\frac{1}{2}x^2$  and satisfies  $P(|X| > x) \leq 2 \exp -\frac{1}{2}x^2$  for every  $x$ . By direct calculation one finds a  $\psi_2$ -norm of  $\sqrt{8/3}$ . In this subsection we study random variables satisfying similar tail bounds.

Hoeffding's inequality asserts a "sub-Gaussian" tail bound for random variables of the form  $X = \sum X_i$  with  $X_1, \dots, X_n$  i.i.d. with zero means and bounded range. (See Proposition A.6.1.) The following special case of Hoeffding's inequality will be needed.

**2.2.7 Lemma (Hoeffding's inequality).** *Let  $a_1, \dots, a_n$  be constants and  $\varepsilon_1, \dots, \varepsilon_n$  be Rademacher random variables; i.e., with  $P(\varepsilon_i = 1) = P(\varepsilon_i = -1) = 1/2$ . Then*

$$P\left(\left|\sum \varepsilon_i a_i\right| > x\right) \leq 2 e^{-\frac{1}{2}x^2/\|a\|^2},$$

for the Euclidean norm  $\|a\|$ . Consequently,  $\|\sum \varepsilon_i a_i\|_{\psi_2} \leq \sqrt{6}\|a\|$ .

**Proof.** For any  $\lambda$  and Rademacher variable  $\varepsilon$ , one has  $Ee^{\lambda\varepsilon} = (e^\lambda + e^{-\lambda})/2 \leq e^{\lambda^2/2}$ , where the last inequality follows after writing out the power series. Thus by Markov's inequality, for any  $\lambda > 0$ ,

$$P\left(\sum_{i=1}^n a_i \varepsilon_i > x\right) \leq e^{-\lambda x} Ee^{\lambda \sum_{i=1}^n a_i \varepsilon_i} \leq e^{(\lambda^2/2) \|a\|^2 - \lambda x}.$$

The best upper bound is obtained for  $\lambda = x/\|a\|^2$  and is the exponential in the probability bound of the lemma. Combination with a similar bound for the lower tail yields the probability bound.

The bound on the  $\psi$ -norm is a consequence of the probability bound in view of Lemma 2.2.1. ■

A stochastic process is called *sub-Gaussian* with respect to the semimetric  $d$  on its index set if

$$P(|X_s - X_t| > x) \leq 2e^{-\frac{1}{2}x^2/d^2(s,t)}, \quad \text{for every } s, t \in T, x > 0.$$

(The constants 2 and  $1/2$  are of no special importance. See Problem 2.2.14.) Any Gaussian process is sub-Gaussian for the standard deviation semimetric  $d(s, t) = \sigma(X_s - X_t)$ . Another example is the *Rademacher process*

$$X_a = \sum_{i=1}^n a_i \varepsilon_i, \quad a \in \mathbb{R}^n,$$

for Rademacher variables  $\varepsilon_1, \dots, \varepsilon_n$ . By Hoeffding's inequality, this is sub-Gaussian for the Euclidean distance  $d(a, b) = \|a - b\|$ .

Sub-Gaussian processes satisfy the increment bound  $\|X_s - X_t\|_{\psi_2} \leq \sqrt{6} d(s, t)$ . Since the inverse of the  $\psi_2$ -function is essentially the square root of the logarithm, the general maximal inequality leads for sub-Gaussian processes to a bound in terms of an entropy integral. (Remember that entropy is defined as the logarithm of packing numbers.) Furthermore, because of the special properties of the logarithm, the statement can be slightly simplified.

**2.2.8 Corollary.** Let  $\{X_t : t \in T\}$  be a separable sub-Gaussian process. Then for every  $\delta > 0$ ,

$$E \sup_{d(s,t) \leq \delta} |X_s - X_t| \leq K \int_0^\delta \sqrt{\log D(\varepsilon, d)} d\varepsilon,$$

for a universal constant  $K$ . In particular, for any  $t_0$ ,

$$E \sup_t |X_t| \leq E|X_{t_0}| + K \int_0^\infty \sqrt{\log D(\varepsilon, d)} d\varepsilon.$$

**Proof.** Apply the general maximal inequality with  $\psi_2(x) = e^{x^2} - 1$  and  $\eta = \delta$ . Since  $\psi_2^{-1}(m) = \sqrt{\log(1+m)}$ , we have  $\psi_2^{-1}(D^2(\delta, d)) \leq$

$\sqrt{2}\psi_2^{-1}(D(\delta, d))$ . Thus the second term in the maximal inequality can first be replaced by  $\sqrt{2}\delta\psi^{-1}(D(\eta, d))$  and next be incorporated in the first (the covering integral) at the cost of increasing the constant. We obtain

$$\left\| \sup_{d(s,t) \leq \delta} |X_s - X_t| \right\|_{\psi_2} \leq K \int_0^\delta \sqrt{\log(1 + D(\varepsilon, d))} d\varepsilon.$$

Here  $D(\varepsilon, d) \geq 2$  for every  $\varepsilon$  that is strictly less than the diameter of  $T$ . Since  $\log(1 + m) \leq 2 \log m$  for  $m \geq 2$ , the 1 inside the logarithm can be removed at the cost of increasing  $K$ . ■

## 2.2.2 Bernstein's Inequality

Since many random variables have larger than normal tails, the results of the previous subsection are not always useful. A sum  $\sum_{i=1}^n Y_i$  of independent variables with mean zero and bounded range is, for large  $n$ , approximately normally distributed with variance  $v = \text{var}(Y_1 + \dots + Y_n)$ . The tails of a normal  $N(0, v)$  variable are of the order  $\exp(-x^2/(2v))$ . If the variables  $Y_i$  have range  $[-M, M]$ , then Bernstein's inequality gives a tail bound  $\exp(-x^2/[2v + (2Mx/3)])$  for the variables  $\sum_{i=1}^n Y_i$ . The extra term,  $2Mx/3$ , may be viewed as a penalty for the nonnormality: for  $n \rightarrow \infty$ , it is typically negligible with respect to  $v = v_n$ .

**2.2.9 Lemma (Bernstein's inequality).** *For independent random variables  $Y_1, \dots, Y_n$  with bounded ranges  $[-M, M]$  and zero means,*

$$P(|Y_1 + \dots + Y_n| > x) \leq 2e^{-\frac{x^2}{2v+Mx/3}},$$

for  $v \geq \text{var}(Y_1 + \dots + Y_n)$ .

**Proof.** See Pollard (1984) or Shorack and Wellner (1986), page 855. ■

For large  $x$ , the upper bound in Bernstein's inequality is essentially of the exponential type  $\exp(-x/M)$ ; for  $x$  close to zero the upper bound behaves like the normal upper bound,  $\exp(-x^2/(2v))$ . This suggests that a maximum of variables that satisfy a Bernstein-type bound can be controlled using a combination of the  $\psi_1$  and  $\psi_2$  Orlicz norms.

**2.2.10 Lemma.** *Let  $X_1, \dots, X_m$  be arbitrary random variables that satisfy the tail bound*

$$P(|X_i| > x) \leq 2e^{-\frac{x^2}{b+ax}},$$

for all  $x$  (and  $i$ ) and fixed  $a, b > 0$ . Then

$$\left\| \max_{1 \leq i \leq m} X_i \right\|_{\psi_1} \leq K \left( a \log(1 + m) + \sqrt{b} \sqrt{\log(1 + m)} \right),$$

for a universal constant  $K$ .

**Proof.** The condition implies the upper bound  $2 \exp(-x^2/(4b))$  on  $P(|X_i| > x)$ , for every  $x \leq b/a$ , and the upper bound  $2 \exp(-x/(4a))$ , for all other positive  $x$ . Consequently, the same upper bounds hold for all  $x > 0$  for the probabilities  $P(|X_i| 1\{|X_i| \leq b/a\} > x)$  and  $P(|X_i| 1\{|X_i| > b/a\} > x)$ , respectively. By Lemma 2.2.1, this implies that the Orlicz norms  $\|X_i 1\{|X_i| \leq b/a\}\|_{\psi_2}$  and  $\|X_i 1\{|X_i| > b/a\}\|_{\psi_1}$  are up to constants bounded by  $\sqrt{b}$  and  $a$ , respectively. Next

$$\|\max_i X_i\|_{\psi_1} \leq \|\max_i X_i 1\{|X_i| \leq b/a\}\|_{\psi_1} + \|\max_i X_i 1\{|X_i| > b/a\}\|_{\psi_1}.$$

Since the  $\psi_p$ -norms are up to constants nondecreasing in  $p$ , the first  $\psi_1$ -norm on the right can be replaced by a  $\psi_2$ -norm. Finally, apply Lemma 2.2.2 to find the bound as stated. ■

A refined form of Bernstein's inequality will be useful in Part 3. A random variable  $Y$  with bounded range  $[-M, M]$  satisfies

$$E|Y|^m \leq M^{m-2} EY^2,$$

for every  $m \geq 2$ . A much weaker inequality than this is the essential element in the proof of Bernstein's inequality.

**2.2.11 Lemma (Bernstein's inequality).** *Let  $Y_1, \dots, Y_n$  be independent random variables with zero mean such that  $E|Y_i|^m \leq m! M^{m-2} v_i / 2$ , for every  $m \geq 2$  (and all  $i$ ) and some constants  $M$  and  $v_i$ . Then*

$$P(|Y_1 + \dots + Y_n| > x) \leq 2 e^{-\frac{1}{2} \frac{x^2}{v+Mx}},$$

for  $v \geq v_1 + \dots + v_n$ .

**Proof.** See Bennett (1962), pages 37–38. ■

The moment condition in the refined form of Bernstein's inequality is somewhat odd. It is implied by

$$E\left(e^{|Y_i|/M} - 1 - \frac{|Y_i|}{M}\right) M^2 \leq \frac{1}{2} v_i.$$

Conversely, if  $Y_i$  satisfies the moment condition of the lemma, then the preceding display is valid with  $M$  replaced by  $2M$  and  $v_i$  replaced by  $2v_i$ . Thus for applications where constants in Bernstein's inequality are unimportant, the preceding display is “equivalent” to the moment condition of the lemma.

### \* 2.2.3 Tightness Under an Increment Bound

In the next chapters we obtain general central limit theorems for empirical processes through the application of maximal inequalities. Independently, the next example shows how a classical, simple sufficient condition for weak convergence follows also from the maximal inequalities.

**2.2.12 Example.** Let  $\{X_n(t): t \in [0, 1]\}$  be a sequence of separable stochastic processes with bounded sample paths and increments satisfying

$$\mathbb{E}|X_n(s) - X_n(t)|^p \leq K|s - t|^{1+r},$$

for constants  $p, K, r > 0$  independent of  $n$ . Assume that the sequences of marginals  $(X_n(t_1), \dots, X_n(t_k))$  converge weakly to the corresponding marginals of a stochastic process  $\{X(t): t \in [0, 1]\}$ . Then there exists a version of  $X$  with continuous sample paths and  $X_n \rightsquigarrow X$  in  $\ell^\infty[0, 1]$ . (Hence also in  $D[0, 1]$  or  $C[0, 1]$ , provided every  $X_n$  has all its sample paths in these spaces.)

To prove this for  $p > 1$ , apply Theorem 2.2.4 with  $\psi(x) = x^p$  and  $d(s, t) = |s - t|^\alpha$ , where  $\alpha = ((1+r)/p) \wedge 1$ . A  $d$ -ball of radius  $\varepsilon$  around some  $t$  is simply the interval  $[t - \varepsilon^{1/\alpha}, t + \varepsilon^{1/\alpha}]$ . Thus the index set  $[0, 1]$  can be covered with  $N(\varepsilon, d) = (1/2)\varepsilon^{-1/\alpha}$  balls of radius  $\varepsilon$ . Since  $\psi^{-1}(x) = x^{1/p}$ , the covering integral can be bounded by a multiple of

$$\int_0^\eta \psi^{-1}(N(\varepsilon, d)) d\varepsilon \leq \int_0^\eta \varepsilon^{-1/(p\alpha)} d\varepsilon.$$

For  $p > 1$ , the integral converges. Since  $\|X_n(s) - X_n(t)\|_p \leq K^{1/p}|s - t|^\alpha$ , the general maximal inequality and Markov's inequality give

$$\mathbb{P}\left(\sup_{|s-t|<\delta} |X_n(s) - X_n(t)| > x\right) \leq \frac{C}{x} \left[ \int_0^\eta \varepsilon^{-1/(p\alpha)} d\varepsilon + \delta \eta^{-2/(p\alpha)} \right],$$

for some constant  $C$  independent of  $n$ . By choosing first  $\eta$  and next  $\delta$ , one can make the right side arbitrarily small. This verifies the asymptotic equicontinuity of  $X_n$ . Its weak convergence and the continuity of the limit process follow from Theorem 1.5.7.

For  $p \leq 1$ , use the inequality  $|(x^+)^{1/q} - (y^+)^{1/q}| \leq |x - y|^{1/q}$  (valid for all reals  $x, y$ , and  $q \geq 1$ ), with  $q = 2/p$ , to derive that

$$\mathbb{E}|X_n^+(s)^{p/2} - X_n^+(t)^{p/2}|^2 \leq K|s - t|^{1+r}, \quad \text{for every } s, t.$$

By the previous argument, it follows that  $(X_n^+)^{p/2} \rightsquigarrow (X^+)^{p/2}$ , where  $(X^+)^{p/2}$  is a process with continuous sample paths. Since the map  $x \mapsto x^q$  from  $\ell^\infty[0, 1]$  to  $\ell^\infty[0, 1]$  is continuous at every  $x \in C[0, 1]$  (Problem 2.2.15), it follows that  $X_n^+ \rightsquigarrow X^+$ . By a similar argument,  $X_n^- \rightsquigarrow X^-$ . Together this yields  $X_n \rightsquigarrow X$ .

\* This section may be skipped at first reading.

## Problems and Complements

1. A standard normal variable  $X$  possesses norm  $\|X\|_{\psi_2} = \sqrt{8/3}$  for  $\psi_2(x) = \exp(x^2) - 1$ .
2. The constant random variable  $X = 1$  possesses norm  $\|X\|_{\psi_p} = (\log 2)^{-1/p}$  for  $\psi_p(x) = e^{x^p} - 1$ .
3. For any constant  $C \geq 1$ , convex, increasing  $\psi$  and random variable  $X$ , one has  $\|X\|_{C\psi} \leq C\|X\|_\psi$ . For  $C \leq 1$  the reverse inequality holds.  
**[Hint:** By convexity of  $\psi$ , one has  $E\psi(|X|/C\|X\|_\psi) \leq 1/C$ , for  $C \geq 1$ .]
4. Show that  $\|X\|_p \leq \lceil p/2 \rceil! \|X\|_{\psi_2}$ . In particular,  $\|X\|_1 \leq \|X\|_2 \leq \|X\|_{\psi_2}$ .  
**[Hint:** For even  $p$ , one has  $x^p \leq (p/2)! \psi_2(x)$ .]
5. (**Comparing  $\psi_p$ -norms**) For any  $X$ , the numbers  $(\log 2)^{1/p}\|X\|_{\psi_p}$  for  $\psi_p(x) = \exp(x^p) - 1$  are nondecreasing in  $p \geq 1$ .  
**[Hint:** The function  $\phi$  for which  $\psi_p(x(\log 2)^{1/p}) = \phi(\psi_q((\log 2)^{1/q}x))$  is concave and satisfies  $\phi(1) = 1$  for  $q \geq p$ . Use Jensen's inequality.]
6. (**Monotone convergence for Orlicz-norms**) Let  $\psi$  be a convex, nondecreasing, nonzero function on  $[0, \infty)$  with  $\psi(0) = 0$ . If  $0 \leq X_n \uparrow X$  almost surely, then  $\|X_n\|_\psi \uparrow \|X\|_\psi$ .  
**[Hint:** By the monotone convergence theorem,  $\lim E\psi(X_n/r\|X\|_\psi) > 1$ , for any  $r < 1$ .]
7. The infimum in the definition of an Orlicz norm is attained (at  $\|X\|_\psi$ ).
8. Let  $\psi$  be any function as in the definition of an Orlicz norm. Then for any random variables  $X_1, \dots, X_m$ , one has  $E \max |X_i| \leq \psi^{-1}(m) \max \|X_i\|_\psi$ .  
**[Hint:** For any  $C$ , one has  $E \max |X_i|/C \leq \psi^{-1}(E \max \psi(|X_i|/C))$  by Jensen's inequality. Bound the maximum by a sum and take  $C = \max \|X_i\|_\psi$ .]
9. If  $\|X\|_\psi < \infty$ , then  $\|X\|_{\sigma\psi} < \infty$  for every  $\sigma > 0$ . Markov's inequality yields the bound  $P(|X| > t) \leq 1/\sigma \cdot 1/\psi(t/\|X\|_{\sigma\psi})$  for every  $t$ . Is there an optimal value of  $\sigma$ ?
10. The condition on the quotient  $\psi(x)\psi(y)/\psi(cxy)$  in Lemma 2.2.2 is not automatic. For instance, the function  $\psi(x) = (x+1)(\log(x+1) - 1) + 1$  is convex and increasing for  $x \geq 0$  but does not satisfy the condition.
11. The function  $\psi_p(x) = \exp(x^p) - 1$  satisfies the hypothesis of Lemma 2.2.2 for every  $0 < p \leq 2$ .
12. Let  $\psi$  be a convex, positive function, with  $\psi(0) = 0$  such that  $\log \psi$  is convex. Then there is a constant  $\sigma$  such that  $\phi(x) = \sigma\psi(x)$  satisfies both  $\phi(1) < 1$  and  $\phi(x)\phi(y) \leq \phi(xy)$ , for all  $x, y \geq 1$ .  
**[Hint:** Define  $\phi(x) = \psi(x)/2\psi(1)$  for  $x \geq 1$ . Then  $\phi(1) = 1/2$  and  $g(x, y) = \phi(x)\phi(y)/\phi(xy)$  satisfies  $g(1, 1) = 1/2$  and is decreasing in both  $x$  and  $y$  for  $x, y \geq 1$  if  $\log \psi$  is convex. The derivative of  $g$  with respect to  $x$  is given by  $\psi(x)\psi(y)/(2\psi(1)\psi(xy))(\psi'/\psi(x) - y\psi'/\psi(xy))$ , which is less than or equal to 0 since  $y \geq 1$  and  $\log \psi$  is convex. By symmetry, the derivative with respect to  $y$  is also bounded above by zero.]

13. The functions  $\psi_p$ , with  $1 \leq p \leq 2$ , and more generally the function  $\psi(x) = \exp(h(x))$ , with  $h(x)$  convex ( $h(x) = x(\log(x)-1)+1$ ), satisfy the conditions in the preceding problem. The functions  $\psi_p$ , for  $0 < p \leq 1$ , do not satisfy these conditions.
14. Suppose that  $P(|X_s - X_t| > x) \leq K \exp(-Cx^2/d^2(s,t))$  for a given stochastic process and certain positive constants  $K$  and  $C$ . Then the process is sub-Gaussian for a multiple of the distance  $d$ .
- [Hint: There exists a constant  $D$  depending only on  $K$  and  $C$  such that  $1 \wedge Ke^{-Cx^2}$  is bounded above by  $2e^{-Dx^2}$ , for every  $x \geq 0$ .]
15. Let  $T$  be a compact, semimetric space and  $q > 0$ . Then the map  $x \mapsto x^q$  from the nonnegative functions in  $\ell^\infty(T)$  to  $\ell^\infty(T)$  is continuous at every continuous  $x$ .
- [Hint: It suffices to show that  $\|x_n - x\|_\infty \rightarrow 0$  implies that  $x_n^q(t_n) - x^q(t_n) \rightarrow 0$  for every sequence  $t_n$ . By compactness of  $T$ , the sequence  $t_n$  may be assumed convergent. Since  $x$  is continuous,  $x(t_n) \rightarrow x(t)$ . Thus  $x_n(t_n) \rightarrow x(t)$ , whence  $x_n^q(t_n) \rightarrow x^q(t)$ .]
16. Suppose the metric space  $T$  has a finite diameter and  $\log N(\varepsilon, T, d) \leq h(\varepsilon)$ , for all sufficiently small  $\varepsilon > 0$  and a fixed, positive, continuous function  $h$ . Then there exists a constant  $K$  with  $\log N(\varepsilon, T, d) \leq Kh(\varepsilon)$  for all  $\varepsilon > 0$ .
- [Hint: For sufficiently large  $\varepsilon$ , the entropy numbers are zero, and for sufficiently small  $\varepsilon$ , the inequality is valid with  $K = 1$ . The function  $\log N(\varepsilon, T, d)/h(\varepsilon)$  is bounded on intervals that are bounded away from zero and infinity.]
17. Let  $X$  be a stochastic process with  $\sup_{\rho(s,t) < \delta} |X(s) - X(t)| \xrightarrow{P^*} 0$  as  $\delta \downarrow 0$ . Then almost all sample paths of  $X$  are uniformly continuous.
- [Hint: Given a decreasing sequence of numbers  $\varepsilon_k \rightarrow 0$ , there are numbers  $\delta_k > 0$  such that  $P^*(\sup_{\rho(s,t) < \delta_k} |X(s) - X(t)| > \varepsilon_k) \leq 2^{-k}$  for every  $k$ . By the Borel-Cantelli lemma,  $|X(s) - X(t)| \leq \varepsilon_k$  whenever  $\rho(s,t) < \delta_k$ , for all sufficiently large  $k$  almost surely.]

## 2.3

# Symmetrization and Measurability

One of the two main approaches toward deriving Glivenko-Cantelli and Donsker theorems is based on the principle of comparing the empirical process to a “symmetrized” empirical process. In this chapter we derive the main symmetrization theorem, as well as a number of technical complements, which may be skipped at first reading.

### 2.3.1 Symmetrization

Let  $\varepsilon_1, \dots, \varepsilon_n$  be i.i.d. Rademacher random variables. Instead of the empirical process

$$f \mapsto (\mathbb{P}_n - P)f = \frac{1}{n} \sum_{i=1}^n (f(X_i) - Pf),$$

consider the symmetrized process

$$f \mapsto \mathbb{P}_n^o f = \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i),$$

where  $\varepsilon_1, \dots, \varepsilon_n$  are independent of  $(X_1, \dots, X_n)$ . Both processes have mean function zero (because  $E(\varepsilon_i f(X_i) | X_i) = 0$  by symmetry of  $\varepsilon_i$ ). It turns out that the law of large numbers or the central limit theorem for one of these processes holds if and only if the corresponding result is true for the other process. One main approach to proving empirical limit theorems is to pass from  $\mathbb{P}_n - P$  to  $\mathbb{P}_n^o$  and next apply arguments conditionally on the original  $X$ 's. The idea is that, for fixed  $X_1, \dots, X_n$ , the symmetrized

empirical measure is a Rademacher process, hence a sub-Gaussian process, to which Corollary 2.2.8 can be applied.

Thus we need to bound maxima and moduli of the process  $\mathbb{P}_n - P$  by those of the symmetrized process. To formulate such bounds, we must be careful about the possible nonmeasurability of suprema of the type  $\|\mathbb{P}_n - P\|_{\mathcal{F}}$ . The result will be formulated in terms of outer expectation, but it does not hold for every choice of an underlying probability space on which  $X_1, \dots, X_n$  are defined. Throughout this part, if outer expectations are involved, it is assumed that  $X_1, \dots, X_n$  are the coordinate projections on the product space  $(\mathcal{X}^n, \mathcal{A}^n, P^n)$ , and the outer expectations of functions  $(X_1, \dots, X_n) \mapsto h(X_1, \dots, X_n)$  are computed for  $P^n$ . Thus “independent” is understood in terms of a product probability space. If auxiliary variables, independent of the  $X$ ’s, are involved, as in the next lemma, we use a similar convention. In that case, the underlying probability space is assumed to be of the form  $(\mathcal{X}^n, \mathcal{A}^n, P^n) \times (\mathcal{Z}, \mathcal{C}, Q)$  with  $X_1, \dots, X_n$  equal to the coordinate projections on the first  $n$  coordinates and the additional variables depending only on the  $(n+1)$ st coordinate.

The following lemma will be used mostly with the choice  $\Phi(x) = x$ .

**2.3.1 Lemma (Symmetrization).** *For every nondecreasing, convex  $\Phi: \mathbb{R} \mapsto \mathbb{R}$  and class of measurable functions  $\mathcal{F}$ ,*

$$\mathbb{E}^* \Phi \left( \|\mathbb{P}_n - P\|_{\mathcal{F}} \right) \leq \mathbb{E}^* \Phi \left( 2 \|\mathbb{P}_n^o\|_{\mathcal{F}} \right),$$

where the outer expectations are computed as indicated in the preceding paragraph.

**Proof.** Let  $Y_1, \dots, Y_n$  be independent copies of  $X_1, \dots, X_n$ , defined formally as the coordinate projections on the last  $n$  coordinates in the product space  $(\mathcal{X}^n, \mathcal{A}^n, P^n) \times (\mathcal{Z}, \mathcal{C}, Q) \times (\mathcal{X}^n, \mathcal{A}^n, P^n)$ . The outer expectations in the statement of the lemma are unaffected by this enlargement of the underlying probability space, because coordinate projections are perfect maps. For fixed values  $X_1, \dots, X_n$ ,

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} = \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n [f(X_i) - \mathbb{E} f(Y_i)] \right| \leq \mathbb{E}_Y^* \sup_{f \in \mathcal{F}} \frac{1}{n} \left| \sum_{i=1}^n [f(X_i) - f(Y_i)] \right|,$$

where  $\mathbb{E}_Y^*$  is the outer expectation with respect to  $Y_1, \dots, Y_n$  computed for  $P^n$  for given, fixed values of  $X_1, \dots, X_n$ . Combination with Jensen’s inequality yields

$$\Phi(\|\mathbb{P}_n - P\|_{\mathcal{F}}) \leq \mathbb{E}_Y \Phi \left( \left\| \frac{1}{n} \sum_{i=1}^n [f(X_i) - f(Y_i)] \right\|_{\mathcal{F}}^{*Y} \right),$$

where  $*Y$  denotes the minimal measurable majorant of the supremum with respect to  $Y_1, \dots, Y_n$ , still with  $X_1, \dots, X_n$  fixed. Because  $\Phi$  is nondecreasing and continuous, the  $*Y$  inside  $\Phi$  can be moved to  $\mathbb{E}_Y^*$  (Problem 1.2.8).

Next take the expectation with respect to  $X_1, \dots, X_n$  to get

$$E^* \Phi(\|\mathbb{P}_n - P\|_{\mathcal{F}}) \leq E_X^* E_Y^* \Phi\left(\frac{1}{n} \left\| \sum_{i=1}^n [f(X_i) - f(Y_i)] \right\|_{\mathcal{F}}\right).$$

Here the repeated outer expectation can be bounded above by the joint outer expectation  $E^*$  by Lemma 1.2.6.

Adding a minus sign in front of a term  $[f(X_i) - f(Y_i)]$  has the effect of exchanging  $X_i$  and  $Y_i$ . By construction of the underlying probability space as a product space, the outer expectation of any function  $f(X_1, \dots, X_n, Y_1, \dots, Y_n)$  remains unchanged under permutations of its  $2n$  arguments. Hence the expression

$$E^* \Phi\left(\left\| \frac{1}{n} \sum_{i=1}^n e_i [f(X_i) - f(Y_i)] \right\|_{\mathcal{F}}\right)$$

is the same for any  $n$ -tuple  $(e_1, \dots, e_n) \in \{-1, 1\}^n$ . Deduce that

$$E^* \Phi(\|\mathbb{P}_n - P\|_{\mathcal{F}}) \leq E_{\varepsilon} E_{X,Y}^* \Phi\left(\left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i [f(X_i) - f(Y_i)] \right\|_{\mathcal{F}}\right).$$

Use the triangle inequality to separate the contributions of the  $X$ 's and the  $Y$ 's and next use the convexity of  $\Phi$  to bound the previous expression by

$$\frac{1}{2} E_{\varepsilon} E_{X,Y}^* \Phi\left(2 \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}}\right) + \frac{1}{2} E_{\varepsilon} E_{X,Y}^* \Phi\left(2 \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(Y_i) \right\|_{\mathcal{F}}\right).$$

By perfectness of coordinate projections, the expectation  $E_{X,Y}^*$  is the same as  $E_X^*$  and  $E_Y^*$  in the two terms, respectively. Finally, replace the repeated outer expectations by a joint outer expectation. ■

The symmetrization lemma is valid for any class  $\mathcal{F}$ . In the proofs of Glivenko-Cantelli and Donsker theorems, it will be applied not only to the original set of functions of interest, but also to several classes constructed from such a set  $\mathcal{F}$  (such as the class  $\mathcal{F}_{\delta}$  of small differences). The next step in these proofs is to apply a maximal inequality to the right side of the lemma, conditionally on  $X_1, \dots, X_n$ . At that point we need to write the joint outer expectation as the repeated expectation  $E_X^* E_{\varepsilon}$ , where the indices  $X$  and  $\varepsilon$  mean expectation over  $X$  and  $\varepsilon$ , respectively, conditionally on the remaining variables. Unfortunately, Fubini's theorem is not valid for outer expectations. To overcome this problem, it is assumed that the integrand in the right side of the lemma is jointly measurable in  $(X_1, \dots, X_n, \varepsilon_1, \dots, \varepsilon_n)$ . Since the Rademacher variables are discrete, this is the case if and only if the maps

$$(2.3.2) \quad (X_1, \dots, X_n) \mapsto \left\| \sum_{i=1}^n e_i f(X_i) \right\|_{\mathcal{F}}$$

are measurable for every  $n$ -tuple  $(e_1, \dots, e_n) \in \{-1, 1\}^n$ . For the intended application of Fubini's theorem, it suffices that this is the case for the completion of  $(\mathcal{X}^n, \mathcal{A}^n, P^n)$ .

**2.3.3 Definition (Measurable class).** A class  $\mathcal{F}$  of measurable functions  $f: \mathcal{X} \mapsto \mathbb{R}$  on a probability space  $(\mathcal{X}, \mathcal{A}, P)$  is called a *P-measurable class* if the function (2.3.2) is measurable on the completion of  $(\mathcal{X}^n, \mathcal{A}^n, P^n)$  for every  $n$  and every vector  $(e_1, \dots, e_n) \in \mathbb{R}^n$ .

In the following, one cannot dispense with measurability conditions of some form. Examples show that the law of large numbers or the central limit theorem may fail only because of the violation of measurability conditions such as the one just introduced. On the other hand, the measurability conditions in the present form are not necessary, but they are suggested by the methods of proof. A trivial, but nevertheless useful, device to relax measurability requirements a little is to replace the class  $\mathcal{F}$  by a class  $\mathcal{G}$  for which measurability can be checked. If  $\|\mathbb{P}_n - P\|_{\mathcal{F}} = \|\mathbb{P}_n - P\|_{\mathcal{G}}$  inner almost surely, then a law of large numbers for  $\mathcal{G}$  clearly implies one for  $\mathcal{F}$ . Similarly, if  $\|\mathbb{P}_n - P\|_{\mathcal{F}_{\delta}} = \|\mathbb{P}_n - P\|_{\mathcal{G}_{\delta}}$  inner almost surely for every  $\delta > 0$ , then asymptotic equicontinuity of the  $\mathcal{F}$ -indexed empirical process follows from the same property for  $\mathcal{G}$ . Furthermore, if in both cases  $\mathcal{G}$  is chosen to be a subclass of  $\mathcal{F}$ , then the other conditions for the uniform law or the central limit theorem typically carry over from  $\mathcal{F}$  onto  $\mathcal{G}$ . Rather than formalizing this principle into a concept such as “nearly measurable class,” we will state the conditions directly in terms of  $\mathcal{F}$ .

**2.3.4 Example (Pointwise measurable classes).** Suppose  $\mathcal{F}$  contains a countable subset  $\mathcal{G}$  such that for every  $f \in \mathcal{F}$  there exists a sequence  $g_m$  in  $\mathcal{G}$  with  $g_m(x) \rightarrow f(x)$  for every  $x$ . Then  $\mathcal{F}$  is *P-measurable* for every  $P$ .

This claim is immediate from the fact that  $\|\sum e_i f(X_i)\|_{\mathcal{F}}$  equals  $\|\sum e_i f(X_i)\|_{\mathcal{G}}$ .

Some examples of this situation are the collection of indicators of cells in Euclidean space, the collection of indicators of balls, and collections of functions that are separable for the supremum norm.

In Section 2.3.3 it will be seen that every separable collection  $\mathcal{F} \subset L_2(P)$  has a “version” that is “almost” pointwise separable. (The name *pointwise separable class* will be reserved for these versions.)

**2.3.5 Example.** Suppose  $\mathcal{F}$  is a Suslin topological space (with respect to an arbitrary topology) and the map  $(x, f) \mapsto f(x)$  is jointly measurable on  $\mathcal{X} \times \mathcal{F}$  for the product  $\sigma$ -field of  $\mathcal{A}$  and the Borel  $\sigma$ -field. [Recall that every Borel subset (even analytic subset) of a Polish space is Suslin for the relative topology.] Then  $\mathcal{F}$  is *P-measurable* for every probability measure  $P$ .

This follows from Example 1.7.5 upon noting that the conditions imply that the process  $(x_1, \dots, x_n, f) \mapsto \sum_{i=1}^n e_i f(x_i)$  is jointly measurable on  $\mathcal{X}^n \times \mathcal{F}$ .

This “measurable Suslin condition” replaces the original problem of showing  $P$ -measurability of  $\mathcal{F}$  by the problem of finding a suitable topology on  $\mathcal{F}$  for which the map  $(x, f) \mapsto f(x)$  is measurable. This is sometimes easy – for instance, if the class  $\mathcal{F}$  is indexed by a parameter in a nice manner – but sometimes as hard as the original problem.

### \* 2.3.2 More Symmetrization

This section continues with additional results on symmetrization. These are not needed for the main results of this part (Chapters 2.2, 2.4, and 2.5), and the reader may wish to skip the remainder of this chapter at first reading.

For future reference it is convenient to generalize the notation. Instead of the empirical distribution, consider sums  $\sum_{i=1}^n Z_i$  of independent stochastic processes  $\{Z_i(f): f \in \mathcal{F}\}$ . Again the processes need not possess any further measurability properties besides measurability of all marginals  $Z_i(f)$ . However, for computing outer expectations, it will be understood that the underlying probability space is a product space  $\prod_{i=1}^n (\mathcal{X}_i, \mathcal{A}_i, P_i) \times (\mathcal{Z}, \mathcal{C}, Q)$  and each  $Z_i$  is a function of the  $i$ th coordinate of  $(x, z) = (x_1, \dots, x_n, z)$  only. For i.i.d. stochastic processes, it is understood in addition that the spaces  $(\mathcal{X}_i, \mathcal{A}_i, P_i)$  as well as the maps  $Z_i(f)$  defined on them are identical. Additional (independent) Rademacher or other variables are understood to be functions of the  $(n+1)$ st coordinate  $z$  only. The empirical distribution corresponds to taking  $Z_i(f) = f(X_i) - Pf$ .

First consider “desymmetrization.” The inequality of the preceding lemma can only be reversed (up to constants) for classes  $\mathcal{F}$  that are centered (Problem 2.3.1). The following lemma shows that the maxima of the processes  $\mathbb{P}_n - P$  and  $n^{-1} \sum \varepsilon_i(\delta_{X_i} - P)$  are fully comparable.

**2.3.6 Lemma.** *Let  $Z_1, \dots, Z_n$  be independent stochastic processes with mean zero. Then*

$$\mathbb{E}^* \Phi \left( \frac{1}{2} \left\| \sum_{i=1}^n \varepsilon_i Z_i \right\|_{\mathcal{F}} \right) \leq \mathbb{E}^* \Phi \left( \left\| \sum_{i=1}^n Z_i \right\|_{\mathcal{F}} \right) \leq \mathbb{E}^* \Phi \left( 2 \left\| \sum_{i=1}^n \varepsilon_i (Z_i - \mu_i) \right\|_{\mathcal{F}} \right),$$

for every nondecreasing, convex  $\Phi: \mathbb{R} \mapsto \mathbb{R}$  and arbitrary functions  $\mu_i: \mathcal{F} \mapsto \mathbb{R}$ .

**Proof.** The inequality on the right follows by similar arguments as in the proof of Lemma 2.3.1. For the inequality on the left, let  $Y_1, \dots, Y_n$  be an independent copy of  $Z_1, \dots, Z_n$  (suitably defined on the probability space

---

\* This section may be skipped at first reading.

$\prod_{i=1}^n (\mathcal{X}_i, \mathcal{A}_i, P_i) \times (\mathcal{Z}, \mathcal{C}, Q) \times \prod_{i=1}^n (\mathcal{X}_i, \mathcal{A}_i, P_i)$  and depending on the last  $n$  coordinates exactly as  $Z_1, \dots, Z_n$  depend on the first  $n$  coordinates). Since  $EY_i(f) = 0$ , the left side of the lemma is an average of expressions of the type

$$E_Z^* \Phi \left( \left\| \frac{1}{2} \sum_{i=1}^n e_i [Z_i(f) - EY_i(f)] \right\|_{\mathcal{F}} \right),$$

where  $(e_1, \dots, e_n)$  ranges over  $\{-1, 1\}^n$ . (cf. Lemma 1.2.7). By Jensen's inequality, this expression is bounded above by

$$E_{Z,Y}^* \Phi \left( \left\| \frac{1}{2} \sum_{i=1}^n e_i [Z_i(f) - Y_i(f)] \right\|_{\mathcal{F}} \right) = E_{Z,Y}^* \Phi \left( \left\| \frac{1}{2} \sum_{i=1}^n [Z_i(f) - Y_i(f)] \right\|_{\mathcal{F}} \right).$$

Finally, apply the triangle inequality and convexity of  $\Phi$ . ■

The most important choice for  $\Phi$  in the preceding lemmas is  $\Phi(x) = x$ . The requirement that  $\Phi$  be convex rules out  $\Phi(x) = 1\{x > a\}$ . Nevertheless, there is an analogous symmetrization inequality for probabilities.

**2.3.7 Lemma (Symmetrization for probabilities).** *For arbitrary stochastic processes  $Z_1, \dots, Z_n$  and arbitrary functions  $\mu_1, \dots, \mu_n: \mathcal{F} \mapsto \mathbb{R}$ ,*

$$\beta_n(x) P^* \left( \left\| \sum_{i=1}^n Z_i \right\|_{\mathcal{F}} > x \right) \leq 2P^* \left( 4 \left\| \sum_{i=1}^n \varepsilon_i (Z_i - \mu_i) \right\|_{\mathcal{F}} > x \right),$$

for every  $x > 0$  and  $\beta_n(x) \leq \inf_f P(|\sum_{i=1}^n Z_i(f)| < x/2)$ . In particular, this is true for i.i.d. mean-zero processes, and  $\beta_n(x) = 1 - (4n/x^2) \sup_f \text{var } Z_1(f)$ . Here the outer expectations are computed as indicated previously.

**Proof.** Let  $Y_1, \dots, Y_n$  be an independent copy of  $Z_1, \dots, Z_n$ , suitably defined on a product space as previously. If  $\|\sum_{i=1}^n Z_i\|_{\mathcal{F}} > x$ , then there is certainly some  $f$  for which  $|\sum_{i=1}^n Z_i(f)| > x$ . Fix a realization,  $Z_1, \dots, Z_n$  and  $f$  for which both are the case. For this fixed realization,

$$\begin{aligned} \beta &\leq P_Y^* \left( \left| \sum_{i=1}^n Y_i(f) \right| < \frac{x}{2} \right) \leq P_Y^* \left( \left| \sum_{i=1}^n Y_i(f) - \sum_{i=1}^n Z_i(f) \right| > \frac{x}{2} \right) \\ &\leq P_Y^* \left( \left\| \sum_{i=1}^n (Y_i - Z_i) \right\|_{\mathcal{F}} > \frac{x}{2} \right). \end{aligned}$$

The far left and far right sides do not depend on the particular  $f$ , and the inequality between them is valid on the set  $\{\|\sum_{i=1}^n Z_i\|_{\mathcal{F}} > x\}$ . Integrate the two sides out with respect to  $Z_1, \dots, Z_n$  over this set to obtain

$$\beta P^* \left( \left\| \sum_{i=1}^n Z_i \right\|_{\mathcal{F}} > x \right) \leq P_Z^* P_Y^* \left( \left\| \sum_{i=1}^n (Y_i - Z_i) \right\|_{\mathcal{F}} > \frac{x}{2} \right).$$

By symmetry, the right side equals

$$E_\varepsilon P_Z^* P_Y^* \left( \left\| \sum_{i=1}^n \varepsilon_i (Y_i - Z_i) \right\|_{\mathcal{F}} > \frac{x}{2} \right).$$

In view of the triangle inequality, this expression is not bigger than  $2P^* \left( \left\| \sum_{i=1}^n \varepsilon_i (Y_i - \mu_i) \right\|_{\mathcal{F}} > x/4 \right)$ .

i.i.d. processes  $Z_1, \dots, Z_n$  with mean zero satisfy the condition for the given  $\beta$  in view of Chebyshev's inequality. ■

Since they only add in signs, Rademacher variables appear to yield an efficient method of symmetrization. Nevertheless, in most arguments they can be replaced by other variables; for instance, standard normal ones. Then the form of the symmetrization inequalities may become more complicated. A discussion is deferred to Chapter 2.9, where it is shown in general that the sequences of processes  $n^{-1/2} \sum_{i=1}^n \xi_i Z_i$  and  $n^{-1/2} \sum_{i=1}^n Z_i$  possess the same asymptotic behavior for most choices of  $\xi_i$ .

Inequalities in terms of the means of suprema of processes are easier to handle than inequalities in terms of probabilities. However, so far the main condition (2.1.8) for the empirical central limit theorem has been stated in terms of probabilities. In view of Markov's inequality, the condition

$$(2.3.8) \quad E^* \sqrt{n} \|P_n - P\|_{\mathcal{F}_{\delta_n}} \rightarrow 0, \quad \text{for every } \delta_n \rightarrow 0,$$

is sufficient for asymptotic equicontinuity (2.1.8). In view of Lemma 2.3.6, this condition is equivalent to the analogous condition in terms of the symmetrized empirical measure  $P_n^o$ . It is therefore of interest that there is no loss of generality in passing from probabilities (2.1.8) to expectations.

For the proof we need a preparatory lemma.

**2.3.9 Lemma.** *Let  $Z_1, Z_2, \dots$  be i.i.d stochastic processes such that  $n^{-1/2} \sum_{i=1}^n Z_i$  converges weakly in  $\ell^\infty(\mathcal{F})$  to a tight Gaussian process. Then*

$$\lim_{x \rightarrow \infty} x^2 \sup_n P^* \left( \frac{1}{\sqrt{n}} \left\| \sum_{i=1}^n Z_i \right\|_{\mathcal{F}} > x \right) = 0.$$

In particular, the random variable  $\|Z_1\|_{\mathcal{F}}^*$  possesses a weak second moment.

**Proof.** Let  $Y_1, \dots, Y_n$  be an independent copy of  $Z_1, \dots, Z_n$ , for the benefit of the outer expectations defined as coordinate projections on an extra factor of a product probability space, as usual. Marginal convergence to a Gaussian process implies that  $E Z_i(f) = 0$  and  $\text{var } Z_i(f) < \infty$  for every  $f$ . Furthermore, the existence of a tight limit ensures that  $\mathcal{F}$  is totally bounded for the semimetric  $\rho(f, g) = \sigma(Z(f) - Z(g))$ . In particular,

$\alpha = \sup_{f \in \mathcal{F}} \text{var } Z(f)$  is finite. By an intermediate step in the proof of the symmetrization Lemma 2.3.7,

$$\left(1 - \frac{4\alpha^2}{x^2}\right) P\left(\frac{1}{\sqrt{n}} \left\| \sum_{i=1}^n Z_i \right\|_{\mathcal{F}}^* > x\right) \leq P\left(\frac{1}{\sqrt{n}} \left\| \sum_{i=1}^n (Z_i - Y_i) \right\|_{\mathcal{F}}^* > \frac{x}{2}\right).$$

It suffices to prove the lemma for  $Z_i - Y_i$  instead of  $Z_i$ . For convenience of notation, suppose that the original  $Z_i$  are symmetric.

Fix  $\varepsilon > 0$ . The norm  $\|\mathbb{G}\|_{\mathcal{F}}$  of the Gaussian limit variable  $\mathbb{G}$  has moments of all orders (see Proposition A.2.3). Thus there exists  $x$  such that  $P(\|\mathbb{G}\|_{\mathcal{F}} \geq x) \leq \varepsilon/x^2 \leq 1/8$ . By the portmanteau theorem, there exists  $N$  such that, for all  $n \geq N$ ,

$$(2.3.10) \quad P^*\left(\left\| \sum_{i=1}^n Z_i \right\|_{\mathcal{F}}^* > x\sqrt{n}\right) \leq 2P(\|\mathbb{G}\|_{\mathcal{F}} \geq x) \leq \frac{2\varepsilon}{x^2}.$$

By Lévy's inequalities A.1.2,

$$P\left(\max_{1 \leq i \leq n} \|Z_i\|_{\mathcal{F}}^* > x\sqrt{n}\right) \leq 2P\left(\left\| \sum_{i=1}^n Z_i \right\|_{\mathcal{F}}^* > x\sqrt{n}\right) \leq \frac{4\varepsilon}{x^2}.$$

In view of Problem 2.3.2, for every  $n \geq N$ ,

$$x^2 n P(\|Z_1\|_{\mathcal{F}} > x\sqrt{n}) \leq 8\varepsilon.$$

This immediately implies the second assertion of the lemma.

To obtain the first assertion, apply the preceding argument to the processes  $Z_i = m^{-1/2} \sum_{j=1}^m Z_{i,j}$  for each  $1 \leq i \leq n$ , where  $Z_{1,1}, \dots, Z_{n,m}$  are i.i.d. copies of  $Z_1$ . The sequence  $n^{-1/2} \sum_{i=1}^n Z_i = (nm)^{-1/2} \sum_{i=1}^n \sum_{j=1}^m Z_{i,j}$  converges to the same limit  $\mathbb{G}$  for each  $m$  as  $n \rightarrow \infty$ . Since the convergence is “accelerated,” there exists  $N$  such that (2.3.10) holds for  $n \geq N$  and all  $m$ , where  $x$  can be chosen dependent on  $\varepsilon$  and  $\mathbb{G}$  only, as before. In other words, there exists  $x$  and  $N$  such that for  $n \geq N$ ,

$$x^2 n P\left(\frac{1}{\sqrt{m}} \left\| \sum_{j=1}^m Z_{1,j} \right\|_{\mathcal{F}}^* > x\sqrt{n}\right) \leq 8\varepsilon,$$

for every  $m$ . This implies the lemma. ■

If a distribution on the real line has tails of the order  $o(x^{-2})$ , then all its moments of order  $0 < r < 2$  exist. One interesting consequence of the preceding lemma is that the sequence of moments  $E^* \|n^{-1/2} \sum_{i=1}^n Z_i\|_{\mathcal{F}}^r$  is uniformly bounded for every  $0 < r < 2$ . Given the weak convergence, this gives that the sequence  $E^* \|n^{-1/2} \sum_{i=1}^n Z_i\|_{\mathcal{F}}^r$  converges to the corresponding moment of the limit variable.

**2.3.11 Lemma.** Let  $Z_1, Z_2, \dots$  be i.i.d. stochastic processes, linear in  $f$ . Set  $\rho_Z(f, g) = \sigma(Z_1(f) - Z_1(g))$  and  $\mathcal{F}_\delta = \{f - g: \rho_Z(f, g) < \delta\}$ . Then the following statements are equivalent:

- (i)  $n^{-1/2} \sum_{i=1}^n Z_i$  converges weakly to a tight limit in  $\ell^\infty(\mathcal{F})$ ;
- (ii)  $(\mathcal{F}, \rho_Z)$  is totally bounded and  $\|n^{-1/2} \sum_{i=1}^n Z_i\|_{\mathcal{F}_{\delta_n}} \xrightarrow{\text{P}^*} 0$  for every  $\delta_n \downarrow 0$ ;
- (iii)  $(\mathcal{F}, \rho_Z)$  is totally bounded and  $\mathbb{E}^* \|n^{-1/2} \sum_{i=1}^n Z_i\|_{\mathcal{F}_{\delta_n}} \rightarrow 0$  for every  $\delta_n \downarrow 0$ .

These conditions imply that the sequence  $\mathbb{E}^* \|n^{-1/2} \sum_{i=1}^n Z_i\|_{\mathcal{F}}^r$  converges to  $\mathbb{E}\|\mathbb{G}\|_{\mathcal{F}}^r$  for every  $0 < r < 2$ .

**Proof.** The equivalence of (i) and (ii) is clear from the general results on weak convergence in  $\ell^\infty(\mathcal{F})$  obtained in Chapter 1.5. Condition (iii) implies (ii) by Markov's inequality.

Suppose (i) and (ii) hold. Then  $\text{P}^*(\|Z_1\|_{\mathcal{F}} > x) = o(x^{-2})$  as  $x$  tends to infinity by Lemma 2.3.9. This implies that

$$\mathbb{E}^* \max_{1 \leq i \leq n} \frac{\|Z_i\|_{\mathcal{F}}}{\sqrt{n}} \rightarrow 0$$

(Problem 2.3.3). In view of the triangle inequality, the same is true with  $\mathcal{F}$  replaced by  $\mathcal{F}_{\delta_n}$ . Convergence to zero in probability of  $\|n^{-1/2} \sum_{i=1}^n Z_i\|_{\mathcal{F}_{\delta_n}}$  implies pointwise convergence to zero of the sequence of their quantile functions. Apply Hoffmann-Jørgensen's inequality (Proposition A.1.5) to obtain (iii).

The last assertion follows from the remark after Lemma 2.3.9. ■

To recover the empirical process, set  $Z_i(f) = f(X_i) - Pf$ .

**2.3.12 Corollary.** Let  $\mathcal{F}$  be a class of measurable functions. Then the following are equivalent:

- (i)  $\mathcal{F}$  is  $P$ -Donsker;
- (ii)  $(\mathcal{F}, \rho_P)$  is totally bounded and (2.1.8) holds;
- (iii)  $(\mathcal{F}, \rho_P)$  is totally bounded and (2.3.8) holds.

**2.3.13 Corollary.** Every  $P$ -Donsker class  $\mathcal{F}$  satisfies  $P(\|f - Pf\|_{\mathcal{F}}^* > x) = o(x^{-2})$  as  $x$  tends to infinity. Consequently, if  $\|Pf\|_{\mathcal{F}} < \infty$ , then  $\mathcal{F}$  possesses an envelope function  $F$  with  $P(F > x) = o(x^{-2})$ .

### \* 2.3.3 Separable Versions

Let  $(\mathcal{F}, \rho)$  be a given separable, semimetric space. A stochastic process  $\{\mathbb{G}(f, \omega): f \in \mathcal{F}\}$  is called *separable* if there exists a null set  $N$  and a countable subset  $\mathcal{G} \subset \mathcal{F}$  such that, for all  $\omega \notin N$  and  $f \in \mathcal{F}$ , there exists a

---

\* This section may be skipped at first reading.

sequence  $g_m$  in  $\mathcal{G}$  with  $g_m \rightarrow f$  and  $\mathbb{G}(g_m, \omega) \rightarrow \mathbb{G}(f, \omega)$ . A stochastic process  $\{\tilde{\mathbb{G}}(f): f \in \mathcal{F}\}$  is a *separable version* of a given process  $\{\mathbb{G}(f): f \in \mathcal{F}\}$  if  $\tilde{\mathbb{G}}$  is separable and  $\tilde{\mathbb{G}}(f) = \mathbb{G}(f)$  almost surely for every  $f$ . (The two processes must be defined on the same probability space, and the exceptional null sets may depend on  $f$ .)

It is well known that every stochastic process possesses a separable version (possibly taking values in the extended real line). By its definition, a separable process has excellent measurability properties. In this subsection it is shown that an empirical process possesses a separable version that is itself an empirical process. This fact is used in Chapter 2.10 (and there only) to remove unnecessary measurability conditions from a theorem.

For any separable process  $\tilde{\mathbb{G}}$  with countable “separant”  $\mathcal{G}$ ,

$$\sup_{\substack{\rho(f,g) < \delta \\ f,g \in \mathcal{F}}} |\tilde{\mathbb{G}}(f) - \tilde{\mathbb{G}}(g)| = \sup_{\substack{\rho(f,g) < \delta \\ f,g \in \mathcal{G}}} |\tilde{\mathbb{G}}(f) - \tilde{\mathbb{G}}(g)|, \quad \text{a.s.}$$

If  $\tilde{\mathbb{G}}$  is a version of  $\mathbb{G}$ , then the countability of  $\mathcal{G}$  implies that the right side changes at most on a null set if  $\tilde{\mathbb{G}}$  is replaced by  $\mathbb{G}$ . Next the expression increases if  $\mathcal{G}$  is replaced by  $\mathcal{F}$ . Thus, for every separable version  $\tilde{\mathbb{G}}$  of a process  $\mathbb{G}$ ,

$$\sup_{\substack{\rho(f,g) < \delta \\ f,g \in \mathcal{F}}} |\tilde{\mathbb{G}}(f) - \tilde{\mathbb{G}}(g)| \leq \sup_{\substack{\rho(f,g) < \delta \\ f,g \in \mathcal{G}}} |\mathbb{G}(f) - \mathbb{G}(g)|, \quad \text{a.s.}$$

It follows that the modulus of continuity of a process decreases when replacing it by a separable version. Consequently, asymptotic equicontinuity of a sequence of processes implies the same for any sequence of separable versions. Since the marginal distributions of a process do not change when passing to a separable version, this yields the following lemma.

**2.3.14 Lemma.** *Let  $\tilde{\mathbb{G}}_n$  be separable versions of a sequence of stochastic processes  $\mathbb{G}_n$  with sample paths in  $\ell^\infty(\mathcal{F})$ . Then the sequence  $\mathbb{G}_n$  converges in distribution to a tight limit in  $\ell^\infty(\mathcal{F})$  if and only if the sequence  $\tilde{\mathbb{G}}_n$  converges to a tight limit and  $\mathbb{G}_n - \tilde{\mathbb{G}}_n$  converges to zero in outer probability in  $\ell^\infty(\mathcal{F})$ .*

Empirical processes allow a special type of separable version, constructed from a separable version of the index class  $\mathcal{F}$ . Let  $\mathcal{F}$  be a class of square integrable measurable functions  $f: \mathcal{X} \mapsto \mathbb{R}$  on a probability space  $(\mathcal{X}, \mathcal{A}, P)$ . Call  $\mathcal{F}$  a *pointwise separable class* if there exists a countable subset  $\mathcal{G} \subset \mathcal{F}$  and for each  $n$  a null set  $N_n \subset \mathcal{X}^n$  (for  $P^n$ ) such that, for all  $(x_1, \dots, x_n) \notin N_n$  and  $f \in \mathcal{F}$ , there exists a sequence  $g_m$  in  $\mathcal{G}$  such that  $g_m \rightarrow f$  in  $L_2(P)$  and  $(g_m(x_1), \dots, g_m(x_n)) \rightarrow (f(x_1), \dots, f(x_n))$ . A class  $\tilde{\mathcal{F}}$  of measurable functions is a *pointwise separable version* of  $\mathcal{F}$  if  $\tilde{\mathcal{F}}$  is pointwise separable and there exists a bijection  $f \leftrightarrow \tilde{f}$  between  $\mathcal{F}$  and  $\tilde{\mathcal{F}}$  such that  $f = \tilde{f}$  almost surely.

Let  $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$  be the empirical process indexed by the class of functions  $\mathcal{F}$ . If  $\tilde{\mathcal{F}}$  is a pointwise separable version of  $\mathcal{F}$ , then

$$\tilde{\mathbb{G}}_n(f) = \mathbb{G}_n(\tilde{f})$$

defines a separable version of  $\mathbb{G}_n$ , for every  $n$ . This separable version is itself an empirical process, and its index class is obtained by changing each element of the original index class on a null set. Furthermore, the difference process  $\mathbb{G}_n - \tilde{\mathbb{G}}_n$  is the empirical process indexed by the class  $\{f - \tilde{f}: f \in \mathcal{F}\}$ . Each of the functions in this class is zero almost surely. It turns out that the empirical process indexed by a zero class converges in probability to zero if and only if the same is true for the class of absolute values. This yields the following theorem.

**2.3.15 Theorem.** *Let  $\tilde{\mathcal{F}}$  be a pointwise separable version of a class of square integrable measurable functions  $\mathcal{F}$ . Then  $\mathcal{F}$  is Donsker if and only if  $\tilde{\mathcal{F}}$  is Donsker and  $\sup_{f \in \mathcal{F}} \sqrt{n} \mathbb{P}_n |f - \tilde{f}| \rightarrow 0$  in outer probability.*

**Proof.** By the preceding lemma,  $\mathcal{F}$  is Donsker if and only if  $\tilde{\mathcal{F}}$  is Donsker and  $\sqrt{n} \|\mathbb{P}_n(f - \tilde{f})\|_{\mathcal{F}} \rightarrow 0$  in outer probability. It suffices to show that for a class of measurable functions  $f: \mathcal{X} \mapsto \mathbb{R}$  such that  $f = 0$  almost surely, for every  $f$ , the two statements  $\sqrt{n} \|\mathbb{P}_n f\|_{\mathcal{F}} \xrightarrow{P_*} 0$  and  $\sqrt{n} \|\mathbb{P}_n |f|\|_{\mathcal{F}} \xrightarrow{P_*} 0$  are equivalent. It is clear that the second statement implies the first.

For a proof in the other direction, fix  $\varepsilon > 0$  and define  $U = \{x \in \mathcal{X}^n: \|\mathbb{P}_n\|_{\mathcal{F}} \leq \varepsilon\}_*$ . It will be shown that there exists a measurable set  $U' \subset U$  of the same measure as  $U$  such that, for all  $x \in U'$  and every subset  $I \subset \{1, 2, \dots, n\}$ ,

$$\frac{1}{n} \left| \sum_{i \in I} f(x_i) \right| \leq \varepsilon.$$

Then it follows that  $P_*(\|\mathbb{P}_n\|_{\mathcal{F}} \leq 2\varepsilon) \geq P_*(\|\mathbb{P}_n\|_{\mathcal{F}} \leq \varepsilon)$ , and the desired result follows.

For  $I \subset \{1, 2, \dots, n\}$ , identify  $x \in \mathcal{X}^n$  with  $(\alpha, \beta)$ , where  $\alpha \in \mathcal{X}^I$  are the coordinates  $x_i$  with  $i \in I$  and  $\beta \in \mathcal{X}^{n-I}$  are the remaining coordinates. Let  $U_\alpha \subset \mathcal{X}^n$  be the “vertical” section  $\{\alpha\} \times \{\beta: (\alpha, \beta) \in U\}$  of  $U$  at “width”  $\alpha$ , and let  $\pi U_\alpha$  be the projection of  $U_\alpha$  on  $\mathcal{X}^{n-I}$  (so that  $U_\alpha = \{\alpha\} \times \pi U_\alpha$ ). Let  $V_I$  be the union of all sections  $U_\alpha$  with projection  $\pi U_\alpha$  equal to a null set (under  $P^{n-I}$ ). By Fubini’s theorem,  $P^n(V_I) = 0$ . Set

$$U' = U - \bigcup_I V_I,$$

where  $I$  ranges over all nonempty, proper subsets of  $\{1, \dots, n\}$ .

Fix  $I$  and  $f$ . Each  $x \in U'$  can be identified with a pair  $(\alpha, \beta)$  as before. By construction  $P^{n-I}(\pi U_\alpha) > 0$  and by assumption  $(f(\beta'_1), \dots, f(\beta'_{n-I})) =$

0 almost surely under  $P^{n-I}$ . Conclude that there must be a  $\beta' \in \pi U_\alpha$  with  $(f(\beta'_1), \dots, f(\beta'_{n-I})) = 0$ . Thus

$$\frac{1}{n} \left| \sum_{i \in I} f(x_i) \right| = \frac{1}{n} \left| \sum_{i \in I} f(\alpha_i) + \sum_{i \notin I} f(\beta'_i) \right| \leq \varepsilon,$$

because  $(\alpha, \beta') \in U$ . This concludes the proof. ■

The preceding theorem shows that the difference between  $\mathbb{G}_n$  and  $\tilde{\mathbb{G}}_n$  must be small due to the fact that the differences  $f - \tilde{f}$  are small: the difference cannot be small through the cancellation of positive terms by negative terms.

A pointwise separable version of a given class of functions can be constructed from a special type of lifting of  $L_\infty(\mathcal{X}, \mathcal{A}, P)$ . A *lifting* is a map  $\rho: L_\infty(\mathcal{X}, \mathcal{A}, P) \mapsto \mathcal{L}_\infty(\mathcal{X}, \mathcal{A}, P)$  that assigns to each equivalence class  $f$  of essentially bounded functions a representative  $\rho f$  in such a way that

- (i)  $\rho 1 = 1$ ;
- (ii)  $\rho$  is positive:  $\rho f(x) \geq 0$  for all  $x$  if  $f \geq 0$  almost surely;
- (iii)  $\rho$  is linear;
- (iv)  $\rho$  is multiplicative:  $\rho(fg) = \rho f \rho g$ .

Any lifting automatically possesses the further property:

- (vi)  $\rho h(f_1, \dots, f_n) \geq h(\rho f_1, \dots, \rho f_n)$  for every lower semicontinuous  $h: \mathbb{R}^n \mapsto \mathbb{R}$  and  $f_1, \dots, f_n$  in  $\mathcal{L}_\infty(\mathcal{X}, \mathcal{A}, P)$  (Problem 2.3.8).

It is well known that every complete probability space  $(\mathcal{X}, \mathcal{A}, P)$  admits a lifting  $\rho$  of the space  $L_\infty(\mathcal{X}, \mathcal{A}, P)$ . For the construction of a separable version of the empirical process, a lifting with a further property is needed. This exists by the following proposition.

**2.3.16 Proposition (Consistent lifting).** *Any complete probability space  $(\mathcal{X}, \mathcal{A}, P)$  admits a lifting of the space  $L_\infty(\mathcal{X}, \mathcal{A}, P)$  such that for each  $n$  the map  $\rho_n$  given by the diagram*

$$\left( (x_1, \dots, x_n) \mapsto f(x_i) \right) \xrightarrow{\rho_n} \left( (x_1, \dots, x_n) \mapsto \rho f(x_i) \right)$$

*is the restriction [to the set of functions of the type  $(x_1, \dots, x_n) \mapsto f(x_i)$ , where  $f$  ranges over  $L_\infty(\mathcal{X}, \mathcal{A}, P)$ ] of a lifting of  $L_\infty(\mathcal{X}^n, \mathcal{A}^n, P^n)$  for each  $1 \leq i \leq n$ .*

This proposition is proved by Talagrand (1982), who calls a lifting  $\rho$  with the given property a *consistent lifting*. There also exist liftings that are not consistent.

Suppose  $\mathcal{F}$  is a class of measurable, uniformly bounded functions. The following theorem shows that a pointwise separable version  $\tilde{\mathcal{F}}$  of  $\mathcal{F}$  can be defined by

$$\tilde{f}(x) = \rho f(x)$$

for a consistent lifting  $\rho$  of  $L_\infty(\mathcal{X}, \mathcal{A}, P)$ . (Interpret  $\rho f$  as the image of the class of  $f$  under  $\rho$ .) This separable version inherits the nice properties of a lifting: the map  $f \mapsto \tilde{f}$  is positive, linear, and multiplicative.

If  $\mathcal{F}$  contains unbounded functions, then a pointwise separable version can be defined by

$$\tilde{f}(x) = \Phi^{-1} \circ \rho(\Phi \circ f)(x),$$

for a consistent lifting as before and a fixed, strictly monotone bijection of  $\bar{\mathbb{R}}$  onto a compact interval. This separable modification  $f \mapsto \tilde{f}$  lacks linearity and multiplicativity but is positive.

**2.3.17 Theorem (Separable modification).** *Let  $(\mathcal{X}, \mathcal{A}, P)$  be a complete probability space and  $\mathcal{F}$  be a class of measurable functions  $f: \mathcal{X} \mapsto \mathbb{R}$ . If  $\mathcal{F}$  is separable in  $L_2(P)$ , then there exists a pointwise separable version of  $\mathcal{F}$ .*

**Proof.** Define  $\tilde{f}$  as indicated, using a consistent lifting, so that the map  $\rho_n$  in the preceding proposition is a lifting. Then  $f = \tilde{f}$  almost surely, because  $\rho\Phi \circ f$  is a representative of the class of  $\Phi \circ f$ .

Fix  $n$  and open sets  $U \subset \mathcal{F}$  and  $G \subset \mathbb{R}^n$ . Define

$$A_f = \left\{ (x_1, \dots, x_n) \in \mathcal{X}^n : (\rho\Phi \circ f(x_1), \dots, \rho\Phi \circ f(x_n)) \in G \right\}.$$

For  $D \subset \mathcal{F}$  set  $A_D = \cup_{f \in D} A_f$ . There exists a countable subset  $D \subset U$  for which  $P^n(A_D)$  is maximal over all countable subsets  $D$  of  $U$ . Any maximizing  $D$  satisfies  $1_{A_f} \leq 1_{A_D}$  almost surely for all  $f \in U$ . By positivity of the lifting  $\rho_n$ , it follows that

$$\rho_n 1_{A_f}(x) \leq \rho_n 1_{A_D}(x), \quad \text{for every } x.$$

Set  $N_n = \{x \in \mathcal{X}^n : 1_{A_D}(x) = 0, \rho_n 1_{A_D}(x) = 1\}$ . By multiplicativity the image of an indicator function under a lifting is an indicator function. It follows that  $N_n$  can also be described as the set of  $x$  such that  $1_{A_D}(x) < \rho_n 1_{A_D}(x)$ . Let  $(\Phi \circ f)_i$  denote the function  $(x_1, \dots, x_n) \mapsto \Phi \circ f(x_i)$ . Since  $G$  is open, property (vi) of the lifting  $\rho_n$  gives

$$\rho_n 1_{A_f} = \rho_n 1_G((\Phi \circ f)_1, \dots, (\Phi \circ f)_n) \geq 1_G(\rho_n(\Phi \circ f)_1, \dots, \rho_n(\Phi \circ f)_n) = 1_{A_f}.$$

Combination of the last two displayed equations yields that  $\rho_n 1_{A_D}(x) \geq 1_{A_f}(x)$  for every  $x$ . If  $x \in A_f$  and  $x \notin N_n$ , then it follows that  $x \in A_D$ .

It has been proved that, for every open  $U \subset \mathcal{F}$  and open  $G \subset \mathbb{R}^n$ , there exists a countable set  $D_{U,G} \subset U$  and a null set  $N_{n,U,G} \subset \mathcal{X}^n$  such that

$$\begin{aligned} x \notin N_{n,U,G} \wedge (\rho\Phi \circ f(x_1), \dots, \rho\Phi \circ f(x_n)) \in G \\ \Rightarrow \exists g \in D_{U,G} : (\rho\Phi \circ g(x_1), \dots, \rho\Phi \circ g(x_n)) \in G. \end{aligned}$$

Repeat the construction for every  $U$  and  $G$  in countable bases for the open sets in  $\mathcal{F}$  and  $\mathbb{R}^n$ , respectively. Let  $N_n$  and  $\mathcal{G}$  be the union of all null

sets  $N_{n,U,G}$  and countable  $D_{U,G}$ . Then for every  $x \notin N_n$  and  $f \in \mathcal{F}$ , there exists a sequence  $g_m$  in  $\mathcal{G}$  with  $(\rho\Phi \circ g_m(x_1), \dots, \rho\Phi \circ g_m(x_n)) \mapsto (\rho\Phi \circ f(x_1), \dots, \rho\Phi \circ f(x_n))$ . Thus the class of functions  $\{\rho(\Phi \circ f) : f \in \mathcal{F}\}$  is pointwise separable. ■

## Problems and Complements

1. There is no universal constant  $K$  such that  $E|\sum \varepsilon_i X_i| \leq K E|\sum (X_i - EX_i)|$  for any i.i.d. random variables  $X_1, \dots, X_n$ .

[Hint: Take  $n = 1$  and  $X_1 \sim N(\mu, 1)$ .]

2. For independent random variables  $\xi_1, \dots, \xi_n$ ,

$$P\left(\max_i |\xi_i| > x\right) \geq \frac{\sum_i P(|\xi_i| > x)}{1 + \sum_i P(|\xi_i| > x)}.$$

In particular, if the left side is less than  $1/2$ , then  $2P(\max_i |\xi_i| > x) \geq \sum_i P(|\xi_i| > x)$ .

[Hint: For  $x \geq 0$ , one has  $1 - x \leq \exp(-x)$  and  $1 - e^{-x} \geq x/(1+x)$ .]

3. Let  $X_{n1}, \dots, X_{nn}$  be an arbitrary array of real random variables.

- (i) If  $\sup_{x > \varepsilon\sqrt{n}} n^{-1} \sum_{i=1}^n x^2 P(|X_{ni}| > x) \rightarrow 0$  for every  $\varepsilon > 0$ , then  $E \max_{1 \leq i \leq n} |X_{ni}| / \sqrt{n} \rightarrow 0$ .  
(ii) If the array is rowwise i.i.d., then  $\max_{1 \leq i \leq n} |X_{ni}| = o_P(\sqrt{n})$  if and only if  $P(|X_{ni}| > \varepsilon\sqrt{n}) = o(n^{-1})$  for every  $\varepsilon > 0$ .

If the variables in the triangular array are i.i.d., then all assertions are equivalent to  $P(|X_1| > x) = o(x^{-2})$ .

[Hint:  $E \max |X_{ni}| \{ |X_{ni}| > \varepsilon\sqrt{n} \}$  is bounded by  $\sum \int_0^\infty P(|X_{ni}| > \varepsilon\sqrt{n}) dt$ . This integral splits into two pieces. The integral over the second part  $[\varepsilon\sqrt{n}, \infty)$  can be bounded by  $\sup_{x > \varepsilon\sqrt{n}} n^{-1} \sum x^2 P(|X_{ni}| > x)$  times  $\int_{\varepsilon\sqrt{n}}^\infty x^{-2} dx$ .]

4. Let  $\xi_1, \dots, \xi_n$  be i.i.d. random variables.

- (i) Show that the following are equivalent.

- (a)  $P(|\xi_1| > x) = o(x^{-1})$ .  
(b)  $\max_{1 \leq i \leq n} |\xi_i|/n$  converges to zero in probability.  
(c)  $\max_{1 \leq i \leq n} |\xi_i|^r/n^r$  converges to zero in mean for every  $r < 1$ .

- (ii) Show that the following are equivalent:

- (a)  $E|\xi_1| < \infty$ .  
(b)  $\max_{1 \leq i \leq n} |\xi_i|/n$  converges to zero almost surely.  
(c)  $\max_{1 \leq i \leq n} |\xi_i|/n$  converges to zero in mean.

- (iii) Let  $r > 1$ . Use (i) and (ii) to show the following:

- (a) Show that  $P(|\xi_1| > x) = o(x^{-r})$  if and only if  $\max_{1 \leq i \leq n} |\xi_i|/n^{1/r}$  converges to zero in probability if and only if  $\max_{1 \leq i \leq n} |\xi_i|/n^{1/r}$  converges to zero in mean.

- (b) Show that  $E|\xi_1|^r < \infty$  if and only if  $\max_{1 \leq i \leq n} |\xi_i|^r/n$  converges to zero almost surely if and only if  $\max_{1 \leq i \leq n} |\xi_i|^r/n$  converges to zero in mean.

5. For  $r > 0$ , suppose that  $\{\xi_i\}$  is a finite sequence of positive independent random variables with  $E|\xi_i|^r < \infty$ , for all  $i$ . Let  $t_0 = \inf\{t > 0 : \sum_i P(\xi_i > t) \leq \lambda\}$ . Then

$$\frac{\lambda}{1+\lambda} t_0^r + \frac{1}{1+\lambda} \sum_i \int_{t_0}^{\infty} P(\xi_i > t) dt^r \leq E \max_i |\xi_i|^r \leq t_0^r + \sum_i \int_{t_0}^{\infty} P(\xi_i > t) dt^r.$$

[Hint: Use Problem 2.3.2.]

6. For i.i.d. random variables  $X_1, X_2, \dots$  with  $E|X_1| < \infty$  and any  $r < 1$ , we have  $E \sup_{n \geq 1} (|X_n|/n)^r < \infty$ .

[Hint: Use Problem 2.3.5.]

7. For i.i.d. random variables  $X_1, X_2, \dots$ , the expectation  $E \sup_{n \geq 1} (|X_n|/n)$  is finite if and only if  $E(|X_1| \log |X_1|)$  is finite. Here  $\log x = 1 \vee (\log x)$ .

[Hint: Use Problem 2.3.5.]

8. Any lifting  $\rho$  of  $L_\infty(\mathcal{X}, \mathcal{A}, P)$  has the following additional properties:

- (i)  $\|\rho f\|_\infty \leq \|f\|_\infty$ ;
- (ii)  $\rho h(f_1, \dots, f_n) = h(\rho f_1, \dots, \rho f_n)$  for every continuous function  $h: \mathbb{R}^n \mapsto \mathbb{R}$  and  $f_1, \dots, f_n$  in  $L_\infty(\mathcal{X}, \mathcal{A}, P)$ ;
- (iii)  $\rho h(f_1, \dots, f_n) \geq h(\rho f_1, \dots, \rho f_n)$  for every lower semicontinuous function  $h: \mathbb{R}^n \mapsto \mathbb{R}$  and  $f_1, \dots, f_n$  in  $L_\infty(\mathcal{X}, \mathcal{A}, P)$ ;
- (iv)  $\rho 1_A = 1_{\tilde{A}}$  for some measurable set  $A$ .

[Hint: For polynomials  $h$ , property (ii) is a consequence of linearity and multiplicativity of a lifting. By the Stone-Weierstrass theorem, every continuous  $h$  can be uniformly approximated by polynomials on any given compact subset of  $\mathbb{R}^n$ . Thus the general form of (ii) can be reduced to polynomials in view of (i).]

Property (iii) is a consequence of (ii), because every lower semicontinuous function can be approximated pointwise from below by a sequence of continuous functions.]

## 2.4

# Glivenko-Cantelli Theorems

In this chapter we prove two types of Glivenko-Cantelli theorems. The first theorem is the simplest and is based on entropy with bracketing. Its proof relies on finite approximation and the law of large numbers for real variables. The second theorem uses random  $L_1$ -entropy numbers and is proved through symmetrization followed by a maximal inequality.

Recall Definition 2.1.6 of the bracketing numbers of a class  $\mathcal{F}$  of functions.

**2.4.1 Theorem.** *Let  $\mathcal{F}$  be a class of measurable functions such that  $N_{[]}(\varepsilon, \mathcal{F}, L_1(P)) < \infty$  for every  $\varepsilon > 0$ . Then  $\mathcal{F}$  is Glivenko-Cantelli.*

**Proof.** Fix  $\varepsilon > 0$ . Choose finitely many  $\varepsilon$ -brackets  $[l_i, u_i]$  whose union contains  $\mathcal{F}$  and such that  $P(u_i - l_i) < \varepsilon$  for every  $i$ . Then, for every  $f \in \mathcal{F}$ , there is a bracket such that

$$(\mathbb{P}_n - P)f \leq (\mathbb{P}_n - P)u_i + P(u_i - f) \leq (\mathbb{P}_n - P)u_i + \varepsilon.$$

Consequently,

$$\sup_{f \in \mathcal{F}} (\mathbb{P}_n - P)f \leq \max_i (\mathbb{P}_n - P)u_i + \varepsilon.$$

The right side converges almost surely to  $\varepsilon$  by the strong law of large numbers for real variables. Combination with a similar argument for  $\inf_{f \in \mathcal{F}} (\mathbb{P}_n - P)f$  yields that  $\limsup \|\mathbb{P}_n - P\|_{\mathcal{F}}^* \leq \varepsilon$  almost surely, for every  $\varepsilon > 0$ . Take a sequence  $\varepsilon_m \downarrow 0$  to see that the limsup must actually be zero almost surely. ■

**2.4.2 Example.** The previous proof generalizes a well-known proof of the Glivenko-Cantelli theorem for the empirical distribution function on the real line. Indeed, the set of indicator functions of cells  $(-\infty, c]$  possesses finite bracketing numbers for any underlying distribution. Simply use the brackets  $[1\{(-\infty, t_i]\}, 1\{(-\infty, t_{i+1}]\}]$  for a grid of points  $-\infty = t_0 < t_1 < \dots < t_m = \infty$  with the property  $P(t_i, t_{i+1}) < \varepsilon$  for each  $i$ . Bracketing numbers of many other classes of functions are discussed in Chapter 2.7.

Both the statement and the proof of the following theorem are more complicated than the previous bracketing theorem. However, its sufficiency condition for the Glivenko-Cantelli property can be checked for many classes of functions by elegant combinatorial arguments, as discussed in a later chapter. Another important note: its random entropy condition is necessary, a fact that is not proved here.

**2.4.3 Theorem.** Let  $\mathcal{F}$  be a  $P$ -measurable class of measurable functions with envelope  $F$  such that  $P^*F < \infty$ . Let  $\mathcal{F}_M$  be the class of functions  $f 1\{F \leq M\}$  when  $f$  ranges over  $\mathcal{F}$ . If  $\log N(\varepsilon, \mathcal{F}_M, L_1(\mathbb{P}_n)) = o_P^*(n)$  for every  $\varepsilon$  and  $M > 0$ , then  $\|\mathbb{P}_n - P\|_{\mathcal{F}}^* \rightarrow 0$  both almost surely and in mean. In particular,  $\mathcal{F}$  is Glivenko-Cantelli.

**Proof.** By the symmetrization Lemma 2.3.1, measurability of the class  $\mathcal{F}$ , and Fubini's theorem,

$$\begin{aligned} E^* \|\mathbb{P}_n - P\|_{\mathcal{F}} &\leq 2E_X E_{\varepsilon} \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}} \\ &\leq 2E_X E_{\varepsilon} \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}_M} + 2P^*F\{F > M\}, \end{aligned}$$

by the triangle inequality, for every  $M > 0$ . For sufficiently large  $M$ , the last term is arbitrarily small. To prove convergence in mean, it suffices to show that the first term converges to zero for fixed  $M$ . Fix  $X_1, \dots, X_n$ . If  $\mathcal{G}$  is an  $\varepsilon$ -net in  $L_1(\mathbb{P}_n)$  over  $\mathcal{F}_M$ , then

$$(2.4.4) \quad E_{\varepsilon} \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}_M} \leq E_{\varepsilon} \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{G}} + \varepsilon.$$

The cardinality of  $\mathcal{G}$  can be chosen equal to  $N(\varepsilon, \mathcal{F}_M, L_1(\mathbb{P}_n))$ . Bound the  $L_1$ -norm on the right by the Orlicz-norm for  $\psi_2(x) = \exp(x^2) - 1$ , and use the maximal inequality Lemma 2.2.2 to find that the last expression does not exceed a multiple of

$$\sqrt{1 + \log N(\varepsilon, \mathcal{F}_M, L_1(\mathbb{P}_n))} \sup_{f \in \mathcal{G}} \left\| \frac{1}{n} \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\psi_2|X} + \varepsilon,$$

where the Orlicz norms  $\|\cdot\|_{\psi_2|X}$  are taken over  $\varepsilon_1, \dots, \varepsilon_n$  with  $X_1, \dots, X_n$  fixed. By Hoeffding's inequality, they can be bounded by  $\sqrt{6/n}(\mathbb{P}_n f^2)^{1/2}$ , which is less than  $\sqrt{6/n}M$ . Thus the last displayed expression is bounded by

$$\sqrt{1 + \log N(\varepsilon, \mathcal{F}_M, L_1(\mathbb{P}_n))} \sqrt{\frac{6}{n}} M + \varepsilon \xrightarrow{P^*} \varepsilon.$$

It has been shown that the left side of (2.4.4) converges to zero in probability. Since it is bounded by  $M$ , its expectation with respect to  $X_1, \dots, X_n$  converges to zero by the dominated convergence theorem.

This concludes the proof that  $\|\mathbb{P}_n - P\|_{\mathcal{F}}^* \rightarrow 0$  in mean. That it also converges almost surely follows from the fact that the sequence  $\|\mathbb{P}_n - P\|_{\mathcal{F}}^*$  is a reverse martingale with respect to a suitable filtration. See the following lemma. ■

**2.4.5 Lemma.** *Let  $\mathcal{F}$  be a class of measurable functions with envelope  $F$  such that  $P^*F < \infty$ . Define a filtration by letting  $\Sigma_n$  be the  $\sigma$ -field generated by all measurable functions  $h: \mathcal{X}^\infty \mapsto \mathbb{R}$  that are permutation-symmetric in their first  $n$  arguments. Then*

$$E\left(\|\mathbb{P}_n - P\|_{\mathcal{F}}^* | \Sigma_{n+1}\right) \geq \|\mathbb{P}_{n+1} - P\|_{\mathcal{F}}^*, \quad \text{a.s.}$$

Furthermore, there exist versions of the measurable cover functions  $\|\mathbb{P}_n - P\|_{\mathcal{F}}^*$  that are adapted to the filtration. Any such versions form a reverse submartingale and converge almost surely to a constant.

**Proof.** Assume without loss of generality that  $Pf = 0$  for every  $f$ . The function  $h: \mathcal{X}^n \mapsto \mathbb{R}$  given by  $h(x_1, \dots, x_n) = \|n^{-1} \sum_{i=1}^n f(x_i)\|_{\mathcal{F}}$  is permutation-symmetric. Its measurable cover  $h^*$  (for  $P^n$ ) can be chosen permutation-symmetric as well (Problem 2.4.4). Then  $h^*(X_1, \dots, X_n)$  is a version of  $\|\mathbb{P}_n\|_{\mathcal{F}}^*$  and is  $\Sigma_n$  measurable.

Let  $\mathbb{P}_n^i$  be the empirical measure of  $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_{n+1}$ . Then  $\mathbb{P}_{n+1}f = (n+1)^{-1} \sum_{i=1}^{n+1} \mathbb{P}_n^i f$  for every  $f$ , whence

$$\|\mathbb{P}_{n+1}\|_{\mathcal{F}}^* \leq \frac{1}{n+1} \sum_{i=1}^{n+1} \|\mathbb{P}_n^i\|_{\mathcal{F}}^*, \quad \text{a.s.}$$

This is true for any version of the measurable covers. Choose the left side  $\Sigma_{n+1}$ -measurable and take conditional expectations with respect to  $\Sigma_{n+1}$  to arrive at

$$\|\mathbb{P}_{n+1}\|_{\mathcal{F}}^* \leq \frac{1}{n+1} \sum_{i=1}^{n+1} E\left(\|\mathbb{P}_n^i\|_{\mathcal{F}}^* | \Sigma_{n+1}\right), \quad \text{a.s..}$$

Each term in the sum on the right side changes on at most a null set if  $\|\mathbb{P}_n^i\|_{\mathcal{F}}^*$  is replaced by another version. For  $h^*$  given in the first paragraph of the proof, choose the version  $h^*(X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_{n+1})$  for  $\|\mathbb{P}_n^i\|_{\mathcal{F}}^*$ .

Since the function  $h^*$  is permutation-symmetric, it follows by symmetry that the conditional expectations  $E(\|\mathbb{P}_n^i\|_{\mathcal{F}}^* | \Sigma_{n+1})$  do not depend on  $i$  (almost surely). Thus the right side of the display equals  $E(\|\mathbb{P}_n^{n+1}\|_{\mathcal{F}}^* | \Sigma_{n+1})$ . ■

The covering numbers of the class  $\mathcal{F}_M$  of truncated functions in the previous theorem are smaller than those of the original class  $\mathcal{F}$ . Thus the conditions  $P^*F < \infty$  and  $\log N(\varepsilon, \mathcal{F}, L_1(\mathbb{P}_n)) = o_{P^*}(n)$  are sufficient for  $\mathcal{F}$  to be Glivenko-Cantelli.

If  $\mathcal{F}$  has a measurable envelope with  $PF < \infty$ , then  $\mathbb{P}_n F = O(1)$  almost surely and the random entropy condition is equivalent to

$$\log N(\varepsilon \|F\|_{\mathbb{P}_n, 1}, \mathcal{F}, L_1(\mathbb{P}_n)) = o_{P^*}(n).$$

In Chapter 2.6 it is shown that the entropy in the left side is uniformly bounded by a constant for so-called Vapnik-Červonenkis classes  $\mathcal{F}$ . Hence any appropriately measurable Vapnik-Červonenkis class is Glivenko-Cantelli, provided its envelope function is integrable.

## Problems and Complements

1. **(Necessity of integrability of the envelope)** Suppose that for a class  $\mathcal{F}$  of measurable functions the empirical measure of an i.i.d. sample satisfies  $\|\mathbb{P}_n - P\|_{\mathcal{F}} \xrightarrow{\text{a.s.}} 0$ . Then  $P^*\|f - Pf\|_{\mathcal{F}} < \infty$ . Consequently, if  $\|P\|_{\mathcal{F}} < \infty$ , then  $P^*F < \infty$  for an envelope function  $F$ .

[Hint: Since  $(1/n)\|f(X_n) - Pf\|_{\mathcal{F}}$  is bounded above by  $\|\mathbb{P}_n - P\|_{\mathcal{F}} + (1 - 1/n)\|\mathbb{P}_{n-1} - P\|_{\mathcal{F}}$ , one has that  $P(\|f(X_n) - Pf\|_{\mathcal{F}}^* \geq n, \text{i.o.}) = 0$ . By the Borel-Cantelli lemma, it follows that the series  $\sum P(\|f(X_n) - Pf\|_{\mathcal{F}}^* \geq n)$  converges. This series is an upper bound for the given first moment, because the  $X_n$  are identically distributed.]

2. The  $L_r(Q)$ -entropy numbers of the class  $\mathcal{F}_M = \{f 1\{F \leq M\}: f \in \mathcal{F}\}$  are smaller than those of  $\mathcal{F}$  for any probability measure  $Q$  and for numbers  $M > 0$  and  $r \geq 1$ .

3. **(Stability of the Glivenko-Cantelli property)** If  $\mathcal{F}$ ,  $\mathcal{F}_1$ , and  $\mathcal{F}_2$  are Glivenko-Cantelli classes of functions, then

- (i)  $\{a_1 f_1 + a_2 f_2: f_i \in \mathcal{F}_i, |a_i| \leq 1\}$  is Glivenko-Cantelli;
- (ii)  $\mathcal{F}_1 \cup \mathcal{F}_2$  is Glivenko-Cantelli;
- (iii) the class of all functions that are both the pointwise limit and the  $L_1(P)$ -limit of a sequence in  $\mathcal{F}$  is Glivenko-Cantelli.

4. Let  $h: \mathcal{X}^n \mapsto \mathbb{R}$  be permutation-symmetric. Then there exists a version  $h^*$  of the measurable cover for  $P^n$  on  $\mathcal{A}^n$  which is permutation-symmetric.

[Hint: Take an arbitrary measurable cover  $\tilde{h} \geq h$ , and set  $h^* = \min_{\sigma} \tilde{h} \circ \sigma$ .]

5. If the underlying probability space is not the product space  $(\mathcal{X}^\infty, \mathcal{A}^\infty, P^\infty)$ , then  $\|\mathbb{P}_n - P\|_{\mathcal{F}}$  may fail to be a reverse martingale, even if  $\|\mathbb{P}_n - P\|_{\mathcal{F}}$  is measurable for each  $n$ .

[Hint: There exists a (nonmeasurable) set  $A \subset [0, 1]$  with inner and outer Lebesgue measures  $\lambda_*(A) = 0$  and  $\lambda^*(A) = 1$ , respectively. Take  $(\mathcal{X}_n, \mathcal{A}_n, P_n)$  equal to  $(A, \mathcal{B} \cap A, \lambda_A)$  for odd values of  $n$  and  $(A^c, \mathcal{B} \cap A^c, \lambda_{A^c})$  for even values of  $n$ , where  $\lambda_A$  is the trace measure defined in Problem 1.2.16. Define  $X_n$  to be the embedding of  $\mathcal{X}_n$  into the unit interval, viewed as a map on the infinite product  $\prod \mathcal{X}_i$ , and let  $\mathcal{F}$  be equal to the set of indicators of all finite subsets of  $A^c$ . Then  $\|\mathbb{P}_n - P\|_{\mathcal{F}}$  equals  $(n - 1)/2n$ , for odd values of  $n$ , and  $1/2$  for  $n$  even.]

6. If the filtration  $\Sigma_n$  in Lemma 2.4.5 is replaced by the filtration consisting of the  $\sigma$ -fields generated by the variables  $\mathbb{P}_k f$  for  $k \geq n$  and  $f$  ranging over  $\mathcal{F}$ , then  $\|\mathbb{P}_n - P\|_{\mathcal{F}}^*$  may fail to be a reverse submartingale.

[Hint: Let  $(\mathcal{X}, \mathcal{A}, P)$  be the unit interval in  $\mathbb{R}$  with Borel sets and Lebesgue measure and  $X_1, X_2, \dots$  the coordinate projections of the infinite product space, as usual. For the collection  $\mathcal{F} = \{1_x : x \in [1/2, 1]\}$ , the variable  $\|\mathbb{P}_n - P\|_{\mathcal{F}}$  is measurable with respect to  $\mathcal{A}^\infty$ , but not  $\mathcal{A}_n$ -measurable for any  $n$ . It is also not measurable with respect to the  $P^\infty$ -completion of  $\mathcal{A}_n$  even though  $\mathcal{F}$  is image-admissible Suslin.]

7. The  $\sigma$ -field  $\Sigma_n$  in Lemma 2.4.5 equals the  $\sigma$ -field generated by the variables  $\mathbb{P}_k f$  with  $k$  ranging over all integers greater than or equal to  $n$  and  $f$  ranging over the set of all measurable, integrable  $f : \mathcal{X} \mapsto \mathbb{R}$ .

8. The sequence of averages  $\bar{Y}_n$  of the first  $n$  elements of a sequence of i.i.d. random variables  $Y_1, Y_2, \dots$  with finite first absolute moment is a reverse martingale with respect to its natural filtration:  $E(\bar{Y}_n | \bar{Y}_{n+1}, \bar{Y}_{n+2}, \dots) = \bar{Y}_{n+1}$ .

[Hint: The conditional expectation equals  $n^{-1} \sum_{i=1}^n E(Y_i | \bar{Y}_{n+1}, \bar{Y}_{n+2}, \dots)$ , and  $E(Y_i | \bar{Y}_{n+1})$  is constant in  $1 \leq i \leq n + 1$  by symmetry.]

9. **(Block-bracketing)** Given a class  $\mathcal{F}$  of functions  $f : \mathcal{X} \mapsto \mathbb{R}$  with integrable envelope, define  $\mathcal{F}^{\oplus m}$  to be the set of functions  $(x_1, \dots, x_m) \mapsto \sum_{i=1}^m f(x_i)$  on  $\mathcal{X}^m$ . Suppose that for every  $\varepsilon > 0$  there exists  $m$  such that  $N_{[]}(\varepsilon m, \mathcal{F}^{\oplus m}, L_1(P^m)) < \infty$ . Then  $\mathcal{F}$  is Glivenko-Cantelli.

[Hint: For every  $\varepsilon > 0$  and  $f$ , there exists  $m$  and a bracket  $[l_m, u_m]$  in  $L_1(P^m)$  such that  $m^{-1} P^m u_m - m^{-1} P^m l_m < \varepsilon$  and

$$\begin{aligned} \frac{1}{nm} \sum_{i=1}^n l_m(Y_{i,m}) &\leq \mathbb{P}_{nm} f \leq \frac{1}{nm} \sum_{i=1}^n u_m(Y_{i,m}), \\ \frac{1}{m} P^m l_m &\leq Pf \leq \frac{1}{m} P^m u_m, \end{aligned}$$

where  $Y_{1,m}, Y_{2,m}, \dots$  are the blocks  $[X_1, \dots, X_m], [X_{m+1}, \dots, X_{2m}], \dots$ . Conclude that for every  $\varepsilon > 0$  there exists  $m$  such that  $\limsup_{n \rightarrow \infty} \|\mathbb{P}_{nm} - P\|_{\mathcal{F}}^* < \varepsilon$ . It suffices to “close the gaps.”]

## 2.5

# Donsker Theorems

In this chapter we present the two main empirical central limit theorems. The first is based on uniform entropy, and its proof relies on symmetrization. The second is based on bracketing entropy.

### 2.5.1 Uniform Entropy

In this section weak convergence of the empirical process will be established under the condition that the envelope function  $F$  be square integrable, combined with the uniform entropy bound

$$(2.5.1) \quad \int_0^\infty \sup_Q \sqrt{\log N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))} d\varepsilon < \infty.$$

Here the supremum is taken over all finitely discrete probability measures  $Q$  on  $(\mathcal{X}, \mathcal{A})$  with  $\|F\|_{Q,2}^2 = \int F^2 dQ > 0$ . These conditions are by no means necessary, but they suffice for many examples. Finiteness of the previous integral will be referred to as the *uniform entropy condition*.

**2.5.2 Theorem.** *Let  $\mathcal{F}$  be a class of measurable functions that satisfies the uniform entropy bound (2.5.1). Let the classes  $\mathcal{F}_\delta = \{f - g: f, g \in \mathcal{F}, \|f - g\|_{P,2} < \delta\}$  and  $\mathcal{F}_\infty^2$  be  $P$ -measurable for every  $\delta > 0$ . If  $P^* F^2 < \infty$ , then  $\mathcal{F}$  is  $P$ -Donsker.*

**Proof.** Let  $\delta_n \downarrow 0$  be arbitrary. By Markov's inequality and the symmetrization Lemma 2.3.1,

$$P^*(\|\mathbb{G}_n\|_{\mathcal{F}_{\delta_n}} > x) \leq \frac{2}{x} E^* \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}_{\delta_n}}.$$

Since the supremum in the right-hand side is measurable by assumption, Fubini's theorem applies and the outer expectation can be calculated as  $E_X E_\varepsilon$ . Fix  $X_1, \dots, X_n$ . By Hoeffding's inequality, the stochastic process  $f \mapsto \{n^{-1/2} \sum_{i=1}^n \varepsilon_i f(X_i)\}$  is sub-Gaussian for the  $L_2(\mathbb{P}_n)$ -seminorm

$$\|f\|_n = \sqrt{\frac{1}{n} \sum_{i=1}^n f^2(X_i)}.$$

Use the second part of the maximal inequality Corollary 2.2.8 to find that

$$(2.5.3) \quad E_\varepsilon \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i f(X_i) \right\|_{\mathcal{F}_{\delta_n}} \lesssim \int_0^\infty \sqrt{\log N(\varepsilon, \mathcal{F}_{\delta_n}, L_2(\mathbb{P}_n))} d\varepsilon.$$

For large values of  $\varepsilon$  the set  $\mathcal{F}_{\delta_n}$  fits in a single ball of radius  $\varepsilon$  around the origin, in which case the integrand is zero. This is certainly the case for values of  $\varepsilon$  larger than  $\theta_n$ , where

$$\theta_n^2 = \sup_{f \in \mathcal{F}_{\delta_n}} \|f\|_n^2 = \left\| \frac{1}{n} \sum_{i=1}^n f^2(X_i) \right\|_{\mathcal{F}_{\delta_n}}.$$

Furthermore, covering numbers of the class  $\mathcal{F}_\delta$  are bounded by covering numbers of  $\mathcal{F}_\infty = \{f - g : f, g \in \mathcal{F}\}$ . The latter satisfy  $N(\varepsilon, \mathcal{F}_\infty, L_2(Q)) \leq N^2(\varepsilon/2, \mathcal{F}, L_2(Q))$  for every measure  $Q$ .

Limit the integral in (2.5.3) to the interval  $(0, \theta_n)$ , make a change of variables, and bound the integrand to obtain the bound

$$\int_0^{\theta_n/\|F\|_n} \sup_Q \sqrt{\log N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))} d\varepsilon \|F\|_n.$$

Here the supremum is taken over all discrete probability measures. The integrand is integrable by assumption. Furthermore,  $\|F\|_n$  is bounded below by  $\|F_*\|_n$ , which converges almost surely to its expectation, which may be assumed positive. Use the Cauchy-Schwarz inequality and the dominated convergence theorem to see that the expectation (with respect to  $X_1, \dots, X_n$ ) of this integral converges to zero provided  $\theta_n \xrightarrow{P^*} 0$ . This would conclude the proof of asymptotic equicontinuity.

Since  $\sup\{Pf^2 : f \in \mathcal{F}_{\delta_n}\} \rightarrow 0$  and  $\mathcal{F}_{\delta_n} \subset \mathcal{F}_\infty$ , it is certainly enough to prove that

$$\|\mathbb{P}_n f^2 - Pf^2\|_{\mathcal{F}_\infty} \xrightarrow{P^*} 0.$$

This is a uniform law of large numbers for the class  $\mathcal{F}_\infty^2$ . This class has integrable envelope  $(2F)^2$  and is measurable by assumption. For any pair  $f, g$  of functions in  $\mathcal{F}_\infty$ ,

$$\mathbb{P}_n |f^2 - g^2| \leq \mathbb{P}_n |f - g| 4F \leq \|f - g\|_n \|4F\|_n.$$

It follows that the covering number  $N(\varepsilon \|2F\|_n^2, \mathcal{F}_\infty^2, L_1(\mathbb{P}_n))$  is bounded by the covering number  $N(\varepsilon \|F\|_n, \mathcal{F}_\infty, L_2(\mathbb{P}_n))$ . By assumption, the latter

number is bounded by a fixed number, so its logarithm is certainly  $o_P^*(n)$ , as required for the uniform law of large numbers, Theorem 2.4.3. This concludes the proof of asymptotic equicontinuity.

Finally, we show that  $\mathcal{F}$  is totally bounded in  $L_2(P)$ . By the result of the last paragraph, there exists a sequence of discrete measures  $P_n$  with  $\|(P_n - P)f^2\|_{\mathcal{F}_\infty}$  converging to zero. Take  $n$  sufficiently large so that the supremum is bounded by  $\varepsilon^2$ . By assumption,  $N(\varepsilon, \mathcal{F}, L_2(P_n))$  is finite. Any  $\varepsilon$ -net for  $\mathcal{F}$  in  $L_2(P_n)$  is a  $\sqrt{2}\varepsilon$ -net in  $L_2(P)$ . ■

**2.5.4 Example (Cells in  $\mathbb{R}^k$ ).** The set  $\mathcal{F}$  of all indicator functions  $1\{(-\infty, t]\}$  of cells in  $\mathbb{R}$  satisfies

$$N(\varepsilon, \mathcal{F}, L_2(Q)) \leq N_{[]}(\varepsilon^2, \mathcal{F}, L_1(Q)) \leq \frac{2}{\varepsilon^2},$$

for any probability measure  $Q$  and  $\varepsilon \leq 1$ . Since  $\int_0^1 \log(1/\varepsilon) d\varepsilon < \infty$ , the class of cells in  $\mathbb{R}$  is Donsker. The covering numbers of the class of cells  $(-\infty, t]$  in higher dimensions satisfy a similar bound, but with a higher power of  $(1/\varepsilon)$ . Thus the class of all cells  $(-\infty, t]$  in  $\mathbb{R}^k$  is Donsker for any dimension.

In the next chapter these classes of sets are shown to be examples of the large collection of VC-classes, which are all Donsker provided they are suitably measurable.

## 2.5.2 Bracketing

The second main empirical central limit theorem uses bracketing entropy rather than uniform entropy. The simplest version of this theorem uses  $L_2(P)$ -brackets and asserts that a class  $\mathcal{F}$  of functions is  $P$ -Donsker if

$$\int_0^\infty \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}, L_2(P))} d\varepsilon < \infty.$$

Note that unlike the uniform entropy condition, this bracketing integral involves only the true underlying measure  $P$ . However, part of this gain is offset by the fact that bracketing numbers can be larger than covering numbers. As a result, the two sufficient conditions for a class to be Donsker are not comparable. Examples of classes of functions that satisfy the bracketing condition are given in Chapter 2.7.

The above assertion will be proved in somewhat greater generality. It is not difficult to see that finiteness of the  $L_2(P)$ -bracketing integral implies that  $P^*F^2 < \infty$  for an envelope function  $F$ . It is known that this is not necessary for a class  $\mathcal{F}$  to be Donsker, though the envelope must possess a weak second moment (meaning that  $P^*(F > x) = o(x^{-2})$  as  $x \rightarrow \infty$ ). Similarly, the  $L_2(P)$ -norm used to measure the size of the brackets can be replaced by a weaker norm, which makes the bracketing numbers smaller

and the convergence of the integral easier. The result below measures the size of the brackets by their  $L_{2,\infty}$ -norm given by

$$\|f\|_{P,2,\infty} = \sup_{x>0} \left( x^2 P(|f| > x) \right)^{1/2}.$$

(Actually this is not really a norm, because it does not satisfy the triangle inequality. It can be shown that there exists a norm that is equivalent to this “norm” up to a constant 2, but this is irrelevant for the present purposes.) Note that  $\|f\|_{P,2,\infty} \leq \|f\|_{P,2}$ , so that the bracketing numbers relative to  $L_{2,\infty}(P)$  are smaller.

The proof of the theorem is based on a chaining argument. Unlike in the proof of the uniform entropy theorem, this will be applied to the original summands, without an initial symmetrization. Because the summands are not appropriately bounded, Hoeffding’s inequality and the sub-Gaussian maximal inequality do not apply. Instead, Bernstein’s inequality is used in the form

$$P(|G_n f| > x) \leq 2 e^{-\frac{x^2}{2 Pf^2 + 1/3 \|f\|_\infty x / \sqrt{n}}}.$$

This is valid for every square integrable, uniformly bounded function  $f$ . Lemma 2.2.10 implies that for a finite set  $\mathcal{F}$  of cardinality  $|\mathcal{F}| \geq 2$ ,

$$(2.5.5) \quad E\|G_n\|_{\mathcal{F}} \lesssim \max_f \frac{\|f\|_\infty}{\sqrt{n}} \log |\mathcal{F}| + \max_f \|f\|_{P,2} \sqrt{\log |\mathcal{F}|}.$$

The chaining argument in the proof of the following theorem is set up in such a way that the two terms on the right are of comparable magnitude. This is the case if  $\|f\|_\infty \sim \|f\|_{P,2}/\sqrt{\log |\mathcal{F}|}$ ; the inequality is applied after truncating the functions  $f$  at this level.

**2.5.6 Theorem.** *Let  $\mathcal{F}$  be a class of measurable functions such that*

$$\int_0^\infty \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}, L_{2,\infty}(P))} d\varepsilon + \int_0^\infty \sqrt{\log N(\varepsilon, \mathcal{F}, L_2(P))} d\varepsilon < \infty.$$

*Moreover, assume that the envelope function  $F$  of  $\mathcal{F}$  possesses a weak second moment. Then  $\mathcal{F}$  is  $P$ -Donsker.*

**Proof.** For each natural number  $q$ , there exists a partition  $\mathcal{F} = \bigcup_{i=1}^{N_q} \mathcal{F}_{qi}$  of  $\mathcal{F}$  into  $N_q$  disjoint subsets such that

$$\begin{aligned} \sum 2^{-q} \sqrt{\log N_q} &< \infty, \\ \left\| \left( \sup_{f,g \in \mathcal{F}_{qi}} |f - g| \right)^* \right\|_{P,2,\infty} &< 2^{-q}, \\ \sup_{f,g \in \mathcal{F}_{qi}} \|f - g\|_{P,2} &< 2^{-q}. \end{aligned}$$

To see this, first cover  $\mathcal{F}$  separately with minimal numbers of  $L_2(P)$ -balls and  $L_{2,\infty}(P)$ -brackets of size  $2^{-q}$ , disjointify, and take the intersection of

the two partitions. The total number of sets will be  $N_q = N_q^1 N_q^2$  if  $N_q^i$  are the number of sets in the two separate partitions. The logarithm turns the product into a sum, and the first condition is satisfied if it is satisfied for both  $N_q^i$ .

The sequence of partitions can, without loss of generality, be chosen as successive refinements. Indeed, first construct a sequence of partitions  $\mathcal{F} = \cup_{i=1}^{\bar{N}_q} \bar{\mathcal{F}}_{qi}$  possibly without this property. Next take the partition at stage  $q$  to consist of all intersections of the form  $\cap_{p=1}^q \bar{\mathcal{F}}_{p,i_p}$ . This gives partitions into  $N_q = \bar{N}_1 \cdots \bar{N}_q$  sets. Using the inequality  $(\log \prod \bar{N}_p)^{1/2} \leq \sum (\log \bar{N}_p)^{1/2}$  and rearranging sums, it is seen that the first of the three displayed conditions is still satisfied.

Choose for each  $q$  a fixed element  $f_{qi}$  from each partitioning set  $\mathcal{F}_{qi}$ , and set

$$\begin{aligned} \pi_q f &= f_{qi} && \text{if } f \in \mathcal{F}_{qi}, \\ \Delta_q f &= \sup_{f,g \in \mathcal{F}_{qi}} |f - g|^*, && \text{if } f \in \mathcal{F}_{qi}. \end{aligned}$$

Note that  $\pi_q f$  and  $\Delta_q f$  run through a set of  $N_q$  functions if  $f$  runs through  $\mathcal{F}$ . In view of Theorem 1.5.6, it suffices to show that the sequence  $\|\mathbb{G}_n(f - \pi_{q_0} f)\|_{\mathcal{F}}$  converges in probability to zero as  $n \rightarrow \infty$  followed by  $q_0 \rightarrow \infty$ .

Define for each fixed  $n$  and (large)  $q \geq q_0$  numbers and indicator functions;

$$a_q = 2^{-q} / \sqrt{\log N_{q+1}},$$

$$\begin{aligned} A_{q-1} f &= 1\{\Delta_{q_0} f \leq \sqrt{n}a_{q_0}, \dots, \Delta_{q-1} f \leq \sqrt{n}a_{q-1}\}, \\ B_q f &= 1\{\Delta_{q_0} f \leq \sqrt{n}a_{q_0}, \dots, \Delta_{q-1} f \leq \sqrt{n}a_{q-1}, \Delta_q f > \sqrt{n}a_q\}, \\ B_{q_0} f &= 1\{\Delta_{q_0} f > \sqrt{n}a_{q_0}\}. \end{aligned}$$

Note that  $A_q f$  and  $B_q f$  are constant in  $f$  on each of the partitioning sets  $\mathcal{F}_{qi}$  at level  $q$ , because the partitions are nested. Now decompose, pointwise in  $x$  (which is suppressed in the notation),

$$f - \pi_{q_0} f = (f - \pi_{q_0} f)B_{q_0} f + \sum_{q_0+1}^{\infty} (f - \pi_q f)B_q f + \sum_{q_0+1}^{\infty} (\pi_q f - \pi_{q-1} f)A_{q-1} f.$$

The idea here is to write the left side as the sum of  $f - \pi_{q_1} f$  and  $\sum_{q_0+1}^{q_1} (\pi_q f - \pi_{q-1} f)$  for the largest  $q_1 = q_1(f, x)$  such that each of the “links”  $\pi_q f - \pi_{q-1} f$  in the “chain” is bounded in absolute value by  $\sqrt{n}a_q$  (note that  $|\pi_q f - \pi_{q-1} f| \leq \Delta_{q-1} f$ ). For a rigorous derivation, note that either all  $B_q f$  are zero or there is a unique  $q_1$  with  $B_{q_1} f = 1$ . In the first case, the first two terms in the decomposition are zero and the third term is an infinite series (all  $A_q f$  equal 1) whose  $q$ th partial sum telescopes out to  $\pi_q f - \pi_{q_0} f$  and converges to  $f - \pi_{q_0} f$  by the definition of the  $A_q f$ . In the second case,  $A_{q-1} f = 1$  if and only if  $q \leq q_1$  and the decomposition is as mentioned,

apart from the separate treatment of the case that  $q_1 = q_0$ , when already the first link fails the test.

Next apply the empirical process  $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$  to each of the three terms separately, and take suprema over  $f \in \mathcal{F}$ . It will be shown that the resulting three variables converge to zero in probability as  $n \rightarrow \infty$  followed by  $q_0 \rightarrow \infty$ .

First, since  $|f - \pi_{q_0} f| B_{q_0} f \leq 2F \{2F > \sqrt{n}a_{q_0}\}$ , one has

$$\mathbb{E}^* \|\mathbb{G}_n(f - \pi_{q_0} f) B_{q_0} f\|_{\mathcal{F}} \leq 4\sqrt{n}P^* F \{2F > \sqrt{n}a_{q_0}\}.$$

The right side converges to zero as  $n \rightarrow \infty$ , for each fixed  $q_0$ , by the assumption that  $F$  has a weak second moment (Problem 2.5.6).

Second, since the partitions are nested,  $\Delta_q f B_q f \leq \Delta_{q-1} f B_q f$ , whence by the inequality of Problem 2.5.5,

$$\sqrt{n}a_q P \Delta_q f B_q f \leq 2 \|\Delta_q f\|_{P,2,\infty}^2 \leq 2 2^{-q}.$$

Since  $\Delta_{q-1} f B_q f$  is bounded by  $\sqrt{n}a_{q-1}$  for  $q > q_0$ , we obtain

$$P(\Delta_q f B_q f)^2 \leq \sqrt{n}a_{q-1} P \Delta_q f \{\Delta_q f > \sqrt{n}a_q\} \leq 2 \frac{a_{q-1}}{a_q} 2^{-2q}.$$

Apply the triangle inequality and inequality (2.5.5) to find

$$\begin{aligned} \mathbb{E}^* \left\| \sum_{q_0+1}^{\infty} \mathbb{G}_n(f - \pi_q f) B_q f \right\|_{\mathcal{F}} \\ \leq \sum_{q_0+1}^{\infty} \mathbb{E}^* \|\mathbb{G}_n \Delta_q f B_q f\|_{\mathcal{F}} + \sum_{q_0+1}^{\infty} 2\sqrt{n} \|P \Delta_q f B_q f\|_{\mathcal{F}} \\ \lesssim \sum_{q_0+1}^{\infty} \left[ a_{q-1} \log N_q + \sqrt{\frac{a_{q-1}}{a_q} 2^{-q} \sqrt{\log N_q}} + \frac{4}{a_q} 2^{-2q} \right]. \end{aligned}$$

Since  $a_q$  is decreasing, the quotient  $a_{q-1}/a_q$  can be replaced by its square. Then in view of the definition of  $a_q$ , the series on the right can be bounded by a multiple of  $\sum_{q_0+1}^{\infty} 2^{-q} \sqrt{\log N_q}$ . This upper bound is independent of  $n$  and converges to zero as  $q_0 \rightarrow \infty$ .

Third, there are at most  $N_q$  functions  $\pi_q f - \pi_{q-1} f$  and at most  $N_{q-1}$  functions  $A_{q-1} f$ . Since the partitions are nested, the function  $|\pi_q f - \pi_{q-1} f| A_{q-1} f$  is bounded by  $\Delta_{q-1} f A_{q-1} f \leq \sqrt{n} a_{q-1}$ . The  $L_2(P)$ -norm of  $|\pi_q f - \pi_{q-1} f|$  is bounded by  $2^{-q+1}$ . Apply inequality (2.5.5) to find

$$\mathbb{E}^* \left\| \sum_{q_0+1}^{\infty} \mathbb{G}_n(\pi_q f - \pi_{q-1} f) A_{q-1} f \right\|_{\mathcal{F}} \lesssim \sum_{q_0+1}^{\infty} \left[ a_{q-1} \log N_q + 2^{-q} \sqrt{\log N_q} \right].$$

Again this upper bound is independent of  $n$  and converges to zero as  $q_0 \rightarrow \infty$ . ■

**2.5.7 Example (Cells in  $\mathbb{R}$ ).** The classical empirical central limit theorem was shown to be a special case of the uniform entropy central limit theorem in the preceding subsection. This classical theorem is a special case of the bracketing central limit theorem as well. This follows since  $N_{[]}(\sqrt{2}\varepsilon, \mathcal{F}, L_2(P)) \leq N_{[]}(\varepsilon^2, \mathcal{F}, L_1(P))$  for every class of functions  $f: \mathcal{X} \mapsto [0, 1]$ . The  $L_1$ -bracketing numbers were bounded above by a power of  $1/\varepsilon$  in the preceding subsection.

## Problems and Complements

1. (Taking the supremum over all  $Q$ ) The supremum in (2.5.1) can be replaced by the supremum over all probability measures  $Q$ , such that  $0 < QF^r < \infty$ , without changing the condition. In fact, for any probability measure  $P$  and  $r > 0$  such that  $0 < PF^r < \infty$ ,

$$D(2\varepsilon\|F\|_{P,r}, \mathcal{F}, L_r(P)) \leq \sup_Q D(\varepsilon\|F\|_{Q,r}, \mathcal{F}, L_r(Q))$$

if the supremum on the right is taken over all discrete probability measures  $Q$ .

[Hint: If the left side is equal to  $m$ , then there exist functions  $f_1, \dots, f_m$  such that  $P|f_i - f_j|^r > 2^r \varepsilon^r PF^r$  for  $i \neq j$ . By the strong law of large numbers,  $\mathbb{P}_n|f_i - f_j|^r \rightarrow P|f_i - f_j|^r$  almost surely. Also  $\mathbb{P}_n F^r \rightarrow PF^r$  almost surely. Thus there exist  $n$  and  $\omega$  such that  $\mathbb{P}_n(\omega)|f_i - f_j|^r > 2^r \varepsilon^r PF^r$  and  $\mathbb{P}_n(\omega)F^r < 2^r PF^r$ .]

2. Can the constant 2 in the preceding problem be replaced by 1? Is a similar inequality valid for covering numbers, and what is the best constant?
3. The presence of the factor  $\|F\|_{Q,2}$  in the uniform entropy condition (2.5.1) is helpful if it is larger than 1 but is detrimental otherwise. However, given any class  $\mathcal{F}$  with a square integrable envelope, there is always a square integrable envelope  $F$  with  $\|F\|_{Q,2} \geq 1$  for all  $Q$ .
4. If  $\mathcal{F}$  is compact, then the function  $\varepsilon \mapsto N(\varepsilon, \mathcal{F}, L_2(P))$  is left continuous.  
[Hint: The function  $f \mapsto \inf_i \|f - f_i\|$  attains its maximum for every finite set  $f_1, \dots, f_p$ .]
5. For each positive random variable  $X$ , one has the inequalities  $\|X\|_{2,\infty}^2 \leq \sup_{t>0} tEX\{X > t\} \leq 2\|X\|_{2,\infty}^2$ .  
[Hint: The first inequality follows from Markov's inequality. For the second, write the expectation in the expression in the middle as  $\int_0^\infty P(X1\{X > t\} > x) dx$ . Next, split the integral in the part from zero to  $t$  and its complement. On the first part, the integrand is constant; on the second, it can be bounded by  $x^{-2}$  times the  $L_{2,\infty}$ -norm.]
6. Each random variable with  $P(|X| > t) = o(t^{-2})$  as  $t \rightarrow \infty$  (a “weak second moment”) satisfies  $E|X|\{|X| > t\} = o(t^{-1})$  as  $t \rightarrow \infty$ .
7. For any random variable  $X$ , one has  $\|X\|_1 \leq 3\|X\|_{2,\infty}$ .

## 2.6

# Uniform Entropy Numbers

In Section 2.5.1 the empirical process was shown to converge weakly for indexing sets  $\mathcal{F}$  satisfying a uniform entropy condition. In particular, if

$$\sup_Q \log N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \leq K \left( \frac{1}{\varepsilon} \right)^{2-\delta},$$

for some  $\delta > 0$ , then the entropy integral (2.5.1) converges and  $\mathcal{F}$  is a Donsker class for any probability measure  $P$  such that  $P^*F^2 < \infty$ , provided measurability conditions are met. Many classes of functions satisfy this condition and often even the much stronger condition

$$\sup_Q N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \leq K \left( \frac{1}{\varepsilon} \right)^V, \quad 0 < \varepsilon < 1,$$

for some number  $V$ . In this chapter this is shown for classes satisfying certain combinatorial conditions. For classes of sets, these were first studied by Vapnik and Červonenkis, whence the name VC-classes. In the second part of this chapter, VC-classes of functions are defined in terms of VC-classes of sets. The remainder of this chapter considers operations on classes that preserve entropy properties, such as taking convex hulls.

### 2.6.1 VC-Classes of Sets

Let  $\mathcal{C}$  be a collection of subsets of a set  $\mathcal{X}$ . An arbitrary set of  $n$  points  $\{x_1, \dots, x_n\}$  possesses  $2^n$  subsets. Say that  $\mathcal{C}$  *picks out* a certain subset from  $\{x_1, \dots, x_n\}$  if this can be formed as a set of the form  $C \cap \{x_1, \dots, x_n\}$  for

a  $C$  in  $\mathcal{C}$ . The collection  $\mathcal{C}$  is said to *shatter*  $\{x_1, \dots, x_n\}$  if each of its  $2^n$  subsets can be picked out in this manner. The *VC-index*  $V(\mathcal{C})$  of the class  $\mathcal{C}$  is the smallest  $n$  for which no set of size  $n$  is shattered by  $\mathcal{C}$ . Clearly, the more refined  $\mathcal{C}$  is, the larger is its index. The index is more formally defined through

$$\Delta_n(\mathcal{C}, x_1, \dots, x_n) = \#\left\{C \cap \{x_1, \dots, x_n\} : C \in \mathcal{C}\right\},$$

$$V(\mathcal{C}) = \inf \left\{ n : \max_{x_1, \dots, x_n} \Delta_n(\mathcal{C}, x_1, \dots, x_n) < 2^n \right\}.$$

Here the infimum over the empty set is taken to be infinity, so that the index is  $\infty$  if and only if  $\mathcal{C}$  shatters sets of arbitrarily large size.<sup>b</sup>

A collection of measurable sets  $\mathcal{C}$  is called a *VC-class* if its index is finite. The main result of this section is the remarkable fact that the covering numbers of any VC-class grow polynomially in  $1/\varepsilon$  as  $\varepsilon \rightarrow 0$  of order dependent on the index of the class. Since the envelope of a class of sets is certainly square integrable, it follows that a VC-class of sets is Donsker for any underlying probability measure, provided measurability conditions are met.

**2.6.1 Example (Cells in  $\mathbb{R}^d$ ).** The collection of all cells of the form  $(-\infty, c]$  in  $\mathbb{R}$  shatters no two-point set  $\{x_1, x_2\}$ , because it fails to pick out the largest of the two points. Hence its VC-index equals 2. The collection of all cells  $(a, b]$  in  $\mathbb{R}$  shatters every two-point set but cannot pick out the subset consisting of the smallest and largest points of any set of three points. Thus its VC-index equals 3. With more effort, it can be seen that the indices of the same type of sets in  $\mathbb{R}^d$  are  $d+1$  and  $2d+1$ , respectively.

Techniques to show that many other collections of sets are VC are discussed in Section 2.6.5.

The following combinatorial result is needed: the number of subsets shattered by a class  $\mathcal{C}$  is at least the number of subsets picked out by  $\mathcal{C}$ . In particular, if the class  $\mathcal{C}$  shatters no set of  $V$  points, then the number of subsets picked out by  $\mathcal{C}$  is at most  $\sum_{j=0}^{V-1} \binom{n}{j}$ , the number of subsets of size  $\leq V-1$ .

**2.6.2 Lemma.** *Let  $\{x_1, \dots, x_n\}$  be arbitrary points. Then the total number of subsets  $\Delta_n(\mathcal{C}, x_1, \dots, x_n)$  picked out by  $\mathcal{C}$  is bounded above by the number of subsets of  $\{x_1, \dots, x_n\}$  shattered by  $\mathcal{C}$ .*

**Proof.** Assume without loss of generality that every  $C$  is a subset of the given set of points, so that  $\Delta_n(\mathcal{C}, x_1, \dots, x_n)$  is the cardinality of  $\mathcal{C}$ .

---

<sup>b</sup> The formal definition appears to leave the index of the collection  $2^{\mathcal{X}}$  of all subsets of a *finite* set  $\mathcal{X}$  undefined. The definition  $V(2^{\mathcal{X}}) = |\mathcal{X}| + 1$  appears to be in agreement with the definition in words. (No sets of size  $|\mathcal{X}| + 1$  are shattered, because there are no such sets.) With this definition the results in this section are true (but uninteresting).

Call  $\mathcal{C}$  *hereditary* if it has the property that  $B \in \mathcal{C}$  whenever  $B \subset C$ , for a set  $C \in \mathcal{C}$ . Each of the sets in a hereditary collection of sets is shattered, whence a hereditary collection shatters at least  $|\mathcal{C}|$  sets and the assertion of the lemma is certainly true for hereditary collections. It will be shown that a general  $\mathcal{C}$  can be transformed into a hereditary collection, without changing its cardinality and without increasing the number of shattered sets.

Given  $1 \leq i \leq n$  and  $C \in \mathcal{C}$ , define the set  $T_i(C)$  to be  $C - \{x_i\}$  if  $C - \{x_i\}$  is not contained in  $\mathcal{C}$  and to be  $C$  otherwise. Thus  $x_i$  is deleted from  $C$  if this creates a “new” set; otherwise it is retained. If  $x_i \notin C$ , then  $C$  is left unchanged.

Since the map  $T_i$  is one-to-one, the collections  $\mathcal{C}$  and  $T_i(\mathcal{C})$  have the same cardinality. Furthermore, every subset  $A \subset \{x_1, \dots, x_n\}$  that is shattered by  $T_i(\mathcal{C})$  is shattered by  $\mathcal{C}$ . If  $x_i \notin A$ , this is clear from the fact that the collection of sets  $C \cap A$  does not change if each  $C$  is replaced by its transformation  $T_i(C)$ . Second, if  $x_i \in A$  and  $A$  is shattered by  $T_i(\mathcal{C})$ , then for every  $B \subset A$ , there is  $C \in \mathcal{C}$  with  $B \cup \{x_i\} = T_i(C) \cap A$ . This implies  $x_i \in T_i(C)$ , whence  $T_i(C) = C$ , whence  $C - \{x_i\} \in \mathcal{C}$ . Thus both  $B \cup \{x_i\}$  and  $B - \{x_i\} = (C - \{x_i\}) \cap A$  are picked out by  $\mathcal{C}$ . One of them equals  $B$ .

It has been shown that the assertion of the lemma is true for  $\mathcal{C}$  if it is true for  $T_i(\mathcal{C})$ . The same is true for the operator  $T_1 \circ T_2 \circ \dots \circ T_n$  playing the role of  $T_i$ . Apply this operator repeatedly, until the collection of sets does not change any more. This happens after at most  $\sum_C |\mathcal{C}|$  steps, because  $\sum_C |T_i(\mathcal{C})| < \sum_C |\mathcal{C}|$  whenever the collections  $T_i(\mathcal{C})$  and  $\mathcal{C}$  are different. The collection  $\mathcal{D}$  obtained in this manner has the property that  $D - \{x_i\}$  is contained in  $\mathcal{D}$  for every  $D \in \mathcal{D}$  and every  $x_i$ . Consequently,  $\mathcal{D}$  is hereditary. ■

**2.6.3 Corollary.** *For a VC-class of sets of index  $V(\mathcal{C})$ , one has*

$$\max_{x_1, \dots, x_n} \Delta_n(\mathcal{C}, x_1, \dots, x_n) \leq \sum_{j=0}^{V(\mathcal{C})-1} \binom{n}{j}.$$

Consequently, the numbers on the left side grow polynomially of order at most  $O(n^{V(\mathcal{C})-1})$  as  $n \rightarrow \infty$ .

**Proof.** A VC-class shatters no set of  $V(\mathcal{C})$  points. All shattered sets are among the sets of size at most  $V(\mathcal{C}) - 1$ . The number of shattered sets gives an upper bound on  $\Delta_n$  by the preceding lemma. ■

**2.6.4 Theorem.** *There exists a universal constant  $K$  such that for any VC-class  $\mathcal{C}$  of sets, any probability measure  $Q$ , any  $r \geq 1$ , and  $0 < \varepsilon < 1$ ,*

$$N(\varepsilon, \mathcal{C}, L_r(Q)) \leq KV(\mathcal{C})(4e)^{V(\mathcal{C})} \left(\frac{1}{\varepsilon}\right)^{r(V(\mathcal{C})-1)}.$$

**Proof.** In view of the equality  $\|1_C - 1_D\|_{Q,r} = Q^{1/r}(C \Delta D)$  for any pair of sets  $C$  and  $D$ , the upper bound for a general  $r$  is an easy consequence of the bound for  $r = 1$ . The proof for  $r = 1$  is long. Problem 2.6.4 gives a short proof of a slightly weaker result.

By Problem 2.6.3, it suffices to consider the case that  $Q$  is an empirical type measure: a measure on a finite set of points  $y_1, \dots, y_k$  such that  $Q\{y_i\} = l_i/n$  for integers  $l_1, \dots, l_k$  that add up to  $n$ . Then it is no loss of generality to assume that each set in the collection  $\mathcal{C}$  is a subset of these points.

Let  $x_1, \dots, x_n$  be the set of points  $y_1, \dots, y_k$  with each  $y_i$  occurring  $l_i$  times. For each subset  $C$  of  $y_1, \dots, y_k$ , form a subset  $\tilde{C}$  of  $x_1, \dots, x_n$  by taking all  $x_i$  that are copies of some  $y_i$  in  $C$ . More precisely, let  $\phi: \{1, \dots, n\} \mapsto \{y_1, \dots, y_k\}$  be an arbitrary, fixed map such that  $\#\{j: \phi(j) = y_i\} = l_i$  for every  $i$ . Let  $\tilde{C} = \{j: \phi(j) \in C\}$ . Now identify  $x_1, \dots, x_n$  with  $1, \dots, n$ .

If a set  $\{x_j: j \in J\}$  is shattered by the collection  $\tilde{\mathcal{C}}$  of sets  $\tilde{C}$ , then every  $x_j$  in this set must correspond to a different  $y_i$  (i.e., the map  $\phi$  must be one-to-one on  $\{x_j: j \in J\}$ ), and the subset  $\{\phi(x_j): j \in J\}$  of  $\{y_1, \dots, y_k\}$  must be shattered by  $\mathcal{C}$ . This shows that the collection  $\tilde{\mathcal{C}}$  is VC of the same index as  $\mathcal{C}$ . By construction,  $Q(C \Delta D) = \tilde{Q}(\tilde{C} \Delta \tilde{D})$  for  $\tilde{Q}$  equal to the discrete uniform measure on  $x_1, \dots, x_n$ . Thus the  $L_1(Q)$ -distance on  $\mathcal{C}$  corresponds to the  $L_1(\tilde{Q})$ -distance on  $\tilde{\mathcal{C}}$ . Conclude that  $N(\varepsilon, \tilde{\mathcal{C}}, L_1(\tilde{Q})) = N(\varepsilon, \mathcal{C}, L_1(Q))$ , and it suffices to prove the upper bound for  $\tilde{\mathcal{C}}$  and  $\tilde{Q}$ . For simplicity of notation, assume that  $Q$  is the discrete uniform measure on a set of points  $x_1, \dots, x_n$ .

Each set  $C$  can be represented by an  $n$ -vector of ones and zeros indicating whether or not the points  $x_i$  are contained in the set. Thus the collection  $\mathcal{C}$  is identified with a subset  $\mathcal{Z}$  of the vertices of the  $n$ -dimensional hypercube  $[0, 1]^n$ . Alternatively,  $\mathcal{C}$  can be identified with an  $(n \times \#\mathcal{C})$ -matrix of zeros and ones, the columns representing the individual sets  $C$ . For a subset  $I$  of rows, let  $\mathcal{Z}_I$  be the matrix obtained by first deleting the other rows and next dropping duplicate columns. Alternatively,  $\mathcal{Z}_I$  is the projection of the point set  $\mathcal{Z}$  onto  $[0, 1]^I$ . In these terms  $\mathcal{Z}_I$  corresponds to the collection of subsets  $C \cap \{x_i: i \in I\}$  of the set of points  $\{x_i: i \in I\}$ , and this set is shattered if the matrix  $\mathcal{Z}_I$  consists of all  $2^I$  possible columns or if, as a subset of the  $I$ -dimensional hypercube,  $\mathcal{Z}_I$  contains all vertices. By assumption, this is possible only for  $I < V(\mathcal{C})$ . Abbreviate  $S = V(\mathcal{C}) - 1$ .

The normalized *Hamming metric* on the point set  $\mathcal{Z}$  is defined by

$$d(w, z) = \frac{1}{n} \sum_{j=1}^n |w_j - z_j|, \quad z, w \in \mathcal{Z}.$$

This corresponds exactly to the  $L_1(Q)$ -metric on the collection  $\mathcal{C}$ : if the sets  $C$  and  $D$  correspond to the vertices  $w$  and  $z$ , then  $Q(C \Delta D) = d(w, z)$  for the discrete uniform measure  $Q$ .

Fix a maximal  $\varepsilon$ -separated collection of sets  $C \in \mathcal{C}$ . For simplicity of notation assume that  $\mathcal{C}$  (or equivalently,  $\mathcal{Z}$ ) itself is  $\varepsilon$ -separated. We shall bound its cardinality, which constitutes a bound on the packing number  $D(\varepsilon, \mathcal{C}, L_1(Q))$ .

Let  $Z$  be a random variable with a discrete uniform distribution on the point set  $\mathcal{Z}$  in  $[0, 1]^n$ . The coordinates of  $Z = (Z_1, \dots, Z_n)$  are (correlated) Bernoulli variables taking values in  $\{0, 1\}$ . Fix an integer  $S \leq m < n$ . Given a subset  $I$  of  $\{1, 2, \dots, n\}$  of size  $m + 1$ , apply Lemma 2.6.6 (below) to the projection  $Z_I$  of  $Z$  onto  $\mathcal{Z}_I$  to conclude that

$$\sum_{i \in I} \mathbb{E} \operatorname{var}(Z_i | Z_{I-\{i\}}) \leq S.$$

Take the sum over all  $\binom{n}{m+1}$  of such subsets  $I$  on both left and right. The double sum on the left can be rearranged. Instead of leaving out one element at a time from every possible subset of size  $m + 1$ , we may add missing elements to every possible set of size  $m$ . Conclude that

$$(2.6.5) \quad \sum_J \mathbb{E} \sum_{i \notin J} \operatorname{var}(Z_i | Z_J) \leq \binom{n}{m+1} S,$$

where the first sum is over all subsets  $J$  of size  $m$ .

Conditionally on  $Z_J = s$ , the vector  $Z$  is uniformly distributed over the set of columns  $z$  in  $\mathcal{Z}$  for which  $z_J = s$ . Suppose there are  $N_s$  of such columns. Let  $W$  and  $\tilde{W}$  be independent random vectors defined on a common probability space distributed uniformly over these columns. By the  $\varepsilon$ -separation of  $\mathcal{Z}$ , the vectors  $W$  and  $\tilde{W}$  are at Hamming distance  $d(W, \tilde{W})$  at least  $\varepsilon$  whenever they are unequal. The latter happens with probability  $1 - 1/N_s$ . Since  $\operatorname{var} W_i = \mathbb{E}(W_i - \tilde{W}_i)^2/2$ ,

$$\sum_{i \notin J} \operatorname{var}(Z_i | Z_J = s) = \frac{1}{2} \sum_i \mathbb{E}(W_i - \tilde{W}_i)^2 = \frac{1}{2} \operatorname{End}(W, \tilde{W}) \geq \frac{1}{2} n \varepsilon \left(1 - \frac{1}{N_s}\right).$$

If we integrate the left side with respect to  $s$  for the distribution of  $Z_J$  and next sum over  $J$ , then we obtain the left side of (2.6.5). Since  $P(Z_J = s) = N_s / \#\mathcal{Z}$ , the resulting expression is bounded below by

$$\sum_J \sum_{s \in \mathcal{Z}_J} \frac{N_s}{\#\mathcal{Z}} \frac{1}{2} \varepsilon n \left(1 - \frac{1}{N_s}\right) = \binom{n}{m} \frac{1}{2} \varepsilon n \left(1 - \frac{\#\mathcal{Z}_J}{\#\mathcal{Z}}\right).$$

This is bounded above by the right side of (2.6.5). Rearrange and simplify the resulting inequality to obtain

$$\#\mathcal{Z} \leq \frac{\overline{\#\mathcal{Z}_J} n \varepsilon (m+1)}{n \varepsilon m + n - 2nS + 2mS} \lesssim \frac{\overline{\#\mathcal{Z}_J} \varepsilon m}{\varepsilon m - 2S}.$$

The number of points in  $\mathcal{Z}_J$  is equal to the number of subsets picked out by  $\mathcal{C}$  from the points  $\{x_i : i \in J\}$ . By Sauer's lemma, this is bounded by  $\sum_{j=0}^S \binom{m}{j}$ , which is smaller than  $(em/S)^S$  for  $m \geq S$ . Thus we obtain

$$\#\mathcal{Z} \leq \left(\frac{e}{S}\right)^S \frac{m^{S+1}\varepsilon}{m\varepsilon - 2S},$$

for every integer  $S \leq m < n$ . The optimal unrestricted choice of  $m \in (0, \infty)$  is  $m = 2(S+1)/\varepsilon$ , for which the upper bound takes the form

$$\left(\frac{e}{S}\right)^S \frac{1}{2} \left(\frac{2(S+1)}{\varepsilon}\right)^{S+1} \leq (S+1)e \left(\frac{2e}{\varepsilon}\right)^S.$$

Since the upper bound evaluated at  $m = 2(S+1)/\varepsilon + 1$  differs from this only up to a universally bounded constant, the discretization of  $m$  causes no problem. Furthermore, for the optimal choice of  $m$ , the inequality  $m > S$  is implied by  $\varepsilon < 2$ , while for  $m = 2(S+1)/\varepsilon \geq n$ , we can use the trivial bound  $\#\mathcal{Z} \leq n \leq m$ , which is certainly bounded by the right side of the preceding display for  $S \geq 1$ . For  $S = 0$ , the collection  $\mathcal{C}$  consists of at most one set, and the theorem is trivial. ■

**2.6.6 Lemma.** *Let  $Z$  be an arbitrary random vector taking values in a set  $\mathcal{Z} \subset \{0, 1\}^n$  that corresponds to a VC-class  $\mathcal{C}$  of subsets of a set of points  $\{x_1, \dots, x_n\}$ . Then*

$$\sum_{i=1}^n \text{E var}(Z_i | Z_j, j \neq i) \leq V(\mathcal{C}) - 1.$$

**Proof.** If the values of all coordinates, except the  $i$ th coordinate, of  $Z$  are given, then the vector  $Z$  can have at most two values. Call these  $v$  and  $w$ , where  $v_i = 0$  and  $w_i = 1$  and the other coordinates of  $v$  and  $w$  agree. Write  $p(z)$  for  $\text{P}(Z = z)$ . Then  $Z_i$  is 1 or 0 with conditional probabilities  $p := p(w)/(p(v) + p(w))$  and  $1 - p$ , respectively. The conditional variance of  $Z_i$  is  $p(1 - p)$ .

Form a graph by connecting any two points that are at minimal Hamming distance. This is a subgraph of the set of edges of the  $n$ -dimensional unit cube. Denote the edge between  $v$  and  $w$  by  $\{v, w\}$ , and let  $\mathcal{E}_i$  and  $\mathcal{E}$  be the set of all edges that cross the  $i$ th dimension and all edges in the graph, respectively. Then

$$\begin{aligned} \sum_{i=1}^n \text{E var}(Z_i | Z_j, j \neq i) &= \sum_{i=1}^n \sum_{\{v, w\} \in \mathcal{E}_i} (p(v) + p(w)) \frac{p(v)}{p(v) + p(w)} \frac{p(w)}{p(v) + p(w)} \\ &\leq \sum_{\{v, w\} \in \mathcal{E}} p(v) \wedge p(w). \end{aligned}$$

Suppose that the edges  $\mathcal{E}$  can be directed in such a way that at each node (point in  $\mathcal{Z}$ ) the number of arrows that are directed away is at most  $V(\mathcal{C}) - 1$ .

Then the last sum can be rearranged as a double sum, the outer sum carried out over the nodes, and the inner sum over the arrows directed away from the particular node. Thus the sum is bounded by  $\sum_{z \in \mathcal{Z}} p(z) \# \text{arrows}(z)$ , which is bounded by  $V(\mathcal{C}) - 1$ .

It may be shown that the edges can always be directed in the given manner (Problem 2.6.5), but it is more direct to proceed in a different manner. Recall from the proof of Lemma 2.6.2 that a class  $\mathcal{C}$  is hereditary if it contains every subset of every set contained in  $\mathcal{C}$ . Since each set in a hereditary collection is shattered, each set in a hereditary VC-class has at most  $V(\mathcal{C}) - 1$  points. Therefore, if  $\mathcal{C}$  is hereditary, the edges of the graph can be directed as desired by decreasing the number of elements: direct the edge between  $v$  and  $w$  in the direction of  $w \rightarrow v$  if the (one) coordinate of  $w$  by which  $w$  differs from  $v$  equals 1. This concludes the proof in the case that  $\mathcal{C}$  is hereditary. It was shown in the proof of Lemma 2.6.2 that an arbitrary collection  $\mathcal{C}$  can be transformed into a hereditary collection by repeated application of certain operators  $T_i$ . It was also shown that the operators are one-to-one and do not increase the VC-dimension. The operators induce a map on the edges of the graph corresponding to  $\mathcal{C}$  as follows. A given edge  $\{v, w\}$  in the graph  $(\mathcal{Z}, \mathcal{E})$  is mapped into an edge  $\{T_i(v), T_i(w)\}$  if this is an edge in the graph  $(T_i(\mathcal{Z}), \mathcal{E}(T_i(\mathcal{Z})))$ ; otherwise it is necessarily the case that  $T_i$  changes one of  $v$  and  $w$ , say  $w$ , and not the other, in which case the edge  $\{v, w\}$  is mapped into  $\{v^0, T_i(w)\}$ , where  $v_i^0 = 0$  and  $v^0$  agrees with  $v = v^1$  in the remaining coordinates. Geometrically, this means that edges that cross the  $i$ th dimension and edges at height zero in the  $i$ th dimension are left unchanged, whereas an edge at height one in the  $i$ th dimension is pushed down if the edge at height zero below it was not in the graph already. This shows that the map is one-to-one (though not necessarily onto). Finally, define a probability measure on  $T_i(\mathcal{Z})$  by shifting mass downward in the following manner: define  $p_i(z^0) = p(z^0) \vee p(z^1)$  and  $p_i(z^1) = p(z^0) \wedge p(z^1)$  for every  $z$ . It may be checked that

$$\sum_{\{v, w\} \in \mathcal{E}(\mathcal{Z})} p(v) \wedge p(w) \leq \sum_{\{v, w\} \in \mathcal{E}(T_i(\mathcal{Z}))} p_i(v) \wedge p_i(w).$$

Note here that  $(a_0 \wedge b_0) + (a_1 \wedge b_1)$  is bounded above by  $(a_0 \vee a_1) \wedge (b_0 \vee b_1) + (a_0 \wedge a_1) \wedge (b_0 \wedge b_1)$  for any numbers  $a_i, b_i$ . Thus by repeated application of the downward shift operation, the original graph and probability measure are changed into a hereditary graph and another probability measure, meanwhile increasing the target function. The theorem follows. ■

## 2.6.2 VC-Classes of Functions

The *subgraph* of a function  $f: \mathcal{X} \mapsto \mathbb{R}$  is the subset of  $\mathcal{X} \times \mathbb{R}$  given by<sup>#</sup>

$$\{(x, t) : t < f(x)\}.$$

A collection  $\mathcal{F}$  of measurable functions on a sample space is called a *VC-subgraph class*, or just a *VC-class*, if the collection of all subgraphs of the functions in  $\mathcal{F}$  forms a VC-class of sets (in  $\mathcal{X} \times \mathbb{R}$ ). Just as for sets, the covering numbers of VC-classes of functions grow at a polynomial rate.

Let  $V(\mathcal{F})$  be the VC-index of the set of subgraphs of functions in  $\mathcal{F}$ .

**2.6.7 Theorem.** *For a VC-class of functions with measurable envelope function  $F$  and  $r \geq 1$ , one has for any probability measure  $Q$  with  $\|F\|_{Q,r} > 0$ ,*

$$N(\varepsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q)) \leq KV(\mathcal{F})(16e)^{V(\mathcal{F})} \left(\frac{1}{\varepsilon}\right)^{r(V(\mathcal{F})-1)},$$

for a universal constant  $K$  and  $0 < \varepsilon < 1$ .

**Proof.** Let  $\mathcal{C}$  be the set of all subgraphs  $C_f$  of functions  $f$  in  $\mathcal{F}$ . By Fubini's theorem,  $Q|f-g| = Q \times \lambda(C_f \Delta C_g)$  if  $\lambda$  is Lebesgue measure on the real line. Renormalize  $Q \times \lambda$  to a probability measure on the set  $\{(x, t) : |t| \leq F(x)\}$  by defining  $P = (Q \times \lambda)/(2QF)$ . Then, by the result for sets in the previous section,

$$N(\varepsilon 2QF, \mathcal{F}, L_1(Q)) = N(\varepsilon, \mathcal{C}, L_1(P)) \leq KV(\mathcal{F}) \left(\frac{4e}{\varepsilon}\right)^{V(\mathcal{F})-1},$$

for a universal constant  $K$ .

This concludes the proof for  $r = 1$ . For  $r > 1$ , note that

$$Q|f - g|^r \leq Q|f - g|(2F)^{r-1} = 2^{r-1} R|f - g|QF^{r-1},$$

for the probability measure  $R$  with density  $F^{r-1}/QF^{r-1}$  with respect to  $Q$ . Thus the  $L_r(Q)$ -distance is bounded by the distance  $2(QF^{r-1})^{1/r} \|f - g\|_{R,1}^{1/r}$ . Elementary manipulations yield

$$N(\varepsilon 2\|F\|_{Q,r}, \mathcal{F}, L_r(Q)) \leq N(\varepsilon^r RF, \mathcal{F}, L_1(R)).$$

This can be bounded by a constant times  $1/\varepsilon$  raised to the power  $r(V(\mathcal{F}) - 1)$  by the result of the previous paragraph. ■

The preceding theorem shows that a VC-class satisfies the uniform entropy condition (2.5.1), with much to spare. Thus Theorem 2.5.2 shows that a suitably measurable VC-class is  $P$ -Donsker for any underlying measure  $P$  for which the envelope is square integrable. The latter condition on the envelope can be relaxed a little to the minimal condition that the envelope has a weak second moment.

---

<sup>#</sup> The form of this definition is different from the definitions given by other authors. However, it can be shown to lead to the same concept. See Problems 2.6.10 and 2.6.11.

**2.6.8 Theorem.** Let  $\mathcal{F}$  be a pointwise separable,  $P$ -pre-Gaussian VC-class of functions with envelope function  $F$  such that  $P^*(F > x) = o(x^{-2})$  as  $x \rightarrow \infty$ . Then  $\mathcal{F}$  is  $P$ -Donsker.

**Proof.** Under the stronger condition that the envelope has a finite second moment, this follows from Theorem 2.5.2. (Then the assumption that  $\mathcal{F}$  is pre-Gaussian is automatically satisfied.) For the refinement, see Alexander (1987c). ■

### 2.6.3 Convex Hulls and VC-Hull Classes

The *symmetric convex hull*  $sconv \mathcal{F}$  of a class of functions is defined as the set of functions  $\sum_{i=1}^m \alpha_i f_i$ , with  $\sum_{i=1}^m |\alpha_i| \leq 1$  and each  $f_i$  contained in  $\mathcal{F}$ .

A set of measurable functions is called a *VC-hull class* if it is in the pointwise sequential closure of the symmetric convex hull of a VC-class of functions. More formally, a collection of functions  $\mathcal{F}$  is VC-hull if there exists a VC-class  $\mathcal{G}$  of functions such that every  $f \in \mathcal{F}$  is the pointwise limit of a sequence of functions  $f_m$  contained in  $sconv \mathcal{G}$ . If the class  $\mathcal{G}$  can be taken equal to a class of indicator functions, then the class  $\mathcal{F}$  is called a *VC-hull class for sets*.

In Chapter 2.10 it is shown that the sequentially closed symmetric convex hull of a Donsker class is Donsker. Since a suitably measurable VC-class of functions is Donsker, provided its envelope has a weak second moment, many VC-hull classes can be seen to be Donsker by this general result. In most cases the same conclusion can also be obtained from the stronger result that any VC-hull class satisfies the uniform entropy condition. This is shown in the following, which gives an upper bound on the entropy.

Even though VC-hull classes are small enough to have a finite uniform entropy integral, they can be considerably larger than VC-classes. Their entropy numbers (logarithms of the covering numbers), rather than their covering numbers, behave polynomially in  $1/\varepsilon$ . More precisely, the entropy numbers of the convex hull of any polynomial class are of lower order than  $(1/\varepsilon)^r$  for some  $r < 2$ . The bound  $r < 2$  is just enough to ensure that the class satisfies the uniform entropy condition (2.5.1).

**2.6.9 Theorem.** Let  $Q$  be a probability measure on  $(\mathcal{X}, \mathcal{A})$ , and let  $\mathcal{F}$  be a class of measurable functions with measurable square integrable envelope  $F$  such that  $QF^2 < \infty$  and

$$N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \leq C \left( \frac{1}{\varepsilon} \right)^V, \quad 0 < \varepsilon < 1.$$

Then there exists a constant  $K$  that depends on  $C$  and  $V$  only such that

$$\log N(\varepsilon \|F\|_{Q,2}, \overline{\text{conv}} \mathcal{F}, L_2(Q)) \leq K \left( \frac{1}{\varepsilon} \right)^{2V/(V+2)}.$$

**Proof.** Every point in the convex hull of  $\mathcal{F}$  is within distance  $\varepsilon$  of the convex hull of an  $\varepsilon$ -net over  $\mathcal{F}$ . Therefore, to prove the assertion of the theorem for a fixed  $\varepsilon$ , it is no loss of generality to assume that  $\mathcal{F}$  is finite.

Set  $W = 1/2 + 1/V$  and  $L = C^{1/V} \|F\|_{Q,2}$ . Then by assumption  $\mathcal{F}$  can be covered by  $n$  balls of radius at most  $Ln^{-1/V}$  for every natural number  $n$ . (Note that the assumption is trivially true for  $1 \leq \varepsilon \leq C^{1/V}$ .) Form sets  $\mathcal{F}_1 \subset \mathcal{F}_2 \subset \dots \subset \mathcal{F}$  such that for each  $n$  the set  $\mathcal{F}_n$  is a maximal,  $Ln^{-1/V}$ -separated net over  $\mathcal{F}$ . Thus  $\mathcal{F}_n$  contains at most  $n$  points. It will be shown by induction that there exist constants  $C_k$  and  $D_k$  depending only on  $C$  and  $V$  such that  $\sup_k C_k \vee D_k < \infty$  and, for  $q \geq 3 + V$ ,

$$(2.6.10) \quad \log N(C_k Ln^{-W}, \text{conv } \mathcal{F}_{nk^q}, L_2(Q)) \leq D_k n, \quad n, k \geq 1.$$

This would imply the theorem. The proof of (2.6.10) consists of a nested induction argument. The outer layer is induction on  $k$ . The case  $k = 1$  is proved for each  $n$  by induction on  $n$ .

Assume  $k = 1$ . For  $n \leq n_0$  and fixed  $n_0$ , the statement is trivially true for sufficiently large  $C_1$ . It suffices to choose  $C_1 Ln_0^{-W} \geq \|F\|_{Q,2}$ , in which case the left side of (2.6.10) vanishes for every  $n \leq n_0$ . For general  $n$ , fix  $m = n/d$  for large enough  $d$  to be chosen. Each  $f \in \mathcal{F}_n - \mathcal{F}_m$  is within distance at most  $Lm^{-1/V}$  of some element  $\Pi_m f$  of  $\mathcal{F}_m$ . Thus each element of  $\text{conv } \mathcal{F}_n$  can be written as

$$\sum_{f \in \mathcal{F}_n} \lambda_f f = \sum_{f \in \mathcal{F}_m} \mu_f f + \sum_{f \in \mathcal{F}_n - \mathcal{F}_m} \lambda_f (f - \Pi_m f),$$

where  $\mu_f \geq 0$  and  $\sum \mu_f = \sum \lambda_f = 1$ . If  $\mathcal{G}_n$  is the set of functions  $0$  and  $f - \Pi_m f$  with  $f$  ranging over  $\mathcal{F}_n - \mathcal{F}_m$ , then it follows that  $\text{conv } \mathcal{F}_n \subset \text{conv } \mathcal{F}_m + \text{conv } \mathcal{G}_n$  for a set  $\mathcal{G}_n$  containing at most  $n$  elements, each of norm smaller than  $Lm^{-1/V}$ . Apply Lemma 2.6.11 (below) to  $\mathcal{G}_n$  with  $\varepsilon$  defined by  $m^{-1/V}\varepsilon = (1/2)C_1 n^{-W}$  to find a  $-(1/2)C_1 Ln^{-W}$ -net over  $\text{conv } \mathcal{G}_n$  consisting of at most

$$(e + en\varepsilon^2)^{2/\varepsilon^2} \leq \left( e + \frac{eC_1^2}{d^{2/V}} \right)^{8d^{2/V}C_1^{-2}n}$$

elements. Apply the induction hypothesis to  $\mathcal{F}_m$  to find a  $C_1 Lm^{-W}$ -net over  $\text{conv } \mathcal{F}_m$  consisting of at most  $e^m$  elements. This defines a partition of  $\text{conv } \mathcal{F}_m$  into  $m$ -dimensional sets of diameter at most  $2C_1 Lm^{-W}$ . Such a set can be isometrically identified with a subset of a ball of radius  $C_1 Lm^{-W}$  in  $\mathbb{R}^m$ . Thus each of these sets can be partitioned in

$$\left( \frac{3C_1 Lm^{-W}}{(1/2)C_1 Ln^{-W}} \right)^m = (6d^W)^{n/d}$$

sets of diameter  $(1/2)C_1 Ln^{-W}$  (Problem 2.1.6). Take a function from each of these sets, and form all sums  $f + g$  of a function  $f$  attached to  $\text{conv } \mathcal{F}_m$

and a function  $g$  attached to  $\text{conv } \mathcal{G}_n$  by the preceding procedure. These form a  $C_1 L n^{-W}$ -net over  $\text{conv } \mathcal{F}_n$  of cardinality bounded by

$$e^{n/d} (6d^W)^{n/d} \left( e + \frac{eC_1^2}{d^{2/V}} \right)^{8d^{2/V} C_1^{-2} n}.$$

This is bounded by  $e^n$  for suitable choices of  $C_1$  and  $d$  depending on  $V$  only. This concludes the proof of (2.6.10) for  $k = 1$  and every  $n$ .

The argument continues by induction on  $k$ . By a similar construction as before,  $\text{conv } \mathcal{F}_{nk^q} \subset \text{conv } \mathcal{F}_{n(k-1)^q} + \text{conv } \mathcal{G}_{n,k}$  for a set  $\mathcal{G}_{n,k}$  containing at most  $nk^q$  elements, each of norm smaller than  $L(n(k-1)^q)^{-1/V}$ . Apply Lemma 2.6.11 to  $\mathcal{G}_{n,k}$  to find an  $Lk^{-2}n^{-W}$ -net over  $\text{conv } \mathcal{G}_{n,k}$  consisting of at most

$$\left( e + ek^{2q/V-4+q} \right)^{2^{2q/V+1} k^{4-2q/V} n}$$

elements. Apply the induction hypothesis to obtain a  $C_{k-1} L n^{-W}$ -net over  $\text{conv } \mathcal{F}_{n(k-1)^q}$  consisting of at most  $e^{D_{k-1} n}$  elements. Combine the nets as before to obtain a  $C_k L n^{-W}$ -net over  $\text{conv } \mathcal{F}_{nk^q}$  consisting of at most  $e^{D_k n}$  elements, for

$$\begin{aligned} C_k &= C_{k-1} + \frac{1}{k^2}, \\ D_k &= D_{k-1} + 2^{2q/V+1} \frac{1 + \log(1 + k^{2q/V-4+q})}{k^{2q/V-4}}. \end{aligned}$$

For  $2q/V - 4 \geq 2$ , the resulting sequences  $C_k$  and  $D_k$  are bounded. ■

Since the symmetric convex hull  $\text{sconv } \mathcal{F}$  of a class  $\mathcal{F}$  is contained in the convex hull of the class  $\mathcal{F} \cup -\mathcal{F} \cup \{0\}$ , the bound of the preceding theorem is valid for  $\text{sconv } \mathcal{F}$  as well. It suffices to note that the covering numbers of the class  $\mathcal{F} \cup -\mathcal{F} \cup \{0\}$  are at most twice the covering numbers of  $\mathcal{F}$  plus 1.

An important ingredient in the proof of the preceding theorem is the following lemma, which gives a useful bound on the covering numbers of the convex hull of an arbitrary finite set of small diameter.

**2.6.11 Lemma.** *Let  $\mathcal{F}$  be an arbitrary set of  $n$  measurable functions  $f: \mathcal{X} \mapsto \mathbb{R}$  of finite  $L_2(Q)$ -diameter  $\text{diam } \mathcal{F}$ . Then for every  $\varepsilon > 0$ ,*

$$N(\varepsilon \text{ diam } \mathcal{F}, \text{conv } \mathcal{F}, L_2(Q)) \leq (e + en\varepsilon^2)^{2/\varepsilon^2}.$$

**Proof.** Write  $\mathcal{F} = \{f_1, \dots, f_n\}$ . For given  $\lambda$  in the  $n$ -dimensional unit simplex and natural number  $k$ , let  $Y_1, \dots, Y_k$  be i.i.d. random elements with  $P(Y_i = f_j) = \lambda_j$  for  $j = 1, \dots, n$ . Then  $EY_i = \sum \lambda_j f_j$  and

$$E\|\bar{Y}_k - EY_1\|_{Q,2}^2 \leq \frac{1}{k} E\|Y_1 - EY_1\|_{Q,2}^2 \leq \frac{1}{k} (\text{diam } \mathcal{F})^2.$$

At least one realization of  $\bar{Y}_k$  must have distance at most  $k^{-1/2} \operatorname{diam} \mathcal{F}$  to the convex combination  $\sum \lambda_j f_j$ . Every realization is an average of the form  $k^{-1} \sum_{i=1}^k f_{i_k}$  (allowing for multiple use of an  $f_i \in \mathcal{F}$ ). There are at most  $\binom{n+k-1}{k}$  of such averages. Conclude that

$$N(k^{-1/2} \operatorname{diam} \mathcal{F}, \operatorname{conv} \mathcal{F}, L_2(Q)) \leq \binom{n+k-1}{k} \leq e^k \left(1 + \frac{n}{k}\right)^k.$$

The last inequality can be proved by using Stirling's inequality with bound. For  $\varepsilon \geq 1$ , the assertion of the lemma is trivial. For  $\varepsilon < 1$ , conclude the proof by choosing the smallest  $k$  such that  $k^{-1/2} \leq \varepsilon$ . ■

**2.6.12 Corollary.** *For any VC-hull class  $\mathcal{F}$  of measurable functions and probability measure  $Q$ ,*

$$\log N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q)) \leq K \left(\frac{1}{\varepsilon}\right)^{2-2V_m^{-1}(\mathcal{F})},$$

for a constant  $K$  that depends only the VC-index  $V_m(\mathcal{F})$  of the VC-subgraph class connected with  $\mathcal{F}$ .

#### 2.6.4 VC-Major Classes

A class of measurable functions is called a *VC-major class* if the sets  $\{x: f(x) > t\}$  with  $f$  ranging over  $\mathcal{F}$  and  $t$  over  $\mathbb{R}$  form a VC-class of sets. Bounded VC-major classes are VC-hull classes.

**2.6.13 Lemma (Bounded VC-major classes).** *A bounded VC-major class is a scalar multiple of a VC-hull class for sets.*

**Proof.** A given function  $f: \mathcal{X} \mapsto [0, 1]$  is the uniform limit of the sequence

$$f_m = \sum_{i=1}^m \frac{1}{m} \mathbf{1}\left\{f > \frac{i}{m}\right\}.$$

Thus a given class of functions  $f: \mathcal{X} \mapsto [0, 1]$  is contained in the pointwise sequential closure of the convex hull of the class of sets of the type  $\{f > t\}$  with  $f$  ranging over  $\mathcal{F}$  and  $t$  over  $\mathbb{R}$ . For a VC-major class, this collection of sets is VC. ■

Since a bounded VC-major class is a VC-hull class, Corollary 2.6.12 shows that it is universally Donsker if suitably measurable. The boundedness can be relaxed considerably by a direct argument.

**2.6.14 Theorem.** Let  $\mathcal{F}$  be a pointwise separable, VC-major class of measurable functions with envelope  $F$  such that  $\int \sqrt{P^*(F > x)} dx < \infty$ . Then  $\mathcal{F}$  is  $P$ -Donsker.

**Proof.** See Dudley and Koltchinskii (1994). ■

## 2.6.5 Examples and Permanence Properties

The first two lemmas of this section give basic methods for generating VC-graph classes. This is followed by a discussion of methods that allow one to build up new classes related to the VC-property from more basic classes.

**2.6.15 Lemma.** Any finite-dimensional vector space  $\mathcal{F}$  of measurable functions  $f: \mathcal{X} \mapsto \mathbb{R}$  is VC-subgraph of index smaller than or equal to  $\dim(\mathcal{F}) + 2$ .

**Proof.** Take any collection of  $n = \dim(\mathcal{F}) + 2$  points  $(x_1, t_1), \dots, (x_n, t_n)$  in  $\mathcal{X} \times \mathbb{R}$ . By assumption, the vectors

$$(f(x_1) - t_1, \dots, f(x_n) - t_n)', \quad f \in \mathcal{F},$$

are contained in a  $\dim(\mathcal{F}) + 1 = (n - 1)$ -dimensional subspace of  $\mathbb{R}^n$ . Any vector  $a \neq 0$  that is orthogonal to this subspace satisfies

$$\sum_{a_i > 0} a_i (f(x_i) - t_i) = \sum_{a_i < 0} (-a_i)(f(x_i) - t_i), \quad \text{for every } f \in \mathcal{F}.$$

Define the sum over an empty set as zero. There exists such a vector  $a$  with at least one strictly positive coordinate. For this vector, the set  $\{(x_i, t_i): a_i > 0\}$  cannot be of the form  $\{(x_i, t_i): t_i < f(x_i)\}$  for some  $f$ , because then the left side of the equation would be strictly positive and the right side nonpositive for this  $f$ . Conclude that the subgraphs of  $\mathcal{F}$  do not pick out the set  $\{(x_i, t_i): a_i > 0\}$ . Hence the subgraphs shatter no set of  $n$  points. ■

**2.6.16 Lemma.** The set of all translates  $\{\psi(x - h): h \in \mathbb{R}\}$  of a fixed monotone function  $\psi: \mathbb{R} \mapsto \mathbb{R}$  is VC of index 2.

**Proof.** By the monotonicity, the subgraphs are linearly ordered by inclusion: if  $\psi$  is nondecreasing, then the subgraph of  $x \mapsto \psi(x - h_1)$  is contained in the subgraph of  $x \mapsto \psi(x - h_2)$  if  $h_1 \geq h_2$ . Any collection of sets with this property has VC-index 2. ■

**2.6.17 Lemma.** Let  $\mathcal{C}$  and  $\mathcal{D}$  be VC-classes of sets in a set  $\mathcal{X}$  and  $\phi: \mathcal{X} \mapsto \mathcal{Y}$  and  $\psi: \mathcal{Z} \mapsto \mathcal{X}$  fixed functions. Then

- (i)  $\mathcal{C}^c = \{C^c : C \in \mathcal{C}\}$  is VC;
- (ii)  $\mathcal{C} \cap \mathcal{D} = \{C \cap D : C \in \mathcal{C}, D \in \mathcal{D}\}$  is VC;
- (iii)  $\mathcal{C} \cup \mathcal{D} = \{C \cup D : C \in \mathcal{C}, D \in \mathcal{D}\}$  is VC;
- (iv)  $\phi(\mathcal{C})$  is VC if  $\phi$  is one-to-one;
- (v)  $\psi^{-1}(\mathcal{C})$  is VC;
- (vi) the sequential closure of  $\mathcal{C}$  for pointwise convergence of indicator functions is VC.

For VC-classes  $\mathcal{C}$  and  $\mathcal{D}$  in sets  $\mathcal{X}$  and  $\mathcal{Y}$ ,

- (vii)  $\mathcal{C} \times \mathcal{D}$  is VC in  $\mathcal{X} \times \mathcal{Y}$ .

**Proof.** The set  $C^c$  picks out the points of a given set  $x_1, \dots, x_n$  that  $C$  does not pick out. Thus if  $\mathcal{C}$  shatters a given set of points, so does  $\mathcal{C}^c$ . This proves (i) (and shows that the indices of  $\mathcal{C}$  and  $\mathcal{C}^c$  are equal). From  $n$  points  $\mathcal{C}$  can pick out  $O(n^{V(\mathcal{C})-1})$  subsets. From each of these subsets,  $\mathcal{D}$  can pick out at most  $O(n^{V(\mathcal{D})-1})$  further subsets. Thus  $\mathcal{C} \cap \mathcal{D}$  can pick out  $O(n^{V(\mathcal{C})+V(\mathcal{D})-2})$  subsets. For large  $n$ , this is certainly smaller than  $2^n$ . This proves (ii). Next (iii) follows from a combination of (i) and (ii), since  $\mathcal{C} \cup \mathcal{D} = (\mathcal{C}^c \cap \mathcal{D}^c)^c$ . For (iv) note first that if  $\phi(\mathcal{C})$  shatters  $\{y_1, \dots, y_n\}$ , then each  $y_i$  must be in the range of  $\phi$  and there exist  $x_1, \dots, x_n$  such that  $\phi$  is a bijection between  $x_1, \dots, x_n$  and  $y_1, \dots, y_n$ . Thus  $\mathcal{C}$  must shatter  $x_1, \dots, x_n$ . For (v) the argument is analogous: if  $\psi^{-1}(\mathcal{C})$  shatters  $z_1, \dots, z_n$ , then all  $\psi(z_i)$  must be different and the restriction of  $\psi$  to  $z_1, \dots, z_n$  is a bijection on its range.

To prove (vi) take any set of points  $x_1, \dots, x_n$  and any set  $\bar{C}$  from the sequential closure. If  $\bar{C}$  is the pointwise limit of a net  $C_\alpha$ , then for sufficiently large  $\alpha$  the equality  $1\{\bar{C}\}(x_i) = 1\{C_\alpha\}(x_i)$  is valid for each  $i$ . For such  $\alpha$  the set  $C_\alpha$  picks out the same subset as  $\bar{C}$ .

For (vii) note first that  $\mathcal{C} \times \mathcal{Y}$  and  $\mathcal{X} \times \mathcal{D}$  are VC-classes. Then by (ii) so is their intersection  $\mathcal{C} \times \mathcal{D}$ . ■

**2.6.18 Lemma.** Let  $\mathcal{F}$  and  $\mathcal{G}$  be VC-subgraph classes of functions on a set  $\mathcal{X}$  and  $g: \mathcal{X} \mapsto \mathbb{R}$ ,  $\phi: \mathbb{R} \mapsto \mathbb{R}$ , and  $\psi: \mathcal{Z} \mapsto \mathcal{X}$  fixed functions. Then

- (i)  $\mathcal{F} \wedge \mathcal{G} = \{f \wedge g : f \in \mathcal{F}, g \in \mathcal{G}\}$  is VC-subgraph;
- (ii)  $\mathcal{F} \vee \mathcal{G}$  is VC-subgraph;
- (iii)  $\{\mathcal{F} > 0\} = \{\{f > 0\} : f \in \mathcal{F}\}$  is VC;
- (iv)  $-\mathcal{F}$  is VC;
- (v)  $\mathcal{F} + g = \{f + g : f \in \mathcal{F}\}$  is VC-subgraph;
- (vi)  $\mathcal{F} \cdot g = \{fg : f \in \mathcal{F}\}$  is VC-subgraph;
- (vii)  $\mathcal{F} \circ \psi = \{f(\psi) : f \in \mathcal{F}\}$  is VC-subgraph;
- (viii)  $\phi \circ \mathcal{F}$  is VC-subgraph for monotone  $\phi$ .

**Proof.** The subgraphs of  $f \wedge g$  and  $f \vee g$  are the intersection and union of the subgraphs of  $f$  and  $g$ , respectively. Hence (i) and (ii) are consequences of the preceding lemma. For (iii) note that the sets  $\{f > 0\}$  are one-to-one

images of the intersections of the (open) subgraphs with the set  $\mathcal{X} \times \{0\}$ . Thus the class  $\{\mathcal{F} > 0\}$  is VC by (ii) and (iv) of the preceding lemma.

The subgraphs of the class  $-\mathcal{F}$  are the images of the open supergraphs of  $\mathcal{F}$  under the map  $(x, t) \mapsto (x, -t)$ . The open supergraphs are the complements of the closed subgraphs, which are VC by Problem 2.6.10. Now (iv) follows from the previous lemma. For (v) it suffices to note that the subgraphs of the class  $\mathcal{F} + g$  shatter a given set of points  $(x_1, t_1), \dots, (x_n, t_n)$  if and only if the subgraphs of  $\mathcal{F}$  shatter the set  $(x_i, t_i - g(x_i))$ . The subgraph of the function  $fg$  is the union of the sets

$$\begin{aligned} C^+ &= \{(x, t): t < f(x)g(x), g(x) > 0\}, \\ C^- &= \{(x, t): t < f(x)g(x), g(x) < 0\}, \\ C^0 &= \{(x, t): t < 0, g(x) = 0\}. \end{aligned}$$

It suffices to show that these sets are VC in  $(\mathcal{X} \cap \{g > 0\}) \times \mathbb{R}$ ,  $(\mathcal{X} \cap \{g < 0\}) \times \mathbb{R}$ , and  $(\mathcal{X} \cap \{g = 0\}) \times \mathbb{R}$ , respectively (Problem 2.6.12). Now, for instance,  $\{i: (x_i, t_i) \in C^-\}$  is the set of indices of the points  $(x_i, t_i/g(x_i))$  picked out by the open supergraphs of  $\mathcal{F}$ . These are the complements of the closed subgraphs and hence form a VC class.

The subgraphs of the class  $\mathcal{F} \circ \psi$  are the inverse images of the subgraphs of functions in  $\mathcal{F}$  under the map  $(z, t) \mapsto (\psi(z), t)$ . Thus (v) of the previous lemma implies (vii).

For (viii) suppose the subgraphs of  $\phi \circ \mathcal{F}$  shatter the set of points  $(x_1, t_1), \dots, (x_n, t_n)$ . Choose  $f_1, \dots, f_m$  from  $\mathcal{F}$  such that the subgraphs of the functions  $\phi \circ f_j$  pick out all  $m = 2^n$  subsets. For each fixed  $i$ , define  $s_i = \max\{f_j(x_i): \phi(f_j(x_i)) \leq t_i\}$ . Then  $s_i < f_j(x_i)$  if and only if  $t_i < \phi \circ f_j(x_i)$ , for every  $i$  and  $j$ , and the subgraphs of  $f_1, \dots, f_m$  shatter the points  $(x_i, s_i)$ . ■

**2.6.19 Lemma.** *If  $\mathcal{F}$  is VC-major, then the class of functions  $h \circ f$ , with  $h$  ranging over the monotone functions  $h: \mathbb{R} \mapsto \mathbb{R}$  and  $f$  over  $\mathcal{F}$ , is VC-major.*

**2.6.20 Lemma.** *If  $\mathcal{F}$  and  $\mathcal{G}$  are VC-hull classes for sets (in particular, uniformly bounded VC-major classes), then  $\mathcal{FG}$  is a VC-hull class for sets.*

**Proofs.** The set  $\{h \circ f > t\}$  can be rewritten as  $\{f \cdot h^{-1}(t)\}$  for a suitable inverse  $h$  and  $\cdot >$  meaning either  $>$  or  $\geq$  depending on  $t$ . Now the sets  $\{f \geq t\}$  with  $f$  ranging over  $\mathcal{F}$  and  $t$  over  $\mathbb{R}$  form a VC-class, since they are in the sequential pointwise closure of the same sets defined with strict inequality.

If  $\mathcal{F}$  and  $\mathcal{G}$  are VC-hull classes for sets, then any function  $fg$  can be approximated by a product of two convex combinations of indicator functions of sets from VC-classes. Since the set of pairwise intersections of sets from two VC-classes is VC, the products of the approximations are convex combinations of indicators of sets from a VC-class. ■

**2.6.21 Example.** The set of all monotone functions  $f: \mathbb{R} \mapsto [0, 1]$  is Donsker for every probability measure.

**2.6.22 Lemma.** Let  $F: \mathcal{X} \mapsto \mathbb{R}$  be a fixed, nonnegative function. Then the class of functions  $x \mapsto t1\{F(x) \geq t\}$  with  $t$  ranging over  $\mathbb{R}$  is VC of index  $V(\mathcal{F}) \leq 3$ .

**Proof.** Consider three arbitrary points  $(x_1, t_1)$ ,  $(x_2, t_2)$ , and  $(x_3, t_3)$  in  $\mathcal{X} \times [0, \infty)$  such that  $F(x_1) \leq F(x_2) \leq F(x_3)$ . Suppose the three-point set is shattered by  $\mathcal{F}$ . Since  $\mathcal{F}$  selects the single point  $(x_1, t_1)$ , there exists  $s_1$  with

$$(t_1 < s_1 \leq F(x_1)) \wedge (s_1 \leq t_2 \text{ or } s_1 > F(x_2)) \wedge (s_1 \leq t_3 \text{ or } s_1 > F(x_3)).$$

Because  $F(x_i)$  is increasing in  $i$ , this can be reduced to

$$(t_1 < s_1 \leq F(x_1)) \wedge (s_1 \leq t_2) \wedge (s_1 \leq t_3).$$

Since  $\mathcal{F}$  selects the single point  $(x_2, t_2)$ , there exists  $s_2$  with

$$(s_2 \leq t_1 \text{ or } s_2 > F(x_1)) \wedge (t_2 < s_2 \leq F(x_2)) \wedge (s_2 \leq t_3 \text{ or } s_2 > F(x_3)).$$

In the first part of this logical statement,  $s_2 \leq t_1$  is impossible because  $t_1 < s_1 \leq t_2 < s_2$  in view of the first two parts of the preceding statement and the middle part of the last statement. Thus the last statement can be reduced to

$$(s_2 > F(x_1)) \wedge (t_2 < s_2 \leq F(x_2)) \wedge (s_2 \leq t_3).$$

This implies that  $t_3 > F(x_1)$ . Then  $\mathcal{F}$  cannot select the two-point set  $\{(x_1, t_1), (x_3, t_3)\}$ . ■

**2.6.23 Example.** For every strictly increasing function  $\phi$  and arbitrary functions  $F \geq 0$  and  $G$ , the set of functions  $x \mapsto \phi(t)G(x)1\{F(x) \geq t\}$  (with  $t$  ranging over  $\mathbb{R}$ ) is VC of index  $V(\mathcal{F}) \leq 3$ .

Indeed, for  $G \equiv 1$  this class is contained in the class of functions  $x \mapsto s1\{\phi \circ F(x) \geq s\}$ . The general case follows from Lemma 2.6.18.

Consequently, the class of functions  $x \mapsto t^{k-1}F(x)1\{F(x) \geq t\}$  is Donsker for every measurable function  $F$  with  $\int F^{2k} dP < \infty$  by the uniform entropy central limit theorem.

## Problems and Complements

1. Every collection of sets that is pre-Gaussian for every underlying measure is VC.

[Hint: Dudley (1984), 11.4.1.]

2. Unlike the case of sets, the VC-subgraph property does not characterize universal Donsker classes of functions. There exist bounded universal Donsker classes that do not even satisfy the uniform entropy condition.

[Hint: Dudley (1987).]

3. Let  $\mathcal{F}$  be a class of measurable functions such that  $D(\varepsilon, \mathcal{F}, L_r(Q)) \leq g(\varepsilon)$  for every empirical type probability measure  $Q$  and a fixed function  $g(\varepsilon)$ . Then the bound is valid for every probability measure.

[Hint: If  $D(\varepsilon, \mathcal{F}, L_r(P)) = m$ , then there are functions  $f_1, \dots, f_m$  such that  $P|f_i - f_j|^r > \varepsilon^r$  for every  $i \neq j$ . By the strong law of large numbers,  $\mathbb{P}_n|f_i - f_j|^r \rightarrow P|f_i - f_j|^r$  almost surely, for every  $(i, j)$ . Thus there exists  $\omega$  and  $n$  such that  $\mathbb{P}_n(\omega)|f_i - f_j|^r > \varepsilon^r$ , for every  $i \neq j$ .]

4. There exists a short proof of the fact that for any VC-class of sets  $\mathcal{C}$  and  $\delta > 0$ ,

$$N(\varepsilon, \mathcal{C}, L_r(Q)) \leq K \left( \frac{1}{\varepsilon} \right)^{r(V(\mathcal{C})-1+\delta)},$$

for any probability measure  $Q$ ,  $r \geq 1$ ,  $0 < \varepsilon < 1$ , and a constant  $K$  depending on  $V(\mathcal{C})$  and  $\delta$  only.

[Hint: Take any subcollection of sets  $C_1, \dots, C_m$  from  $\mathcal{C}$  such that  $Q(C_i \Delta C_j) > \varepsilon$  for every pair  $i \neq j$ . Generate a sample  $X_1, \dots, X_n$  from  $Q$ . Two sets  $C_i$  and  $C_j$  pick out the same subset from a realization of the sample if and only if no  $X_k$  falls in the symmetric difference  $C_i \Delta C_j$ . If every symmetric difference contains a point of the sample, then all  $C_i$  pick out a different subset from the sample. In that case  $\mathcal{C}$  picks out at least  $m$  subsets from  $X_1, \dots, X_n$ . The probability that this event does not occur is bounded by

$$\begin{aligned} \sum_{i < j} Q(X_k \notin C_i \Delta C_j \text{ for every } k) &\leq \binom{m}{2} (1 - Q(C_i \Delta C_j))^n \\ &\leq \binom{m}{2} (1 - \varepsilon)^n. \end{aligned}$$

For sufficiently large  $n$ , the last expression is strictly less than 1. For such  $n$  there exists a set of  $n$  points from which  $\mathcal{C}$  picks out at least  $m$  subsets. In other words, for  $n > -\log \binom{m}{2} / \log(1 - \varepsilon)$ , one has the first inequality in

$$m \leq \max_{x_1, \dots, x_n} \Delta_n(\mathcal{C}, x_1, \dots, x_n) \leq K n^{V(\mathcal{C})-1}.$$

The last inequality follows from the previous corollary and the constant  $K$  depends only on  $V(\mathcal{C})$ . Since  $-\log(1 - \varepsilon) > \varepsilon$ , we can take  $n = 3(\log m)/\varepsilon$  and obtain

$$m \leq K \left( \frac{3 \log m}{\varepsilon} \right)^{V(\mathcal{C})-1}.$$

Since  $\log m$  is bounded by a constant times  $m^\delta$ , it follows that  $m^{1-\delta}$  is bounded by a constant times  $(3/\varepsilon)^{V(C)-1}$ .]

5. The edges of the graph with nodes  $\mathcal{Z} \subset \{0,1\}^n$  representing a VC-class  $\mathcal{C}$  of subsets of a set of points  $\{x_1, \dots, x_n\}$  can be directed in such a manner that at most  $V(\mathcal{C}) - 1$  arrows are positively incident with each node.

[Hint: The number of edges of a hereditary collection  $\mathcal{C}$  can be enumerated by listing for each node  $v$  the edges pointing inward: edges  $\{v, w\}$  such that  $w$  has a zero in the (one) position in which it differs from  $v$ . Since every subset of hereditary class has at most  $V(\mathcal{C}) - 1$  points, it follows that  $\#\mathcal{E}/\#\mathcal{Z} \leq V(\mathcal{C}) - 1$ . By repeated application of the operators  $T_i$ , an arbitrary collection of sets can be transformed into a hereditary class. The operations do not decrease the quotient  $\#\mathcal{E}/\#\mathcal{Z}$ : the number of edges may increase, while the number of nodes remains the same. Conclude that for any VC-collection of sets,  $\#\mathcal{E}/\#\mathcal{Z} \leq V(\mathcal{C}) - 1$ . Since a subcollection of a given VC-class is VC of no greater index, it follows that  $\#\mathcal{E}' \leq \#\mathcal{Z}'(V(\mathcal{C}) - 1)$  for every subgraph  $(\mathcal{Z}', \mathcal{E}')$  as well.]

Now apply Hall's marriage lemma [e.g. Dudley (1989), Theorem 11.6.1, page 318], representing directing an edge  $\{v, w\}$  as marrying the edge to one of the nodes  $v$  and  $w$ . In fact, let the eligible partners for  $\{v, w\}$  be the collection of  $V(\mathcal{C}) - 1$  copies of  $v$  and  $V(\mathcal{C}) - 1$  copies of  $w$ . Then the total number of partners for a given set of edges is at least  $V(\mathcal{C}) - 1$  times the number of edges. By the marriage lemma, successful marriage is possible.]

6. For every pair of integers  $S, r \geq 1$ , and  $n = Sr$ , there exists a subset  $\mathcal{Z} \subset \{0,1\}^n$  such that, for all  $1 \leq k \leq n$  and  $S = V(\mathcal{C}) - 1$ ,

$$D(k/n, \mathcal{Z}, d) \geq \left( \frac{n}{2e(k+S)} \right)^S.$$

[Hint: Start with the subset  $\mathcal{W} = \{(0, 0, \dots, 0), (1, 0, \dots, 0), (1, 1, \dots, 0), \dots, (1, 1, 1, \dots, 1)\}$  of  $\{0,1\}^r$ , and let  $\mathcal{Z} = \mathcal{W}^S$  be the set of all vectors in  $\{0,1\}^n$  obtained by concatenating  $S$  vectors from  $\mathcal{W}$ .]

7. Every VC-class  $\mathcal{C}$  of sets satisfies  $\Delta_n(\mathcal{C}, x_1, \dots, x_n) \leq (ne/(V(\mathcal{C}) - 1))^{V(\mathcal{C})-1}$  for  $n \geq V(\mathcal{C}) - 1$ .

[Hint: Consider a random variable  $Y$  with a binomial distribution with parameters  $n$  and  $1/2$ . Bound  $P(Y \leq k)$  by  $Er^{Y-k}$  for a simple (not optimal) choice of  $r$ .]

8. Let  $Q$  be a finitely discrete probability measure supported on  $x_1, \dots, x_n$ . Then for any collection of sets, one has  $\Delta_n(\mathcal{C}, x_1, \dots, x_n) = N(\varepsilon, \mathcal{C}, L_r(Q))$  for all  $\varepsilon \leq \inf Q\{x\}^{1/r}$ .

[Hint: The  $L_r$ -distance equals  $Q^{1/r}(C \Delta D)$  and can be less than a sufficiently small  $\varepsilon$  only if  $C \Delta D$  does not contain a support point.]

9. If a collection of sets  $\mathcal{C}$  is a VC-class, then the collection of indicators of sets in  $\mathcal{C}$  is a VC-subgraph class of the same index.

- 10. (Open and closed subgraphs)** For a set  $\mathcal{F}$  of measurable functions, define “closed” and “open” subgraphs by  $\{(x, t): t \leq f(x)\}$  and  $\{(x, t): t < f(x)\}$ , respectively. Then the collection of “closed” subgraphs has the same VC-index as the collection of “open” subgraphs. Consequently, “closed” and “open” are equivalent in the definition of a VC-subgraph class.

[**Hint:** Suppose the “closed” subgraphs shatter the set  $(x_1, t_1), \dots, (x_n, t_n)$ . Choose  $f_1, \dots, f_m$  whose “closed” subgraphs pick out all  $m = 2^n$  subsets. Set  $2\varepsilon = \inf\{t_i - f_j(x_i): t_i - f_j(x_i) > 0\}$ . Then the “open” subgraphs shatter the set  $(x_1, t_1 - \varepsilon), \dots, (x_n, t_n - \varepsilon)$ . The converse can be argued in a similar manner.]

- 11. (Between graphs)** For a set  $\mathcal{F}$  of measurable functions, define the “between” graphs as the sets  $\{(x, t): 0 \leq t \leq f(x) \text{ or } f(x) \leq t \leq 0\}$ . The “between” graphs form a VC-class of sets if and only if  $\mathcal{F}$  is a VC-subgraph class.

[**Hint:** The “closed” subgraphs intersected with the set  $\{t \geq 0\}$  yields a VC-class of sets which is the positive half of the “between” graphs. The “open” subgraphs intersected with the set  $\{t \leq 0\}$  and next complemented within  $\{t \leq 0\}$  give the lower parts of the “between” graphs.]

- 12.** If  $\mathcal{X}$  is the union of finitely many disjoint sets  $\mathcal{X}_i$ , and  $\mathcal{C}_i$  is a VC-class of subsets of  $\mathcal{X}_i$  for each  $i$ , then  $\sqcup \mathcal{C}_i$  is a VC-class in  $\sqcup \mathcal{X}_i$  of index  $\sum V(\mathcal{C}_i)$ .

- 13.** If  $\mathcal{F}$  is a VC-major class, then the sets  $\{x: f(x) \geq t\}$ , with  $f$  ranging over  $\mathcal{F}$  and  $t$  over  $\mathbb{R}$ , form a VC-class.

[**Hint:** The set  $\{x: f(x) \geq t\}$  is the intersection of the sets  $\{x: f(x) \geq t - n^{-1}\}$ . Use Lemma 2.6.17(vi).]

- 14.** The collection of all *half-spaces* in  $\mathbb{R}^d$  is a VC-class of index  $d + 2$ . A half-space is a set of the form  $\{x \in \mathbb{R}^d: \langle x, u \rangle \leq c\}$  for fixed  $u \in \mathbb{R}^d$  and  $c \in \mathbb{R}$ . The collection of all closed balls in  $\mathbb{R}^d$  is a VC-class of index  $d + 2$ .

[**Hint:** Use Lemma 2.6.15. See Dudley (1979).]

- 15.** The class of all closed convex subsets in  $\mathbb{R}^d$  is not VC for  $d \geq 2$ . The same is true for the collection of all open convex sets.

[**Hint:** Any set of  $n$  points on the rim of the unit ball is shattered by the closed convex sets. The closed convex sets are in the sequential closure of the open convex sets.]

- 16.** The set of all open polygons with extreme points on the rim of the unit circle in  $\mathbb{R}^2$  is not VC.

[**Hint:** Form a regular polygon with its  $n$  extreme points on the rim of the unit circle. Choose  $n$  points in the  $n$  gaps between the polygon and the unit circle.]

17. If  $\mathcal{C}$  is a VC-class of subsets of  $\mathcal{X}$  and  $\phi: \mathcal{X} \mapsto \mathcal{Y}$  is an arbitrary map, then  $\phi(\mathcal{C})$  need not be VC.

[Hint: Take  $\mathcal{Y}$  any infinite set and let  $\mathcal{X} = \mathcal{Y} \times \mathbb{N}$  with  $\phi(y, n) = y$ . Let  $\mathcal{C}$  be the collection of all subsets of exactly one point of  $\mathcal{Y} \times 1$ , all subsets of exactly 2 points of  $\mathcal{Y} \times 2$ , etcetera. Then  $\phi(\mathcal{C})$  consists of all finite subsets of  $\mathcal{Y}$ , but no subset of two points in  $\mathcal{X}$  is shattered.]

18. The set of all monotone functions  $f: \mathbb{R} \mapsto [0, 1]$  is VC-hull but not VC-subgraph.

[Hint: Any set of points  $(x_i, t_i)$  with both  $x_i$  and  $t_i$  strictly increasing in  $i$  is shattered.]

19. For a VC-subgraph class  $\mathcal{F}$ , the class  $\{f - Pf: f \in \mathcal{F}\}$  is not necessarily VC-subgraph.

[Hint: For any countable collection of functions  $g_n: \mathcal{X} \mapsto [0, 1]$ , the subgraphs of the collection  $g_n + n$  form a linearly ordered set. Hence the functions  $g_n + n$  form a VC-subgraph class.]

20. The class of functions of the form  $x \mapsto c1_{(a,b]}(x)$  with  $a, b$ , and  $c > 0$  ranging over  $\mathbb{R}$  is VC of index 3.

[Hint: Any  $f$  whose subgraph picks out the subset  $\{(x_1, t_1), (x_3, t_3)\}$  from three points  $(x_i, t_i)$  with  $x_1 \leq x_2 \leq x_3$  also picks out  $(x_2, t_2)$  unless  $t_2 > t_1 \vee t_3$ . If a set of four points with nondecreasing  $x_i$  is shattered, it follows that  $t_2 > t_1 \vee t_2$ , but also  $t_3 > t_2 \vee t_4$ ,  $t_2 > t_1 \vee t_4$ , and  $t_3 > t_1 \vee t_4$ .]

21. The “Box-Cox family of transformations”  $\mathcal{F} = \{f_\lambda: (0, \infty) \mapsto \mathbb{R}: \lambda \in \mathbb{R} - \{0\}\}$ , with  $f_\lambda(x) = (x^\lambda - 1)/\lambda$ , is a VC-subgraph class.

[Hint: The “between” graph class of sets (as defined in Exercise 2.6.11) is the class of subsets  $\mathcal{C} = \{C_\lambda: \lambda \neq 0\}$  of  $\mathbb{R}^2$ , where  $C_\lambda = \{(x, t): 0 \leq t \leq (x^\lambda - 1)/\lambda\}$ . Examine the “dual class” of subsets of  $\mathbb{R}$  given by  $\mathcal{D} = \{D_{(x,t)}: (x, t) \in (0, \infty) \times (0, \infty)\}$ , where

$$D_{(x,t)} = \{\lambda \neq 0: (x, t) \in C_\lambda\} = \{\lambda \neq 0: 0 \leq t \leq (x^\lambda - 1)/\lambda\}.$$

Show that  $\mathcal{D}$  is a VC-class of sets, and apply Assouad (1983), page 246, Proposition 2.12.]

## 2.7

# Bracketing Numbers

While the VC-theory gives control over the entropy numbers of many interesting classes through simple combinatorial arguments, results on bracketing numbers can be found in approximation theory. This section gives examples. In some cases, the bracketing numbers are actually uniform in the underlying measure.

### 2.7.1 Smooth Functions and Sets

For  $\alpha > 0$ , we study the class of all functions on a bounded set  $\mathcal{X}$  in  $\mathbb{R}^d$  that possess uniformly bounded partial derivatives up to order  $\underline{\alpha}$  (the greatest integer smaller than  $\alpha$ ) and whose highest partial derivatives are Lipschitz of order  $\alpha - \underline{\alpha}$ . A simple example is Lipschitz functions of some order  $0 < \alpha \leq 1$  (for which  $\underline{\alpha} = 0$  even if  $\alpha = 1$ !). For a more precise description, define for any vector  $k = (k_1, \dots, k_d)$  of  $d$  integers the differential operator

$$D^k = \frac{\partial^k}{\partial x_1^{k_1} \cdots \partial x_d^{k_d}},$$

where  $k_+ = \sum k_i$ . Then for a function  $f: \mathcal{X} \mapsto \mathbb{R}$ , let

$$\|f\|_\alpha = \max_{k_+ \leq \underline{\alpha}} \sup_x |D^k f(x)| + \max_{k_+ = \underline{\alpha}} \sup_{x,y} \frac{|D^k f(x) - D^k f(y)|}{\|x - y\|^{\alpha - \underline{\alpha}}},$$

where the suprema are taken over all  $x, y$  in the interior of  $\mathcal{X}$  with  $x \neq y$ . Let  $C_M^\alpha(\mathcal{X})$  be the set of all continuous functions  $f: \mathcal{X} \mapsto \mathbb{R}$  with  $\|f\|_\alpha \leq M$ .<sup>†</sup>

---

<sup>†</sup> Note the inconsistency in notation. The notation  $\|\cdot\|_\alpha$  is used only with  $\alpha < \infty$  to define the present classes of functions. The notation  $\|\cdot\|_\infty$  always denotes the supremum norm.

Bounds on the entropy numbers of the classes  $C_1^\alpha(\mathcal{X})$  with respect to the supremum norm  $\|\cdot\|_\infty$  were among the first known after the introduction of the concept of covering numbers. These readily yield bounds on the  $L_r(Q)$ -bracketing numbers for the present classes of functions, as well as for the class of subgraphs corresponding to them. The latter are classes of sets with “smooth boundaries.”

**2.7.1 Theorem.** *Let  $\mathcal{X}$  be a bounded, convex subset of  $\mathbb{R}^d$  with nonempty interior. There exists a constant  $K$  depending only on  $\alpha$  and  $d$  such that*

$$\log N(\varepsilon, C_1^\alpha(\mathcal{X}), \|\cdot\|_\infty) \leq K \lambda(\mathcal{X}^1) \left( \frac{1}{\varepsilon} \right)^{d/\alpha},$$

for every  $\varepsilon > 0$ , where  $\lambda(\mathcal{X}^1)$  is the Lebesgue measure of the set  $\{x : \|x - \mathcal{X}\| < 1\}$ .

**Proof.** Write  $\lesssim$  for “less than equal a constant times,” where the constant depends on  $\alpha$  and  $d$  only. Let  $\beta$  denote the greatest integer strictly smaller than  $\alpha$ .

Since the functions in  $C_1^\alpha(\mathcal{X})$  are continuous on  $\mathcal{X}$  by assumption, it may be assumed without loss of generality that  $\mathcal{X}$  is open, so that Taylor’s theorem applies everywhere on  $\mathcal{X}$ .

Fix  $\delta = \varepsilon^{1/\alpha} \leq 1$ , and form a  $\delta$ -net for  $\mathcal{X}$  of points  $x_1, \dots, x_m$  contained in  $\mathcal{X}$ . The number  $m$  of points can be taken to satisfy  $m \lesssim \lambda(\mathcal{X}^1)/\delta^d$ . For each vector  $k = (k_1, \dots, k_d)$  with  $k_* \leq \beta$ , form for each  $f$  the vector

$$A_k f = \left( \left\lfloor \frac{D^k f(x_1)}{\delta^{d-k_*}} \right\rfloor, \dots, \left\lfloor \frac{D^k f(x_m)}{\delta^{d-k_*}} \right\rfloor \right).$$

Then the vector  $\delta^{d-k_*} A_k f$  consists of the values  $D^k f(x_i)$  discretized on a grid of mesh-width  $\delta^{d-k_*}$ .

If a given pair of functions  $f$  and  $g$  satisfy  $A_k f = A_k g$  for each  $k$  with  $k_* \leq \beta$ , then  $\|f - g\|_\infty \lesssim \varepsilon$ . Indeed, for each  $x$  there exists an  $x_i$  with  $\|x - x_i\| \leq \delta$ . By Taylor’s theorem,

$$(f - g)(x) = \sum_{k_* \leq \beta} D^k(f - g)(x_i) \frac{(x - x_i)^k}{k!} + R,$$

where  $|R| \lesssim \|x - x_i\|^\alpha$ , because the highest-order partial derivatives are uniformly bounded. In the multidimensional case, the notation  $h^k/k!$  is short for  $\prod_{i=1}^d h_i^{k_i}/k_i!$ . Thus

$$|f - g|(x) \lesssim \sum_{k_* \leq \beta} \delta^{d-k_*} \frac{\delta^k}{k!} + \delta^\alpha \leq \delta^\alpha (e^d + 1).$$

It follows that there exists a constant  $C$  depending on  $\alpha$  and  $d$  only such that the covering number  $N(C\varepsilon, C_1^\alpha(\mathcal{X}), \|\cdot\|_\infty)$  is bounded by the number of different matrices

$$Af = \begin{pmatrix} A_{0,0,\dots,0}f \\ A_{1,0,\dots,0}f \\ \vdots \\ A_{0,0,\dots,\beta}f \end{pmatrix},$$

when  $f$  ranges over  $C_1^\alpha(\mathcal{X})$ . Each row in the matrix  $Af$  is one of the vectors  $A_k f$  for  $k \leq \beta$ . Hence the number of rows is certainly smaller than  $(\beta+1)^d$ . By definition of each  $A_k f$  and the fact that  $|D^k f(x_i)| \leq 1$  for each  $i$ , the number of possible values of each element in the row  $A_k f$  is bounded by  $2/\delta^{\alpha-k} + 1$ , which does not exceed  $2\delta^{-\alpha} + 1$ . Thus each column of the matrix can have at most  $(2\delta^{-\alpha} + 1)^{(\beta+1)^d}$  different values.

Assume without loss of generality that  $x_1, \dots, x_m$  have been chosen and ordered in such a way that for each  $j > 1$  there is an index  $i < j$  such that  $\|x_i - x_j\| < 2\delta$ . Then use the crude bound obtained previously for the first column only. For each later column, indexed by  $x_j$ , there exists a previous  $x_i$  with  $\|x_i - x_j\| < 2\delta$ . By Taylor's theorem,

$$D^k f(x_j) = \sum_{k+l \leq \beta} D^{k+l} f(x_i) \frac{(x_i - x_j)^l}{l!} + R,$$

where  $|R| \lesssim \|x_i - x_j\|^{\alpha-k}$ . Thus with  $B_k f = \delta^{\alpha-k} \cdot A_k f$ ,

$$\begin{aligned} & \left| D^k f(x_j) - \sum_{k+l \leq \beta} B_{k+l} f(x_i) \frac{(x_i - x_j)^l}{l!} \right| \\ & \lesssim \sum_{k+l \leq \beta} |B_{k+l} f(x_i) - D^{k+l} f(x_i)| \frac{|x_i - x_j|^l}{l!} + \delta^{\alpha-k}. \\ & \leq \sum_{k+l \leq \beta} \delta^{\alpha-k-l} \frac{\delta^l}{l!} + \delta^{\alpha-k} \lesssim \delta^{\alpha-k}. \end{aligned}$$

Thus given the values in the  $i$ th column of  $Af$ , the values  $D^k f(x_j)$  range over an interval of length proportional to  $\delta^{\alpha-k}$ . It follows that the values in the  $j$ th column of  $Af$  range over integers in an interval of length proportional to  $\delta^{k-\alpha} \delta^{\alpha-k} = 1$ . Consequently, there exists a constant  $C$  depending only on  $\alpha$  and  $d$  such that

$$\#Af \leq (2\delta^{-\alpha} + 1)^{(\beta+1)^d} C^{m-1}.$$

The theorem follows upon replacing  $\delta$  by  $\varepsilon^{1/\alpha}$  and  $m$  by its upper bound  $\lambda(\mathcal{X}^1) \varepsilon^{-d/\alpha}$ , respectively, taking logarithms, and bounding  $\log(1/\varepsilon)$  by a constant times  $(1/\varepsilon)^{d/\alpha}$ . ■

**2.7.2 Corollary.** Let  $\mathcal{X}$  be a bounded, convex subset of  $\mathbb{R}^d$  with nonempty interior. There exists a constant  $K$  depending only on  $\alpha$ ,  $\text{diam } \mathcal{X}$ , and  $d$  such that

$$\log N_{[]}(\varepsilon, C_1^\alpha(\mathcal{X}), L_r(Q)) \leq K \left( \frac{1}{\varepsilon} \right)^{d/\alpha},$$

for every  $r \geq 1$ ,  $\varepsilon > 0$ , and probability measure  $Q$  on  $\mathbb{R}^d$ .

**Proof.** Let  $f_1, \dots, f_p$  be the centers of  $\|\cdot\|_\infty$ -balls of radius  $\varepsilon$  that cover  $C_1^\alpha(\mathcal{X})$ . Then the brackets  $[f_i - \varepsilon, f_i + \varepsilon]$  cover  $C_1^\alpha(\mathcal{X})$ . Each bracket has  $L_r(Q)$ -size at most  $2\varepsilon$ . By the previous theorem,  $\log p$  can be chosen smaller than the given polynomial in  $1/\varepsilon$ . ■

The corollary, together with either the bracketing central limit theorem or the uniform entropy theorem, implies that  $C_1^\alpha[0, 1]^d$  is universally Donsker for  $\alpha > d/2$ . For instance, on the unit interval in the line, uniformly bounded and uniformly Lipschitz of order  $> 1/2$  suffices and in the unit square it suffices that the partial derivatives exist and satisfy a Lipschitz condition.

For the collection of subgraphs to be Donsker, a smoothness condition on the underlying measure is needed, in addition to sufficient smoothness of the graphs. In this case smoothness of the graphs is of little help if the probability mass is distributed in an erratic manner. The following result implies that the subgraphs (contained in  $\mathbb{R}^{d+1}$ ) of the functions  $C_1^\alpha[0, 1]^d$  are  $P$ -Donsker for Lebesgue-dominated measures  $P$  with bounded density, provided  $\alpha > d$ . For instance, for the sets cut out in the plane by functions  $f: [0, 1] \mapsto [0, 1]$ , a uniform Lipschitz condition of any order on the derivatives suffices.

**2.7.3 Corollary.** Let  $\mathcal{C}_{\alpha,d}$  be the collection of subgraphs of  $C_1^\alpha[0, 1]^d$ . There exists a constant  $K$  depending only on  $\alpha$  and  $d$  such that

$$\log N_{[]}(\varepsilon, \mathcal{C}_{\alpha,d}, L_r(Q)) \leq K \|q\|_\infty^{d/\alpha} \left( \frac{1}{\varepsilon} \right)^{dr/\alpha},$$

for every  $r \geq 1$ ,  $\varepsilon > 0$ , and probability measure  $Q$  with bounded Lebesgue density  $q$  on  $\mathbb{R}^{d+1}$ .

**Proof.** Let  $f_1, \dots, f_p$  be the centers of  $\|\cdot\|_\infty$ -balls of radius  $\varepsilon$  that cover  $C_1^\alpha[0, 1]^d$ . For each  $i$ , define the sets  $C_i$  and  $D_i$  as the subgraphs of  $f_i - \varepsilon$  and  $f_i + \varepsilon$ , respectively. Then the pairs  $[C_i, D_i]$  form brackets that cover  $\mathcal{C}_{\alpha,d}$ . (More precisely, their indicator functions bracket the set of indicator functions of the sets in  $\mathcal{C}_{\alpha,d}$ .) Their  $L_1(Q)$ -size equals

$$Q(C_i \Delta D_i) = \int_{[0,1]^d} \int_{\mathbb{R}} 1\{f_i(x) - \varepsilon \leq t < f_i(x) + \varepsilon\} dQ(t, x) \leq 2\varepsilon \|q\|_\infty.$$

The  $L_r(Q)$ -size of the brackets is the  $(1/r)$ th power of this. It follows that the bracketing number  $N_{[]}((2\varepsilon \|q\|_\infty)^{1/r}, \mathcal{C}_{\alpha,d}, L_r(Q))$  is bounded by  $p$ . Finally, apply the previous theorem and make a change of variable. ■

The previous results are restricted to bounded subsets of Euclidean space. Under appropriate conditions on the tails of the underlying distributions they can be extended to classes of functions on the whole of Euclidean space. The general conclusion is that under a weak tail condition, the same amount of smoothness suffices for the class to be Donsker. For instance, for any distribution on the line with a finite  $2 + \delta$  moment, the class of all uniformly bounded and uniformly Lipschitz functions of order  $\alpha > 1/2$  has a finite bracketing integral and hence is Donsker. A more general result is an easy corollary to the first theorem of this section also.

**2.7.4 Corollary.** *Let  $\mathbb{R}^d = \cup_{j=1}^{\infty} I_j$  be a partition of  $\mathbb{R}^d$  into bounded, convex sets with nonempty interior, and let  $\mathcal{F}$  be a class of functions  $f: \mathbb{R}^d \mapsto \mathbb{R}$  such that the restrictions  $\mathcal{F}|_{I_j}$  belong to  $C_{M_j}^{\alpha}(I_j)$  for every  $j$ . Then there exists a constant  $K$  depending only on  $\alpha, V, r$ , and  $d$  such that*

$$\log N_{[]}(\varepsilon, \mathcal{F}, L_r(Q)) \leq K \left( \frac{1}{\varepsilon} \right)^V \left( \sum_{j=1}^{\infty} \lambda(I_j^1)^{\frac{r}{V+r}} M_j^{\frac{Vr}{V+r}} Q(I_j)^{\frac{V}{V+r}} \right)^{\frac{V+r}{r}},$$

for every  $\varepsilon > 0$ ,  $V \geq d/\alpha$ , and probability measure  $Q$ .

**Proof.** Let  $a_j$  be any sequence of numbers in  $(0, \infty]$ . For each  $j \in \mathbb{N}$ , take an  $\varepsilon a_j$ -net  $f_{j,1}, \dots, f_{j,p_j}$  for  $C_{M_j}^{\alpha}(I_j)$  for the uniform norm on  $I_j$ . By Theorem 2.7.1,  $p_j$  can be chosen to satisfy

$$\log p_j \leq K \lambda(I_j^1) \left( \frac{M_j}{\varepsilon a_j} \right)^{d/\alpha},$$

for a constant  $K$  depending on  $d$  and  $\alpha$  only. It is clear that, for  $\varepsilon a_j > M_j$ , the value  $p_j$  can be chosen equal to 1. Now form the brackets

$$\left[ \sum_{j=1}^{\infty} (f_{j,i_j} - \varepsilon a_j) 1\{I_j\}, \sum_{j=1}^{\infty} (f_{j,i_j} + \varepsilon a_j) 1\{I_j\} \right],$$

where the sequence  $i_1, i_2, \dots$  ranges over all possible values: for each  $j$  the integer  $i_j$  ranges over  $\{1, 2, \dots, p_j\}$ . The number of brackets is bounded by  $\prod p_j$ , and each bracket has  $L_r(Q)$ -size equal to  $2\varepsilon (\sum a_j^r Q(I_j))^{1/r}$ . It follows that

$$\begin{aligned} \log N_{[]} \left( 2\varepsilon (\sum a_j^r Q(I_j))^{1/r}, \mathcal{F}, L_r(Q) \right) &\lesssim K \left( \frac{1}{\varepsilon} \right)^{d/\alpha} \sum_{\substack{j=1 \\ a_j \varepsilon \leq M_j}}^{\infty} \lambda(I_j^1) \left( \frac{M_j}{a_j} \right)^{d/\alpha} \\ &\leq K \left( \frac{1}{\varepsilon} \right)^V \sum_{j=1}^{\infty} \lambda(I_j^1) \left( \frac{M_j}{a_j} \right)^V, \end{aligned}$$

for every  $V \geq d/\alpha$ . The choice  $a_j^{V+r} = \lambda(I_j^1) M_j^V / Q(I_j)$  reduces both series in this expression to essentially the series in the statement of the corollary. Simplify to obtain the result. ■

The preceding upper bound may be applied to obtain a simple sufficient condition for classes of smooth functions on the whole space to have a finite bracketing integral. As an example, consider the class  $C_1^\alpha(\mathbb{R})$ . If  $\alpha > 1/2$  and  $\sum_{j=1}^{\infty} P(I_j)^s < \infty$  for a partition of  $\mathbb{R}$  into intervals of fixed length and some  $s < 1/2$ , then for sufficiently small  $\delta > 0$ ,

$$\log N_{[]}(\varepsilon, C_1^\alpha(\mathbb{R}), L_2(P)) \leq K \left( \frac{1}{\varepsilon} \right)^{2-\delta}$$

(for a constant  $K$  depending on  $P$ ,  $\alpha$ , and  $\delta$ ). Consequently, the class  $C_1^\alpha(\mathbb{R})$  has a finite bracketing integral and is  $P$ -Donsker. Convergence of the series is implied by a tail condition: the series converges for some  $s < 1/2$  if  $\int |x|^{2+\delta} dP(x)$  for some  $\delta > 0$ .

The bound given by the preceding corollary can be proved to be the best in terms of a power of  $(1/\varepsilon)$ , but it is not sharp in terms of lower-order terms. As a consequence, the best conditions for the class  $\mathcal{F}$  as in the corollary to be Glivenko-Cantelli or Donsker cannot be obtained from the corollary. The class is known to be Glivenko-Cantelli or Donsker if and only if  $\sum_{j=1}^{\infty} M_j P(I_j) < \infty$  or  $\sum_{j=1}^{\infty} M_j P^{1/2}(I_j) < \infty$ , respectively. See Section 2.10.4 and the Notes at the end of Part 2 for further details.

## 2.7.2 Monotone Functions

The class of all uniformly bounded, monotone functions on the real line is Donsker. This may be proved in many ways; for instance, by verifying that the class possesses a finite bracketing integral. The following theorem shows that the bracketing entropy is of the order  $1/\varepsilon$  uniformly in the underlying measure.

**2.7.5 Theorem.** *The class  $\mathcal{F}$  of monotone functions  $f: \mathbb{R} \mapsto [0, 1]$  satisfies*

$$\log N_{[]}(\varepsilon, \mathcal{F}, L_r(Q)) \leq K \left( \frac{1}{\varepsilon} \right),$$

*for every probability measure  $Q$ , every  $r \geq 1$ , and a constant  $K$  that depends on  $r$  only.*

**Proof.** The proof of the theorem is long. It can be shown by a much shorter and straightforward proof that the bracketing entropy is bounded above by a constant times  $(1/\varepsilon) \log(1/\varepsilon)$ . For this, construct the brackets as piecewise constant functions on a regular grid (in the uniform case).

It suffices to establish the bound for the class of monotonically increasing functions. It also suffices to prove the bound for  $Q$  equal to the uniform measure  $\lambda$  on  $[0, 1]$ . To see the latter, note first that if  $Q^{-1}(u) = \inf\{x: Q(x) \geq u\}$  denotes the quantile function of  $Q$ , then the class  $\mathcal{F} \circ Q^{-1}$  consists of functions  $f \circ Q^{-1}: [0, 1] \mapsto [0, 1]$  that are monotone. Since  $Q^{-1} \circ Q(x) \leq x$  for every  $x$  and  $u \leq Q \circ Q^{-1}(u)$  for every  $u$ , an

$\varepsilon$ -bracket  $[l, u]$  for  $f \circ Q^{-1}$  for  $\lambda$  yields a function  $l \circ Q$  with the properties  $l \circ Q \leq f \circ Q^{-1} \circ Q \leq f$  and

$$\|f - l \circ Q\|_{Q,r} = \|f \circ Q^{-1} - l \circ Q \circ Q^{-1}\|_{\lambda,r} \leq \|f \circ Q^{-1} - l\|_{\lambda,r} < \varepsilon.$$

Thus  $l \circ Q$  is a “lower bracket.” If  $Q$  is strictly increasing, then  $u \circ Q$  can be used as an upper bracket. More generally, we can repeat the preceding construction with  $\bar{Q}^{-1}(u) = \sup\{x: Q(x) \leq u\}$  and obtain upper brackets  $\bar{u} \circ \bar{Q}$ . A set of full brackets can next be formed by taking every pair  $[l \circ Q, \bar{u} \circ \bar{Q}]$  of functions attached to a same function  $f$ .

Call a function  $g$  a left bracket for  $f$  if  $g \leq f$  and  $\|f - g\|_{\lambda,r} \leq \varepsilon$ . It suffices to construct a set of left brackets of the given cardinality. Right brackets may next be constructed from left brackets for the class of functions  $h(x) = 1 - f(1 - x)$ .

Fix  $\varepsilon$  and set  $c = (1/2)^{1/r}$ . Fix a function  $f$ . Let  $\mathcal{P}_0$  be the partition  $0 = x_0 < x_1 = 1$  of the unit interval. Given a partition  $\mathcal{P}_i$  of the form  $0 = x_0 < x_1 < \dots < x_n = 1$ , define  $\varepsilon_i = \varepsilon_i(f)$  by

$$\varepsilon_i = \max_j (f(x_j) - f(x_{j-1}))(x_j - x_{j-1})^{1/r}.$$

Form a partition  $\mathcal{P}_{i+1}$  by dividing the intervals  $[x_{j-1}, x_j]$  in  $\mathcal{P}_i$  for which

$$(f(x_j) - f(x_{j-1}))(x_j - x_{j-1})^{1/r} \geq c\varepsilon_i$$

into two halves of equal length. It is clear that  $\varepsilon_0 \leq 1$  and that  $\varepsilon_{i+1} \leq c\varepsilon_i \leq 2\varepsilon_{i+1}$ . Let  $n_i = n_i(f)$  be the number of intervals in  $\mathcal{P}_i$  and  $s_i = n_{i+1} - n_i$  the number of members of  $\mathcal{P}_i$  that are divided to obtain  $\mathcal{P}_{i+1}$ . Then by the definitions of  $s_i$  and  $\varepsilon_i$ ,

$$\begin{aligned} s_i(c\varepsilon_i)^{r/(r+1)} &\leq \sum_j (f(x_j) - f(x_{j-1}))^{r/(r+1)}(x_j - x_{j-1})^{1/(r+1)} \\ &\leq \left(\sum_j (f(x_j) - f(x_{j-1}))\right)^{r/(r+1)} \left(\sum_j (x_j - x_{j-1})\right)^{1/(r+1)}, \end{aligned}$$

by Hölder’s inequality. This is bounded by  $(f(1) - f(0))^{r/(r+1)} 1 \leq 1$ . Consequently, the sum of the numbers of intervals up to the  $i$ th partition satisfies

$$\begin{aligned} (2.7.6) \quad \sum_{j=1}^i n_j &= i + \sum_{j=1}^i j s_{i-j} \leq 2 \sum_{j=1}^i j (c\varepsilon_{i-j})^{-r/(r+1)} \\ &\lesssim \sum_{j=1}^i j c^{rj/(r+1)} \varepsilon_i^{-r/(r+1)} \lesssim \varepsilon_i^{-r/(r+1)}, \end{aligned}$$

where  $\lesssim$  denotes smaller than up to a constant that depends on  $r$  only.

Each function  $f$  generates a sequence of partitions  $\mathcal{P}_0 \subset \mathcal{P}_1 \subset \dots$ . Call two functions equivalent at stage  $i$  if their partitions up to the  $i$ -th

are the same. For each  $i$  this yields a partitioning of the class  $\mathcal{F}$  in equivalence classes, which can for increasing  $i$  be visualized as a tree structure with the different equivalence classes at stage  $i$  as the nodes at level  $i$ . Continue the partitioning of a certain branch until the first level  $k$  such that  $\varepsilon_k(f)^r \leq \varepsilon^{r+1}$  for every  $f$  in that branch. This defines a partitioning of the class  $\mathcal{F}$  into finitely many subsets, each corresponding to some sequence of partitions  $\mathcal{P}_0 \subset \dots \subset \mathcal{P}_k$  of the unit interval. For each subset and  $i$ , define

$$\tilde{\varepsilon}_i = \sup_f \varepsilon_i(f),$$

where the supremum is taken over all functions  $f$  in the subset. While  $\tilde{\varepsilon}_i = \tilde{\varepsilon}_i(f)$  may be thought of as depending on  $f$ , it really depends only on the subset of  $f$  in the final partitioning, unlike  $\varepsilon_i$ . Since the numbers  $n_1 \leq n_2 \leq \dots \leq n_i$  also depend on the sequence  $\mathcal{P}_0 \subset \dots \subset \mathcal{P}_i$  only, inequality (2.7.6) remains valid if  $\varepsilon_i$  is replaced by  $\tilde{\varepsilon}_i$ .

For a fixed function  $f$ , define a left-bracketing function  $f_i$ , which is constant on each interval  $[x_{j-1}, x_j]$  in the partition  $\mathcal{P}_i$ , recursively as follows. First  $f_0 = 0$ . Next given  $f_{i-1}$ , define  $f_i$  on the interval  $[x_{j-1}, x_j]$  by

$$f_i(x_{j-1}) = f_{i-1}(x_{j-1}) + l_j \frac{\tilde{\varepsilon}_i}{(x_j - x_{j-1})^{1/r}},$$

where  $l_j \geq 0$  is the largest integer such that  $f_i \leq f$ . Thus to construct  $f_i$ , the left bracket  $f_{i-1}$  is raised at  $x_{j-1}$  by as many steps of size  $\tilde{\varepsilon}_i(x_j - x_{j-1})^{-1/r}$  as possible.

We claim that the set of functions  $f_k$ , when  $f$  ranges over  $\mathcal{F}$  and  $k = k(f)$  is the final level of partitioning for  $f$ , constitutes a set of left brackets of size proportional to  $\varepsilon$  of the required cardinality.

First, it is immediate from the construction of  $f_i$  that, for each of the intervals  $[x_{j-1}, x_j]$  in the  $i$ th partition,

$$(f(x_j) - f_i(x_{j-1}))(x_j - x_{j-1})^{1/r} \leq \tilde{\varepsilon}_i.$$

Combining this with the definition of  $\varepsilon_i$  and the monotonicity of  $f$ , we obtain

$$(2.7.7) \quad (f(x) - f_i(x_{j-1}))(x_j - x_{j-1})^{1/r} \leq \varepsilon_i + \tilde{\varepsilon}_i \leq 2\tilde{\varepsilon}_i, \quad x \in [x_{j-1}, x_j].$$

Consequently,

$$\|f - f_i\|_{\lambda, r}^r \leq n_i(2\tilde{\varepsilon}_i)^r \lesssim \tilde{\varepsilon}_i^{r^2/(r+1)},$$

since  $n_i \lesssim \tilde{\varepsilon}_i^{-r/(r+1)}$  by (2.7.6). For  $i = k(f)$ , this is bounded up to a constant by  $\varepsilon^r$ . Thus the brackets have the correct size.

Second, we count the number of functions  $f_k$  obtained when  $f$  ranges over  $\mathcal{F}$ . In view of the definition of  $k = k(f)$ , we have that  $\varepsilon_{k-1}(g)^r > \varepsilon^{r+1}$ , for some function  $g$  which is equivalent to  $f$  at stage  $k-1$ . This implies that  $\varepsilon_{k-1}(g)^{-r/(r+1)} < 1/\varepsilon$ . Combining this with (2.7.6), we find  $\sum_{j=1}^{k-1} n_j \lesssim 1/\varepsilon$

and trivially  $n_k(f) \leq 2n_{k-1} \lesssim 1/\varepsilon$ . (Note that the numbers  $n_j$  for  $j \leq k-1$  are the same for  $f$  and  $g$ .) The number of sequences  $1 = n_0 \leq n_1 \leq \dots \leq n_k \leq C/\varepsilon$  is equal to the number of ways we can choose  $k$  integers from the set  $\{2, 3, \dots, \lfloor C/\varepsilon \rfloor\}$ . It does not exceed  $2^{C/\varepsilon}$ . Given a sequence of this type, the number of ways to obtain  $\mathcal{P}_{i+1}$  from  $\mathcal{P}_i$  is  $\binom{n_i}{s_i}$ , which is bounded by  $2^{n_i}$ . We conclude that the total number of different final partitions  $\mathcal{P}_0 \subset \dots \subset \mathcal{P}_k$  generated when  $f$  ranges over  $\mathcal{F}$  is bounded by

$$2^{C/\varepsilon} 2^{n_0} 2^{n_1} \dots 2^{n_{k-1}} \leq 2^{2C/\varepsilon}.$$

Given a sequence of partitions  $\mathcal{P}_0 \subset \dots \subset \mathcal{P}_k$ , let  $\mathcal{F}_i$  be the set of all left brackets  $f_i$  constructed on the partition  $\mathcal{P}_i$  when  $f$  ranges over the subset of  $\mathcal{F}$  corresponding to this sequence of partitions. Then  $\mathcal{F}_0 = \{0\}$ , and since  $\tilde{\varepsilon}_i$  is fixed for the given sequence of partitions,

$$\#\mathcal{F}_i \leq \prod_{j=1}^{n_i} m_j \#\mathcal{F}_{i-1},$$

where  $m_j$  is the number of nonnegative integers that can occur in the definition of a function  $f_i \in \mathcal{F}_i$  from a function  $f_{i-1} \in \mathcal{F}_{i-1}$ . Let the interval  $[x_{j-1}, x_j]$  occur in the  $i$ th partition. By (2.7.7) we have that  $f(x_{j-1}) - f_{i-1}(x_{j-1}) \leq 2\tilde{\varepsilon}_{i-1}(x_{j,i} - x_{j-1,i})^{-1/r}$ , where  $[x_{j-1,i}, x_{j,i}]$  is the interval in the partition  $\mathcal{P}_{i-1}$  that contains  $[x_{j-1}, x_j]$ . It follows that

$$0 \leq m_j \leq \frac{2\tilde{\varepsilon}_{i-1}(x_{j,i} - x_{j-1,i})^{-1/r}}{\tilde{\varepsilon}_i(x_j - x_{j-1})^{-1/r}} + 2 \lesssim \frac{2}{c} \cdot 1 + 2 =: L.$$

Conclude that for a given sequence of partitions  $\mathcal{P}_0 \subset \dots \subset \mathcal{P}_k$ , the total number of left brackets  $f_k$  is bounded by

$$L^{n_k} L^{n_{k-1}} \dots L^{n_1} \leq L^{2C/\varepsilon}.$$

Multiply this with the upper bound on the number of partitions to conclude that the total number of left brackets does not exceed  $(2L)^{2C/\varepsilon}$ . ■

### 2.7.3 Closed Convex Sets and Convex Functions

For a pair of subsets  $C$  and  $D$  of a metric space, the *Hausdorff distance* is defined as

$$h(C, D) = \sup_{x \in C} d(x, D) \vee \sup_{x \in D} d(x, C).$$

Restricted to the closed subsets, this defines a metric (which may be infinite). The following lemma gives the entropy of the collection of all compact, convex subsets of a fixed, bounded subset of  $\mathbb{R}^d$  with respect to the Hausdorff metric.

**2.7.8 Lemma.** *For the class  $\mathcal{C}$  of all compact, convex subsets of a fixed, bounded subset of  $\mathbb{R}^d$ , with  $d \geq 2$ , one has*

$$K_1 \left( \frac{1}{\varepsilon} \right)^{(d-1)/2} \leq \log N(\varepsilon, \mathcal{C}, h) \leq K_2 \left( \frac{1}{\varepsilon} \right)^{(d-1)/2},$$

for constants  $K_i$  that depend on  $d$  and the bounded set only.

**Proof.** See Bronštein (1976) or Dudley (1984). ■

As a consequence we obtain  $L_r(Q)$ -bracketing numbers for Lebesgue absolutely continuous probability measures  $Q$ . Combination with the bracketing central limit theorem shows that the closed convex sets of the unit ball in two dimensions form a Donsker class for, for instance, the uniform measure. For any dimension, this class is Glivenko-Cantelli.

For an extension to unbounded convex sets, see Section 2.10.4.

**2.7.9 Corollary.** *For the class  $\mathcal{C}$  of all compact, convex subsets of a fixed, bounded subset of  $\mathbb{R}^d$  with  $d \geq 2$ , one has the entropy bound*

$$\log N_{[]}(\varepsilon, \mathcal{C}, L_r(Q)) \leq K \left( \frac{1}{\varepsilon} \right)^{(d-1)r/2},$$

for every Lebesgue absolutely continuous probability measure  $Q$  with bounded density and a constant  $K$  that depends on the bounded set,  $Q$ , and  $d$  only.

**Proof.** For a given set  $C$ , let  ${}_\varepsilon C = \{x: d(x, C^c) > \varepsilon\}$  be the points that are at least a distance  $\varepsilon$  inside  $C$ , and let  $C^\varepsilon$  be the points within distance  $\varepsilon$  of  $C$ . Then there exists a constant  $K$  depending only on  $\mathcal{C}$  such that the Lebesgue measure  $\lambda$  satisfies

$$\lambda(C^\varepsilon - {}_\varepsilon C) \leq K\varepsilon,$$

for every  $0 < \varepsilon < 1$ .<sup>†</sup>

If  $h(C, D) < \varepsilon$ , then it must be that  ${}_\varepsilon C \subset D \subset C^\varepsilon$ . Thus if  $C_1, \dots, C_p$  are the centers of Hausdorff balls of radius  $\varepsilon$  that cover  $\mathcal{C}$ , then the pairs  $[{}_\varepsilon C_i, C_i^\varepsilon]$  form a class of brackets that cover  $\mathcal{C}$ . Their sizes in  $L_r(Q)$  are bounded by  $Q^{1/r}(C^\varepsilon - {}_\varepsilon C)$ , which is bounded by  $\|q\|_\infty^{1/r} (K\varepsilon)^{1/r}$ . Now the corollary follows from the previous lemma. ■

Next consider the set of all convex functions  $f: C \mapsto \mathbb{R}$  defined on a compact convex subset of  $\mathbb{R}^d$ . If this class is restricted to functions that are also uniformly Lipschitz, then the entropy with respect to the uniform metric can be derived from the entropy of the class of their supergraphs (which are convex sets in  $\mathbb{R}^{d+1}$ ) for the Hausdorff metric. The assumption that the functions are Lipschitz is unpleasant, though it may be noted that every function  $f$  that is convex on  $C^\eta$  for some  $\eta > 0$  is automatically Lipschitz on  $C$ , with Lipschitz constant  $2 \sup\{|f(x)|: x \in C^\eta\}/\eta$  (Problem 2.7.4).

---

<sup>†</sup> This is not trivial, although it is intuitively clear.

**2.7.10 Corollary.** Let  $\mathcal{F}$  be the class of all convex functions  $f: C \mapsto [0, 1]$  defined on a compact, convex subset  $C \subset \mathbb{R}^d$  such that  $|f(x) - f(y)| \leq L\|x - y\|$  for every  $x, y$ . Then

$$\log N(\varepsilon, \mathcal{F}, \|\cdot\|_\infty) \leq K(1 + L)^{d/2} \left(\frac{1}{\varepsilon}\right)^{d/2},$$

for a constant  $K$  that depends on the dimension  $d$  and  $C$  only.

**Proof.** Let  $C_f$  be the supergraph  $\{(x, t): f(x) \leq t\}$  of a function  $f$ . For every pair of points  $x$  and  $y$  and functions  $f$  and  $g$ ,

$$|f(x) - g(x)| \leq |f(x) - g(y)| + L\|y - x\|.$$

Fix a point  $x$  such that  $f(x) < g(x)$ . Then the boundary of the supergraph of  $f$  is below the supergraph of  $g$  at  $x$ , and the projection (closest point) of the point  $(x, f(x))$  on the supergraph of  $g$  has the form  $(y, g(y))$  for some point  $y$ . The distance  $d((x, f(x)), (y, g(y)))$  between the two points in  $\mathbb{R}^{d+1}$  is bounded above by the Hausdorff distance between the two supergraphs. On the other hand, the distance between the two points is bounded below by a multiple of

$$\|x - y\| + |f(x) - g(x)| \geq (1 + L)^{-1}|f(x) - g(x)|.$$

Conclude that  $\|f - g\|_\infty \lesssim (1 + L) h(C_f, C_g)$ . Finally, apply Lemma 2.7.8 to the set of supergraphs intersected with the set  $C \times [0, 1]$ . ■

## 2.7.4 Classes That Are Lipschitz in a Parameter

A simple, but useful, application of bracketing is to classes of functions  $x \mapsto f_t(x)$  that are Lipschitz in the index parameter  $t \in T$ . Suppose that

$$|f_s(x) - f_t(x)| \leq d(s, t) F(x),$$

for some metric  $d$  on the index set, function  $F$  on the sample space, and every  $x$ . Then  $(\text{diam } T)F$  is an envelope function for the class  $\{f_t - f_{t_0}: t \in T\}$  for any fixed  $t_0$ . The bracketing numbers of this class are bounded by the covering numbers of  $T$ .

**2.7.11 Theorem.** Let  $\mathcal{F} = \{f_t: t \in T\}$  be a class of functions satisfying the preceding display for every  $s$  and  $t$  and some fixed function  $F$ . Then, for any norm  $\|\cdot\|$ ,

$$N_{[]} (2\varepsilon\|F\|, \mathcal{F}, \|\cdot\|) \leq N(\varepsilon, T, d).$$

**Proof.** Let  $t_1, \dots, t_p$  be an  $\varepsilon$ -net for  $d$  for  $T$ . Then the brackets  $[f_{t_i} - \varepsilon F, f_{t_i} + \varepsilon F]$  cover  $\mathcal{F}$ . They are of size  $2\varepsilon\|F\|$ . ■

## Problems and Complements

- (Lower bound) There exists a constant  $K$  depending only on  $d$  and  $\alpha$  such that  $\log N(\varepsilon, C_1^\alpha [0, 1]^d, \|\cdot\|_\infty) \geq K(1/\varepsilon)^{d/\alpha}$  for every  $\varepsilon > 0$ .  
 [Hint: See Kolmogorov and Tikhomirov (1961).]
- Let  $\|\cdot\|$  be a norm on a vector space of real-valued functions. Then given a subclass  $\mathcal{F}$  of functions, the entropy numbers of the class  $M\mathcal{F} = \{Mf : f \in \mathcal{F}\}$  satisfy  $N(\varepsilon M, \mathcal{F}, \|\cdot\|) = N(\varepsilon/M, \mathcal{F}, \|\cdot\|)$ . The same relationship holds for bracketing numbers.
- Let  $\mathcal{F}$  be a class of measurable functions  $x \mapsto f(x, r)$  indexed by  $0 \leq r \leq 1$  such that  $r \mapsto f(x, r)$  is monotone for each  $x$ . If the envelope function of  $\mathcal{F}$  is square integrable, then the bracketing numbers of  $\mathcal{F}$  are polynomial.
- Let  $\varepsilon > 0$ ,  $C$  a convex subset of a normed space, and  $f: C^\varepsilon \mapsto \mathbb{R}$  a bounded and convex function. Then  $|f(x) - f(y)| \leq 2\varepsilon^{-1} \|f\|_{C^\varepsilon} \|x - y\|$  for every  $x, y$  in  $C$ .  
 [Hint: Take  $x \neq y$  in  $C$  and define  $z = y + \eta(y - x)/\|y - x\|$  for fixed  $\eta < \varepsilon$ . Then  $z \in C^\varepsilon$  and  $y$  is a convex combination of  $x$  and  $z$ . By convexity,  $f(y) \leq (1 - \lambda)f(x) + \lambda f(z)$ , whence  $f(y) - f(x) \leq \lambda(f(z) - f(x))$ . Calculate  $\lambda$ .]
- Let  $\varepsilon > 0$ ,  $C$  a bounded, convex subset of  $\mathbb{R}^k$ , and  $\mathcal{F}$  a class of convex functions  $f: C^\varepsilon \mapsto \mathbb{R}$  such that  $\{f(x) : f \in \mathcal{F}\}$  is bounded above for every  $x \in C^\varepsilon$  and bounded below for at least one  $x \in C^\varepsilon$ . Then  $\mathcal{F}$  is uniformly bounded and uniformly Lipschitz.  
 [Hint: The function  $\sup_{f \in \mathcal{F}} f(x)$  is finite and convex. Therefore it is continuous and hence bounded on compact sets. For a lower bound on  $\mathcal{F}$  suppose that  $\{f(y) : f \in \mathcal{F}\}$  is bounded below. For any  $x \in C^\varepsilon$  define  $z$  as in the preceding problem. By convexity of  $f$  it follows that (for  $\lambda$  depending on  $x$ ),  $f(x) \geq (1 - \lambda)^{-1} f(y) - \lambda(1 - \lambda)^{-1} f(z)$ , which is uniformly bounded below.]
- Let  $\varepsilon > 0$  and let  $D$  be a compact, convex subset of  $\mathbb{R}^k$ . Then there exists a finite set  $D_0 \subset D^\varepsilon$  and a constant  $c$  depending only on  $\varepsilon$  and  $D$ , such that  $\|f\|_D$  is bounded by  $c\|f\|_{D_0}$  for every convex function  $f: D^\varepsilon \mapsto \mathbb{R}$ .  
 [Hint: Since we can cover  $D$  by finitely many cubes that are contained in  $D^\varepsilon$ , it suffices to bound the supremum norm of  $f$  over a given closed cube  $C \subset D^\varepsilon$ . By convexity, the maximum of  $f$  over  $C$  is bounded above by the maximum of  $|f|$  over the corners. Next,  $f$  can be bounded below by a supporting hyperplane at an arbitrary point in the interior of  $C$ . This takes the form  $f(c) + \langle \nabla f(c), x \rangle$ , which is bounded below by  $f(c) - \|\nabla f(c)\| \|x\|$  and  $\nabla f(c)_i$  can be bounded by the maximum of the values  $|f(c + \varepsilon e_i) - f(c)|$ .]
- Let  $\varepsilon > 0$ ,  $C$  a bounded, convex subset of a normed space, and  $\mathcal{F}$  a class of convex functions  $f: C^\varepsilon \mapsto \mathbb{R}$  that is uniformly bounded above and such that  $\{f(x) : f \in \mathcal{F}\}$  is bounded below for at least one  $x \in C^\varepsilon$ . Then  $\mathcal{F}$  is uniformly bounded and uniformly Lipschitz.

## 2.8

# Uniformity in the Underlying Distribution

The previous chapters present empirical laws of large numbers and central limit theorems for observations from a fixed underlying distribution  $P$ . Many of the sufficient conditions given there are actually satisfied by very large classes of underlying measures; typically, the only limitation is finiteness of some appropriate moment of the envelope function. For instance, classes satisfying the uniform entropy condition are, up to measurability, Glivenko-Cantelli or Donsker for all  $P$  with  $P^*F < \infty$  or  $P^*F^2 < \infty$ , respectively. In particular, many bounded classes of functions are universally Donsker: Donsker for every probability measure on the sample space.

In this chapter we note that even stronger results are typically true. For example, not only does the central limit theorem hold for all underlying measures, or very large classes of measures, it is also valid uniformly in the underlying measure, when this ranges over large classes of measures. Essentially, the main empirical limit theorems are valid uniformly in the underlying measure if the conditions hold in a uniform sense, which appears to be the case frequently. This type of uniformity is certainly of interest in statistical applications.

### 2.8.1 Glivenko-Cantelli Theorems

We start with a uniform Glivenko-Cantelli theorem. Convergence almost surely to zero of a sequence of random variables  $X_n$  is equivalent to convergence in probability of  $\sup_{m \geq n} |X_m|$  to zero. The latter characterization can be used to define “almost sure convergence uniformly in the underlying measure.” A class  $\mathcal{F}$  of measurable functions on a measurable space  $(\mathcal{X}, \mathcal{A})$

is said to be *Glivenko-Cantelli uniformly in  $P \in \mathcal{P}$*  for a given class  $\mathcal{P}$  of probability measures on  $(\mathcal{X}, \mathcal{A})$  if

$$\sup_{P \in \mathcal{P}} P_P^* \left( \sup_{m \geq n} \|\mathbb{P}_m - P\|_{\mathcal{F}} > \varepsilon \right) \rightarrow 0,$$

for every  $\varepsilon > 0$  as  $n \rightarrow \infty$ .

The set  $\mathcal{Q}_n$  in the statement of the following theorem is the collection of all possible realizations of empirical measures of  $n$  observations.

**2.8.1 Theorem.** *Let  $\mathcal{F}$  be a  $P$ -measurable class of functions on a measurable space for every probability measure  $P$  in a class  $\mathcal{P}$ . Suppose that, for some measurable envelope function  $F$ ,*

$$\lim_{M \rightarrow \infty} \sup_{P \in \mathcal{P}} PF\{F > M\} = 0,$$

$$\sup_{Q \in \mathcal{Q}_n} \log N(\varepsilon \|F\|_{Q,1}, \mathcal{F}, L_1(Q)) = o(n), \quad \text{for every } \varepsilon > 0,$$

where the supremum is taken over the set  $\mathcal{Q}_n$  of all discrete probability measures with atoms of size integer multiples of  $1/n$ . Then  $\mathcal{F}$  is Glivenko-Cantelli uniformly in  $P \in \mathcal{P}$ .

**Proof.** For fixed  $M$ , let  $\mathcal{F}_M$  be the class of all functions  $f 1\{f \leq M\}$  as  $f$  ranges over  $\mathcal{F}$ . Then

$$\|\mathbb{P}_n - P\|_{\mathcal{F}} \leq \|(\mathbb{P}_n - P)f 1\{F \leq M\}\|_{\mathcal{F}} + \mathbb{P}_n F 1\{F > M\} + PF\{F > M\}.$$

By the uniform strong law for real random variables (Proposition A.5.1), the second term on the right converges uniformly in  $P$  almost surely to  $PF\{F > M\}$ . By the first condition of the theorem, this quantity can be made arbitrarily small uniformly in  $P$  by choosing large  $M$ . Conclude that it suffices to show that the first term on the right, which equals  $\|\mathbb{P}_n - P\|_{\mathcal{F}_M}$ , converges almost surely to zero uniformly in  $P$  for every fixed  $M$ .

The class  $\mathcal{F}_M$  satisfies the entropy condition of the theorem relative to the envelope function  $M$ . This follows from the inequality

$$\sup_{Q \in \mathcal{Q}_n} N(\varepsilon M, \mathcal{F}_M, L_1(Q)) \leq \sup_{n \leq k \leq 2n} \sup_{Q \in \mathcal{Q}_k} N(\varepsilon \|F\|_{Q,1}, \mathcal{F}, L_1(Q)),$$

which may be proved as follows. Suppose that  $k$  out of the  $n$  support points  $x_1, \dots, x_n$  of a given measure  $Q \in \mathcal{Q}_n$  (with multiple appearance of a given point permitted) satisfy  $F(x_i) \leq M$ . If  $Q_k \in \mathcal{Q}_k$  is the discrete measure on these points, then  $Q|f\{F \leq M\}| = (k/n)Q_k|f|$  for any function  $f$ . Since  $Q_k F \leq M$ , it follows that an  $\varepsilon Q_k F$ -net for  $\mathcal{F}$  in  $L_1(Q_k)$  yields an  $\varepsilon(k/n)M$  net for  $\mathcal{F}_M$  in  $L_1(Q)$ . Thus  $N(\varepsilon M, \mathcal{F}_M, L_1(Q)) \leq N(\varepsilon \|F\|_{Q_k,1}, \mathcal{F}, L_1(Q_k))$ . Since  $\mathcal{Q}_k \subset \mathcal{Q}_{2k} \subset \dots$ , the right side is bounded by the right side of the preceding display. This completes the proof that  $\mathcal{F}_M$  satisfies the entropy condition of the theorem relative to the envelope function  $M$ .

Fix  $\eta > 0$  and values  $X_1, \dots, X_n$ , and take a minimal  $\eta M$ -net  $\mathcal{F}_{nX}$  in  $\mathcal{F}_M$  for the  $L_1(\mathbb{P}_n)$  semimetric. It has just been proved that the cardinality  $N(\eta M, \mathcal{F}_M, L_1(\mathbb{P}_n))$  of  $\mathcal{F}_{nX}$  is bounded by a deterministic sequence  $N_n(\eta)$  (uniformly in  $X_1, \dots, X_n$ ) satisfying  $\log N_n(\eta) = o(n)$ . Let  $\mathbb{P}_n^o$  be the symmetrized empirical measure as defined in Chapter 2.3. Then

$$\|\mathbb{P}_n^o\|_{\mathcal{F}_M} \leq \|\mathbb{P}_n^o\|_{\mathcal{F}_{nX}} + \eta M.$$

The measurability of  $\mathcal{F}$  implies the measurability of  $\mathcal{F}_M$ , so that the probability  $P_P(\|\mathbb{P}_n^o\|_{\mathcal{F}_M} > \varepsilon)$  can be written as  $E_{P,X} P_\varepsilon(\|\mathbb{P}_n^o\|_{\mathcal{F}_M} > \varepsilon)$ . By Hoeffding's inequality, Lemma 2.2.7 applied for fixed  $X_1, \dots, X_n$ ,

$$E_{P,X} P_\varepsilon\left(\|\mathbb{P}_n^o f\|_{\mathcal{F}_{nX}} > \varepsilon - \eta M\right) \leq E_{P,X} N_n(\eta) 2e^{-\frac{1}{2}n(\varepsilon - \eta M)^2/M^2}.$$

The integrand on the right side is independent of  $X_1, \dots, X_n$ . Since  $N_n(\eta) = \exp o(n)$ , the integrand is bounded by  $2 \exp(-n\varepsilon^2/4M^2)$  for sufficiently large  $n$  and small  $\eta$ . Conclude that  $\sum_{m \geq n} P(\|\mathbb{P}_m^o\|_{\mathcal{F}_M} > \varepsilon) \rightarrow 0$  for every  $\varepsilon > 0$ . In view of the symmetrization Lemma 2.3.7 for probabilities, the same is true for the empirical measure replacing the symmetrized empirical. This concludes the proof that  $\|\mathbb{P}_n - P\|_{\mathcal{F}_M}$  converges to zero outer almost surely uniformly in  $P$ . ■

Vapnik-Červonenkis classes of sets or functions satisfy

$$\sup_Q \log N(\varepsilon \|F\|_{Q,1}, \mathcal{F}, L_1(Q)) < \infty,$$

for the supremum taken over all probability measures  $Q$ . This trivially implies the entropy condition of the previous theorem. It follows that a suitably measurable VC-class is uniformly Glivenko-Cantelli over any class of underlying measures for which its envelope function is uniformly integrable. In particular, a uniformly bounded VC-class is uniformly Glivenko-Cantelli over all probability measures, provided it satisfies the measurability condition.

### 2.8.2 Donsker Theorems

Next consider uniform in  $P$  central limit theorems. Let  $\mathcal{F}$  be a class of measurable functions  $f: \mathcal{X} \mapsto \mathbb{R}$  that is Donsker for every  $P$  in a set  $\mathcal{P}$  of probability measures on  $(\mathcal{X}, \mathcal{A})$ . Thus the empirical process  $\mathbb{G}_{n,P} = \sqrt{n}(\mathbb{P}_n - P)$  converges weakly in  $\ell^\infty(\mathcal{F})$  to a tight, Borel measurable version of the Brownian bridge  $\mathbb{G}_P$ . According to Chapter 1.12, this is equivalent to

$$\sup_{h \in \text{BL}_1} |\mathbb{E}_P^* h(\mathbb{G}_{n,P}) - \mathbb{E} h(\mathbb{G}_P)| \rightarrow 0.$$

(Here  $\text{BL}_1$  is the set of all functions  $h: \ell^\infty(\mathcal{F}) \mapsto \mathbb{R}$  which are uniformly bounded by 1 and satisfy  $|h(z_1) - h(z_2)| \leq \|z_1 - z_2\|_{\mathcal{F}}$ .) We shall call  $\mathcal{F}$  *Donsker uniformly in  $P \in \mathcal{P}$*  if this convergence is uniform in  $P$ .

The weak convergence  $\mathbb{G}_{n,P} \rightsquigarrow \mathbb{G}_P$  under a fixed  $P$  is equivalent to asymptotic equicontinuity of the sequence  $\mathbb{G}_{n,P}$  and total boundedness of  $\mathcal{F}$  under the seminorm  $\rho_P(f) = \|f - Pf\|_{P,2}$ . It is to be expected that uniform versions of these two conditions are sufficient for the uniform Donsker property. The sequence  $\mathbb{G}_{n,P}$  is called *asymptotically equicontinuous uniformly in  $P \in \mathcal{P}$*  if, for every  $\varepsilon > 0$ ,

$$\lim_{\delta \downarrow 0} \limsup_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} P_P^* \left( \sup_{\rho_P(f,g) < \delta} |\mathbb{G}_{n,P}(f) - \mathbb{G}_{n,P}(g)| > \varepsilon \right) = 0.$$

Furthermore, the class  $\mathcal{F}$  is  $\rho_P$ -totally bounded uniformly in  $P \in \mathcal{P}$  if its covering numbers satisfy  $\sup_{P \in \mathcal{P}} N(\varepsilon, \mathcal{F}, \rho_P) < \infty$ .

These uniform versions of the two conditions actually imply more than the uniform Donsker property. In the situation for a fixed  $P$ , the asymptotic equicontinuity of the sequence  $\mathbb{G}_{n,P}$  has its counterpart in the continuity of the sample paths of the limit process  $\mathbb{G}_P$ . Requiring the asymptotic equicontinuity uniformly in  $P$  yields, apart from uniform weak convergence, also uniformity in the continuity of the limit process. The latter is expressed in the following. The class  $\mathcal{F}$  is *pre-Gaussian uniformly in  $P \in \mathcal{P}$*  if the Brownian bridges satisfy the two conditions

$$\sup_{P \in \mathcal{P}} E\|G_P\|_{\mathcal{F}} < \infty; \quad \lim_{\delta \downarrow 0} \sup_{P \in \mathcal{P}} E \sup_{\rho_P(f,g) < \delta} |\mathbb{G}_P(f) - \mathbb{G}_P(g)| = 0.$$

(Some authors include the requirement of uniform pre-Gaussianity in the concept of a uniform Donsker class.)

Throughout this section it is assumed that the class  $\mathcal{F}$  possesses a measurable envelope function  $F$  with the property

$$\lim_{M \rightarrow \infty} \sup_{P \in \mathcal{P}} PF^2\{F > M\} \rightarrow 0.$$

Such an envelope function is called square integrable uniformly in  $P \in \mathcal{P}$ .

**2.8.2 Theorem.** *Let  $\mathcal{F}$  be a class of measurable functions with envelope function  $F$  that is square integrable uniformly in  $P \in \mathcal{P}$ . Then the following statements are equivalent:*

- (i)  $\mathcal{F}$  is Donsker and pre-Gaussian, both uniformly in  $P \in \mathcal{P}$ ;
- (ii) the sequence  $\mathbb{G}_{n,P}$  is asymptotically  $\rho_P$ -equicontinuous uniformly in  $P \in \mathcal{P}$  and  $\sup_{P \in \mathcal{P}} N(\varepsilon, \mathcal{F}, \rho_P) < \infty$  for every  $\varepsilon > 0$ .

Furthermore, uniform pre-Gaussianity implies uniform total boundedness of  $\mathcal{F}$ .

**Proof.** The last assertion is a consequence of Sudakov's minorization inequality A.2.5 for Gaussian processes, which gives the upper bound  $3E\|G_P\|_{\mathcal{F}}$  for  $\varepsilon\sqrt{\log N(\varepsilon, \mathcal{F}, \rho_P)}$  for every  $\varepsilon > 0$  and  $P$ .

(i)  $\Rightarrow$  (ii). For each fixed  $\delta > 0$  and  $P$ , the truncated modulus function

$$z \mapsto h_P(z) = \sup_{\rho_P(f,g) < \delta} |z(f) - z(g)| \wedge 1$$

is contained in  $2BL_1$ . If  $\mathcal{F}$  is uniformly Donsker, then  $E_P^* h_P(\mathbb{G}_{n,P}) \rightarrow Eh_P(\mathbb{G}_P)$  uniformly in  $P$ . Given uniform pre-Gaussianity, the expressions  $Eh_P(\mathbb{G}_P)$  can be made uniformly small by choice of a sufficiently small  $\delta$ . Thus  $\limsup E_P^* h_P(\mathbb{G}_{n,P})$  is uniformly small for sufficiently small  $\delta$ . This implies that the sequence  $\mathbb{G}_{n,P}$  is uniformly asymptotically  $\rho_P$ -equicontinuous.

(ii)  $\Rightarrow$  (i). Fix an arbitrary finite subset  $\mathcal{G}$  of  $\mathcal{F}$  and  $P \in \mathcal{P}$ . The sequence of random vectors  $\{\mathbb{G}_{n,P}(g) : g \in \mathcal{G}\}$  converges weakly to the Gaussian vector  $\{\mathbb{G}_P(g) : g \in \mathcal{G}\}$ . By the portmanteau theorem,

$$\begin{aligned} P\left(\sup_{\substack{\rho_P(f,g) < \delta \\ f,g \in \mathcal{G}}} |\mathbb{G}_P(f) - \mathbb{G}_P(g)| > \varepsilon\right) \\ \leq \liminf_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} P_P^*\left(\sup_{\substack{\rho_P(f,g) < \delta \\ f,g \in \mathcal{G}}} |\mathbb{G}_{n,P}(f) - \mathbb{G}_{n,P}(g)| > \varepsilon\right), \end{aligned}$$

for every  $\delta > 0$ . On the right side  $\mathcal{G}$  can be replaced by  $\mathcal{F}$ . By assumption the resulting expression can be made arbitrarily small by choice of  $\delta$ . Conclude that, for every  $\varepsilon, \eta > 0$ , there exists  $\delta > 0$  such that the left side of the display is smaller than  $\eta$  for every  $P$ . Let  $\mathcal{G}$  increase to a  $\rho_P$ -countable dense subset of  $\mathcal{F}$ , and use the continuity of the sample paths of  $\mathbb{G}_P$  to see that the same statement is true (with the same  $\varepsilon, \eta$ , and  $\delta$ ) with  $\mathcal{G}$  replaced by  $\mathcal{F}$ .

This gives a probability form of the second requirement of uniform pre-Gaussianity. The expectation form follows with the help of Borell's inequality A.2.1, which allows one to bound the moment  $E\|\mathbb{G}_P\|_{\mathcal{F}_\delta}$  by a universal constant times the median of  $\|\mathbb{G}_P\|_{\mathcal{F}_\delta}$  (Problem A.2.2).

Fix  $\delta > 0$ . By the uniform total boundedness, there exists for each  $P$  a  $\delta$ -net  $\mathcal{G}_P$  for  $\rho_P$  over  $\mathcal{F}$  such that the cardinalities of the  $\mathcal{G}_P$  are uniformly bounded by a constant  $k$ . Then

$$\|\mathbb{G}_P\|_{\mathcal{F}} \leq \|\mathbb{G}_P\|_{\mathcal{G}_P} + \sup_{\rho_P(f,g) < \delta} |\mathbb{G}_P(f) - \mathbb{G}_P(g)|.$$

By the result of the preceding paragraph, the expectation of the second term on the right can be made arbitrarily small uniformly in  $P$  by choice of  $\delta$ ; it is certainly finite. The first term on the right is a maximum over at most  $k$  Gaussian random variables, each having mean zero and variance not exceeding  $\sup_{P \in \mathcal{P}} PF^2 < \infty$ . Thus  $E\|\mathbb{G}_P\|_{\mathcal{G}_P}$  is uniformly bounded. This concludes the proof of uniform pre-Gaussianity.

For each  $P$  let  $\Pi_P : \mathcal{F} \mapsto \mathcal{G}_P$  map each  $f$  into a  $\rho_P$ -closest element of  $\mathcal{G}_P$ . Write  $z \circ \Pi_P$  for the discretized element of  $\ell^\infty(\mathcal{F})$  taking the value

$z(\Pi_P f)$  at each  $f \in \mathcal{F}$ . By the uniform central limit theorem for  $\mathbb{R}^k$ , the random vectors  $\{\mathbb{G}_{n,P}(f): f \in \mathcal{G}_P\}$  in  $\mathbb{R}^k$  (with zero coordinates added if  $|\mathcal{G}_P| < k$ ) converge weakly to the random vectors  $\{\mathbb{G}_P(f): f \in \mathcal{G}_P\}$  uniformly in  $P \in \mathcal{P}$  (Proposition A.5.2). This implies that

$$\sup_{h \in \text{BL}_1} |\mathbb{E}_P^* h(\mathbb{G}_{n,P} \circ \Pi_P) - \mathbb{E} h(\mathbb{G}_P \circ \Pi_P)| \rightarrow 0.$$

Next, since every  $h \in \text{BL}_1$  satisfies the inequality  $|h(z_1) - h(z_2)| \leq 2 \wedge \|z_1 - z_2\|_{\mathcal{F}}$ , we have, for every  $\varepsilon > 0$ ,

$$\sup_{h \in \text{BL}_1} |\mathbb{E}_P^* h(\mathbb{G}_{n,P} \circ \Pi_P) - \mathbb{E}^* h(\mathbb{G}_{n,P})| \leq \varepsilon + 2\mathbb{P}_P^*(\|\mathbb{G}_{n,P} \circ \Pi_P - \mathbb{G}_{n,P}\|_{\mathcal{F}} > \varepsilon).$$

By construction of  $\mathcal{G}_P$ , the random variable  $\|\mathbb{G}_{n,P} \circ \Pi_P - \mathbb{G}_{n,P}\|_{\mathcal{F}}$  is bounded by the modulus of continuity  $\|\mathbb{G}_{n,P}\|_{\mathcal{F}_{\delta}}$  of the process  $\mathbb{G}_{n,P}$ . In view of the uniform asymptotic equicontinuity,  $\limsup \mathbb{P}_P^*(\|\mathbb{G}_{n,P} \circ \Pi_P - \mathbb{G}_{n,P}\|_{\mathcal{F}} > \varepsilon)$  can be made arbitrarily small by choice of  $\delta$ , uniformly in  $P$ , for every fixed  $\varepsilon$ . Thus the limsup of the left side of the last display can be made arbitrarily small by choice of  $\delta$ .

Using the uniform pre-Gaussianity, we can obtain the analogous result for the limit processes:

$$\sup_{h \in \text{BL}_1} |\mathbb{E}^* h(\mathbb{G}_P \circ \Pi_P) - \mathbb{E}^* h(\mathbb{G}_P)| \rightarrow 0, \quad \delta \downarrow 0.$$

Combination of the last three displayed equations shows that  $\mathbb{E}_P^* h(\mathbb{G}_{n,P})$  converges to  $\mathbb{E} h(\mathbb{G}_P)$  uniformly in  $h \in \text{BL}_1$  and  $P$ . ■

In the situation that  $\mathcal{P}$  consists of all probability measures on the underlying measurable space, there is an interesting addendum to the previous theorem. In that case, the uniform pre-Gaussianity implies the uniform Donsker property (under suitable measurability conditions on  $\mathcal{F}$ ). We do not reproduce the proof of this result here. (See the Notes for further comments and references.) Instead we derive the uniform versions of the two main empirical limit theorems: the central limit theorem under the uniform entropy condition and the central limit theorem with bracketing.

The uniform entropy condition (2.5.1) implies the uniform Donsker property.

**2.8.3 Theorem.** *Let  $\mathcal{F}$  be a class of measurable functions with measurable envelope function  $F$  such that  $\mathcal{F}_{\delta,P} = \{f - g: f, g \in \mathcal{F}, \|f - g\|_{P,2} < \delta\}$  and  $\mathcal{F}_{\infty}^2 = \{(f - g)^2: f, g \in \mathcal{F}\}$  are  $P$ -measurable for every  $\delta > 0$  and  $P \in \mathcal{P}$ . Furthermore, suppose that, as  $M \rightarrow \infty$ ,*

$$\begin{aligned} \sup_{P \in \mathcal{P}} P F^2\{F > M\} &\rightarrow 0, \\ \int_0^\infty \sup_Q \sqrt{\log N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))} d\varepsilon &< \infty, \end{aligned}$$

where  $Q$  ranges over all finitely discrete probability measures. Then  $\mathcal{F}$  is Donsker and pre-Gaussian uniformly in  $P \in \mathcal{P}$ .

**Proof.** It may be assumed that the envelope function satisfies  $F \geq 1$ , because replacing a given  $F$  by  $F + 1$  does not affect the uniform integrability and makes the uniform entropy condition weaker. Given an arbitrary sequence  $\delta_n \downarrow 0$ , set

$$\theta_{n,P}^2 = \left\| \frac{1}{n} \sum_{i=1}^n f^2(X_i) \right\|_{\mathcal{F}_{\delta_n,P}}^2.$$

Exactly as in the proof of Theorem 2.5.2, it can be shown that for some universal constant  $K$ ,

$$\begin{aligned} & P_P^* \left( \|\mathbb{G}_{n,P}\|_{\mathcal{F}_{\delta_n,P}} > \varepsilon \right)^2 \\ & \leq K E_P^* \int_0^{\theta_{n,P}/\|F\|_n} \sup_Q \sqrt{\log N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))} d\varepsilon E_P \|F\|_n^2. \end{aligned}$$

Here  $\|F\|_n \geq 1$  and  $E_P \|F\|_n^2 = PF^2$  is uniformly bounded in  $P$ . Conclude that the right side is up to a constant uniformly bounded by

$$\int_0^\eta \sup_Q \sqrt{\log N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))} d\varepsilon + \sup_{P \in \mathcal{P}} P_P^*(\theta_{n,P} > \eta),$$

for every  $\eta > 0$ . To establish uniform asymptotic equicontinuity, it suffices to show that the second term converges to zero for all  $\eta$ . This follows as in the proof of Theorem 2.5.2, except that presently we use the uniform law of large numbers, Theorem 2.8.1.

By assumption,  $\sup_Q N(\varepsilon, \mathcal{F}, \rho_Q)$  is finite for every  $\varepsilon > 0$  if  $Q$  ranges over all finitely discrete measures. That the supremum is still finite if  $Q$  also ranges over  $\mathcal{P}$  can be proved as in the proof of Theorem 2.5.2. ■

The uniform entropy condition requires the supremum over all probability measures of the root entropy to be integrable. In terms of entropy with bracketing, it suffices to consider the supremum over the class of interest.

**2.8.4 Theorem.** Let  $\mathcal{F}$  be a class of measurable functions such that

$$\begin{aligned} & \lim_{M \rightarrow \infty} \sup_{P \in \mathcal{P}} PF^2\{F > M\} = 0, \\ & \int_0^\infty \sup_{P \in \mathcal{P}} \sqrt{\log N_{[]}(\varepsilon \|F\|_{P,2}, \mathcal{F}, L_2(P))} d\varepsilon < \infty. \end{aligned}$$

Then  $\mathcal{F}$  is Donsker and pre-Gaussian uniformly in  $P \in \mathcal{P}$ .

**Proof.** The first condition implies that  $\sup_P \|F\|_{P,2} < \infty$ . Finiteness of the integral implies finiteness of the integrand. Therefore,

$$\sup_P N_{[]}(\varepsilon \|F\|_{P,2}, \mathcal{F}, L_2(P)) < \infty,$$

and  $\mathcal{F}$  is totally bounded uniformly in  $P$ . Uniform asymptotic equicontinuity follows by making the proof of Theorem 2.5.6 uniform in  $P$ . Actually, the proof of this theorem contains the essence of the proof of the maximal inequality given by Theorem 2.14.2, which implies that

$$E_P^* \|\mathbb{G}_{n,P}\|_{\mathcal{F}_{\delta,P}} \lesssim \int_0^\delta \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}, L_2(P))} d\varepsilon + \frac{PF^2\{F > \sqrt{n}b(\delta, P)\}}{b(\delta, P)}.$$

Here

$$b(\delta, P) = \delta / \sqrt{1 + \log N_{[]}(\delta, \mathcal{F}, L_2(P))}.$$

The right side of the first display converges to zero uniformly in  $P$  as  $n \rightarrow \infty$  followed by  $\delta \downarrow 0$ . ■

### 2.8.3 Central Limit Theorem Under Sequences

As an application of the uniform central limit theorem, consider the empirical process based on triangular arrays. For each  $n$ , let  $X_{n1}, \dots, X_{nn}$  be i.i.d. according to a probability measure  $P_n$ , and set

$$\mathbb{P}_n = \sum_{i=1}^n \delta_{X_{ni}}.$$

If the sequence of underlying measures  $P_n$  converges to a measure  $P_0$  in a suitable sense, then we may hope to derive that the sequence  $\mathbb{G}_{n,P_n} = \sqrt{n}(\mathbb{P}_n - P_n)$  converges in distribution to  $\mathbb{G}_{P_0}$  in  $\ell^\infty(\mathcal{F})$ . According to the general results on weak convergence to Gaussian processes, this is equivalent to marginal convergence plus  $\mathcal{F}$  being totally bounded and the sequence  $\mathbb{G}_{n,P_n}$  being asymptotically uniformly equicontinuous, both with respect to  $\rho_{P_0}$ . Suppose that the  $\rho_{P_n}$  semimetrics converge uniformly to  $\rho_{P_0}$  in the sense that

$$(2.8.5) \quad \sup_{f,g \in \mathcal{F}} |\rho_{P_n}(f, g) - \rho_{P_0}(f, g)| \rightarrow 0.$$

Then asymptotic equicontinuity for  $\rho_{P_0}$  follows from asymptotic equicontinuity “for the sequence  $\rho_{P_n}$ ,” defined as

$$\lim_{\delta \downarrow 0} \limsup_{n \rightarrow \infty} P_{P_n}^* \left( \sup_{\rho_{P_n}(f,g) < \delta} |\mathbb{G}_{n,P_n}(f) - \mathbb{G}_{n,P_n}(g)| > \varepsilon \right) = 0.$$

In particular, it suffices that  $\mathcal{F}$  be asymptotically equicontinuous uniformly in the set  $\{P_m\}$ . This is in turn implied by  $\mathcal{F}$  being Donsker and pre-Gaussian uniformly in  $\{P_m\}$ .

For marginal convergence, we can apply the Lindeberg central limit theorem. Pointwise convergence of  $\rho_{P_n}$  to  $\rho_{P_0}$  means exactly that the covariance functions of the  $\mathbb{G}_{n,P_n}$  converge to the desired limit. The Lindeberg condition for the sequence  $\mathbb{G}_{n,P_n} f$  is certainly implied by

$$(2.8.6) \quad \limsup_{n \rightarrow \infty} P_n F^2 \{F \geq \varepsilon \sqrt{n}\} = 0, \quad \text{for every } \varepsilon > 0.$$

One possible conclusion is that  $\mathcal{F}$  being uniformly Donsker and pre-Gaussian in the sequence  $\{P_m\}$  together with (2.8.5) and (2.8.6) is sufficient for the convergence  $\mathbb{G}_{n,P_n} \rightsquigarrow \mathbb{G}_{P_0}$ . It is of interest that under the same conditions the Gaussian limit distributions are continuous in the underlying measure.

**2.8.7 Lemma.** *Let  $\mathcal{F}$  be Donsker and pre-Gaussian uniformly in the sequence  $\{P_m\}$ , and let (2.8.5) and (2.8.6) hold. Then  $\mathbb{G}_{n,P_n} \rightsquigarrow \mathbb{G}_{P_0}$  in  $\ell^\infty(\mathcal{F})$ .*

**2.8.8 Lemma.** *Let  $\mathcal{F}$  be pre-Gaussian uniformly in the sequence  $\{P_m\}$ , and let (2.8.5) hold. Then  $\mathbb{G}_{P_n} \rightsquigarrow \mathbb{G}_{P_0}$  in  $\ell^\infty(\mathcal{F})$ .*

**Proofs.** The first lemma was argued earlier (under slightly weaker conditions). (Inspection of the proof shows that the implication (i)  $\Rightarrow$  (ii) of Theorem 2.8.2 is valid without uniform integrability of the envelope.)

For the second lemma, first note that pointwise convergence of the covariance functions and zero-mean normality of each  $\mathbb{G}_P$  imply marginal convergence. The uniform pre-Gaussianity gives

$$\lim_{\delta \downarrow 0} \limsup_{n \rightarrow \infty} P_{P_n} \left( \sup_{\rho_{P_n}(f,g) < \delta} |\mathbb{G}_{P_n}(f) - \mathbb{G}_{P_n}(g)| > \varepsilon \right) = 0.$$

By the uniform convergence of  $\rho_{P_n}$  to  $\rho_{P_0}$ , this is then also true with  $\rho_{P_n}$  replaced by  $\rho_{P_0}$ . The uniform pre-Gaussianity implies uniform total boundedness of  $\mathcal{F}$  by Theorem 2.8.3, which together with (2.8.5) implies total boundedness for  $\rho_{P_0}$ . Finally, apply Theorems 1.5.4 and 1.5.7. ■

The uniform Donsker and pre-Gaussian property in the first lemma could be established by the uniform entropy condition or by a bracketing condition. For easy reference we formulate two easily applicable Donsker theorems for sequences. They are proved as the uniform Donsker theorems 2.8.3 and 2.8.4. Alternatively, compare Theorems 2.11.1 and 2.11.9 or Example 2.14.4.

**2.8.9 Theorem.** *Let  $\mathcal{F}$  be a class of measurable functions with a measurable envelope function  $F$  such that  $\mathcal{F}_{\delta,P_n} = \{f-g: f, g \in \mathcal{F}, \|f-g\|_{P_n,2} < \delta\}$  and  $\mathcal{F}_\infty^2 = \{(f-g)^2: f, g \in \mathcal{F}\}$  are  $P_n$ -measurable for every  $\delta > 0$  and  $n$ . Furthermore, suppose that  $\mathcal{F}$  satisfies the uniform entropy condition, that  $P_n F^2 = O(1)$  and that (2.8.5) and (2.8.6) hold. Then  $\mathbb{G}_{n,P_n} \rightsquigarrow \mathbb{G}_{P_0}$  in  $\ell^\infty(\mathcal{F})$ .*

**2.8.10 Theorem.** Let  $\mathcal{F}$  be a class of measurable functions, with a measurable envelope function  $F$ , that is totally bounded for  $\rho_{P_0}$ , satisfies (2.8.5) and (2.8.6), and is such that

$$\int_0^{\delta_n} \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}, L_2(P_n))} d\varepsilon \rightarrow 0, \quad \text{for every } \delta_n \downarrow 0.$$

Then  $\mathbb{G}_{n, P_n} \rightsquigarrow \mathbb{G}_{P_0}$  in  $\ell^\infty(\mathcal{F})$ .

## Problems and Complements

1. Suppose  $\rho_{P_n} \rightarrow \rho_{P_0}$  uniformly on  $\mathcal{F} \times \mathcal{F}$ . Then  $\mathcal{F}$  is totally bounded for  $\rho_{P_0}$  if and only if, for every  $\varepsilon > 0$ , there exists an  $N$  such that  $\sup_{n \geq N} N(\varepsilon, \mathcal{F}, \rho_{P_n}) < \infty$ .

## 2.9

# Multiplier Central Limit Theorems

With the notation  $Z_i = \delta_{X_i} - P$ , the empirical central limit theorem can be written

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n Z_i \rightsquigarrow \mathbb{G},$$

in  $\ell^\infty(\mathcal{F})$ , where  $\mathbb{G}$  is a (tight) Brownian bridge. Given i.i.d. real-valued random variables  $\xi_1, \dots, \xi_n$ , which are independent of  $Z_1, \dots, Z_n$ , the *multiplier central limit theorem* asserts that

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i Z_i \rightsquigarrow \mathbb{G}.$$

If the  $\xi_i$  have mean zero, have variance 1, and satisfy a moment condition, then the multiplier central limit theorem is true if and only if  $\mathcal{F}$  is Donsker: in that case, the two displays are equivalent.

A more refined and deeper result is the *conditional multiplier central limit theorem*, according to which

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i Z_i \rightsquigarrow \mathbb{G},$$

given almost every sequence  $Z_1, Z_2, \dots$ . This more interesting assertion turns out to be true under only slightly stronger conditions. Next to this almost sure version, we also discuss a version in probability of the conditional central limit theorem.

The techniques developed in this chapter have applications to statistical problems, especially to the study of various ways of bootstrapping the empirical process (see Chapter 3.6). The results are also used in the proofs of Chapter 2.10.

It is unnecessary to make measurability assumptions on  $\mathcal{F}$ . However, as usual, outer expectations are understood relative to a product space. Throughout the section let  $X_1, X_2, \dots$  be defined as the coordinate projections on the “first”  $\infty$  coordinates in the probability space  $(\mathcal{X}^\infty \times \mathcal{Z}, \mathcal{A}^\infty \times \mathcal{C}, P^\infty \times Q)$ , and let  $\xi_1, \xi_2, \dots$  depend on the last coordinate only.

The (unconditional) multiplier central limit theorem is a corollary of a symmetrization inequality, which complements the symmetrization inequalities for Rademacher variables obtained in Chapter 2.3. For a random variable  $\xi$ , set

$$\|\xi\|_{2,1} = \int_0^\infty \sqrt{P(|\xi| > x)} dx.$$

In spite of the notation, this is not a norm (but there exists a norm that is equivalent to  $\|\cdot\|_{2,1}$ ). Finiteness of  $\|\xi\|_{2,1}$  requires slightly more than a finite second moment, but it is implied by a finite  $2 + \varepsilon$  absolute moment (Problem 2.9.1).

In Chapter 2.3 it is shown that the norms of a given process  $\sum Z_i$  and its symmetrized version  $\sum \varepsilon_i Z_i$  are comparable in magnitude. The following lemma extends this comparison to the norm of a general multiplier process  $\sum \xi_i Z_i$ .

**2.9.1 Lemma (Multiplier inequalities).** *Let  $Z_1, \dots, Z_n$  be i.i.d. stochastic processes with  $E^* \|Z_i\|_{\mathcal{F}} < \infty$  independent of the Rademacher variables  $\varepsilon_1, \dots, \varepsilon_n$ . Then for every i.i.d. sample  $\xi_1, \dots, \xi_n$  of mean-zero random variables independent of  $Z_1, \dots, Z_n$ , and any  $1 \leq n_0 \leq n$ ,*

$$\begin{aligned} \frac{1}{2} \|\xi\|_1 E^* \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i Z_i \right\|_{\mathcal{F}} &\leq E^* \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i Z_i \right\|_{\mathcal{F}} \\ &\leq 2(n_0 - 1) E^* \|Z_1\|_{\mathcal{F}} E \max_{1 \leq i \leq n} \frac{|\xi_i|}{\sqrt{n}} \\ &\quad + 2\sqrt{2} \|\xi\|_{2,1} \max_{n_0 \leq k \leq n} E^* \left\| \frac{1}{\sqrt{k}} \sum_{i=n_0}^k \varepsilon_i Z_i \right\|_{\mathcal{F}}. \end{aligned}$$

For symmetrically distributed variables  $\xi_i$ , the constants  $1/2$ ,  $2$ , and  $2\sqrt{2}$  can all be replaced by  $1$ .

**Proof.** Define  $\varepsilon_1, \dots, \varepsilon_n$  as independent of  $\xi_1, \dots, \xi_n$  (on “their own” factor of a product probability space). If the  $\xi_i$  are symmetrically distributed, then

the variables  $\varepsilon_i|\xi_i|$  possess the same distribution as the  $\xi_i$ . In that case, the inequality on the left follows from

$$\mathbb{E}^* \left\| \sum_{i=1}^n \varepsilon_i \mathbb{E}_\xi |\xi_i| Z_i \right\|_{\mathcal{F}} \leq \mathbb{E}^* \left\| \sum_{i=1}^n \varepsilon_i |\xi_i| Z_i \right\|_{\mathcal{F}}.$$

For the general case, let  $\eta_1, \dots, \eta_n$  be an independent copy of  $\xi_1, \dots, \xi_n$ . Then  $\|\xi_i\|_1 = \mathbb{E}|\xi_i - \mathbb{E}\eta_i| \leq \|\xi_i - \eta_i\|_1$ , so that  $\|\xi_i\|_1$  can be replaced by  $\|\xi_i - \eta_i\|_1$  in the left-hand side. Next, apply the inequality for symmetric variables to the variables  $\xi_i - \eta_i$ , and then use the triangle inequality to see that

$$\mathbb{E}^* \left\| \sum_{i=1}^n (\xi_i - \eta_i) Z_i \right\|_{\mathcal{F}} \leq 2 \mathbb{E}^* \left\| \sum_{i=1}^n \xi_i Z_i \right\|_{\mathcal{F}}.$$

This concludes the proof of the inequality on the left.

Again assume that the  $\xi_i$  are symmetrically distributed. Let  $\tilde{\xi}_1 \geq \dots \geq \tilde{\xi}_n$  be the (reversed) order statistics of  $|\xi_1|, \dots, |\xi_n|$ . By the definition of  $Z_1, \dots, Z_n$  as fixed functions of the coordinates on the product space  $(\mathcal{X}^n, \mathcal{B}^n)$ , it follows that for any fixed  $\xi_1, \dots, \xi_n$ ,

$$\mathbb{E}_\varepsilon \mathbb{E}_Z^* \left\| \sum_{i=1}^n \varepsilon_i |\xi_i| Z_i \right\|_{\mathcal{F}} = \mathbb{E}_\varepsilon \mathbb{E}_Z^* \left\| \sum_{i=1}^n \varepsilon_i \tilde{\xi}_i Z_i \right\|_{\mathcal{F}}.$$

In view of Lemma 1.2.7, the joint outer expectation  $\mathbb{E}^*$  can be replaced by  $\mathbb{E}_\xi \mathbb{E}_Z^*$ . Thus

$$\begin{aligned} \mathbb{E}^* \left\| \sum_{i=1}^n \xi_i Z_i \right\|_{\mathcal{F}} &= \mathbb{E}_{\xi, \varepsilon} \mathbb{E}_Z^* \left\| \sum_{i=1}^n \varepsilon_i |\xi_i| Z_i \right\|_{\mathcal{F}} = \mathbb{E}_{\xi, \varepsilon} \mathbb{E}_Z^* \left\| \sum_{i=1}^n \varepsilon_i \tilde{\xi}_i Z_i \right\|_{\mathcal{F}} \\ &\leq (n_0 - 1) \mathbb{E} \tilde{\xi}_1 \mathbb{E}^* \|Z_1\|_{\mathcal{F}} + \mathbb{E}^* \left\| \sum_{i=n_0}^n \varepsilon_i \tilde{\xi}_i Z_i \right\|_{\mathcal{F}}. \end{aligned}$$

Substitute  $\tilde{\xi}_i = \sum_{k=i}^n (\tilde{\xi}_k - \tilde{\xi}_{k+1})$  in the second term (with  $\tilde{\xi}_{n+1} = 0$ ) and change the order of summation in the resulting double sum to see that it is equal to

$$\begin{aligned} \mathbb{E}^* \left\| \sum_{i=n_0}^n \varepsilon_i \tilde{\xi}_i Z_i \right\|_{\mathcal{F}} &= \mathbb{E}^* \left\| \sum_{k=n_0}^n (\tilde{\xi}_k - \tilde{\xi}_{k+1}) \sum_{i=n_0}^k \varepsilon_i Z_i \right\|_{\mathcal{F}} \\ &\leq \mathbb{E} \sum_{k=n_0}^n \sqrt{k} (\tilde{\xi}_k - \tilde{\xi}_{k+1}) \max_{n_0 \leq k \leq n} \mathbb{E}^* \left\| \frac{1}{\sqrt{k}} \sum_{i=n_0}^k \varepsilon_i Z_i \right\|_{\mathcal{F}}. \end{aligned}$$

For  $\tilde{\xi}_{k+1} < t \leq \tilde{\xi}_k$ , one has  $k = \#\{i: |\xi_i| \geq t\}$ . Thus the first expectation in the product can be written as

$$\mathbb{E} \sum_{k=n_0}^n \int_{\tilde{\xi}_{k+1}}^{\tilde{\xi}_k} \sqrt{k} dt \leq \int_0^\infty \mathbb{E} \sqrt{\#\{i: |\xi_i| \geq t\}} dt \leq \int_0^\infty \sqrt{n \mathbb{P}(|\xi_i| \geq t)} dt,$$

by Jensen's inequality. This concludes the proof of the upper bound for symmetric variables.

For possibly asymmetric variables, first note that

$$\mathbb{E}^* \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i Z_i \right\|_{\mathcal{F}} \leq \mathbb{E}^* \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n (\xi_i - \eta_i) Z_i \right\|_{\mathcal{F}}.$$

Next apply the bound for symmetric variables to the variables  $\xi_i - \eta_i$ , and then use the triangle inequality and the “corrected” triangle inequality  $\|\xi - \eta\|_{2,1} \leq 2\sqrt{2}\|\xi\|_{2,1}$  (Problem 2.9.2). ■

For  $n_0 = 1$ , the preceding lemma gives the inequalities

$$\begin{aligned} \frac{1}{2} \|\xi\|_1 \mathbb{E}^* \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i Z_i \right\|_{\mathcal{F}} &\leq \mathbb{E}^* \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i Z_i \right\|_{\mathcal{F}} \\ &\leq 2\sqrt{2} \|\xi\|_{2,1} \max_{1 \leq k \leq n} \mathbb{E}^* \left\| \frac{1}{\sqrt{k}} \sum_{i=1}^k \varepsilon_i Z_i \right\|_{\mathcal{F}}. \end{aligned}$$

For variables  $\xi_i$  with range  $[-1, 1]$ , the maximum in the right may be replaced by only its  $n$ th term (Problem 2.9.3), but this appears to be untrue in general. This simpler inequality obtained for  $n_0 = 1$  is too weak to yield the following theorem.

**2.9.2 Theorem.** *Let  $\mathcal{F}$  be a class of measurable functions. Let  $\xi_1, \dots, \xi_n$  be i.i.d. random variables with mean zero, variance 1, and  $\|\xi\|_{2,1} < \infty$ , independent of  $X_1, \dots, X_n$ . Then the sequence  $n^{-1/2} \sum_{i=1}^n \xi_i (\delta_{X_i} - P)$  converges to a tight limit process in  $\ell^\infty(\mathcal{F})$  if and only if  $\mathcal{F}$  is Donsker. In that case, the limit process is a  $P$ -Brownian bridge.*

**Proof.** Since both the empirical processes  $n^{-1/2} \sum_{i=1}^n (\delta_{X_i} - P)$  and the multiplier processes  $n^{-1/2} \sum_{i=1}^n \xi_i (\delta_{X_i} - P)$  do not change if indexed by the class of functions  $\{f - Pf: f \in \mathcal{F}\}$  instead of  $\mathcal{F}$ , it may be assumed without loss of generality that  $Pf = 0$  for every  $f$ . Marginal convergence of both sequences of processes is equivalent to  $\mathcal{F} \subset \mathcal{L}_2(P)$ . It suffices to show that the asymptotic equicontinuity conditions for the empirical and the multiplier processes are equivalent.

If  $\mathcal{F}$  is Donsker, then  $P^*(F > x) = o(x^{-2})$  as  $x \rightarrow \infty$  by Lemma 2.3.9. By the same lemma convergence of the multiplier processes to a tight limit implies that  $P^*(|\xi F| > x) = o(x^{-2})$ . In particular,  $P^* F < \infty$  in both cases.

Since the assumption  $\|\xi\|_{2,1} < \infty$  implies the existence of a second moment, we have  $\mathbb{E}^* \max_{1 \leq i \leq n} |\xi_i| / \sqrt{n} \rightarrow 0$ . Combination with the multiplier

inequalities, Lemma 2.9.1, gives

$$\begin{aligned} \frac{1}{2}\|\xi\|_1 \limsup_{n \rightarrow \infty} E^* \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \varepsilon_i Z_i \right\|_{\mathcal{F}_\delta} &\leq \limsup_{n \rightarrow \infty} E^* \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i Z_i \right\|_{\mathcal{F}_\delta} \\ &\leq 2\sqrt{2}\|\xi\|_{2,1} \sup_{n_0 \leq k} E^* \left\| \frac{1}{\sqrt{k}} \sum_{i=1}^k \varepsilon_i Z_i \right\|_{\mathcal{F}_\delta}, \end{aligned}$$

for every  $n_0$  and  $\delta > 0$ . By Lemma 2.3.6, the Rademacher variables can be deleted in this statement at the cost of changing the constants. Conclude that  $E^* \|n^{-1/2} \sum_{i=1}^n Z_i\|_{\mathcal{F}_{\delta_n}} \rightarrow 0$  if and only if  $E^* \|n^{-1/2} \sum_{i=1}^n \xi_i Z_i\|_{\mathcal{F}_{\delta_n}} \rightarrow 0$ . These are the mean versions of the asymptotic equicontinuity conditions. By Lemma 2.3.11, these are equivalent to the probability versions. ■

Under the conditions of the preceding theorem, the sequence

$$\left( \frac{1}{\sqrt{n}} \sum_{i=1}^n (\delta_{X_i} - P), \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i (\delta_{X_i} - P) \right)$$

is jointly asymptotically tight. Since the two coordinates are uncorrelated and the joint marginals converge to multivariate normal distributions, the sequence converges jointly to a vector of two independent Brownian bridges.

**2.9.3 Corollary.** *Under the conditions of the preceding theorem, the sequence of processes  $(n^{-1/2} \sum_{i=1}^n (\delta_{X_i} - P), n^{-1/2} \sum_{i=1}^n \xi_i (\delta_{X_i} - P))$  converges in  $\ell^\infty(\mathcal{F}) \times \ell^\infty(\mathcal{F})$  in distribution to a vector  $(\mathbb{G}, \mathbb{G}')$  of independent (tight) Brownian bridges  $\mathbb{G}$  and  $\mathbb{G}'$ .*

An additional corollary applies when the multipliers  $\xi_i$  do not have mean zero. Let  $E\xi_i = \mu$  and  $\text{var } \xi_i = \sigma^2$ . Simple algebra gives

$$\begin{aligned} \frac{1}{\sqrt{n}} \sum_{i=1}^n (\xi_i \delta_{X_i} - \mu P) \\ = \frac{\mu}{\sqrt{n}} \sum_{i=1}^n (\delta_{X_i} - P) + \frac{\sigma}{\sqrt{n}} \sum_{i=1}^n \left( \frac{\xi_i - \mu}{\sigma} \right) (\delta_{X_i} - P) + \sqrt{n}(\bar{\xi}_n - \mu)P. \end{aligned}$$

The first two terms on the right converge weakly if and only if  $\mathcal{F}$  is Donsker; the third is in  $\ell^\infty(\mathcal{F})$  if and only if  $\|P\|_{\mathcal{F}} < \infty$ ; in that case, it converges.

**2.9.4 Corollary.** *Let  $\mathcal{F}$  be Donsker with  $\|P\|_{\mathcal{F}} < \infty$ . Let  $\xi_1, \dots, \xi_n$  be i.i.d. random variables with mean  $\mu$ , variance  $\sigma^2$ , and  $\|\xi_i\|_{2,1} < \infty$ , independent of  $X_1, \dots, X_n$ . Then*

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n (\xi_i \delta_{X_i} - \mu P) \rightsquigarrow \mu \mathbb{G} + \sigma \mathbb{G}' + \sigma Z P,$$

where  $\mathbb{G}$  and  $\mathbb{G}'$  are independent (tight) Brownian bridges and are independent of the random variable  $Z$ , which is standard normally distributed. The limit process  $\mu\mathbb{G} + \sigma\mathbb{G}' + \sigma Z P$  is a Gaussian process with zero mean and covariance function  $(\sigma^2 + \mu^2)Pfg - \mu^2PfPg$ .

Next consider conditional multiplier central limit theorems. For finite  $\mathcal{F}$ , the almost sure version is a simple consequence of the Lindeberg theorem.

**2.9.5 Lemma.** Let  $Z_1, Z_2, \dots$  be i.i.d. Euclidean random vectors with  $EZ_i = 0$  and  $E\|Z_i\|^2 < \infty$  independent of the i.i.d. sequence  $\xi_1, \xi_2, \dots$ , with  $E\xi_i = 0$  and  $E\xi_i^2 = 1$ . Then, conditionally on  $Z_1, Z_2, \dots$ ,

$$\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i Z_i \rightsquigarrow N(0, \text{cov } Z_1),$$

for almost every sequence  $Z_1, Z_2, \dots$

**Proof.** According to the Lindeberg central limit theorem, the statement is true for every sequence  $Z_1, Z_2, \dots$  such that, for every  $\varepsilon > 0$ ,

$$\frac{1}{n} \sum_{i=1}^n Z_i Z'_i \rightarrow \text{cov } Z_1; \quad \frac{1}{n} \sum_{i=1}^n \|Z_i\|^2 E\xi_i^2 \{ |\xi_i| \|Z_i\| > \varepsilon\sqrt{n} \} \rightarrow 0.$$

The first condition is true for almost all sequences by the law of large numbers. Furthermore, a finite second moment,  $E\|Z_i\|^2 < \infty$ , implies that  $\max_{1 \leq i \leq n} \|Z_i\|/\sqrt{n} \rightarrow 0$  for almost all sequences. For the intersection of the two sets of sequences, both conditions are satisfied. ■

The preceding lemma takes care of marginal convergence in the conditional multiplier central limit theorems. For convergence in  $\ell^\infty(\mathcal{F})$ , it suffices to check asymptotic equicontinuity.

Conditional weak convergence in probability must be formulated in terms of a metric on conditional laws. Since “conditional laws” do not exist without proper measurability, we utilize the bounded dual Lipschitz distance based on outer expectations. In Chapter 1.12 it is shown that weak convergence,  $G_n \rightsquigarrow G$ , of a sequence of random elements,  $G_n$  in  $\ell^\infty(\mathcal{F})$ , to a separable limit,  $G$ , is equivalent to

$$\sup_{h \in \text{BL}_1} |E^*h(G_n) - Eh(G)| \rightarrow 0.$$

Here  $\text{BL}_1$  is the set of all functions  $h: \ell^\infty(\mathcal{F}) \mapsto [0, 1]$  such that  $|h(z_1) - h(z_2)| \leq \|z_1 - z_2\|_{\mathcal{F}}$  for every  $z_1, z_2$ .

**2.9.6 Theorem.** Let  $\mathcal{F}$  be a class of measurable functions. Let  $\xi_1, \dots, \xi_n$  be i.i.d. random variables with mean zero, variance 1, and  $\|\xi\|_{2,1} < \infty$ , independent of  $X_1, \dots, X_n$ . Let  $\mathbb{G}'_n = n^{-1/2} \sum_{i=1}^n \xi_i (\delta_{X_i} - P)$ . Then the following assertions are equivalent:

- (i)  $\mathcal{F}$  is Donsker;
- (ii)  $\sup_{h \in \text{BL}_1} |\mathbb{E}_\xi h(\mathbb{G}'_n) - \mathbb{E} h(\mathbb{G})| \rightarrow 0$  in outer probability, and the sequence  $\mathbb{G}'_n$  is asymptotically measurable.

**Proof.** (i)  $\Rightarrow$  (ii). For a Donsker class  $\mathcal{F}$ , the sequence  $\mathbb{G}'_n$  converges in distribution to a Brownian bridge process by the unconditional multiplier theorem, Theorem 2.9.2. Thus, it is asymptotically measurable.

A Donsker class is totally bounded for  $\rho_P$ . For each fixed  $\delta > 0$  and  $f \in \mathcal{F}$ , let  $\Pi_\delta f$  denote a closest element in a given, finite  $\delta$ -net for  $\mathcal{F}$ . By continuity of the limit process  $\mathbb{G}$ , we have  $\mathbb{G} \circ \Pi_\delta \rightarrow \mathbb{G}$  almost surely, as  $\delta \downarrow 0$ . Hence, it is certainly true that

$$\sup_{h \in \text{BL}_1} |\mathbb{E} h(\mathbb{G} \circ \Pi_\delta) - \mathbb{E} h(\mathbb{G})| \rightarrow 0.$$

Second, by the preceding lemma, as  $n \rightarrow \infty$  and for every fixed  $\delta > 0$ ,

$$\sup_{h \in \text{BL}_1} |\mathbb{E}_\xi h(\mathbb{G}'_n \circ \Pi_\delta) - \mathbb{E} h(\mathbb{G} \circ \Pi_\delta)| \rightarrow 0,$$

for almost every sequence  $X_1, X_2, \dots$ . (To see this, define  $A: \mathbb{R}^p \mapsto \ell^\infty(\mathcal{F})$  by  $Ay(f) = y_i$  if  $\Pi_\delta f = f_i$ . Then  $h(\mathbb{G} \circ \Pi_\delta) = g(\mathbb{G}(f_1), \dots, \mathbb{G}(f_p))$ , for the function  $g$  defined by  $g(y) = h(Ay)$ . If  $h$  is bounded Lipschitz on  $\ell^\infty(\mathcal{F})$ , then  $g$  is bounded Lipschitz on  $\mathbb{R}^p$  with a smaller bounded Lipschitz norm.) Since  $\text{BL}_1(\mathbb{R}^p)$  is separable for the topology of uniform convergence on compacta, the supremum in the preceding display can be replaced by a countable supremum. It follows that the variable in the display is measurable, because  $h(\mathbb{G}'_n \circ \Pi_\delta)$  is measurable. Third,

$$\begin{aligned} \sup_{h \in \text{BL}_1} |\mathbb{E}_\xi h(\mathbb{G}'_n \circ \Pi_\delta) - \mathbb{E}_\xi h(\mathbb{G}'_n)| &\leq \sup_{h \in \text{BL}_1} \mathbb{E}_\xi |h(\mathbb{G}'_n \circ \Pi_\delta) - h(\mathbb{G}'_n)| \\ &\leq \mathbb{E}_\xi \|\mathbb{G}'_n \circ \Pi_\delta - \mathbb{G}'_n\|_{\mathcal{F}}^* \leq \mathbb{E}_\xi \|\mathbb{G}'_n\|_{\mathcal{F}_\delta}^*, \end{aligned}$$

where  $\mathcal{F}_\delta$  is the class  $\{f - g: \rho_P(f - g) < \delta\}$ . Thus, the outer expectation of the left side is bounded above by  $\mathbb{E}^* \|\mathbb{G}'_n\|_{\mathcal{F}_\delta}$ . As in the proof of Theorem 2.9.2, we can use the multiplier inequalities of Lemma 2.9.1 to see that the latter expression converges to zero as  $n \rightarrow \infty$  followed by  $\delta \rightarrow 0$ . Combine this with the previous displays to obtain one-half of the theorem.

(ii)  $\Rightarrow$  (i). Letting  $h(\mathbb{G}'_n)^*$  and  $h(\mathbb{G}'_n)_*$  denote measurable majorants and minorants with respect to  $(\xi_1, \dots, \xi_n, X_1, \dots, X_n)$  jointly, we have, by the triangle inequality and Fubini's theorem,

$$|\mathbb{E}^* h(\mathbb{G}'_n) - \mathbb{E} h(\mathbb{G})| \leq |\mathbb{E}_X \mathbb{E}_\xi h(\mathbb{G}'_n)^* - \mathbb{E}_X^* \mathbb{E}_\xi h(\mathbb{G}'_n)| + \mathbb{E}_X^* |\mathbb{E}_\xi h(\mathbb{G}'_n) - \mathbb{E} h(\mathbb{G})|.$$

If (ii) holds, then, by dominated convergence, the second term on the right side converges to zero for every  $h \in \text{BL}_1$ . The first term on the right is bounded above by  $E_X E_\xi h(\mathbb{G}'_n)^* - E_X E_\xi h(\mathbb{G}'_n)_*$ , which converges to zero if  $\mathbb{G}'_n$  is asymptotically measurable. Thus, (ii) implies that  $\mathbb{G}'_n \rightsquigarrow \mathbb{G}$  unconditionally. Then  $\mathcal{F}$  is Donsker by the converse part of the unconditional multiplier theorem, Theorem 2.9.2. ■

It may be noted that the functions  $(\xi_1, \dots, \xi_n) \mapsto h(\mathbb{G}'_n)$  are (Lipschitz) continuous, hence Borel measurable, for every given sequence  $X_1, X_2, \dots$  (under the assumption that  $\|f(x) - Pf\|_{\mathcal{F}} < \infty$  for every  $x$ , which is implicitly made to ensure that our processes take their values in  $\ell^\infty(\mathcal{F})$ ). Thus, the expectations  $E_\xi h(\mathbb{G}'_n)$  in the statement of the preceding theorem make sense even though they may be nonmeasurable functions of  $(X_1, \dots, X_n)$ . The second assumption in (ii) implies that these functions are asymptotically measurable and appears to be necessary for the “easy” implication, (ii)  $\Rightarrow$  (i).

The almost sure version of the conditional central limit theorem asserts weak convergence of the sequence  $\mathbb{G}'_n = n^{-1/2} \sum_{i=1}^n \xi_i (\delta_{X_i} - P)$  given almost every sequence  $X_1, X_2, \dots$ . It was seen in Example 1.9.4 that almost sure convergence without proper measurability may not mean much. Thus, it is more attractive to define almost sure conditional convergence by strengthening part (ii) of the preceding theorem. It will be shown that

$$\sup_{h \in \text{BL}_1} |E_\xi h(\mathbb{G}'_n) - Eh(\mathbb{G})| \xrightarrow{\text{as*}} 0.$$

This implies that  $E_\xi h(\mathbb{G}'_n) \rightarrow Eh(\mathbb{G})$  for every  $h \in \text{BL}_1$ , for almost every sequence  $X_1, X_2, \dots$ . By the portmanteau theorem, this is then also true for every continuous, bounded  $h$ . Thus, the sequence  $\mathbb{G}'_n$  converges in distribution to  $\mathbb{G}$  given almost every sequence  $X_1, X_2, \dots$  also in a more naive sense of almost sure conditional convergence.

**2.9.7 Theorem.** Let  $\mathcal{F}$  be a class of measurable functions. Let  $\xi_1, \dots, \xi_n$  be i.i.d. random variables with mean zero, variance 1, and  $\|\xi\|_{2,1} < \infty$ , independent of  $X_1, \dots, X_n$ . Define  $\mathbb{G}'_n = n^{-1/2} \sum_{i=1}^n \xi_i (\delta_{X_i} - P)$ . Then the following assertions are equivalent:

- (i)  $\mathcal{F}$  is Donsker and  $P^* \|f - Pf\|_{\mathcal{F}}^2 < \infty$ ;
- (ii)  $\sup_{h \in \text{BL}_1} |E_\xi h(\mathbb{G}'_n) - Eh(\mathbb{G})| \rightarrow 0$  outer almost surely, and the sequence  $E_\xi h(\mathbb{G}'_n)^* - E_\xi h(\mathbb{G}'_n)_*$  converges almost surely to zero for every  $h \in \text{BL}_1$ . Here  $h(\mathbb{G}'_n)^*$  and  $h(\mathbb{G}'_n)_*$  denote measurable majorants and minorants with respect to  $(\xi_1, \dots, \xi_n, X_1, \dots, X_n)$  jointly.

**Proof.** (i)  $\Rightarrow$  (ii). For the first assertion of (ii), the proof of the probability conditional central limit theorem applies, except that it must be argued that  $E_\xi \|\mathbb{G}'_n\|_{\mathcal{F}_\delta}^*$  converges to zero outer almost surely, as  $n \rightarrow \infty$  followed

by  $\delta \downarrow 0$ . By Corollary 2.9.9,

$$\limsup_{n \rightarrow \infty} E_\xi \|G'_n\|_{\mathcal{F}_\delta}^* \leq 6\sqrt{2} \limsup_{n \rightarrow \infty} E^* \|G'_n\|_{\mathcal{F}_\delta}, \quad \text{a.s.}$$

The right-hand side decreases to zero as  $\delta \downarrow 0$ , because  $G'_n \rightsquigarrow G$  unconditionally, by Theorem 2.9.2.

To see that the sequence  $E_\xi h(G'_n)$  is strongly asymptotically measurable, obtain first by the same proof, but with a star added, that

$$|E_\xi h(G'_n)^* - Eh(G)| \xrightarrow{\text{as*}} 0.$$

The same proof also shows that this is true with a lower star. Thus, the sequence  $E_\xi h(G'_n)^* - E_\xi h(G'_n)_*$  converges to zero almost surely.

(ii)  $\Rightarrow$  (i). The class  $\mathcal{F}$  is Donsker in view of the in probability conditional central limit theorem. We need to prove the condition on its envelope function. As in the proof of the portmanteau theorem, there exists a sequence of Lipschitz functions  $h_m: \mathbb{R} \mapsto \mathbb{R}$  such that  $1 \geq h_m(\|z\|_{\mathcal{F}}) \downarrow 1\{\|z\|_{\mathcal{F}} \geq t\}$  pointwise. By Lemma 1.2.2(vi),

$$E_\xi (\|G'_n\|_{\mathcal{F}}^* > t) = E_\xi 1\{\|G'_n\|_{\mathcal{F}} > t\}^* \leq E_\xi h_m(\|G'_n\|_{\mathcal{F}})^*.$$

Under assumption (ii), the right side converges to  $Eh_m(\|G\|_{\mathcal{F}})$  almost surely as  $n \rightarrow \infty$ , for every fixed  $m$ . Conclude that

$$\limsup_{n \rightarrow \infty} P_\xi (\|G'_n\|_{\mathcal{F}}^* > t) \leq P(\|G\|_{\mathcal{F}} \geq t), \quad \text{a.s.}$$

Set  $Z_i = \delta_{X_i} - P$ . By the triangle inequality,  $\|\xi_n Z_n\|_{\mathcal{F}}^* \leq \sqrt{n} \|G'_n\|_{\mathcal{F}}^* + \sqrt{n-1} \|G'_{n-1}\|_{\mathcal{F}}^*$ . Thus, for sufficiently large  $t$ ,

$$\limsup_{n \rightarrow \infty} P_\xi (\|\xi_n Z_n\|_{\mathcal{F}}^* > t\sqrt{n}) < 1, \quad \text{a.s.}$$

Since  $\xi_n$  is distributed as  $\xi_1$ , it follows that  $\limsup \|Z_n\|_{\mathcal{F}}^*/\sqrt{n} < \infty$  almost surely. This implies  $E^* \|Z_1\|_{\mathcal{F}} < \infty$  (Problem 2.9.4). ■

The proof of the following lemma is based on isoperimetric inequalities for product measures, which is treated in Appendix A.4. The Problems section gives alternative proofs of similar results using more conventional methods.

**2.9.8 Lemma.** Let  $Z_1, Z_2, \dots$  be i.i.d. stochastic processes such that  $E^* \|Z_1\|_{\mathcal{F}}^2 < \infty$ . Let  $\xi_1, \xi_2, \dots$  be i.i.d. random variables with mean zero, independent of  $Z_1, Z_2, \dots$ . Then there exists a constant  $K$  such that, for every  $t > 0$ ,

$$\sum_n P \left( E_\xi \left\| \sum_{i=1}^{2^n} \xi_i Z_i \right\|_{\mathcal{F}}^* > 6E^* \left\| \sum_{i=1}^{2^n} \xi_i Z_i \right\|_{\mathcal{F}} + t2^{n/2} \right) \leq K \frac{E^* \|Z_1\|_{\mathcal{F}}^2}{t^2}.$$

Here the star on the right denotes a measurable majorant with respect to the variables  $(\xi_1, \dots, \xi_n, Z_1, \dots, Z_n)$  jointly.

**2.9.9 Corollary.** *In the situation of the preceding lemma,*

$$\limsup_{n \rightarrow \infty} E_\xi \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i Z_i \right\|_{\mathcal{F}}^* \leq 6\sqrt{2} \limsup_{n \rightarrow \infty} E^* \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i Z_i \right\|_{\mathcal{F}}, \quad \text{a.s.},$$

$$E \sup_n E_\xi \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i Z_i \right\|_{\mathcal{F}}^* \leq C \sup_n E^* \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i Z_i \right\|_{\mathcal{F}} + C \sqrt{E^* \|Z_1\|_{\mathcal{F}}^2},$$

for a universal constant  $C$ .

**Proofs.** Abbreviate  $S_n = E_\xi \left\| \sum_{i=1}^n \xi_i Z_i \right\|_{\mathcal{F}}^*$ . Fix  $n$ , and let  $Z_{[1]}, \dots, Z_{[2^n]}$  be the processes  $Z_1, \dots, Z_{2^n}$  ordered by decreasing norms  $\|Z_{[i]}\|_{\mathcal{F}}^*$ . Define the control number  $f(A, A, A, Z_1, \dots, Z_{2^n})$  as in Appendix A.4 for the event  $A = \{S_{2^n} \leq 2ES_{2^n}\}$ . Then on the event  $f(A, A, A, \vec{Z}) < k$ ,

$$S_{2^n} \leq \sum_{i=1}^{k-1} \|Z_{[i]}\|^* + 6ES_{2^n}.$$

(See the explanation after the proof of Proposition A.4.1.) By Markov's inequality, the event  $A$  has probability at least  $1/2$ . Combining this with Proposition A.4.1 yields, for every  $k \geq 3$ ,

$$\begin{aligned} P(S_{2^n} > 6ES_{2^n} + t2^{n/2}) &\leq \left(\frac{2}{3}\right)^k + P\left(\sum_{i=1}^{k-1} \|Z_{[i]}\|^* > t2^{n/2}\right) \\ &\leq \left(\frac{2}{3}\right)^k + P\left(\|Z_{[1]}\|^* > \frac{t}{2} 2^{n/2}\right) + P\left(k\|Z_{[2]}\|^* > \frac{t}{2} 2^{n/2}\right) \\ &\lesssim \left(\frac{1}{k}\right)^8 + \gamma_n\left(\frac{t}{2}\right) + \gamma_n^2\left(\frac{t}{2k}\right) \wedge 1. \end{aligned}$$

Here  $\gamma_n(t) = 2^n P(\|Z_1\|_{\mathcal{F}}^* > t2^{n/2})$ , and the square arises from the binomial inequality of Problem 2.9.6. Let  $\beta_n(t) = \sum_{l=0}^{\infty} \gamma_{n-l}(t)2^{-l}$  be the convolution of the sequences  $\gamma_n(t)$  and  $(1/2)^n$ . Then  $\beta_n(t) \geq \gamma_{n-l}(t)(1/2)^l$  for any natural number  $l$ , whence  $\gamma_n(t2^{-l}) = 2^{2l} \gamma_{n-2l}(t) \leq 2^{4l} \beta_n(t)$ . This readily yields that  $\gamma_n(t/k) \lesssim k^4 \beta_n(t)$  for every number  $k \geq 3$ . Use this inequality with the choice  $k = \lfloor \beta_n(t)^{-1/8} \rfloor \vee 3$  to bound the last line of the preceding display up to a constant by

$$\left(\frac{1}{k}\right)^8 + \beta_n(t) + k^8 \beta_n^2(t) \wedge 1 \lesssim \beta_n(t).$$

The proof of the lemma is complete upon noting that  $\sum \beta_n(t) \leq 2 \sum_{n=-\infty}^{\infty} \gamma_n(t)$ , which in turn is bounded by a multiple of  $E^* \|Z_1\|_{\mathcal{F}}^2 / t^2$ .

Since the random variables  $S_n$  are nondecreasing in  $n$  by Jensen's inequality, we obtain

$$\sup_{n \geq m} \frac{S_n}{\sqrt{n}} \leq \sqrt{2} \sup_{n \geq 2 \log m} \frac{S_{2^n}}{2^{n/2}}.$$

In view of the Borel-Cantelli lemma, Lemma 2.9.8 gives the inequality  $\limsup_n (S_{2^n} - 6\mathbb{E}S_{2^n})/2^{n/2} \leq 0$  almost surely. The first assertion of the corollary follows. Next

$$\begin{aligned}\mathbb{E} \sup_n \frac{S_{2^n} - 6\mathbb{E}S_{2^n}}{2^{n/2}} &\leq \int_0^\infty \sum_n \mathbb{P}(S_{2^n} - 6\mathbb{E}S_{2^n} > t2^{n/2}) dt \\ &\lesssim \int_0^\infty \left( \frac{\mathbb{E}^* \|Z_1\|_{\mathcal{F}}^2}{t^2} \wedge 1 \right) dt.\end{aligned}$$

This is bounded by the square root of  $\mathbb{E}^* \|Z_1\|_{\mathcal{F}}^2$ . The second assertion now follows similarly from the monotonicity of the variables  $S_n$ . ■

## Problems and Complements

1. For any random variable  $\xi$  and  $r > 2$ , one has  $(1/2)\|\xi\|_2 \leq \|\xi\|_{2,1} \leq r/(r-2)\|\xi\|_r$ .
2. For any pair of random variables  $\xi$  and  $\eta$ , one has  $\|\xi + \eta\|_{2,1}^2 \leq 4\|\xi\|_{2,1}^2 + 4\|\eta\|_{2,1}^2$ .
3. In the situation of Lemma 2.9.1, if the variables  $\xi_i$  are symmetric and possess bounded range contained in  $[-M, M]$ , then  $\mathbb{E}^* \left\| \sum_{i=1}^n \xi_i Z_i \right\|_{\mathcal{F}} \leq M \mathbb{E}^* \left\| \sum_{i=1}^n \xi_i Z_i \right\|_{\mathcal{F}}$ .  
**[Hint:** Use the contraction principle, Proposition A.1.10.]
4. If  $X_1, X_2, \dots$  are i.i.d. random variables with  $\limsup |X_n|/\sqrt{n} < \infty$  almost surely, then  $\mathbb{E}X_1^2 < \infty$ .  
**[Hint:** By the Kolmogorov 0-1 law,  $\limsup |X_n|/\sqrt{n}$  is constant. Thus, there exists a constant  $M$  with  $\mathbb{P}(\limsup\{|X_n| > M\sqrt{n}\}) = 1$ . By the Borel-Cantelli lemma,  $\sum \mathbb{P}(|X_n|^2 > nM^2) < \infty$ .]
5. Let  $\mathcal{F}$  be Donsker with  $\|P\|_{\mathcal{F}} < \infty$ . Let  $\xi_1, \dots, \xi_n$  be i.i.d. random variables with mean  $\mu$ , variance  $\sigma^2$ , and  $\|\xi_i\|_{2,1} < \infty$ , independent of  $X_1, \dots, X_n$ . Let  $N_n$  be Poisson-distributed with mean  $n$  and independent of the  $\xi_i$  and  $X_i$ . Then the sequence  $n^{-1/2} \sum_{i=1}^{N_n} \xi_i \delta_{X_i} - \sqrt{n}\mu P$  converges in distribution in  $\ell^\infty(\mathcal{F})$  to  $(\sigma^2 + \mu^2)^{1/2}$  times a Brownian motion process.
6. If  $Z_{[1]}, \dots, Z_{[n]}$  are the reversed order statistics of an i.i.d. sample  $Z_1, \dots, Z_n$ , then  $\mathbb{P}(Z_{[k]} > x) \leq \binom{n}{k} \mathbb{P}(Z_1 > x)^k$  for every  $x$ .  
**[Hint:** Note that  $\mathbb{P}(Z_{[k]} > x) \leq \binom{n}{k} \mathbb{P}(Z_1 > x, \dots, Z_k > x)$ .]

**7. (Alternative proof of Corollary 2.9.9)** Let  $Z_1, Z_2, \dots$  be i.i.d. stochastic processes with sample paths in  $\ell^\infty(\mathcal{F})$  and  $E^* \|Z_i\|_{\mathcal{F}}^2 < \infty$ . Let  $\xi_1, \xi_2, \dots$  be i.i.d. random variables with mean zero, independent of  $Z_1, Z_2, \dots$ . Then there exists a constant  $K$  with

$$\limsup_{n \rightarrow \infty} E_\xi \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i Z_i \right\|_{\mathcal{F}}^* \leq K \limsup_{n \rightarrow \infty} E^* \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i Z_i \right\|_{\mathcal{F}}, \quad \text{a.s.}$$

Here the star on the left side denotes a measurable majorant with respect to the variables  $(\xi_1, \dots, \xi_n, Z_1, \dots, Z_n)$  jointly.

[Hint: Without loss of generality, assume that  $E^* \left\| n^{-1/2} \sum_{i=1}^n \xi_i Z_i \right\|_{\mathcal{F}} \leq 1/2$ , for every  $n$ , and  $E|\xi_1| = 1$ . Since  $E\xi = 0$ , the random variables  $E_\xi \left\| \sum_{i=1}^n \xi_i Z_i \right\|_{\mathcal{F}}^*$  are nondecreasing in  $n$  by Jensen's inequality. Therefore, it suffices to consider the limsup along the subsequence indexed by  $2^n$ . Since  $E^* \|Z_1\|_{\mathcal{F}}^2 < \infty$ ,

$$\sum_{n=1}^{\infty} P \left( \max_{1 \leq i \leq 2^n} \|Z_i\|_{\mathcal{F}}^* > 2^{n/2} \right) \leq \sum_{n=1}^{\infty} 2^n P \left( \|Z_1\|_{\mathcal{F}}^* > 2^{n/2} \right) < \infty.$$

It follows by the Borel-Cantelli lemma that each variable  $\|Z_i\|_{\mathcal{F}}^*$ , for  $1 \leq i \leq 2^n$ , is bounded by  $2^{n/2}$  eventually, almost surely. Consequently, if  $\tilde{Z}_i = Z_i 1\{\|Z_i\|_{\mathcal{F}}^* \leq 2^{n/2}\}$ , then the variables  $E_\xi \left\| \sum_{i=1}^{2^n} \xi_i Z_i \right\|_{\mathcal{F}}^*$  and  $E_\xi \left\| \sum_{i=1}^{2^n} \xi_i \tilde{Z}_i \right\|_{\mathcal{F}}^*$  are equal eventually, almost surely. (We abuse notation, since  $\tilde{Z}_i$  depends on  $n$ .) It suffices to show that the random variable  $\limsup 2^{-n/2} E_\xi \left\| \sum_{i=1}^{2^n} \xi_i \tilde{Z}_i \right\|_{\mathcal{F}}^*$  is bounded by some fixed number almost surely. By the Borel-Cantelli lemma, this is certainly the case if

$$\sum_{n=1}^{\infty} P \left( E_\xi \left\| \sum_{i=1}^{2^n} \xi_i \tilde{Z}_i \right\|_{\mathcal{F}}^* > 5 \cdot 2^{n/2} \right) < \infty.$$

The probabilities in this series can be replaced by their squares at the cost of decreasing 5 to 2. Indeed, the random variables  $S_{k,l} = E_\xi \left\| \sum_{i=k}^l \xi_i \tilde{Z}_i \right\|_{\mathcal{F}}^*$  are monotonely decreasing in  $k$  and increasing in  $l$ . Let  $T$  be the first  $l$  such that  $S_{1,l} > 2 \cdot 2^{n/2}$ . Then  $T \leq 2^n$  if and only if  $S_{1,2^n} > 2 \cdot 2^{n/2}$ . Furthermore, if  $T = k$  and  $S_{1,2^n} > 5 \cdot 2^{n/2}$ , then  $S_{k+1,2^n} > 2 \cdot 2^{n/2}$ , because  $S_{k,k} = \|\tilde{Z}_k\|_{\mathcal{F}}^* \leq 2^{n/2}$ . Thus

$$\begin{aligned} P(S_{1,2^n} > 5 \cdot 2^{n/2}) &\leq \sum_{k=1}^{2^n} P(T = k, S_{1,2^n} > 5 \cdot 2^{n/2}) \\ &\leq \sum_{k=1}^{2^n} P(T = k, S_{k+1,2^n} > 2 \cdot 2^{n/2}) \\ &\leq \sum_{k=1}^{2^n} P(T = k) \max_{1 \leq k \leq 2^n} P(S_{k,2^n} > 2 \cdot 2^{n/2}) \\ &\leq P(S_{1,2^n} > 2 \cdot 2^{n/2})^2. \end{aligned}$$

In view of the contraction principle stated in Proposition A.1.10, we have  $E^* \left\| \sum_{i=1}^{2^n} \xi_i \tilde{Z}_i \right\|_{\mathcal{F}} \leq E^* \left\| \sum_{i=1}^{2^n} \xi_i Z_i \right\|_{\mathcal{F}} \leq 2^{n/2}$ . Conclude that it suffices to

prove that

$$(2.9.10) \quad \sum_{n=1}^{\infty} P\left(E_{\xi}\left\|\sum_{i=1}^{2^n} \xi_i \tilde{Z}_i\right\|_{\mathcal{F}}^* - E^*\left\|\sum_{i=1}^{2^n} \xi_i \tilde{Z}_i\right\|_{\mathcal{F}} > 2^{n/2}\right)^2 < \infty.$$

Let  $E^i$  denote conditional expectation with respect to the  $\sigma$ -field generated by  $Z_1, \dots, Z_i$ . More precisely, let  $E^0$  denote unconditional expectation, and if  $Z_i$  is the projection on the  $i$ th coordinate of the product measurable space  $(\mathcal{X}^\infty \times \mathcal{Z}, \mathcal{A}^\infty \times \mathcal{C})$ , let  $E^i$  denote conditional expectation given  $\mathcal{A}^i \times \mathcal{X}^{\infty-i} \times \mathcal{C}$ . Set

$$D_i = (E^i - E^{i-1})\left(E_{\xi}\left\|\sum_{i=1}^{2^n} \xi_i \tilde{Z}_i\right\|_{\mathcal{F}}^*\right) = (E^i - E^{i-1})(C_i),$$

where

$$C_i = E_{\xi}\left\|\sum_{j=1}^{2^n} \xi_j \tilde{Z}_j\right\|_{\mathcal{F}}^* - E_{\xi}\left\|\sum_{j=1, j \neq i}^{2^n} \xi_j \tilde{Z}_j\right\|_{\mathcal{F}}^*.$$

The random variables inside the probabilities in (2.9.10) can be written as the (telescoping) sums  $\sum_{i=1}^{2^n} D_i$ . By the triangle inequality,  $|D_i| \leq 2C_i \leq 2\|\tilde{Z}_i\|^*$ . Furthermore, in view of Problem 2.9.9, the expectation of  $C_i$  can be bounded by  $EC_i \leq 2^{-n}E^*\left\|\sum_{j=1}^{2^n} \xi_j \tilde{Z}_j\right\|_{\mathcal{F}} \leq 2^{-n/2}$ . The random variables  $D_1, \dots, D_{2^n}$  are uncorrelated and have mean zero. By Chebyshev's inequality, (2.9.10) is bounded by

$$\sum_{n=1}^{\infty} \left(2^{-n} \sum_{i=1}^{2^n} ED_i^2\right)^2 \leq \sum_{n=1}^{\infty} \max_{1 \leq i \leq 2^n} E|D_i|^3 E|D_i| \lesssim \sum_{n=1}^{\infty} 2^{-n/2} E^*\|\tilde{Z}_1\|_{\mathcal{F}}^3.$$

Substitute  $E^*\|\tilde{Z}_1\|_{\mathcal{F}}^3 = \sum_{k=-\infty}^{\infty} E^*\|Z_1\|_{\mathcal{F}}^3 \{2^{(k-1)/2} < \|Z_1\|_{\mathcal{F}}^* \leq 2^{k/2}\}$ ; next bound one factor of  $\|Z_1\|_{\mathcal{F}}^3$  by  $2^{k/2}$ ; and, finally, change the order of summation to bound the previous display by

$$\sum_{k=-\infty}^{\infty} \sum_{n=k}^{\infty} E^*\|Z_1\|_{\mathcal{F}}^2 \{2^{(k-1)/2} < \|Z_1\|_{\mathcal{F}}^* \leq 2^{k/2}\} 2^{k/2-n/2}.$$

This equals  $E^*\|Z_1\|_{\mathcal{F}}^2 \sqrt{2}/(\sqrt{2}-1)$ . The proof is complete.]

8. **(Alternative proof of Corollary 2.9.9)** Let  $Z_1, Z_2, \dots$  be i.i.d. stochastic processes with sample paths in  $\ell^\infty(\mathcal{F})$ . Let  $\xi_1, \xi_2, \dots$  be i.i.d. random variables with mean zero, independent of  $Z_1, Z_2, \dots$ . Suppose  $M = \sup_n E^*\left\|n^{-1/2} \sum_{i=1}^n \xi_i Z_i\right\|_{\mathcal{F}} < \infty$ . Then there exists a universal constant  $K$  with

$$P\left(\sup_{n \geq 1} E_{\xi}\left\|\frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i Z_i\right\|_{\mathcal{F}}^* > \sqrt{2}(2M + 3t)\right) \leq K \left(\frac{2}{t^2} + \frac{M}{t^3}\right) E^*\|Z_1\|_{\mathcal{F}}^2,$$

for every  $t > 0$ . Here the star on the left side denotes a measurable majorant with respect to  $(\xi_1, \dots, \xi_n, Z_1, \dots, Z_n)$  jointly.

[Hint: The supremum over all  $n \geq 1$  inside the probability can be replaced by

$\sqrt{2}$  times the supremum over the numbers  $2^n$ . Set  $\tilde{Z}_i = Z_i 1\{\|Z_i\|_{\mathcal{F}}^* \leq t2^{n/2}\}$ . The given probability is bounded by

$$\begin{aligned} & \sum_n P\left(\max_{1 \leq i \leq 2^n} \|Z_i\|_{\mathcal{F}}^* > t2^{n/2}\right) + \sum_n P\left(E_\xi \left\| \sum_{i=1}^{2^n} \xi_i \tilde{Z}_i \right\|_{\mathcal{F}}^* > 2^{n/2}(2M + 3t)\right) \\ & \leq \sum_n 2^n P\left(\|Z_1\|_{\mathcal{F}}^* > t2^{n/2}\right) + \sum_n P\left(E_\xi \left\| \sum_{i=1}^{2^n} \xi_i \tilde{Z}_i \right\|_{\mathcal{F}}^* > 2^{n/2}(M + t)\right)^2. \end{aligned}$$

Here the square probability is obtained by the same argument as before, but with  $T$  equal to the first  $l$  such that  $S_{1,l} > 2^{n/2}(M + t)$ . The proof can be finished as before.]

9. Given i.i.d. stochastic processes  $Z_1, Z_2, \dots$  with sample paths in  $\ell^\infty(\mathcal{F})$ , define  $S_n = \sum_{i=1}^n Z_i$ . Then  $E^* \|S_n\|_{\mathcal{F}} - E^* \|S_{n-1}\|_{\mathcal{F}} \leq n^{-1} E^* \|S_n\|_{\mathcal{F}}$ .

[Hint: The inequality is equivalent to  $E^* \|S_n/n\|_{\mathcal{F}} \leq E^* \|S_{n-1}/(n-1)\|_{\mathcal{F}}$ . The sequence  $\|S_n/n\|_{\mathcal{F}}^*$  forms a reversed submartingale with respect to the filtration  $\Sigma_n$  generated by the symmetric  $\sigma$ -fields, as in Lemma 2.4.5.]

# 2.10

## Permanence of the Donsker Property

In this chapter we consider a number of operations that preserve the Donsker property and allow the formation of many new Donsker classes from given examples. For instance, unions, convex hulls, and certain closures of Donsker classes are Donsker. The main result of this chapter concerns Lipschitz transformations of Donsker classes and is discussed in Section 2.10.2. Section 2.10.3 covers the preservation of the uniform-entropy condition, and in Section 2.10.4 new Donsker classes are formed through union of sample spaces.

### 2.10.1 Closures and Convex Hulls

Given a class  $\mathcal{F}$  of measurable functions, let  $\bar{\mathcal{F}}$  denote the set of all  $f: \mathcal{X} \mapsto \mathbb{R}$  for which there exists a sequence  $f_m$  in  $\mathcal{F}$  with  $f_m \rightarrow f$  both pointwise and in  $L_2(P)$ . Let  $\text{sconv } \mathcal{F}$  denote the set of convex combinations  $\sum_{i=1}^{\infty} \lambda_i f_i$  of functions  $f_i$  in  $\mathcal{F}$  where  $\sum |\lambda_i| \leq 1$  and the series converges both pointwise and in  $L_2(P)$ .<sup>b</sup>

**2.10.1 Theorem.** *If  $\mathcal{F}$  is Donsker and  $\mathcal{G} \subset \mathcal{F}$ , then  $\mathcal{G}$  is Donsker.*

**2.10.2 Theorem.** *If  $\mathcal{F}$  is Donsker, then  $\bar{\mathcal{F}}$  is Donsker.*

**2.10.3 Theorem.** *If  $\mathcal{F}$  is Donsker, then  $\text{sconv } \mathcal{F}$  is Donsker.*

---

<sup>b</sup> Pointwise convergence means convergence for every argument, not convergence almost surely.

**Proofs.** The first two results are immediate consequences of the characterization of weak convergence in  $\ell^\infty(\mathcal{F})$  as marginal convergence plus asymptotic equicontinuity. In both cases the modulus of continuity does not increase when passing from  $\mathcal{F}$  to the new class.

For the third result, let  $\{\psi_i\}$  be an orthonormal base of a subspace of  $L_2(P)$  that contains  $\mathcal{F}$ , and let  $Z_1, Z_2, \dots$  be i.i.d. standard normally distributed random variables. Since  $\mathcal{F}$  is pre-Gaussian, the series  $\sum_{i=1}^{\infty} (P(f - Pf)\psi_i)Z_i$  is uniformly convergent in  $\ell^\infty(\mathcal{F})$  and represents a tight Brownian bridge process (Problem 2.10.1). Thus the sequence of partial sums  $\mathbb{G}_k = \sum_{i=1}^k (P(f - Pf)\psi_i)Z_i$  satisfies

$$\|\mathbb{G}_k - \mathbb{G}_l\|_{\text{sconv } \mathcal{F}} = \|\mathbb{G}_k - \mathbb{G}_l\|_{\mathcal{F}} \rightarrow 0, \quad \text{a.s.,}$$

as  $k, l \rightarrow \infty$ . It follows that the sequence of partial sums forms a Cauchy sequence in  $\ell^\infty(\text{sconv } \mathcal{F})$  almost surely and hence converges almost surely to a limit  $\mathbb{G}$ . Since each of the partial sums is linear and  $\rho_P$ -uniformly continuous on  $\text{sconv } \mathcal{F}$ , so is the limit  $\mathbb{G}$ , with probability 1. This proves that the symmetric convex hull of  $\mathcal{F}$  is pre-Gaussian.

According to the almost sure representation theorem, Theorem 1.10.4, there exists a probability space  $(\Omega, \mathcal{U}, P)$  and perfect maps  $\phi_n: \Omega \mapsto \mathcal{X}^n$  and  $\phi$  such that

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n (f(\phi_n(\omega)_i) - Pf) - \mathbb{G}(f, \phi(\omega)) \right\|_{\mathcal{F}} \xrightarrow{\text{as*}} 0.$$

The norm does not increase if  $\mathcal{F}$  is replaced by  $\text{sconv } \mathcal{F}$ . Thus there exists a version of the empirical process that converges outer almost surely in  $\ell^\infty(\text{sconv } \mathcal{F})$  to a process with uniformly continuous sample paths. This implies that  $\text{sconv } \mathcal{F}$  is Donsker. (Note that by perfectness of  $\phi_n$ ,  $Eh^*(\{\mathbb{G}_nf: f \in \text{sconv } \mathcal{F}\}) = (P^n)^*h(\{\mathbb{G}_nf: f \in \text{sconv } \mathcal{F}\})$  for every function  $h$  on  $\ell^\infty(\text{sconv } \mathcal{F})$ .) ■

**2.10.4 Example.** The class of all functions  $x \mapsto \mu(-\infty, x]$  with  $\mu$  ranging over all signed measures on  $\mathbb{R}^k$  with total variation bounded by 1 is universally Donsker.

This can be deduced by applying the preceding results several times. The given class is in the convex hull of the set of cumulative distribution functions of probability measures. In view of the classical Glivenko-Cantelli theorem, any cumulative distribution function is the uniform limit of a sequence of finitely discrete cumulative distribution functions. The class of functions  $x \mapsto \sum_i p_i 1\{t_i \leq x\}$ , with  $p_i \geq 0$  and  $\sum p_i = 1$ , is in the convex hull of the class of indicator functions of cells  $[t, \infty)$ . This is Vapnik-Cervonenkis and suitably measurable, hence universally Donsker.

### 2.10.2 Lipschitz Transformations

For given classes  $\mathcal{F}_1, \dots, \mathcal{F}_k$  of real functions defined on some set  $\mathcal{X}$  and a fixed map  $\phi: \mathbb{R}^k \mapsto \mathbb{R}$ , let  $\phi \circ (\mathcal{F}_1, \dots, \mathcal{F}_k)$  denote the class of functions  $x \mapsto \phi(f_1(x), \dots, f_k(x))$  as  $f = (f_1, \dots, f_k)$  ranges over  $\mathcal{F}_1 \times \dots \times \mathcal{F}_k$ . It will be shown that the Donsker property is preserved if  $\phi$  satisfies

$$(2.10.5) \quad |\phi \circ f(x) - \phi \circ g(x)|^2 \leq \sum_{l=1}^k (f_l(x) - g_l(x))^2,$$

for every  $f, g \in \mathcal{F}_1 \times \dots \times \mathcal{F}_k$  and  $x$ . This is certainly the case for Lipschitz functions  $\phi$ .

**2.10.6 Theorem.** Let  $\mathcal{F}_1, \dots, \mathcal{F}_k$  be Donsker classes with  $\|P\|_{\mathcal{F}_i} < \infty$  for each  $i$ . Let  $\phi: \mathbb{R}^k \mapsto \mathbb{R}$  satisfy (2.10.5). Then the class  $\phi \circ (\mathcal{F}_1, \dots, \mathcal{F}_k)$  is Donsker, provided  $\phi \circ (f_1, \dots, f_k)$  is square integrable for at least one  $(f_1, \dots, f_k)$ .

**2.10.7 Example.** If  $\mathcal{F}$  and  $\mathcal{G}$  are Donsker classes and  $\|P\|_{\mathcal{F} \cup \mathcal{G}} < \infty$ , then the pairwise infima  $\mathcal{F} \wedge \mathcal{G}$ , the pairwise suprema  $\mathcal{F} \vee \mathcal{G}$ , and pairwise sums  $\mathcal{F} + \mathcal{G}$  are Donsker classes.

Since  $\mathcal{F} \cup \mathcal{G}$  is contained in the sum of  $\mathcal{F} \cup \{0\}$  and  $\mathcal{G} \cup \{0\}$ , the union of  $\mathcal{F}$  and  $\mathcal{G}$  is Donsker as well.

**2.10.8 Example.** If  $\mathcal{F}$  and  $\mathcal{G}$  are uniformly bounded Donsker classes, then the pairwise products  $\mathcal{F} \cdot \mathcal{G}$  form a Donsker class. The function  $\phi(f, g) = fg$  is Lipschitz on bounded subsets of  $\mathbb{R}^2$ , but not on the whole plane. The condition that  $\mathcal{F}$  and  $\mathcal{G}$  be uniformly bounded cannot be omitted in general. (In fact, the envelope function of the product of two Donsker classes need not be weak  $L_2$ .)

**2.10.9 Example.** If  $\mathcal{F}$  is Donsker with  $\|P\|_{\mathcal{F}} < \infty$  and  $f \geq \delta$  for some constant  $\delta > 0$  for every  $f \in \mathcal{F}$ , then  $1/\mathcal{F} = \{1/f: f \in \mathcal{F}\}$  is Donsker.

**2.10.10 Example.** If  $\mathcal{F}$  is a Donsker class with  $\|P\|_{\mathcal{F}} < \infty$  and  $g$ , a uniformly bounded, measurable function, then  $\mathcal{F} \cdot g$  is Donsker. Although the function  $\phi(f, g) = fg$  is not Lipschitz on the domain of interest, we have

$$|\phi(f_1(x), g(x)) - \phi(f_2(x), g(x))| \leq \|g\|_{\infty} |f_1(x) - f_2(x)|,$$

for all  $x$ . Thus, condition (2.10.5) is satisfied and the theorem applies.

**2.10.11 Example.** If  $\mathcal{F}$  is Donsker with integrable envelope function  $F$ , then the class  $\mathcal{F}_{\leq M} = \{f1\{F \leq M'\}: f \in \mathcal{F}, M' \leq M\}$  is Donsker. Indeed, the class of sets  $\{F \leq M\}$  is linearly indexed by inclusion, so that it is VC. Furthermore, the function  $\phi(f, g) = fg$  is Lipschitz on the set  $\{(f, 0): f \in \mathbb{R}\} \cup \{(f, 1): |f| \leq M\}$ .

Consequently, the class  $\mathcal{F}^M = \{f1\{F > M\}: f \in \mathcal{F}\}$  is Donsker because it is contained in  $\mathcal{F} - \mathcal{F}_{\leq M}$ .

Many transformations of interest are Lipschitz, but not uniformly so. Consider a map  $\phi: \mathbb{R}^k \mapsto \mathbb{R}$  such that

$$(2.10.12) \quad |\phi \circ f(x) - \phi \circ g(x)|^2 \leq \sum_{l=1}^k L_{\alpha,l}^2(x) (f_l(x) - g_l(x))^2,$$

for given measurable functions  $L_{\alpha,1}, \dots, L_{\alpha,k}$  and every  $f, g \in \mathcal{F}_1 \times \dots \times \mathcal{F}_k$  and  $x$ . Then the class  $\phi \circ (\mathcal{F}_1, \dots, \mathcal{F}_k)$  is Donsker if each of the classes  $L_{\alpha,i} \mathcal{F}_i$ , consisting of the functions  $x \mapsto L_{\alpha,i}(x)f(x)$  with  $f$  ranging over  $\mathcal{F}_i$ , is Donsker. By Example 2.10.10, the class  $g\mathcal{F}$  is Donsker whenever  $g$  is bounded and  $\mathcal{F}$  is Donsker, but this is not useful to generalize the preceding theorem. Typically, it is possible to relax the requirement that  $g$  is bounded at the cost of more stringent conditions on  $\mathcal{F}$ . For instance, in the next subsection it is shown that  $g\mathcal{F}$  is Donsker for every combination of a Donsker class  $\mathcal{F}$  that satisfies the uniform entropy condition and for a function  $g$  such that  $Pg^2 F^2 < \infty$ .

**2.10.13 Corollary.** Let  $\phi: \mathbb{R}^k \mapsto \mathbb{R}$  satisfy (2.10.12). Let each of the classes  $L_{\alpha,i} \mathcal{F}_i$  be Donsker with  $\|P\|_{L_{\alpha,i} \mathcal{F}_i} < \infty$ . Then the class  $\phi \circ (\mathcal{F}_1, \dots, \mathcal{F}_k)$  is Donsker, provided  $\phi \circ (f_1, \dots, f_k)$  is square integrable for at least one  $(f_1, \dots, f_k)$ .

**Proof.** Without loss of generality, assume that each of the functions  $L_{\alpha,i}$  is positive. For  $f \in L_{\alpha} \mathcal{F} = L_{\alpha,1} \mathcal{F}_1 \times \dots \times L_{\alpha,k} \mathcal{F}_k$ , define  $\psi(f) = \phi(f/L_{\alpha})$ . Then  $\psi$  is uniformly Lipschitz in the sense of (2.10.5). By the preceding theorem,  $\psi(L_{\alpha} \mathcal{F}) = \phi(\mathcal{F})$  is Donsker. ■

The main tool in the proof of Theorem 2.10.6 is “Gaussianization.” Let  $\xi_1, \xi_2, \dots$  be i.i.d. random variables with a standard normal distribution independent of  $X_1, X_2, \dots$ . By Chapter 2.9, the (conditional as well as unconditional) asymptotic behavior of the empirical process is related to the behavior of the processes

$$\mathbb{Z}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \delta_{X_i}.$$

Given fixed values  $X_1, \dots, X_n$ , the process  $\{\mathbb{Z}_n(f) : f \in L_2(P)\}$  is Gaussian with zero mean and standard deviation metric

$$\sigma_\xi(\mathbb{Z}_n(f) - \mathbb{Z}_n(g)) = \left( \frac{1}{n} \sum_{i=1}^n (f(X_i) - g(X_i))^2 \right)^{1/2}$$

equal to the  $L_2(\mathbb{P}_n)$  semimetric. This observation permits the use of several comparison principles for Gaussian processes, including Slepian's lemma, Proposition A.2.6.

Three lemmas precede the proof of Theorem 2.10.6. The first is a comparison principle of independent interest.

**2.10.14 Lemma.** *Let  $\mathcal{F}$  be a Donsker class with  $\|P\|_{\mathcal{F}} < \infty$ . Then the class  $\mathcal{F}^2 = \{f^2 : f \in \mathcal{F}\}$  is Glivenko-Cantelli in probability:  $\|\mathbb{P}_n - P\|_{\mathcal{F}^2}^* \xrightarrow{P} 0$ . If, in addition,  $P^* F^2 < \infty$  for some envelope function  $F$ , then  $\mathcal{F}^2$  is also Glivenko-Cantelli almost surely and in mean.*

**Proof.** Let  $\xi_1, \xi_2, \dots$  be i.i.d. random variables with a standard normal distribution independent of  $X_1, X_2, \dots$ . Given fixed values  $X_1, \dots, X_n$  the stochastic process  $\mathbb{Z}_n(f) = n^{-1/2} \sum_{i=1}^n \xi_i f(X_i)$  is a Gaussian process with zero mean and variance  $\text{var}_\xi(\mathbb{Z}_n(f)) = \mathbb{P}_n f^2$ . Thus

$$\|\mathbb{P}_n f^2\|_{\mathcal{F}}^{1/2} = \sqrt{\frac{\pi}{2}} \left\| \mathbb{E}_\xi |\mathbb{Z}_n(f)| \right\|_{\mathcal{F}} \leq \sqrt{\frac{\pi}{2}} \mathbb{E}_\xi \|\mathbb{Z}_n\|_{\mathcal{F}}.$$

Since  $\mathcal{F}$  is Donsker and  $\|P\|_{\mathcal{F}} < \infty$ , the sequence  $\mathbb{Z}_n$  converges (unconditionally) to a tight Gaussian process  $\mathbb{Z}$  by the unconditional multiplier theorem, Theorem 2.9.2. By Lemma 2.3.11, this implies that  $\mathbb{E}^* \|\mathbb{Z}_n\|_{\mathcal{F}} = O(1)$ . Conclude that for every  $\varepsilon > 0$  there exist constants  $M$  and  $N$  such that with inner probability at least  $1 - \varepsilon$ :  $\mathbb{P}_n f^2 \leq M$  for every  $f \in \mathcal{F}$  and every  $n \geq N$ . Since  $\mathcal{F}$  is bounded in  $L_2(P)$ , one can also ensure that  $Pf^2 \leq M$  for every  $f \in \mathcal{F}$ .

For fixed  $M$  and any  $\sigma^2, \tau^2 \leq M$ , the bounded Lipschitz distance between the normal distributions  $N(0, \sigma^2)$  and  $N(0, \tau^2)$  is bounded below by  $K |\sigma^2 - \tau^2|$  for a constant  $K$  depending on  $M$  only (Problem 2.10.2). Conclude that with inner probability at least  $1 - \varepsilon$  and  $n \geq N$ ,

$$K |\mathbb{P}_n f^2 - Pf^2| \leq \sup_{h \in \text{BL}_1(\mathbb{R})} |\mathbb{E}_\xi h(\mathbb{Z}_n(f)) - \mathbb{E} h(\mathbb{Z}(f))|.$$

The right side is bounded by the bounded Lipschitz distance between the processes  $\mathbb{Z}_n$  and  $\mathbb{Z}$ . Conclude that with inner probability at least  $1 - \varepsilon$ ,

$$K \|\mathbb{P}_n f^2 - Pf^2\|_{\mathcal{F}} \leq \sup_{h \in \text{BL}_1} |\mathbb{E}_\xi h(\mathbb{Z}_n) - \mathbb{E} h(\mathbb{Z})|.$$

The right side converges to zero in outer probability by the conditional multiplier theorem, Theorem 2.9.6.

Under the condition  $P^*F^2 < \infty$ , the submartingale argument in the proof of Theorem 2.4.3 (based on Lemma 2.4.5) applies and shows that the convergence in probability can be strengthened to convergence in mean and almost surely. ■

**2.10.15 Lemma.** *If  $\mathcal{F}$  is pre-Gaussian, then  $\varepsilon^2 \log N(\varepsilon, \mathcal{F}, \rho_P) \rightarrow 0$  as  $\varepsilon \rightarrow 0$ . If, in addition,  $\|P\|_{\mathcal{F}} < \infty$ , then also  $\varepsilon^2 \log N(\varepsilon, \mathcal{F}, L_2(P)) \rightarrow 0$ .*

**Proof.** For any partition  $\mathcal{F} = \cup_{i=1}^m \mathcal{F}_i$  and  $\varepsilon > 0$ , one has  $N(\varepsilon, \mathcal{F}, \rho_P) \leq \sum_{i=1}^m N(\varepsilon, \mathcal{F}_i, \rho_P)$ . Bound the sum by  $m$  times the maximum of its terms, and take logarithms to obtain

$$\varepsilon \sqrt{\log N(\varepsilon, \mathcal{F}, \rho_P)} \leq \varepsilon \sqrt{\log m} + \varepsilon \max_{i \leq m} \sqrt{\log N(\varepsilon, \mathcal{F}_i, \rho_P)}.$$

For fixed  $\delta > 0$ , choose a partition into  $m = N(\delta, \mathcal{F}, \rho_P)$  sets of diameter at most  $\delta$ . Take an arbitrary element  $f_i$  from each partitioning set. Let  $\mathbb{G}$  be a separable Brownian bridge process. According to Sudakov's inequality A.2.5,  $\varepsilon \sqrt{\log N(\varepsilon, \mathcal{G}, \rho_P)} \leq 3E^* \|\mathbb{G}\|_{\mathcal{G}}$  for any set of functions  $\mathcal{G}$  and  $\varepsilon$ . Since  $N(\varepsilon, \mathcal{F}_i, \rho_P) = N(\varepsilon, \mathcal{F}_i - f_i, \rho_P)$ , the preceding display can be bounded by

$$\frac{3\varepsilon}{\delta} E^* \|\mathbb{G}\|_{\mathcal{F}} + 3 \max_{i \leq m} E^* \|\mathbb{G}\|_{\mathcal{F}_i - f_i} \leq \frac{3\varepsilon}{\delta} E^* \|\mathbb{G}\|_{\mathcal{F}} + 3E^* \|\mathbb{G}\|_{\mathcal{F}_{\delta}},$$

where  $\mathcal{F}_{\delta} = \{f - g : \rho_P(f - g) < \delta, f, g \in \mathcal{F}\}$ . Since  $\mathcal{F}$  is pre-Gaussian,  $\mathbb{G}$  can be chosen bounded and uniformly continuous. This implies that  $E^* \|\mathbb{G}\|_{\mathcal{F}} < \infty$  and  $E^* \|\mathbb{G}\|_{\mathcal{F}_{\delta}} \rightarrow 0$  as  $\delta \rightarrow 0$ . Thus, the first term on the right is finite and converges to zero as  $\varepsilon \downarrow 0$ . The second term can be made arbitrarily small by choosing  $\delta$  sufficiently small.

The second assertion of the lemma is proved in a similar way, now using a Brownian motion process  $\mathbb{G} + \xi P$  (where  $\xi$  is standard normally distributed and is independent of  $\mathbb{G}$ ) instead of  $\mathbb{G}$  (cf. Problem 2.1.4). ■

**2.10.16 Lemma.** *Let  $Z_1, \dots, Z_m$  be separable, mean-zero Gaussian processes indexed by arbitrary sets  $T_i$ . Then*

$$E \max_{1 \leq i \leq m} \|Z_i\|_{T_i} \leq K \left( \max_{1 \leq i \leq m} E \|Z_i\|_{T_i} + \sqrt{\log m} \max_{1 \leq i \leq m} \|\sigma(Z_{i,t})\|_{T_i} \right),$$

for a universal constant  $K$ .

**Proof.** Without loss of generality, assume that  $m \geq 2$ . Set  $\sigma_i = \|\sigma(Z_{i,t})\|_{T_i}$  and  $Y_i = \|Z_{i,t}\|_{T_i}$ . By Borell's inequality A.2.1,

$$(2.10.17) \quad P(|Y_i - M_i| > x) \leq e^{-\frac{1}{2}x^2/\sigma_i^2},$$

where  $M_i$  is a median of the random variable  $Y_i$ . By Lemma 2.2.1, the variable  $Y_i - M_i$  has Orlicz norm  $\|Y_i - M_i\|_{\psi_2}$  bounded by a multiple of  $\sigma_i$ . In view of the triangle inequality,

$$\mathbb{E} \max_i \|Z_i\|_{T_i} \leq \mathbb{E} \max_i |Y_i - M_i| + \max_i M_i \lesssim \sqrt{\log m} \max_i \sigma_i + \max_i M_i,$$

by Lemma 2.2.2. Integration of (2.10.17) (or another application of the inequality  $\|Y_i - M_i\|_{\psi_2} \lesssim \sigma_i$ ) yields  $\mathbb{E}|Y_i - M_i| \lesssim \sigma_i$ . The lemma follows upon using this inequality to bound  $\max_i M_i$  in the last displayed equation. ■

**Proof of Theorem 2.10.6.** The square integrability of  $\phi \circ (f_1, \dots, f_k)$  for one element of  $\mathcal{F}_1 \times \dots \times \mathcal{F}_k$  and the Lipschitz condition imply square integrability of every function of this type. This ensures marginal convergence.

Let  $\xi_1, \xi_2, \dots, \xi_n$  be i.i.d. random variables with a standard normal distribution, independent of  $X_1, \dots, X_n$ . Given fixed values  $X_1, \dots, X_n$ , the process

$$H_n(f) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i \phi \circ f(X_i), \quad f = (f_1, \dots, f_k),$$

is zero-mean Gaussian with “intrinsic” semimetric  $\sigma_\xi(H_n(f) - H_n(g))$  equal to the  $L_2(\mathbb{P}_n)$  semimetric  $\|\phi \circ f - \phi \circ g\|_n$ . Because each  $\mathcal{F}_l$  is Donsker and each  $\|P\|_{\mathcal{F}_l} < \infty$ , the set  $\mathcal{F} = \mathcal{F}_1 \times \dots \times \mathcal{F}_k$  is totally bounded in the product space  $L_2(P)^k$ . For given  $\delta > 0$ , let  $\mathcal{F} = \cup_{j=1}^m \mathcal{H}_j$  be a partition of  $\mathcal{F}$  into sets of diameter less than  $\delta$  with respect to the supremum product metric. The number of sets in the partition can be chosen as  $m \leq \prod_{l=1}^k N(\delta, \mathcal{F}_l, L_2(P))$ . Choose an arbitrary element  $h_j$  from each  $\mathcal{H}_j$ . Given the same  $X_1, \dots, X_n$  as before, the processes  $\{H_n(f) - H_n(h_j) : f \in \mathcal{H}_j\}$  possess the same intrinsic semimetric as  $H_n$ . By the preceding lemma,

$$(2.10.18) \quad \begin{aligned} \mathbb{E}_\xi \sup_j \|H_n - H_n(h_j)\|_{\mathcal{H}_j} &\lesssim \max_{j \leq m} \mathbb{E}_\xi \|H_n - H_n(h_j)\|_{\mathcal{H}_j} \\ &\quad + \sqrt{\log m} \max_{j \leq m} \sup_{f \in \mathcal{H}_j} \|\phi \circ f - \phi \circ h_j\|_n. \end{aligned}$$

For each  $l$ , let  $\xi_{l1}, \dots, \xi_{ln}$  be i.i.d. random variables with a standard normal distribution, chosen independently for different  $l$ . Define  $G_n(f) = \sum_{l=1}^k n^{-1/2} \sum_{i=1}^n \xi_{li} f_l(X_i)$ . In view of (2.10.5),

$$\text{var}_\xi(H_n(f) - H_n(g)) \leq \sum_{l=1}^k \frac{1}{n} \sum_{i=1}^n (f_l - g_l)^2(X_i) = \text{var}_\xi(G_n(f) - G_n(g)).$$

By Slepian’s lemma A.2.6, the expectation  $\mathbb{E}_\xi \|H_n - H_n(h_j)\|_{\mathcal{H}_j}$  is bounded above by two times the same expression, but with  $G_n$  instead of  $H_n$ . Make

this substitution in the right side of (2.10.18), and conclude that the left side of (2.10.18) is bounded up to a constant by

$$\mathbb{E}_\xi \sup_{e_P(f,h) < \delta} |G_n(f) - G_n(h)| + \sqrt{\log m} \sup_{e_P(f,h) < \delta} \|f - h\|_n.$$

(Here  $e_P$  and  $\|f - g\|_n$  denote product metrics corresponding to  $\|\cdot\|_{P,2}$  and  $\|\cdot\|_n$ .) Since each of the classes  $\mathcal{F}_l$  is Donsker and satisfies  $\|P\|_{\mathcal{F}_l} < \infty$ , the sequence  $G_n$  is asymptotically equicontinuous with respect to  $e_P$ . (See Theorem 2.9.2 and Problem 2.1.2.) Hence the outer expectation of the first term converges to zero as  $n \rightarrow \infty$  followed by  $\delta \downarrow 0$ . For the second term, note that the classes  $\mathcal{F}_l - \mathcal{F}_l$  are Donsker by Theorem 2.10.3, so that their squares  $(\mathcal{F}_l - \mathcal{F}_l)^2$  are Glivenko-Cantelli in probability, by Lemma 2.10.14. Thus, as  $n \rightarrow \infty$ ,

$$\begin{aligned} \sqrt{\log m} \sup_{e_P(f,h) < \delta} \|f - h\|_n &\xrightarrow{P^*} \sqrt{\log m} \sup_{e_P(f,h) < \delta} e_P(f, h) \\ &\leq \sum_{l=1}^k \sqrt{\log N(\delta, \mathcal{F}_l, L_2(P))} \delta. \end{aligned}$$

This converges to zero as  $\delta \downarrow 0$ , by Lemma 2.10.15. Combination of the preceding displays gives that the left side of (2.10.18) converges to zero in probability as  $n \rightarrow \infty$  and next as  $\delta \downarrow 0$ .

Suppose that the processes  $H_n$  are separable, so that the variables  $\|H_n - H_n(h_j)\|_{\mathcal{H}_j}$  are measurable. Then it can be concluded that for all positive numbers  $\varepsilon$  and  $\eta$  there exists a finite partition  $\mathcal{F} = \cup_{j=1}^m \mathcal{H}_j$  such that

$$\limsup_{n \rightarrow \infty} P\left(\sup_j \|H_n - H_n(h_j)\|_{\mathcal{H}_j} > \varepsilon\right) < \eta.$$

This implies that the sequence  $H_n$  is asymptotically tight in  $\ell^\infty(\mathcal{F})$ , by Theorem 1.5.6.

Define a map  $T: \phi \circ \mathcal{F} \mapsto \mathcal{F}$  by assigning to each function  $g \in \phi \circ \mathcal{F}$  a unique (but arbitrary) vector  $f = Tg \in \mathcal{F}$  such that  $\phi(f) = g$ . The map  $z \mapsto z \circ T$  from  $\ell^\infty(\mathcal{F})$  to  $\ell^\infty(\phi \circ \mathcal{F})$  is (Lipschitz) continuous. Hence, by the continuous mapping theorem, the sequence of multiplier empirical processes  $\mathbb{G}'_n = H_n \circ T$  converges weakly to a tight limit in  $\ell^\infty(\phi \circ \mathcal{F})$ .

Since  $\|P\|_{\phi \circ \mathcal{F}} < \infty$ , it follows that the sequence of processes  $n^{-1/2} \sum_{i=1}^n \xi_i (\delta_{X_i} - P)$  converges weakly to a tight limit as well. By Theorem 2.9.2, the same is true for the sequence of empirical processes  $\mathbb{G}_n$ .

Finally, we remove the assumption that the processes  $H_n$  are separable. According to Chapter 2.3.3, there exist pointwise-separable versions  $\tilde{\mathcal{F}}_l$  of the classes  $\mathcal{F}_l$ . Almost every value  $\tilde{f}_l(x)$  is the limit of a sequence  $\tilde{g}_{m,l}(x)$  with each  $\tilde{g}_{m,l}$  contained in a countable ‘separant’ of  $\tilde{\mathcal{F}}_l$ . Each element  $\tilde{g}_l$  of the separant is almost surely equal to an element  $g_l$  of the original class  $\mathcal{F}_l$ . It follows that, perhaps possibly on a null set of  $x$ , the function  $\phi$  is defined at each vector  $\tilde{f}(x)$  or can be extended to this vector by continuity.

The extended function satisfies (2.10.5) on  $\tilde{\mathcal{F}} \cup \mathcal{F}$  except perhaps for  $x$  in a fixed null set. Then the class of functions  $\phi \circ \tilde{\mathcal{F}}$  is (essentially) well defined and is a pointwise-separable version of  $\phi \circ \mathcal{F}$ . By the preceding argument,  $\phi \circ \tilde{\mathcal{F}}$  is Donsker. By (2.10.5),

$$\sqrt{n} \|\mathbb{P}_n \phi(f) - \mathbb{P}_n \phi(\tilde{f})\|_{\mathcal{F}} \lesssim \sum_{l=1}^k \sqrt{n} \|\mathbb{P}_n |f_l - \tilde{f}_l|\|_{\mathcal{F}_l}.$$

According to Theorem 2.3.15, this converges to zero in probability. Thus  $\phi \circ \mathcal{F}$  is Donsker also. ■

### 2.10.3 Permanence of the Uniform Entropy Bound

The main theorem of the preceding subsection combines the minimal condition that the classes  $\mathcal{F}_1, \dots, \mathcal{F}_k$  are Donsker with the strong condition that  $\phi$  is uniformly Lipschitz. It may be expected that stronger assumptions on the classes  $\mathcal{F}_i$  combined with weaker conditions on  $\phi$  will yield the same conclusion that the class  $\phi(\mathcal{F}_1, \dots, \mathcal{F}_k)$  is Donsker. Obviously, a concrete formulation of this principle is useful only if it is sufficiently simple, because otherwise it is preferable to deal with the class  $\phi(\mathcal{F}_1, \dots, \mathcal{F}_k)$  directly. Uniform entropy, integral-type conditions allow such a simple statement.

Let  $\mathcal{F}_1, \dots, \mathcal{F}_k$  be classes of measurable functions and  $\phi: \mathbb{R}^k \mapsto \mathbb{R}$  a given map that is Lipschitz of orders  $\alpha_1, \dots, \alpha_k \in (0, 1]$  in the sense that

$$(2.10.19) \quad |\phi \circ f(x) - \phi \circ g(x)|^2 \leq \sum_{i=1}^k L_{\alpha,i}^2(x) |f_i - g_i|^{2\alpha_i}(x),$$

for all  $f = (f_1, \dots, f_k)$  and  $g = (g_1, \dots, g_k)$  in  $\mathcal{F} = \mathcal{F}_1 \times \dots \times \mathcal{F}_k$  and for every  $x$ . In the special case that  $k = 1$ , we can set

$$L_\alpha = \sup_{f,g} \frac{|\phi(f) - \phi(g)|}{|f - g|^\alpha}.$$

This relaxes condition (2.10.5) in two ways: the functions  $L_{\alpha,i}$  need not be bounded, and a lower-order Lipschitz condition suffices. In the following theorem, the first weakening of (2.10.5) is compensated by a moment condition on  $L_\alpha$ ; the second by a strengthening of the entropy condition on the  $\mathcal{F}_i$ .

If  $\mathcal{F}_i$  has envelope function  $F_i$ , then for every pair  $f$  and  $f_0$  in  $\mathcal{F}$ ,

$$|\phi \circ f - \phi \circ f_0|^2 \leq 4 \sum_{i=1}^k L_{\alpha,i}^2 F_i^{2\alpha_i}.$$

Thus, the function  $2(\sum_{i=1}^k L_{\alpha,i}^2 F_i^{2\alpha_i})^{1/2}$ , which will be denoted  $2L_\alpha \cdot F^\alpha$ , is an envelope function of the class  $\phi(\mathcal{F}) - \phi(f_0)$ . We use this function to standardize the entropy numbers of the class  $\phi \circ \mathcal{F}$ .

**2.10.20 Theorem.** Let  $\mathcal{F}_1, \dots, \mathcal{F}_k$  be classes of measurable functions with measurable envelopes  $F_i$ , and let  $\phi: \mathbb{R}^k \mapsto \mathbb{R}$  be a map that satisfies (2.10.19). Then, for every  $\delta > 0$ ,

$$\begin{aligned} & \int_0^\delta \sup_Q \sqrt{\log N(\varepsilon \|L_\alpha \cdot F^\alpha\|_{Q,2}, \phi(\mathcal{F}), L_2(Q))} d\varepsilon \\ & \leq \sum_{i=1}^k \int_0^{\delta^{1/\alpha_i}} \sup_Q \sqrt{\log N(\varepsilon \|F_i\|_{Q,2\alpha_i}, \mathcal{F}_i, L_{2\alpha_i}(Q))} \frac{d\varepsilon}{\varepsilon^{1-\alpha_i}}. \end{aligned}$$

Here the supremum is taken over all finitely discrete probability measures  $Q$ . Consequently, if the right side is finite and  $P^*(L_\alpha \cdot F^\alpha)^2 < \infty$ , then the class  $\phi(\mathcal{F})$  is Donsker provided its members are square integrable and the class is suitably measurable.

**Proof.** For a given finitely discrete probability measure  $Q$ , define measures  $R_i$  by  $dR_i = L_{\alpha,i}^2 dQ$ . Then

$$\|\phi \circ f - \phi \circ g\|_{Q,2}^2 \leq \sum R_i |f_i - g_i|^{2\alpha_i}.$$

For each  $i$ , construct a minimal  $\varepsilon^{1/\alpha_i} \|F_i\|_{R_i,2\alpha_i}$ -net  $\mathcal{G}_i$  for  $\mathcal{F}_i$  with respect to the  $L_{2\alpha_i}(R_i)$ -norm. Then the set of points  $\phi(g_1, \dots, g_k)$  with  $(g_1, \dots, g_k)$  ranging over all possible combinations of  $g_i$  from  $\mathcal{G}_i$  forms a net for  $\phi(\mathcal{F})$  of  $L_2(Q)$ -size bounded by the square root of

$$\sum \varepsilon^2 \|F_i\|_{R_i,2\alpha_i}^{2\alpha_i} = \varepsilon^2 Q \sum L_{\alpha,i}^2 F_i^{2\alpha_i}.$$

This equals  $\varepsilon^2 \|L_\alpha \cdot F^\alpha\|_{Q,2}^2$ . Conclude that

$$N(\varepsilon \|L_\alpha \cdot F^\alpha\|_{Q,2}, \phi(\mathcal{F}), L_2(Q)) \leq \prod_{i=1}^k N(\varepsilon^{1/\alpha_i} \|F_i\|_{R_i,2\alpha_i}, \mathcal{F}_i, L_{2\alpha_i}(R_i)).$$

Since expressions of the type  $N(\varepsilon \|F\|_{cR,r}, \mathcal{F}, L_r(cR))$  are constant in  $c$ , the bound is also valid with  $R_i$  replaced by the probability measure  $R_i/R_i 1$ .

If  $Q$  ranges over all finitely discrete measures, then each of the  $R_i$  runs through finitely discrete measures. Substitute the bound of the preceding display in the left side of the theorem, and next make a change of variables  $\varepsilon^{1/\alpha_i} \mapsto \varepsilon$  to conclude the proof. ■

**2.10.21 Example.** If  $\mathcal{F}$  satisfies the uniform entropy condition and  $P^* F^{2k} < \infty$  for  $k > 1$ , then  $\mathcal{F}^k$  is Donsker if suitably measurable. In comparison with Theorem 2.10.6, this result has traded a stronger entropy condition for a weaker (almost optimal) condition on the envelope function: Theorem 2.10.6 shows that  $\mathcal{F}^k$  is Donsker if  $\mathcal{F}$  is Donsker and uniformly bounded.

The present result follows since  $|f^k - g^k| \leq |f - g|(2F)^{k-1}$ , so that (2.10.19) applies with  $\alpha = 1$  and  $L = (2F)^{k-1}$ .

**2.10.22 Example.** Let the functions in  $\mathcal{F}$  be nonnegative with lower envelope function  $\underline{F}$ , and consider the class  $\sqrt{\mathcal{F}}$ . If  $\mathcal{F}$  satisfies the uniform-entropy condition and  $P^* F^2 / \underline{F} < \infty$ , then  $\sqrt{\mathcal{F}}$  is Donsker if suitably measurable. This follows since  $|\sqrt{f} - \sqrt{g}| \leq |f - g|/\sqrt{\underline{F}}$ .

The conclusion that  $\sqrt{\mathcal{F}}$  is Donsker can also be obtained under other combinations of moment and entropy conditions. First, Theorem 2.10.6 shows that  $\sqrt{\mathcal{F}}$  is Donsker if  $\mathcal{F}$  is Donsker and  $\underline{F} \geq \delta > 0$ . This combines a stronger assumption on the lower envelope with a weaker entropy condition. Second, the present theorem can be applied with a different Lipschitz order. For  $1/2 \leq \alpha \leq 1$ ,

$$\frac{|\sqrt{f} - \sqrt{g}|}{|f - g|^\alpha} = \frac{|\sqrt{f} - \sqrt{g}|^{1-\alpha}}{|\sqrt{f} + \sqrt{g}|^\alpha} \leq (2\sqrt{\underline{F}})^{1-2\alpha}.$$

Thus, a suitably measurable class  $\sqrt{\mathcal{F}}$  is Donsker if, for some  $\alpha \geq 1/2$ , the upper and lower envelopes of  $\mathcal{F}$  satisfy  $P^* F^{2\alpha} \underline{F}^{1-2\alpha} < \infty$  and, in addition,  $\mathcal{F}$  satisfies

$$\int_0^\infty \sup_Q \sqrt{\log N(\varepsilon \|F\|_{Q,2\alpha}, \mathcal{F}, L_{2\alpha}(Q))} \frac{d\varepsilon}{\varepsilon^{1-\alpha}} < \infty.$$

It appears that finiteness of the integral is more restrictive than a finite uniform  $L_2$ -entropy integral (though the supremum in the integrand is increasing in  $\alpha$ ). For polynomial classes  $\mathcal{F}$  and more generally classes with uniform  $L_1$ -entropy of the order  $(1/\varepsilon)^r$  with  $r < 1$ , the integral is certainly finite for  $\alpha = 1/2$ . For such classes, the class  $\sqrt{\mathcal{F}}$  is Donsker provided its envelope  $\sqrt{\underline{F}}$  is square integrable.

**2.10.23 Example.** Let  $\mathcal{F}$  and  $\mathcal{G}$  satisfy the uniform-entropy condition and be suitably measurable. Then  $\mathcal{FG}$  is Donsker provided the envelopes  $F$  and  $G$  satisfy  $P^* F^2 G^2 < \infty$ .

This follows from the theorem with  $L_\alpha = (G, F)$  and Lipschitz orders  $\alpha = (1, 1)$ .

#### 2.10.4 Partitions of the Sample Space

Let  $\mathcal{X} = \cup_{j=1}^\infty \mathcal{X}_j$  be a partition of  $\mathcal{X}$  into measurable sets, and let  $\mathcal{F}_j$  be the class of functions  $f1_{\mathcal{X}_j}$  when  $f$  ranges over  $\mathcal{F}$ . If the class  $\mathcal{F}$  is Donsker, then each class  $\mathcal{F}_j$  is Donsker. This follows from Theorem 2.10.6, because a restriction is a contraction in the sense that  $|((f1_{\mathcal{X}_j})(x) - (g1_{\mathcal{X}_j})(x))| \leq |f(x) - g(x)|$  for every  $x$ . Consider the converse of this statement. While the sum of infinitely many Donsker classes need not be Donsker, it is clear that if each  $\mathcal{F}_j$  is Donsker and the classes  $\mathcal{F}_j$  become suitably small as  $j \rightarrow \infty$ , then  $\mathcal{F}$  is Donsker.

**2.10.24 Theorem.** For each  $j$ , let the  $\mathcal{F}_j$ , the restriction of  $\mathcal{F}$  to  $\mathcal{X}_j$ , be Donsker and satisfy

$$\mathbb{E}^* \|\mathbb{G}_n\|_{\mathcal{F}_j} \leq C c_j,$$

for a constant  $C$  not depending on  $j$  or  $n$ . If  $\sum_{j=1}^{\infty} c_j < \infty$  and  $P^* F < \infty$  for some envelope function, then the class  $\mathcal{F}$  is Donsker.

**Proof.** We can assume without loss of generality that the class  $\mathcal{F}$  contains the constant function 1.

Since  $\mathcal{F}_j$  is Donsker, the sequence of empirical processes indexed by  $\mathcal{F}_j$  converges in distribution in  $\ell^\infty(\mathcal{F}_j)$  to a tight brownian bridge  $\mathbb{H}_j$  for each  $j$ . This implies that  $\mathbb{E}^* \|\mathbb{G}_n\|_{\mathcal{F}_j} \rightarrow \mathbb{E} \|\mathbb{H}_j\|_{\mathcal{F}_j}$ . Thus,  $\mathbb{E} \|\mathbb{H}_j\|_{\mathcal{F}_j} \leq C c_j$ . Let  $Z_j$  be a standard normal variable independent of  $\mathbb{H}_j$ , constructed on the same probability space. Since  $\sup_{f \in \mathcal{F}_j} |Pf| < \infty$ , the process  $f \mapsto \mathbb{Z}_j(f) = \mathbb{H}_j(f) + Z_j P f$  is well defined on  $\mathcal{F}_j$  and takes its values in  $\ell^\infty(\mathcal{F}_j)$ . We can construct these processes for different  $j$  as independent random elements on a single probability space. Then the series  $\mathbb{Z}(f) = \sum_{j=1}^{\infty} \mathbb{Z}_j(f 1_{\mathcal{X}_j})$  converges in second mean for every  $f$ , and satisfies  $\mathbb{E} \mathbb{Z}(f) \mathbb{Z}(g) = P f g$ , for every  $f$  and  $g$ . Thus, the series defines a version of a brownian motion process. Since each of the processes  $\{\mathbb{Z}_j(f 1_{\mathcal{X}_j}): f \in \mathcal{F}\}$  has bounded and uniformly continuous sample paths with respect to the  $L_2(P)$ -seminorm, so have the partial sums  $\mathbb{Z}_{\leq k} = \{\sum_{j=1}^k \mathbb{Z}_j(f 1_{\mathcal{X}_j}): f \in \mathcal{F}\}$ . Furthermore,

$$\begin{aligned} \mathbb{E} \sup_{f \in \mathcal{F}} \left| \sum_{j>k} \mathbb{Z}_j(f 1_{\mathcal{X}_j}) \right| &\leq \sum_{j>k} (\mathbb{E} \|\mathbb{H}_j\|_{\mathcal{F}_j} + \mathbb{E} |Z_j| P^* F 1_{\mathcal{X}_j}) \\ &\leq C \sum_{j>k} c_j + \sqrt{\frac{2}{\pi}} P^* F 1_{\cup_{j>k} \mathcal{X}_j}. \end{aligned}$$

This converges to zero as  $k \rightarrow \infty$ . Thus the series  $\mathbb{Z}(f) = \sum_{j=1}^{\infty} \mathbb{Z}_j(f 1_{\mathcal{X}_j})$  converges in mean in the space  $\ell^\infty(\mathcal{F})$ . By the Ito-Nisio theorem, it also converges almost surely. Conclude that almost all sample paths of the process  $\mathbb{Z}$  are bounded and uniformly continuous. Since  $\mathbb{E} \|\mathbb{Z}\|_{\mathcal{F}} < \infty$ , the class  $\mathcal{F}$  is totally bounded in  $L_2(P)$ , by Sudakov's inequality. Hence  $\mathbb{Z}$  is a tight version of a brownian motion process indexed by  $\mathcal{F}$ . The process  $\mathbb{G}(f) = \mathbb{Z}(f) - \mathbb{Z}(1) P f$  defines a tight brownian bridge process indexed by  $\mathcal{F}$ . We have proved that  $\mathcal{F}$  is pre-Gaussian.

For each  $k$  set  $\mathbb{G}_{n,\leq k}(f) = \mathbb{G}_n(f 1_{\cup_{j \leq k} \mathcal{X}_j})$ . The continuity modulus of the process  $\{\mathbb{G}_{n,\leq k}(f); f \in \mathcal{F}\}$  is bounded by the continuity modulus of the empirical process indexed by the sum class  $\sum_{j \leq k} \mathcal{F}_j$ . Since the sum of finitely many Donsker classes is Donsker we can conclude that the sequence  $(\mathbb{G}_{n,\leq k})_{n=1}^{\infty}$  is asymptotically tight in  $\ell^\infty(\mathcal{F})$ . Considering the marginal distributions, we can conclude that the sequence converges in distribution to

$$\mathbb{G}_{\leq k}(f) = \mathbb{Z}_{\leq k}(f) - d_k \mathbb{Z}_{\leq k}(1) P f 1_{\cup_{j \leq k} \mathcal{X}_j},$$

as  $n \rightarrow \infty$  for each fixed  $k$ , for  $d_k$  a solution to the equation  $d^2 P(\cup_{j \leq k} \mathcal{X}_j) - 2d = -1$ .

As  $k \rightarrow \infty$ , the sequence  $\mathbb{G}_{\leq k}$  converges almost surely, whence in distribution to  $\mathbb{G}$  in  $\ell^\infty(\mathcal{F})$ . Since weak convergence to a tight limit is metrizable, and the array  $\mathbb{G}_{n,\leq k}$  converges along every row to limits that converge to  $\mathbb{G}$ , there exist integers  $k_n \rightarrow \infty$  such that the sequence  $\mathbb{G}_{n,\leq k_n}$  converges in distribution to  $\mathbb{G}$ . We also have that the sequence  $\mathbb{G}_n - \mathbb{G}_{n,\leq k_n}$  converges in outer probability to zero, since

$$\mathbb{E}^* \|\mathbb{G}_n - \mathbb{G}_{n,\leq k_n}\|_{\mathcal{F}} \leq C \sum_{j > k_n} c_j \rightarrow 0.$$

An application of Slutsky's lemma completes the proof. ■

Methods to obtain maximal inequalities for the  $L_1$ -norm are discussed in Section 2.14.1. We give three applications of the preceding theorem.

**2.10.25 Example (Smooth functions).** Let  $\mathbb{R}^d = \cup_{j=1}^\infty \mathcal{X}_j$  be a partition of  $\mathbb{R}^d$  into uniformly bounded, convex sets with nonempty interior. Consider the class  $\mathcal{F}$  of functions such that the class  $\mathcal{F}_j$  of restrictions is contained in  $C_{M_j}^\alpha(\mathcal{X}_j)$  for each  $j$  for given constants  $M_j$ .

If  $\alpha > d/2$  and  $\sum_{j=1}^\infty M_j P(\mathcal{X}_j)^{1/2} < \infty$ , then the class  $\mathcal{F}$  is  $P$ -Donsker.

**2.10.26 Example (Closed convex sets).** Let  $\mathbb{R}^2 = \cup_{j=1}^\infty \mathcal{X}_j$  be a partition into squares of fixed size. Let  $\mathcal{C}$  be the class of closed convex subsets of  $\mathbb{R}^2$ .

Then  $\mathcal{C}$  is  $P$ -Donsker for every probability measure  $P$  with a density  $p$  such that  $\sum \|p\|_{\mathcal{X}_j}^{1/2} < \infty$ . In particular, this is the case if  $(1+|xy|^{2+\delta})p(x,y)$  is bounded for some  $\delta > 0$ .

**2.10.27 Example (Monotone functions).** Consider the class  $\mathcal{F}$  of all nondecreasing functions  $f: \mathbb{R} \rightarrow \mathbb{R}$ , such that  $0 \leq f \leq F$ , for a given nondecreasing function  $F$ . This class is Donsker provided  $\|F\|_{2,1} < \infty$ .

To see this, first reduce the problem to the case of uniform  $[0,1]$  observations by the quantile transformation. If  $P^{-1}$  is the quantile function corresponding to the underlying measure  $P$ , then the class  $\mathcal{F} \circ P^{-1}$  is Donsker with respect to the uniform measure if and only if  $\mathcal{F}$  is  $P$ -Donsker. The class  $\mathcal{G} = \mathcal{F} \circ P^{-1}$  consists of monotone functions  $g: [0, 1] \rightarrow \mathbb{R}$ , with  $0 \leq g \leq F \circ P^{-1}$ . Furthermore,

$$\int_0^1 \frac{F \circ P^{-1}}{\sqrt{1-u}} du \leq \int \frac{F(x)}{\sqrt{1-P(-\infty, x)}} dP(x).$$

By partial integration the latter can be shown to be finite if and only if  $\|F\|_{2,1}$  is finite. For simplicity of notation assume now that  $P$  is the uniform measure.

Use the theorem with the partition into the intervals  $\mathcal{X}_j = (x_j, x_{j+1}]$ , with  $x_j = 1 - 2^{-j}$  for each integer  $j \geq 0$ . The condition of the theorem involves the series

$$\sum_{j=0}^{\infty} \|F_j\|_{P,2} = \sum_{j=1}^{\infty} \frac{F(1 - 2^{-j})}{\sqrt{1 - (1 - 2^{-j})}} 2^{-j} \leq 2 \int_0^1 \frac{F(u)}{\sqrt{1-u}} du.$$

The result follows by the quantile transformation.

## Problems and Complements

1. Let  $\{\psi_i\}$  be an orthonormal base of a subspace of  $L_2(P)$  that contains  $\mathcal{F}$  and the constant functions, and let  $Z_1, Z_2, \dots$  be i.i.d. standard normally distributed random variables. If  $\mathcal{F}$  is  $P$ -pre-Gaussian, then the series  $\sum_{i=1}^{\infty} Z_i (P(f - Pf)\psi_i)$  is almost surely convergent in  $\ell^{\infty}(\mathcal{F})$  and represents a tight Brownian bridge process.

[Hint: The series converges in second mean for every single  $f$ , and the limit process (series)  $\{\mathbb{G}(f): f \in L_2(P)\}$  is a well-defined stochastic process that is continuous in second mean with respect to  $\rho_P$ . Since  $\mathcal{F}$  is separable, the latter implies that  $\{\mathbb{G}(f): f \in \mathcal{F}\}$  is a separable process. A separable version of a uniformly continuous process is automatically uniformly continuous. Thus,  $\{\mathbb{G}(f): f \in \mathcal{F}\}$  defines a tight Brownian bridge.]

Let  $\mathbb{G}_n$  be the  $n$ th partial sum of the series, and set  $\mathcal{F}_{\delta} = \{f - g: f, g \in \mathcal{F}, \rho_P(f - g) < \delta\}$ . By Jensen's inequality,  $E\|\mathbb{G}_n\|_{\mathcal{F}_{\delta}} \leq E\|\mathbb{G}\|_{\mathcal{F}_{\delta}}$ . Since  $\mathcal{F}$  is pre-Gaussian, the right side converges to zero as  $\delta \downarrow 0$ . Hence the sequence  $\mathbb{G}_n$  is uniformly tight in  $\ell^{\infty}(\mathcal{F})$ . Since  $\mathbb{G}_n - \mathbb{G} \rightsquigarrow 0$  marginally, it follows that  $\mathbb{G}_n - \mathbb{G} \rightsquigarrow 0$  in  $\ell^{\infty}(\mathcal{F})$ . By the Itô-Nisio theorem A.1.3, it converges almost surely also.]

2. The bounded Lipschitz distance between the normal distributions  $N(0, \sigma^2)$  and  $N(0, \tau^2)$  on the line is bounded below by  $|\sigma - \tau| g(1 \wedge \tau^{-1} \wedge \sigma^{-1})$  for  $g(t) = t^2 \phi(t)$  and  $\phi$  the standard normal density.

[Hint: For  $0 \leq c \leq 1$ , the function  $h(x) = (c - |x|) \vee 0$  is contained in  $BL_1(\mathbb{R})$ . For  $\sigma < \tau$ , the integral  $|\int (h(x\sigma) - h(x\tau)) \pi(x) dx|$  is bounded below by  $\phi(c/\tau)$  times  $\int_{-c/\tau}^{c/\tau} (h(x\sigma) - h(x\tau)) dx = (\tau - \sigma)(c/\tau)^2$ . Take  $c = \tau \wedge 1$ .]

3. For any class  $\mathcal{F}$  of functions with envelope function  $F$  and  $0 < r < s < \infty$ ,

$$\sup_Q N(2\varepsilon \|F\|_{Q,s}, \mathcal{F}, L_s(Q)) \leq \sup_Q N(2\varepsilon^{r/s} \|F\|_{Q,r}, \mathcal{F}, L_r(Q))$$

if the supremum is taken over all finitely discrete measures.

[Hint: If  $dR = (2F)^{s-r} dQ$ , then  $\|f - g\|_{Q,s} \leq \|f - g\|_{R,r}^{r/s}$ .]

4. For any class  $\mathcal{F}$  of functions with strictly positive envelope function  $F$ , the expression  $\sup_Q N(\varepsilon \|F\|_{Q,r}, \mathcal{F}, L_r(Q))$ , where the supremum is taken over all finitely discrete measures, is nondecreasing in  $0 < r < \infty$ . (Here  $L_r(Q)$  is equipped with  $(Q|f|^r)^{1/r}$  even though this is not a norm for  $r < 1$ .)

[Hint: Given  $r < s$  and  $Q$  define  $dR = F^{r-s} dQ$ . Then  $Q|f - g|^r \leq \|f - g\|_{R,s}^r \|F\|_{R,s}^{s-r}$ , by Hölder's inequality, with  $(p, q) = (s/r, s/(s-r))$ .]

5. The expression  $N(\varepsilon \|aF\|_{bQ,r}, a\mathcal{F}, L_r(bQ))$  is constant in  $a$  and  $b$ .
6. If  $\mathcal{F}$  is  $P$ -pre-Gaussian and  $A: L_2(P) \mapsto L_2(P)$  is continuous, then the class  $\{Af: f \in \mathcal{F}\}$  is  $P$ -pre-Gaussian.
7. If  $\mathcal{F}$  is  $P$ -Donsker and  $A: \mathcal{L}_2(P) \mapsto \mathcal{L}_2(P)$  is an orthogonal projection, then the class  $\{Af: f \in \mathcal{F}\}$  is not necessarily  $P$ -Donsker.
8. If both  $\mathcal{F}$  and  $\mathcal{G}$  are Donsker, then the class  $\mathcal{F} \cdot \mathcal{G}$  of products  $x \mapsto f(x)g(x)$  is Glivenko-Cantelli in probability.

## 2.11

# The Central Limit Theorem for Processes

So far we have focused on limit theorems and inequalities for the empirical process of independent and identically distributed random variables. Most of the methods of proof apply more generally. In this chapter we indicate some extensions to the case of independent but not identically distributed processes.

We consider the general situation of sums of independent stochastic processes  $\{Z_{ni}(f): f \in \mathcal{F}\}$  with bounded sample paths indexed by an arbitrary set  $\mathcal{F}$ .

### 2.11.1 Random Entropy

For each  $n$ , let  $Z_{n1}, \dots, Z_{nm_n}$  be independent stochastic processes indexed by a common (arbitrary) semimetric space  $(\mathcal{F}, \rho)$ . As usual, for computation of outer expectations, the independence is understood in the sense that the processes are defined on a product probability space  $\prod_{i=1}^{m_n} (\mathcal{X}_{ni}, \mathcal{A}_{ni}, P_{ni})$  with each  $Z_{ni}(f) = Z_{ni}(f, x)$  depending only on the  $i$ th coordinate of  $x = (x_1, \dots, x_{m_n})$ . In the first theorem it is assumed that every one of the maps

$$(x_1, \dots, x_{m_n}) \mapsto \sup_{\rho(f,g) < \delta} \left| \sum_{i=1}^{m_n} e_i (Z_{ni}(f) - Z_{ni}(g)) \right|,$$
$$(x_1, \dots, x_{m_n}) \mapsto \sup_{\rho(f,g) < \delta} \left| \sum_{i=1}^{m_n} e_i (Z_{ni}(f) - Z_{ni}(g))^2 \right|,$$

is measurable, for every  $\delta > 0$ , every vector  $(e_1, \dots, e_{m_n}) \in \{-1, 0, 1\}^{m_n}$ , and every natural number  $n$ . (Measurability for the completion of the probability spaces  $\prod_{i=1}^{m_n} (\mathcal{X}_{ni}, \mathcal{A}_{ni}, P_{ni})$  suffices.)

Define a random semimetric by

$$d_n^2(f, g) = \sum_{i=1}^{m_n} (Z_{ni}(f) - Z_{ni}(g))^2.$$

The simplest result on marginal convergence of the processes  $\sum_{i=1}^{m_n} Z_{ni}$  is the Lindeberg central limit theorem. The following theorem gives a uniform central limit theorem under a Lindeberg condition on norms combined with a condition on entropies with respect to the random semimetric  $d_n$ , a “random entropy condition.”

**2.11.1 Theorem.** *For each  $n$ , let  $Z_{n1}, \dots, Z_{nm_n}$  be independent stochastic processes indexed by a totally bounded semimetric space  $(\mathcal{F}, \rho)$ . Assume that the sums  $\sum_{i=1}^{m_n} e_i Z_{ni}$  are measurable as indicated and that*

$$\begin{aligned} & \sum_{i=1}^{m_n} \mathbb{E}^* \|Z_{ni}\|_{\mathcal{F}}^2 \{\|Z_{ni}\|_{\mathcal{F}} > \eta\} \rightarrow 0, \quad \text{for every } \eta > 0, \\ & \sup_{\rho(f,g) < \delta_n} \sum_{i=1}^{m_n} \mathbb{E} (Z_{ni}(f) - Z_{ni}(g))^2 \rightarrow 0, \quad \text{for every } \delta_n \downarrow 0, \\ (2.11.2) \quad & \int_0^{\delta_n} \sqrt{\log N(\varepsilon, \mathcal{F}, d_n)} d\varepsilon \xrightarrow{P^*} 0, \quad \text{for every } \delta_n \downarrow 0. \end{aligned}$$

Then the sequence  $\sum_{i=1}^{m_n} (Z_{ni} - \mathbb{E} Z_{ni})$  is asymptotically  $\rho$ -equicontinuous. It converges in distribution in  $\ell^\infty(\mathcal{F})$  provided the sequence of covariance functions converges pointwise on  $\mathcal{F} \times \mathcal{F}$ .

**Proof.** The Lindeberg condition for norms implies the Lindeberg condition for marginals. Together with the assumption that the covariance function converges, this gives marginal weak convergence (to a Gaussian process).

Set  $Z_{ni}^o = Z_{ni} - \mathbb{E} Z_{ni}$ , and let  $\delta_n \downarrow 0$  be arbitrary. For fixed  $t$  and sufficiently large  $n$  Chebyshev's inequality and the second condition give the bound  $P(|\sum_{i=1}^{m_n} Z_{ni}^o(f) - Z_{ni}^o(g)| > t/2) \leq 1/2$  for every pair  $f, g$  with  $\rho(f, g) < \delta_n$ . By the symmetrization lemma, Lemma 2.3.7, for sufficiently large  $n$ ,

$$\begin{aligned} & P^* \left( \sup_{\rho(f,g) < \delta_n} \left| \sum_{i=1}^{m_n} Z_{ni}^o(f) - Z_{ni}^o(g) \right| > t \right) \\ & \leq 4P \left( \sup_{\rho(f,g) < \delta_n} \left| \sum_{i=1}^{m_n} \varepsilon_i (Z_{ni}(f) - Z_{ni}(g)) \right| > \frac{t}{4} \right). \end{aligned}$$

For fixed values of the processes  $Z_{n1}, \dots, Z_{nm_n}$ , define the subset  $A_n \subset \mathbb{R}^{m_n}$  as the set of all vectors  $(Z_{n1}(f) - Z_{n1}(g), \dots, Z_{nm_n}(f) - Z_{nm_n}(g))$  when the pairs  $(f, g)$  range over the set  $\{(f, g) \in \mathcal{F} \times \mathcal{F}: \rho(f, g) < \delta_n\}$ . By Hoeffding's

inequality, Lemma 2.2.7, the stochastic process  $\{\sum \varepsilon_i a_i : a \in A_n\}$  is sub-Gaussian for the Euclidean metric on  $A_n$ . By Theorem 2.2.8,

$$\mathbb{E}_\varepsilon \sup_{\rho(f,g) < \delta_n} \left| \sum_{i=1}^{m_n} \varepsilon_i (Z_{ni}(f) - Z_{ni}(g)) \right| \lesssim \int_0^\infty \sqrt{\log N(\varepsilon, A_n, \|\cdot\|)} d\varepsilon,$$

where  $\|\cdot\|$  is the Euclidean norm. The integrand can be bounded using the inequality  $N(\varepsilon, A_n, \|\cdot\|) \leq N^2(\varepsilon/2, \mathcal{F}, d_n)$ . Furthermore, if

$$\theta_n^2 = \sup_{a \in A_n} \sum_{i=1}^{m_n} a_i^2 = \left\| \sum_{i=1}^{m_n} a_i^2 \right\|_{A_n},$$

then, for  $\varepsilon > \theta_n$ , the set  $A_n$  fits in the ball of radius  $\varepsilon$  around the origin and the integrand vanishes. Conclude from this and the entropy condition (2.11.2) that the integral converges to zero in outer probability if  $\theta_n \rightarrow 0$  in probability. Under the measurability assumption, this implies the asymptotic equicontinuity of the sequence  $\sum_{i=1}^{m_n} Z_{ni}^o$ .

By the Lindeberg condition, there exists a sequence of numbers  $\eta_n \downarrow 0$  such that  $\mathbb{E}^* \|\sum a_i^2 \{ \|Z_{ni}\|_{\mathcal{F}}^* > \eta_n \}\|_{A_n}$  converges to zero. Thus, for showing that  $\theta_n \xrightarrow{P^*} 0$ , it is not a loss of generality to assume that each  $Z_{ni}$  satisfies  $\|Z_{ni}\|_{\mathcal{F}} \leq \eta_n$ . Fix  $Z_{n1}, \dots, Z_{nm_n}$ , and take an  $\varepsilon$ -net  $B_n$  for  $A_n$  for the Euclidean norm. For every  $a \in A_n$ , there exists  $b \in B_n$  with

$$|\sum \varepsilon_i a_i^2| = |\sum \varepsilon_i (a_i - b_i)^2 + 2\sum \varepsilon_i (a_i - b_i)b_i + \sum \varepsilon_i b_i^2| \leq \varepsilon^2 + 2\varepsilon\|b\| + |\sum \varepsilon_i b_i^2|.$$

By Hoeffding's inequality, Lemma 2.2.7, the variable  $\sum \varepsilon_i b_i^2$  has Orlicz norm for  $\psi_2$  bounded by a multiple of  $(\sum b_i^4)^{1/2} \leq \eta_n (\sum b_i^2)^{1/2}$ . Apply Lemma 2.2.2 to the third term on the right, and replace suprema over  $B_n$  by suprema over  $A_n$  to obtain

$$\mathbb{E}_\varepsilon \|\sum \varepsilon_i a_i^2\|_{A_n} \lesssim \varepsilon^2 + 2\varepsilon \|\sum a_i^2\|_{A_n}^{1/2} + \sqrt{1 + \log |B_n|} \eta_n \|\sum a_i^2\|_{A_n}^{1/2}.$$

The size of the  $\varepsilon$ -net can be chosen  $|B_n| \leq N^2(\varepsilon/2, \mathcal{F}, d_n)$ . By the entropy condition (2.11.2), this variable is bounded in probability (Problem 2.11.1). Conclude that for some constant  $K$ ,

$$\begin{aligned} \mathbb{P} \left( \|\sum \varepsilon_i a_i^2\|_{A_n} > t \right) \\ \leq P^*(|B_n| > M) + \frac{K}{t} \left[ \varepsilon^2 + (\varepsilon + \eta_n \sqrt{\log M}) \mathbb{E} \|\sum a_i^2\|_{A_n}^{1/2} \right]. \end{aligned}$$

For  $M$  and  $t$  sufficiently large, the right side is smaller than  $1 - v_1$  for the constant  $v_1$  in Hoffmann-Jørgensen's Proposition A.1.5. More precisely, this can be achieved for  $M$  such that  $P^*(|B_n| > M) \leq (1 - v_1)/2$  and  $t$  equal to  $(1 - v_1)/2$  times the numerator of the second term. Then  $t$  is bigger than the  $v_1$ -quantile of the variable  $\|\sum \varepsilon_i a_i^2\|_{A_n}$ . Thus, Proposition A.1.5 yields

$$\begin{aligned} \mathbb{E} \|\sum \varepsilon_i a_i^2\|_{A_n} &\lesssim \mathbb{E} \|\max a_i^2\|_{A_n} + t \\ &\lesssim \eta_n^2 + \varepsilon^2 + (\varepsilon + \eta_n \sqrt{\log M}) \left( \mathbb{E} \|\sum a_i^2\|_{A_n} \right)^{1/2}. \end{aligned}$$

By the second assumption of the theorem,  $\|\sum \mathbf{E} a_i^2\|_{A_n} \rightarrow 0$ . Combine this with Lemma 2.3.6 to see that

$$\mathbf{E}\|\sum a_i^2\|_{A_n} \leq \mathbf{E}\|\sum \varepsilon_i a_i^2\|_{A_n} + o(1) \leq \zeta + \zeta (\mathbf{E}\|\sum a_i^2\|_{A_n})^{1/2},$$

for  $\zeta > \varepsilon^2 \vee \varepsilon$  and sufficiently large  $n$ . The inequality  $c \leq \zeta + \zeta \sqrt{c}$  for a nonnegative number  $c$  implies that  $c \leq (\zeta + \sqrt{\zeta^2 + 4\zeta})^2$ . Apply this to  $c = \mathbf{E}\|\sum a_i^2\|_{A_n}$  and conclude that  $\mathbf{E}\|\sum a_i^2\|_{A_n} \rightarrow 0$  as  $n$  tends to infinity. ■

**2.11.3 Example (I.i.d. observations).** The empirical process of a sample  $X_1, \dots, X_n$  studied in the earlier sections of this part can be recovered by setting  $Z_{ni}(f) = n^{-1/2}f(X_i)$ . In this situation,  $\mathcal{F}$  is a class of measurable functions on the sample space.

Theorem 2.11.1 implies that the class  $\mathcal{F}$  is Donsker if it is suitably measurable, is totally bounded in  $L_2(P)$ , possesses a square-integrable envelope, and satisfies for every sequence  $\delta_n \downarrow 0$

$$\int_0^{\delta_n} \sqrt{\log N(\varepsilon, \mathcal{F}, L_2(\mathbb{P}_n))} d\varepsilon \xrightarrow{P} 0.$$

The latter random-entropy condition is certainly satisfied if  $\mathcal{F}$  satisfies the uniform-entropy condition (2.5.1) and the envelope function is square-integrable. Thus, Theorem 2.11.1 is a generalization of Theorem 2.5.2.

**2.11.4 Example (Truncation).** The Lindeberg condition on the norms is certainly not necessary for the central limit theorem. In combination with truncation, the preceding theorem applies to more general processes. Consider stochastic processes  $Z_{n1}, \dots, Z_{nm_n}$  such that

$$\sum_{i=1}^{m_n} P^*(\|Z_{ni}\|_{\mathcal{F}} > \eta) \rightarrow 0, \quad \text{for every } \eta > 0.$$

Then the truncated processes  $Z_{ni,\eta}(f) = Z_{ni}(f)1\{\|Z_{ni}\|_{\mathcal{F}} \leq \eta\}$  satisfy

$$\sum_{i=1}^{m_n} Z_{ni} - \sum_{i=1}^{m_n} Z_{ni,\eta} \xrightarrow{P^*} 0, \quad \text{in } \ell^\infty(\mathcal{F}).$$

Since this is true for every  $\eta > 0$ , it is also true for every sequence  $\eta_n \downarrow 0$  that converges to zero sufficiently slowly. The processes  $Z_{ni,\eta_n}$  certainly satisfy the Lindeberg condition. If they (or the centered processes  $Z_{ni,\eta_n} - \mathbf{E} Z_{ni,\eta_n}$ ) also satisfy the other conditions of the theorem, then the sequence  $\sum_{i=1}^{m_n} (Z_{ni} - \mathbf{E} Z_{ni,\eta_n})$  converges weakly in  $\ell^\infty(\mathcal{F})$ . The random semimetrics  $d_n$  decrease by the truncation. Hence the conditions for the truncated processes are weaker than those for the original processes.

### 2.11.1.1 Measurelike Processes

The preceding theorem is valid for arbitrary index sets  $\mathcal{F}$ . Consider the special case that  $\mathcal{F}$  is a set of measurable functions  $f: \mathcal{X} \mapsto \mathbb{R}$  on a measurable space  $(\mathcal{X}, \mathcal{A})$  that satisfies a uniform-entropy condition:

$$(2.11.5) \quad \int_0^\infty \sup_{Q \in \mathcal{Q}} \sqrt{\log N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))} d\varepsilon < \infty.$$

Then the preceding theorem readily yields a central limit theorem for processes with increments that are bounded by a (random)  $L_2$ -metric on  $\mathcal{F}$ . Call the processes  $Z_{ni}$  *measurelike* with respect to (random) measures  $\mu_{ni}$  if

$$(Z_{ni}(f) - Z_{ni}(g))^2 \leq \int (f - g)^2 d\mu_{ni}, \quad \text{every } f, g \in \mathcal{F}.$$

For measurelike processes, the random semimetric  $d_n$  is bounded by the  $L_2(\sum \mu_{ni})$ -semimetric, and the entropy condition (2.11.2) can be related to the uniform-entropy condition.

**2.11.6 Lemma.** *Let  $\mathcal{F}$  be a class of measurable functions with envelope function  $F$ . Let  $Z_{n1}, \dots, Z_{nm_n}$  be measurelike processes indexed by  $\mathcal{F}$ . If  $\mathcal{F}$  satisfies the uniform-entropy condition (2.11.5) for a set  $\mathcal{Q}$  that contains the measures  $\mu_{ni}$  and  $\sum_{i=1}^{m_n} \mu_{ni} F^2 = O_P^*(1)$ , then the entropy condition (2.11.2) is satisfied.*

**Proof.** Set  $\mu_n = \sum_{i=1}^{m_n} \mu_{ni}$ . Since  $d_n$  is bounded by the  $L_2(\mu_n)$ -semimetric, we have

$$\begin{aligned} & \int_0^{\delta_n} \sqrt{\log N(\varepsilon, \mathcal{F}, d_n)} d\varepsilon \\ & \leq \int_0^{\delta_n/\|F\|_{\mu_n}} \sqrt{\log N(\varepsilon \|F\|_{\mu_n}, \mathcal{F}, L_2(\mu_n))} d\varepsilon \|F\|_{\mu_n}, \end{aligned}$$

on the set where  $\|F\|_{\mu_n}^2 = \mu_n F^2$  is finite. Abbreviate

$$J(\delta) = \int_0^\delta \sup_Q \sqrt{\log N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))} d\varepsilon.$$

On the set where  $\|F\|_{\mu_n} > \eta$ , the right side of the next-to-last displayed equation is bounded by  $J(\delta_n/\eta) O_P(1)$ . This converges to zero in probability for every  $\eta > 0$ . On the set where  $\|F\|_{\mu_n} \leq \eta$ , we have the bound  $J(\infty) \eta$ . This can be made arbitrarily small by the choice of  $\eta$ . Thus, the entropy condition (2.11.2) is satisfied. ■

**2.11.7 Example.** Suppose the independent processes  $Z_{n1}, \dots, Z_{nm_n}$  are measurelike. Let  $\mathcal{F}$  satisfy the uniform-entropy condition. Assume that, for some probability measure  $P$  with  $P^*F^2 < \infty$ ,

$$\begin{aligned} E^* \sum_{i=1}^{m_n} \mu_{ni} F^2 \{\mu_{ni} F^2 > \eta\} &\rightarrow 0, \quad \text{for every } \eta > 0, \\ \sup_{\|f-g\|_{P,2} < \delta_n} E^* \sum_{i=1}^{m_n} \mu_{ni} (f - g)^2 &\rightarrow 0, \quad \text{for every } \delta_n \downarrow 0, \\ \sum_{i=1}^{m_n} \mu_{ni} F^2 &= O_P^*(1). \end{aligned}$$

Then the preceding lemma verifies the entropy condition (2.11.2), while the first two conditions of the display ensure the other conditions of Theorem 2.11.1. Note that  $\|Z_{ni}\|_{\mathcal{F}} \leq |Z_{ni}(f)| + 4\mu_{ni}F^2$  for any  $f$ .

Thus, the sequence  $\sum_{i=1}^{m_n} (Z_{ni} - EZ_{ni})$  converges in  $\ell^\infty(\mathcal{F})$  provided that the sequence of covariance functions converges pointwise and that measurability conditions are met.

**2.11.8 Example (Weighted empirical processes).** For each  $n$ , let  $X_{n1}, \dots, X_{nm_n}$  be independent random elements in a measurable space  $(\mathcal{X}, \mathcal{A})$ . Let  $X_{ni}$  have law  $P_{ni}$  and let  $P_{ni}f$  exist for each element  $f$  of a class  $\mathcal{F}$  of measurable functions  $f: \mathcal{X} \mapsto \mathbb{R}$ . Given a triangular array of constants  $c_{ni}$ , consider the weighted empirical process  $\mathbb{G}_n(f) = \sum_{i=1}^{m_n} c_{ni}(f(X_{ni}) - P_{ni}f)$ . Suppose  $\mathcal{F}$  satisfies the uniform-entropy condition and that

$$\begin{aligned} \max_{1 \leq i \leq m_n} |c_{ni}| &\rightarrow 0, \\ \sum_{i=1}^{m_n} c_{ni}^2 P_{ni} &\leq P, \end{aligned}$$

for a probability measure  $P$  with  $P^*F^2 < \infty$ .

Then under measurability conditions, the sequence  $\mathbb{G}_n$  converges weakly to a Gaussian process in  $\ell^\infty(\mathcal{F})$ , provided the sequence converges marginally. Furthermore, there is a version of the limit process with uniformly continuous sample paths with respect to the  $L_2(P)$ -semimetric.

This follows from the preceding example applied to the processes  $Z_{ni} = c_{ni}\delta_{X_{ni}}$ , which are measurelike for the measures  $\mu_{ni} = c_{ni}^2\delta_{X_{ni}}$ .

## 2.11.2 Bracketing

For each  $n$ , let  $Z_{n1}, \dots, Z_{nm_n}$  be independent stochastic processes indexed by a common index set  $\mathcal{F}$ . In this chapter we aim at generalizations of the bracketing central limit theorem, Theorem 2.5.6, in two directions: we

prove a version for the non-i.i.d. case, and we also weaken the entropy conditions to conditions involving the existence of either majorizing measures or certain Gaussian processes.

For the first generalization, define, for every  $n$ , the bracketing number  $N_{[]}(\varepsilon, \mathcal{F}, L_2^n)$  as the minimal number of sets  $N_\varepsilon$  in a partition  $\mathcal{F} = \cup_{j=1}^{N_\varepsilon} \mathcal{F}_{\varepsilon j}^n$  of the index set into sets  $\mathcal{F}_{\varepsilon j}^n$  such that, for every partitioning set  $\mathcal{F}_{\varepsilon j}^n$

$$\sum_{i=1}^{m_n} \mathbb{E}^* \sup_{f,g \in \mathcal{F}_{\varepsilon j}^n} |Z_{ni}(f) - Z_{ni}(g)|^2 \leq \varepsilon^2.$$

Note that the partitions are allowed to depend on  $n$ .

**2.11.9 Theorem (Bracketing central limit theorem).** *For each  $n$ , let  $Z_{n1}, \dots, Z_{nm_n}$  be independent stochastic processes with finite second moments indexed by a totally bounded semimetric space  $(\mathcal{F}, \rho)$ . Suppose*

$$\begin{aligned} \sum_{i=1}^{m_n} \mathbb{E}^* \|Z_{ni}\|_{\mathcal{F}} \{ \|Z_{ni}\|_{\mathcal{F}} > \eta \} &\rightarrow 0, && \text{for every } \eta > 0, \\ \sup_{\rho(f,g) < \delta_n} \sum_{i=1}^{m_n} \mathbb{E} (Z_{ni}(f) - Z_{ni}(g))^2 &\rightarrow 0, && \text{for every } \delta_n \downarrow 0, \\ \int_0^{\delta_n} \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}, L_2^n)} d\varepsilon &\rightarrow 0, && \text{for every } \delta_n \downarrow 0. \end{aligned}$$

Then the sequence  $\sum_{i=1}^{m_n} (Z_{ni} - \mathbb{E} Z_{ni})$  is asymptotically tight in  $\ell^\infty(\mathcal{F})$  and converges in distribution provided it converges marginally. If the partitions can be chosen independent of  $n$ , then the middle of the displayed conditions is unnecessary.

A central limit theorem for the empirical process of i.i.d. observations can be recovered from the preceding theorem by setting  $Z_{ni}(f) = n^{-1/2} f(X_i)$ . Then the bracketing numbers in the theorem may be reduced to  $N_{[]}(\varepsilon, \mathcal{F}, L_2(P))$ . The resulting theorem is weaker than the bracketing central limit theorem, Theorem 2.5.6, in that the latter theorem uses a combination of  $L_2(P)$ - and  $L_{2,\infty}(P)$ -entropies. In the present theorem the  $L_2$ -entropies may also be replaced by smaller numbers: the preceding theorem remains true if  $N_{[]}(\varepsilon, \mathcal{F}, L_2^n)$  is replaced by the minimal number of sets  $N_\varepsilon$  in a partition  $\mathcal{F} = \cup_{j=1}^{N_\varepsilon} \mathcal{F}_{\varepsilon j}^n$  of the index set in sets  $\mathcal{F}_{\varepsilon j}^n$  such that, for every  $j$  and every  $n$ ,

$$\begin{aligned} \sup_{f,g \in \mathcal{F}_{\varepsilon j}^n} \sum_{i=1}^{m_n} \mathbb{E} (Z_{ni}(f) - Z_{ni}(g))^2 &\leq \varepsilon^2, \\ \sup_{t>0} \sum_{i=1}^{m_n} t^2 \mathbb{P}^* \left( \sup_{f,g \in \mathcal{F}_{\varepsilon j}^n} |Z_{ni}(f) - Z_{ni}(g)| > t \right) &\leq \varepsilon^2. \end{aligned}$$

These bracketing numbers appear to be rather hard to work with.

Another refinement of the theorem is to replace its entropy assumption by a majorizing measure condition. This entails the existence of partitions  $\mathcal{F} = \cup_j \mathcal{F}_{\varepsilon_j}^n$  as before and discrete probability measures  $\mu_n$  on  $\mathcal{F}$  such that

$$(2.11.10) \quad \sup_f \int_0^{\delta_n} \sqrt{\log \frac{1}{\mu_n(\mathcal{F}_\varepsilon^n f)}} d\varepsilon \rightarrow 0, \quad \text{for every } \delta_n \downarrow 0.$$

Here  $\mathcal{F}_\varepsilon^n f$  is defined as the partitioning set at level  $\varepsilon$  to which  $f$  belongs. This majorizing measure condition is a weakening of finiteness of the entropy integral. Entropies correspond to certain uniform majorizing measures; the majorizing measure criterion permits one to measure the size of  $\mathcal{F}$  in a nonuniform way.<sup>#</sup>

Majorizing measures are also hard to work with. The resulting central limit theorem can be expressed more elegantly in terms of the existence of Gaussian semimetrics. For simplicity, we utilize only a single semimetric, although it may be fruitful to allow the semimetric to depend on  $n$ . Call a semimetric  $\rho$  *Gaussian* if it is the standard deviation semimetric

$$\rho(f, g) = \left( E(G(f) - G(g))^2 \right)^{1/2}$$

of a tight, zero-mean, Gaussian random element  $G$  in  $\ell^\infty(\mathcal{F})$ . The connection with majorizing measures is the characterization of continuity of Gaussian processes by majorizing measures. See Appendix A.2.3 for a discussion. In this chapter this deep characterization is used only in the proof of the following theorem to translate its condition into the majorizing measure condition (2.11.10). The proof itself is next based on the existence of the majorizing measure.

Call a semimetric  $\rho$  *Gaussian-dominated* if it is bounded above (on  $\mathcal{F} \times \mathcal{F}$ ) by a Gaussian semimetric. Any semimetric  $\rho$  such that

$$\int_0^\infty \sqrt{\log N(\varepsilon, \mathcal{F}, \rho)} d\varepsilon < \infty$$

is Gaussian-dominated (see Problem 2.11.4).

**2.11.11 Theorem (Bracketing by Gaussian hypotheses).** *For each  $n$ , let  $Z_{n1}, \dots, Z_{nm_n}$  be independent stochastic processes indexed by an arbitrary index set  $\mathcal{F}$ . Suppose that there exists a Gaussian-dominated semimetric  $\rho$  on  $\mathcal{F}$  such that*

$$\sum_{i=1}^{m_n} E^* \|Z_{ni}\|_{\mathcal{F}} \{ \|Z_{ni}\|_{\mathcal{F}} > \eta \} \rightarrow 0, \quad \text{for every } \eta > 0,$$

$$\sum_{i=1}^{m_n} E(Z_{ni}(f) - Z_{ni}(g))^2 \leq \rho^2(f, g), \quad \text{for every } f, g,$$

---

<sup>#</sup> See the Problems and Appendix A.2.3.

$$\sup_{t>0} \sum_{i=1}^{m_n} t^2 P^* \left( \sup_{f,g \in B(\varepsilon)} |Z_{ni}(f) - Z_{ni}(g)| > t \right) \leq \varepsilon^2,$$

for every  $\rho$ -ball  $B(\varepsilon) \subset \mathcal{F}$  of radius less than  $\varepsilon$  and for every  $n$ . Then the sequence  $\sum_{i=1}^{m_n} (Z_{ni} - E Z_{ni})$  is asymptotically tight in  $\ell^\infty(\mathcal{F})$ . It converges in distribution provided it converges marginally.

The specialization of the preceding theorem to the case of i.i.d. observations is of interest because of the use of a Gaussian semimetric instead of entropy integrals. For instance, the theorem contains the classical Chibisov-O'Reilly theorem on the weighted empirical distribution function on the real line (see Example 2.11.15).

To recover the i.i.d. case, let  $X_1, \dots, X_n$  be a random sample in a measurable space  $(\mathcal{X}, \mathcal{A})$ . For a class  $\mathcal{F}$  of measurable functions  $f: \mathcal{X} \mapsto \mathbb{R}$ , set  $Z_{ni}(f) = f(X_i)/\sqrt{n}$ . Recall that  $\|f\|_{P,2,\infty}$  is the weak  $L_2$ -pseudonorm, defined as the square root of  $\sup_{t>0} t^2 P(|f| > t)$ .

**2.11.12 Corollary.** *Let  $\mathcal{F}$  be a pre-Gaussian class of measurable functions whose envelope function possesses a weak second moment. Suppose there exists a Gaussian-dominated semimetric  $\rho$  on  $\mathcal{F}$  such that, for every  $\varepsilon > 0$ ,*

$$\left\| \sup_{f,g \in B(\varepsilon)} |f - g| \right\|_{P,2,\infty} \leq \varepsilon,$$

for every  $\rho$ -ball  $B(\varepsilon) \subset \mathcal{F}$  of radius  $\varepsilon$ . Then  $\mathcal{F}$  is  $P$ -Donsker.

**Proof.** Since  $\mathcal{F}$  is  $P$ -pre-Gaussian and  $\|P\|_{\mathcal{F}} \leq P^* F < \infty$ , there exists a tight version of Brownian motion. This means that the  $L_2(P)$  semimetric  $e_P$  is Gaussian. The semimetric  $d = \sqrt{e_P^2 + \rho^2}$  is the standard deviation metric of the sum of Brownian motion and an independent copy of the process corresponding to  $\rho$ . Thus, this semimetric is Gaussian also. The conditions of the preceding theorem are satisfied for  $d$  playing the role of  $\rho$ . ■

**2.11.13 Example (Jain-Marcus theorem).** For each natural number  $n$ , let  $Z_{n1}, \dots, Z_{n,m_n}$  be independent stochastic processes indexed by an arbitrary index set  $\mathcal{F}$  such that

$$|Z_{ni}(f) - Z_{ni}(g)| \leq M_{ni} \rho(f, g), \quad \text{for every } f, g,$$

for independent random variables  $M_{n1}, \dots, M_{n,m_n}$  and a semimetric  $\rho$  such that

$$\int_0^\infty \sqrt{\log N(\varepsilon, \mathcal{F}, \rho)} d\varepsilon < \infty,$$

$$\sum_{i=1}^{m_n} E M_{ni}^2 = O(1).$$

If the triangular array of norms  $\|Z_{ni}\|_{\mathcal{F}}$  satisfies the Lindeberg condition, then the sequence  $\sum_{i=1}^{m_n}(Z_{ni} - EZ_{ni})$  converges in distribution in  $\ell^\infty(\mathcal{F})$  to a tight Gaussian process provided the sequence of covariance functions converges pointwise on  $\mathcal{F} \times \mathcal{F}$ .

This follows from Theorems 2.11.9 and 2.11.11. The entropy condition implies that  $\rho$  is Gaussian-dominated.

In view of the preceding discussion, the conditions may be relaxed in two ways: the  $L_2$ -norm can be partly replaced by the weak  $L_2$ -norm, and the entropy criterion can be replaced by the majorizing measure criterion.

**2.11.14 Example.** Let  $Z, Z_1, Z_2, \dots$  be i.i.d. stochastic processes indexed by the unit interval  $\mathcal{F} = [0, 1] \subset \mathbb{R}$ , with  $\|Z\|_{\mathcal{F}} \leq 1$ , such that

$$\mathbb{E}|Z(f) - Z(g)| \leq K|f - g|,$$

for some constant  $K$ . Then the sequence  $n^{-1/2} \sum_{i=1}^n (Z_i - EZ_i)$  converges in distribution to a tight Gaussian process in  $\ell^\infty[0, 1]$ . Note that the condition bounds the mean of the increments, not the increments themselves as required by the Jain-Marcus theorem. On the other hand, the index set must be a compact interval in the real line and the processes uniformly bounded.

The result can be deduced from Theorem 2.11.11. For every interval  $[a, a + \varepsilon]$ ,

$$\sup_{a < f \leq g < a + \varepsilon} |Z(f) - Z(g)| \leq \lim_{n \rightarrow \infty} \sum_{k=1}^{2^n} |Z(a + \varepsilon k 2^{-n}) - Z(a + \varepsilon(k-1) 2^{-n})|.$$

(The process is separable, with any dense subset of  $[0, 1]$  as the separant, because it is continuous in probability; the sums on the right side are increasing in  $n$  and bound dyadic increments.) The right-hand side has mean bounded by  $K\varepsilon$ . Since the processes are bounded by 1

$$\mathbb{E} \sup_{a < f \leq g < a + \varepsilon} |Z(f) - Z(g)|^2 \leq 2K\varepsilon.$$

Every set of diameter  $\varepsilon$  for the semimetric  $\rho(f, g) = |f - g|^{1/2}$  is contained in an interval  $[a, a + \varepsilon^2]$ . Conclude that the condition of Theorem 2.11.11 is up to a constant satisfied for this semimetric. This semimetric is Gaussian-dominated, because it has a finite entropy integral.

**2.11.15 Example (Weighted empirical distribution function).** Let  $X_1, X_2, \dots$  be i.i.d. random variables with the uniform distribution  $P$  on  $[0, 1] \subset \mathbb{R}$ . For a fixed function  $q: (0, 1/2] \mapsto \mathbb{R}^+$ , consider the class of functions

$$\mathcal{F} = \left\{ \frac{1_{(0,t]}}{q(t)} : 0 < t \leq \frac{1}{2} \right\}.$$

For simplicity, assume that the function  $1/q$  is decreasing. The empirical process indexed by this class is the classical weighted empirical process

(restricted to  $0 < t \leq 1/2$  for convenience). According to the Chibisov-O'Reilly theorem<sup>†</sup> the following statements are equivalent:

- (i)  $\mathcal{F}$  is Donsker;
- (ii)  $\mathcal{F}$  is pre-Gaussian and  $1/q(t) = o(1/\sqrt{t})$ ;
- (iii)  $\int_0^{1/2} t^{-1} \exp(-\varepsilon q^2(t)/t) dt < \infty$ , for every  $\varepsilon > 0$ .

The equivalence of (ii) and (iii) can be argued from properties of Brownian motion. Here we deduce that (ii) implies (i) from Theorem 2.11.11.

In view of the second condition of (ii), the envelope function  $F = (1/q)1_{(0,1/2]}$  has a weak second moment and  $\|P\|_{\mathcal{F}} = \sup\{t/q(t): 0 < t \leq 1/2\}$  is finite. Hence the pre-Gaussianity implies the existence of a tight version of Brownian motion; its standard deviation metric is Gaussian. Since for  $s < t$ ,

$$\frac{1_{(0,s]}}{q(s)} - \frac{1_{(0,t]}}{q(t)} = \frac{1_{(s,t]}}{q(t)} + \left(\frac{1}{q(t)} - \frac{1}{q(s)}\right)1_{(0,s]},$$

the square of this metric equals

$$\rho^2(s, t) = \frac{t-s}{q^2(t)} + \left|\frac{1}{q(t)} - \frac{1}{q(s)}\right|^2 s, \quad s \leq t.$$

If  $\rho(s, t) < \varepsilon$ , then both terms on the right are bounded by  $\varepsilon^2$ . Conclude that for every fixed  $s$ ,

$$\sup_{\substack{s < t \\ \rho(s,t) < \varepsilon}} \left| \frac{1_{(0,s]}}{q(s)} - \frac{1_{(0,t]}}{q(t)} \right| \leq \sup_{t > s} \frac{\varepsilon 1_{(s,t]}}{\sqrt{t-s}} + \frac{\varepsilon 1_{(0,s]}}{\sqrt{s}} = \frac{\varepsilon 1_{(s,1]}}{\sqrt{1-s}} + \frac{\varepsilon 1_{(0,s]}}{\sqrt{s}}.$$

The  $L_{2,\infty}(P)$ -norm of the function  $x \mapsto 1_{(s,1]}/\sqrt{x-s}$  equals 1 for every  $s$ . It follows that the  $L_{2,\infty}(P)$ -norm of the left side of the preceding display is bounded by a multiple of  $\varepsilon$ .

**2.11.16 Example (Monotone processes).** Let  $Z$  be a stochastic process indexed by an interval  $[a, b] \subset \bar{\mathbb{R}}$ , whose sample paths  $t \mapsto Z(t)$  are non-decreasing. If  $EZ^2(a) < \infty$  and  $EZ^2(b) < \infty$ , then  $Z$  satisfies the central limit theorem in  $\ell^\infty[a, b]$ . More precisely, if  $Z_1, Z_2, \dots$  are i.i.d. copies of  $Z$ , then the sequence  $n^{-1/2} \sum_{i=1}^n (Z_i - EZ_i)$  converges in  $\ell^\infty[a, b]$  to a tight Gaussian process.

We shall derive this from Theorem 2.11.9. The condition that  $Z(a)$  and  $Z(b)$  have finite second moments is necessary for the convergence of the processes evaluated at  $a$  and  $b$ , respectively. If we would be interested in the convergence of the processes in  $\ell^\infty(a, b)$ , then the square integrability could be relaxed considerably, as is clear from the preceding example.

Since we can replace  $Z(t)$  by  $Z(t) - Z(a)$ , we may assume that  $Z(a) = 0$ . The envelope function of  $Z$  is  $\|Z\| = Z(b)$  and is square integrable by assumption. It suffices to verify the entropy condition of Theorem 2.11.9, where we shall choose the partitions independent of  $n$ . Define

---

<sup>†</sup> Shorack and Wellner (1986), page 462.

$F(t) = EZ(t)Z(b)$ . This function is right-continuous with left limits, and is nondecreasing from  $F(a) = 0$  to  $F(b) = EZ^2(b)$ . Given  $\varepsilon > 0$  choose a partition  $a = t_0 < t_1 < \dots < t_N = b$  such that  $F(t_i-) - F(t_i) < \varepsilon^2$  for every  $i$ . Then, for every  $i$ ,

$$E \sup_{t_{i-1} \leq s, t < t_i} |Z(s) - Z(t)|^2 \leq E(Z(t_i-) - Z(t_{i-1}))Z(b) < \varepsilon^2.$$

The number of points in the partition can be chosen smaller than a constant times  $1/\varepsilon^2$ . (Make sure that the points where  $F$  jumps more than  $\varepsilon^2$  are among  $t_0, \dots, t_N$ .) Thus the entropy condition of Theorem 2.11.9 is satisfied easily for the partition  $[a, b] = \cup(t_{i-1}, t_i) \cup \{a\}$ .

The proof of the theorems is based on Bernstein's inequality combined with a chaining argument that utilizes majorizing measures. The following lemma is used to control the finite suprema over the links at a fixed level in the chain.

**2.11.17 Lemma.** *Let  $X$  be an arbitrary random variable such that*

$$P(|X| > x) \leq 2e^{-\frac{1}{2}\frac{x^2}{b+ax}},$$

*for every  $x > 0$ . Then there exists a universal constant  $K$  such that*

$$E|X|1_A \leq K \left( a \log \frac{1}{\mu} + \sqrt{b} \sqrt{\log \frac{1}{\mu}} \right) (\mu + P(A)),$$

*for every measurable set  $A$  and every constant  $0 < \mu < e^{-1}$ .*

**Proof.** The condition implies the upper bound  $2 \exp(-x^2/(4b))$  on  $P(|X| > x)$ , for every  $x \leq b/a$ , and the upper bound  $2 \exp(-x/(4a))$ , for all other positive  $x$ . Consequently, the same upper bounds hold for all  $x > 0$  for the probabilities  $P(|X|1\{|X| \leq b/a\} > x)$  and  $P(|X|1\{|X| > b/a\} > x)$ , respectively. By Lemma 2.2.1, this implies that the Orlicz norms  $\|X1\{|X| \leq b/a\}\|_{\psi_2}$  and  $\|X1\{|X| > b/a\}\|_{\psi_1}$  are up to constants bounded by  $\sqrt{b}$  and  $a$ , respectively. For any random variable  $Y$ , Jensen's inequality yields for any convex, increasing, nonnegative function  $\psi$ ,

$$\begin{aligned} E|Y|1_A &\leq \psi^{-1} \left( E\psi \left( \frac{|Y|}{\|Y\|_\psi} \right) \frac{1_A}{P(A)} \right) \|Y\|_\psi P(A) \\ &\leq \psi^{-1} \left( \frac{1}{P(A)} \right) \|Y\|_\psi P(A). \end{aligned}$$

Apply this to the random variables  $|X|1\{|X| \leq b/a\}$  and  $|X|1\{|X| > b/a\}$  separately and use the triangle inequality to find that

$$E|X|1_A \lesssim \psi_2^{-1} \left( \frac{1}{P(A)} \right) \sqrt{b} P(A) + \psi_1^{-1} \left( \frac{1}{P(A)} \right) a P(A).$$

Finally, apply the inequality  $p \log^k(1 + 1/p) \lesssim (p + \mu) \log^k(1/\mu)$ , which is valid for small  $\mu > 0$  and  $k = 1/2, 1$ . ■

**Proof of Theorems 2.11.9 and 2.11.11.** There exists a sequence of numbers  $\eta_n \downarrow 0$  such that  $\sum E^* \|Z_{ni}\|_{\mathcal{F}} \{\|Z_{ni}\|_{\mathcal{F}} > \eta_n\} \rightarrow 0$ . Therefore, it is no loss of generality to assume that  $\|Z_{ni}\|_{\mathcal{F}} \leq \eta_n$  for every  $i$  and  $n$ . Otherwise, replace  $Z_{ni}$  by  $Z_{ni} 1\{\|Z_{ni}\|_{\mathcal{F}}^* \leq \eta_n\}$ ; the conditions of the theorems remain valid, and the difference between the centered sums and the centered truncated sums converges to zero.

Under the conditions of the theorems, there exists for every  $n$  a sequence of nested partitions  $\mathcal{F} = \cup_j \mathcal{F}_{qj}^n$  and discrete subprobability measures  $\mu_n$  on  $\mathcal{F}$  such that, for every  $j$  and  $n$ ,

$$(2.11.18) \quad \begin{aligned} & \lim_{q_0 \rightarrow \infty} \limsup_{n \rightarrow \infty} \sup_f \sum_{q > q_0} 2^{-q} \sqrt{\log \frac{1}{\mu_n(\mathcal{F}_q^n f)}} = 0, \\ & \sup_{f,g \in \mathcal{F}_{qj}^n} \sum_{i=1}^{m_n} E(Z_{ni}(f) - Z_{ni}(g))^2 \leq 2^{-2q}, \\ & \sup_{t>0} \sum_{i=1}^{m_n} t^2 P^* \left( \sup_{f,g \in \mathcal{F}_{qj}^n} |Z_{ni}(f) - Z_{ni}(g)| > t \right) \leq 2^{-2q}. \end{aligned}$$

Here  $\mathcal{F}_q^n f$  is the set in the  $q$ th partition to which  $f$  belongs. Under the conditions of Theorem 2.11.9, this is true with  $1/\mu_n(\mathcal{F}_q^n f)$  replaced by the number  $N_q^n$  of sets in the  $q$ -th partition. The measures  $\mu_n$  can then be constructed as  $\mu_n = \sum_q 2^{-q} \mu_{n,q}$ , where  $\mu_{n,q}(\mathcal{F}_q^n f) = (N_q^n)^{-1}$ , for every  $f$  and  $q$ . (Alternatively, the expression  $1/\mu_n(\mathcal{F}_q^n f)$  can be replaced by  $N_q^n$  throughout the proof.) Under the conditions of Theorem 2.11.11, a single measure  $\mu = \mu_n$  can be derived from a majorizing measure corresponding to the Gaussian semimetric  $\rho$ . The existence of this measure is indicated in Proposition A.2.17. The construction of a sequence of partitions (in sets of  $\rho$ -diameter at most  $2^{-q}$ ) is carried out in Lemma A.2.19. In the following, it is not a loss of generality to assume that  $\mu_n(\mathcal{F}_q^n f) \leq 1/4$  for every  $q$  and  $f$ . Most of the argument is carried out for a fixed  $n$ ; this index will be suppressed in the notation.

The remainder of the proof is similar to the proof of Theorem 2.5.6, except that Lemma 2.11.17 is substituted for Lemma 2.2.10 in the chaining argument, which now uses majorizing measures. Choose an element  $f_{qj}$  from each partitioning set  $\mathcal{F}_{qj}$  and define

$$\begin{aligned} \pi_q f &= f_{qj}, \\ (\Delta_q f)_{ni} &= \sup_{f,g \in \mathcal{F}_{qj}} |Z_{ni}(f) - Z_{ni}(g)|, \quad \text{if } f \in \mathcal{F}_{qj}, \\ a_q f &= 2^{-q} / \sqrt{\log \frac{1}{\mu(\mathcal{F}_{q+1} f)}}. \end{aligned}$$

Next, for  $q > q_0$ , define indicator functions

$$\begin{aligned}(A_{q-1}f)_{ni} &= 1\{(\Delta_{q_0}f)_{ni} \leq a_{q_0}f, \dots, (\Delta_{q-1}f)_{ni} \leq a_{q-1}f\}, \\(B_qf)_{ni} &= 1\{(\Delta_{q_0}f)_{ni} \leq a_{q_0}f, \dots, (\Delta_{q-1}f)_{ni} \leq a_{q-1}f, (\Delta_qf)_{ni} > a_qf\}, \\(B_{q_0}f)_{ni} &= 1\{(\Delta_{q_0}f)_{ni} > a_{q_0}f\}.\end{aligned}$$

The indicator functions  $A_qf$  and  $B_qf$  are constant in  $f$  on each of the partitioning sets  $\mathcal{F}_{qj}$  at level  $q$ , because the partitions are nested. Now decompose:

$$\begin{aligned}Z_{ni}(f) - Z_{ni}(\pi_{q_0}f) &= (Z_{ni}(f) - Z_{ni}(\pi_{q_0}f))(B_{q_0}f)_{ni} \\(2.11.19) \quad &\quad + \sum_{q_0+1}^{\infty} (Z_{ni}(f) - Z_{ni}(\pi_qf))(B_qf)_{ni} \\&\quad + \sum_{q_0+1}^{\infty} (Z_{ni}(\pi_qf) - Z_{ni}(\pi_{q-1}f))(A_{q-1}f)_{ni}.\end{aligned}$$

Center each term at zero expectation; take the sum over  $i$ ; and, finally, take the supremum over  $f$  for all three terms on the right separately. It will be shown that the resulting expressions converge to zero in mean as  $n \rightarrow \infty$  followed by  $q_0 \rightarrow \infty$ , whence the centered processes  $Z_{ni}^o$  satisfy

$$(2.11.20) \quad \lim_{q_0 \rightarrow \infty} \limsup_{n \rightarrow \infty} E^* \left\| \sum_{i=1}^{m_n} (Z_{ni}^o(f) - Z_{ni}^o(\pi_{q_0}^n f)) \right\|_{\mathcal{F}} = 0.$$

In view of Theorem 1.5.6, this gives a complete proof in the case that the partitions do not depend on  $n$ . The case that the partitions depend on  $n$  requires an additional argument, given at the end of this proof.

Since  $(\Delta_qf)_{ni} \leq 2\eta_n$ , the first term in (2.11.19) is zero as soon as  $2\eta_n \leq \inf_f a_{q_0}f$ . For every fixed  $q_0$ , this is true for sufficiently large  $n$ , because, by assumption (2.11.18),  $\inf_f a_qf$  (which depends on  $n$ ) is bounded away from zero as  $n \rightarrow \infty$  for every fixed  $q_0$ .

By the nesting of the partitions,  $(\Delta_qf)_{ni} \leq (\Delta_{q-1}f)_{ni}$ , which is bounded by  $a_{q-1}f$  on the set where  $(B_{nq}f)_{ni} = 1$ . It follows that

$$\begin{aligned}|Z_{ni}(f) - Z_{ni}(\pi_qf)|(B_qf)_{ni} &\leq (\Delta_qf)_{ni}(B_qf)_{ni} \leq a_{q-1}f, \\ \text{var} \left( \sum_{i=1}^{m_n} (\Delta_qf)_{ni}(B_qf)_{ni} \right) &\leq \sum_{i=1}^{m_n} a_{q-1}f E((\Delta_qf)_{ni} \{ (\Delta_qf)_{ni} > a_qf \}) \\ &\leq 2 \frac{a_{q-1}f}{a_qf} 2^{-2q},\end{aligned}$$

by Problem 2.5.6. Combination with Bernstein's inequality 2.2.9 shows that the random variable  $\sum_{i=1}^{m_n} (\Delta_qf)_{ni}(B_qf)_{ni}$  centered at its expectation satisfies the condition of Lemma 2.11.17 with  $b = 2(a_{q-1}f/a_qf)2^{-2q}$  and

$a = 2a_{q-1}f$ . Thus, the lemma implies that, for every  $f$  and measurable set  $A$ ,

$$\begin{aligned} \mathbb{E} \left| \sum_{i=1}^{m_n} (\Delta_q f)_{ni} (B_q f)_{ni} - \mathbb{E}(\Delta_q f)_{ni} (B_q f)_{ni} \right| 1_A \\ \lesssim \left( a_{q-1} f \log \frac{1}{\mu(\mathcal{F}_q f)} + \sqrt{\frac{a_{q-1} f}{a_q f}} 2^{-q} \sqrt{\log \frac{1}{\mu(\mathcal{F}_q f)}} \right) (\mathbb{P}(A) + \mu(\mathcal{F}_q f)) \\ \lesssim 2^{-q} \sqrt{\log \frac{1}{\mu(\mathcal{F}_q f)}} (\mathbb{P}(A) + \mu(\mathcal{F}_q f)), \end{aligned}$$

since  $\mu(\mathcal{F}_{q+1} f) \leq \mu(\mathcal{F}_q f)$ . For each  $q$ , let  $\Omega = \cup_j \Omega_{qj}$  be a partition of the underlying probability space such that the maximum  $\|\sum_{i=1}^{m_n} (\Delta_q f)_{ni} (B_q f)_{ni} - \mathbb{E}(\Delta_q f)_{ni} (B_q f)_{ni}\|_{\mathcal{F}}$  is achieved at  $f_{qj}$  on the set  $\Omega_{qj}$ . For every  $q$ , there are as many sets  $\Omega_{qj}$  as there are sets  $\mathcal{F}_{qj}$  in the  $q$ th partition. Then

$$\begin{aligned} \mathbb{E} \left\| \sum_{q>q_0} \sum_{i=1}^{m_n} (\Delta_q f)_{ni} (B_q f)_{ni} - \mathbb{E}(\Delta_q f)_{ni} (B_q f)_{ni} \right\|_{\mathcal{F}} \\ \leq \sum_{q>q_0} \sum_j \mathbb{E} \left| \sum_{i=1}^{m_n} (\Delta_q f_{qj})_{ni} (B_q f_{qj})_{ni} - \mathbb{E}(\Delta_q f)_{ni} (B_q f_{qj})_{ni} \right| 1_{\Omega_{qj}} \\ \leq \sum_j \sup_f \sum_{q>q_0} 2^{-q} \sqrt{\log \frac{1}{\mu(\mathcal{F}_q f)}} (\mathbb{P}(\Omega_{qj}) + \mu(\mathcal{F}_q f_{qj})) \\ \leq 2 \sup_f \sum_{q>q_0} 2^{-q} \sqrt{\log \frac{1}{\mu(\mathcal{F}_q f)}}. \end{aligned}$$

This converges to zero as  $n \rightarrow \infty$  followed by  $q_0 \rightarrow \infty$  and takes care of the second term resulting from the decomposition (2.11.19), apart from the centering. The centering is bounded by

$$\begin{aligned} \sum_{q>q_0} \sum_{i=1}^{m_n} \mathbb{E}(\Delta_q f)_{ni} (B_q f)_{ni} &\leq \sum_{q>q_0} \sum_{i=1}^{m_n} \mathbb{E}(\Delta_q f)_{ni} \{(\Delta_q f)_{ni} > a_q f\} \\ &\leq \sup_f \sup_{t>0} \sum_{q>q_0} \frac{2^{-2q}}{a_q f} \lesssim \sup_f \sum_{q>q_0} 2^{-q} \sqrt{\log \frac{1}{\mu(\mathcal{F}_q f)}}. \end{aligned}$$

This concludes the proof that the second term in the decomposition resulting from (2.11.19) converges to zero as  $n \rightarrow \infty$  followed by  $q_0 \rightarrow \infty$ .

The third term can be handled in a similar manner. The proof of (2.11.20) is complete.

Finally, if the partitions depend on  $n$ , then the discrepancies at the ends of the chains must be analyzed separately. If the  $q_0$ th partition consists of

$N_{q_0}^n$  sets, then the combination of Bernstein's inequality and Lemma 2.2.10 yields

$$\mathbb{E} \sup_{\rho(f,g) < \delta_n} \left| \sum_{i=1}^{m_n} Z_{ni}^o(\pi_{q_0}^n f) - Z_{ni}^o(\pi_{q_0}^n g) \right| \lesssim \log N_{q_0}^n \eta_n + \sqrt{\log N_{q_0}^n} (2^{-q_0} + \delta_n).$$

The entropy condition of Theorem 2.11.9 implies that  $2^{-q_0} \log N_{q_0}^n \rightarrow 0$  as  $n \rightarrow \infty$  followed by  $q_0 \rightarrow \infty$ . Hence the expression in the display converges to zero as  $n \rightarrow \infty$  followed by  $q_0 \rightarrow \infty$ . Combine this with (2.11.20) to see that the sequence  $\sum_{i=1}^{m_n} Z_{ni}^o$  is asymptotically equicontinuous with respect to  $\rho$ . Finally, Theorem 1.5.7 shows that the sequence is asymptotically tight. ■

### 2.11.3 Classes of Functions Changing with $n$

Let  $X_1, X_2, \dots$  be a sequence of independent random elements with common law  $P$  on a measurable space  $(\mathcal{X}, \mathcal{A})$ , and let  $x \mapsto f_{n,t}(x)$  be functions from  $\mathcal{X}$  to  $\mathbb{R}$  indexed by  $n \in \mathbb{N}$  and a fixed, totally bounded semimetric space  $(T, \rho)$ . We wish to derive conditions for the stochastic processes

$$\left\{ n^{-1/2} \sum_{i=1}^n (f_{n,t}(X_i) - P f_{n,t}): t \in T \right\}$$

to converge in distribution in the space  $\ell^\infty(T)$ . These are the empirical processes  $\{\mathbb{G}_n f_{n,t}: t \in T\}$  indexed by classes of functions  $\mathcal{F}_n = \{f_{n,t}: t \in T\}$  changing with  $n$ .

This situation fits in the general set-up of this section upon setting  $Z_{ni}(t) = f_{n,t}(X_i)/\sqrt{n}$ .

Given envelope functions  $F_n$ , assume that

$$(2.11.21) \quad \begin{aligned} P^* F_n^2 &= O(1), \\ P^* F_n^2 \{F_n > \eta \sqrt{n}\} &\rightarrow 0, \quad \text{for every } \eta > 0, \\ \sup_{\rho(s,t) < \delta_n} P(f_{n,s} - f_{n,t})^2 &\rightarrow 0, \quad \text{for every } \delta_n \downarrow 0. \end{aligned}$$

Then the central limit theorem holds under an entropy condition. The following theorems impose a uniform-entropy condition and a bracketing entropy condition, respectively.

**2.11.22 Theorem.** *For each  $n$ , let  $\mathcal{F}_n = \{f_{n,t}: t \in T\}$  be a class of measurable functions indexed by a totally bounded semimetric space  $(T, \rho)$  such that the classes  $\mathcal{F}_{n,\delta} = \{f_{n,s} - f_{n,t}: \rho(s, t) < \delta\}$  and  $\mathcal{F}_{n,\delta}^2$  are  $P$ -measurable for every  $\delta > 0$ . Suppose (2.11.21) holds, as well as*

$$\sup_Q \int_0^{\delta_n} \sqrt{\log N(\varepsilon \|F_n\|_{Q,2}, \mathcal{F}_n, L_2(Q))} d\varepsilon \rightarrow 0, \quad \text{for every } \delta_n \downarrow 0.$$

Then the sequence  $\{\mathbb{G}_n f_{n,t}: t \in T\}$  is asymptotically tight in  $\ell^\infty(T)$  and converges in distribution to a Gaussian process provided the sequence of covariance functions  $Pf_{n,s}f_{n,t} - Pf_{n,s}Pf_{n,t}$  converges pointwise on  $T \times T$ .

**2.11.23 Theorem.** For each  $n$ , let  $\mathcal{F}_n = \{f_{n,t}: t \in T\}$  be a class of measurable functions indexed by a totally bounded semimetric space  $(T, \rho)$ . Suppose (2.11.21) holds, as well as

$$\int_0^{\delta_n} \sqrt{\log N_{[]}(\varepsilon \|F_n\|_{P,2}, \mathcal{F}_n, L_2(P))} d\varepsilon \rightarrow 0, \quad \text{for every } \delta_n \downarrow 0.$$

Then the sequence  $\{\mathbb{G}_n f_{n,t}: t \in T\}$  is asymptotically tight in  $\ell^\infty(T)$  and converges in distribution to a tight Gaussian process provided the sequence of covariance functions  $Pf_{n,s}f_{n,t} - Pf_{n,s}Pf_{n,t}$  converges pointwise on  $T \times T$ .

**Proofs.** The random distance given in Theorem 2.11.1 reduces to

$$d_n^2(s, t) = \frac{1}{n} \sum_{i=1}^n (f_{n,s} - f_{n,t})^2(X_i) = \mathbb{P}_n(f_{n,s} - f_{n,t})^2.$$

It follows that  $N(\varepsilon, T, d_n) = N(\varepsilon, \mathcal{F}_n, L_2(\mathbb{P}_n))$ , for every  $\varepsilon > 0$ . If  $F_n$  is replaced by  $F_n \vee 1$ , then the conditions of the theorem still hold. Hence, assume without loss of generality that  $F_n \geq 1$ . Insert the bound on the covering numbers and next make a change of variables to bound the entropy integral (2.11.2) by

$$\int_0^{\delta_n} \sqrt{\log N(\varepsilon \|F_n\|_{\mathbb{P}_n,2}, \mathcal{F}_n, L_2(\mathbb{P}_n))} d\varepsilon \|F_n\|_{\mathbb{P}_n,2}.$$

This converges to zero in probability for every  $\delta_n \downarrow 0$ . Apply Theorem 2.11.1 to obtain the first theorem.

The second theorem is an immediate consequence of Theorem 2.11.9. ■

**2.11.24 Example.** The uniform-entropy condition of the first theorem is certainly satisfied if, for each  $n$ , the set of functions  $\mathcal{F}_n = \{f_{n,t}: t \in T\}$  is a VC-class with VC-index bounded by some constant independent of  $n$ .

## Problems and Complements

1. The random entropy condition (2.11.2) implies that the sequence  $N(\varepsilon, \mathcal{F}, d_n)$  is bounded in probability for every  $\varepsilon > 0$ .

[Hint: For every  $\delta_n \leq \varepsilon$ ,

$$\mathbb{P}\left(N(\varepsilon, \mathcal{F}, d_n) \geq M_n\right) \leq \mathbb{P}\left(\int_0^{\delta_n} \sqrt{\log N(\varepsilon, \mathcal{F}, d_n)} d\varepsilon \geq \delta_n \sqrt{\log M_n}\right).$$

Given  $M_n \rightarrow \infty$ , choose  $\delta_n \downarrow 0$  such that  $\delta_n \sqrt{\log M_n}$  is bounded away from zero.]

2. Let  $\mathcal{F} = \cup_j \mathcal{F}_{qj}$  be a sequence ( $j \in \mathbb{N}$ ) of nested partitions of an arbitrary set  $\mathcal{F}$ . Take arbitrary  $f_{qj} \in \mathcal{F}_{qj}$  for every partitioning set, and define  $\pi_q f = f_{qj}$  and  $\mathcal{F}_q f = \mathcal{F}_{qj}$  if  $f \in \mathcal{F}_{qj}$ . Let  $\rho(f, g)$  be  $2^{-q_0}$  for the first value  $q_0$  (counting from 1) such that  $f$  and  $g$  do not belong to the same partitioning set at level  $q_0$ . (Set  $\rho(f, g) = 0$  if  $f$  and  $g$  are never separated.) Then  $\rho$  defines a semimetric on  $\mathcal{F}$  inducing open balls

$$B(f, 2^{-q}) = \mathcal{F}_q f, \quad \text{for every } q.$$

Furthermore, if  $G(f) = \sum_q 2^{-q} \xi_{q, \pi_q f}$  for i.i.d. standard normal variables  $\xi_{q, f_{qj}}$ , then

$$\text{var}(G(f) - G(g)) = \frac{8}{3} \rho^2(f, g),$$

for every  $f, g$ .

[Hint: We have  $\rho(f, g) < 2^{-q}$  if and only if  $f$  and  $g$  are in the same partitioning set at level  $q$  if and only if  $\mathcal{F}_q f = \mathcal{F}_q g$ . Of course,  $g \in \mathcal{F}_q g$  for every  $g$ .]

3. In the situation of the preceding problem, let the number  $N_q$  of partitioning sets  $\mathcal{F}_{qj}$  at level  $q$  satisfy

$$\sum_q 2^{-q} \sqrt{\log N_q} < \infty.$$

Then the entropy numbers for the semimetric  $\rho$  satisfy

$$\int_0^\infty \sqrt{\log N(\varepsilon, \mathcal{F}, \rho)} d\varepsilon < \infty.$$

Conclude that the Gaussian process  $G$  defined in the preceding problem has a version with bounded, uniformly  $\rho$ -continuous sample paths. Hence  $\rho$  is a Gaussian semimetric.

[Hint: The convergence of the integral is immediate from the fact that  $N(2^{-q}, \mathcal{F}, \rho) = N_q$ . The continuity follows from Corollary 2.2.8 and Problem 2.2.17. Also see Appendix A.2.3.]

4. Any semimetric  $d$  on an arbitrary set  $\mathcal{F}$  such that  $\int_0^\infty \sqrt{\log N(\varepsilon, \mathcal{F}, d)} d\varepsilon < \infty$  is Gaussian-dominated.

[Hint: Construct a sequence of nested partitions of  $\mathcal{F} = \cup_j \mathcal{F}_{qj}$  in sets of  $\delta$ -diameter at most  $2^{-q}$  for each  $q$ . The number  $N_q$  of sets in the  $q$ th partition

can be chosen such that  $\sum_q 2^{-q} \sqrt{\log N_q} < \infty$ . As in the preceding problems, the semimetric  $\rho$  defined from this sequence of partitions is Gaussian by the preceding problem and has the property that each ball  $B_\rho(f, 2^{-q})$  equals a partitioning set; hence these balls have  $d$ -diameter at most  $2^{-q}$  for every  $q$ . The latter implies that  $d \leq 2\rho$ .]

5. Any semimetric  $d$  on an arbitrary set  $\mathcal{F}$  for which there exists a Borel probability measure  $\mu$  with

$$\limsup_{\delta \downarrow 0} \int_0^\delta \sqrt{\log 1/\mu(B(f, \varepsilon))} d\varepsilon = 0$$

is Gaussian-dominated. Here  $B(f, \varepsilon)$  is the  $d$ -ball of radius  $\varepsilon$ .

[Hint: By Lemma A.2.19, there exists a sequence of nested partitions of  $\mathcal{F} = \cup_i \mathcal{F}_{qi}$  in sets of  $\delta$ -diameter at most  $2^{-q}$  for each  $q$  and a Borel probability measure  $m$  such that

$$\lim_{q_0 \rightarrow \infty} \sup_f \sum_{q > q_0} 2^{-q} \sqrt{\log \frac{1}{m(\mathcal{F}_q f)}} = 0.$$

As in the preceding problems, the semimetric  $\rho$  defined from this sequence of partitions has the property that each ball  $B_\rho(f, 2^{-q})$  equals a partitioning set. Hence

$$\limsup_{\delta \downarrow 0} \int_0^\delta \sqrt{\log 1/m(B_\rho(f, \varepsilon))} d\varepsilon = 0,$$

where  $B_\rho(f, \varepsilon)$  is a ball with respect to  $\rho$ . By Proposition A.2.17, the Gaussian process  $G$  defined in the first problem has a version with bounded, uniformly  $\rho$ -continuous sample paths. The inequalities  $\text{diam } B(f, 2^{-q}) \leq 2^{-q}$  imply that  $d \leq 2\rho$ .]

6. Let  $\mathcal{F}$  be a class of measurable functions such that

$$\int \sqrt{\log N_{[]}(\varepsilon, \mathcal{F}, L_{2,\infty}(P))} d\varepsilon$$

is finite. Then there exists a Gaussian semimetric  $\rho$  on  $\mathcal{F}$  such that

$$\left\| \sup_{f,g \in B(\varepsilon)} |f - g| \right\|_{P,2,\infty} \leq \varepsilon,$$

for every  $\rho$ -ball  $B(\varepsilon) \subset \mathcal{F}$  of radius  $\varepsilon$ .

[Hint: The finiteness of the integral implies the existence of a sequence of partitions of  $\mathcal{F}$ , as in the preceding problems, with the further property

$$\left\| \sup_{f,g \in \mathcal{F}_{qj}} |f - g| \right\|_{P,2,\infty} \leq 2^{-q}.$$

The semimetric  $\rho$  in the preceding problems is Gaussian, and every ball of radius  $2^{-q}$  coincides with a partitioning set. Thus the inequality is valid for  $\varepsilon = 2^{-q}$  and hence up to a constant 2 for every  $\varepsilon$ . Change the metric to remove the 2.]

7. Corollary 2.11.12 is a refinement of Theorem 2.5.6.
8. If the Lindeberg condition on norms in Theorems 2.11.1 and 2.11.9 is replaced by the two assumptions

$$\sup_f \left| \sum_{i=1}^{m_n} EZ_{ni}(f) \{ \|Z_{ni}\|_{\mathcal{F}}^* > \eta \} \right| \rightarrow 0,$$

$$\sum_{i=1}^{m_n} P^*(\|Z_{ni}\|_{\mathcal{F}} > \eta) \rightarrow 0,$$

then the conclusion that the sequence of processes  $\sum_{i=1}^{m_n} (Z_{ni} - EZ_{ni})$  is asymptotically tight is still valid.

## 2.12

# Partial-Sum Processes

The name “Donsker class of functions” was chosen in honor of Donsker’s theorem on weak convergence of the empirical distribution function. A second famous theorem by Donsker concerns the partial-sum process

$$\mathbb{Z}_n(s) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor ns \rfloor} Y_i = \frac{1}{\sqrt{n}} \sum_{i=1}^k Y_i, \quad \frac{k}{n} \leq s < \frac{k+1}{n},$$

for i.i.d. random variables  $Y_1, \dots, Y_n$  with zero mean and variance 1. Donsker essentially proved that the sequence of processes  $\{\mathbb{Z}_n(t): 0 \leq t \leq 1\}$  converges in distribution in the space  $\ell^\infty[0, 1]$  to a standard Brownian motion process [Donsker (1951)].

In this chapter we discuss generalizations of this result in two directions. The first is to replace the  $Y_i$ ’s by processes  $\{f(X_i): f \in \mathcal{F}\}$  and consider convergence in  $\ell^\infty([0, 1] \times \mathcal{F})$ . The second is to consider real variables on a lattice. Partial sums are then obtained by intersecting the lattice with sets in a given collection.

### 2.12.1 The Sequential Empirical Process

Let  $X_1, \dots, X_n$  be i.i.d. random elements with law  $P$  in the measurable space  $(\mathcal{X}, \mathcal{A})$ , and let  $\mathcal{F}$  be a collection of square-integrable, measurable functions  $f: \mathcal{X} \mapsto \mathbb{R}$ . The *sequential empirical process* is defined as

$$\mathbb{Z}_n(s, f) = \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor ns \rfloor} (f(X_i) - Pf) = \sqrt{\frac{\lfloor ns \rfloor}{n}} \mathbb{G}_{\lfloor ns \rfloor}(f),$$

where  $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$  is the empirical process indexed by  $\mathcal{F}$ . The index  $(s, f)$  ranges over  $[0, 1] \times \mathcal{F}$ . The covariance function of  $\mathbb{Z}_n$  is given by

$$\text{cov}(\mathbb{Z}_n(s, f), \mathbb{Z}_n(t, g)) = \frac{\lfloor ns \rfloor \wedge \lfloor nt \rfloor}{n} (Pfg - PfPg).$$

By the multivariate central limit theorem, the marginals of the sequence of processes  $\{\mathbb{Z}_n(s, f) : (s, f) \in [0, 1] \times \mathcal{F}\}$  converge to the marginals of a Gaussian process  $\mathbb{Z}$ . The latter is known as the *Kiefer-Müller process*; it has mean zero and covariance function

$$\text{cov}(\mathbb{Z}(s, f), \mathbb{Z}(t, g)) = (s \wedge t)(Pfg - PfPg).$$

The aim of this section is to show that the weak convergence  $\mathbb{Z}_n \rightsquigarrow \mathbb{Z}$  is uniform with respect to the semimetric of  $\ell^\infty([0, 1] \times \mathcal{F})$  for every Donsker class of functions  $\mathcal{F}$ . More precisely, for every such  $\mathcal{F}$ , the sequence  $\mathbb{Z}_n$  is asymptotically tight and there exists a tight, Borel measurable version of the Kiefer-Müller process  $\mathbb{Z}$ .

In accordance with the general results on Gaussian processes, tightness and measurability of the limit process  $\mathbb{Z}$  are equivalent to the existence of a version of with all sample paths  $(s, f) \mapsto \mathbb{Z}(s, f)$  uniformly bounded and uniformly continuous with respect to the semimetric whose square is given by

$$\mathbb{E}(\mathbb{Z}(s, f) - \mathbb{Z}(t, g))^2 = |s - t| [\rho_P^2(f)1_{s>t} + \rho_P^2(g)1_{s\leq t}] + (s \wedge t) \rho_P^2(f - g).$$

This intrinsic semimetric is somewhat complicated. However, it is up to a constant bounded (on  $[0, 1] \times \mathcal{F}$ ) by the natural semimetric  $|s - t| + \rho_P(f - g)$  if  $\mathcal{F}$  is bounded for  $\rho_P$ ; in particular, if  $\mathcal{F}$  is a Donsker class. By the addendum of Theorem 1.5.7, asymptotic equicontinuity with respect to the intrinsic semimetric is equivalent to asymptotic equicontinuity with respect to the natural semimetric whenever  $[0, 1] \times \mathcal{F}$  is totally bounded under the natural semimetric, particularly if  $\mathcal{F}$  is a Donsker class.

Call  $\mathcal{F}$  a *functional Donsker class* if and only if the sequence  $\mathbb{Z}_n$  converges in distribution in  $\ell^\infty([0, 1] \times \mathcal{F})$  to a tight limit. Then the result can be expressed as follows.

**2.12.1 Theorem.** *A class of measurable functions is functionally Donsker if and only if it is Donsker.*

**Proof.** The empirical process  $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$  can be recovered from the sequential process as  $\mathbb{G}_n(f) = \mathbb{Z}_n(1, f)$ . Since a restriction map is continuous, a class is certainly Donsker if it is functionally Donsker.

For the converse, it suffices to establish the asymptotic equicontinuity of the sequence  $\mathbb{Z}_n$ . With the usual notation,  $\mathcal{F}_\delta = \{f - g : f, g \in \mathcal{F}, \rho_P(f -$

$g) < \delta\}$ , the triangle inequality yields

$$(2.12.2) \quad \begin{aligned} & \sup_{|s-t|+\rho_P(f,g)<\delta} |\mathbb{Z}_n(s, f) - \mathbb{Z}_n(t, g)| \\ & \leq \sup_{|s-t|<\delta} \|\mathbb{Z}_n(s, f) - \mathbb{Z}_n(t, f)\|_{\mathcal{F}} + \sup_{0 \leq t \leq 1} \|\mathbb{Z}_n(t, f)\|_{\mathcal{F}_{\delta}}. \end{aligned}$$

In the second term on the right the parameter  $t$  may be restricted to the points  $k/n$  with  $k$  ranging over  $1, 2, \dots, n$ . Thus, this term is equal to  $\max_{k \leq n} \|\sqrt{k/n} \mathbb{G}_k\|_{\mathcal{F}_{\delta}}$ . By Ottaviani's inequality A.1.1,

$$P^* \left( \max_{k \leq n} \sqrt{k/n} \|\mathbb{G}_k\|_{\mathcal{F}_{\delta}} > 2\varepsilon \right) \leq \frac{P^*(\|\mathbb{G}_n\|_{\mathcal{F}_{\delta}} > \varepsilon)}{1 - \max_{k \leq n} P^*(\sqrt{k/n} \|\mathbb{G}_k\|_{\mathcal{F}_{\delta}} > \varepsilon)}.$$

If the class  $\mathcal{F}$  is Donsker, then the sequence  $\mathbb{G}_n$  is asymptotically equicontinuous. Thus, the numerator converges to zero as  $n \rightarrow \infty$  followed by  $\delta \downarrow 0$ . The terms of the maximum in the denominator indexed by  $k \leq n_0$  can be controlled with the help of the inequality  $\sqrt{k} \|\mathbb{G}_k\|_{\mathcal{F}_{\delta}} \leq 2 \sum_{i=1}^{n_0} F(X_i) + 2n_0 P^* F$  for an envelope function  $F$ . For sufficiently large  $n_0$ , the terms indexed by  $k > n_0$  are bounded away from 1 by the asymptotic equicontinuity of  $\mathbb{G}_n$ . Conclude that the denominator is bounded away from zero. Hence the second term on the right in (2.12.2) converges to zero in probability as  $n \rightarrow \infty$  followed by  $\delta \downarrow 0$ .

To prove the same for the first term on the right in (2.12.2), it suffices to prove convergence to zero of

$$P^* \left( \max_{0 \leq j\delta \leq 1} \sup_{j\delta \leq s \leq (j+1)\delta} \|\mathbb{Z}_n(s, f) - \mathbb{Z}_n(j\delta, f)\|_{\mathcal{F}} > 2\varepsilon \right).$$

By the stationarity of the increments of  $\mathbb{Z}_n$  in  $s$ , the at most  $\lceil 1/\delta \rceil$  terms in the maximum are identically distributed. Thus, the probability can be bounded by

$$\left\lceil \frac{1}{\delta} \right\rceil P^* \left( \sup_{0 \leq s \leq \delta} \|\mathbb{Z}_n(s, f)\|_{\mathcal{F}} > 2\varepsilon \right).$$

Again replace the continuous index  $s$  by a discrete one and conclude from Ottaviani's inequality that the preceding display is not smaller than

$$\left\lceil \frac{1}{\delta} \right\rceil P^* \left( \max_{k \leq n\delta} \sqrt{k/n} \|\mathbb{G}_k\|_{\mathcal{F}} > 2\varepsilon \right) \leq \frac{\lceil 1/\delta \rceil P^* \left( \sqrt{\lceil n\delta \rceil/n} \|\mathbb{G}_{\lfloor n\delta \rfloor}\|_{\mathcal{F}} > \varepsilon \right)}{1 - \max_{k \leq n\delta} P^*(\sqrt{k/n} \|\mathbb{G}_k\|_{\mathcal{F}} > \varepsilon)}.$$

By the portmanteau theorem, the  $\limsup$  as  $n \rightarrow \infty$  of the probability in the numerator is bounded by  $P(\|\mathbb{G}\|_{\mathcal{F}} \geq \varepsilon/\delta^{1/2})$ . Since the norm  $\|\mathbb{G}\|_{\mathcal{F}}$  of a Brownian bridge has moments of all orders (cf. Proposition A.2.3), the latter probability converges to zero faster than any power of  $\delta$  as  $\delta \downarrow 0$ . Conclude that the numerator converges to zero as  $n \rightarrow \infty$  followed by  $\delta \downarrow 0$ . By a similar argument as before, but now also using the fact that  $P(\|\mathbb{G}\|_{\mathcal{F}} > \varepsilon) < 1$  for every  $\varepsilon > 0$  (cf. Problem A.2.5), the denominator remains bounded away from zero. ■

### 2.12.2 Partial-Sum Processes on Lattices

For each positive integer  $n$ , let  $Y_{n1}, \dots, Y_{nm_n}$  be independent, real-valued random variables, and let  $Q_{n1}, \dots, Q_{nm_n}$  be deterministic probability measures on a measurable space  $(\mathcal{X}, \mathcal{A})$ . Given a collection  $\mathcal{C}$  of measurable subsets of  $\mathcal{X}$ , consider the stochastic process

$$\mathbb{S}_n(C) = \sum_{i=1}^{m_n} Y_{ni} Q_{ni}(C).$$

Thus,  $\mathbb{S}_n$  is a randomly weighted sum of the measures  $Q_{ni}$ . Both the classical partial-sum process and its smoothed version are of this type.

**2.12.3 Example.** Let  $\mathcal{X} = [0, 1]$  be the unit interval in the real line and  $\mathcal{C}$  the collection of cells  $[0, s]$  with  $0 \leq s \leq 1$ . If  $Q_{ni}$  is the Dirac measure at the point  $i/n$  (for  $1 \leq i \leq n$ ), then

$$\mathbb{S}_n([0, s]) = \sum_{i/n \leq s} Y_{ni} = \sum_{i=1}^k Y_{ni}, \quad \text{if } \frac{k}{n} \leq s < \frac{k+1}{n}.$$

The choice  $Y_{ni} = Y_i/\sqrt{n}$  for an i.i.d. sequence  $Y_1, Y_2, \dots$  gives the classical partial-sum process.

**2.12.4 Example.** Let  $\mathcal{X} = [0, 1]$  be the unit interval in the real line and  $\mathcal{C}$  the collection of cells  $[0, s]$  with  $0 \leq s \leq 1$ . If  $Q_{ni}$  is the uniform measure on the interval  $[i/n, (i+1)/n)$ , then

$$\mathbb{S}_n([0, s]) = \sum_{i=1}^k Y_{ni} + \frac{s - k/n}{1/n} Y_{n,k+1}, \quad \text{if } \frac{k}{n} \leq s < \frac{k+1}{n}.$$

This is a linear interpolation of the partial-sum process in the previous example.

**2.12.5 Example.** If each  $Q_{ni}$  equals the Dirac measure at some point  $x_{ni} \in \mathcal{X}$ , then the general process reduces to

$$\mathbb{S}_n(C) = \sum_{i: x_{ni} \in C} Y_{ni}.$$

This may be visualized as random weights  $Y_{ni}$  being located at fixed points  $x_{ni}$ . Each set  $C$  is charged with the sum of the weights that it carries.

The process  $\mathbb{S}_n = \sum Y_{ni} Q_{ni}$  is the sum of the independent processes  $Z_{ni} = Y_{ni} Q_{ni}$ . Since

$$(Z_{ni}(C) - Z_{ni}(D))^2 \leq Y_{ni}^2 Q_{ni}(C \Delta D),$$

the processes  $Z_{ni}$  are measurelike with respect to the measures  $\mu_{ni} = Y_{ni}^2 Q_{ni}$ . Thus for a collection  $\mathcal{C}$  that satisfies the uniform-entropy condition, Theorem 2.11.1 combined with Lemma 2.11.6 readily yields a central limit theorem for the sequence  $\mathbb{S}_n$ . This is true in particular for VC-classes of sets.

**2.12.6 Theorem.** For each  $n$ , let  $Q_{n1}, \dots, Q_{nm_n}$  be deterministic probability measures on some measurable space, and let  $Y_{n1}, \dots, Y_{nm_n}$  be independent, real-valued random variables with mean zero, satisfying

$$\begin{aligned} \sum_{i=1}^{m_n} EY_{ni}^2 &= O(1), \\ \sum_{i=1}^{m_n} EY_{ni}^2 \{ |Y_{ni}| > \eta \} &\rightarrow 0, \quad \text{for every } \eta > 0. \end{aligned}$$

Let  $\mathcal{C}$  be a class of measurable sets that satisfies the uniform-entropy condition, and assume that for some probability measure  $Q$ ,

$$(2.12.7) \quad \sup_{Q(C \Delta D) < \delta_n} \sum_{i=1}^{m_n} EY_{ni}^2 (Q_{ni}(C) - Q_{ni}(D))^2 \rightarrow 0, \quad \text{for every } \delta_n \downarrow 0.$$

Finally, suppose that the covariance function  $E\mathbb{S}_n(C)\mathbb{S}_n(D)$  converges pointwise on  $\mathcal{C} \times \mathcal{C}$ . Then the sequence  $\mathbb{S}_n = \sum_{i=1}^{m_n} Y_{ni} Q_{ni}$  converges weakly in  $\ell^\infty(\mathcal{C})$  to a tight Gaussian process with uniformly continuous sample paths with respect to the semimetric  $Q(C \Delta D)$ .

**Proof.** Apply Theorem 2.11.1 with  $Z_{ni} = Y_{ni} Q_{ni}$  and  $\rho$  the  $L_1(Q)$ -semimetric. Then  $\|Z_{ni}\|_{\mathcal{F}}$  can be taken equal to  $|Y_{ni}|$ .

The entropy condition (2.11.2) is satisfied in view of Lemma 2.11.6. Since  $\mathcal{C}$  satisfies the uniform-entropy condition, it is totally bounded under the  $L_1(Q)$ -semimetric (for any probability measure  $Q$ ).

The measurability conditions listed before Theorem 2.11.1 are satisfied, because the suprema can be replaced by countable suprema. Indeed, since  $\mathcal{C}$  satisfies the uniform-entropy condition, it is totally bounded and hence separable in  $L_1(\sum_{i=1}^{m_n} Q_{ni} + Q)$ . Thus, there exists a countable subcollection that contains, for every  $C \in \mathcal{C}$ , a sequence  $C_k$  with  $(\sum_{i=1}^{m_n} Q_{ni} + Q)(C_k \Delta C) \rightarrow 0$  as  $k \rightarrow \infty$  (and  $n$  fixed). For this sequence,  $\mathbb{S}_n(C) = \lim_{k \rightarrow \infty} \mathbb{S}_n(C_k)$ . Similar arguments applied to the multiplier processes show that the suprema that are assumed measurable in Theorem 2.11.1 can be replaced by countable suprema.

The theorem follows from Theorem 2.11.1. ■

The covariance function of the process  $\mathbb{S}_n$  is given by

$$E\mathbb{S}_n(C)\mathbb{S}_n(D) = \sum_{i=1}^{m_n} EY_{ni}^2 Q_{ni}(C)Q_{ni}(D).$$

In the special case that each  $Q_{ni}$  is a Dirac measure, we have the equality  $Q_{ni}(C)Q_{ni}(D) = Q_{ni}(C \cap D)$  for every pair of sets, and the covariance function reduces to  $Q_n(C \cap D)$  for the measure

$$Q_n = \sum_{i=1}^{m_n} EY_{ni}^2 Q_{ni}.$$

Then pointwise convergence of  $Q_n$  to a limit on the collection of sets  $\{C \cap D : C, D \in \mathcal{C}\}$  verifies convergence of the sequence of covariance functions.

For any measures  $Q_{ni}$ , the sum on the left side of (2.12.7) is bounded by  $Q_n(C \Delta D)$ . Thus, consideration of the measures  $Q_n$  can also be helpful for verification of the other, more involved, condition of the theorem. A simple sufficient condition for (2.12.7) is that  $Q_n$  converges uniformly to  $Q$  on the collection of sets  $\{C \Delta D : C, D \in \mathcal{C}\}$ . This puts further restrictions on the class  $\mathcal{C}$ . A simple further sufficient condition is that the sequence  $Q_n$  converges weakly to  $Q$  and, for every  $\varepsilon > 0$ , the collection  $\mathcal{C}$  can be covered by finitely many brackets  $[C_l, C^u]$  of size  $Q(C^u - C_l) < \varepsilon$  and having  $P$ -continuity sets  $C_l$  and  $C^u$  as boundaries (Problem 2.12.1).

**2.12.8 Example.** Let  $\mathcal{X} = [0, 1]^d$  be the unit cube and  $\mathcal{C}$  the collection of all cells  $[0, t]$  with  $0 \leq t \leq 1$ . Let the collection of measures  $Q_{ni}$  be the  $n^d$  Dirac measures located at nodes of the regular grid consisting of the  $n^d$  points  $\{1/n, 2/n, \dots, 1\}^d$ . This gives a higher-dimensional version of the classical partial-sum process. If  $Y_{ni} = Y_i/n^{d/2}$  for an i.i.d. mean-zero sequence  $Y_1, Y_2, \dots$  with unit variance, then the measure  $Q_n$  reduces to

$$Q_n = \frac{1}{n^d} \sum_{i=1}^{n^d} Q_{ni}.$$

This is the discrete uniform measure on the grid points. The sequence  $Q_n$  converges weakly to Lebesgue measure, and the collection of cells  $[0, t]$  satisfies the bracketing condition for Lebesgue measure. Thus, (2.12.7) is satisfied for  $Q$  equal to Lebesgue measure, and the sequence of covariance functions converges. The limiting Gaussian process has covariance function  $\text{ES}(C)\text{S}(D) = \lambda(C \cap D)$  for Lebesgue measure  $\lambda$  (a standard *Brownian sheet*).

**2.12.9 Example.** Let  $\mathcal{X} = [0, 1]^d$  be the unit cube and  $\mathcal{C}$  a VC-class. Partition  $[0, 1]^d$  into  $n^d$  cubes of volume  $n^{-d}$ , and let the collection  $Q_{ni}$  be the set of uniform probability measures on the cubes. This gives a smoothed version of the partial-sum process in higher dimensions. If each of the  $Y_{ni}$  has variance  $n^{-d/2}$ , then the measure  $Q_n$  reduces to Lebesgue measure. Thus (2.12.7) is satisfied for  $Q$  equal to Lebesgue measure.

If the cubes in the partition are denoted  $C_{ni}$ , then the covariance function can be written

$$\frac{1}{n^d} \sum_{i=1}^{n^d} \frac{\lambda(C_{ni} \cap C)}{\lambda(C_{ni})} \frac{\lambda(C_{ni} \cap D)}{\lambda(C_{ni})} = \text{EE}(1_C(U) | \mathcal{A}_n) \text{E}(1_D(U) | \mathcal{A}_n),$$

where  $U$  is the identity map on  $[0, 1]^d$  equipped with Lebesgue measure and  $\mathcal{A}_n$  is the  $\sigma$ -field generated by the partition. For  $n \rightarrow \infty$ , the sequence  $\text{E}(1_C(U) | \mathcal{A}_n)$  converges in probability to  $1_C(U)$  for every set  $C$ . (This

follows from a martingale convergence theorem or a direct argument: it is clear that  $E(h|\mathcal{A}_n) \rightarrow h$  in quadratic mean for every continuous function  $h$ ; each indicator function  $1_C$  can be approximated arbitrarily closely by a continuous function.) Thus, the sequence of covariance functions converges to the limit  $E1_C(U)1_D(U) = \lambda(C \cap D)$ .

## Problems and Complements

- Suppose the sequence of Borel measures  $Q_n$  converges weakly to a probability measure  $Q$ . Let  $\mathcal{C}$  be a collection of Borel sets that for every  $\varepsilon > 0$  can be covered with finitely many brackets  $[C_l, C^u]$  of size  $Q(C^u - C_l) < \varepsilon$  and consisting of  $Q$ -continuity sets  $C_l$  and  $C^u$ . Then  $Q_n \rightarrow Q$  uniformly in  $C \in \mathcal{C} \Delta \mathcal{C}$  and  $C \in \mathcal{C} \cap \mathcal{C}$ .

**[Hint:** The following problem is helpful. Compare the proof of Theorem 2.4.1.]

- If  $\mathcal{C}$  has the bracketing property as in the preceding problem, then so do the collections  $\mathcal{C} \Delta \mathcal{C}$  and  $\mathcal{C} \cap \mathcal{C}$ .

- If  $Z_{ni}(s, f) = f(X_i)1\{i/n \leq t\}$ , then the sequential empirical process can be written  $\mathbb{Z}_n = n^{-1/2}(\sum_{i=1}^n Z_{ni} - EZ_{ni})$ . Next the results of Chapter 2.11 can be used to deduce that the sequence  $\mathbb{Z}_n$  converges in distribution. Compare the result with the present chapter.

- (Time reversal) If  $\mathbb{Z}$  is a Kiefer-Müller process indexed by  $[1, \infty) \times \mathcal{F}$ , then the process  $(t, f) \mapsto t\mathbb{Z}(1/t, f)$  is a Kiefer-Müller process indexed by  $(0, 1] \times \mathcal{F}$ .

- If  $\mathbb{Z}_n$  is the sequential empirical process, then the sequence of processes  $(t, f) \mapsto \mathbb{Z}_n(t, f)/(t \vee 1)$  converges weakly in  $\ell^\infty(\mathbb{R}^+ \times \mathcal{F})$ .

**[Hint:** The process equals  $\sqrt{n}([nt]/n(t \vee 1))(\mathbb{P}_{[nt]} - P)$ . Use the reverse martingale property of  $\|\mathbb{P}_n - P\|_{\mathcal{F}}^*$  given in Lemma 2.4.5.]

- Combination of the preceding problems yields that  $\sup_{m \geq n} \sqrt{n}\|\mathbb{P}_m - P\|_{\mathcal{F}}$  converges weakly to  $\sup_{0 \leq t \leq 1} \|\mathbb{Z}(t, f)\|_{\mathcal{F}}$ . This generalizes results of Müller (1968). See Hjort and Fenstad (1992), Section 4, for applications of this result.

- A standard Brownian sheet  $\mathbb{S}$  indexed by  $\mathbb{R}^+ \times [0, 1]$  is a zero-mean Gaussian process with continuous sample paths and covariance function

$$\text{cov}(\mathbb{S}(s, u), \mathbb{S}(t, v)) = (s \wedge t)(u \wedge v).$$

Then the process  $\mathbb{S}(s, u) - u\mathbb{S}(s, 1)$  is a classical Kiefer-Müller process (a Kiefer-Müller process for  $\mathcal{X} = [0, 1]$  with uniform measure and  $\mathcal{F}$  the collection of indicators of cells  $[0, u]$ ).

- With  $\mathbb{S}$  a standard Brownian sheet, consider the processes  $\mathbb{S}(s, u) - su\mathbb{S}(1, 1)$  and  $\mathbb{S}(s, u) - u\mathbb{S}(s, 1) - v\mathbb{S}(1, u) + us\mathbb{S}(1, 1)$ . On which sides of the unit square are these processes zero almost surely? (The second process appears in connection with tests for independence; see Chapter 3.8 and Appendix A.2.2.)

## 2.13

# Other Donsker Classes

In this section we consider some Donsker classes of interest, that do not fit well in the framework of the preceding chapters.

### 2.13.1 Sequences

A countable class of functions may be shown to be Donsker by any of the criteria we have discussed so far. A very simple sufficient condition is that the sequence converges to zero sufficiently fast.

**2.13.1 Theorem (Sequences).** *Any sequence  $\{f_i\}$  of square-integrable, measurable functions with the property  $\sum_{i=1}^{\infty} P(f_i - Pf_i)^2 < \infty$  is  $P$ -Donsker.*

**Proof.** For a fixed natural number  $m$ , define a partition  $\{f_i\} = \cup_{i=1}^{m+1} \mathcal{F}_i$  by letting  $\mathcal{F}_i$  consist of the single function  $f_i$  for  $i \leq m$  and  $\mathcal{F}_{m+1} = \{f_{m+1}, f_{m+2}, \dots\}$ . Since the variation over the first  $m$  sets in the partition is zero,

$$\begin{aligned} P\left(\sup_i \sup_{f,g \in \mathcal{F}_i} |\mathbb{G}_n(f - g)| > \varepsilon\right) &\leq P\left(\sup_{f \in \mathcal{F}_{m+1}} |\mathbb{G}_n f| > \frac{\varepsilon}{2}\right) \\ &\leq \frac{4}{\varepsilon^2} \sum_{i=m+1}^{\infty} P(f_i - Pf_i)^2 < \infty, \end{aligned}$$

by Chebyshev's inequality. For sufficiently large  $m$ , this is smaller than any prescribed  $\eta > 0$ . The result follows from Theorem 1.5.6. ■

### 2.13.2 Elliptical Classes

The preceding theorem may be combined with Theorem 2.10.3 to conclude that the class

$$\left\{ \sum_{i=1}^{\infty} c_i f_i : \sum |c_i| \leq 1, \text{ and the series converges pointwise} \right\}$$

is Donsker for any given sequence  $f_i$  with  $\sum_{i=1}^{\infty} Pf_i^2 < \infty$ . Under the additional condition that the functions  $f_i$  are orthogonal, this can be improved.

**2.13.2 Theorem (Elliptical classes).** *Let  $\{f_i\}$  be a sequence of measurable functions such that  $Pf_i f_j = 0$  for every  $i \neq j$  and  $\sum_{i=1}^{\infty} Pf_i^2 < \infty$ . Then the class of all pointwise converging series  $\sum_{i=1}^{\infty} c_i f_i$ , such that  $\sum_{i=1}^{\infty} c_i^2 \leq 1$ , is  $P$ -Donsker.*

**Proof.** By the condition on  $c$ , each of the series  $\sum_{i=1}^{\infty} c_i f_i$  converges pointwise as well as in  $L_2(P)$ . The class  $\mathcal{F}$  of all these series is totally bounded in  $L_2(P)$ , because it is bounded and can be approximated arbitrarily closely by a finite-dimensional set, since

$$P \left( \sum_{i>m} c_i f_i \right)^2 = \sum_{i>m} c_i^2 Pf_i^2 \leq \max_{i>m} Pf_i^2 \rightarrow 0, \quad m \rightarrow \infty.$$

It suffices to show that the sequence of empirical processes  $\mathbb{G}_n$  indexed by  $\mathcal{F}$  is asymptotically equicontinuous with respect to the  $L_2(P)$ -seminorm. For  $f = \sum c_i f_i$ ,  $g = \sum d_i f_i$ , and any natural number  $k$ ,

$$\begin{aligned} |\mathbb{G}_n(f) - \mathbb{G}_n(g)|^2 &= \left| \sum_{i=1}^{\infty} (c_i - d_i) \mathbb{G}_n(f_i) \right|^2 \\ &\leq 2 \sum_{i=1}^k (c_i - d_i)^2 Pf_i^2 \sum_{i=1}^k \frac{\mathbb{G}_n^2(f_i)}{Pf_i^2} + 2 \sum_{i=k+1}^{\infty} (c_i - d_i)^2 \sum_{i=k+1}^{\infty} \mathbb{G}_n^2(f_i). \end{aligned}$$

In view of the assumption that  $\|c - d\|_2 \leq \|c\|_2 + \|d\|_2 \leq 2$ , this expression is bounded by

$$2\|f - g\|_{P,2}^2 \sum_{i=1}^k \frac{\mathbb{G}_n^2(f_i)}{Pf_i^2} + 8 \sum_{i=k+1}^{\infty} \mathbb{G}_n^2(f_i).$$

Take the supremum over all pairs of series  $f$  and  $g$  with  $\|f - g\|_{P,2} < \delta$ . Since  $E\mathbb{G}_n^2(f_i) \leq Pf_i^2$ , the expectation of this supremum is bounded by  $2\delta^2 k + 8 \sum_{i=k+1}^{\infty} Pf_i^2$ . This expression can be made arbitrarily small by first choosing  $k$  large and next  $\delta$  small. ■

In terms of an orthonormal sequence  $\{\psi_i\}$  in  $\mathcal{L}_2(P)$ , the preceding theorem can be stated in the following manner. For a given sequence of numbers  $b_i$ , the *elliptical class*

$$\mathcal{F} = \left\{ \sum_{i=1}^{\infty} c_i \psi_i : \sum \frac{c_i^2}{b_i^2} \leq 1 \text{ and the series converges pointwise} \right\}$$

is  $P$ -Donsker if  $\sum_{i=1}^{\infty} b_i^2 < \infty$ . The latter condition is also known to be necessary. In fact, it is already necessary for the class to be pre-Gaussian.<sup>†</sup> The pointwise convergence of the series forming  $\mathcal{F}$  appears not to be automatic. In view of the Cauchy-Schwarz inequality, a simple sufficient condition for absolute convergence at  $x$  is that  $\sum \psi_i(x)^2 b_i^2 < \infty$ .

Elliptical classes are of some interest, because some well-known test statistics, including the Cramér-von Mises and the Anderson-Darling statistic, arise as the (generalized) Kolmogorov-Smirnov statistic  $\|\mathbb{G}_n\|_{\mathcal{F}}$  indexed by an elliptical class. This is shown in the examples ahead. The Kolmogorov-Smirnov statistics corresponding to an elliptical class can be represented as a series of uncorrelated variables. Indeed, for  $\mathcal{F}$  equal to the class of functions in the previous display,

$$\|\mathbb{G}_n\|_{\mathcal{F}}^2 = \sup_{\mathcal{F}} \left| \sum_{i=1}^{\infty} c_i \mathbb{G}_n(\psi_i) \right|^2 = \sum_{i=1}^{\infty} b_i^2 \mathbb{G}_n^2(\psi_i),$$

in view of the Cauchy-Schwarz inequality. If the  $\psi_i$  are uncorrelated and of unit variance, then the sequence  $\|\mathbb{G}_n\|_{\mathcal{F}}^2$  is asymptotically distributed as  $\sum b_i^2 Z_i^2$  for an i.i.d. sequence  $Z_1, Z_2, \dots$  of standard normal variables. This follows from the series representation for  $\|\mathbb{G}_n\|_{\mathcal{F}}^2$ ; it can also be obtained from the fact that  $\|\mathbb{G}_n\|_{\mathcal{F}}^2$  converges in distribution to the square norm of a Brownian bridge  $\mathbb{G}$  and a series representation for  $\|\mathbb{G}\|_{\mathcal{F}}^2$ , which can be obtained in an analogous manner.

**2.13.3 Example (Cramér-von Mises).** Let  $\mathbb{P}_n$  be the empirical distribution of an i.i.d. sample of size  $n$  from the uniform distribution on the unit interval  $[0, 1] \subset \mathbb{R}$ . Write  $\mathbb{G}_n(t) = \sqrt{n}(\mathbb{P}_n - P)[0, t]$  for the classical empirical process ( $t \in [0, 1]$ ). Let  $\mathcal{F}$  be the class of functions

$$\mathcal{F} = \left\{ \sum_{j=1}^{\infty} c_j \sqrt{2} \cos \pi j t : \sum c_j^2 \pi^2 j^2 \leq 1 \right\}.$$

Since the cosines are bounded, the series defining the elements of  $\mathcal{F}$  are uniformly convergent in this case. The classical Cramér-von Mises statistic equals the square of the Kolmogorov-Smirnov statistic over the elliptical class  $\mathcal{F}$ :

$$\int_0^1 \mathbb{G}_n^2(t) dt = \|\mathbb{G}_n\|_{\mathcal{F}}^2.$$

---

<sup>†</sup> Dudley (1967b), Proposition 6.3.

To see this, note that the Cramér-von Mises statistic is the square of the  $L_2[0, 1]$ -norm of the function  $t \mapsto \mathbb{G}_n(t)$ . Since the functions  $\{\sqrt{2} \sin \pi jt: j = 1, 2, \dots\}$  form an orthonormal base of  $L_2[0, 1]$ , Parseval's formula yields

$$\int_0^1 \mathbb{G}_n^2(t) dt = \sum_{j=1}^{\infty} \left[ \int_0^1 \mathbb{G}_n(t) \sqrt{2} \sin(\pi jt) dt \right]^2 = \sum_{j=1}^{\infty} \frac{1}{\pi^2 j^2} \mathbb{G}_n^2(\sqrt{2} \cos \pi jt).$$

(Note that  $-\int f' d\mathbb{G}_n = \int f d\mathbb{G}_n = \mathbb{G}_n f$  by the definitions of  $t \mapsto \mathbb{G}_n(t)$  and  $\mathbb{G}_n f$ .) The functions  $\{\sqrt{2} \cos \pi jt: j = 1, 2, \dots\}$  form an orthonormal system in  $L_2[0, 1]$ , so that the result follows from the general series representation of  $\|\mathbb{G}_n\|_{\mathcal{F}}^2$  for elliptical classes.

**2.13.4 Example (Watson).** In the situation of the preceding example, let  $\mathcal{F}$  be the elliptical class

$$\left\{ \sum_{j=1}^{\infty} (c_{2j-1} \sqrt{2} \cos 2\pi jt + c_{2j} \sqrt{2} \sin 2\pi jt) : \sum (c_{2j-1}^2 + c_{2j}^2) \pi^2 j^2 \leq 1 \right\}.$$

The series are again uniformly convergent. The Watson statistic equals the square of the Kolmogorov-Smirnov statistic over the elliptical class  $\mathcal{F}$ :

$$\int_0^1 [\mathbb{G}_n(t) - \int \mathbb{G}_n(s) ds]^2 dt = \|\mathbb{G}_n\|_{\mathcal{F}}^2.$$

This can be seen by a similar argument. The Watson statistic is the square of the  $L_2[0, 1]$ -norm of the projection of the function  $t \mapsto \mathbb{G}_n(t)$  on the mean-zero functions. The functions  $\{\sqrt{2} \sin 2\pi jt, \sqrt{2} \cos 2\pi jt: j = 1, 2, \dots\}$  form an orthonormal base of the mean-zero functions. Application of Parseval's formula followed by partial integration as in the preceding example yields the result.

**2.13.5 Example (Anderson-Darling).** In the situation of the preceding example, let  $\mathcal{F}$  be the elliptical class

$$\left\{ \sum_{j=1}^{\infty} c_j \sqrt{2} p_j(2t-1) : \sum c_j^2 j(j+1) \leq 1 \text{ and pointwise convergence} \right\},$$

where the functions  $p_0(u) = (1/2)\sqrt{2}$ ,  $p_1(u) = (1/2)\sqrt{6}u$ ,  $p_2(u) = (1/4)\sqrt{10}(3u^2 - 1)$ ,  $p_3(u) = (1/4)\sqrt{14}(5u^3 - 3u)$ , ... are the orthonormalized Legendre polynomials in  $L_2[-1, 1]$ . (This is the orthonormal system obtained by applying the Gram-Schmidt procedure to the functions  $1, u, u^2, \dots$ .) The Anderson-Darling statistic equals the square of the Kolmogorov-Smirnov statistic over the elliptical class  $\mathcal{F}$ :

$$\int_0^1 \frac{\mathbb{G}_n^2(t)}{t(1-t)} dt = \|\mathbb{G}_n\|_{\mathcal{F}}^2.$$

The argument is slightly more involved than in the preceding examples, but it is based on the same idea. The normalized Legendre polynomials satisfy the differential equations  $p_j''(u)(1-u^2) - 2up_j'(u) = -j(j+1)p_j(u)$  for  $u \in [-1, 1]$  (Problem 2.13.1). By a change of variables and partial integration, it can be deduced that

$$\begin{aligned} 2 \int_0^1 p_i'(2t-1)p_j'(2t-1)t(1-t) dt \\ = -\frac{1}{4} \int_{-1}^1 p_i(u) [p_j''(u)(1-u^2) - 2up_j'(u)] du \\ = \frac{1}{4} j(j+1)\delta_{ij}. \end{aligned}$$

It follows that the functions  $2\sqrt{2}p_i'(2t-1)\sqrt{t(1-t)}/\sqrt{j(j+1)}$  with  $j$  ranging over  $\{1, 2, \dots\}$  form an orthonormal base of  $L_2[0, 1]$ . By Parseval's formula, the Anderson-Darling statistic equals

$$\sum_{j=1}^{\infty} [\int_0^1 \mathbb{G}_n(t)p_i'(2t-1) dt]^2 \frac{8}{j(j+1)} = \sum_{j=1}^{\infty} \frac{2}{j(j+1)} \mathbb{G}_n^2(p_j(2t-1)).$$

This equals  $\|\mathbb{G}_n\|_{\mathcal{F}}^2$  by the general series representation of  $\|\mathbb{G}_n\|_{\mathcal{F}}^2$  for elliptical classes.

### 2.13.3 Classes of Sets

The preceding chapters give a wide variety of relatively easy-to-apply sufficient conditions for the Donsker property. They have little to say about necessary conditions beyond the most abstract level or in concrete cases. For classes of sets, simple necessary conditions, up to measurability, are known.

Recall from Chapter 2.6 that for a given collection  $\mathcal{C}$  of sets and given points,  $\Delta_n(\mathcal{C}, X_1, \dots, X_n)$  denotes the number of subsets of  $\{X_1, \dots, X_n\}$  picked out by  $\mathcal{C}$ . Also define  $K_n(\mathcal{C}, X_1, \dots, X_n)$  to be the cardinality of a maximal subset of  $\{X_1, \dots, X_n\}$  shattered by  $\mathcal{C}$ :

$$K_n(\mathcal{C}, X_1, \dots, X_n) = \max \left\{ \#A : \Delta_n(\mathcal{C}, A) = 2^{\#A} \right\},$$

where  $A$  ranges over all possible subsets of  $\{X_1, \dots, X_n\}$ .

**2.13.6 Theorem.** *For every pointwise-separable collection of measurable sets, the following statements are equivalent:*

- (i)  $\log \Delta_n(\mathcal{C}, X_1, \dots, X_n) = o_P^*(\sqrt{n})$  and  $\mathcal{C}$  is  $P$ -pre-Gaussian;
- (ii)  $K_n(\mathcal{C}, X_1, \dots, X_n) = o_P^*(\sqrt{n})$  and  $\mathcal{C}$  is  $P$ -pre-Gaussian;
- (iii)  $\mathcal{C}$  is  $P$ -Donsker;
- (iv)  $\log N(\varepsilon n^{-1/2}, \mathcal{C}, L_1(\mathbb{P}_n)) = o_P^*(\sqrt{n})$  for every  $\varepsilon > 0$  and  $\mathcal{C}$  is  $P$ -pre-Gaussian.

**Proof.** Giné and Zinn (1984) and Talagrand (1988). ■

## Problems and Complements

1. The Legendre polynomials on  $[-1, 1]$  satisfy the differential equation

$$p_j''(u)(1 - u^2) - 2up_j'(u) = -j(j + 1)p_j(u).$$

[**Hint:** It suffices to prove the same relationship for the polynomials  $p_j(u)$  defined as  $u^j$  minus the projection of  $u^j$  on the linear space spanned by  $1, u, \dots, u^{j-1}$ . The left side of the differential equation is a polynomial of degree  $j$  with leading term  $-j(j + 1)u^j$ . By definition, the right side has the same property. It suffices to show that the left side is orthogonal to the functions  $1, u, \dots, u^{j-1}$ . By partial integration,

$$\begin{aligned} \int p_j'(u) 2u u^k du &= 2p_j(1) \pm 2p_j(-1) - \int p_j(u) 2(k+1)u^k du \\ &= 2p_j(1) \pm 2p_j(-1), \end{aligned}$$

where  $\pm$  is  $+$  if  $k < j$  is even and  $-$  otherwise. By using partial integration twice, this can be seen to be equal to  $\int p_j''(u)(1 - u^2) u^k du$ .]

2. The representation of the Watson statistic as the Kolmogorov-Smirnov statistic of an elliptic class is not unique. Consider the class

$$\mathcal{F} = \left\{ \sum_{j=1}^{\infty} c_j \sqrt{2} \sin \pi j t : \sum c_j^2 \pi^2 j^2 \leq 1 \right\}.$$

Then the Watson statistic equals  $\|\mathbb{G}_n\|_{\mathcal{F}}^2$ . (The terms in the accompanying series representation are not asymptotically independent; hence this representation is of less interest.)

[**Hint:** The functions  $\{\sqrt{2} \cos \pi j t : j = 1, 2, \dots\}$  form an orthonormal base of the mean-zero functions in  $L_2[0, 1]$ .]

## 2.14

# Tail Bounds

In this chapter we derive moment and tail bounds for the supremum  $\|\mathbb{G}_n\|_{\mathcal{F}}$  of the empirical process. Throughout this chapter,  $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$  denotes the empirical process of an i.i.d. sample  $X_1, \dots, X_n$  from a probability measure  $P$ , defined as the coordinate projections of a product probability space  $(\mathcal{X}^\infty, \mathcal{A}^\infty, P^\infty)$ .

In the first subsection we consider classes of functions such that either the uniform-entropy integral or the bracketing integral converges. Then both  $L_p$ -moments and  $\psi_p$ -moments can be bounded by a multiple of the entropy integral times the corresponding moment of the envelope function. In view of general results that bound the higher order  $L_p$  and the  $\psi_p$ -norms in terms of the  $L_1$ -norm, the main job is to derive upper bounds for the expectation  $E^*\|\mathbb{G}_n\|_{\mathcal{F}}$ . We derive such bounds also with a view toward statistical applications in Part 3.

Bounds on moments imply rates of decrease for the tail probabilities  $P^*(\|\mathbb{G}_n\|_{\mathcal{F}} > t)$  as  $t \rightarrow \infty$ . For uniformly bounded classes  $\mathcal{F}$ , the tails of  $\|\mathbb{G}_n\|_{\mathcal{F}}$  are exponentially small. In Section 2.14.2, we derive exponential bounds that give the correct constant in the exponent for such classes.

Section 2.14.3 is concerned with the related problem of bounding deviation probabilities of the type  $P^*(\|\mathbb{G}_n\|_{\mathcal{F}} > C(E^*\|\mathbb{G}_n\|_{\mathcal{F}} + t))$  for a universal constant  $C$ . Here the choice  $C = 1$  would be desirable, but it is unattainable by the present methods. However, even with unknown  $C$ , the bounds appear to be of interest, in particular because they are valid without any conditions on the size of  $\mathcal{F}$ .

Most of the tail bounds are uniform in  $n$ .

### 2.14.1 Finite Entropy Integrals

In this subsection we derive bounds on moments and tail probabilities of  $\|\mathbb{G}_n\|_{\mathcal{F}}$  for classes  $\mathcal{F}$  that possess a finite uniform-entropy or bracketing entropy integral. Such classes permit tail bounds of the order  $\exp(-Ct^p)$  or  $t^{-p}$ , uniformly in  $n$ , depending on the envelope function  $F$ . While the methods used are too simple to obtain sharp bounds, at least the powers of  $t$  in the bounds appear correct.

The tail bounds will be implicitly expressed in terms of Orlicz norms. By Markov's inequality, a finite  $L_p$ -norm  $\|X\|_p$  yields a polynomial bound

$$\mathbf{P}(|X| > t) \leq \frac{1}{t^p} \|X\|_p^p.$$

Alternatively, a bound on the Orlicz norm  $\|X\|_{\psi_p}$  for  $\psi_p(x) = \exp x^p - 1$  and  $p \geq 1$  leads to an exponential bound

$$\mathbf{P}(|X| > t) \leq 2e^{-t^p/\|X\|_{\psi_p}^p}.$$

It is useful to employ also exponential Orlicz norms for  $0 < p < 1$ . The fact that the functions  $x \mapsto \exp x^p - 1$  are convex only on the interval  $(c_p, \infty)$  for  $c_p = (1/p - 1)^{1/p}$  leads to the inconvenience that  $\|\cdot\|_{\psi_p}$  (if defined exactly as before) does not satisfy the triangle inequality. The usual solution is to define  $\psi_p(x)$  as  $\exp x^p - 1$  for  $x \geq c_p$  and to be linear and continuous on  $(0, c_p)$ . Since we are interested in using these norms to measure tail probabilities, the particular adaptation of the definition of  $\psi_p$  is not important.

In Theorem 2.14.5 (ahead), it is shown that a general Orlicz norm of  $\|\mathbb{G}_n\|_{\mathcal{F}}^*$  is bounded by its  $L_1$ -norm plus the corresponding Orlicz norm of the envelope function  $F$ . In the first results, we therefore focus on bounding the  $L_1$ -norm, or the  $L_p$ -norm if this does not complicate the result.

The first result is for classes satisfying a uniform-entropy bound. Set

$$J(\delta, \mathcal{F}) = \sup_Q \int_0^\delta \sqrt{1 + \log N(\varepsilon \|F\|_{Q,2}, \mathcal{F}, L_2(Q))} d\varepsilon,$$

where the supremum is taken over all discrete probability measures  $Q$  with  $\|F\|_{Q,2} > 0$ . Certainly  $J(1, \mathcal{F}) < \infty$  if  $\mathcal{F}$  satisfies the uniform-entropy condition (2.5.1). For Vapnik-Cervonenkis classes  $\mathcal{F}$ , the function  $J(\delta, \mathcal{F})$  is of the order  $O(\delta \sqrt{\log(1/\delta)})$  as  $\delta \downarrow 0$ .

**2.14.1 Theorem.** *Let  $\mathcal{F}$  be a  $P$ -measurable class of measurable functions with measurable envelope function  $F$ . Then*

$$\|\|\mathbb{G}_n\|_{\mathcal{F}}^*\|_{P,p} \lesssim \|J(\theta_n, \mathcal{F}) \|F\|_n\|_{P,p} \lesssim J(1, \mathcal{F}) \|F\|_{P,2 \vee p}, \quad 1 \leq p.$$

Here  $\theta_n = \|\|\cdot\|_n\|_{\mathcal{F}}^*/\|F\|_n$ , where  $\|\cdot\|_n$  is the  $L_2(\mathbb{P}_n)$ -seminorm and the inequalities are valid up to constants depending only on the  $p$  involved in the statement.

**Proof.** Set  $\mathbb{G}_n^o = n^{-1/2} \sum_{i=1}^n \varepsilon_i f(X_i)$  for the symmetrized empirical process. In view of the symmetrization lemma, Lemma 2.3.1, Orlicz norms of  $\|\mathbb{G}_n\|_{\mathcal{F}}^*$  are bounded by two times the corresponding Orlicz norms of  $\|\mathbb{G}_n^o\|_{\mathcal{F}}$ .

Given  $X_1, \dots, X_n$ , the process  $\mathbb{G}_n^o$  is sub-Gaussian for the  $L_2(\mathbb{P}_n)$ -seminorm  $\|\cdot\|_n$  by Hoeffding's inequality. The value  $\eta_n = \|\|f\|_n\|_{\mathcal{F}}$  is an upper bound for the radius of  $\mathcal{F} \cup \{0\}$  with respect to this norm. The maximal inequality Theorem 2.2.4 gives

$$\left\| \|\mathbb{G}_n^o\|_{\mathcal{F}} \right\|_{\psi_2|X} \lesssim \int_0^{\eta_n} \sqrt{1 + \log N(\varepsilon, \mathcal{F}, L_2(\mathbb{P}_n))} d\varepsilon,$$

where  $\|\cdot\|_{\psi_2|X}$  is the conditional Orlicz norm for  $\psi_2$ , given  $X_1, X_2, \dots$ . Make a change of variable and bound the random entropy by a supremum to see that the right side is bounded by  $J(\theta_n, \mathcal{F}) \|F\|_n$ . Every  $L_p$ -norm is bounded by a multiple of the  $\psi_2$ -Orlicz norm. Hence

$$E_\varepsilon \|\mathbb{G}_n^o\|_{\mathcal{F}}^p \lesssim J(\theta_n, \mathcal{F})^p \|F\|_n^p.$$

Take the expectation over  $X_1, \dots, X_n$  to obtain the left inequality of the theorem. Since  $\theta_n \leq 1$ , the right side of the preceding display is bounded by  $J(1, \mathcal{F})^p \|F\|_n^p$ . For  $p \geq 2$ , this is further bounded by  $J(1, \mathcal{F})^p n^{-1} \sum F^p(X_i)$ , by Jensen's inequality. This gives the inequality on the right side of the theorem. ■

For a given norm  $\|\cdot\|$ , define a bracketing integral of a class of functions  $\mathcal{F}$  as

$$J_{[]}(\delta, \mathcal{F}, \|\cdot\|) = \int_0^\delta \sqrt{1 + \log N_{[]}(\varepsilon \|F\|, \mathcal{F}, \|\cdot\|)} d\varepsilon.$$

The basic bracketing maximal inequality uses the  $L_2(P)$ -norm.

**2.14.2 Theorem.** Let  $\mathcal{F}$  be a class of measurable functions with measurable envelope function  $F$ . For a given  $\eta > 0$ , set

$$a(\eta) = \eta \|F\|_{P,2} / \sqrt{1 + \log N_{[]}(\eta \|F\|_{P,2}, \mathcal{F}, L_2(P))}.$$

Then for every  $\eta > 0$ ,

$$\begin{aligned} \|\|\mathbb{G}_n\|_{\mathcal{F}}^*\|_{P,1} &\lesssim J_{[]}(\eta, \mathcal{F}, L_2(P)) \|F\|_{P,2} + \sqrt{n} P F\{F > \sqrt{n} a(\eta)\} \\ &\quad + \|\|f\|_{P,2}\|_{\mathcal{F}} \sqrt{1 + \log N_{[]}(\eta \|F\|_{P,2}, \mathcal{F}, L_2(P))}. \end{aligned}$$

Consequently, if  $\|f\|_{P,2} < \delta \|F\|_{P,2}$  for every  $f$  in the class  $\mathcal{F}$ , then

$$\|\|\mathbb{G}_n\|_{\mathcal{F}}^*\|_{P,1} \lesssim J_{[]}(\delta, \mathcal{F}, L_2(P)) \|F\|_{P,2} + \sqrt{n} P F\{F > \sqrt{n} a(\delta)\}.$$

In particular, for any class  $\mathcal{F}$ ,

$$\|\|\mathbb{G}_n\|_{\mathcal{F}}^*\|_{P,1} \lesssim J_{[]} (1, \mathcal{F}, L_2(P)) \|F\|_{P,2}.$$

The constants in the inequalities are universal.

In some applications it is useful to use a stronger norm than the  $L_2$ -norm. Then the bracketing integral is larger, but the stronger norm gives better control over the links left over at the end of the chains by the chaining argument. In the preceding theorem, these are bounded in a rather crude manner. The following lemma is more complicated, but will be useful.

**2.14.3 Lemma.** *Let  $\mathcal{F}$  be an arbitrary class of measurable functions and  $\|\cdot\|$  a norm that dominates the  $L_2(P)$ -norm. Then, for every  $\delta > 3\gamma \geq 0$ , there exist deterministic functions  $e_n: \mathcal{F} \mapsto \mathbb{R}$  with  $\|e_n\|_{\mathcal{F}} \leq 8\gamma\sqrt{n}$  such that*

$$\begin{aligned} \left\| \sup_f (\mathbb{G}_n - e_n)^{+*} \right\|_{P,p} &\lesssim \int_{\gamma}^{\delta} \sqrt{1 + \log N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)} d\varepsilon \\ &\quad + \left\| \sup_i |\mathbb{G}_n f_i| \right\|_{P,p} + \left\| \sup_i |\mathbb{G}_n \sup_{f \in \mathcal{F}_i} |f - f_i||^* \right\|_{P,p}, \end{aligned}$$

for a minimal partition  $\mathcal{F} = \cup_{i=1}^m \mathcal{F}_i$  into sets of  $\|\cdot\|$ -diameter at most  $\delta$  and any choice of  $f_i \in \mathcal{F}_i$ . The constants in the inequalities are universal.

**Proofs.** For the proof of the lemma, fix integers  $q_0$  and  $q_2$  such that  $2^{-q_0} < \delta \leq 2^{-q_0+1}$  and  $2^{-q_2-2} < \gamma \leq 2^{-q_2-1}$ . For  $q \geq q_0$ , construct a nested sequence of partitions  $\mathcal{F} = \cup_{i=1}^{N_q} \mathcal{F}_{qi}$  such that the  $q_0$ th partition is the partition given in the statement of the lemma and for, each  $q > q_0$ ,

$$\left\| \left( \sup_{f,g \in \mathcal{F}_{qi}} |f - g| \right)^* \right\| < 2^{-q}.$$

By the choice of  $q_0$  the number of sets in the  $q_0$ th partition satisfies  $N_{q_0} \leq N_{[]} (2^{-q_0}, \mathcal{F}, \|\cdot\|)$  and the preceding display is valid for  $q = q_0$  up to a factor 2. The number  $N_q$  of subsets in the  $q$ th partition can be chosen to satisfy

$$\log N_q \leq \sum_{r=q_0}^q \log N_{[]} (2^{-r}, \mathcal{F}, \|\cdot\|).$$

Now adopt the same notation as in the proof of Theorem 2.5.6 (except define  $a_q$  with  $1 + \log$  rather than  $\log$ ) and decompose, pointwise in  $x$  (which is suppressed in the notation),

$$\begin{aligned} f - \pi_{q_0} f &= (f - \pi_{q_0} f) B_{q_0} f + \sum_{q_0+1}^{q_2} (f - \pi_q f) B_q f \\ &\quad + \sum_{q_0+1}^{q_2} (\pi_q f - \pi_{q-1} f) A_{q-1} f + (f - \pi_{q_2} f) A_{q_2} f. \end{aligned}$$

See the proof of Theorem 2.5.6 for the notation and motivation.

Apply the empirical process  $\mathbb{G}_n$  to each of the three terms separately and take suprema over  $f \in \mathcal{F}$ . The  $L_p(P)$ -norm of the first term resulting from this procedure can be bounded by

$$\left\| \mathbb{G}_n(f - \pi_{q_0}f)B_{q_0}f \right\|_{\mathcal{F}}^* \lesssim \left\| \mathbb{G}_n\Delta_{q_0}f \right\|_{\mathcal{F}}^* + \sqrt{n} \left\| P\Delta_{q_0}f B_{q_0}f \right\|_{\mathcal{F}}.$$

The first term on the right arises as the last term in the upper bound of the lemma. The second term on the right can be bounded by

$$\begin{aligned} a_{q_0}^{-1} P\Delta_{q_0}^2 f &\lesssim \delta \sqrt{1 + \log N_{[]}(\delta/2, \mathcal{F}, \|\cdot\|)} \\ &\lesssim \int_{\delta/3}^{\delta} \sqrt{1 + \log N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)} d\varepsilon. \end{aligned}$$

This is bounded by the right side of the lemma.

The second and third terms in the decomposition can be bounded in exactly the same manner as in the proof of Theorem 2.5.6, where it may be noted that Lemma 2.2.10 bounds the  $\psi_1$ -norm, hence every  $L_p$ -norm up to a constant. This shows that inequality (2.5.5) is also valid with the  $L_p$ -norm of  $\|\mathbb{G}_n\|_{\mathcal{F}}$  on the left. Thus the second and third terms give contributions to the  $L_p(P)$ -norm of  $\|\mathbb{G}_n\|_{\mathcal{F}}^*$  bounded up to a constant by

$$\sum_{q_0+1}^{q_2+1} 2^{-q} \sqrt{1 + \log N_q} \lesssim \int_{2^{-q_2-1}}^{2^{-q_0}} \sqrt{1 + \log N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)} d\varepsilon.$$

This is bounded by the right side of the lemma.

For the final term, define  $e_n(f) = 2\sqrt{n}P\Delta_{q_2}f$ . Then  $\mathbb{G}_n(f - \pi_{q_2}f)A_{q_2}f - e_n(f)$  is bounded above by  $\mathbb{G}_n\Delta_{q_2}f A_{q_2}f$ , and the fourth term in the decomposition shifted by  $e_n$  yields

$$\begin{aligned} \left\| \sup_f (\mathbb{G}_n(f - \pi_{q_2}f)A_{q_2}f - e_n(f))^+ \right\|_{P,p}^* &\leq \left\| \mathbb{G}_n\Delta_{q_2}f A_{q_2}f \right\|_{\mathcal{F}}^* \\ &\lesssim a_{q_2-1}(1 + \log N_{q_2}) + 2^{-q_2} \sqrt{1 + \log N_{q_2}}. \end{aligned}$$

This is bounded by the contribution of the second and third parts of the decomposition. This concludes the proof of the lemma.

For the proof of the theorem, apply the preceding argument with the  $L_2(P)$ -norm substituted for  $\|\cdot\|$  and  $q_2 = \infty$ . Then the fourth term in the decomposition of  $f - \pi_{q_0}f$  vanishes, and the contributions of the second and third terms to  $E^*\|\mathbb{G}_n\|_{\mathcal{F}}$  can be bounded as before. The first term yields the contribution

$$E^* \left\| \mathbb{G}_n(f - \pi_{q_0}f)B_{q_0}f \right\|_{\mathcal{F}} \lesssim \sqrt{n}P^*F\{2F > \sqrt{n}a_{q_0}\}.$$

For  $\delta = 8\eta\|F\|_{P,2}$ , the right side is bounded by  $\sqrt{n}P^*F\{F > \sqrt{n}a(\eta)\}$ . Finally,

$$E^* \left\| \mathbb{G}_n\pi_{q_0}f \right\|_{\mathcal{F}} \lesssim E^* \left\| \mathbb{G}_n\pi_{q_0}f \{F \leq \sqrt{n}a(\eta)\} \right\|_{\mathcal{F}} + \sqrt{n}P^*F\{F > \sqrt{n}a(\eta)\}.$$

Each function  $\pi_{q_0} f \mathbf{1}\{F \leq \sqrt{n}a(\eta)\}$  is uniformly bounded by  $\sqrt{n}a(\eta)$ . Inequality (2.5.5) yields a bound for the first term on the left equal to a multiple of

$$(1 + \log N_{q_0})a(\eta) + \sqrt{1 + \log N_{q_0}} \|\|f\|_{P,2}\|_{\mathcal{F}}.$$

The first term is bounded by a multiple of  $J_{[]}(\eta, \mathcal{F}, L_2(P))\|F\|_{P,2}$ . The second arises as the last term in the first inequality of the theorem.

For the second inequality of the theorem, it suffices to note that the integrand in the bracketing integral is decreasing so that

$$\delta \sqrt{1 + \log N_{[]}(\delta\|F\|_{P,2}, \mathcal{F}, L_2(P))} \leq J_{[]}(\delta, \mathcal{F}, L_2(P)).$$

For the final inequality of the theorem, choose  $\delta = 2$  in the second. Since the class  $\mathcal{F}$  fits in the single bracket  $[-F, F]$ , it follows that  $a(2) = 2\|F\|_{P,2}$ . ■

**2.14.4 Example (Alternative proofs of the Donsker theorems).** The proofs of the preceding maximal inequalities are modelled after the proofs of the uniform-entropy central limit theorem and the bracketing central limit theorem. The inequalities are sufficiently strong to yield short proofs of these results.

First consider the central limit theorem under the conditions of Theorem 2.5.2. Under the uniform-entropy bound (2.5.1), the entropy integral  $J(\delta, \mathcal{F})$  is uniformly bounded and  $J(\delta, \mathcal{F}) \rightarrow 0$  as  $\delta \downarrow 0$ . The entropy integral of the class  $\mathcal{F} - \mathcal{F}$  (with envelope  $2F$ ) is bounded by a multiple of  $J(\delta, \mathcal{F})$ . Application of the  $L_1$ -inequality of Theorem 2.14.1 followed by the Cauchy-Schwarz inequality yields, with  $\theta_n^2 = \|\mathbb{P}_n f^2\|_{\mathcal{F}_\delta}^*/\mathbb{P}_n F^2$ ,

$$\mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}_\delta} \lesssim \mathbb{E}^* J(\theta_n, \mathcal{F})\|F\|_n \leq \left(\mathbb{E}^* J^2(\theta_n, \mathcal{F}) PF^2\right)^{1/2}.$$

The sequence of empirical processes indexed by  $\mathcal{F}$  is asymptotically tight if the right side converges to zero as  $n \rightarrow \infty$  followed by  $\delta \downarrow 0$ . In view of the dominated convergence theorem, this is true if  $\theta_n \rightarrow 0$  in probability.

Without loss of generality, assume that  $F \geq 1$ , so that  $\theta_n^2 \leq \|\mathbb{P}_n f^2\|_{\mathcal{F}_\delta}$ . The desired conclusion follows if  $\|\mathbb{P}_n f^2 - Pf^2\|_{\mathcal{F}_\delta}$  converges in probability to zero. This is certainly the case if the class  $\mathcal{G} = (\mathcal{F} - \mathcal{F})^2$  is Glivenko-Cantelli in probability. Let  $\mathcal{G}_M$  be the functions  $g\mathbf{1}\{F \leq M\}$  when  $g$  ranges over  $\mathcal{G}$ . Since  $N(4\varepsilon M\|F\|_{Q,2}, \mathcal{G}_M, L_2(Q)) \leq N(\varepsilon\|F\|_{Q,2}, \mathcal{F}, L_2(Q))^2$ , the entropy integral of  $\mathcal{G}_M$  with envelope  $4MF$  (!) is bounded by a multiple of  $J(\delta, \mathcal{F})$ . A second application of Theorem 2.14.1 gives

$$\mathbb{E}^* \|\mathbb{P}_n f^2 - Pf^2\|_{\mathcal{F} - \mathcal{F}} \lesssim \frac{1}{\sqrt{n}} J(1, \mathcal{F}) M (PF^2)^{1/2} + PF^2 \{F > M\},$$

for every  $M$ . This converges to zero as  $n \rightarrow \infty$  followed by  $M \rightarrow \infty$ .

Next, consider the simplified version of the bracketing central limit Theorem 2.5.6, which asserts that  $\mathcal{F}$  is Donsker if the bracketing integral

$J_{[]}(\delta, \mathcal{F}, L_2(P))$  is finite. This theorem is a corollary from the second maximal inequality given by Theorem 2.14.2, which shows that  $E^* \|\mathbb{G}\|_{\mathcal{F}_\delta} \rightarrow 0$  as  $n \rightarrow \infty$  followed by  $\delta \downarrow 0$ .

The preceding maximal inequalities can be extended to other Orlicz norms. In fact, Orlicz norms of sums of independent processes can in general be bounded by their  $L_1$ -norm plus a maximum or sum of Orlicz norms of the individual terms. Thus, bounds on  $E^* \|\mathbb{G}_n\|_{\mathcal{F}}$  can be turned into bounds on more general Orlicz norms of  $\|\mathbb{G}_n\|_{\mathcal{F}}^*$ .

**2.14.5 Theorem.** *Let  $\mathcal{F}$  be a class of measurable functions with measurable envelope function  $F$ . Then*

$$\begin{aligned} \|\|\mathbb{G}_n\|_{\mathcal{F}}^*\|_{P,p} &\lesssim \|\|\mathbb{G}_n\|_{\mathcal{F}}^*\|_{P,1} + n^{-1/2+1/p} \|F\|_{P,p} & (p \geq 2), \\ \|\|\mathbb{G}_n\|_{\mathcal{F}}^*\|_{P,\psi_p} &\lesssim \|\|\mathbb{G}_n\|_{\mathcal{F}}^*\|_{P,1} + n^{-1/2}(1 + \log n)^{1/p} \|F\|_{P,\psi_p} & (0 < p \leq 1), \\ \|\|\mathbb{G}_n\|_{\mathcal{F}}^*\|_{P,\psi_p} &\lesssim \|\|\mathbb{G}_n\|_{\mathcal{F}}^*\|_{P,1} + n^{-1/2+1/q} \|F\|_{P,\psi_p} & (1 < p \leq 2). \end{aligned}$$

Here  $1/p + 1/q = 1$  and the constants in the inequalities  $\lesssim$  depend only on the type of norm involved in the statement.

**Proof.** Specialization of Proposition A.1.6 to the present situation gives the inequalities

$$\begin{aligned} \|\|\mathbb{G}_n\|_{\mathcal{F}}^*\|_{P,p} &\lesssim \|\|\mathbb{G}_n\|_{\mathcal{F}}^*\|_{P,1} + n^{-1/2} \|\max_{1 \leq i \leq n} F(X_i)\|_{P,p} & (p \geq 1), \\ \|\|\mathbb{G}_n\|_{\mathcal{F}}^*\|_{P,\psi_p} &\lesssim \|\|\mathbb{G}_n\|_{\mathcal{F}}^*\|_{P,1} + n^{-1/2} \|\max_{1 \leq i \leq n} F(X_i)\|_{P,\psi_p} & (0 < p \leq 1). \end{aligned}$$

Next, Lemma 2.2.2 can be used to further bound the maxima appearing on the right. This gives the first two lines of the theorem. The third line is immediate from Proposition A.1.6. ■

To obtain tail bounds for the random variable  $\|\mathbb{G}_n\|_{\mathcal{F}}^*$ , the last theorem may be combined with the bounds on  $E^* \|\mathbb{G}_n\|_{\mathcal{F}}$  given by the preceding theorems. While the last theorem is valid for any class  $\mathcal{F}$  of functions, the bounds on  $E^* \|\mathbb{G}_n\|_{\mathcal{F}}$  rely on the finiteness of an entropy integral. If either the uniform-entropy integral or the bracketing entropy integral of the class  $\mathcal{F}$  is finite, then

$$E^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim \|F\|_{P,2},$$

where the constant depends on the entropy integral. The  $L_2(P)$ -norm of the envelope function can in turn be bounded by an exponential Orlicz norm. Combination with the last theorem shows that

$$\|\|\mathbb{G}_n\|_{\mathcal{F}}^*\|_{P,\psi_p} \lesssim \|F\|_{P,\psi_p} \quad (0 < p \leq 2).$$

Conclude that  $\|\mathbb{G}_n\|_{\mathcal{F}}^*$  has tails of the order  $\exp(-Ct^p)$  for some  $0 < p \leq 2$ , whenever the envelope function  $F$  has tails of this order. Similar reasoning

can be applied using the  $L_p(P)$ -norms. It may be noted that, for large  $n$  and  $0 < p < 2$ , the size of the  $\psi_p$ -norm of  $\|\mathbb{G}_n\|_{\mathcal{F}}^*$  is mostly determined by the entropy integral and the  $L_2$ -norm of the envelope: according to the preceding theorem, the  $\psi_p$ -norm of the envelope enters the upper bound multiplied by a factor that converges to zero as  $n \rightarrow \infty$ .

To see some of the strength of the preceding theorems, it is instructive to apply them to a class  $\mathcal{F}$  consisting of a single function  $f$ . Then  $J(\delta, \{f\}) = \delta$  and the envelope function is  $|f|$ . For  $\mathcal{F} = \{f\}$ , Theorem 2.14.1 gives

$$\|\mathbb{G}_n f\|_{P,p} \lesssim \left\| \left( \frac{1}{n} \sum_{i=1}^n f^2(X_i) \right)^{1/2} \right\|_{P,p} \leq \|f\|_{P,p}, \quad p \geq 2.$$

This is the upper half of the Marcinkiewicz-Zygmund inequality.<sup>b</sup> Theorem 2.14.5 yields a tightening of the inequality between the far left and the right sides. More interestingly, it gives bounds for the exponential Orlicz norms

$$\begin{aligned} \|\mathbb{G}_n f\|_{P,\psi_p} &\lesssim \|f\|_{P,2} + n^{-1/2}(1 + \log n)^{1/p} \|f\|_{P,\psi_p} & (0 < p \leq 1), \\ \|\mathbb{G}_n f\|_{P,\psi_p} &\lesssim \|f\|_{P,2} + n^{-1/2+1/q} \|f\|_{P,\psi_p} & (1 < p \leq 2). \end{aligned}$$

The last line for  $p = q = 2$  shows that the random variable  $\mathbb{G}_n f$  is sub-Gaussian whenever  $\|f\|_{P,\psi_2}$  is finite. Apart from a universal constant, this generalizes Hoeffding's inequality, which makes the same statement for uniformly bounded functions  $f$ . Hoeffding's inequality for linear combinations of Rademacher variables is given by Lemma 2.2.7 and is the essential ingredient in the proof of Theorem 2.14.1.

## 2.14.2 Uniformly Bounded Classes

In this subsection we refine the tail bounds on  $P^*(\|\mathbb{G}_n\|_{\mathcal{F}} > t)$  in the case of uniformly bounded classes  $\mathcal{F}$ . We present both Hoeffding- and Bernstein-type bounds. Throughout this subsection it is assumed that  $0 \leq f \leq 1$  for every  $f \in \mathcal{F}$ . Without further mention, it is also assumed that  $X_1, X_2, \dots$  are defined as the coordinate projections on a product space  $(\mathcal{X}^\infty, \mathcal{A}^\infty, P^\infty)$  and that the class  $\mathcal{F}$  is pointwise separable.

The first theorem is valid for classes  $\mathcal{F}$  with polynomial covering numbers or polynomial bracketing numbers: for some constants  $V$  and  $K$ ,

$$(2.14.6) \quad \sup_Q N(\varepsilon, \mathcal{F}, L_2(Q)) \leq \left( \frac{K}{\varepsilon} \right)^V, \quad \text{for every } 0 < \varepsilon < K,$$

or

$$(2.14.7) \quad N_{[]}(\varepsilon, \mathcal{F}, L_2(P)) \leq \left( \frac{K}{\varepsilon} \right)^V, \quad \text{for every } 0 < \varepsilon < K.$$

---

<sup>b</sup> E.g., Chow and Teicher (1978), page 356.

In the first inequality, the supremum is taken over all probability measures  $Q$ . In particular, the theorem is valid for Vapnik-Červonenkis classes: according to Theorem 2.6.7, a VC-class of index  $V(\mathcal{F})$  with envelope function  $F = 1$  satisfies (2.14.6) for  $V = 2V(\mathcal{F}) - 2$  and a constant  $K$  that depends on  $V$  only.

The second theorem applies to classes such that, for some constants  $0 < W < 2$  and  $K$ ,

$$(2.14.8) \quad \sup_Q \log N(\varepsilon, \mathcal{F}, L_2(Q)) \leq K \left( \frac{1}{\varepsilon} \right)^W, \quad \text{for every } \varepsilon > 0.$$

for some  $0 < W < 2$  and constant  $K$ . Again the supremum is taken over all probability measures  $Q$ .

**2.14.9 Theorem.** *Let  $\mathcal{F}$  be a class of measurable functions  $f: \mathcal{X} \mapsto [0, 1]$  that satisfies (2.14.6) or (2.14.7). Then, for every  $t > 0$ ,*

$$P^*(\|\mathbb{G}_n\|_{\mathcal{F}} > t) \leq \left( \frac{Dt}{\sqrt{V}} \right)^V e^{-2t^2},$$

for a constant  $D$  that depends on  $K$  only.

**2.14.10 Theorem.** *Let  $\mathcal{F}$  be a class of measurable functions  $f: \mathcal{X} \mapsto [0, 1]$  that satisfies (2.14.8). Then, for every  $\delta > 0$  and  $t > 0$ ,*

$$P^*(\|\mathbb{G}_n\|_{\mathcal{F}} > t) \leq Ce^{Dt^{U+\delta}} e^{-2t^2},$$

where  $U = W(6 - W)/(2 + W)$  and the constants  $C$  and  $D$  depend on  $K$ ,  $W$ , and  $\delta$  only.

We note that the exponent  $U$  in the second theorem increases from 0 to 2 as  $W$  increases from 0 to 2.

While the constant 2 in the exponential  $e^{-2t^2}$  in Theorem 2.14.9 is sharp, the power of the additional term is not. For instance, the empirical distribution function on the line satisfies the exponential bound  $De^{-2t^2}$ , whereas the bound obtained from Theorem 2.14.9 is of the form  $Dt^2e^{-2t^2}$ , since (2.14.6) holds with  $V = 2$  in this case. We shall consider two improvements for  $\mathcal{F}$  equal to a class of indicators of sets.

The exponential bounds for the suprema  $\|\mathbb{G}_n\|_{\mathcal{F}}$  are based on exponential bounds for the individual variables  $\mathbb{G}_n f$ . For sets  $C$  with probability bounded away from zero and one, an improved exponential bound can be derived from the exponential bound of Talagrand for the tail of a binomial random variable

$$P(|\mathbb{G}_n(C)| > t) \leq \frac{K}{t} e^{-2t^2}, \quad \text{for every } t > 0$$

(see Proposition A.6.4). This suggests that it might be possible to improve on Theorem 2.14.9 in the case of sets. In fact, this improvement is possible. We replace (2.14.6) and (2.14.7) by their corresponding  $L_1$  versions.

Suppose that  $\mathcal{C}$  is a class of sets such that, for given constants  $K$  and  $V$ , either

$$(2.14.11) \quad \sup_Q N(\varepsilon, \mathcal{C}, L_1(Q)) \leq \left( \frac{K}{\varepsilon} \right)^V, \quad \text{for every } 0 < \varepsilon < K,$$

or

$$(2.14.12) \quad N_{[]}(\varepsilon, \mathcal{C}, L_1(P)) \leq \left( \frac{K}{\varepsilon} \right)^V, \quad \text{for every } 0 < \varepsilon < K.$$

The supremum is taken over all probability measures  $Q$ . We note that the present  $V$  is  $1/2$  times the exponent  $V$  in conditions (2.14.6) and (2.14.7).

**2.14.13 Theorem.** *Let  $\mathcal{C}$  be a class of sets that satisfies (2.14.11) or (2.14.12). Then*

$$P^*(\|\mathbb{G}_n\|_c > t) \leq \frac{D}{t} \left( \frac{DKt^2}{V} \right)^V e^{-2t^2},$$

for every  $t > 0$  and a constant  $D$  that depends on  $K$  only.

Considering the special case of the empirical distribution function on the line once again, we see that Theorem 2.14.13 still does not yield the known exponential bound  $D e^{-2t^2}$ . Since the collection of left half-lines satisfies (2.14.11) with  $V = 1$ , Theorem 2.14.13 yields a bound of the form  $D t e^{-2t^2}$ .

To obtain the “correct” power of  $t$  in the bounds requires consideration of the size of the sets in a neighborhood of the collection  $\{C \in \mathcal{C}: P(C) = 1/2\}$  for which the variance of  $\mathbb{G}_n C$  is maximal. To this end, define

$$\mathcal{C}_\delta = \left\{ C \in \mathcal{C}: |P(C) - 1/2| \leq \delta \right\}.$$

It is the dimensionality of such neighborhoods that really governs which power of  $t$  is possible. The following theorem gives a refinement of Theorem 2.14.13 to clarify this dependence.

**2.14.14 Theorem.** *Let  $\mathcal{C}$  be a class of sets that satisfies either (2.14.11) or (2.14.12), and suppose moreover that*

$$(2.14.15) \quad N(\varepsilon, \mathcal{C}_\delta, L_1(P)) \leq K' \delta^W \varepsilon^{-V'}, \quad \text{for every } \delta \geq \varepsilon > 0,$$

for some constant  $K'$ . Then

$$P^*(\|\mathbb{G}_n\|_c > t) \leq D t^{2V' - 2W} e^{-2t^2},$$

for every  $t > K\sqrt{W}$  and a constant  $D$  that depends on  $K, K', W, V$ , and  $V'$  only.

For  $\mathcal{C}$  equal to the collection of left half-lines in  $\mathbb{R}$ , inequality (2.14.15) holds with  $V' = 1$  and  $W = 1$  (at worst) for any probability measure  $P$ .

Hence Theorem 2.14.14 gives a bound of the form  $D e^{-2t^2}$  in agreement with the bound of Dvoretzky, Kiefer, and Wolfowitz (1956).

When  $P$  is uniform on  $[0, 1]^d$  and  $\mathcal{C}$  is the collection of lower-left subrectangles of  $[0, 1]^d$ , (2.14.15) holds with  $V' = d$ ,  $W = 1$ , and the bound given by Theorem 2.14.14 is of the form  $D t^{2(d-1)} e^{-2t^2}$ . However, this bound is not uniform in  $P$ . See Smith and Dudley (1992) and Adler and Brown (1986).

The preceding theorems give tail bounds in terms of entropy and envelope but do not show the dependence on the variances of the functions in  $\mathcal{F}$ . In many applications, especially when an entire collection  $\mathcal{F}$  is replaced by a subcollection, as in the study of oscillation moduli, it is useful to introduce the maximal variance

$$\sigma_{\mathcal{F}}^2 = \|P(f - Pf)^2\|_{\mathcal{F}}$$

into the bound. If every  $f$  takes its values in the interval  $[0, 1]$ , this variance is maximally  $1/4$ . Thus a bound of the order  $\exp(-(1/2)t^2/\sigma_{\mathcal{F}}^2)$  could be better than the bounds given by the preceding theorems. This bound is valid in the limit as  $n \rightarrow \infty$ , but for finite sample size a correction is necessary.

**2.14.16 Theorem.** Let  $\mathcal{F}$  be a class of measurable functions  $f: \mathcal{X} \mapsto [0, 1]$  that satisfies (2.14.6). Then for every  $\sigma_{\mathcal{F}}^2 \leq \sigma^2 \leq 1$  and every  $\delta > 0$ ,

$$P^*(\|\mathbb{G}_n\|_{\mathcal{F}} > t) \leq C \left(\frac{1}{\sigma}\right)^{2V} \left(1 \vee \frac{t}{\sigma}\right)^{3V+\delta} e^{-\frac{1}{2} \frac{t^2}{\sigma^2 + (3+t)/\sqrt{n}}},$$

for every  $t > 0$  and a constant  $C$  that depends on  $K$ ,  $V$ , and  $\delta$  only.

**2.14.17 Theorem.** Let  $\mathcal{F}$  be a class of measurable functions  $f: \mathcal{X} \mapsto [0, 1]$  that satisfies (2.14.8). Then for every  $\sigma_{\mathcal{F}}^2 \leq \sigma^2 \leq 1$ , every  $\delta > 0$ ,  $0 < p(1-u) < 2$ , and  $0 < u < 1$ ,

$$P^*(\|\mathbb{G}_n\|_{\mathcal{F}} > t) \leq C e^{D \left(\frac{1}{\sigma}\right)^W \left(\frac{t}{\sigma}\right)^{p(1-u)+\delta} + 5 \left(\frac{t}{\sigma}\right)^{2u} e^{-\frac{1}{2} \frac{t^2}{\sigma^2 + (3(t/\sigma)^u + t)/\sqrt{n}}}},$$

for every  $t > 0$ , where  $p = W(6-W)/(2-W)$  and the constants  $C$  and  $D$  depend on  $K$ ,  $W$ , and  $\delta$  only.

We do not include proofs of all of Theorems 2.14.9 – 2.14.17. Theorems 2.14.10, 2.14.16, and 2.14.17 are proved completely in this subsection. For Theorem 2.14.9, we give a proof assuming (2.14.6), but with the power of  $t$  in the theorem equal to  $3V + \delta$  instead of  $V$ . After developing more tools needed to prove the tighter bounds in the following subsection, we give a complete proof of Theorem 2.14.13 under the assumption (2.14.11) in Subsection 2.14.4. For complete proofs of Theorems 2.14.9, 2.14.13, and 2.14.14, we refer the interested reader to Talagrand (1994).

The proofs of the tail bounds stated in Theorems 2.14.10, 2.14.16, and 2.14.17 rely on two key lemmas. The first lemma bounds the tail probability

of interest by a corresponding tail probability for sampling  $n$  observations without replacement from the empirical measure based on a sample of size  $N = mn$  (with large  $m$  chosen dependent on  $t$  in the proof). This lemma may be considered an alternative to the symmetrization inequalities involving Rademacher variables derived previously: the present bound is more complicated, but it appears to be more economic.

Sampling without replacement can be described in terms of a random permutation. Let  $(R_1, \dots, R_N)$  be uniformly distributed on the set of permutations of  $\{1, \dots, N\}$ , and let  $X_1, \dots, X_N$  be an independent i.i.d. sample from  $P$ . Define  $n' = N - n$  and

$$\tilde{\mathbb{P}}_{n,N} = \frac{1}{n} \sum_{i=1}^n \delta_{X_{R_i}}; \quad \mathbb{P}_N = \frac{1}{N} \sum_{i=1}^N \delta_{X_i}.$$

Then the tail probabilities of the empirical process of  $n$  observations can be bounded by the tail probabilities of  $\tilde{\mathbb{P}}_{n,N} - \mathbb{P}_N$ .

**2.14.18 Lemma.** *For all  $0 < a < 1$ , any  $\sigma^2 \geq \|P(f - Pf)^2\|_{\mathcal{F}}$ , and every  $t > 0$ ,*

$$P^*(\|\mathbb{P}_n - P\|_{\mathcal{F}} > t) \leq \left[ 1 - \frac{\sigma^2}{(1-a)^2 t^2 n'} \right]^{-1} P^*\left(\|\tilde{\mathbb{P}}_{n,N} - \mathbb{P}_N\|_{\mathcal{F}} > ta \frac{n'}{N}\right).$$

**Proof.** See Devroye (1982) and Shorack and Wellner (1986), pages 829–830. ■

The motivation for this step is the availability of several relatively sharp bounds for the tail probabilities of the mean of a sample of size  $n$  without replacement. In the proof of the main results, these are applied conditionally on the sample  $X_1, \dots, X_N$ . For given real numbers  $c_1, \dots, c_N$ , set

$$\bar{c}_N = \frac{1}{N} \sum_{i=1}^N c_i; \quad \sigma_N^2 = \frac{1}{N} \sum_{i=1}^N (c_i - \bar{c}_N)^2; \quad \Delta_N = \max_{1 \leq i \leq N} c_i - \min_{1 \leq i \leq N} c_i.$$

The following lemma collects four exponential bounds on the tail probabilities of a mean of a sample without replacement.

**2.14.19 Lemma.** *Let  $U_1, \dots, U_n$  be a random sample without replacement from the real numbers  $\{c_1, \dots, c_N\}$ . Then for every  $t > 0$ ,*

$$P(|\bar{U}_n - \bar{c}_N| > t) \leq \begin{cases} 2 \exp\left(-\frac{2nt^2}{\Delta_N^2}\right) & (\text{Hoeffding}), \\ 2 \exp\left(-\frac{2nt^2}{(1-(n-1)/N)\Delta_N^2}\right) & (\text{Serfling}), \\ 2 \exp\left(-\frac{nt^2}{2\sigma_N^2 + t\Delta_N}\right) & (\text{Hoeffding-Bernstein}), \\ 2 \exp\left(-\frac{nt^2}{m\sigma_N^2}\right) & \text{if } N = mn \quad (\text{Massart}). \end{cases}$$

**Proof.** For a proof of the first three inequalities, see Shorack and Wellner (1986). We prove the fourth.

The sample without replacement can be generated in two steps. First, partition the set of points  $\{c_1, \dots, c_N\}$  randomly into  $n$  subsets  $J_1, \dots, J_n$  of  $m$  elements each. Next, randomly choose one element of each subset. Let  $E_2$  denote expectation with respect to the second step only, treating the partition obtained in the first step as fixed. Given the partition, the variables  $U_1, \dots, U_n$  are independent with means  $\bar{c}_1, \dots, \bar{c}_n$  equal to the averages over the  $c_i$  in each partitioning set. The average of these averages is the grand average  $\bar{c}_N$ . Thus

$$E_2 e^{s(\bar{U}_n - \bar{c}_N)} = \prod_{i=1}^n E_2 e^{s(U_i - \bar{c}_i)/n} \leq \prod_{i=1}^n e^{s^2 \Delta_i^2 / 8n^2},$$

by Hoeffding's inequality (Problem 2.14.2), where the numbers  $\Delta_i = \max_{j \in J_i} c_j - \min_{j \in J_i} c_j$  are the ranges of the partitioning sets. Since  $\Delta_i^2 \leq 2 \sum_{j \in J_i} (c_j - \bar{c}_N)^2$ , the sum of the  $\Delta_i^2$  is bounded by  $2N\sigma_N^2$ . Conclude that, for every  $s$ ,

$$P(\bar{U}_n - \bar{c}_N > t) \leq e^{-st} E E_2 e^{s(\bar{U}_n - \bar{c}_N)} \leq e^{-st + s^2 m \sigma_N^2 / 4n}.$$

The choice  $s = 2nt/m\sigma_N^2$  yields an upper bound equal to half the upper bound given by the lemma. The probability  $P(-\bar{U}_n + \bar{c}_N > t)$  can be bounded similarly. ■

**Proof of Theorems 2.14.9, 2.14.10, 2.14.16, and 2.14.17.** (Recall that Theorem 2.14.9 will only be proved with  $3V + \delta$  instead of  $V$ .) It suffices to prove the inequalities for sufficiently large  $t$ , since their validity for small  $t$  can be ensured by choice of the constant  $C$ . Here "sufficiently large" means larger than some value depending only on  $K$ ,  $V$ ,  $W$ , and  $\delta$  as in the statements of the theorems.

For a given  $N = mn$  and a sequence of positive constants  $\varepsilon_q \downarrow 0$ , take a minimal  $\varepsilon_q$ -net for  $\mathcal{F}$  for the  $L_2(\mathbb{P}_N)$ -semimetric. For each  $f$ , let  $\pi_q f$  be a closest element in this net. Fix  $q_0$ . The series

$$(\tilde{\mathbb{P}}_{n,N} - \mathbb{P}_N)\pi_{q_0} f + \sum_{q>q_0} (\tilde{\mathbb{P}}_{n,N} - \mathbb{P}_N)(\pi_q - \pi_{q-1})f$$

is convergent with limit  $(\tilde{\mathbb{P}}_{n,N} - \mathbb{P}_N)f$ . To see this, note first that the  $q$ th partial sum of the series telescopes out to  $(\tilde{\mathbb{P}}_{n,N} - \mathbb{P}_N)\pi_q f$ . Since  $\tilde{\mathbb{P}}_{n,N} f \leq m\mathbb{P}_N f$  for every nonnegative  $f$ , we have

$$|(\tilde{\mathbb{P}}_{n,N} - \mathbb{P}_N)(f - \pi_q f)|^2 \leq 2(\tilde{\mathbb{P}}_{n,N} + \mathbb{P}_N)(f - \pi_q f)^2 \leq 2(m+1)\varepsilon_q^2 \rightarrow 0,$$

and the conclusion follows. The triangle inequality can now be applied to obtain

$$\|\tilde{\mathbb{P}}_{n,N} - \mathbb{P}_N\|_{\mathcal{F}} \leq \|(\tilde{\mathbb{P}}_{n,N} - \mathbb{P}_N)\pi_{q_0} f\|_{\mathcal{F}} + \sum_{q>q_0} \|(\tilde{\mathbb{P}}_{n,N} - \mathbb{P}_N)(\pi_q - \pi_{q-1})f\|_{\mathcal{F}}.$$

If the  $\varepsilon_q$ -net has  $N_q$  elements, then the suprema are over at most  $N_{q_0}$  in the first term on the right and  $N_q N_{q-1}$  elements in each of the terms of the series, respectively. Conclude that for any positive numbers  $\eta_q$  and  $b$  with  $\sum_{q>q_0} \eta_q + b \leq 1$ :

$$(2.14.20) \quad \begin{aligned} & P_R(\|\tilde{\mathbb{P}}_{n,N} - \mathbb{P}_N\|_{\mathcal{F}} > t) \\ & \leq N_{q_0} \left\| P_R(|(\tilde{\mathbb{P}}_{n,N} - \mathbb{P}_N)\pi_{q_0} f| > tb) \right\|_{\mathcal{F}} \\ & + \sum_{q>q_0} N_q^2 \left\| P_R(|(\tilde{\mathbb{P}}_{n,N} - \mathbb{P}_N)(\pi_q - \pi_{q-1})f| > t\eta_q) \right\|_{\mathcal{F}}. \end{aligned}$$

The probabilities on the right will be bounded with the help of the various exponential bounds of Lemma 2.14.19. Lemma 2.14.18 turns the resulting upper bounds into bounds on the tails of the distribution of  $\|\mathbb{P}_n - P\|_{\mathcal{F}}$ . For appropriate choices of tuning constants, this gives the desired results. In all four cases the second term on the right is negligible with respect to the first.

For the proofs of Theorems 2.14.9 and 2.14.10 (with  $3V + \delta$  instead of  $V$ ), use Hoeffding's inequality on the first term in the right of (2.14.20) and Massart's inequality on each of the terms of the series. Thus, (2.14.20) is bounded by

$$(2.14.21) \quad N_{q_0} 2 \exp(-2nt^2b^2) + \sum_{q>q_0} N_q^2 \exp\left(\frac{-nt^2\eta_q^2}{m4\varepsilon_{q-1}^2}\right).$$

Note that  $\mathbb{P}_N(\pi_q f - \pi_{q-1} f)^2 \leq 2\varepsilon_q^2 + 2\varepsilon_{q-1}^2$  is bounded by  $4\varepsilon_{q-1}^2$ .

For the proof of Theorem 2.14.9, choose  $\alpha$  large and

$$\begin{aligned} a &= 1 - t^{-2}; & b &= 1 - t^{-2}; & m &= \lfloor t^2 \rfloor; & q_0 &= 2 + \lfloor t^{2/(\alpha-1)} \rfloor; \\ \sqrt{m} \varepsilon_q &= q^{-\alpha-1}; & \eta_q &= (q-1)^{-\alpha}. \end{aligned}$$

Combination of Lemma 2.14.18 and (2.14.6) with (2.14.20) and (2.14.21) yields as upper bound for  $P^*(\|\mathbb{G}_n\|_{\mathcal{F}} > t)$  a constant times

$$\begin{aligned} & \left(1 - \frac{n}{4(1-a)^2 t^2 n'}\right)^{-1} \left[ (tq_0^{\alpha+1})^V 2 \exp(-2t^2 b^2 a^2 n'^2 / N^2) \right. \\ & \quad \left. + \sum_{q>q_0} (tq^{\alpha+1})^{2V} 2 \exp(-\frac{1}{4} t^2 (q-1)^2 a^2 n'^2 / N^2) \right]. \end{aligned}$$

For sufficiently large  $t$ , the leading multiplicative factor outside the square brackets is bounded by 2. The quotient  $n'/n = 1 - 1/m$  is bounded below by  $1 - 2t^{-2}$ . Furthermore, the series is bounded by its  $q_0$ th term. To see this, write it as  $\sum_{q>q_0} \exp(-\psi(q))$  for  $\psi(q) = (1/4)t^2(q-1)^2 a^2 n'^2 / N^2 -$

$2V(\alpha + 1) \log q$  and apply Problem 2.14.3. Thus, for sufficiently large  $t$ , the last display is up to a constant bounded by

$$\begin{aligned} t^V \left( 3 + t^{\frac{2}{\alpha-1}} \right)^{V(\alpha+1)} \exp(-2t^2(1-t^{-2})^6) \\ + t^{2V} (2+t^2)^{2V(\alpha+1)} \exp\left(-\frac{1}{4}t^2 t^{\frac{4}{\alpha-1}} (1-2t^{-2})^4\right), \end{aligned}$$

the second term being an upper bound for the  $q_0$ th term of the series. For large  $t$ , the second term is certainly bounded by  $e^{-2t^2}$ . For sufficiently large  $\alpha$ , the first term is bounded by a multiple of  $(1 \vee t)^{3V+\delta} e^{-2t^2}$ . This concludes the proof of the first theorem.

For the proof of Theorem 2.14.10, choose  $\alpha$  large and

$$\begin{aligned} a = 1 - t^{-2}; \quad b = 1 - t^{-r}; \quad m = \lfloor t^2 \rfloor; \quad q_0 = 2 + \lfloor t^{r/(\alpha-1)} \rfloor; \\ \sqrt{m}\varepsilon_q = q^{-\alpha-\beta}; \quad \eta_q = (q-1)^{-\alpha}; \quad \beta = \frac{W\alpha}{2-W}; \quad 2-r = W + Wr\frac{\alpha+\beta}{\alpha-1}. \end{aligned}$$

Combination of Lemma 2.14.18 and (2.14.8) with (2.14.20) and (2.14.21) yields that the probability  $P^*(\|\mathbb{G}_n\|_{\mathcal{F}} > t)$  is bounded by a constant times (2.14.22)

$$\begin{aligned} \left(1 - \frac{n}{4(1-a)^2 t^2 n'}\right)^{-1} \left[ \exp K(tq_0^{\alpha+\beta})^W 2 \exp(-2t^2 a^2 b^2 n'^2 / N^2) \right. \\ \left. + \sum_{q>q_0} \exp 2K(tq^{\alpha+\beta})^W 2 \exp\left(-\frac{1}{4}t^2(q-1)^{2\beta} a^2 n'^2 / N^2\right) \right]. \end{aligned}$$

The multiplicative factor on the left is bounded by 2 for sufficiently large  $t$ . The first term inside the square brackets is bounded by

$$2 \exp\left(Kt^W (2+t^{r/(\alpha-1)})^{(\alpha+\beta)W} - 2t^2(1-2t^{-r})^6\right).$$

Since  $2t^2(1-2t^{-r})^6 \geq 2t^2 - 24t^{2-r}$ , this is further bounded by

$$2 \exp\left(Dt^{W+W r \frac{\alpha+\beta}{\alpha-1}} + 24t^{2-r} - 2t^2\right).$$

The constant  $r$  has been chosen so that the first two terms are both of the order  $t^{2-r}$ . For  $\alpha \rightarrow \infty$  (and hence  $\beta \rightarrow \infty!$ ), the exponent  $2-r$  decreases to  $U$  as in the statement of the theorem.

The constant  $\beta$  has been chosen so that  $2\beta = W(\alpha + \beta)$ , whence the series in (2.14.22) can be written in the form  $\sum_{q>q_0} e^{-\psi(q)}$  for a function of the form  $\psi(q) = c(q-1)^{2\beta} - dq^{2\beta}$ . For sufficiently large  $t$ , we have  $c > d$  and  $\psi$  is increasing and convex for  $q \geq q_0$ . Then Problem 2.14.3 applies and the series is bounded by its  $q_0$ th term. This is of much lower order than  $e^{-2t^2}$ . The second theorem is proved.

For the proofs of Theorems 2.14.16 and 2.14.17, we replace the first term of the bound (2.14.21) for (2.14.20) by a bound based on the Hoeffding-Bernstein inequality. On the set  $C_n$  where  $\mathbb{P}_N(f - \mathbb{P}_N f)^2$  is bounded by

$P(f - Pf)^2 + s$  for every  $f$ , the right side of (2.14.20) is bounded by

$$(2.14.23) \quad N_{q_0} 2 \exp\left(-\frac{nt^2 b^2}{2(\sigma^2 + s) + tb}\right) + \sum_{q > q_0} N_q^2 \exp\left(-\frac{nt^2 \eta_q^2}{m4\varepsilon_{q-1}^2}\right).$$

For the proof of Theorem 2.14.16, choose  $\alpha$  large and

$$\begin{aligned} a &= 1 - (t/\sigma)^{-2}; & b &= 1 - (t/\sigma)^{-2}; & m &= \lfloor (t/\sigma)^2 \rfloor; \\ q_0 &= 2 + \lfloor (t/\sigma)^{2/(\alpha-1)} \rfloor; \\ \sqrt{m}\varepsilon_q &= \sigma q^{-\alpha-1}; & \eta_q &= (q-1)^{-\alpha}; & \sqrt{2N}s &= 3t/\sigma. \end{aligned}$$

Combination of Lemma 2.14.18 and (2.14.6) with (2.14.20) and (2.14.23) yields as upper bound for  $P^*(\|\mathbb{G}_n\|_{\mathcal{F}} > t)$  a constant times

$$\begin{aligned} &\left(1 - \frac{n\sigma^2}{(1-a)t^2 n'}\right)^{-1} \left[ \left(\frac{t}{\sigma} \frac{q_0^{\alpha+1}}{\sigma}\right)^V 2 \exp\left(-\frac{t^2 b^2 a^2 n'^2 / N^2}{2(\sigma^2 + s) + tba n' / N\sqrt{n}}\right) \right. \\ &\quad \left. + \sum_{q > q_0} \left(\frac{t}{\sigma} \frac{q^{\alpha+1}}{\sigma}\right)^{2V} 2 \exp\left(-\frac{1}{4} \frac{t^2}{\sigma^2} (q-1)^2 a^2 \frac{n'^2}{N^2}\right) \right] + P^*(C_n). \end{aligned}$$

Once again the multiplicative factor is bounded and the series is bounded by its  $q_0$ th term. If  $t$  is sufficiently large, then so is  $t/\sigma \geq t$ ,  $m \geq (1/2)(t/\sigma)^2$ , and  $s \leq 3/\sqrt{n}$ . The preceding expression is up to a constant bounded by

$$\begin{aligned} &\left(\frac{1}{\sigma} \frac{t}{\sigma}\right)^V \left(2 + \left(\frac{t}{\sigma}\right)^{2/(\alpha-1)}\right)^{V(\alpha+1)} \exp\left(-\frac{1}{2} \frac{t^2(1-2\sigma^2/t^2)^6}{\sigma^2 + (3+t)/\sqrt{n}}\right) \\ &\quad + \frac{\exp(-\frac{1}{2}t^2/\sigma^2)}{\sigma^{2V}} + P^*(C_n). \end{aligned}$$

The first two terms are bounded as desired. The last term can be bounded with the help of Theorems 2.14.9 and 2.14.10. Since each  $|f|$  is bounded by 1,

$$s1\{C_n\} \leq \|\mathbb{P}_N(f - \mathbb{P}_N)^2 - P(f - Pf)^2\|_{\mathcal{F}} \leq \|\mathbb{P}_N - P\|_{\mathcal{F}^2} + 2\|\mathbb{P}_N - P\|_{\mathcal{F}}.$$

The covering numbers  $N(\varepsilon, \mathcal{F}^2, L_2(Q))$  of the squares are bounded by  $N(\varepsilon/2, \mathcal{F}, L_2(Q))$ . Thus, under (2.14.8) the probability  $P^*(C_n)$  is bounded by  $\psi(s\sqrt{N}/3) \exp(-2Ns^2/9)$  for a function  $\psi$  of the form  $\psi(s) = e^{Ds^U}$  for some  $U < 2$ . For the present choice of  $s$ , this is bounded by a multiple of  $\exp(-t^2/(2\sigma^2))$ . This concludes the proof of the third theorem.

For the proof of Theorem 2.14.17, choose  $\alpha$  large and

$$\begin{aligned} a &= 1 - (t/\sigma)^{-r}; & b &= 1 - (t/\sigma)^{-r}; & m &= \lfloor (t/\sigma)^r \rfloor; \\ q_0 &= 2 + \lfloor (t/\sigma)^{r/(\alpha-1)} \rfloor; & \beta &= \frac{W\alpha}{2-W}; & 2-r &= 2u; \\ \sqrt{m}\varepsilon_q &= \sigma q^{-\alpha-\beta}; & \eta_q &= (q-1)^{-\alpha}; & \sqrt{2N}s &= 3t/\sigma. \end{aligned}$$

Combination of Lemma 2.14.18 and (2.14.6) with (2.14.20) and (2.14.23) yields as upper bound for  $P^*(\|\mathbb{G}_n\|_{\mathcal{F}} > t)$  a constant times

$$\begin{aligned} & \left(1 - \frac{n\sigma^2}{(1-a)^2 t^2 n'}\right)^{-1} \left[ P^*(C_n) \right. \\ & + \exp K \left( \left(\frac{t}{\sigma}\right)^{r/2} \left(\frac{q_0^{\alpha+\beta}}{\sigma}\right)^W \right) 2 \exp \left( -\frac{t^2 b^2 a^2 n'^2 / N^2}{2(\sigma^2 + s) + tba n' / N\sqrt{n}} \right) \\ & \left. + \sum_{q>q_0} \exp 2K \left( \left(\frac{t}{\sigma}\right)^{r/2} \left(\frac{q^{\alpha+\beta}}{\sigma}\right)^W \right) 2 \exp \left( -\frac{1}{4} \frac{t^2}{\sigma^2} (q-1)^{2\beta} a^2 \frac{n'^2}{N^2} \right) \right]. \end{aligned}$$

Since  $(\alpha + \beta)W = 2\beta$ , the series can be written in the form  $\sum_{q>q_0} e^{-\psi(q)}$  for a function of the form  $\psi(q) = c(q-1)^{2\beta} - dq^{2\beta}$ . The constants  $c$  and  $d$  depend on  $t$ . Elementary calculations show that  $\psi$  is convex and increasing for  $q \geq q_0$  if  $t^{2-Wr/2} \geq C\sigma^{2-Wr/2-W}$  for a constant  $C$  depending only on  $K$ . This is certainly the case for large  $t$  if  $2-Wr/2-W \geq 0$ . In that case the series can be bounded by its  $q_0$ th term. By similar arguments as before, the preceding display can be bounded by a multiple of

$$\begin{aligned} & \exp \left( D \left( \frac{1}{\sigma} \right)^W \left( \frac{t}{\sigma} \right)^{\frac{Wr}{2} + Wr \frac{\alpha+\beta}{\alpha-1}} + 5 \left( \frac{t}{\sigma} \right)^{2-r} - \frac{\frac{1}{2} t^2}{\sigma^2 + 3((t/\sigma)^{1-r/2} + t)/\sqrt{n}} \right) \\ & + \exp \left( 2D \left( \frac{1}{\sigma} \right)^W \left( \frac{t}{\sigma} \right)^{\frac{Wr}{2} + Wr \frac{\alpha+\beta}{\alpha-1}} - \frac{1}{4} \left( \frac{t}{\sigma} \right)^{2+\frac{2\beta r}{\alpha-1}} \left( 1 - 2 \left( \frac{\sigma}{t} \right)^r \right)^4 \right). \end{aligned}$$

For  $\alpha \rightarrow \infty$ , the exponent

$$\frac{Wr}{2} + Wr \frac{\alpha + \beta}{\alpha - 1} = \frac{Wr}{2} + \frac{2Wr}{2 - W} \frac{\alpha}{\alpha - 1}$$

converges to  $rp/2 = (1-u)p$ . The upper bound is valid if  $2-Wr/2-W \geq 0$ , which is certainly the case if  $rp/2 < 2$ . ■

### 2.14.3 Deviations from the Mean

Borell's inequality Proposition A.2.1 gives an exponential bound for probabilities of deviations from the mean for suprema of separable Gaussian processes, which is valid without conditions. This also suggests for empirical processes to split a tail bound in a bound on the mean  $\mu_n = E^*\|\mathbb{G}_n\|_{\mathcal{F}}$  and a bound on deviations from the mean. Entropy conditions might come in for bounding the mean, but would hopefully not play a role in bounding the probabilities of deviations from the mean.

The following theorem is valid for any class of uniformly bounded functions. The size of the class enters only through the  $L_1$ -norms  $\mu_n = E^*\|\mathbb{G}_n\|_{\mathcal{F}}$ . For a Donsker class  $\mathcal{F}$ , the sequence  $\mu_n$  converges to the expectation of the norm of a Brownian bridge process. For each fixed  $n$ , it can be bounded in terms of entropy integrals. The following theorem is even useful for classes of functions such that  $\mu_n \rightarrow \infty$ . Set  $\bar{\mu}_n = \mu_n \vee n^{-1/2}$ .

**2.14.24 Theorem.** There exist universal constants  $C$  and  $D$  such that, for every class  $\mathcal{F}$  of measurable functions  $f: \mathcal{X} \mapsto [0, 1]$  such that  $\mathcal{F}$  and  $\mathcal{F}^2$  are  $P$ -measurable,

$$\begin{aligned} & P^*(\|\mathbb{G}_n\|_{\mathcal{F}} > Ct) \\ & \leq \begin{cases} D \exp - \frac{t^2 \sqrt{n}}{\bar{\mu}_n + \sqrt{n}\sigma_{\mathcal{F}}^2}, & \mu_n \leq t \leq \bar{\mu}_n + \sqrt{n}\sigma_{\mathcal{F}}^2, \\ D \exp - t\sqrt{n} \left( \log \frac{et}{\bar{\mu}_n + \sqrt{n}\sigma_{\mathcal{F}}^2} \right)^{1/2}, & t \geq \bar{\mu}_n + \sqrt{n}\sigma_{\mathcal{F}}^2. \end{cases} \end{aligned}$$

The theorem obtains a simpler appearance by stating it directly in terms of deviations from the mean. In the following corollary, the bound is also written in a similar form as in Bernstein's inequality: of the order  $\exp(-Ct^2/\sigma_{\mathcal{F}}^2)$  for  $t$  close to zero and  $\exp(-tC\sqrt{n}/M)$  for large  $t$ .

The constants resulting from the proof below are not sharp. In analogy with Borell's inequality, one might conjecture that in the following theorem the constant  $C = 1$  would work, but at the present time this has not been established.

**2.14.25 Theorem.** There exist universal constants  $C$  and  $D$  such that, for every class  $\mathcal{F}$  of measurable functions  $f: \mathcal{X} \mapsto [-M, M]$  such that  $\mathcal{F}$  and  $\mathcal{F}^2$  are  $P$ -measurable,

$$P^*(\|\mathbb{G}_n\|_{\mathcal{F}} > C(\mu_n + t)) \leq \exp - D \left( \frac{t^2}{\sigma_{\mathcal{F}}^2} \wedge \frac{t\sqrt{n}}{M} \right).$$

**Proofs.** Without loss of generality, assume that  $Pf = 0$  for every  $f \in \mathcal{F}$ . Otherwise replace  $f$  by  $f - Pf$ , which takes its values in  $[-1, 1]$ . For  $C \geq 4$  and  $t \geq \mu_n$ , the probability  $P(|\mathbb{G}_n(f)| < Ct/2)$  is at least  $1/2$  in view of Markov's inequality, for every  $f$ . Let  $\mathbb{G}_n^o = n^{-1/2} \sum_{i=1}^n \varepsilon_i f(X_i)$  be the symmetrized empirical process. By Lemma 2.3.7,

$$P^*(\|\mathbb{G}_n\|_{\mathcal{F}} > Ct) \leq 4P^*\left(\|\mathbb{G}_n^o\|_{\mathcal{F}} > \frac{Ct}{4}\right).$$

The probability on the right can be further bounded by

$$\begin{aligned} & E_X^* P_\varepsilon \left( \|\mathbb{G}_n^o\|_{\mathcal{F}} - E_\varepsilon \|\mathbb{G}_n^o\|_{\mathcal{F}} > \frac{Ct}{8}, \|n\mathbb{P}_n f^2\|_{\mathcal{F}} \leq v \right) \\ & + P \left( E_\varepsilon \|\mathbb{G}_n^o\|_{\mathcal{F}} > Ct/8 \right) + P^* \left( \|n\mathbb{P}_n f^2\|_{\mathcal{F}} > v \right). \end{aligned}$$

The proof proceeds by application of Proposition A.3.1 to the conditional probability in the first term and Lemma A.4.3 to the second and third terms. In view of Lemma 2.3.6, the means of the variables in the second and third terms can be bounded as follows:

$$\begin{aligned} E\mathbb{E}_\varepsilon \|\sqrt{n}\mathbb{G}_n^o\|_{\mathcal{F}} & \leq 2\sqrt{n}\mu_n, \\ E^* \|n\mathbb{P}_n f^2\|_{\mathcal{F}} & \leq n\sigma_{\mathcal{F}}^2 + 2E^* \|\mathbb{P}_n^o f^2\|_{\mathcal{F}} \leq n\sigma_{\mathcal{F}}^2 + 16\sqrt{n}\mu_n, \end{aligned}$$

where the last inequality follows by Proposition A.3.2 applied with  $\phi_i(x) = x^2/2$ . Conclude that the left side of the theorem is bounded up to a constant by

$$e^{-\frac{C^2 t^2 n}{576v}} + e^{-\frac{Ct\sqrt{n}}{16} \log \frac{Ct\sqrt{n}}{192(n\sigma_{\mathcal{F}}^2 + \bar{\mu}_n\sqrt{n})}} + e^{-\frac{1}{2}v \log \frac{v}{192(n\sigma_{\mathcal{F}}^2 + \bar{\mu}_n\sqrt{n})}}.$$

For  $0 < t < n^{-1/2}$ , the bound of the theorem is trivial for large enough  $D$ . For  $t > n^{-1/2}$  in the range  $[\mu_n, \bar{\mu}_n + \sqrt{n}\sigma_{\mathcal{F}}^2)$ , choose  $v = 192e(n\sigma_{\mathcal{F}}^2 + \sqrt{n}\bar{\mu}_n)$ ; for  $t$  in the range  $[\bar{\mu}_n + \sqrt{n}\sigma_{\mathcal{F}}^2, \infty)$ , choose  $v = 192e\sqrt{nt}/\sqrt{\log et}/(\bar{\mu}_n + \sqrt{n}\sigma_{\mathcal{F}}^2)$  to complete the proof of the first theorem.

For the proof of the second theorem, it is no loss of generality to assume that  $M = 1$  and that the functions  $f$  take their values in the unit interval. Replace  $t$  by  $\bar{\mu}_n + t$  in the first theorem. We have

$$\frac{(\bar{\mu}_n + t)^2}{\bar{\mu}_n + \sqrt{n}\sigma_{\mathcal{F}}^2} \geq \frac{2\bar{\mu}_n t + t^2}{\bar{\mu}_n + \sqrt{n}\sigma_{\mathcal{F}}^2} \geq \frac{t^2}{\sqrt{n}\sigma_{\mathcal{F}}^2}, \quad t \leq 2\sqrt{n}\sigma_{\mathcal{F}}^2.$$

Furthermore, the second branch of the inequality is always smaller than  $D \exp -t\sqrt{n}$ . Conclude that there exist constants  $C$  and  $K$  such that

$$P^*(\|\mathbb{G}_n\|_{\mathcal{F}} > C(\bar{\mu}_n + t)) \leq K \exp -\left(\frac{t^2}{\sigma_{\mathcal{F}}^2} \wedge t\sqrt{n}\right).$$

We can finish the proof by “moving” the constant  $K$  into the exponent and replacing  $\bar{\mu}_n$  by  $\mu_n$ . For  $t$  larger than  $L(\sigma_{\mathcal{F}} \vee n^{-1/2})$  for a sufficiently large constant  $L$  depending on  $K$ , the constant  $K$  can be bounded above by the expression  $\exp[(t^2/\sigma_{\mathcal{F}}^2 \wedge t\sqrt{n})/2]$ . On the other hand, for  $t$  smaller than  $L(\sigma_{\mathcal{F}} \vee n^{-1/2})$ , the upper bound given on the right side of the theorem is bounded away from zero and is larger than the probability on the left side for sufficiently large  $C$ , because by Markov’s inequality

$$P^*(\|\mathbb{G}_n\|_{\mathcal{F}} > C(\bar{\mu}_n + t)) \leq \frac{\mu_n}{C(\bar{\mu}_n + t)} \leq \frac{1}{C}, \quad t > 0.$$

This concludes the proof if  $\bar{\mu}_n = \mu_n$ . For  $t > 2n^{-1/2}$ , we have  $\bar{\mu}_n + t/2 < \bar{\mu}_n + t$ , and the left side of the theorem is bounded by  $P^*(\|\mathbb{G}_n\|_{\mathcal{F}} > C(\bar{\mu}_n + t/2))$ , to which we can apply the bound as obtained previously. For  $t \leq 2n^{-1/2}$ , the right side of the theorem is bounded away from zero and we can apply the same argument as before. ■

The following corollary is an example of how the preceding theorems can be applied. It can be regarded as a uniform version of Kiefer’s inequality for a binomial variable, Corollary A.6.3. It will be one of the key tools used in the proof of Theorem 2.14.13. Set  $\mu_n = E^*\|\mathbb{G}_n\|_{\mathcal{F}}$  and  $\bar{\mu}_n = \mu_n \vee n^{-1/2}$ .

**2.14.26 Lemma (Uniform small variance exponential bound).** *There exist universal constants  $D$ ,  $K_0$ , and  $\sigma_0$  such that, for every class  $\mathcal{F}$  of measurable functions  $f: \mathcal{X} \mapsto [0, 1]$  with  $\sigma_{\mathcal{F}}^2 \leq \sigma_0^2$  and  $K_0 \bar{\mu}_n \leq \sqrt{n}$  and such that  $\mathcal{F}$  and  $\mathcal{F}^2$  are  $P$ -measurable,*

$$P^*(\|\mathbb{G}_n\|_{\mathcal{F}} > t) \leq D \exp(-11t^2), \quad \text{for every } t \geq K_0 \bar{\mu}_n.$$

**Proof.** Since  $|f(x) - Pf| \leq 1$  for all  $x$ , the probability on the left side of the above display is zero for all  $t \geq \sqrt{n}$ . Hence any nonnegative bound will be trivially true for  $t > \sqrt{n}$ , and we may restrict attention to  $t \leq \sqrt{n}$ .

For  $\sigma_{\mathcal{F}}^2 \leq \sigma_0^2$  and  $K_0 \bar{\mu}_n \leq \sqrt{n}$ , we have

$$\frac{t^2 \sqrt{n}}{\bar{\mu}_n + \sqrt{n} \sigma_{\mathcal{F}}^2} \geq \frac{t^2}{K_0^{-1} + \sigma_0^2}.$$

This shows that the first expression on the right in Theorem 2.14.24 is bounded by  $\exp -11C^2t^2$  for every  $t$  if  $K_0^{-1} + \sigma_0^2$  is chosen sufficiently small. To bound the second expression, fix  $\varepsilon > 0$ . For  $\bar{\mu}_n + \sqrt{n} \sigma_{\mathcal{F}}^2 \leq t \leq \varepsilon \sqrt{n}$ , we have

$$t \sqrt{n} \left( \log \frac{et}{\bar{\mu}_n + \sqrt{n} \sigma_{\mathcal{F}}^2} \right)^{1/2} \geq \frac{t^2}{\varepsilon} (\log e)^{1/2},$$

while for  $\varepsilon \sqrt{n} \leq t \leq \sqrt{n}$ ,  $K_0 \bar{\mu}_n \leq \sqrt{n}$ , and  $\sigma_{\mathcal{F}}^2 \leq \sigma_0^2$ ,

$$t \sqrt{n} \left( \log \frac{et}{\bar{\mu}_n + \sqrt{n} \sigma_{\mathcal{F}}^2} \right)^{1/2} \geq t^2 \left( \log \frac{e\varepsilon}{K_0^{-1} + \sigma_0^2} \right)^{1/2}.$$

Choose sufficiently small  $\varepsilon$  and next sufficiently small  $K_0^{-1} + \sigma_0^2$  to bound the right-hand sides of the last two displays by  $11C^2t^2$ . Finally, also choose  $K_0 \geq C$  to ensure that the present  $t \geq K_0 \bar{\mu}_n$  is in the range of  $t$  that is permitted in Theorem 2.14.24. ■

#### 2.14.4 Proof of Theorem 2.14.13

A first step in the direction of the proof of Theorem 2.14.13 is the following “basic inequality,” which controls the supremum of the empirical process  $\mathbb{G}_n$  over any class  $\mathcal{C}$  containing a set  $C_0$  with  $\sigma_0^2 \leq P(C_0) \leq 1 - \sigma_0^2$  for the constant  $\sigma_0^2$  in Lemma 2.14.26.

**2.14.27 Theorem.** *Let  $\mathcal{C}$  be a separable class of measurable subsets of  $\mathcal{X}$ . Suppose that  $C_0 \in \mathcal{C}$  has  $\sigma_0^2 \leq P(C_0) \leq 1 - \sigma_0^2$  where  $\sigma_0^2$  is the constant determined in Lemma 2.14.26. Let  $\mu_n^0 = E^* \|\mathbb{G}_n\|_{\mathcal{C} \triangle C_0}$ ,  $\bar{\mu}_n^0 = \mu_n^0 \vee n^{-1/2}$ , and  $a = \sup_{C \in \mathcal{C}} P(C \triangle C_0)$ . Then, for  $t \geq 4/\sigma_0^2$ ,*

$$P^*(\|\mathbb{G}_n\|_{\mathcal{C}} > t) \leq \frac{K}{t} e^{-2t^2} \exp(Kat^2 + K\bar{\mu}_n^0 t - \frac{t^4}{4n}).$$

To prove this theorem, define variables  $W_1, W_2$ , and  $W = W_1 + W_2$  through

$$W_1 = n\mathbb{P}_n(C_0) \left\| \frac{\mathbb{P}_n(C_0 - C)}{\mathbb{P}_n(C_0)} - \frac{P(C_0 - C)}{P(C_0)} \right\|_c,$$

$$W_2 = n\mathbb{P}_n(C_0^c) \left\| \frac{\mathbb{P}_n(C - C_0)}{\mathbb{P}_n(C_0^c)} - \frac{P(C - C_0)}{P(C_0^c)} \right\|_c.$$

Also, define a collection of functions by

$$\varphi_r(t) = \frac{t^2}{r} 1\{t \leq r\} + t \left( \log\left(\frac{et}{r}\right)\right)^{1/2} 1\{t \geq r\}.$$

Theorem 2.14.24 can be reformulated in terms of these functions and asserts that there exist universal constants  $C, D$  such that with  $S = \sqrt{n}\bar{\mu}_n + n\sigma_{\mathcal{F}}^2$

$$P^*(\|\mathbb{G}_n\|_{\mathcal{F}} > t) \leq D \exp -\varphi_S\left(\frac{t\sqrt{n}}{C}\right), \quad t \geq C\mu_n.$$

**2.14.28 Lemma.** With the notation of Theorem 2.14.27,  $S = na + \sqrt{n}\bar{\mu}_n^0$ ,  $h(u) = 2u^2 + (1/4)u^4$ , and  $\varphi(w) = \varphi_S(w/C)1\{w \geq C\sqrt{n}\mu_n\}$ ,

$$n\|\mathbb{P}_n - P\|_c \leq n|\mathbb{P}_n(C_0) - P(C_0)|\left(1 + \frac{2a}{\sigma_0^2}\right) + W \equiv nU\left(1 + \frac{2a}{\sigma_0^2}\right) + W,$$

where for all  $w \geq 0$ ,  $u > t > 0$ ,

$$P^*(U \geq t, W \geq w) \leq \frac{K}{\sqrt{nu}} \exp(-nh(u) - \varphi(w) + 5nu(u - t)).$$

**Proof.** For any set  $C \in \mathcal{C}$ , we can decompose

$$\mathbb{P}_n(C) = \mathbb{P}_n(C_0) + \mathbb{P}_n(C - C_0) - \mathbb{P}_n(C_0 - C),$$

and similarly with  $\mathbb{P}_n$  replaced by  $P$ . Consequently, the difference  $|\mathbb{P}_n(C) - P(C)|$  can be bounded by a sum of three terms. The first term directly yields a term that involves  $U$ . The second term can be bounded by

$$|\mathbb{P}_n(C_0^c)| \left| \frac{\mathbb{P}_n(C - C_0)}{\mathbb{P}_n(C_0^c)} - \frac{P(C - C_0)}{P(C_0^c)} \right| + |\mathbb{P}_n(C_0^c)| \left| \frac{P(C - C_0)}{P(C_0^c)} - \frac{P(C - C_0)}{\mathbb{P}_n(C_0^c)} \right|.$$

This is bounded by  $1/n$  times  $W_2 + Ua/\sigma_0^2$ . The third term can be handled similarly.

Now consider the exponential bound. Since  $nU$  is distributed as the absolute deviation from the mean of a binomial  $(n, P(C_0))$  variable, Talagrand's inequality for the binomial tail (Proposition A.6.4) yields

$$P(U \geq t) \leq \frac{2K}{\sqrt{nu}} \exp(-nh(u)) \exp 5nu(u - t).$$

For  $I \subset \{1, \dots, n\}$ , let  $\Omega_I$  be the set such that  $X_i \in C_0$  for  $i \in I$  and  $X_i \notin C_0$  otherwise. Define  $P_1$  on  $\Omega_I$  by  $P_1(A) = P(A \cap C_0)/P(C_0)$ . Then for

$\#I = k$ , conditionally on  $\Omega_I$ ,  $W_1$  has the same distribution as  $\sqrt{k}\|\mathbb{G}_k^Y\|_{C_1}$  where  $Y_1, \dots, Y_k$  are i.i.d.  $P_1$  and  $C_1 = \{C_0 - C: C \in \mathcal{C}\}$ . Similarly, if we define  $P_2$  on  $\Omega_I$  by  $P_2(A) = P(A \cap C_0^c)/P(C_0^c)$ , then conditionally on  $\Omega_I$ ,  $W_2$  has the same distribution as  $\sqrt{n-k}\|\mathbb{G}_{n-k}^Z\|_{C_2}$ , where  $Z_1, \dots, Z_{n-k}$  are i.i.d.  $P_2$  and  $C_2 = \{C - C_0: C \in \mathcal{C}\}$ . Also, note that for any class  $\mathcal{F}$  of functions,

$$\sqrt{k}\mathbb{E}^*\|\mathbb{G}_k^Y\|_{\mathcal{F}} \leq K\sqrt{n}\mathbb{E}^*\|\mathbb{G}_n\|_{\mathcal{F}}$$

for an absolute constant  $K$  (Problem 2.14.6). A similar inequality is valid for  $\sqrt{n-k}\mathbb{E}^*\|\mathbb{G}_{n-k}^Z\|_{\mathcal{F}}$ . By Theorem 2.14.24 applied twice conditionally on  $\Omega_I$ , for  $w \geq 2CK\sqrt{n}\mu_n^0$ ,

$$\mathbb{P}^*(W \geq w | \Omega_I) \leq 2D \exp -\varphi_{S'}\left(\frac{w}{2C}\right) \leq 2D \exp -\varphi_S\left(\frac{w}{K}\right),$$

with

$$\begin{aligned} S' &= \left[ \sqrt{k}\mathbb{E}^*\|\mathbb{G}_k^Y\|_{C_1} \vee 1 + k\|P_1(f - P_1f)^2\|_{C_1} \right] \\ &\quad \vee \left[ \sqrt{n-k}\mathbb{E}^*\|\mathbb{G}_{n-k}^Z\|_{C_2} \vee 1 + (n-k)\|P_2(f - P_2f)^2\|_{C_2} \right] \\ &\leq K\sqrt{n}\bar{\mu}_n^0 + n\frac{a}{\sigma_0^2} \\ &\leq K'(na + \sqrt{n}\bar{\mu}_n^0) \equiv K'S. \end{aligned}$$

Therefore, it follows that, for  $u > t > 0$  and  $w > 0$ ,

$$\begin{aligned} \mathbb{P}^*(U \geq t, W \geq w) &= \mathbb{E}^*\left[\sum_I 1\{\Omega_I\} 1\{U \geq t\} \mathbb{P}^*(W \geq w | \Omega_I)\right] \\ &\leq 2D \exp -\varphi_S(w/K) \mathbb{P}(U \geq t) \\ &\leq \frac{2DK}{\sqrt{nu}} \exp(-nh(u) - \varphi(w) + 5nu(u-t)). \end{aligned}$$

This concludes the proof. ■

**Proof of Theorem 2.14.27.** Consider the function  $\varphi$  defined in Lemma 2.14.28 and note that  $\varphi(t)/t$  increases. Define  $u$  by

$$u(1 + 2\frac{a}{\sigma_0^2}) = \frac{t}{\sqrt{n}}.$$

Then  $u \leq 1$  since  $t \leq \sqrt{n}$  yields

$$u = \frac{t/\sqrt{n}}{1 + 2a/\sigma_0^2} \leq \frac{1}{1 + 2a/\sigma_0^2} \leq 1.$$

Let  $d \geq 1/u$  be the smallest number satisfying  $\varphi(d)/d \geq 11u$ . Suppose that  $nU(1 + 2a/\sigma_0^2) + W \geq t\sqrt{n}$ . Let  $l \geq 0$  be the smallest integer such that

$$nU(1 + 2a/\sigma_0^2) \geq t\sqrt{n} - (l+1)d.$$

Then if  $l > 0$ ,

$$nU(1 + 2a/\sigma_0^2) \leq t\sqrt{n} - ld,$$

so  $W \geq ld$ . Thus, since  $W \geq 0$ , we can find  $l \geq 0$  such that

$$W \geq ld \quad \text{and} \quad U \geq u - \frac{(l+1)d}{n}.$$

In other words,

$$\left\{ nU(1 + \frac{2a}{\sigma_0^2}) + W \geq t\sqrt{n} \right\} \subset \bigcup_{l=1}^{\infty} \left\{ W \geq ld, U \geq u - \frac{(l+1)d}{n} \right\}.$$

By Lemma 2.14.28, this implies that

$$\begin{aligned} P^*(\|\mathbb{G}_n\|_C > t) &\leq \sum_{l=0}^{\infty} P\left(U \geq u - \frac{(l+1)d}{n}, W \geq ld\right) \\ &\leq \frac{K}{\sqrt{nu}} \sum_{l=0}^{\infty} \exp\left(-nh(u) - \varphi(ld) + 5nu(l+1)\frac{d}{n}\right) \\ &= \frac{K}{\sqrt{nu}} \exp(-nh(u)) \sum_{l=0}^{\infty} \exp(5u(l+1)d - \varphi(ld)). \end{aligned}$$

For  $l \geq 1$ ,

$$\varphi(ld) \geq l\varphi(d) \geq l \cdot 11ud \geq 5u(l+1)d + lud.$$

Hence  $5u(l+1)d - \varphi(ld) \leq -lud \leq -l$  and

$$\sum_{l=1}^{\infty} e^{5u(l+1)d - \varphi(ld)} \leq \sum_{l=1}^{\infty} e^{-l} \equiv K.$$

Thus

$$P^*(\|\mathbb{G}_n\|_C > t) \leq \frac{K}{\sqrt{nu}} e^{-nh(u)} e^{5ud}.$$

It remains to bound  $ud$  and  $u$  in terms of  $t$ .

By Problem 2.14.5 with  $t = K(C)u\sqrt{S}$ , it follows that  $d \leq d_0 \equiv K(C)uS$ . Since  $d \geq 1/u$  and  $d \geq C\sqrt{n}\mu_n^0$ , we deduce that that  $d \leq \max\{1/u, C\sqrt{n}\mu_n^0, K(C)uS\}$ , or

$$ud \leq \max\{1, C\sqrt{n}\mu_n^0 u, K(C)u^2 S\}.$$

Hence, using  $t/\sqrt{n} \leq 1$ ,

$$ud \leq K\{1 + \mu_n^0 t + n^{-1}t^2 S\} \leq \tilde{K}\{1 + \bar{\mu}_n^0 t + at^2\}.$$

Finally, we bound  $h(u)$ . By convexity of  $h$ , we have  $h'(u) \leq 5u$  and

$$\frac{t}{\sqrt{n}} - u = \frac{t}{\sqrt{n}} \left(1 - \frac{1}{1 + 2a/\sigma_0^2}\right).$$

It follows that

$$\begin{aligned} h(u) &\geq h\left(\frac{t}{\sqrt{n}}\right) + \left(u - \frac{t}{\sqrt{n}}\right)h'\left(\frac{t}{\sqrt{n}}\right) \\ &\geq h\left(\frac{t}{\sqrt{n}}\right) - 5\frac{t}{\sqrt{n}}\left(\frac{t}{\sqrt{n}} - u\right) \\ &= h\left(\frac{t}{\sqrt{n}}\right) - 5\frac{t^2}{n}\frac{2a/\sigma_0^2}{1+2a/\sigma_0^2}. \end{aligned}$$

Putting this together yields the claimed inequality. ■

We are now ready to start the proof of Theorem 2.14.13. We first decompose  $\mathcal{C}$  as  $\mathcal{C} = \mathcal{C}_0 \cup \mathcal{C}_1$ , where

$$\mathcal{C}_0 = \{C \in \mathcal{C}: P(C) \leq \sigma_0^2 \text{ or } P(C) \geq 1 - \sigma_0^2\}$$

and  $\mathcal{C}_1 = \mathcal{C} - \mathcal{C}_0$ . Then  $\|\mathbb{G}_n\|_{\mathcal{C}} = \|\mathbb{G}_n\|_{\mathcal{C}_0} \vee \|\mathbb{G}_n\|_{\mathcal{C}_1}$  so that

$$(2.14.29) \quad \mathbb{P}^*(\|\mathbb{G}_n\|_{\mathcal{C}} > t) \leq \mathbb{P}^*(\|\mathbb{G}_n\|_{\mathcal{C}_0} > t) + \mathbb{P}^*(\|\mathbb{G}_n\|_{\mathcal{C}_1} > t).$$

The first term is bounded by  $Ct^{-1}e^{-2t^2}$  for  $t \geq \tilde{K}\sqrt{V \log K}$  by Lemma 2.14.26 (cf. Problem 2.14.9). Thus, it suffices to bound the second term in (2.14.29), where  $\sigma_0^2 \leq P(C) \leq 1 - \sigma_0^2$  for all  $C \in \mathcal{C}_1$ . Without loss of generality, we may assume that these inequalities hold for all  $C \in \mathcal{C}$ , and relabel  $\mathcal{C}_1$  as  $\mathcal{C}$ .

To prove the desired inequality for the supremum over  $\mathcal{C}$ , we partition and then apply Theorem 2.14.27. The strategy is to decompose  $\mathcal{C}$  into pieces  $\mathcal{D}_j$ ,  $j = 1, \dots, M$ , satisfying

$$a_j \equiv \sup_{C \in \mathcal{D}_j} P(C) \leq a \equiv \frac{V}{t^2}.$$

Next, Theorem 2.14.27 is applied to each  $\mathcal{D}_j$  separately. The main work is to further bound

$$(2.14.30) \quad \mu_n \equiv \max_{1 \leq j \leq M} \mu_{nj}^0,$$

where

$$(2.14.31) \quad \mu_{nj}^0 \equiv \mathbb{E}^* \|\mathbb{G}_n\|_{\mathcal{D}_j \triangle \mathcal{D}_j}, \quad j = 1, \dots, M.$$

Here  $D_j$  is an arbitrary fixed element of  $\mathcal{D}_j$  for each  $j \leq M$ . To bound  $\mu_n$ , we will use the following partitioning lemma.

**2.14.32 Lemma (Partitioning).** Let  $(T, \rho)$  be a metric space, and let  $p, q$  be nonnegative integers with  $p < q$ . Consider a partition  $\mathcal{P}_q$  of  $T$  such that each set of  $\mathcal{P}_q$  has diameter  $\leq 4^{-q}$ , and suppose that  $k_l$ ,  $p \leq l \leq q$ , are given positive integers. Then there is an increasing sequence  $\{\mathcal{P}_l\}_{p \leq l \leq q}$  of partitions of  $T$  with the following properties:

- (i) each set of  $\mathcal{P}_l$  has diameter less than or equal to  $4^{-l+1}$ ;
  - (ii) each atom of  $\mathcal{P}_l$  contains at most  $k_l$  atoms of  $\mathcal{P}_{l+1}$ ;
  - (iii) for all  $l < q$ ,  $\#\mathcal{P}_l \leq N(4^{-l}, T, \rho) + \#\mathcal{P}_{l+1}/k_l$ ;
  - (iv) for  $m \geq l$  and  $T_i \in \mathcal{P}_l$ ,  $N(4^{-m}, T_i, \rho) \leq \prod_{j=l+1}^m k_j$ .
- If  $N(4^{-l}, T, \rho) \leq (K4^l)^V$  for  $l \geq p$ ,  $\#\mathcal{P}_q \leq (K4^l)^V$ , and if  $k_l \geq 2 \cdot 4^V$  for all  $l$ , then

$$\#\mathcal{P}_l \leq 2(K4^l)^V \quad \text{for } p \leq l \leq q.$$

**Proof.** The construction proceeds by decreasing induction on  $l$ . Given  $\mathcal{P}_{l+1}$ , we will construct  $\mathcal{P}_l$ .

Set  $N = N(4^{-l}, T, \rho)$ . Consider points  $\{t_i\}_{1 \leq i \leq N}$  of  $T$  so that each point of  $T$  is within  $\rho$ -distance  $4^{-l}$  of at least one  $t_i$ ,  $i = 1, \dots, N$ . Define

$$A_i = \cup\{T_{l+1,j} : T_{l+1,j} \cap B(t_i, 4^{-l}) \neq \emptyset, T_{l+1,j} \in \mathcal{P}_{l+1}\}.$$

Since each set  $T_{l+1,j}$  has diameter  $\leq 4^{-l}$ ,  $A_i$  has diameter at most  $2 \cdot 4^{-l} + 2 \cdot 4^{-l} = 4^{-l+1}$ . Now disjointify the sets  $\{A_i\}$  by defining  $C_i = A_i - \cup_{j < i} A_j$  for every  $i = 1, \dots, N$ . The collection  $\{C_i\}$  forms a partition  $\mathcal{Q}$  of  $T$  that is coarser than  $\mathcal{P}_{l+1}$ . Certain elements of  $\mathcal{Q}$  might contain more than  $\kappa_l$  elements of  $\mathcal{P}_{l+1}$ ; partition any such element of  $\mathcal{Q}$  into sets all of which except one contain exactly  $k_l$  elements of  $\mathcal{P}_{l+1}$  and one being the union of at most  $k_l$  sets of  $\mathcal{P}_{l+1}$ . This constructs  $\mathcal{P}_l$  satisfying (i) and (ii); (iii) follows easily from the construction.

To prove the last assertion, note that by (iii) and the hypotheses we have

$$\#\mathcal{P}_l \leq (K4^l)^V + \frac{\#\mathcal{P}_{l+1}}{2 \cdot 4^V}.$$

Induction then yields the claim. ■

Now we return to the proof of Theorem 2.14.13. Given  $a = V/t^2$ , let  $p$  be the smallest integer such that  $4^{-p+1} \leq a$ . Let  $q \geq p$ ; the value of  $q$  will be determined later in the proof. The partitions  $\{\mathcal{P}\}_{l=p}^q$  of  $T = \mathcal{C}_1$  given by Lemma 2.14.32 will be constructed with  $k_l = \lfloor 3 \cdot 4^V \rfloor \geq 2 \cdot 4^V$ . Thus

$$M = \#\mathcal{P}_p \leq 2\left(\frac{Kt^2}{V}\right)^V.$$

Consider the atoms  $\{\mathcal{D}_j\}_{j=1}^M$  of  $\mathcal{P}_p$ . To bound  $\mu_n$ , we first show that if  $3(q-p)V \leq n4^{-q}$ , then

$$(2.14.33) \quad \begin{aligned} \mu_n &\leq \tilde{K} \left[ \sqrt{V} \left( \sqrt{4^{-p}} + \sqrt{q4^{-q}} + \sqrt{4^{-q} \log K} \right) \right. \\ &\quad \left. + n^{-1/2} (qV + V \log K) \right]. \end{aligned}$$

To prove (2.14.33), first recall (2.14.30) and (2.14.31). By a chaining argument based on the partitions  $\mathcal{P}_l$ , we find that

$$(2.14.34) \quad \mu_{n,j}^0 \leq \sum_{p < l \leq q} \mathbb{E}^* \|\mathbb{G}_n\|_{\mathcal{F}_l} + \mathbb{E}^* \sup_{1 \leq k \leq m} \|\mathbb{G}_n\|_{\mathcal{G}_k}.$$

Here each  $f \in \mathcal{F}_l$  is of the form  $1_C - 1_{C'}$  where  $P(C \Delta C') \leq 4^{-l+1} = a_l$  and  $\#\mathcal{F}_l \leq (3 \cdot 4^V)^{l-p}$ . Moreover, for some fixed sets  $C_1, \dots, C_k$  in  $\mathcal{C}_l$ ,

$$\begin{aligned} \mathcal{C}_k &= \{C \in \mathcal{C}_l : P(C \Delta C_k) \leq 4^{-q+1}\}, \\ \mathcal{G}_k &= \{1_C - 1_{C_k} : C \in \mathcal{C}_k\}, \quad k = 1, \dots, m, \end{aligned}$$

and  $m \leq (3 \cdot 4^V)^{q-p}$ . To bound the terms involving a supremum over  $\mathcal{F}_l$ , we use Theorem 2.14.24 (in the form given in the proof of Lemma 2.14.26; or even Bernstein's inequality) to show that for  $f \in \mathcal{F}_l$ ,

$$\mathbb{P}(|\sqrt{n}\mathbb{G}_n| \geq t) \leq \exp -\varphi_S(t/C) \quad \text{for } t \geq D,$$

with  $C$  as in Theorem 2.14.24,  $D \leq K\sqrt{na_l}$ , and  $S \leq na_l + \sqrt{na_l}$ . Since  $V \geq 1$ , we have  $\log \#\mathcal{F}_l \leq 3(l-p)V$ . Hence if

$$\log \#\mathcal{F}_l \leq 3(l-p)V \leq na_l \leq na_l + \sqrt{na_l} \equiv S$$

and  $na_l \geq 1$  (so that  $\sqrt{na_l} \leq na_l$ ), then we have (see Exercise 2.14.10)

$$(2.14.35) \quad \begin{aligned} \mathbb{E}\|\mathbb{G}_n\|_{\mathcal{F}_l} &\leq K \left( \sqrt{na_l} + \sqrt{na_l} \sqrt{\log \#\mathcal{F}_l} \right) \\ &\leq K\sqrt{na_l} \sqrt{(l-p)V}. \end{aligned}$$

Hence the contribution of the first term of (2.14.34) is bounded by a constant times  $\sqrt{V} \sum_{l \geq p+1} 2^{-l} \sqrt{l-p} \leq K\sqrt{V4^{-p}}$ .

To bound the term involving suprema over  $\mathcal{G}_k$ , let  $a_q = 4^{-q+1}$ , and note that

$$\left\| \sum_{i=1}^n \varepsilon_i (1_C(X_i) - 1_{C_k}(X_i)) \right\|_{\mathcal{C}_k} \sim \left\| \sum_{i=1}^n \varepsilon_i 1_{C \Delta C_k}(X_i) \right\|_{\mathcal{C}_k}.$$

Hence by Lemma 2.3.6, Theorem 2.14.24, and Problem 2.14.8, the variables

$$Z_k = \left\| \sum_{i=1}^n (1_C - 1_{C_k})(X_i) - n(P(C) - P(C_k)) \right\|_{\mathcal{C}_k}$$

satisfy the bound of Problem 2.14.10 with  $C$  as in Theorem 2.14.24,  $r = na_q + D$ , and

$$D = K\sqrt{nV} \left[ \left( a_q + \frac{V}{n} \log \frac{K}{a_q} \right) \log \frac{K}{a_q} \right]^{1/2}.$$

Thus, by Exercise 2.14.10, it follows that

$$\begin{aligned} \mathbb{E}^* \sup_{1 \leq k \leq m} \|\mathbb{G}_n\|_{\mathcal{G}_k} &\leq \frac{K}{\sqrt{n}} \left[ D + C \sqrt{(na_q + D)(q-p)V} \right] \\ &\leq K \left[ \frac{D}{\sqrt{n}} + \sqrt{V(q-p) \frac{D}{\sqrt{n}}} + \sqrt{V(q-p)a_q} \right], \end{aligned}$$

provided that  $3(q-p)V \leq na_q + D$ . Since  $a_q \leq 4 \cdot 4^{p-q}$ , we have  $\log(K/a_q) \geq (q-p)/K_0$ , and hence

$$V(q-p) \vee \sqrt{V(q-p)a_q} \leq K_1 \frac{D}{\sqrt{n}}.$$

Therefore, it follows that

$$\begin{aligned} \mathbb{E}^* \sup_{1 \leq k \leq m} \|\mathbb{G}_n\|_{\mathcal{G}_k} &\leq K_2 \frac{D}{\sqrt{n}} \\ (2.14.36) \quad &\leq K_3 \left[ \sqrt{Vq4^{-q}} + \sqrt{V4^{-q} \log K} \right. \\ &\quad \left. + n^{-1/2}(Vq + V \log K) \right]. \end{aligned}$$

Combining the inequalities (2.14.35) and (2.14.36) with (2.14.34) completes the proof of (2.14.33).

Now we need to choose  $q \geq p$  so that

$$(2.14.37) \quad 3(q-p) \leq n4^{-q}$$

and also bound the sum of the second and third terms in the exponential in the conclusion of Theorem 2.14.27. Namely, after using (2.14.33) and  $4^{-p} \leq V/t^2$ , we need to bound

$$Q = \tilde{K}t \left[ \sqrt{Vq4^{-q}} + \sqrt{V4^{-q} \log K} + n^{-1/2}(Vq + V \log K) \right] - \frac{t^4}{4n}.$$

To do this, let  $q$  be the largest integer so that (2.14.37) holds. Then

$$3(q-p)4^{q-p} \leq \frac{n}{V}4^{-p} \leq \frac{n}{V} \frac{V}{t^2} = \frac{n}{t^2}$$

and hence  $q-p \leq K_4 \log(K_4 n/t^2)$  for some constant  $K_4$ . Furthermore, by the definition of  $q$ ,

$$n4^{-q-1} \leq 3(q+1-p)V \leq 3qV,$$

and hence, using  $\sqrt{q \log K} \leq 2(q + \log K)$ , it follows that

$$\begin{aligned} Q &\leq \tilde{K} \frac{t}{\sqrt{n}} \left[ 2Vq + \sqrt{3V^2 q \log K} + V \log K \right] - \frac{t^4}{4n} \\ &\leq K_5 \frac{tV}{\sqrt{n}} [q + \log K] - \frac{t^4}{4n} \\ &= K_5 \frac{tV}{\sqrt{n}} q - \frac{t^4}{8n} + K_5 \frac{tV}{\sqrt{n}} \log K - \frac{t^4}{8n} \\ &\equiv Q_1 + Q_2. \end{aligned}$$

But from the definition of  $p$  it follows that  $4^{-(p-1)+1} \geq V/t^2$  and hence  $4^{-p} \geq V/16t^2$  or  $p \leq K_6 \log(16t^2/V)$ . Then, by writing  $q = p + (q - p)$ , it follows that, for  $t \geq \tilde{K}\sqrt{V \log K}$ , we have

$$q \leq K_6 \log\left(\frac{16t^2}{V}\right) + K_4 \log\left(\frac{K_4 n}{t^2}\right) \leq K_7 \log\left(\frac{K_7 n}{V}\right),$$

and this yields

$$\begin{aligned} Q_1 &\leq K_8 \frac{tV}{\sqrt{n}} \log\left(\frac{K_7 n}{V}\right) - \frac{t^4}{8n} \leq \sup_{0 \leq t \leq \sqrt{n}} K_8 \frac{tV}{\sqrt{n}} \log\left(\frac{K_7 n}{V}\right) - \frac{t^4}{8n} \\ &\leq K_9 V. \end{aligned}$$

Furthermore, if  $t \leq \sqrt{n}/\log K$ , then  $Q_2 \leq K_5 V$ , while

$$\sup_{n \geq 1} Q_2 \leq K_{10} \frac{V^2}{t^2} (\log K)^2,$$

so that  $Q \leq K_{11} V$  if  $t \leq \sqrt{n}/\log K$ , and

$$Q \leq K_{12} \left[ V + \frac{V^2}{t^2} (\log K)^2 \right]$$

in any case. Thus, it follows from Theorem 2.14.27 and our choices for  $a$ ,  $p$ , and  $q$  that

$$\begin{aligned} \mathbb{P}^*(\|\mathbb{G}_n\|_{C_1} \geq t) &\leq \mathbb{P}^*\left(\max_{1 \leq j \leq M} \|\mathbb{G}_n\|_{D_j} \geq t\right) \\ &\leq \sum_{j=1}^M \mathbb{P}^*(\|\mathbb{G}_n\|_{D_j} \geq t) \\ &\leq M \frac{K}{t} e^{-2t^2} \exp\left(Kat^2 + K\mu_n t - \frac{t^4}{4n}\right) \\ &\leq 2 \left(\frac{Kt^2}{V}\right)^V \frac{K}{t} e^{-2t^2} \exp\left(KV + K_{12}V + K_{12} \frac{V^2}{t^2} (\log K)^2\right) \\ &\leq K_{13} \left(\frac{Kt^2}{V}\right)^V \frac{1}{t} e^{-2t^2} \exp(3K_{14}V) \\ &\leq \frac{D}{t} \left(\frac{DKt^2}{V}\right)^V e^{-2t^2} \end{aligned}$$

if  $t \geq \sqrt{V \log K}$  or  $n \geq t^2(\log K)^2$ . Choosing  $D = D(K)$  so large that the last bound  $\geq 1$  for  $t \leq \sqrt{V \log K}$  and combining with (2.14.29) completes the proof of Theorem 2.14.13 when (2.14.11) holds.

## Problems and Complements

1. (**Alternative definition of  $\|\cdot\|_{\psi_p}$  for small  $p$** ) For  $0 < p < 1$ , define

$$\psi_p(x) = e^{-c_p^p} (e^{x^p} - 1); \quad \tilde{\psi}_p(x) = (\psi_p(x) - \psi_p(c_p)) 1\{x \geq c_p\}.$$

Then  $\tilde{\psi}_p$  is convex and there exists a constant  $C_p$  such that  $\|X\|_{\tilde{\psi}_p} \leq \|X\|_{\psi_p} \leq C_p \|X\|_{\tilde{\psi}_p}$ . Thus  $\|X\|_{\tilde{\psi}_p}$  defines a true Orlicz norm and can be used interchangeably with  $\|X\|_{\psi_p}$  up to constants. A bound on the latter pseudonorm leads to the tail bound

$$P(|X| > t) \leq (1 + e^{c_p^p}) e^{-t^p / \|X\|_{\psi_p}^p}.$$

[**Hint:** Note that  $\tilde{\psi}_p \leq \psi_p \leq \tilde{\psi}_p + \psi_p(c_p)$ . Also  $E\tilde{\psi}_p(Y/C) \leq E\tilde{\psi}_p(Y)/C$  for every  $C > 1$  and  $Y$  by the convexity of  $\tilde{\psi}_p$ .]

2. (**Hoeffding's inequality**) If the random variable  $X$  has zero mean and takes its values in the interval  $[u, v]$ , then  $E \exp(sX) \leq \exp(s^2(v-u)^2/8)$  for every  $s \in \mathbb{R}$ .

3. If  $\psi: [q_0, \infty) \mapsto \mathbb{R}$  is convex and increasing, then

$$\sum_{q > q_0} \exp -\psi(q) \leq \psi'(q_0)^{-1} \exp -\psi(q_0).$$

Here  $\psi'$  is the right derivative of  $\psi$ .

[**Hint:** The tail series can be bounded by  $\psi'(q_0)^{-1} \int_{q_0}^{\infty} \exp -\psi(x) d\psi(x)$ .]

4. Let  $\mathcal{F}$  be a class of measurable functions  $f: \mathcal{X} \mapsto [l, u]$  satisfying one of the following two conditions:

$$\sup_Q N(\varepsilon(u-l), \mathcal{F}, L_2(Q)) \leq \left(\frac{K}{\varepsilon}\right)^V,$$

$$\sup_Q \log N(\varepsilon(u-l), \mathcal{F}, L_2(Q)) \leq K \left(\frac{1}{\varepsilon}\right)^W,$$

for every  $\varepsilon > 0$ , where the supremum is taken over all discrete probability measures  $Q$ . Then the tail probabilities  $P^*(\|\mathbb{G}_n\|_{\mathcal{F}} > t)$  can be bounded as in the theorems, but with  $t$  replaced by  $t/(u-l)$ .

5. For the function  $\varphi_r$  as in Lemma 2.14.26, there is an absolute constant  $K = K(C)$  such that for all  $t \leq K(C)\sqrt{r}$ ,

$$\varphi_r\left(\frac{K(C)\sqrt{rt}}{C}\right) \geq 11t^2.$$

[**Hint:** Consider the two regions  $K(C)\sqrt{rt} \geq rC$  and  $K(C)\sqrt{rt} \leq rC$ . On the first region,  $11C^2 \exp(11C^2 - 1)$  works for  $K^2(C)$ , while  $11C^2$  works for the second region. Hence  $K^2(C)$  can be chosen to be the larger of these two constants.]

6. Let  $Y_1, \dots, Y_k$  and  $X_1, \dots, X_n$  be i.i.d. samples from  $P_1$  and  $P$ , respectively, that are related by  $P_1(C) = P(C \cap C_0)/P(C_0)$ . If  $\sigma_0^2 \leq P(C_0) \leq 1 - \sigma_0^2$ , then for any class of functions  $\mathcal{F}$

$$\mathbb{E}^* \left\| \sum_{i=1}^k (f(Y_i) - P_1 f) \right\|_{\mathcal{F}} \leq K \mathbb{E}^* \left\| \sum_{i=1}^n (f(X_i) - P f) \right\|_{\mathcal{F}},$$

for a constant  $K$  that depends only on  $\sigma_0^2$ .

7. Suppose that  $x_1, \dots, x_n$  are points in a set  $\mathcal{X}$  and  $\varepsilon_1, \dots, \varepsilon_n$  i.i.d. Rademacher variables. If  $\mathcal{C}$  is a class of subsets of  $\mathcal{X}$  satisfying (2.14.11), then there exists a constant  $\tilde{K}$  such that, with  $N = \left\| \sum_{i=1}^n 1_C(x_i) \right\|_{\mathcal{C}}$ ,

$$\mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i 1_C(x_i) \right\|_{\mathcal{C}} \leq \tilde{K} \sqrt{NV \log(Kn/N)}.$$

[Hint: Use Corollary 2.2.8 with  $d(C, D) = \sqrt{n\|1_C - 1_D\|_{L_1(Q_n)}}$ , where  $Q_n$  is the empirical measure on the points  $x_i$ . The diameter of  $\mathcal{C}$  for  $d$  is at most  $\sqrt{2N}$ .]

8. Suppose that  $X_1, \dots, X_n$  are i.i.d. random elements with distribution  $P$  on  $(\mathcal{X}, \mathcal{A})$ , and let  $\varepsilon_1, \dots, \varepsilon_n$  i.i.d. Rademacher variables. If  $\mathcal{C}$  is a measurable class of measurable subsets of  $\mathcal{X}$  satisfying (2.14.11), then there exists a constant  $\tilde{K}$  such that, with  $a = \sup_{C \in \mathcal{C}} P(C)$ ,

$$\mathbb{E}^* \left\| \sum_{i=1}^n 1_C(X_i) \right\|_{\mathcal{C}} \leq 2na + \tilde{K}V \log \frac{K}{a}$$

and

$$\mathbb{E}^* \left\| \sum_{i=1}^n \varepsilon_i 1_C(X_i) \right\|_{\mathcal{C}} \leq \tilde{K} \sqrt{Vn} \left[ \left( a + \frac{V}{n} \log \frac{V}{a} \right) \log \frac{K}{a} \right]^{1/2}.$$

9. Suppose that  $\mathcal{C}$  is a class of sets satisfying (2.14.11) and  $\sup_{C \in \mathcal{C}} P(C) \leq \sigma_0^2$ . Then for some constants  $D$  and  $\tilde{K}$

$$\mathbb{P}^* (\|\mathbb{G}_n\|_{\mathcal{C}} > t) \leq \left( \frac{D}{t} \right) e^{-2t^2},$$

for all  $t \geq \tilde{K} \sqrt{V \log K}$  and  $n \geq 1$ .

[Hint: Use Lemma 2.14.26 and the second result proved in Exercise 2.14.8.]

10. Suppose that  $Z_1, \dots, Z_m$  are random variables satisfying

$$\mathbb{P}(|Z_j| \geq t) \leq \exp -\varphi_r(t/C),$$

for all  $t \geq D$ , where  $\varphi_r(t)$  is the function defined in the proof of Lemma 2.14.26. Then there exists a constant such that, for  $\log m \leq r$ ,

$$\mathbb{E} \max_{1 \leq j \leq m} |Z_j| \leq K \left( D + C \sqrt{r \log m} \right).$$

[**Hint:** Recall the proof of Lemma 2.2.2; compute the left side as the integral  $\int_0^\infty P(\max_{1 \leq j \leq m} |Z_j| > t) dt$ ; and split the region of integration at  $\tau \equiv D \vee C\sqrt{r \log m}$ .]

11. Suppose that  $\mathcal{C}$  is a class of subsets of a given set  $C_0$  satisfying (2.14.12). There exists a constant  $K_1$  such that if  $P(C_0) = d \in (0, 1)$ , then

$$E^* \|\mathbb{G}_n\|_{\mathcal{C}} \leq K_1 \sqrt{V d \log \left( \frac{K}{d} \right)}.$$

[**Hint:** Relate  $N_{[]}(\varepsilon, \mathcal{C}, L_1(P))$  to  $N_{[]}(\varepsilon, \mathcal{C}, L_2(P))$ , and then use Theorem 2.14.2; compare with Exercise 2.14.7.]

## 2

# Notes

**2.2.** The maximal inequality Theorem 2.2.4 uses the special properties of the Orlicz norm only through the application of Lemma 2.2.2. Hence an identical result can be proved using other norms provided the analogue of the lemma is available. For instance, Ledoux and Talagrand (1991) show that the lemma holds for weak  $L_p$ -norms with  $\psi$  taken equal to  $x^p$ . Second, it may be remarked that often the full strength of Theorem 2.2.4 is not needed and the weakened version of this theorem, where the Orlicz norm on the left is replaced by  $\text{Esup}_{s,t} |X_s - X_t|$ , suffices. This weakened version can be proved along exactly the same lines, provided a similarly weakened version of Lemma 2.2.2 is available. The latter is contained in Problem 2.2.8 for *any* Orlicz norm, a result due to Pisier (1983).

The use of “entropy” in the study of the continuity of Gaussian processes goes back to Dudley (1967b) and Sudakov (1969). Dudley (1967b) credits V. Strassen with the idea and basic results, but see the review of Sudakov (1976) by Dudley (1978b). Strassen and Dudley (1969) apparently made the first use of entropy in connection with empirical processes. The use of entropy in maximal inequalities will probably persist because of its simplicity. However, it is known that entropy inequalities are not sharp in all cases, while a stronger tool, majorizing measures, gives sharp results, at least for Gaussian processes. See Ledoux and Talagrand (1991) and Appendix A.2.3 for an exposition.

The original papers by Bernstein are apparently unavailable. Bennett (1962) discusses the history of Bernstein’s inequality, provides proofs, and compares Bernstein’s inequality to a number of other inequalities.

**2.3.** The method of symmetrization goes back to Kahane (1968) and Hoffmann-Jørgensen (1973, 1976). The construction of separable versions using liftings is adapted from Talagrand (1987a).

**2.4.** Generalizations of the classical Glivenko (1933) and Cantelli (1933) theorem for sets under a bracketing condition were first obtained by Blum (1955) and Dehardt (1971). Dudley (1984) gives Theorem 2.4.1 in its present formulation. Vapnik and Červonenkis (1968, 1971, 1981) give generalizations with conditions in terms of the logarithm of the  $L_1$ -covering numbers. Pollard (1982) formulates the reverse submartingale property of  $\|\mathbb{P}_n - P\|_{\mathcal{F}}$ . Measurability difficulties regarding the choice of filtration are pointed out by Strobl (1992). For more on Glivenko-Cantelli theorems, see Chapter 2.8 and Dudley, Giné, and Zinn (1991).

**2.5.** The first uniform entropy central limit theorems were due to Pollard (1982) and Kolčinskii (1981) following path-breaking work by Dudley (1978a), who, among other things, established the uniform central limit theorem for the empirical process indexed by a (suitably measurable) VC-class of sets. Bracketing empirical central limit theorems were obtained by Dudley (1978a, 1984), Ossiander (1987) and Andersen, Giné, Ossiander, and Zinn (1988). Ossiander (1987) obtained the result under the assumption that the  $L_2(P)$ -bracketing integral is finite. The result presented in this section is more general than this, but less general than the theorems obtained by Andersen et al. These authors relax finiteness of the two integrals to continuous majorizing measure conditions on  $L_2(P)$ -balls and  $L_{2,\infty}(P)$ -brackets, respectively. See Chapter 2.11 for a statement.

The chaining argument in the proof of Theorem 2.5.6 may be refined so that it is not necessary to construct a *nested* sequence of partitions. See Pollard (1989a). The present organization and notation of the chaining argument is borrowed from Arcones and Giné (1993), who chain by exponential bounds for probabilities. The use of means in combination with Lemma 2.2.10 in the proof is new.

**2.6.** The study of VC-classes apparently originated in Vapnik and Červonenkis (1971), who were motivated by problems in pattern recognition. See also Vapnik and Červonenkis (1981) and Vapnik (1982). Sauer's lemma and its generalizations can be found in Sauer (1972) and Frankl (1983). Theorem 2.6.4 is due to Haussler (1995), who improved an earlier result of Dudley (1978a). Our rendering of the proof owes its origins to conversations with David Pollard. Pollard (1984) showed how to obtain the result about functions from the bound for sets.

The behavior of entropies under the formation of convex hulls as described in Theorem 2.6.9 is an extension of results of Dudley (1987), who obtained the bound up to an additional  $\delta$ , and Ball and Pajor (1990), who obtained the bound under the assumption that  $\mathcal{F}$  is a sequence of functions decreasing at the rate  $\|f_i\|_{Q,2} \leq i^{-1/V}$ . The present formulation and

its induction proof are new. Lemma 2.6.11 is given by Pisier (1981) but is apparently due to B. Maurey.

Many results in Section 2.6.5 are due to Assouad (1981, 1983), Pollard (1984), and Dudley (1984). Assouad (1981, 1983) gives explicit bounds for the VC-indices of many of the classes in Lemma 2.6.17 and formulates a number of results concerning “dual density,” which we have not included here (see especially Assouad (1983), pages 245–247). Lemma 2.6.22 and Example 2.6.23 improve upon an example of Dudley (1985), who shows that the class of functions  $x \mapsto t^{k-1}F(x)\mathbf{1}\{F(x) \geq t\}$  is VC-hull. We learned Exercise 2.6.21 from A. Quiroz: see Quiroz, Nakamura, and Perez (1995).

Notable among the results not included in this section are those of Stengle and Yukich (1989) and Pisier (1984). Stengle and Yukich show how new VC classes can be generated by “semialgebraic sets.” Pisier (1984) links up VC theory with probability in Banach spaces by showing that a class of sets is VC if and only if a certain operator is type-2. See Ledoux and Talagrand (1991) for an exposition.

**2.7.** Bounds on entropies for classes of smooth functions are obtained by Kolmogorov and Tikhomirov (1961), Lorentz (1966), and Birman and Solomjak (1967). Dudley (1984), Chapter 7, gives a useful introduction to this literature. We adapted the proof of the bound for entropy with bracketing for monotone functions on  $\mathbb{R}$  from Van de Geer (1991), who adapts techniques from Birman and Solomjak (1967). The results for convex sets are due to Bronštein (1976), who improved on results of Dudley (1974). These bounds were first exploited in the context of empirical processes by Bolthausen (1978). Van der Vaart (1994b) gives improvements of Corollary 2.7.4 and shows that the classes of functions considered there are  $P$ -Glivenko-Cantelli if and only if  $\sum M_j P(I_j) < \infty$ . He also shows that the bracketing integral converges if the constants  $M_j P(I_j)^{1/2}$  are regularly varying at infinity and  $\sum M_j P(I_j)^{1/2} < \infty$ . Giné and Zinn (1986b) first proved that the class  $C_1^1(\mathbb{R})$  is Donsker under the latter condition. Van der Vaart (1993) generalized their result to the classes of Corollary 2.7.4 in general.

For more recent work on entropy numbers, see Carl and Stephani (1990) and Ball and Pajor (1990) and the references stated therein. These authors study the entropy of the image of the unit ball under a (compact) operator  $T$  between Banach spaces. More precisely, they study entropy numbers denoted  $\varepsilon_n(T)$  and  $e_n(T)$ , which as functions of  $n$  are roughly the inverses of the covering numbers and entropy numbers as in the present manuscript. For instance,  $e_n(T)$  is defined as the smallest  $\varepsilon$  such that the image  $T(U)$  can be covered by  $2^{n-1}$  balls of radius  $\varepsilon$ .

**2.8.** Theorem 2.8.1 is due to Dudley, Giné, and Zinn (1991), who give a complete treatment of the conditions under which a class of functions is universal or uniform Glivenko-Cantelli.

The development of uniform in  $P$  Glivenko-Cantelli and Donsker theorems began with the work of Dvoretzky, Kiefer, and Wolfowitz (1956), Kiefer and Wolfowitz (1959), and Kiefer (1961). The exponential bounds developed in these papers yield uniform in  $P$ -Glivenko-Cantelli theorems for the classical empirical distribution function of random variables in  $\mathbb{R}^d$ . Moreover, these authors essentially establish the uniform Donsker property for the classical distribution function in the course of proving the asymptotic minimaxity of the empirical distribution function as an estimator of the true distribution function. Uniformity in  $P$  is implicit in the work of Vapnik and Červonenkis (1971) on the Glivenko-Cantelli theorems. The introduction of uniform entropy conditions by Pollard (1982) and Kolčinskii (1981) continued this development. Massart (1986), Theorem 5.10, page 4.11, establishes rates of convergence for weak-approximation versions of Theorem 2.8.3 under growth rate hypotheses on the uniform entropy and additional hypotheses on the envelope function of the class. (It is plausible that his rates are in fact uniform over the collection  $\mathcal{P}$  for which the moment hypotheses on  $\mathcal{F}$  hold uniformly.) Dudley (1987) unifies and extends results for the “universal Donsker property,” which entail that  $\mathcal{F}$  is Donsker for all probability measures  $P$  on the sample space. A universal Donsker class  $\mathcal{F}$  is essentially bounded:  $\sup_f \text{diam } f < \infty$  (where  $\text{diam } f = \sup_{x,y} |f(x) - f(y)|$ ). For classes with this property, Giné and Zinn (1991) use Gaussian comparison methods to characterize classes that are uniformly Donsker in all possible underlying measures.

Theorem 2.8.2 is contained in Sheehy and Wellner (1992) along with other equivalences; the present proof is new.

**2.9.** The multiplier central limit, Theorem 2.9.2, is implicit in Giné and Zinn (1984) and is stated explicitly in Giné and Zinn (1986a), who attribute the multiplier inequality to Pisier and Fernique. Alexander (1985), solving a problem posed by Hoffmann-Jørgensen, shows that no “universal multiplier moment” exists: there is no function  $\psi: \mathbb{R}^2 \mapsto \mathbb{R}$  so that  $\xi Z$  satisfies the central limit theorem whenever  $E\xi Z = 0$ ,  $E\psi(|\xi|, \|Z\|) < \infty$  and  $Z$  satisfies the central limit theorem, for independent real- and Banach-space-valued random elements  $\xi$  and  $Z$ . On the other hand, Ledoux and Talagrand (1986) show that the  $L_{2,1}$ -hypothesis on the multipliers  $\xi_i$  cannot be relaxed in the sense that, for every  $\xi$  with  $\|\xi\|_{2,1} = \infty$ , there exists a Banach space valued  $Z$  that satisfies the central limit theorem, but  $\xi Z$  does not satisfy the central limit theorem. Ledoux and Talagrand (1986) and Ledoux and Talagrand (1991), Proposition 10.4 on page 279, give a different proof of the basic multiplier inequality.

The almost sure conditional multiplier central limit theorem is due to Ledoux and Talagrand (1988), with contributions by J. Zinn. The present proof, based on isoperimetric methods, is adapted from Ledoux and Talagrand (1991), Theorem 10.14, on page 293. The proof using martingale

difference methods originating in Yurinskii (1974), given in the “Problems and Complements” section, is taken from Ledoux and Talagrand (1988).

**2.10.** Alexander (1987c) gives Theorem 2.10.1 and derives that  $\mathcal{F} + \mathcal{G}$  and  $\mathcal{F} \cup \mathcal{G}$  are Donsker if  $\mathcal{F}$  and  $\mathcal{G}$  are Donsker from his more general Proposition 2.6. Theorem 2.10.2 and Theorem 2.10.3 are due to Dudley (1985).

Versions of Theorem 2.10.6 are established by Giné and Zinn (1986a) (one-dimensional  $\phi$  under measurability conditions) and Talagrand (1987a) (uniformly bounded classes). Lemma 2.10.14 is established by Giné and Zinn (1986a); the proof given here is new. Section 2.10.4 is based on Van der Vaart (1993). The result of Example 2.10.25 specialized to the class  $C_1^1(\mathbb{R})$  was first obtained by Giné and Zinn (1986b) by a different method. Pollard (1990), Section 5, gives a variety of results similar to those in Section 2.10.3.

**2.11.** The limit theory in this section has its antecedents in Koul (1970), Shorack (1973, 1979), and Van Zuijlen (1978). Also see Shorack and Beirlant (1986), Marcus and Zinn (1984), and Shorack and Wellner (1986), Section 3.3, pages 108–109.

The first subsection, based on random and uniform-entropy, follows Alexander (1987b), but with a simplified proof for the main Theorem 2.11.1. See Alexander’s paper for a discussion of the minimal moment conditions on the envelope within the context of this theorem. The improvement over Theorem 2.5.2 in the i.i.d. case was already obtained by Giné and Zinn (1984). Pollard (1990) develops a combinatorial theory for the non-i.i.d. case paralleling the VC-theory of Chapter 2.6.

The bracketing central limit theorem, Theorem 2.11.11, and its corollary are due to Andersen, Giné, Ossiander, and Zinn (1988), building on Ossiander (1987) and work on majorizing measures and Gaussian processes due to Fernique (1974) and Talagrand (1987c). Theorem 2.11.9 does not seem to be a corollary of their result, but appears to be the natural generalization of Ossiander’s result to the case of non-i.i.d. observations. The proof of Theorem 2.11.11 shows that the theorem remains true if the single majorizing measure, which is guaranteed to exist by the assumption of Gaussian domination, is replaced by certain sequences of majorizing measures (depending on  $n$ ). The proof given here is greatly simplified in comparison to the original proof.

Jain and Marcus proved their theorem, as given here, in Jain and Marcus (1975). Example 2.11.14 is taken from Giné and Zinn (1986a). Dudley (1985), Section 6, gives the first satisfactory integration of the famous “Chibisov-O’Reilly theorem” treated in Example 2.11.15 with modern empirical process theory.

**2.12.** Partial-sum processes have a long and honorable history in probability theory and statistics, with rapid development of the classical theory

occurring in the late 1940s and early 1950s. The classical work of Erdős and Kac (1946), Doob (1949), and Donsker (1951, 1952) was beautifully summarized by Billingsley (1968).

Study of the partial-sum or sequential empirical process seems to have begun with the work of Müller (1968) on the distributional invariance principle for the Glivenko-Cantelli theorem, Pyke (1968) on random-sample-size limit theorems for empirical processes, and Kiefer (1969) on embedding questions. The Hungarian school's work on embedding theorems in the 1970s was followed by embedding theorems (explicit construction) for general empirical processes in Dudley and Philipp (1983). The present form seems to have been first stated by Sheehy and Wellner (1992), along with other equivalences of Dudley and Philipp (1983) and Dudley (1984). The present proof is analogous to the proof by Billingsley (1968) of a classical partial-sum convergence theorem: Theorem 10.1, pages 68–70.

The subsection on partial-sum processes on lattices follows Alexander and Pyke (1986), Bass and Pyke (1985), and Pyke (1983).

**2.13.** Theorem 2.13.1 is taken from Dudley (1984), who also shows that, in a certain sense, the condition that the sequence converges is sharp. Giné and Zinn (1986a) give further results for sequences of functions bounded in uniform norm.

Dudley (1987) notes that the square of the supremum of the empirical process over an elliptical class  $\mathcal{F}$  equals a weighted sum of the squares of the empirical process acting on the basis functions defining the elliptical class, and he points out the practical importance of such classes.

The equivalence of (iii) and (iv) in Theorem 2.13.6 is due to Giné and Zinn (1984); the equivalence of (i), (ii), and (iii) was proved by Talagrand (1988).

**2.14.** Bounds for expectations of suprema of Gaussian processes were developed by Dudley (1967b). In the case of empirical processes, Pollard (1989b) was apparently the first to recognize the usefulness of moment bounds for suprema of the form given in Theorems 2.14.1 and 2.14.2; also see Pollard (1990) and Kim and Pollard (1990). Ledoux and Talagrand (1991) systematically developed the use of Orlicz norms; see their Notation and Chapters 6 and 11. Theorem 2.14.5 is a straightforward consequence of the general domination of Orlicz norms of sums of independent random elements by  $L_1$ -norms plus the Orlicz norm of maximal summand proved by Talagrand (1989); see Ledoux and Talagrand (1991), Chapter 6.

Dvoretzky, Kiefer, and Wolfowitz (1956) proved the first exponential bound for the supremum distance between the empirical distribution function and the true distribution function in the case of  $\mathcal{X} = \mathbb{R}$ . Massart (1990) found the best constant  $D$  multiplying the exponential for this classical case. Related bounds for the more difficult case  $\mathcal{X} = \mathbb{R}^k$  were established by Kiefer (1961). Vapnik and Červonenkis (1971) proved exponential bounds

of the same form as those given in Theorem 2.14.13, but with no powers of  $t$  in front and the constant  $D$  depending on (and increasing with)  $n$ . Alexander (1984) gives bounds of the same type as those in Theorems 2.14.9, 2.14.13, and 2.14.16 and also has sharper bounds than the bound given in 2.14.16.

Theorems 2.14.9, 2.14.13, and 2.14.14 are due to Talagrand (1994), while Theorems 2.14.10, 2.14.16, and 2.14.17 are from Massart (1986). For still more bounds, see Smith and Dudley (1992) and Adler and Brown (1986).

The developments in Subsection 2.14.3 and Exercises 2.14.5 - 2.14.11 are based on Talagrand (1994) with one exception: the reformulation of Theorem 2.14.24 given in Theorem 2.14.25 is due to Birgé and Massart (1994).

PART 3

# Statistical Applications

## 3.1

# Introduction

The empirical process methods and techniques developed in Part 2 have many applications in statistics. The present part illustrates this in some detail by applications ranging from  $M$ -estimation (limit theory and rates of convergence in infinite-dimensional applications), bootstrapping, permutation tests, tests of independence, the functional delta-method, and continuity theory.

In agreement with its importance in statistics, a treatment of  $M$ -estimators opens this part. In Chapter 3.2 we consider estimators  $\hat{\theta}_n$  defined as the value of  $\theta$  maximizing (or nearly maximizing) a random criterion function  $\theta \mapsto \mathbb{M}_n(\theta)$ . A number of results are formulated for general, abstract criterion functions, but we are particularly interested in “empirical criterion functions” of the form  $\mathbb{M}_n(\theta) = \mathbb{P}_n m_\theta$ .

The limiting distribution of such  $M$ -estimators may be derived either from a continuous mapping theorem for the argmax functional or from a linearization argument. In both approaches it appears useful to split the derivation into three steps:

- establish consistency;
- establish a rate of convergence;
- derive the limiting distribution.

The most successful methods in the first step are the (generalized) method of Wald (which we do not discuss) and a method based on uniform convergence of the criterion functions. The latter method, when applied with a empirical criterion function, boils down to proving that certain classes of functions are Glivenko-Cantelli and is discussed briefly.

Rates of convergence are discussed in some detail and generality in Chapter 3.4, but in Chapter 3.2 we present a preliminary approach, which appears to be useful for the study of  $M$ -estimators of Euclidean parameters. When applied to empirical criterion functions of the form  $\theta \mapsto \mathbb{P}_n m_\theta$  this approach is based on controlling the entropy of the classes of functions  $\{m_\theta - m_{\theta_0} : \|\theta - \theta_0\| < \delta\}$  (for small  $\delta$ ) and expresses the rate of convergence in terms of the  $L_2$ -norms of the envelope functions of these classes.

The argmax continuous mapping theorem asserts that given a sequence of criterion functions  $\mathbb{M}_n$  converging (in an appropriate distributional sense) to a limit  $\mathbb{M}$ , the point of maximum, or argmax, of  $\mathbb{M}_n$ , which is the  $M$ -estimator, converges in distribution to the argmax of the limit process  $\mathbb{M}$ . In obtaining a limit distribution of a sequence of  $M$ -estimators, this theorem is usually not applied with the original criterion functions  $\theta \mapsto \mathbb{M}_n(\theta)$ , but to (a multiple of) “localized” criterion functions of the form

$$h \mapsto \tilde{\mathbb{M}}_n(h) = \mathbb{M}_n\left(\theta_0 + \frac{h}{r_n}\right) - \mathbb{M}_n(\theta_0).$$

Here  $r_n$  is the rate of convergence and  $\theta_0$  the true parameter. These localized criterion functions usually converge functionally in distribution to a limit only if the local parameter  $h = r_n(\hat{\theta} - \theta_0)$  is restricted to compacta. This is the reason that it is necessary to obtain the rate of convergence first. Once it is known that  $\hat{h}_n = r_n(\hat{\theta}_n - \theta_0)$  is uniformly tight, convergence in distribution of the processes  $\{\mathbb{M}_n(h) : h \in K\}$  in the space  $\ell^\infty(K)$  for each compact  $K$  suffices for successful application of the argmax theorem. Actually, this type of convergence is too strong, but it fits well within the framework of empirical processes developed in Part 2.

Using this approach, we obtain simple, but general, conditions for the asymptotic normality of maximum likelihood estimators in smooth parametric models, but we also treat a number of nonstandard problems.

An alternative approach to obtaining the limit distribution of  $M$ -estimators is to derive a characterization as the solution of a family of estimating equations of the form  $\Psi_n(\hat{\theta}_n) = 0$ . In Chapter 3.3 we discuss the asymptotic behavior of solutions of estimating equations, which we call  $Z$ -estimators, in general. The framework in this chapter allows an infinite-dimensional parameter space  $\Theta$  and a corresponding infinite-dimensional system of estimating equations. This covers not only the classical estimating equation framework with a finite-dimensional parameter space, but also a wide range of problems in nonparametric and semiparametric inference involving estimation of parameters with values in some (Banach) space of functions.

Establishing a rate of convergence of  $M$ -estimators based on empirical criterion functions is quite amenable to empirical process methods. Chapter 3.4 develops these methods in some detail and applies them to maximum likelihood estimation, least-squares estimation, and least-absolute-deviation estimation. In this chapter we are mostly interested in infinite-dimensional

(nonparametric or semiparametric) estimation. The basic result is that a modulus of continuity of the empirical process determines the rate of convergence. Using the results of Part 2, this modulus can be bounded in terms of entropies, and the rate of convergence can be expressed in terms of entropy integrals of the parameter set. For instance, the rate of convergence  $r_n$  of a maximum likelihood estimator can be found as the solution of the equation

$$r_n^2 \tilde{J}_{[]} \left( \frac{1}{r_n}, \mathcal{P}_n, h \right) \leq \sqrt{n}.$$

Here  $\mathcal{P}_n$  is the class of densities over which the likelihood is maximized and the bracketing entropies are calculated with respect to the Hellinger distance  $h$ .

In applications it frequently occurs that the sample size is not fixed but is, in fact, random—perhaps even depending on the data  $X_1, \dots, X_n$  in some complicated way. The effect of random sample size on the empirical processes  $\mathbb{G}_n$  is investigated in Chapter 3.5. Theorem 3.5.1 asserts that if  $\mathcal{F}$  is a Donsker class of functions and  $N_n$  is a sequence of random samples sizes satisfying  $N_n/c_n \xrightarrow{\text{P}} \nu$  with  $c_n \rightarrow \infty$ , then  $\mathbb{G}_{N_n} \rightsquigarrow \mathbb{G}$  in  $\ell^\infty(\mathcal{F})$  provided only that  $\text{P}(\nu > 0) = 1$ . In this chapter we give special attention to the case that the random sample size  $N_n$  is Poisson-distributed with mean  $n$  independent of the data. In this case the *Kac empirical point process*  $\mathbb{N}_n = \sum_{i=1}^{N_n} \delta_{X_i}$  is a Poisson point process with intensity measure  $nP$ , and, for  $\mathcal{F}$  with  $\|P\|_{\mathcal{F}} < \infty$ , the *Kac empirical process*  $\mathbb{Z}_n = n^{-1/2}(\mathbb{N}_n - nP)$  converges in distribution in  $\ell^\infty(\mathcal{F})$  to a Brownian motion process if and only if  $\mathcal{F}$  is *P*-Donsker.

Nonparametric bootstrap methods have become popular in statistics since their introduction by Efron (1979). This method is based on sampling from the empirical measure  $\mathbb{P}_n$ . Given the original sample  $X_1, \dots, X_n$ , let  $\hat{X}_1, \dots, \hat{X}_n$  be an i.i.d. sample from  $\mathbb{P}_n$ . The *bootstrap empirical measure* and the *bootstrap empirical process* are given by

$$\hat{\mathbb{P}}_n = \frac{1}{n} \sum_{i=1}^n \delta_{\hat{X}_i} \quad \text{and} \quad \hat{\mathbb{G}}_n = \sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n),$$

respectively. These are important tools for the study of a wide variety of statistical methods. Letting  $M_{ni}$  be the number of times that  $X_i$  is “redrawn” from the original sample, the bootstrap empirical measure and process can also be written as

$$(3.1.1) \quad \hat{\mathbb{P}}_n = \frac{1}{n} \sum_{i=1}^n M_{ni} \delta_{X_i} \quad \text{and} \quad \hat{\mathbb{G}}_n = \frac{1}{\sqrt{n}} \sum_{i=1}^n (M_{ni} - 1) \delta_{X_i},$$

respectively. The vector  $M_n = (M_{n1}, \dots, M_{nn})$  has a multinomial distribution with  $n$  cells,  $n$  trials, and success probabilities  $1/n$  for each of the  $n$  cells. Since the variables  $M_{n1}, M_{n2}, \dots$  converge in distribution to a sequence of i.i.d. Poisson variables  $Y_1, Y_2, \dots$  with mean 1, it is apparent that

the limit theory for the bootstrap empirical process  $\hat{G}_n$  is closely linked to the conditional multiplier central limit theorems developed in Chapter 2.9. This is indeed the case, and our proofs of two bootstrap limit theorems due to Giné and Zinn (1990) rely on this connection. In Theorem 3.6.1 it is shown that the bootstrap empirical process satisfies  $\hat{G}_n \rightsquigarrow G$  “in outer probability” if and only if  $\mathcal{F}$  is  $P$ -Donsker, while in Theorem 3.6.2 it is shown that  $\hat{G}_n \rightsquigarrow G$  “outer almost surely” if and only if  $\mathcal{F}$  is  $P$ -Donsker and  $P^* \|f - Pf\|_{\mathcal{F}}^2 < \infty$ . The statistical meaning of these results is that the “distribution” of  $\hat{G}_n$  (which depends on the original data only) is a consistent estimator for the distribution of  $G_n$  (which depends on the underlying distribution  $P$ ), in the sense that the difference converges to zero as  $n \rightarrow \infty$ . The two types of theorems yield consistency in probability and consistency in an almost sure sense, respectively. Our proofs rely heavily on the Poissonization of the bootstrap sample size. Other bootstrap methods can be based on replacing the multinomial weights in (3.1.1) by other exchangeable weights. For example, resampling  $k$  from the original sample of  $n$  without replacement can be treated by an appropriate choice of  $M_n$ . Conditional limit theorems for these alternative bootstrap methods are also presented.

Chapter 3.7 introduces general versions of the two-sample Kolmogorov-Smirnov statistics and shows how the corresponding two-sample tests can be implemented using either bootstrap or permutation methods. Many statistics used for testing independence can be viewed as functionals of the *independence empirical process*, which we define and study in Chapter 3.8.

One of the most important basic tools of large sample theory in statistics is the *delta-method*: if  $r_n(X_n - \theta) \rightsquigarrow Z$  for some constant  $\theta$  and  $r_n \rightarrow \infty$ , and  $\phi$  is differentiable at  $\theta$ , then  $r_n(\phi(X_n) - \phi(\theta)) \rightsquigarrow \phi'(\theta)Z$ . For Euclidean vectors  $X_n$ , this result is standard and  $\phi'(\theta)$  is the usual derivative from calculus. Extensions to functions  $\phi$  of infinite-dimensional statistics  $X_n$ , for instance the empirical measure  $\mathbb{P}_n$ , have become increasingly useful and important. The formulation of the *functional delta-method* requires careful definition of the type of derivative. A map  $\phi: \mathbb{D} \mapsto \mathbb{E}$  is called *Fréchet-differentiable* at  $\theta$  if there is a continuous linear map  $\phi'_\theta: \mathbb{D} \mapsto \mathbb{E}$  such that

$$\|\phi(\theta + h) - \phi(\theta) - \phi'_\theta(h)\| = o(\|h\|), \quad \|h\| \rightarrow 0.$$

Such a map  $\phi$  is called *Hadamard-differentiable* at  $\theta$  if there is a continuous linear map  $\phi'_\theta: \mathbb{D} \mapsto \mathbb{E}$  such that

$$\frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} \rightarrow \phi'_\theta(h),$$

for every sequence of real numbers  $t_n \rightarrow 0$  and every sequence  $h_n \rightarrow h$ . Hadamard differentiability is a weaker requirement than Fréchet differentiability and, fortunately, is sufficient for the functional delta-method. This

asserts that if  $r_n(X_n - \theta) \rightsquigarrow Z$  for a separable variable  $Z$  and  $\phi$  is Hadamard-differentiable at  $\theta$ , then

$$r_n(\phi(X_n) - \phi(\theta)) \rightsquigarrow \phi'_\theta(Z).$$

Examples treated in some detail in Chapter 3.9 include the Nelson-Aalen estimator from right-censored data, quantile and copula functions, the product integral, multivariate trimming, and  $M$ -functionals.

When specialized to a Hadamard-differentiable map  $\phi$  on  $\mathbb{D} = \ell^\infty(\mathcal{F})$  for a Donsker class  $\mathcal{F}$ , the functional delta-method gives

$$\sqrt{n}(\phi(\mathbb{P}_n) - \phi(P)) \rightsquigarrow \phi'_P(\mathbb{G}).$$

In this case it makes sense to ask if this continues to hold for the empirical bootstrap: is it true that

$$\sqrt{n}(\phi(\hat{\mathbb{P}}_n) - \phi(\mathbb{P}_n)) \rightsquigarrow \phi'_P(\mathbb{G})$$

in outer probability (or outer almost surely)? Several results asserting that this is true are proved in Section 3.9.3. Thus, the consistency of the bootstrap of the abstract empirical process implies the consistency of the bootstrap for many statistical functionals. This is probably the most important motivation for studying the abstract bootstrap process.

The theory of *contiguous* families of probability measures provides a notion of asymptotic absolute continuity that has proved to be of fundamental importance for asymptotic theory in statistics. Chapter 3.10 presents the basic elements of contiguity theory and also applies it to the empirical process. One result of this type concerns the empirical measure  $\mathbb{P}_n$  of independent and identically distributed variables  $X_{n1}, \dots, X_{nn}$  with common distribution  $P_n$  satisfying

$$\int \left[ \sqrt{n}(dP_n^{1/2} - dP^{1/2}) - \frac{1}{2}h dP^{1/2} \right]^2 \rightarrow 0,$$

for some measurable function  $h: \mathcal{X} \mapsto \mathbb{R}$ . If  $\mathcal{F}$  is  $P$ -Donsker, then under  $P_n$

$$\sqrt{n}(\mathbb{P}_n - P) \rightsquigarrow \mathbb{G} + s_h \quad \text{in } \ell^\infty(\mathcal{F})$$

where the shift  $s_h: \mathcal{F} \mapsto \mathbb{R}$  is given by  $s_h f = Pf h$ .

Part 3 ends in Chapter 3.11 with infinite-dimensional versions of the Hájek convolution and local asymptotic minimax theorems. These results apply to many statistical models, but we discuss particularly the asymptotic efficiency of the empirical distribution in the nonparametric situation.

## Problems and Complements

1. Suppose that for each natural number  $n$  the vector  $M_n = (M_{n1}, \dots, M_{nn})$  has a multinomial distribution with  $n$  cells,  $n$  trials, and success probabilities  $1/n$  for each of the  $n$  cells. Then the variables  $M_{n1}, M_{n2}, \dots$  converge in distribution to a sequence of i.i.d. Poisson variables  $Y_1, Y_2, \dots$  with mean 1 in  $\mathbb{R}^\infty$ . Furthermore, if  $k(n) = o(n)$ ,  $P_{n,k(n)}$  denotes the joint distribution of  $(M_{n1}, \dots, M_{nk(n)})$  and  $Q_{n,k(n)}$  denotes the joint distribution of  $(Y_1, \dots, Y_{k(n)})$ , then the total variation distance from  $P_{n,k(n)}$  to  $Q_{n,k(n)}$  converges to 0, i.e.,  $\sup_A |P_{n,k(n)}(A) - Q_{n,k(n)}(A)| \rightarrow 0$ , for the supremum taken over the Borel sets in  $\mathbb{R}^{k(n)}$ .

[**Hint:** This is closely related to Runnenburg and Vervaat (1969) and to Wellner (1977).]

## 3.2

# M-Estimators

The most important method of constructing statistical estimators is to choose the estimator to maximize a certain criterion function. We shall call such estimators *M-estimators* (from “maximum” or “minimum”). In the case of i.i.d. observations  $X_1, \dots, X_n$ , a common type of criterion function is of the form

$$\theta \mapsto \mathbb{P}_n m_\theta = \frac{1}{n} \sum_{i=1}^n m_\theta(X_i),$$

for known given functions  $m_\theta$  on the sample space. In particular, the method of maximum likelihood estimation corresponds to the choice  $m_\theta = \log p_\theta$ , where  $p_\theta$  is the density of the observations.<sup>†</sup> The theory of empirical processes comes in naturally when studying the asymptotic properties of these estimators. In this chapter we present several results that give the asymptotic distribution of *M-estimators*. Some results are of a general nature, while others presume the set-up of i.i.d. observations.

In many situations estimators that maximize a certain map also solve a system of equations. In particular, in the case of i.i.d. observations many estimators are a zero of a map of the type

$$\theta \mapsto \mathbb{P}_n \psi_\theta,$$

for given maps  $\psi_\theta$  on the sample space. We shall refer to solutions of estimating equations as *Z-estimators* (from “zero”). We note that in some

---

<sup>†</sup> For some purposes, such as for proving consistency or rates of convergence, particularly in the case of models that are convex in the parameter, it is technically convenient to choose a different function.

of the literature the name  $M$ -estimator is (also) used for what we call  $Z$ -estimator and the distinction between the different types of estimators is not always made.

The asymptotic behavior of  $Z$ -estimators is studied in Chapter 3.3. The approach given there provides an alternative to the approach in the present chapter, although generally the asymptotic properties of  $M$ -estimators are most economically studied from their characterization as a point of maximum. In this chapter we discuss two approaches that use this characterization directly.

The natural “parameter” (or functional) is denoted  $\theta$  and is assumed to run through a subset  $\Theta$  of a metric space; for instance, Euclidean space. It is often fruitful to separate the derivation of the limit behavior of a sequence of estimators  $\hat{\theta}_n$  into subproblems, such as proving consistency, deriving the convergence rate, and finally establishing the limit distribution. In this scheme certain theorems are applied to “local” (pseudo-) parameters (such as  $h = \sqrt{n}(\theta - \theta_0)$  if  $\theta_0$  is the “true” parameter). For convenience, theorems that are typically applied to local parameters are stated in terms of a parameter  $h$ .

### 3.2.1 The Argmax Theorem

Consider finding a point of maximum as a functional called “argmax”. When applied to a random criterion function this yields an  $M$ -estimator. If the argmax functional were continuous with respect to some metric on the space of criterion functions, then convergence in distribution of the criterion functions would imply the convergence in distribution of their points of maximum, the  $M$ -estimators, to the point of maximum of the limit criterion function. This is simply a special case of the continuous mapping theorem for weak convergence.

In keeping with the main development in this book—the convergence of stochastic processes in spaces of the type  $\ell^\infty(T)$ , such as the empirical process—we shall develop this idea using the uniform topology on the space of criterion functions. Actually, the proofs ahead show that the (joint) convergence of suprema over certain sets is crucial for the convergence of a point of maximum, not the convergence of the processes with respect to the uniform metric. For simplicity, we shall not pursue such refinements.

Consider a sequence  $\{\mathbb{M}_n(h): h \in H\}$  of stochastic processes indexed by a metric space  $H$ . For each  $n$  let the “estimator”  $\hat{h}_n$  be a point of (near) maximum of the “criterion function”

$$h \mapsto \mathbb{M}_n(h).$$

Formally,  $\hat{h}_n$  may be any  $H$ -valued map defined on the same probability space as the process  $\mathbb{M}_n$ ; its (outer) distribution is evaluated accordingly. Let  $\{\mathbb{M}(h): h \in H\}$  be a “limit process”, and assume that it possesses the

same relation to a random point  $\hat{h}$ , except that  $\hat{h}$  is implicitly assumed to be Borel measurable.

The argmax functional is continuous at functions  $\mathbb{M}$  that have a unique, “well-separated” point of maximum: the function  $h \mapsto \mathbb{M}(h)$  should be strictly smaller than  $\mathbb{M}(\hat{h})$  on the complement of every neighborhood of the point  $\hat{h}$ . This requirement, which appears to be natural, is the first condition in the following lemma.

This lemma contains the crux of the (simple) argument; to obtain the best results, it may be necessary to tailor the underlying idea to particular examples.

**3.2.1 Lemma.** *Let  $\mathbb{M}_n, \mathbb{M}$  be stochastic processes indexed by a metric space  $H$ . Let  $A$  and  $B$  be arbitrary subsets of  $H$ . Suppose there exists a random element  $\hat{h}$  such that almost surely*

$$\mathbb{M}(\hat{h}) > \sup_{h \notin G, h \in A} \mathbb{M}(h),$$

*for every open set  $G$  that contains  $\hat{h}$ . Suppose the sequence  $\hat{h}_n$  satisfies*

$$\mathbb{M}_n(\hat{h}_n) \geq \sup_h \mathbb{M}_n(h) - o_P(1).$$

*If  $\mathbb{M}_n \rightsquigarrow \mathbb{M}$  in  $\ell^\infty(A \cup B)$ , then, for every closed set  $F$ ,*

$$\limsup_{n \rightarrow \infty} P^*(\hat{h}_n \in F \cap A) \leq P(\hat{h} \in F \cup B^c).$$

If  $\mathbb{M}_n \rightsquigarrow \mathbb{M}$  in  $\ell^\infty(H)$ , then the lemma could be applied with  $A = B = H$ . In view of the portmanteau theorem, it would yield the conclusion that  $\hat{h}_n \rightsquigarrow \hat{h}$ .

Unfortunately, the assumption that  $\mathbb{M}_n \rightsquigarrow \mathbb{M}$  uniformly in the whole (local) parameter space is often too strong. This does not mean that the weak convergence  $\hat{h}_n \rightsquigarrow \hat{h}$  is not true, since the uniform convergence of the criterion functions is much stronger than the convergence of the locations of maxima. Therefore, it may be beneficial to establish additional properties of the sequence  $\hat{h}_n$  before invoking the continuous mapping theorem for the argmax functional. The following theorem requires uniform tightness, in which case uniform convergence of the criterion functions on compacta suffices.

On compacta, a unique maximum of an upper semicontinuous function  $h \mapsto \mathbb{M}(h)$  is automatically well separated, as required by the preceding lemma. We obtain the following theorem.

**3.2.2 Theorem (Argmax continuous mapping).** *Let  $\mathbb{M}_n, \mathbb{M}$  be stochastic processes indexed by a metric space  $H$  such that  $\mathbb{M}_n \rightsquigarrow \mathbb{M}$  in  $\ell^\infty(K)$  for every compact  $K \subset H$ . Suppose that almost all sample paths  $h \mapsto \mathbb{M}(h)$  are upper semicontinuous and possess a unique maximum at a (random) point  $\hat{h}$ , which as a random map in  $H$  is tight. If the sequence  $\hat{h}_n$  is uniformly tight and satisfies  $\mathbb{M}_n(\hat{h}_n) \geq \sup_h \mathbb{M}_n(h) - o_P(1)$ , then  $\hat{h}_n \rightsquigarrow \hat{h}$  in  $H$ .*

**Proofs.** For the proof of the lemma, note that by the continuous mapping theorem, the sequence  $\sup_{h \in F \cap A} \mathbb{M}_n(h) - \sup_{h \in B} \mathbb{M}_n(h)$  converges in distribution to the same expression, but with  $\mathbb{M}$  instead of  $\mathbb{M}_n$ . Thus

$$\begin{aligned} & \limsup_{n \rightarrow \infty} P^*(\hat{h}_n \in F \cap A) \\ & \leq \limsup P^*\left(\sup_{h \in F \cap A} \mathbb{M}_n(h) \geq \sup_{h \in B} \mathbb{M}_n(h) - o_P(1)\right) \\ & \leq P\left(\sup_{h \in F \cap A} \mathbb{M}(h) \geq \sup_{h \in B} \mathbb{M}(h)\right), \end{aligned}$$

by Slutsky's lemma and the portmanteau theorem. By the condition on the sample paths of  $\mathbb{M}$ , the event in the last probability is contained in the set  $\{\hat{h} \in F\} \cup \{\hat{h} \notin B\}$ , since the set  $G = F^c$  is open. The lemma follows.

For the proof of the theorem, take  $A = B = K$  equal to a compact set  $K$ . Then almost surely

$$\mathbb{M}(\hat{h}) > \sup_{h \notin G, h \in K} \mathbb{M}(h),$$

for every open set  $G$  around  $\hat{h}$ . If this were not true, then there would exist a sequence  $h_m \subset G^c \cap K$  with  $\mathbb{M}(h_m) \rightarrow \mathbb{M}(\hat{h})$ . Since  $K$  is compact the sequence may be chosen to be convergent; by upper semicontinuity the value  $\mathbb{M}(h)$  at the limit would be at least  $\mathbb{M}(\hat{h})$ . This contradicts the fact that  $\hat{h}$  is unique, for  $h$  is contained in the closed set  $G^c$  and hence cannot equal  $\hat{h}$ .

Thus, the lemma may be applied and yields

$$\limsup_{n \rightarrow \infty} P^*(\hat{h}_n \in F) \leq P(\hat{h} \in F) + P(\hat{h} \notin K) + \limsup_{n \rightarrow \infty} P^*(\hat{h}_n \notin K),$$

for every closed set  $F$ . The last two terms on the right can be made arbitrarily small by the choice of  $K$ . Apply the portmanteau theorem to conclude that  $\hat{h}_n \rightsquigarrow \hat{h}$ . ■

The preceding results are stated in terms of the parameter  $h$ , because they are typically applied to a local parameter. However, they may also be applied to the original parameter, in which case the limit criterion function is typically nonrandom and the approach turns into a consistency proof.

**3.2.3 Corollary (Consistency).** Let  $\mathbb{M}_n$  be stochastic processes indexed by a metric space  $\Theta$ , and let  $\mathbb{M}: \Theta \mapsto \mathbb{R}$  be a deterministic function.

(i) Suppose that  $\|\mathbb{M}_n - \mathbb{M}\|_\Theta \rightarrow 0$  in outer probability and that there exists a point  $\theta_0$  such that

$$\mathbb{M}(\theta_0) > \sup_{\theta \notin G} \mathbb{M}(\theta),$$

for every open set  $G$  that contains  $\theta_0$ . Then any sequence  $\hat{\theta}_n$ , such that  $\mathbb{M}_n(\hat{\theta}_n) \geq \sup_\theta \mathbb{M}_n(\theta) - o_P(1)$ , satisfies  $\hat{\theta}_n \rightarrow \theta_0$  in outer probability.

- (ii) Suppose that  $\|\mathbb{M}_n - \mathbb{M}\|_K \rightarrow 0$  in outer probability for every compact  $K \subset \Theta$  and that the map  $\theta \mapsto \mathbb{M}(\theta)$  is upper semicontinuous with a unique maximum at  $\theta_0$ . Then the same conclusion is true provided the sequence  $\hat{\theta}_n$  is uniformly tight.

If the estimator  $\hat{\theta}_n$  maximizes the criterion function  $\theta \mapsto \mathbb{M}_n(\theta)$ , then the preceding theorem will typically be applied to a multiple of a “rescaled” criterion function

$$h \mapsto \mathbb{M}_n\left(\theta + \frac{h}{r_n}\right) - \mathbb{M}_n(\theta),$$

where  $\theta$  is the “true” parameter and  $r_n \rightarrow \infty$  is the “rate of convergence” of the estimator. We then obtain a distributional limit result for the sequence  $\hat{h}_n = r_n(\hat{\theta}_n - \theta)$ . On the other hand, the corollary will be applied to (a multiple of) the original criterion functions  $\mathbb{M}_n$ .

A good approach to obtain the limiting distribution of  $M$ -estimators of Euclidean parameters appears to consist of three steps:

- establish the consistency of the sequence  $\hat{\theta}_n$  for a value  $\theta_0$ ;
- establish the rate of convergence  $r_n$  of the sequence  $\hat{\theta}_n$  or, equivalently, establish the tightness of the sequence of local parameters  $\hat{h}_n = r_n(\hat{\theta}_n - \theta_0)$ ;
- show that suitably rescaled versions of the criterion functions converge in distribution to a limit process  $\mathbb{M}$  in the space  $\ell^\infty(h: \|h\| \leq K)$  for every  $K$ .

If the sample paths  $h \mapsto \mathbb{M}(h)$  of the limit process are upper semicontinuous and possess a unique maximum  $\hat{h}$ , then the final conclusion is that the sequence  $r_n(\hat{\theta}_n - \theta_0)$  converges in distribution to  $\hat{h}$ .

**3.2.4 Example (Parametric maximum likelihood).** For the maximum likelihood estimator based on i.i.d. observations from a density  $p_\theta$  we may choose  $\mathbb{M}_n(\theta) = \sum_{i=1}^n \log p_\theta(X_i)$ . If the map  $\theta \mapsto p_\theta$  is sufficiently smooth, then the sequence of statistical models is locally asymptotically normal:

$$\sum_{i=1}^n \log \frac{p_{\theta+h/\sqrt{n}}}{p_\theta}(X_i) = h' \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_\theta(X_i) - \frac{1}{2} h' I_\theta h + o_{P_\theta}(1).$$

Here  $\dot{\ell}_\theta$  is the score function of the model and  $I_\theta$  is the Fisher information matrix. The sequence of stochastic processes on the right (indexed by  $h \in \mathbb{R}^k$ ) converges marginally in distribution to the Gaussian process

$$h \mapsto h' \Delta - \frac{1}{2} h' I_\theta h,$$

for a  $N_k(0, I_\theta)$ -distributed random variable  $\Delta$ . If  $\theta$  is an inner point of the parameter set, then the sequence  $\hat{h}_n = \sqrt{n}(\hat{\theta}_n - \theta)$  typically converges in distribution to the maximizer  $\hat{h}$  of this process when  $h$  ranges over the full Euclidean space. The classical results on asymptotic normality of maximum

likelihood estimators precisely make this statement, for the maximizer is  $\hat{\theta} = I_{\theta}^{-1} \Delta_{\theta}$  and possesses a normal  $N(0, I_{\theta}^{-1})$ -distribution.

Asymptotic normality may be proved by means of the preceding theorem provided the marginal convergence of the processes is suitably strengthened. Simple, but already powerful, conditions are given in Example 3.2.24.

In the case of i.i.d. data and an empirical criterion function of the form  $M_n(\theta) = P_n m_{\theta}$ , the uniform convergence in Corollary 3.2.3 (to  $M(\theta) = P m_{\theta}$ ) is valid if and only if the class of functions  $\{m_{\theta}: \theta \in \Theta\}$  is Glivenko-Cantelli. The main Glivenko-Cantelli theorems can be found in Chapter 2.4. The rescaled processes  $\{P_n(m_{\theta+h/r_n} - m_{\theta}): h \in K\}$  are up to centering at mean zero a multiple of the empirical process indexed by the classes of functions  $\{m_{\theta+h/r_n} - m_{\theta}: h \in K\}$ . Because these classes are changing with  $n$ , the convergence in distribution of these processes is somewhat more complicated than establishing a Donsker property. However, sufficient conditions for the convergence in distribution are given in Section 2.11.3.

In the next section we present a method to establish the rate of convergence.

### 3.2.2 Rate of Convergence

If  $\theta_0$  is a point of maximum of the map  $\theta \mapsto M(\theta)$ , then the first derivative must vanish at  $\theta_0$  and the second derivative should be negative definite. Thus, it is natural to assume that for  $\theta$  in a neighborhood of  $\theta_0$ ,<sup>†</sup>

$$M(\theta) - M(\theta_0) \lesssim -d^2(\theta, \theta_0).$$

Consider estimators  $\hat{\theta}_n$  that (nearly) maximize maps  $\theta \mapsto M_n(\theta)$ . An upper bound for the rate of convergence of  $\hat{\theta}_n$  can be obtained from the continuity modulus of the difference  $M_n - M$ .

**3.2.5 Theorem (Rate of convergence).** *Let  $M_n$  be stochastic processes indexed by a semimetric space  $\Theta$  and  $M: \Theta \mapsto \mathbb{R}$  a deterministic function, such that for every  $\theta$  in a neighborhood of  $\theta_0$ ,*

$$M(\theta) - M(\theta_0) \lesssim -d^2(\theta, \theta_0).$$

*Suppose that, for every  $n$  and sufficiently small  $\delta$ , the centered process  $M_n - M$  satisfies*

$$E^* \sup_{d(\theta, \theta_0) < \delta} |(M_n - M)(\theta) - (M_n - M)(\theta_0)| \lesssim \frac{\phi_n(\delta)}{\sqrt{n}},$$

---

<sup>†</sup> The notation  $\lesssim$  means “is bounded above up to a universal constant”. This condition is often satisfied but is unnecessarily restrictive. Theorem 3.2.5 remains true even if the numbers  $d^2(\theta, \theta_0)$  do not arise from a distance but simply define an arbitrary function from  $\Theta$  to the nonnegative reals.

for functions  $\phi_n$  such that  $\delta \mapsto \phi_n(\delta)/\delta^\alpha$  is decreasing for some  $\alpha < 2$  (not depending on  $n$ ). Let

$$r_n^2 \phi_n\left(\frac{1}{r_n}\right) \leq \sqrt{n}, \quad \text{for every } n.$$

If the sequence  $\hat{\theta}_n$  satisfies  $\mathbb{M}_n(\hat{\theta}_n) \geq \mathbb{M}_n(\theta_0) - O_P(r_n^{-2})$  and converges in outer probability to  $\theta_0$ , then  $r_n d(\hat{\theta}_n, \theta_0) = O_P^*(1)$ . If the displayed conditions are valid for every  $\theta$  and  $\delta$ , then the condition that  $\hat{\theta}_n$  is consistent is unnecessary.

**Proof.** Assume for simplicity that  $\hat{\theta}_n$  truly maximizes the map  $\theta \mapsto \mathbb{M}_n(\theta)$ . For each  $n$ , the parameter space (minus the point  $\theta_0$ ) can be partitioned into the “shells”  $S_{j,n} = \{\theta : 2^{j-1} < r_n d(\theta, \theta_0) \leq 2^j\}$  with  $j$  ranging over the integers. If  $r_n d(\hat{\theta}_n, \theta_0)$  is larger than  $2^M$  for a given integer  $M$ , then  $\hat{\theta}_n$  is in one of the shells  $S_{j,n}$  with  $j \geq M$ . In that case the supremum of the map  $\theta \mapsto \mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0)$  over this shell is nonnegative by the property of  $\hat{\theta}_n$ . Conclude that, for every  $\eta > 0$ ,

$$\begin{aligned} P^*\left(r_n d(\hat{\theta}_n, \theta_0) > 2^M\right) &\leq \sum_{\substack{j \geq M \\ 2^j \leq \eta r_n}} P^*\left(\sup_{\theta \in S_{j,n}} (\mathbb{M}_n(\theta) - \mathbb{M}_n(\theta_0)) \geq 0\right) \\ &\quad + P^*(2d(\hat{\theta}_n, \theta_0) \geq \eta). \end{aligned}$$

If the sequence  $\hat{\theta}_n$  is consistent for  $\theta_0$ , then the second probability on the right converges to 0 as  $n \rightarrow \infty$  for every  $\eta > 0$ . Choose  $\eta > 0$  small enough that the first condition of the theorem holds for every  $d(\theta, \theta_0) \leq \eta$  and the second for every  $\delta \leq \eta$ . Then for every  $j$  involved in the sum, we have, for every  $\theta \in S_{j,n}$ ,

$$\mathbb{M}(\theta) - \mathbb{M}(\theta_0) \leq -d^2(\theta, \theta_0) \lesssim \frac{-2^{2j-2}}{r_n^2}.$$

In terms of the centered process  $W_n = \mathbb{M}_n - \mathbb{M}$ , the series may be bounded by

$$\begin{aligned} \sum_{\substack{j \geq M \\ 2^j \leq \eta r_n}} P^*\left(\|W_n(\theta) - W_n(\theta_0)\|_{S_{j,n}} \geq \frac{2^{2j-2}}{r_n^2}\right) &\lesssim \sum_{j \geq M} \frac{\phi_n(2^j/r_n) r_n^2}{\sqrt{n} 2^{2j}} \\ &\lesssim \sum_{j \geq M} 2^{j\alpha - 2j}, \end{aligned}$$

by Markov’s inequality, the definition of  $r_n$ , and the fact that  $\phi_n(c\delta) \leq c^\alpha \phi_n(\delta)$  for every  $c > 1$  by the assumption on  $\phi_n$ . This expression converges to zero for every  $M = M_n \rightarrow \infty$ .

The second assertion follows by a minor simplification of the preceding argument. ■

In the case of i.i.d. data and criterion functions of the form  $\mathbb{M}_n(\theta) = \mathbb{P}_n m_\theta$  and  $\mathbb{M}(\theta) = Pm_\theta$ , the centered and scaled process  $\sqrt{n}(\mathbb{M}_n - \mathbb{M}) = \mathbb{G}_n m_\theta$  equals the empirical process at  $m_\theta$ . The second condition of the theorem involves the suprema of the empirical process indexed by classes of functions

$$\mathcal{M}_\delta = \{m_\theta - m_{\theta_0} : d(\theta, \theta_0) < \delta\}.$$

It is not unreasonable to assume that these suprema are bounded uniformly in  $n$ . This leads to the following result.

**3.2.6 Corollary.** *In the i.i.d. case assume that, for every  $\theta$  in a neighborhood of  $\theta_0$ ,*

$$P(m_\theta - m_{\theta_0}) \lesssim -d^2(\theta, \theta_0).$$

*Furthermore, assume that there exists a function  $\phi$  such that  $\delta \mapsto \phi(\delta)/\delta^\alpha$  is decreasing for some  $\alpha < 2$  and, for every  $n$ ,*

$$\mathbb{E}^* \|\mathbb{G}_n\|_{\mathcal{M}_\delta} \lesssim \phi(\delta).$$

*If the sequence  $\hat{\theta}_n$  satisfies  $\mathbb{P}_n m_{\hat{\theta}_n} \geq \mathbb{P}_n m_{\theta_0} - O_P(r_n^{-2})$  and converges in outer probability to  $\theta_0$ , then  $r_n d(\hat{\theta}_n, \theta_0) = O_P^*(1)$  for every sequence  $r_n$  such that  $r_n^2 \phi(1/r_n) \leq \sqrt{n}$  for every  $n$ . Thus the ‘‘continuity modulus’’ of the empirical process gives an upper bound on the rate. For instance, the modulus  $\phi(\delta) = \delta^\alpha$  gives a rate of at least  $n^{1/(4-2\alpha)}$ ; the ‘‘usual’’ rate  $\sqrt{n}$  corresponds to  $\phi(\delta) = \delta$ .*

For a Euclidean parameter space, the first condition of the theorem is satisfied if the map  $\theta \mapsto Pm_\theta$  is twice continuously differentiable at the point of maximum  $\theta_0$  with nonsingular second-derivative matrix.

The preceding theorem appears to give the correct rate in fair generality, the main problem being to derive sharp bounds on the continuity modulus of the empirical process. A simple, but not necessarily efficient, method is to apply the maximal inequalities given by Theorems 2.14.1 and 2.14.2. These yield bounds in terms of the uniform entropy integral  $J(1, \mathcal{M}_\delta)$  or the bracketing integral  $J_{[]} (1, \mathcal{M}_\delta, L_2(P))$  of the class  $\mathcal{M}_\delta$  given by

$$\begin{aligned} \mathbb{E}_P^* \|\mathbb{G}_n\|_{\mathcal{M}_\delta} &\lesssim J(1, \mathcal{M}_\delta) (P^* M_\delta^2)^{1/2}, \\ \mathbb{E}_P^* \|\mathbb{G}_n\|_{\mathcal{M}_\delta} &\lesssim J_{[]} (1, \mathcal{M}_\delta, L_2(P)) (P^* M_\delta^2)^{1/2}. \end{aligned}$$

Here  $M_\delta$  is an envelope function of the class  $\mathcal{M}_\delta$ . These bounds are pessimistic in that they depend on the size of the functions  $m_\theta - m_{\theta_0}$ , which are likely to be small for small  $\delta$ , mostly through the envelope of the class. The assumption that the entropy integrals  $J(1, \mathcal{M}_\delta)$  and  $J_{[]} (1, \mathcal{M}_\delta, L_2(P))$  are bounded as  $\delta \downarrow 0$  appears reasonable. (This is not automatically the case, because they are defined relative to an envelope function.) In that case the

preceding corollary may be applied with the upper bound  $\phi^2(\delta) = P^* M_\delta^2$  and leads to a rate of convergence  $r_n$  of at least the solution of

$$(3.2.7) \quad r_n^4 P^* M_{1/r_n}^2 \sim n.$$

In many examples involving finite-dimensional parameters this gives the correct result. In Chapter 3.4 we discuss rates of convergence in more generality.

In the remainder of this section we pursue the special case that the rate is given by (3.2.7) and present a theorem concerning the limit distribution of the sequence  $r_n(\hat{\theta}_n - \theta_0)$ . As is clear from the examples given in the next section, this theorem should not be viewed as the only approach, but it does give a good illustration of the combination of the argmax theorem, Theorem 3.2.2, and the rate theorem, Theorem 3.2.5.

We specialize to a Euclidean parameter set  $\Theta$  and the case of i.i.d. observations. To derive the limit distribution of  $r_n(\hat{\theta}_n - \theta_0)$  using the argmax theorem, we need to establish the convergence of a multiple of the processes  $h \mapsto \mathbb{P}_n(m_{\theta_0+h/r_n} - m_{\theta_0})$  in  $\ell^\infty(h: \|h\| \leq K)$  for every  $K$ . Theorems 2.11.22 and 2.11.23 give conditions for the convergence in distribution of the centered processes

$$\begin{aligned} h &\mapsto \frac{r_n^2}{\sqrt{n}} \mathbb{G}_n(m_{\theta_0+h/r_n} - m_{\theta_0}) \\ &= r_n^2 \mathbb{P}_n(m_{\theta_0+h/r_n} - m_{\theta_0}) - r_n^2 P(m_{\theta_0+h/r_n} - m_{\theta_0}). \end{aligned}$$

These are the empirical processes with indexed classes  $(r_n^2/\sqrt{n})\mathcal{M}_{K/r_n}$ . Theorems 2.11.22 and 2.11.23 require that the envelope functions of these classes are bounded in square expectation. In the present case this is satisfied by the definition of  $r_n$  in (3.2.7). (Of course, it was the motivation for multiplying the empirical measure by  $r_n^2$ .) We shall also translate the other conditions to the present situation.

Assume that either the uniform entropy or bracketing integrals of the classes  $\mathcal{M}_\delta$  are uniformly bounded as  $\delta$  tends to 0: for some  $\delta_0 > 0$

$$(3.2.8) \quad \int_0^\infty \sup_{\delta < \delta_0} \sup_Q \sqrt{\log N(\varepsilon \|M_\delta\|_{Q,2}, \mathcal{M}_\delta, L_2(Q))} d\varepsilon < \infty$$

or

$$(3.2.9) \quad \int_0^\infty \sup_{\delta < \delta_0} \sqrt{\log N_{[]}(\varepsilon \|M_\delta\|_{P,2}, \mathcal{M}_\delta, L_2(P))} d\varepsilon < \infty.$$

The first type of entropy condition can be exploited only under some measurability conditions. In the following it will be understood, that (3.2.8) includes the condition that the classes  $\mathcal{M}_\delta$  satisfy the conditions of Theorem 2.14.1.

**3.2.10 Theorem.** For each  $\theta$  in an open subset of Euclidean space, let  $m_\theta$  be a measurable function such that  $\theta \mapsto Pm_\theta$  is twice continuously differentiable at a point of maximum  $\theta_0$ , with nonsingular second-derivative matrix  $V$ .<sup>b</sup> Let the entropy condition (3.2.8) or (3.2.9) hold. Assume that for some continuous function  $\phi$ , such that  $\phi^2(\delta) \geq P^*M_\delta^2$  and such that  $\delta \mapsto \phi(\delta)/\delta^\alpha$  is decreasing for some  $\alpha < 2$ , and for every  $\eta > 0$ ,

$$(3.2.11) \quad \begin{aligned} \lim_{\delta \downarrow 0} \frac{P^*M_\delta^2 \{ M_\delta > \eta\delta^{-2}\phi^2(\delta) \}}{\phi^2(\delta)} &= 0, \\ \lim_{\varepsilon \downarrow 0} \limsup_{\delta \downarrow 0} \sup_{\substack{\|h-g\| < \varepsilon \\ \|h\| \vee \|g\| \leq K}} \frac{P(m_{\theta_0+\delta g} - m_{\theta_0+\delta h})^2}{\phi^2(\delta)} &= 0, \\ \lim_{\delta \downarrow 0} \frac{P(m_{\theta_0+\delta g} - m_{\theta_0+\delta h})^2}{\phi^2(\delta)} &= E(G(g) - G(h))^2, \end{aligned}$$

for all  $K$  and some zero-mean Gaussian process  $G$  such that  $G(g) = G(h)$  almost surely only if  $h = g$ .<sup>#</sup> Then there exists a version of  $G$  with bounded, uniformly continuous sample paths on compacta. Define  $r_n$  as the solution of  $r_n^2 \phi(1/r_n) = \sqrt{n}$ . If  $\hat{\theta}_n$  nearly maximizes the map  $\theta \mapsto \mathbb{P}_n m_\theta$  for every  $n$  and converges in outer probability to  $\theta_0$ , then the sequence  $r_n(\hat{\theta}_n - \theta_0)$  converges in distribution to the unique maximizer  $\hat{h}$  of the process  $h \mapsto G(h) + \frac{1}{2}h'Vh$ .

**Proof.** By the discussion preceding the theorem and the nonsingularity of the second-derivative matrix  $V$ , the sequence  $r_n(\hat{\theta}_n - \theta_0)$  is uniformly tight.

The conditions of Theorems 2.11.22 and 2.11.23 for asymptotic tightness of the sequence  $(r_n^2/\sqrt{n})\mathbb{G}_n(m_{\theta_0+h/r_n} - m_{\theta_0})$  in the space  $\ell^\infty(h: \|h\| \leq K)$  are an entropy condition and

$$\begin{aligned} \frac{r_n^4}{n} P^*M_{K/r_n}^2 &= O(1); \quad \frac{r_n^4}{n} P^*M_{K/r_n}^2 \{ r_n^2 M_{K/r_n} > \eta n \} = o(1), \\ \sup_{\|h-g\| < \eta_n} \frac{r_n^4}{n} P(m_{\theta_0+g/r_n} - m_{\theta_0+h/r_n})^2 &= o(1). \end{aligned}$$

These conditions are implied by the conditions of the theorem. By the two-times differentiability of the map  $\theta \mapsto Pm_\theta$ ,

$$r_n^2 P(m_{\theta_0+h/r_n} - m_{\theta_0}) \rightarrow \frac{1}{2}h'Vh,$$

uniformly in  $h$  ranging over bounded sets. Conclude that the sequence of processes  $h \mapsto r_n^2 \mathbb{P}_n(m_{\theta_0+h/r_n} - m_{\theta_0})$  is asymptotically tight in the

<sup>b</sup> It suffices that a two-term Taylor expansion  $Pm_{\theta_0+h} = Pm_{\theta_0} + \frac{1}{2}h'Vh + o(\|h\|^2)$  is valid.

<sup>#</sup> Condition (3.2.11) may be replaced by marginal convergence of the processes  $h \mapsto r_n^2 \mathbb{P}_n(m_{\theta_0+h/r_n} - m_{\theta_0})$  to a process  $h \mapsto G(h)$ .

space  $\ell^\infty(h: \|h\| \leq K)$  as well. Its marginals are triangular arrays of i.i.d. random vectors that satisfy the Lindeberg condition. In view of the preceding display and (3.2.11), the mean and covariance function converge to  $\frac{1}{2}h'Vh$  and  $EG(g)G(h)$ , respectively. Thus, the sequence of processes  $h \mapsto r_n^2 \mathbb{P}_n(m_{\theta_0+h/r_n} - m_{\theta_0})$  converges in distribution in the space  $\ell^\infty(h: \|h\| \leq K)$  to the Gaussian process  $h \mapsto G(h) + \frac{1}{2}h'Vh$ , for every  $K$ .

By Proposition A.2.20, almost all sample paths of this process attain their supremum at a unique point  $\hat{h}$ . Finally, apply Theorem 3.2.2. ■

The second and third lines of the display in the theorem are implied by the single condition: for every  $g_\delta \rightarrow g$  and  $h_\delta \rightarrow h$ ,

$$\lim_{\delta \downarrow 0} \frac{P(m_{\theta_0+\delta g_\delta} - m_{\theta_0+\delta h_\delta})^2}{\phi^2(\delta)} = E(G(g) - G(h))^2.$$

Instead of the variances of the differences, the condition could also be stated in terms of the covariances of  $m_{\theta_1}$  and  $m_{\theta_2}$ . See Problem 3.2.1 for further simplifications of the theorem.

### 3.2.3 Examples

This section presents some examples that show the scope of combining the argmax theorem, Theorem 3.2.2, and the rate theorem, Theorem 3.2.5. In some examples we simply apply Theorem 3.2.10; in other examples we give the complete argument.

**3.2.12 Example (Lipschitz in parameter).** A simple method to verify the conditions of the preceding theorems is by a pointwise Lipschitz condition on the maps  $\theta \mapsto m_\theta$ . Let  $X_1, \dots, X_n$  be i.i.d. random variables with common law  $P$ , and let  $m_\theta$  be measurable functions such that, for every  $\theta_1, \theta_2$  in a neighborhood of  $\theta_0$ ,

$$|m_{\theta_1}(x) - m_{\theta_2}(x)| \leq \dot{m}(x)\|\theta_1 - \theta_2\|^\alpha,$$

for a square-integrable function  $\dot{m}$ . Then the first two displayed conditions of the theorem are satisfied for  $\phi(\delta) = \delta^\alpha$ . The envelope function  $M_\delta$  can be taken equal to  $\delta^\alpha \dot{m}$ , and the bracketing numbers of the class  $\mathcal{M}_\delta$  are bounded by the covering numbers of Euclidean balls of radius  $\delta$  for the Euclidean metric to the power  $\alpha$ , hence polynomial. See Section 2.7.4.

Let  $\hat{\theta}_n$  maximize  $\theta \mapsto \mathbb{P}_n m_\theta$  for every  $n$  and be consistent for  $\theta_0$ . If in addition the map  $\theta \mapsto Pm_\theta$  is twice continuously differentiable at its point of maximum  $\theta_0$  with a nonsingular second-derivative matrix, then Theorem 3.2.5 gives a rate of convergence of at least  $n^{1/(4-2\alpha)}$ .

The limit distribution of the sequence  $n^{1/(4-2\alpha)}(\hat{\theta}_n - \theta_0)$  follows by the preceding theorem under the further condition that for a zero-mean Gaussian process  $G$ ,

$$P(m_{\theta_0+\delta g} - m_{\theta_0})(m_{\theta_0+\delta h} - m_{\theta_0}) \sim \delta^{2\alpha} EG(g)G(h).$$

In this case the sequence  $n^{1/(4-2\alpha)}(\hat{\theta}_n - \theta_0)$  converges in distribution to the maximizer of the process  $h \mapsto G(h) + \frac{1}{2}h'Vh$  for a version of  $G$  with continuous sample paths.

For  $\alpha = 1$ , it appears not unreasonable to assume that, for some vector-valued function  $\dot{m}_{\theta_0}$ ,

$$P\left[\frac{m_{\theta_0+\delta h} - m_{\theta_0}}{\delta} - h'\dot{m}_{\theta_0}\right]^2 \rightarrow 0.$$

This yields the covariance structure  $EG(g)G(h) = g'Wh$  for the matrix  $W = P\dot{m}_{\theta_0}\dot{m}'_{\theta_0}$ . A process  $G$  with a covariance structure of this type can be represented as  $G(h) = h'\Delta$  for a  $N(0, W)$ -distributed random vector  $\Delta$ . The maximizer of the process  $h \mapsto h'\Delta + \frac{1}{2}h'Vh$  is  $\hat{h} = -V^{-1}\Delta$  and is normally distributed with mean zero and covariance matrix  $V^{-1}WV^{-1}$ . Thus, in this case the sequence  $\sqrt{n}(\hat{\theta}_n - \theta)$  converges in distribution to a normal distribution.

Example 3.2.22 ahead establishes the same conclusion by a more classical linearization argument.

**3.2.13 Example.** Let  $X_1, \dots, X_n$  be i.i.d., real-valued random variables with common law  $P$  with a differentiable Lebesgue density  $p$ . Define an estimator  $\hat{\theta}_n$  of location as the maximizer of the function

$$\theta \mapsto \mathbb{P}_n[\theta - 1, \theta + 1].$$

Thus  $\hat{\theta}_n$  is the center of an interval of length 2 that contains the largest possible fraction of the observations.

By the classical Glivenko-Cantelli theorem, the sequence  $\sup_{\theta} |(\mathbb{P}_n - P)[\theta - 1, \theta + 1]|$  converges in probability to zero. Consequently, if the map  $\theta \mapsto P[\theta - 1, \theta + 1]$  has a unique “well-separated” point of maximum  $\theta_0$ , then the sequence  $\hat{\theta}_n$  converges in probability to  $\theta_0$ . The second derivative of  $\theta \mapsto P[\theta - 1, \theta + 1]$  equals  $p'(\theta + 1) - p'(\theta - 1)$ , which is likely to be strictly negative at  $\theta_0$ .

The functions  $m_{\theta} = 1_{[\theta-1, \theta+1]}$  are not Lipschitz in the parameter. Nevertheless, the classes of functions  $\mathcal{M}_{\delta}$  satisfy the conditions of the preceding theorem. These classes are Vapnik-Červonenkis with envelope functions

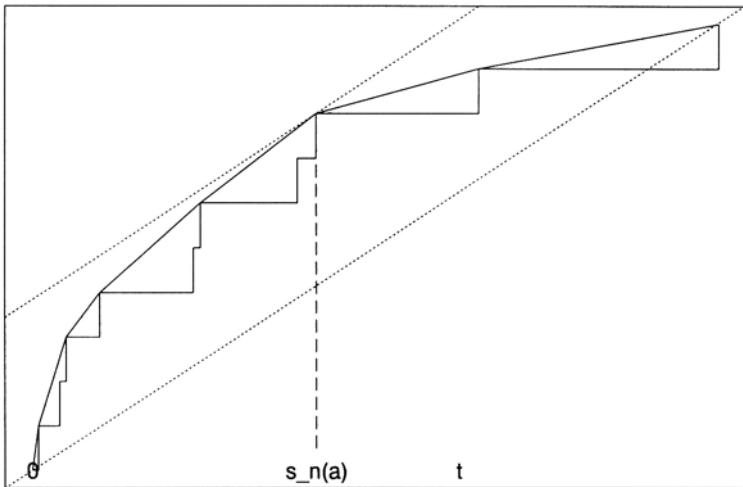
$$\sup_{|\theta - \theta_0| < \delta} |1_{[\theta-1, \theta+1]} - 1_{[\theta_0-1, \theta_0+1]}| \leq 1_{[\theta_0-1-\delta, \theta_0-1+\delta]} + 1_{[\theta_0+1-\delta, \theta_0+1+\delta]}.$$

The  $L_2(P)$ -norm of these functions is bounded above by a constant times  $\sqrt{\delta}$ . Thus, the conditions of the theorem are satisfied with  $\phi(\delta) = c\sqrt{\delta}$  for a constant  $c$ , leading to a rate of convergence  $n^{1/3}$ . The sequence  $n^{1/3}(\hat{\theta}_n - \theta_0)$  converges in distribution to the maximizer of the process  $h \mapsto G(h) + \frac{1}{2}h^2(p'(\theta_0 + 1) - p'(\theta_0 - 1))$ . Here  $G$  is zero-mean Gaussian with continuous sample paths and variance function

$$E(G(g) - G(h))^2 = (p(\theta_0 - 1) + p(\theta_0 + 1)) |g - h|.$$

Up to a scale factor,  $G$  is a *two-sided Brownian motion* originating from zero.

**3.2.14 Example (Monotone densities).** Let  $X_1, \dots, X_n$  be an i.i.d. sample from a Lebesgue density  $f$  on  $[0, \infty)$  that is known to be decreasing. The maximum likelihood estimator  $\hat{f}_n$  of  $f$  is the step function equal to the left derivative of the least concave majorant of the empirical distribution function  $\mathbb{F}_n$ . For a fixed value  $t > 0$ , we shall derive the limit distribution of the sequence  $n^{1/3}(\hat{f}_n(t) - f(t))$  under the assumption that the true density is differentiable at  $t$  with derivative  $f'(t) < 0$ .



**Figure 3.1.** If  $\hat{f}_n(t) \leq a$ , then a line of slope  $a$  moved down vertically from  $+\infty$  first hits  $\mathbb{F}_n$  to the left of  $t$ . The point where the line hits is the point where  $\mathbb{F}_n$  is farthest above the line of slope  $a$  through the origin.

Define a stochastic process  $\{\hat{s}_n(a): a > 0\}$  by

$$\hat{s}_n(a) = \operatorname{argmax}_s \{\mathbb{F}_n(s) - as\},$$

where the largest value is chosen when multiple maximizers exist. The function  $\hat{s}_n$  is the inverse of the function  $\hat{f}_n$  in the sense that  $\hat{f}_n(t) \leq a$  if and only if  $\hat{s}_n(a) \leq t$  for every  $t$  and  $a$ . This is explained in Figure 3.1. It follows that

$$P\left(n^{1/3}(\hat{f}_n(t) - f(t)) \leq x\right) = P\left(\hat{s}_n(f(t) + xn^{-1/3}) \leq t\right).$$

Hence the desired result can be deduced from the limiting behavior of the process of points of maximum  $\hat{s}_n(f(t) + xn^{-1/3})$ . By the change of variable  $s \mapsto t + hn^{-1/3}$  in the definition of  $\hat{s}_n$ , we have

$$\begin{aligned} & \hat{s}_n(f(t) + xn^{-1/3}) - t \\ &= n^{-1/3} \operatorname{argmax}_h \left\{ \mathbb{F}_n(t + hn^{-1/3}) - (f(t) + xn^{-1/3})(t + hn^{-1/3}) \right\}. \end{aligned}$$

It follows that the probability of interest is  $P(\hat{h}_n \leq 0)$  for  $\hat{h}_n$  equal to the argmax  $h$  appearing on the right. The location of the maximum of a function does not change when the function is multiplied by a positive constant or shifted vertically. Thus, the argmax  $\hat{h}_n$  is also a point of maximum of the process

$$\begin{aligned} h \mapsto n^{2/3}(\mathbb{P}_n - P)(1_{[0,t+hn^{-1/3}]} - 1_{[0,t]}) \\ + n^{2/3}[F(t + hn^{-1/3}) - F(t) - f(t)hn^{-1/3}] - xh. \end{aligned}$$

By Theorem 2.11.22 or 2.11.23, this sequence of processes converges, for every  $K$ , in the space  $\ell^\infty(-K, K)$  to the process

$$h \mapsto \sqrt{f(t)} \mathbb{Z}(h) + \frac{1}{2}f'(t)h^2 - xh,$$

where  $\mathbb{Z}$  is a two-sided Brownian motion originating from zero: a mean-zero Gaussian process with  $\mathbb{Z}(0) = 0$  and  $E(\mathbb{Z}(g) - \mathbb{Z}(h))^2 = |g - h|$  for every  $g, h$ . If it can be proved that  $\hat{h}_n = O_P(1)$ , then the argmax continuous mapping theorem yields that the sequence  $\hat{h}_n$  converges in distribution to the maximizer  $\hat{h}$  of the limit process. In particular,

$$P\left(n^{1/3}(\hat{f}_n(t) - f(t)) \leq x\right) = P(\hat{h}_n \leq 0) \rightarrow P(\hat{h} \leq 0).$$

Elementary properties of Brownian motion allow one to rewrite the limit probability  $P(\hat{h} \leq 0)$  in a simple form. See Problem 3.2.5. The final conclusion is that

$$n^{1/3}(\hat{f}_n(t) - f(t)) \rightsquigarrow |4f'(t)f(t)|^{1/3} \operatorname{argmax}_h \{\mathbb{Z}(h) - h^2\}.$$

The sequence  $\hat{h}_n$  can be shown to be uniformly tight by Theorem 3.2.5 with criterion and centering function

$$\begin{aligned} \mathbb{M}_n(g) &= (\mathbb{P}_n - P)(1_{[0,t+g]} - 1_{[0,t]}) + F(t+g) - F(t) - f(t)g - xgn^{-1/3}, \\ \mathbb{M}(g) &= F(t+g) - F(t) - f(t)g. \end{aligned}$$

By its definition,  $\hat{g}_n = n^{-1/3}\hat{h}_n$  maximizes  $g \mapsto \mathbb{M}_n(g)$ . In order to show that the sequence  $n^{1/3}\hat{g}_n$  is bounded in probability, we shall apply Theorem 3.2.5. By assumption  $\mathbb{M}(g) = \frac{1}{2}f'(t)g^2 + o(g^2)$  as  $g \rightarrow 0$ . The bracketing entropy integral of the class of functions  $\{1_{[0,t+g]} - 1_{[0,t]} : |g| < \delta\}$  is bounded uniformly in  $\delta$  and the envelopes satisfy  $PM_\delta^2 \lesssim \delta$ . It follows that we can set  $\phi_n(\delta) = \sqrt{\delta} + \delta n^{1/6}$ , leading to a rate of convergence  $n^{1/3}$  for  $\hat{g}_n$  as desired, provided  $\hat{g}_n$  is consistent. The consistency can be proved by a direct argument, but also by Theorem 3.2.5 applied with  $d(g, 0) = -\mathbb{M}(g)$ . For this choice of “distance” the conditions of Theorem 3.2.5 are valid for every  $g$  and  $\delta$ , so that by the last assertion of the theorem,  $-n^{2/3}\mathbb{M}(\hat{g}_n) = O_P^*(1)$ . The concavity of  $\mathbb{M}$  shows that  $\hat{g}_n$  converges to zero in probability.

**3.2.15 Example (Current status).** Let  $X_1, \dots, X_n$  and  $T_1, \dots, T_n$  be independent i.i.d. samples from distribution functions  $F$  and  $G$  on the nonnegative half-line, respectively. Define  $\Delta_i = 1\{X_i \leq T_i\}$ . Each  $X_i$  is interpreted as the (unobserved) time of onset of a disease; each  $T_i$  is a check-up time at which the patient is observed to be ill or not:  $\Delta_i = 1$  or 0. The observations consist of the pairs  $(\Delta_1, T_1), \dots, (\Delta_n, T_n)$ . The maximum likelihood estimator  $\hat{F}_n$  for  $F$  maximizes the (partial) likelihood function

$$F \mapsto \sum_{i=1}^n \left( \Delta_i \log F(T_i) + (1 - \Delta_i) \log(1 - F(T_i)) \right).$$

This function depends on  $F$  only through the values  $F(T_i)$ . Hence the maximum likelihood estimator is not unique. In the following,  $\hat{F}_n$  is assumed constant on the intervals  $[T_{(i-1)}, T_{(i)}]$ .

If the observation times are initially ordered so that the observation times are increasing, then finding the values  $\hat{F}_n(T_i)$  is equivalent to the “isotonic” maximization problem of maximizing the function

$$(y_1, \dots, y_n) \mapsto \sum_{i=1}^n \left( \delta_i \log y_i + (1 - \delta_i) \log(1 - y_i) \right),$$

subject to  $0 \leq y_1 \leq y_2 \leq \dots \leq y_n \leq 1$ . In view of Theorem 1.5.1 of Robertson, Wright, and Dykstra (1988) applied with the convex function  $\Phi(y) = y \log y + (1 - y) \log(1 - y)$  the solution vector  $\hat{y}$  to this problem is the “isotonic regression” of the vector  $(\delta_1, \dots, \delta_n)$ : the vector  $\hat{y}$  that minimizes  $y \mapsto \sum_{i=1}^n (\delta_i - y_i)^2$  subject to the same condition as before. By Theorem 1.2.1 of the same reference, the isotonic regression  $\hat{y}_i$  can be graphically represented as the slope in the interval  $(i-1, i)$  of the greatest convex minorant of the “cumulative-sum diagram” of the points  $n^{-1}(i, \sum_{j \leq i} \delta_j)$ . The cumulative-sum diagram is the function  $c: [0, 1] \mapsto \mathbb{R}$ , which equals  $n^{-1} \sum_{j \leq i} \delta_j$  on the interval  $n^{-1}(i-1, i]$  and equals 0 at zero. The graphical representation shows that  $\hat{y}_i \leq a$  if and only if  $\operatorname{argmin}_s \{c(s) - as\} \geq i/n$ . (In case of multiple points of minimum, take the largest.) Let  $\mathbb{P}_n$  be the empirical measure of the pairs  $(X_1, T_1), \dots, (X_n, T_n)$ , and let  $A = \{(x, t): x \leq t\}$ . Define the stochastic processes

$$\begin{aligned} V_n(t) &= \mathbb{P}_n 1_A 1_{\mathbb{R} \times [0, t]} = \frac{1}{n} \sum \Delta_i 1\{T_i \leq t\}, \\ G_n(t) &= \mathbb{P}_n 1_{\mathbb{R} \times [0, t]} = \frac{1}{n} \sum 1\{T_i \leq t\}. \end{aligned}$$

Then the function  $s \mapsto V_n \circ G_n^{-1}(s)$  equals the cumulative-sum diagram, whence for every observation point  $t_i$ ,

$$\hat{F}_n(T_i) \leq a \quad \text{if and only if} \quad \operatorname{argmin}_s \{V_n(s) - aG_n(s)\} \geq T_i.$$

Thus, the limit distribution of  $\hat{F}_n(t)$  can be derived by studying the locations of the minima of the sequence of processes  $s \mapsto V_n(s) - aG_n(s)$ . Assume that  $F$  and  $G$  are differentiable at  $t > 0$  with derivatives  $f(t)$  and  $g(t) > 0$ . For each  $t$ , the value  $\hat{F}_n(t)$  equals  $\hat{F}_n(T_i)$  for some  $i$ ; hence  $\hat{F}_n(t) \leq a$  if and only if the argmin appearing in the display is larger than the observation time  $T_i$  that is just left of  $t$ . Under the condition that  $g(t) > 0$ , the difference between this observation time and  $t$  is asymptotically negligible. By the change of variables  $s \mapsto t + n^{-1/3}h$ ,

$$\begin{aligned} n^{1/3} \left( \underset{s}{\operatorname{argmin}} \left\{ V_n(s) - (F(t) + xn^{-1/3})G_n(s) \right\} - t \right) \\ = \underset{h}{\operatorname{argmin}} \left\{ n^{2/3} (\mathbb{P}_n - P)(1_A - F(t)) (1_{\mathbb{R} \times [0, t + hn^{-1/3}]} - 1_{\mathbb{R} \times [0, t]}) \right. \\ \quad + n^{2/3} P(1_A - F(t)) (1_{\mathbb{R} \times [0, t + hn^{-1/3}]} - 1_{\mathbb{R} \times [0, t]}) \\ \quad \left. - xn^{1/3} \mathbb{P}_n (1_{\mathbb{R} \times [0, t + hn^{-1/3}]} - 1_{\mathbb{R} \times [0, t]}) \right\}. \end{aligned}$$

Let  $\mathbb{Z}$  be a two-sided Brownian motion process, originating from zero. By Theorems 2.11.22 and 2.11.23, the sequence of processes inside the argmin converges, for every  $K$ , in the space  $\ell^\infty(-K, K)$  to the process

$$h \mapsto \sqrt{F(1 - F)g(t)} \mathbb{Z}(h) + \frac{1}{2} h^2 f(t)g(t) - xg(t)h.$$

The locations of the minima converge to the location of the minimum  $\hat{h}$  of this limit process provided the sequence of locations of minima is bounded in probability. The final conclusion is

$$n^{1/3} (\hat{F}_n(t) - F(t)) \rightsquigarrow \left( \frac{4F(1 - F)f}{g}(t) \right)^{1/3} \underset{h}{\operatorname{argmin}} \{ \mathbb{Z}(h) + h^2 \}.$$

Here we have used Problem 3.2.5 to rewrite the probability  $P(\hat{h} \geq 0)$  in a simple form.

Uniform tightness of the locations of the minima can be shown with the help of Theorem 3.2.5. Define the processes

$$\begin{aligned} \mathbb{M}_n(h) &= (\mathbb{P}_n - P)(1_A - F(t)) (1_{\mathbb{R} \times [0, t+h]} - 1_{\mathbb{R} \times [0, t]}) \\ &\quad + P(1_A - F(t)) (1_{\mathbb{R} \times [0, t+h]} - 1_{\mathbb{R} \times [0, t]}) \\ &\quad - xn^{-1/3} \mathbb{P}_n (1_{\mathbb{R} \times [0, t+h]} - 1_{\mathbb{R} \times [0, t]}), \\ \mathbb{M}(h) &= P(1_A - F(t)) (1_{\mathbb{R} \times [0, t+h]} - 1_{\mathbb{R} \times [0, t]}). \end{aligned}$$

The conditions of Theorem 3.2.5 are satisfied with  $\phi_n(\delta) = \sqrt{\delta} + x\delta n^{1/6}$ . Thus, the location of the maximum of  $h \mapsto \mathbb{M}_n(h)$  has rate of convergence  $n^{1/3}$ , provided the sequence converges to zero in probability. The latter can be shown by a standard consistency proof.

### 3.2.4 Linearization

A (more) classical approach toward proving asymptotic normality of  $M$ -estimators is through a Taylor expansion of the criterion function. Assume that the limit criterion function  $\theta \mapsto Pm_\theta$  takes its maximum at a point  $\theta_0$ ; then its first derivative must vanish at  $\theta_0$ , and the second derivative  $V$  must be negative definite. If the function  $\theta \mapsto m_\theta$  is several times differentiable, then since  $\mathbb{P}_n = P + n^{-1/2}\mathbb{G}_n$ ,

$$\begin{aligned} n\mathbb{P}_n(m_\theta - m_{\theta_0}) &= nP(m_\theta - m_{\theta_0}) + \sqrt{n}\mathbb{G}_n(m_\theta - m_{\theta_0}) \\ &\approx \frac{1}{2}\sqrt{n}(\theta - \theta_0)'V\sqrt{n}(\theta - \theta_0)' + \sqrt{n}(\theta - \theta_0)'\mathbb{G}_n\dot{m}_{\theta_0} \\ &\quad + n o(\|\theta - \theta_0\|^2) + \sqrt{n}o_P(\|\theta - \theta_0\|). \end{aligned}$$

If the two remainder terms are neglected, then the maximum of the right-hand side is taken for  $\sqrt{n}(\theta - \theta_0) = -V^{-1}\mathbb{G}_n\dot{m}_{\theta_0}$ . Thus, it may be expected that the  $M$ -estimator  $\hat{\theta}_n$  that maximizes the left-hand side satisfies

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V^{-1}\mathbb{G}_n\dot{m}_{\theta_0} + o_P(1).$$

This derivation can be made rigorous by various methods. One possibility is to impose “classical” smoothness conditions on the maps  $\theta \mapsto m_\theta(x)$  (for every  $x$ ) in combination with domination of derivatives. More general results can be obtained under the assumption that the criterion functions are differentiable in a stochastic sense.

The basic result does not require the i.i.d. set-up and will be stated for general criterion functions. Let  $\hat{\theta}_n$  maximize the random function  $\theta \mapsto \mathbb{M}_n(\theta)$  which is “centered” at the deterministic function  $\theta \mapsto \mathbb{M}(\theta)$ . Assume that the sequence  $\hat{\theta}_n$  converges to a point of maximum  $\theta_0$  of  $\theta \mapsto \mathbb{M}(\theta)$ .

**3.2.16 Theorem.** *Let  $\mathbb{M}_n$  be stochastic processes indexed by an open subset  $\Theta$  of Euclidean space and  $\mathbb{M}: \Theta \mapsto \mathbb{R}$  a deterministic function. Assume that  $\theta \mapsto \mathbb{M}(\theta)$  is twice continuously differentiable at a point of maximum  $\theta_0$  with nonsingular second-derivative matrix  $V$ .<sup>†</sup> Suppose that*

$$\begin{aligned} r_n(\mathbb{M}_n - \mathbb{M})(\tilde{\theta}_n) - r_n(\mathbb{M}_n - \mathbb{M})(\theta_0) \\ = (\tilde{\theta}_n - \theta_0)'Z_n + o_P^*(\|\tilde{\theta}_n - \theta_0\| + r_n\|\tilde{\theta}_n - \theta_0\|^2 + r_n^{-1}), \end{aligned}$$

for every random sequence  $\tilde{\theta}_n = \theta_0 + o_P^*(1)$  and a uniformly tight sequence of random vectors  $Z_n$ . If the sequence  $\hat{\theta}_n$  converges in outer probability to  $\theta_0$  and satisfies  $\mathbb{M}_n(\hat{\theta}_n) \geq \sup_\theta \mathbb{M}_n(\theta) - o_P(r_n^{-2})$  for every  $n$ , then

$$r_n(\hat{\theta}_n - \theta_0) = -V^{-1}Z_n + o_P^*(1).$$

If it is known that the sequence  $r_n(\hat{\theta}_n - \theta_0)$  is uniformly tight, then the displayed condition needs to be verified for sequences  $\tilde{\theta}_n = \theta_0 + O_P^*(r_n^{-1})$  only.

---

<sup>†</sup> It suffices that a two-term Taylor expansion is valid at  $\theta_0$ .

**Proof.** The stochastic differentiability condition of the theorem together with the two-times differentiability of the map  $\theta \mapsto \mathbb{M}(\theta)$  yields for every sequence  $\tilde{h}_n = o_P^*(1)$

$$(3.2.17) \quad \begin{aligned} \mathbb{M}_n(\theta_0 + \tilde{h}_n) - \mathbb{M}_n(\theta_0) &= \frac{1}{2}\tilde{h}'_n V \tilde{h}_n + r_n^{-1}\tilde{h}'_n Z_n \\ &\quad + o_P^*(\|\tilde{h}_n\|^2 + r_n^{-1}\|\tilde{h}_n\| + r_n^{-2}). \end{aligned}$$

For  $\tilde{h}_n$  chosen equal to  $\hat{h}_n = \hat{\theta}_n - \theta_0$ , the left side (and hence the right side) is at least  $-o_P(r_n^{-2})$  by the definition of  $\hat{\theta}_n$ . In the right side the term  $\tilde{h}'_n V \tilde{h}_n$  can be bounded above by  $-c\|\tilde{h}_n\|^2$  for a positive constant  $c$ , since the matrix  $V$  is strictly negative definite. Conclude that

$$-o_P(r_n^{-2}) \leq -c\|\hat{h}_n\|^2 + r_n^{-1}\|\hat{h}_n\|O_P(1) + o_P(\|\hat{h}_n\|^2 + r_n^{-2}).$$

Complete the square to see that this implies that

$$(c + o_P(1)) \left( \|\hat{h}_n\| - O_P(r_n^{-1}) \right)^2 \leq O_P(r_n^{-2}).$$

This can be true only if  $\|\hat{h}_n\| = O_P^*(r_n^{-1})$ .

For any sequence  $\tilde{h}_n$  of the order  $O_P^*(r_n^{-1})$ , the three parts of the remainder term in (3.2.17) are of the order  $o_P(r_n^{-2})$ . Apply this with the choices  $\hat{h}_n$  and  $-r_n^{-1}V^{-1}Z_n$  to conclude that

$$\begin{aligned} \mathbb{M}_n(\theta_0 + \hat{h}_n) - \mathbb{M}_n(\theta_0) &= \frac{1}{2}\hat{h}'_n V \hat{h}_n + r_n^{-1}\hat{h}'_n Z_n + o_P^*(r_n^{-2}), \\ \mathbb{M}_n(\theta_0 - r_n^{-1}V^{-1}Z_n) - \mathbb{M}_n(\theta_0) &= -\frac{1}{2}r_n^{-2}Z'_n V^{-1}Z_n + o_P^*(r_n^{-2}). \end{aligned}$$

The left-hand side of the first equation is larger than the second, up to an  $o_P^*(r_n^{-2})$ -term. Subtract the second equation from the first to find that

$$\frac{1}{2}(\hat{h}_n + r_n^{-1}V^{-1}Z_n)'V(\hat{h}_n + r_n^{-1}V^{-1}Z_n) \geq -o_P(r_n^{-2}).$$

Since  $V$  is strictly negative definite, this yields the first assertion of the theorem.

If it is known that the sequence  $\hat{\theta}_n$  is  $r_n$ -consistent, then the middle part of the preceding proof is unnecessary and we can proceed to inserting  $\hat{h}_n$  and  $-r_n^{-1}V^{-1}Z_n$  in (3.2.17) immediately. The latter equation is then needed for sequences  $\tilde{h}_n = O_P^*(r_n^{-1})$  only. ■

The main condition of the theorem could be described as stochastic differentiability of the standardized process  $r_n(\mathbb{M}_n - \mathbb{M})$ , requiring a linear approximation with remainder of order

$$o_P(\|\tilde{\theta}_n - \theta_0\| + r_n\|\tilde{\theta}_n - \theta_0\|^2 + r_n^{-1}).$$

To require a remainder of the order  $o_P(\|\tilde{\theta}_n - \theta_0\|)$  would be more natural, but also more stringent: if  $\tilde{\theta}_n$  converges to  $\theta_0$  slower or faster than  $O_P(r_n^{-1})$ , then the sum in the previous display is dominated by its second and third

terms, respectively. Thus, in these cases the differentiability as required by the theorem is less stringent than the “natural” condition.

The last statement of the theorem is of some interest, because it also allows one in this approach to separate establishing the rate of convergence from the derivation of the limit distribution. The rate of convergence may be derived with the help of Theorem 3.2.5. For sequences  $\tilde{\theta}_n = \theta_0 + O_P(r_n^{-1})$  the stochastic differentiability condition reduces to<sup>†</sup>

$$r_n(\mathbb{M}_n - \mathbb{M})(\tilde{\theta}_n) - r_n(\mathbb{M}_n - \mathbb{M})(\theta_0) = (\tilde{\theta}_n - \theta_0)'Z_n + o_P^*(r_n^{-1}).$$

This condition together with  $r_n$ -consistency of  $\tilde{\theta}_n$  yield the conclusion of the theorem if the map  $\theta \mapsto \mathbb{M}(\theta)$  is two times differentiable with nonsingular second derivative matrix.

In the case of i.i.d. observations, the theorem can be applied with criterion functions  $\mathbb{M}_n(\theta) = \mathbb{P}_n m_\theta$  and  $\mathbb{M}(\theta) = Pm_\theta$  and rate of convergence  $r_n = \sqrt{n}$ . In this case  $\sqrt{n}(\mathbb{M}_n - \mathbb{M})(\theta) = \mathbb{G}_n m_\theta$  for the empirical process  $\mathbb{G}_n$ . Then the stochastic differentiability can be ascertained by empirical process methods. It is natural to require that each  $Z_n$  will be a centered and scaled sum, in which case the stochastic differentiability condition can be recast in terms of the existence of a vector-valued function  $\dot{m}_{\theta_0}$  (not necessarily a pointwise partial derivative) such that, for every  $\tilde{\theta}_n = \theta_0 + o_P^*(1)$ ,

$$(3.2.18) \quad \begin{aligned} \mathbb{G}_n(m_{\tilde{\theta}_n} - m_{\theta_0}) &= (\tilde{\theta}_n - \theta_0)' \mathbb{G}_n \dot{m}_{\theta_0} \\ &+ o_P^*(\|\tilde{\theta}_n - \theta_0\| + \sqrt{n}\|\tilde{\theta}_n - \theta_0\|^2 + n^{-1/2}). \end{aligned}$$

The following lemma gives a simple sufficient condition for this type of differentiability.

**3.2.19 Lemma.** Suppose that there exists a vector-valued function  $\dot{m}_{\theta_0}$  such that, for some  $\delta > 0$ ,

$$\left\{ \frac{m_\theta - m_{\theta_0} - (\theta - \theta_0)' \dot{m}_{\theta_0}}{\|\theta - \theta_0\|}: \quad \|\theta - \theta_0\| < \delta \right\} \quad \text{is } P\text{-Donsker,}$$

$$P[m_\theta - m_{\theta_0} - (\theta - \theta_0)' \dot{m}_{\theta_0}]^2 = o(\|\theta - \theta_0\|^2).$$

Then (3.2.18) is satisfied for every sequence  $\tilde{\theta}_n = \theta_0 + o_P^*(1)$ . Consequently, if the map  $\theta \mapsto Pm_\theta$  is twice continuously differentiable at  $\theta_0$  with nonsingular second derivative matrix<sup>‡</sup> and  $\tilde{\theta}_n$  converges to  $\theta_0$  in outer probability, then  $\sqrt{n}(\tilde{\theta}_n - \theta_0) = -V^{-1}\mathbb{G}_n \dot{m}_{\theta_0} + o_P^*(1)$ .

**Proof.** Without loss of generality, assume that  $\tilde{\theta}_n$  takes its values in  $\Theta_\delta = \{\theta: \|\theta - \theta_0\| < \delta\}$ . Define a function  $f: \ell^\infty(\Theta_\delta) \times \Theta_\delta \mapsto \mathbb{R}^d$  by  $f(z, \theta) = z(\theta)$ .

---

<sup>†</sup> The presence of  $r_n^{-1}$  in the remainder term of the stochastic differentiability condition of the theorem ensures that one need not worry about sequences  $\tilde{\theta}_n$  that converge to  $\theta_0$  too fast.

<sup>‡</sup> It suffices that a two-term Taylor expansion is valid at  $\theta_0$ .

This function is continuous at every point  $(z, \theta_0)$  such that  $\theta \mapsto z(\theta)$  is continuous at  $\theta_0$ .

Define a stochastic process  $Z_n$  indexed by  $\Theta_\delta$  by

$$Z_n(\theta) = \mathbb{G}_n \frac{m_\theta - m_{\theta_0} - (\theta - \theta_0)' \dot{m}_{\theta_0}}{\|\theta - \theta_0\|}.$$

By assumption, the sequence  $Z_n$  converges in  $\ell^\infty(\Theta_\delta)$  to a tight Gaussian process  $Z$ . This process has continuous sample paths with respect to the semimetric  $\rho$  given by

$$\rho^2(\theta_1, \theta_2) = \mathbb{E}(Z(\theta_1) - Z(\theta_2))^2.$$

By assumption,  $\rho(\theta, \theta_0) \rightarrow 0$  if  $\theta \rightarrow \theta_0$ . Thus, almost all sample paths of  $Z$  are continuous at  $\theta_0$ .

By Slutsky's lemma  $(Z_n, \tilde{\theta}_n) \rightsquigarrow (Z, \theta_0)$ . By the continuous mapping theorem,  $Z_n(\tilde{\theta}_n) = f(Z_n, \tilde{\theta}_n) \rightsquigarrow f(Z, \theta_0) = 0$ . Since convergence in distribution to a constant is the same as convergence in probability, it follows that (3.2.18) holds with remainder term  $o_P^*(\|\tilde{\theta}_n - \theta_0\|)$ . ■

The conditions of the lemma require some type of differentiability of the map  $\theta \mapsto m_\theta$  but are weak enough to allow the treatment of slightly irregular functions, such as the absolute value  $x \mapsto |x - \theta|$ .

The presence of the factor  $\|\theta - \theta_0\|$  in the denominator appears a bit awkward. It would be unnecessary to handle this term if the rate of convergence could first be established by a different method. Thus, that may motivate separating also in this approach deriving the limit distribution from establishing the rate of convergence. If it is known that the maximizing sequence  $\hat{\theta}_n$  is  $\sqrt{n}$ -consistent, then (3.2.18) need hold only for sequences  $\tilde{\theta}_n = \theta_0 + O_P^*(n^{-1/2})$  and can be replaced by

$$(3.2.20) \quad \mathbb{G}_n \sqrt{n}(m_{\tilde{\theta}_n} - m_{\theta_0}) = \sqrt{n}(\tilde{\theta}_n - \theta_0)' \mathbb{G}_n \dot{m}_{n, \theta_0} + o_P^*(1).$$

The function  $\dot{m}_{n, \theta_0}$  may depend on  $n$ ; typically it would be equal almost everywhere to a derivative of  $m_\theta$  and fixed, but we could even try functions such as  $\sqrt{n}(m_{\theta_0 + e_i n^{-1/2}} - m_{\theta_0})$ .<sup>#</sup>

The following lemma gives a simple sufficient condition for this type of differentiability.

<sup>#</sup> If any functions  $\dot{m}_{n, \theta_0}$  work, then so do the functions  $\sqrt{n}(m_{\theta_0 + e_i n^{-1/2}} - m_{\theta_0})$ .

**3.2.21 Lemma.** Suppose that the classes of functions  $\mathcal{M}_\delta = \{m_\theta - m_{\theta_0} : \|\theta - \theta_0\| < \delta\}$  satisfy the entropy condition (3.2.8) or (3.2.9) and have envelope functions  $M_\delta$  such that  $P^* M_\delta^2 \{M_\delta > \eta\} = o(\delta^2)$  for every  $\eta > 0$ , and

$$\lim_{\varepsilon \downarrow 0} \limsup_{\delta \downarrow 0} \sup_{\substack{\|h-g\| < \varepsilon \\ \|g\| \vee \|h\| \leq K}} \frac{P(m_{\theta_0+\delta g} - m_{\theta_0+\delta h})^2}{\delta^2} = 0,$$

$$\lim_{\delta \downarrow 0} \frac{P(m_{\theta_0+\delta g} - m_{\theta_0+\delta h})^2}{\delta^2} = E(G(h) - G(g))^2,$$

for all  $K$  and some linear, zero-mean Gaussian process  $G$ . Then (3.2.20) holds for every sequence  $\hat{\theta}_n = \theta_0 + o_P^*(n^{-1/2})$ . Consequently, if the map  $\theta \mapsto Pm_\theta$  is twice continuously differentiable at  $\theta_0$  with nonsingular derivative and  $\hat{\theta}_n$  converges to  $\theta_0$  in outer probability, then  $\sqrt{n}(\hat{\theta}_n - \theta_0) = -V^{-1}\mathbb{G}_n \dot{m}_{n,\theta_0} + o_P^*(1)$ .

**Proof.** For the first assertion it suffices to show that the sequence of processes

$$h \mapsto \mathbb{G}_n(\sqrt{n}(m_{\theta_0+h/\sqrt{n}} - m_{\theta_0}) - h' \dot{m}_{n,\theta_0})$$

converges in probability to zero in the space  $\ell^\infty(h : \|h\| \leq K)$ , for every  $K$ . The sequence of processes  $h \mapsto \mathbb{G}_n \sqrt{n}(m_{\theta_0+h/\sqrt{n}} - m_{\theta_0})$  can be shown to be asymptotically tight by application of Theorem 2.11.22 or 2.11.23. A fortiori the sequence  $\mathbb{G}_n \dot{m}_{n,\theta_0}$  and hence the sequence of processes  $h \mapsto h' \dot{m}_{n,\theta_0}$  are asymptotically tight. It now suffices to show that the processes in the preceding display converge marginally to zero. By the definition of  $\dot{m}_{n,\theta_0}$ ,

$$P\left(\sqrt{n}(m_{\theta_0+h/\sqrt{n}} - m_{\theta_0}) - h' \dot{m}_{n,\theta_0}\right)^2 \rightarrow E\left(G(h) - \sum h_i G(e_i)\right)^2.$$

This is zero for every  $h$  by the linearity of  $G$ . An application of Markov's inequality yields marginal convergence to zero and concludes the proof of the first assertion.

For the final assertion of the lemma, note first that the rate of convergence of  $\hat{\theta}_n$  is  $\sqrt{n}$  by Theorem 3.2.5 (or Theorem 3.2.10). Thus, the conclusion follows by Theorem 3.2.16. ■

The second and third lines of the display in the lemma are implied by the single condition: for every  $g_\delta \rightarrow g$  and  $h_\delta \rightarrow h$ ,

$$\lim_{\delta \downarrow 0} \frac{1}{\delta^2} P(m_{\theta_0+\delta g_\delta} - m_{\theta_0+\delta h_\delta})^2 = E(G(g) - G(h))^2.$$

Instead of the variances of the differences, the condition could also be stated in terms of the covariances of  $m_{\theta_1}$  and  $m_{\theta_2}$ . In particular, the condition is implied by twice differentiability of the map  $(\theta_1, \theta_2) \mapsto Pm_{\theta_1} m_{\theta_2}$  at  $(\theta_0, \theta_0)$ . See Problem 3.2.1.

**3.2.22 Example (Lipschitz in parameter).** Let  $X_1, \dots, X_n$  be i.i.d. random variables with common law  $P$ , and let  $m_\theta$  be measurable functions indexed by a parameter  $\theta$  ranging over an open subset of Euclidean space. Assume that, for every  $\theta_1, \theta_2$  in a neighborhood of  $\theta_0$ ,

$$\begin{aligned} |m_{\theta_1}(x) - m_{\theta_2}(x)| &\leq \dot{m}(x)\|\theta_1 - \theta_2\|, \\ P[m_\theta - m_{\theta_0} - (\theta - \theta_0)' \dot{m}_{\theta_0}]^2 &= o(\|\theta - \theta_0\|^2), \end{aligned}$$

for functions  $\dot{m}$  and  $m_{\theta_0}$  with  $P\dot{m}^2(x) < \infty$ . Furthermore, assume that the map  $\theta \mapsto Pm_\theta$  is twice continuously differentiable at a point of maximum  $\theta_0$  with nonsingular second-derivative matrix  $V$ . If  $\hat{\theta}_n$  maximizes  $\theta \mapsto \mathbb{P}_n m_\theta$  up to an  $o_P(1/n)$ -term and is consistent for  $\theta_0$ , then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -V^{-1}\mathbb{G}_n \dot{m}_{\theta_0} + o_P^*(1).$$

This follows from the preceding lemma. By the Lipschitz condition, the  $L_2(P)$ -covering numbers of the classes  $\mathcal{M}_\delta$  are polynomial and the envelopes can be taken equal to  $M = \dot{m}\delta$ . In view of the differentiability of  $\theta \mapsto m_\theta$  in square mean, the process  $G$  can be taken as  $G(h) = h'\Delta$  for normally distributed vector  $\Delta$  with mean zero and covariance matrix  $P\dot{m}_{\theta_0} \dot{m}'_{\theta_0}$ .

It may be noted that  $P\dot{m}_{\theta_0} = 0$ , because this is the first derivative at  $\theta_0$  of the map  $\theta \mapsto Pm_\theta$ . Hence the sequence  $\mathbb{G}_n \dot{m}_{\theta_0}$  is asymptotically zero-mean normal.

**3.2.23 Example (Least absolute deviation regression).** Given independent i.i.d. random vectors  $X_1, \dots, X_n$  and  $e_1, \dots, e_n$  in  $\mathbb{R}^d$  and  $\mathbb{R}$ , respectively, let  $Y_i = \theta'_0 X_i + e_i$ . The least-absolute-deviation estimator  $\hat{\theta}_n$  minimizes the function

$$\theta \mapsto \frac{1}{n} \sum_{i=1}^n |Y_i - \theta' X_i| = \mathbb{P}_n m_\theta,$$

where  $\mathbb{P}_n$  is the empirical measure of the pairs  $(X_i, Y_i)$  and  $m_\theta(x, y) = |y - \theta' x|$ .

The parameter  $\theta_0$  is a point of minimum of the map  $\theta \mapsto P|Y - \theta' X|$  if the distribution of the errors  $e_i$  has median zero. (Write the expectation as  $E_X E_{e_i} |e_i - (\theta - \theta_0)' X_i|$  and note that the inner integral is minimized by  $\theta_0$  for every fixed value of  $X$ .) Furthermore, the maps  $\theta \mapsto m_\theta$  satisfy the conditions in Example 3.2.22:

$$\begin{aligned} ||y - \theta'_1 x| - |y - \theta'_2 x|| &\leq \|\theta_1 - \theta_2\| \|x\|, \\ P[|Y - \theta' X| - |Y - \theta'_0 X| - (\theta - \theta_0)' X \operatorname{sign}(Y - \theta'_0 X)]^2 &= o(\|\theta - \theta_0\|^2). \end{aligned}$$

The consistency of the least-absolute-deviation estimator can be argued from the convexity of the map  $\theta \mapsto |y - \theta' x|$ . The map  $\theta \mapsto P|Y - \theta' X|$

is twice differentiable at  $\theta_0$  if the distribution of the errors has a positive density at its median.

The final conclusion is that the sequence  $\sqrt{n}(\hat{\theta}_n - \theta_0)$  is asymptotically normal with mean zero and covariance matrix  $V^{-1}E(XX')\text{sign}(Y - \theta'_0 X)V^{-1}$  for  $V$ , the second-derivative matrix of  $\theta \mapsto P|Y - \theta'X|$  at  $\theta_0$ .

**3.2.24 Example (Maximum likelihood).** Let  $X_1, \dots, X_n$  be a sample from a density  $p_\theta$  with respect to a measure  $\mu$  on a measurable space  $(\mathcal{X}, \mathcal{A})$ . The parameter  $\theta$  ranges over an open subset of Euclidean space. Assume that, for  $\theta_1, \theta_2$  in a neighborhood of  $\theta_0$ ,

$$|\log p_{\theta_1}(x) - \log p_{\theta_2}(x)| \leq \dot{\ell}(x) \|\theta_1 - \theta_2\|,$$

for a function  $\dot{\ell}$  with  $P_{\theta_0}\dot{\ell}^2 < \infty$ . Furthermore, assume that the map  $\theta \mapsto P_{\theta_0} \log p_\theta$  is twice continuously differentiable at  $\theta_0$  and that as  $\theta \mapsto \theta_0$

$$\int \left[ p_\theta^{1/2} - p_{\theta_0}^{1/2} - \frac{1}{2}(\theta - \theta_0)' \dot{\ell}_{\theta_0} p_{\theta_0}^{1/2} \right]^2 d\mu = o(\|\theta - \theta_0\|^2),$$

for some measurable vector-valued function  $\dot{\ell}_{\theta_0}$ . Under these conditions, if the maximum likelihood estimator  $\hat{\theta}_n$  for  $\theta_0$  is consistent, then

$$\sqrt{n}(\hat{\theta}_n - \theta_0) = -I_{\theta_0}^{-1} \frac{1}{\sqrt{n}} \sum_{i=1}^n \dot{\ell}_{\theta_0}(X_i) + o_{P_{\theta_0}}^*(1).$$

Here the Fisher information matrix  $I_{\theta_0} = P_{\theta_0} \dot{\ell}_{\theta_0} \dot{\ell}'_{\theta_0}$  is assumed to be non-singular.

Since the maximum likelihood estimator is an  $M$ -estimator for  $m_\theta = \log p_\theta$ , its asymptotic normality also follows under the conditions of Example 3.2.22 for general Lipschitz criterion functions. The present statement is better in two respects. First, the differentiability of the log density in quadratic mean is replaced by the differentiability of the root density, which has long been known to be the “right” condition for asymptotic theory. Second, the present statement gives the inverse of the Fisher information matrix as the asymptotic covariance matrix, rather than a matrix  $V^{-1}I_{\theta_0}V^{-1}$ , which would next have to be checked to be equal to  $I_{\theta_0}^{-1}$  under additional conditions.

The claim can be proved as follows. Define a zero-mean stochastic process  $\mathbb{Z}_n$  by

$$\mathbb{Z}_n(h) = n(\mathbb{P}_n - P_{\theta_0}) \left( \log \frac{p_{\theta_0+h/\sqrt{n}}}{p_{\theta_0}} - \frac{h' \dot{\ell}_{\theta_0}}{\sqrt{n}} \right).$$

The differentiability of the root density implies that the score is mean-zero:  $P_{\theta_0} \dot{\ell}_{\theta_0} = 0$  as well as local asymptotic normality. The latter means that for every  $h$ ,

$$\begin{aligned} \mathbb{Z}_n(h) &= -\frac{1}{2} h' I_{\theta_0} h - n P_{\theta_0} \log \frac{p_{\theta_0+h/\sqrt{n}}}{p_{\theta_0}} + o_{P_{\theta_0}}(1) \\ &= -\frac{1}{2} h' I_{\theta_0} h - \frac{1}{2} h' V h + o_{P_{\theta_0}}(1). \end{aligned}$$

Here  $V$  is the second-derivative matrix of the map  $\theta \mapsto P_{\theta_0} \log p_\theta$  at  $\theta = \theta_0$ . The second moment of  $\mathbb{Z}_n(h)$  is equal to its variance and satisfies

$$\mathbb{E}\mathbb{Z}_n^2(h) \leq P_{\theta_0} \left( \sqrt{n} \log \frac{p_{\theta_0+h/\sqrt{n}}}{p_{\theta_0}} - h' \dot{\ell}_{\theta_0} \right)^2 \lesssim P_{\theta_0} (\dot{\ell}^2 + \|\dot{\ell}_{\theta_0}\|^2) \|h\|^2.$$

Since this is uniformly bounded in  $n$ , it follows that the sequence of means  $\mathbb{E}\mathbb{Z}_n(h) = 0$  converges to the mean of the limit in probability,  $-\frac{1}{2}h'Vh - \frac{1}{2}h'Vh$  of the sequence  $\mathbb{Z}_n(h)$ . Since this is true for every  $h$ , it follows that  $V = -I_{\theta_0}$ .

In view of Corollary 3.2.6, Theorem 3.2.10, or Example 3.2.12, the rate of the maximum likelihood estimator is  $\sqrt{n}$ . In view of the discussion following Theorem 3.2.16, it now suffices to verify condition (3.2.20) with  $\dot{m}_{n,\theta_0} = \dot{\ell}_{\theta_0}$ . In the present situation this is implied by the sequence of random variables  $\sup_{\|h\| \leq K} |\mathbb{Z}_n(h)|$  converging to zero in distribution for every  $K$ . Since it was seen that the sequence of processes  $\{\mathbb{Z}_n(h): \|h\| \leq K\}$  converges marginally in probability to zero, it suffices to show that the sequence is asymptotically tight in the space of bounded functions. In view of the Lipschitz condition on the maps  $\theta \mapsto \log p_\theta$ , the bracketing numbers  $N_{[]}(\varepsilon \|M_n\|_{P_{\theta_0,2}}, \mathcal{M}_n, L_2(P_{\theta_0}))$  of the classes of functions

$$\mathcal{M}_n = \left\{ \sqrt{n} \log \frac{p_{\theta_0+h/\sqrt{n}}}{p_{\theta_0}} : \|h\| \leq K \right\}, \quad M_n = K\dot{\ell},$$

are uniformly bounded by the covering numbers of a Euclidean ball. The asymptotic tightness follows from Theorem 2.11.23.

## Problems and Complements

- Assume one of the entropy conditions (3.2.8) or (3.2.9) for the class of functions  $\mathcal{M}_\delta = \{m_\theta - m_{\theta_0} : \|\theta - \theta_0\| < \delta\}$  with envelope function  $M_\delta$ . Furthermore, suppose that, for every  $\eta > 0$ ,

$$P^* M_\delta^2 \leq \delta^2; \quad P^* M_\delta^2 \{M_\delta > \eta\} = o(\delta^2).$$

Finally, suppose that the maps  $\theta \mapsto Pm_\theta$  and  $(\theta_1, \theta_2) \mapsto Pm_{\theta_1}m_{\theta_2}$  are twice continuously differentiable at  $\theta_0$  and  $(\theta_0, \theta_0)$ , respectively, the first one with nonsingular second-derivative matrix  $V$ . Then a (near) maximizer  $\hat{\theta}_n$  of  $\theta \mapsto \mathbb{P}_n m_\theta$  satisfies  $\sqrt{n}(\hat{\theta}_n - \theta_0) = -V^{-1}\mathbb{G}_n \dot{m}_{n,\theta_0} + o_P^*(1)$  provided it converges to  $\theta_0$  in outer probability. Here  $\mathbb{G}_n \dot{m}_{n,\theta_0}$  converges in distribution to a normal distribution with mean zero and covariance matrix the derivative  $\partial^2 / \partial \theta_1 \partial \theta_2 P m_{\theta_1} m_{\theta_2}$  evaluated at  $(\theta_0, \theta_0)$ .

- If  $M: \Theta \mapsto \mathbb{R}$  is upper semicontinuous on a compact metric space  $\Theta$ , then it achieves its maximum value. If it achieves its maximum value at a unique point  $\theta_0$ , then  $M(\theta_0) > \sup_{\theta \notin G} M(\theta)$  for every open set  $G$  around  $\theta_0$ .

- 3. (Consistency and Lipschitz criterion functions)** Let  $\Theta$  be a measurable subset of Euclidean space and  $M_n$  be separable stochastic processes that converge pointwise in probability to a fixed function  $M$ . Suppose that

$$|M_n(\theta_1) - M_n(\theta_2)| \leq \dot{M}_n \|\theta_1 - \theta_2\|,$$

for random variables such that  $\sup_n \dot{M}_n < \infty$  almost surely. Let  $\theta \mapsto M(\theta)$  be upper semicontinuous and possess a unique maximum at  $\theta_0$ . If  $\hat{\theta}_n$  (nearly) maximizes  $\theta \mapsto M_n(\theta)$  and  $\hat{\theta}_n = O_P(1)$ , then  $\hat{\theta}_n \rightarrow \theta_0$  in probability.

[Hint: Almost every sequence of sample paths  $\theta \mapsto M_n(\theta, \omega)$  is uniformly bounded and uniformly equicontinuous on compacta. By the Arzelà-Ascoli theorem almost every sequence is relatively compact for the topology of uniform convergence on compacta. Conclude that the sequence  $\|M_n - M\|_K$  converges to zero in probability for every compact  $K$ . Apply part (ii) of Corollary 3.2.3.]

- 4. (Consistency and concave criterion functions)** Let  $\Theta$  be a subset of Euclidean space and  $M_n$  be separable stochastic processes that converge pointwise in probability to a fixed function  $M$ . Let every  $M_n$  and  $M$  be strictly concave, and let the unique maximizer  $\theta_0$  of  $M$  be an interior point of  $\Theta$ . If  $\hat{\theta}_n$  (nearly) maximizes  $\theta \mapsto M_n(\theta)$ , then  $\hat{\theta}_n$  converges in probability to  $\theta_0$ .

[Hint: A sequence of concave functions  $M_n$  that converges pointwise on an open set automatically converges uniformly on every compact subset. In particular, it converges uniformly on a sufficiently small ball around the unique maximizer  $\theta_0$  of the limit  $M$ . Since  $M$  is continuous, the value  $M(\theta_0)$  is strictly larger than its maximum value on the edge of the ball. This implies that for each sufficiently large  $n$  the functions  $M_n$  have a local maximum in this ball. A concave function has no point of local maxima besides the point of global maximum. Conclude that the sequence  $\hat{\theta}_n$  is  $O_P(1)$  and apply part (ii) of Corollary 3.2.3. (The uniform convergence on compacta is a consequence of the fact that a concave function on an open set is automatically Lipschitz on every closed set inside the domain with Lipschitz constant depending on the supremum of the function over the open set and the distance from the closed set to the complement of the open set. Thus, the concavity in this problem is mostly used to obtain the uniform tightness of  $\hat{\theta}_n$ . The consistency follows from the more general argument of the preceding problem.)]

5. Let  $\{Z(h): h \in \mathbb{R}\}$  be a standard two-sided Brownian motion with  $Z(0) = 0$ . (The process is zero-mean Gaussian and the increment  $Z(g) - Z(h)$  has variance  $|g - h|$ .) Then  $\operatorname{argmax}_h \{aZ(h) - bh^2 - ch\}$  is equal in distribution to  $(a/b)^{2/3} \operatorname{argmax}_g \{Z(g) - g^2\} - \frac{1}{2}c/b$ . Here  $a, b$ , and  $c$  are positive constants.

[Hint: The process  $z \mapsto Z(\sigma h - \mu)$  is equal in distribution to the process  $h \mapsto \sqrt{\sigma}Z(h) - Z(\mu)$ . Apply the change of variable  $h \mapsto (a/b)^{2/3}g - \frac{1}{2}c/b$  and note that the location of a maximum does not change by multiplication by a positive constant or a vertical shift.]

## 3.3

# Z-Estimators

Let the parameter set  $\Theta$  be a subset of a Banach space, and let

$$\Psi_n: \Theta \mapsto \mathbb{L}, \quad \Psi: \Theta \mapsto \mathbb{L}$$

be random maps and a deterministic map, respectively, with values in another Banach space  $\mathbb{L}$ . Here “random maps” means that each  $\Psi_n(\theta)$  is defined on the product of  $\Theta$  and some probability space. The dependence on the probability space is suppressed in the notation.

In this chapter we prove asymptotic normality of “estimators”  $\hat{\theta}_n$  that (approximately) satisfy the equation

$$\Psi_n(\hat{\theta}_n) = 0.$$

Such estimators will be referred to as *Z-estimators*.

If  $\mathbb{L}$  is an  $\ell^\infty(\mathcal{H})$ -space, as can be assumed without loss of generality, the equation is equivalent to the collection of (real-valued) estimating equations  $\Psi_n(\hat{\theta}_n)h = 0$ , when  $h$  runs through  $\mathcal{H}$ . If  $\theta$  is finite-dimensional, then the number of estimating equations is typically chosen equal to the dimension, and the space  $\ell^\infty(\mathcal{F})$  can be identified with Euclidean space. In the case of an infinite-dimensional parameter, infinitely many estimating equations may be used. Letting  $\Psi$  be the “asymptotic version” of  $\Psi_n$  as  $n$  tends to infinity, we may hope that  $\hat{\theta}_n$  tends to a value  $\theta_0$  satisfying

$$\Psi(\theta_0) = 0.$$

In the following, “consistency”  $\hat{\theta}_n \xrightarrow{P^*} \theta_0$  is assumed from the start. We study conditions under which  $\sqrt{n}(\hat{\theta}_n - \theta_0)$  converges in distribution.

The set  $\Theta$  should be chosen to contain the “true value”  $\theta_0$  and the possible values of the estimators  $\hat{\theta}_n$ , at least asymptotically. The estimators must be defined on the same probability space as the criterion maps  $\Psi_n$ . They need not be defined by the estimating equation and need not be measurable. Also,  $\theta_0$  need not be an interior point of the parameter set  $\Theta$  and  $\Theta$  need not be a linear space.

We assume that the map  $\Psi$  is Fréchet-differentiable at  $\theta_0$ . This entails that

$$\|\Psi(\theta) - \Psi(\theta_0) - \dot{\Psi}_{\theta_0}(\theta - \theta_0)\| = o(\|\theta - \theta_0\|),$$

as  $\theta \rightarrow \theta_0$  within  $\Theta$ , for a continuous, linear, one-to-one map  $\dot{\Psi}_{\theta_0}: \text{lin } \Theta \mapsto \mathbb{L}$ . We also make the crucial assumption that the inverse  $\dot{\Psi}_{\theta_0}^{-1}$  of the derivative exists and is continuous on the range of  $\dot{\Psi}_{\theta_0}$ . A linear operator on a finite-dimensional space is automatically continuous; in this case the assumption is simply that  $\theta \mapsto \Psi(\theta)$  is differentiable at  $\theta_0$  with a nonsingular derivative matrix  $\dot{\Psi}_{\theta_0}$ . In the infinite-dimensional case, the assumption that the derivative  $\dot{\Psi}_{\theta_0}$  is continuously invertible is harder to ascertain. If the inverse  $\dot{\Psi}_{\theta_0}^{-1}$  is continuous on the range of  $\dot{\Psi}_{\theta_0}$ , then it has a unique continuous extension to the closure of this range. This will also be denoted  $\dot{\Psi}_{\theta_0}^{-1}$  and is the inverse of the unique continuous extension of  $\dot{\Psi}_{\theta_0}$  to the closure of  $\text{lin } \Theta$ .

**3.3.1 Theorem.** Let  $\Psi_n$  and  $\Psi$  be random maps and a fixed map, respectively, from  $\Theta$  into a Banach space such that

$$(3.3.2) \quad \sqrt{n}(\Psi_n - \Psi)(\hat{\theta}_n) - \sqrt{n}(\Psi_n - \Psi)(\theta_0) = o_P^*(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|),$$

and such that the sequence  $\sqrt{n}(\Psi_n - \Psi)(\theta_0)$  converges in distribution to a tight random element  $Z$ . Let  $\theta \mapsto \Psi(\theta)$  be Fréchet-differentiable at  $\theta_0$  with a continuously invertible derivative  $\dot{\Psi}_{\theta_0}$ . If  $\Psi(\theta_0) = 0$  and  $\hat{\theta}_n$  satisfies  $\Psi_n(\hat{\theta}_n) = o_P^*(n^{-1/2})$  and converges in outer probability to  $\theta_0$ , then

$$\sqrt{n}\dot{\Psi}_{\theta_0}(\hat{\theta}_n - \theta_0) = -\sqrt{n}(\Psi_n - \Psi)(\theta_0) + o_{P*}(1).$$

Consequently,  $\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow -\dot{\Psi}_{\theta_0}^{-1}Z$ . If it is known that the sequence  $\sqrt{n}\|\hat{\theta}_n - \theta_0\|$  is asymptotically tight, then the first conclusion is valid without the assumption of continuous invertibility of  $\dot{\Psi}_{\theta_0}$ . If it is known that  $\sqrt{n}(\hat{\theta}_n - \theta_0)$  is asymptotically tight, then it suffices that  $\Psi$  is Hadamard-differentiable.

**Proof.** By the definition of  $\hat{\theta}_n$

$$(3.3.3) \quad \begin{aligned} \sqrt{n}(\Psi(\hat{\theta}_n) - \Psi(\theta_0)) &= \sqrt{n}(\Psi(\hat{\theta}_n) - \Psi_n(\hat{\theta}_n)) + o_P^*(1), \\ &= -\sqrt{n}(\Psi_n - \Psi)(\theta_0) + o_P^*(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|), \end{aligned}$$

by the first displayed assumption of the theorem. Since the derivative  $\dot{\Psi}_{\theta_0}$  is continuously invertible, there exists a positive constant  $c$  such that

$\|\dot{\Psi}_{\theta_0}(\theta - \theta_0)\| \geq c\|\theta - \theta_0\|$  for every  $\theta$  and  $\theta_0$ . This can be combined with the differentiability of  $\Psi$  to yield

$$\|\Psi(\theta) - \Psi(\theta_0)\| \geq c\|\theta - \theta_0\| + o(\|\theta - \theta_0\|).$$

Apply this to (3.3.3) to find that

$$\sqrt{n}\|\hat{\theta}_n - \theta_0\|(c + o_P(1)) \leq O_P(1) + o_P(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|).$$

This implies that the sequence  $\hat{\theta}_n$  is  $\sqrt{n}$ -consistent for  $\theta_0$  in norm. By the differentiability of  $\Psi$ , the left side of (3.3.3) can be replaced by

$$\sqrt{n}\dot{\Psi}_{\theta_0}(\hat{\theta}_n - \theta_0) + o_P^*(\sqrt{n}\|\hat{\theta}_n - \theta_0\|).$$

The last term is  $o_P(1)$  as is the remainder on the right side of (3.3.3). The second assertion of the theorem follows. Next the continuity of  $\dot{\Psi}_{\theta_0}^{-1}$  and the continuous mapping theorem give the weak convergence of the sequence  $\sqrt{n}(\hat{\theta}_n - \theta_0)$ .

Under the additional assumptions on the sequence  $\hat{\theta}_n$ , the preceding proof can be simplified. ■

The preceding theorem separates the conditions for asymptotic normality into the stochastic condition (3.3.2), which requires that the remainder in a Taylor expansion is negligible, and analytical conditions on the asymptotic criterion function  $\Psi$ .

In the case of i.i.d. observations the theorem may be applied with  $\Psi_n(\theta)h = \mathbb{P}_n\psi_{\theta,h}$  and  $\Psi(\theta)h = P\psi_{\theta,h}$  for given measurable functions  $\psi_{\theta,h}$  indexed by  $\Theta$  and an arbitrary index set  $\mathcal{H}$ . In this case  $\sqrt{n}(\Psi_n - \Psi)(\theta) = \{\mathbb{G}_n\psi_{\theta,h}: h \in \mathcal{H}\}$  is the empirical process indexed by the class of functions  $\{\psi_{\theta,h}: h \in \mathcal{H}\}$ . Then the stochastic condition reduces to

$$(3.3.4) \quad \|\mathbb{G}_n(\psi_{\hat{\theta}_n,h} - \psi_{\theta_0,h})\|_{\mathcal{H}} = o_P^*(1 + \sqrt{n}\|\hat{\theta}_n - \theta_0\|).$$

It is a stronger condition to require that the left-hand side is  $o_P^*(1)$ . The following lemma gives a simple sufficient condition.

**3.3.5 Lemma.** Suppose that the class of functions

$$\left\{ \psi_{\theta,h} - \psi_{\theta_0,h}: \|\theta - \theta_0\| < \delta, h \in \mathcal{H} \right\}$$

is  $P$ -Donsker for some  $\delta > 0$  and that

$$\sup_{h \in \mathcal{H}} P(\psi_{\theta,h} - \psi_{\theta_0,h})^2 \rightarrow 0, \quad \theta \rightarrow \theta_0.$$

If  $\hat{\theta}_n$  converges in outer probability to  $\theta_0$ , then (3.3.4) is satisfied.

**Proof.** Without loss of generality, assume that  $\hat{\theta}_n$  takes its values in  $\Theta_\delta = \{\theta: \|\theta - \theta_0\| < \delta\}$ . Define a function  $f: \ell^\infty(\Theta_\delta \times \mathcal{H}) \times \Theta_\delta \mapsto \ell^\infty(\mathcal{H})$  by

$f(z, \theta)h = z(\theta, h)$ . This function is continuous at every point  $(z, \theta_0)$  such that  $\|z(\theta, h) - z(\theta_0, h)\|_{\mathcal{H}} \rightarrow 0$  as  $\theta \rightarrow \theta_0$ .

Define a stochastic process  $Z_n$  indexed by  $\Theta_\delta \times \mathcal{H}$  by

$$Z_n(\theta, h) = \mathbb{G}_n(\psi_{\theta, h} - \psi_{\theta_0, h}).$$

By assumption, the sequence  $Z_n$  converges in  $\ell^\infty(\Theta_\delta \times \mathcal{H})$  to a tight Gaussian process  $Z$ . This process has continuous sample paths with respect to the semimetric  $\rho$  given by

$$\rho^2((\theta_1, h_1), (\theta_2, h_2)) = P(\psi_{\theta_1, h_1} - \psi_{\theta_0, h_1} - \psi_{\theta_2, h_2} + \psi_{\theta_0, h_2})^2.$$

By assumption,  $\sup_h \rho(\theta, h), (\theta_0, h) \rightarrow 0$  if  $\theta \rightarrow \theta_0$ . Conclude that the function  $f$  is continuous at almost all sample paths of  $Z$ .

By Slutsky's lemma,  $(Z_n, \hat{\theta}_n) \rightsquigarrow (Z, \theta_0)$ . By the continuous mapping theorem,  $Z_n(\hat{\theta}_n) = f(Z_n, \hat{\theta}_n) \rightsquigarrow f(Z, \theta_0) = 0$  in  $\ell^\infty(\mathcal{H})$ . ■

It appears that the conditions of the preceding lemma are relatively weak. In the case of infinite-dimensional parameters, verification of the analytical conditions may require more effort. The assertion of the theorem becomes

$$(3.3.6) \quad \sqrt{n}(\hat{\theta}_n - \theta_0) = -\dot{\Psi}_{\theta_0}^{-1} \mathbb{G}_n \psi_{\theta_0} + o_P^*(1),$$

where  $\dot{\Psi}_{\theta_0}$  is the derivative of the map  $\theta \mapsto P\psi_\theta$ . Note that  $P\psi_{\theta_0} = 0$  by the "definition" of  $\theta_0$ .

**3.3.7 Example (Lipschitz functions).** Let  $X_1, \dots, X_n$  be i.i.d. random variables with common law  $P$ , and let  $\psi_\theta$  be measurable vector-valued functions indexed by a Euclidean parameter  $\theta$ . (Thus, we take  $\mathcal{H}$  to be a finite set, whose cardinality is the dimension of  $\theta$ .) Assume that, for every  $\theta_1, \theta_2$  in a neighborhood of  $\theta_0$ ,

$$\|\psi_{\theta_1}(x) - \psi_{\theta_2}(x)\| \leq \dot{\psi}(x)\|\theta_1 - \theta_2\|,$$

for some measurable function  $\dot{\psi}$  with  $P\dot{\psi}^2 < \infty$ . Let the map  $\theta \mapsto P\psi_\theta$  be differentiable at its zero  $\theta_0$  with a nonsingular derivative. Then the (near) zeros  $\hat{\theta}_n$  of the map  $\theta \mapsto \mathbb{P}_n \psi_\theta$  satisfy (3.3.6), provided the sequence  $\hat{\theta}_n$  converges in outer probability to  $\theta_0$ .

This is a consequence of the preceding theorem. The class of functions  $\{\psi_\theta - \psi_{\theta_0}: \|\theta - \theta_0\| < \delta\}$  is Donsker for sufficiently small  $\delta > 0$ , because its bracketing integral is finite. Furthermore,  $P\|\psi_\theta - \psi_{\theta_0}\|^2 \leq P\dot{\psi}^2 \|\theta - \theta_0\|^2$  converges to zero as  $\theta \rightarrow \theta_0$ .

The pointwise local Lipschitz condition in the preceding example is satisfied if for each  $x$  the map  $\theta \mapsto \psi_\theta(x)$  is differentiable in a neighborhood of  $\theta_0$  with derivative matrices  $\dot{\psi}_\theta(x)$  such that, for some  $\delta > 0$ ,

$$P^* \sup_{\|\theta - \theta_0\| < \delta} \|\dot{\psi}_\theta\|^2 < \infty.$$

This is comparable to the classical condition of domination of derivatives<sup>†</sup>, but it concerns the second moment of the first derivative rather than the first moment of the second derivative. To obtain complete analogy with the classical result, it appears to be necessary to exploit the possibility of the extra term  $\sqrt{n}(\hat{\theta}_n - \theta_0)$  in the remainder of (3.3.2). See the following example.

**3.3.8 Example (Classical smoothness).** Let  $X_1, \dots, X_n$  be i.i.d. random variables with common law  $P$ , and let  $\psi_\theta$  be measurable vector-valued functions indexed by a parameter  $\theta$  ranging over an open subset of Euclidean space. Assume that  $P\psi_{\theta_0} = 0$  and that the map  $\theta \mapsto \psi_\theta(x)$  is twice continuously differentiable with respect to  $\theta$  in a neighborhood of  $\theta_0$ , for each  $x$ , with derivatives satisfying

$$P\|\dot{\psi}_{\theta_0}\|^2 < \infty; \quad P^* \sup_{\|\theta - \theta_0\| < \delta} \|\ddot{\psi}_\theta\| < \infty.$$

Then the map  $\theta \mapsto P\psi_\theta$  is differentiable at  $\theta_0$  with derivative  $P\dot{\psi}_{\theta_0}$  and the sequence  $\sqrt{n}(\hat{\theta}_n - \theta_0)$  is asymptotically normally distributed provided  $\hat{\theta}_n$  is consistent for  $\theta_0$ .

A Taylor expansion gives for each  $\theta$  that is sufficiently close to  $\theta_0$

$$\mathbb{G}_n(\psi_\theta - \psi_{\theta_0}) = \mathbb{G}_n \dot{\psi}_{\theta_0} (\theta - \theta_0) + \frac{1}{2}(\theta - \theta_0)' \mathbb{G}_n \ddot{\psi}_{\tilde{\theta}} (\theta - \theta_0),$$

for a point  $\tilde{\theta} = \tilde{\theta}(\theta, x)$  on the line connecting  $\theta$  and  $\theta_0$ . In the quadratic term, the empirical process is applied to the function  $x \mapsto \ddot{\psi}_{\tilde{\theta}(\theta, x)}(x)$ . Let the function  $M$  bound  $\|\ddot{\psi}_\theta\|$  for every  $\theta$  in a neighborhood of  $\theta_0$ . Then  $\|\mathbb{G}_n(\psi_{\hat{\theta}_n} - \psi_{\theta_0})\|$  is bounded by

$$\|\mathbb{G}_n \dot{\psi}_{\theta_0}\| \|\hat{\theta}_n - \theta_0\| + \frac{1}{2} \|\hat{\theta}_n - \theta_0\|^2 \sqrt{n} (\mathbb{P}_n + P) M = o_P(1) + o_P(1) \sqrt{n} \|\hat{\theta}_n - \theta_0\|,$$

since the sequence  $(\mathbb{P}_n + P)M$  is bounded in probability. Thus (3.3.4) is satisfied.

Our emphasis on Lipschitz conditions on the criterion functions in the preceding examples may give the (wrong) impression that smoothness is necessary for the estimators to be asymptotically normal or the theory of empirical processes to be of help. This is certainly a false impression, as

---

<sup>†</sup> Cf., for instance, Lehmann (1983).

should be clear from Lemmas 3.3.5 and 3.2.19. The point is that smoothness permits an easy general method to check the more abstract empirical process conditions, such as that the functions  $\psi_{\theta,h}$  are in a Donsker class. In particular examples we can apply all methods and results of Part 2 to verify these conditions; for instance, the stability properties of Chapter 2.10. We include one example to illustrate this.

**3.3.9 Example (Median and median deviation).** The median and median deviation are  $M$ -estimators for the location and scale parameter  $(\mu, \sigma)$  relative to the functions

$$\psi_{\mu,\sigma}(x) = \left( \text{sign}\left(\frac{x-\mu}{\sigma}\right), \text{sign}\left(\left|\frac{x-\mu}{\sigma}\right| - \beta\right) \right).$$

The constant  $\beta$  is usually chosen such that the median deviation converges to the standard deviation of some distribution of interest; for instance,  $\Phi^{-1}(3/4)$  for the normal distribution.

The dependence  $(\mu, \sigma) \mapsto \psi_{\mu,\sigma}(x)$  is not Lipschitz. However, the class of functions  $\{\psi_{\mu,\sigma,1}, \psi_{\mu,\sigma,2} : \mu \in \mathbb{R}, \sigma > 0\}$  can easily be shown to be a Donsker class. The class of functions  $x \mapsto (x - \mu)/\sigma$  belongs to a finite-dimensional vector space and hence is a VC-class. Next the stability properties of VC-classes show that the components of  $\psi_{\mu,\sigma}$  also run through VC-classes. Since they are uniformly bounded and pointwise separable, they are universally Donsker.

The preceding examples are concerned with finite-dimensional parameters or functionals. The theory of empirical processes is used to carry out the usual linearization argument under better conditions. An important advantage is that cases that require special and separate treatment in the classical approach can be dealt with in a unified manner by using the theory of empirical processes.

The next example shows that empirical processes can also be used successfully to derive properties of estimators in infinitely dimensional models. It is just one out of a number of examples that have been studied recently. These examples cannot be treated by classical methods, and the development of abstract empirical process theory seems to have occurred at the right time for the development of this part of semiparametric statistics.

**3.3.10 Example.** Suppose one observes a sample  $(X_1, Y_1), \dots, (X_n, Y_n)$  from a distribution given formally by the density

$$\int p_\theta(x|z) d\eta(z) d\eta(y).$$

Here the map  $x \rightarrow p_\theta(x|z)$  is for each  $z$  a probability density with respect to a measure  $\mu$  on some measurable space  $(\mathcal{X}, \mathcal{A})$ , which is known up to a parameter  $\theta$  ranging over an open subset  $\Theta$  of Euclidean space. The second

parameter  $\eta$  is a completely unknown probability distribution on a measurable space  $(\mathcal{Z}, \mathcal{C})$ . Write  $p_\theta(x|\eta)$  for the mixture density  $\int p_\theta(x|z) d\eta(z)$ .

Consider the asymptotic behavior of the maximum likelihood estimator  $(\hat{\theta}, \hat{\eta})$  of  $(\theta, \eta)$  defined as the maximizer of the “likelihood”

$$(\theta, \eta) \rightarrow \prod_{i=1}^n p_\theta(X_i|\eta) \prod_{i=1}^n \eta\{Y_i\}.$$

This combines the usual likelihood of the first sample, the joint density of the observations, with the “nonparametric” likelihood of the second sample, the product of the point masses  $\eta\{Y_i\}$  of the observations. In this semi-parametric set-up, we are mostly interested in estimating  $\theta$ .

A proof of the asymptotic normality of the sequence of maximum likelihood estimators may proceed by showing that a maximum likelihood estimator solves a collection of (likelihood) equations and next apply Theorem 3.3.1. This method applies to many different kernels. Consider in particular the case that  $p_\theta(x|z) = \phi((x-z)/\theta)/\theta$ . Then the observations are a sample  $X_1, \dots, X_n$  from the model  $X = Z + \theta e$  for an (unobserved) standard normal error variable  $e$  and an (unobserved) independent variable  $Z$  with distribution  $\eta$  and in addition a sample  $Y_1, \dots, Y_n$  without measurement error; that is with the same distribution as  $Z$ . We can think of this as observing one sample from the distribution of  $Z$  itself and an additional sample corrupted by normal measurement error. In this special example take the parameter set  $\Theta$  and  $\mathcal{Z}$  to be compact intervals in  $(0, \infty)$ .

Likelihood equations corresponding to  $\theta$  can be obtained in the usual manner by partial differentiation of the loglikelihood with respect to  $\theta$  at  $\hat{\theta}$ . For simplicity, assume that  $\theta$  is one-dimensional. Then we obtain

$$\mathbb{P}_n \dot{\ell}_{\hat{\theta}, \hat{\eta}} = 0,$$

where  $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{(X_i, Y_i)}$  is the empirical measure of the observations and  $\dot{\ell}_{\theta, \eta}$  is the score function for  $\theta$  of the mixture model. This can be expressed in the score functions  $\dot{\ell}_\theta(x|z) = \partial/\partial\theta \log p_\theta(x|z)$  of the conditional models by

$$\dot{\ell}_{\theta, \eta}(x, y) = \frac{\int \dot{\ell}_\theta(x|z) p_\theta(x|z) d\eta(z)}{p_\theta(x|\eta)}.$$

Likelihood equations corresponding to the infinite-dimensional parameter  $\eta$  can be obtained by inserting one-dimensional submodels  $t \rightarrow \hat{\eta}_t$  passing through  $\hat{\eta}$  in the loglikelihood and differentiating with respect to  $t$ . In particular, given a bounded, measurable function  $h$  and every sufficiently small number  $|t|$ , we can define a probability measure  $\hat{\eta}_t$  by

$$d\hat{\eta}_t = (1 + t(h - \int h d\hat{\eta})) d\hat{\eta}.$$

This leads to the likelihood equation

$$\mathbb{P}_n A_{\hat{\theta}, \hat{\eta}} h - P_{\hat{\theta}, \hat{\eta}} A_{\hat{\theta}, \hat{\eta}} h = 0,$$

where  $A_{\theta,\eta}$  are the “score operators” given by

$$A_{\theta,\eta}h(x,y) = B_{\theta,\eta}h(x) + h(y) = \frac{\int h(z)p_\theta(x|z)d\eta(z)}{p_\theta(x|\eta)} + h(y).$$

The operators  $B_{\theta,\eta}$  are the score operators for the mixture part of the model.

Let  $H$  be the set of all functions  $h: \mathcal{Z} \rightarrow [0, 1]$  with  $\|h\|_{lip} \leq 1$ , the positive part of the unit ball of the space  $C^1(\mathcal{Z})$  considered in Chapter 2.7. For each  $h \in H$  the maximum likelihood estimator satisfies the equation as derived previously. Let  $\Psi_n(\theta, \eta) = (\Psi_{n1}(\theta, \eta), \Psi_{n2}(\theta, \eta))$  be the element of  $\mathbb{R} \times \ell^\infty(H)$  given by

$$\Psi_{n1}(\theta, \eta) = \mathbb{P}_n \dot{\ell}_{\theta,\eta}; \quad \Psi_{n2}(\theta, \eta)h = \mathbb{P}_n A_{\theta,\eta}h - P_{\theta,\eta} A_{\theta,\eta}h.$$

The map  $h \rightarrow \Psi_{n2}(\theta, \eta)h$  is indeed uniformly bounded on  $H$ , because the conditional expectation operator  $B_{\theta,\eta}$  retains boundedness:  $0 \leq B_{\theta,\eta}h \leq 1$  for every  $h \in H$ . The expectation of  $\Psi_n(\theta, \eta)$  under the true distribution  $P_0 = P_{\theta_0, \eta_0}$  is the element  $\Psi(\theta, \eta) = (\Psi_1(\theta, \eta), \Psi_2(\theta, \eta))$  of  $\mathbb{R} \times \ell^\infty(H)$  given by

$$\Psi_1(\theta, \eta) = P_0 \dot{\ell}_{\theta,\eta}; \quad \Psi_2(\theta, \eta)h = P_0 A_{\theta,\eta}h - P_{\theta,\eta} A_{\theta,\eta}h.$$

The maximum likelihood estimator  $(\hat{\theta}_n, \hat{\eta}_n)$  and the true value  $(\theta_0, \eta_0)$  are zeros of these maps, respectively:

$$\Psi_n(\hat{\theta}_n, \hat{\eta}_n) \equiv 0; \quad \Psi(\theta_0, \eta_0) \equiv 0.$$

We identify each probability measure  $\eta$  with an element of  $\ell^\infty(H)$  through  $\eta h = \int h d\eta$ . Then both  $\Psi_n$  and  $\Psi$  can be viewed as maps from the space  $\mathbb{R} \times \ell^\infty(H)$  into itself whose domain is the product of  $\Theta$  and the set of probability measures in  $\ell^\infty(H)$  under the given identification. The present choice of  $H$  is appropriate for our special example (and many other examples) but should be replaced by a different choice whenever convenient.

The remainder of this example is concerned with checking the conditions of Theorem 3.3.1, particularly for our special example involving a normal measurement error model.

The consistency of the sequence  $(\hat{\theta}_n, \hat{\eta}_n)$  may be proved by the method of Wald or by a method using empirical processes. This requires some regularity conditions on the kernel  $z \mapsto p_\theta(x|z)$ . We omit a discussion.

In the present case the process  $\sqrt{n}(\Psi_n - \Psi)$  takes the form

$$\sqrt{n}(\Psi_n - \Psi)(\theta, \eta) = \left( \sqrt{n}(\mathbb{P}_n - P_0)\dot{\ell}_{\theta,\eta}, \sqrt{n}(\mathbb{P}_n - P_0)A_{\theta,\eta} \cdot \right).$$

The right side is the empirical process indexed by the class of functions  $\{A_{\theta,\eta}h: h \in H\} \cup \{\dot{\ell}_{\theta,\eta}\}$ . In view of the discussion following the proof of

Theorem 3.3.1, condition (3.3.2) of this theorem is certainly satisfied if, for some  $\delta > 0$ ,

$$(3.3.11) \quad \{A_{\theta,\eta}h : h \in H, \|\theta - \theta_0\| + \|\eta - \eta_0\| < \delta\} \quad \text{is } P_0\text{-Donsker},$$

$$\sup_{h \in H} P_0(A_{\theta,\eta}h - A_0h)^2 + P_0(\ell_{\theta,\eta} - \ell_0)^2 \rightarrow 0, \quad \theta \rightarrow \theta_0, \eta \rightarrow \eta_0.$$

In view of the dominated convergence theorem, the last condition is valid if  $A_{\theta,\eta}h \rightarrow A_0h$  and  $\ell_{\theta,\eta} \rightarrow \ell_0$  pointwise, uniformly in  $h$ .

It is hard to characterize condition (3.3.11) directly in terms of the kernel  $p_\theta(x|z)$ . In fact, different kernels may require different approaches. Consider the case that the kernel is a smooth function on Euclidean space. Then the functions  $x \mapsto B_{\theta,\eta}h(x)$  will be smooth functions as well and Theorem 2.10.24 may be appropriate. If the tail of  $P_0$  is not too large and  $\mathcal{X}$  is a subset of  $\mathbb{R}^d$ , then a uniform  $\alpha$ -smoothness condition on the functions  $x \mapsto B_{\theta,\eta}h(x)$  for some  $\alpha > d/2$  is sufficient. Under appropriate conditions on the map  $x \rightarrow p_\theta(x|z)$ , straightforward differentiation yields

$$\frac{\partial}{\partial x_i} B_{\theta,\eta}h(x) = \text{cov}_x \left( h(Z), \frac{\partial}{\partial x_i} \log p_\theta(x|Z) \right),$$

where for each  $x$  the covariance is computed for the random variable  $Z$  having the (conditional) density  $z \rightarrow p_\theta(x|z) d\eta(z)/p_\theta(x|\eta)$ . Thus, for a given bounded function  $h$ ,

$$\left| \frac{\partial}{\partial x_i} B_{\theta,\eta}h(x) \right| \leq \|h\|_\infty \frac{\int |\partial/\partial x_i \log p_\theta(x|z)| p_\theta(x|z) d\eta(z)}{\int p_\theta(x|z) d\eta(z)}.$$

Depending on the function  $\partial/\partial x_i \log p_\theta(x|z)$ , this leads to a bound on the first derivative and hence on the Lipschitz norm of order 1 of the function  $x \rightarrow B_{\theta,\eta}h(x)$ . If  $\mathcal{X}$  is equal to  $\mathbb{R}$ , then this is sufficient for the applicability of Theorem 2.10.24. In our example we have  $|\partial/\partial x \log p_\theta(x|z)| = |x - z|/\theta^2$ , which is bounded by a constant times  $|x|$  for  $\theta$  in a neighborhood of  $\theta_0$  and  $\eta$  ranging over probability measures on a fixed interval. Since the tails of  $P_0$  are Gaussian, the function  $|x|$  has the required integrability properties and Theorem 2.10.24 implies that (3.3.11) is satisfied.

If  $\mathcal{X}$  is a subset of a higher-dimensional Euclidean space, then the same conclusion may be reached by consideration of higher-order derivatives.

The remaining conditions of Theorem 3.3.1 are the differentiability of the map  $\Psi$  and the continuity of the inverse of the derivative. For convenience of notation, we introduce the Hilbert-space adjoint  $B_{\theta,\eta}^*$  of the operator  $B_{\theta,\eta} : L_2(\eta) \rightarrow L_2(p_\theta(\cdot|\eta))$  given by

$$B_{\theta,\eta}^* g(z) = \int g(x) p_\theta(x|z) d\mu(x).$$

The range of the operator  $A_{\theta,\eta}$  is contained in the subset  $G$  of  $L_2(p_\theta(\cdot|\eta) \times \eta)$  consisting of functions of the form  $(x,y) \rightarrow g_1(x) + g_2(y) + c$ . The

representation of a function of this type is unique if both  $g_1$  and  $g_2$  are taken to be zero-mean functions. The adjoint of the operator  $A_{\theta,\eta}: L_2(\eta) \rightarrow G$  is given by  $A_{\theta,\eta}^*(g_1 + g_2 + c) = B_{\theta,\eta}^*g_1 + g_2 + 2c$ . On the set of zero-mean functions in  $L_2(\eta)$ , we have the identity  $A_{\theta,\eta}^*A_{\theta,\eta} = I + B_{\theta,\eta}^*B_{\theta,\eta}$ .

Informally, the derivative  $\dot{\Psi} = (\dot{\Psi}_1, \dot{\Psi}_2)$  of the map  $\Psi$  at  $(\theta_0, \eta_0)$  can be derived as follows. First,

$$\begin{aligned}\Psi_1(\theta, \eta) - \Psi_1(\theta_0, \eta_0) &= P_0(\dot{\ell}_{\theta,\eta} - \dot{\ell}_0) \\ &\approx (\theta - \theta_0)P_0\ddot{\ell}_0 + \iint (\dot{\ell}_0(x|z) - \dot{\ell}_0(x)) p_0(x|z) d\mu(x) d(\eta - \eta_0)(z).\end{aligned}$$

Under regularity conditions,  $P_0\ddot{\ell}_0 = -I_0$  is minus the Fisher information for  $\theta$  in the situation when  $\eta = \eta_0$  is known and the expectation  $\int \dot{\ell}_0(x|z)p_0(x|z) d\mu(x) = 0$  for every  $z$ . Then the last line can be rewritten as

$$-I_0(\theta - \theta_0) - \int B_0^*\dot{\ell}_0 d(\eta - \eta_0).$$

The derivative of the second component of  $\Psi$  can be obtained in a similar way. Uniformly in  $h$ ,

$$\begin{aligned}\Psi_2(\theta, h)h - \Psi_2(\theta_0, \eta_0)h &= - \int A_{\theta,\eta}h d(P_{\theta,\eta} - P_0) \\ &\approx - \int A_0h d(P_{\theta,\eta} - P_0) \\ &\approx -(\theta - \theta_0) \int A_0h \dot{\ell}_0 dP_0 - \int (I + B_0^*B_0)h d(\eta - \eta_0).\end{aligned}$$

The combination of the preceding displays suggests that the derivative of  $\Psi$  at  $(\theta_0, \eta_0)$  is given by the map

$$(\theta - \theta_0, \eta - \eta_0) \rightarrow \begin{pmatrix} \dot{\Psi}_{11} & \dot{\Psi}_{12} \\ \dot{\Psi}_{21} & \dot{\Psi}_{22} \end{pmatrix} \begin{pmatrix} \theta - \theta_0 \\ \eta - \eta_0 \end{pmatrix},$$

where

$$\begin{aligned}\dot{\Psi}_{11}(\theta - \theta_0) &= -I_0(\theta - \theta_0), \\ \dot{\Psi}_{12}(\eta - \eta_0) &= - \int B_0^*\dot{\ell}_0 d(\eta - \eta_0), \\ \dot{\Psi}_{21}(\theta - \theta_0)h &= -P_0A_0h\dot{\ell}_0(\theta - \theta_0), \\ \dot{\Psi}_{22}(\eta - \eta_0)h &= - \int (I + B_0^*B_0)h d(\eta - \eta_0)(z).\end{aligned}$$

The derivation can be made rigorous under regularity conditions. We omit the somewhat tedious verification in our concrete example.

Theorem 3.3.1 requires that the derivative operator be continuously invertible on the linear span of the domain of  $\Psi$ . This can be verified by

ascertaining the continuous invertibility of the two operators  $\dot{\Psi}_{11}$  and  $\dot{V} = \dot{\Psi}_{22} - \dot{\Psi}_{21}\dot{\Psi}_{11}^{-1}\dot{\Psi}_{12}$ . In that case we have

$$\dot{\Psi}^{-1} = \begin{pmatrix} \dot{\Psi}_{11}^{-1}(\dot{\Psi}_{11} + \dot{\Psi}_{12}\dot{V}^{-1}\dot{\Psi}_{21})\dot{\Psi}_{11}^{-1} & -\dot{\Psi}_{11}^{-1}\dot{\Psi}_{12}\dot{V}^{-1} \\ -\dot{V}^{-1}\dot{\Psi}_{21}\dot{\Psi}_{11}^{-1} & \dot{V}^{-1} \end{pmatrix}.$$

The operator  $\dot{\Psi}_{11}$  is continuously invertible provided the Fisher information for  $\theta$  is positive. The second operator has the form

$$\dot{V}(\eta - \eta_0)h = - \int \left[ (I + B_0^*B_0)h - \frac{P_0A_0h\dot{\ell}_0}{I_0}B_0^*\dot{\ell}_0 \right] d(\eta - \eta_0).$$

This operator is certainly continuously invertible if there exists a positive number  $\epsilon$  such that

$$\left\{ (I + B_0^*B_0)h - \frac{P_0A_0h\dot{\ell}_0}{I_0}B_0^*\dot{\ell}_0 : h \in H \right\} \supset \epsilon H.$$

In many examples, including our concrete example, this can be ascertained by using the theory of Fredholm operators. An operator of the form  $I + K$  from a Banach space in itself, where  $K$  is compact, is continuously invertible if and only if it is one-to-one.<sup>†</sup> In the present situation, consider the operator  $K$  given by

$$Kh = B_0^*B_0h - \frac{P_0A_0h\dot{\ell}_0}{I_0}B_0^*\dot{\ell}_0.$$

If the maps  $z \rightarrow p_{\theta_0}(x|z)$  are sufficiently smooth, as is the case in our example, then the operator  $B_0^*: C^1(\mathcal{Z}) \rightarrow C^1(\mathcal{Z})$  can be seen to be compact by the Arzelà-Ascoli theorem. Consequently,  $B_0^*B_0$  is a compact operator from  $C^1(\mathcal{Z})$  into itself. The second part of  $K$  is compact, because it has a one-dimensional range. Finally, it suffices to show that  $I + K$  is one-to-one. The spectrum of the self-adjoint operator  $I + B_0^*B_0: L_2(\eta_0) \rightarrow L_2(\eta_0)$  is contained in  $[1, \infty)$ . Hence this operator is continuously invertible in the Hilbert-space sense. The equation

$$(3.3.12) \quad h + B_0^*B_0h - \frac{P_0A_0h\dot{\ell}_0}{I_0}B_0^*\dot{\ell}_0 = 0$$

can be solved to give either

$$P_0A_0h\dot{\ell}_0 = 0 \quad \text{or} \quad P_0(A_0(I + B_0^*B_0)^{-1}B_0^*\dot{\ell}_0)\dot{\ell}_0 = I_0.$$

The efficient Fisher information for  $\theta$  is by definition the square length of the projection of  $\dot{\ell}_0$  on the orthocomplement of the range of  $A_0$ . This is positive in most examples, including the present one. The function  $A_0(I + B_0^*B_0)^{-1}B_0^*\dot{\ell}_0 = A_0(A_0^*A_0)^{-1}A_0^*\dot{\ell}_0$  is the projection of  $\dot{\ell}_0$  onto the range of  $A_0$ . Hence the difference of the right and the left sides of the second equation in the preceding display is the efficient Fisher information, which

---

<sup>†</sup> See, for instance, Rudin (1973), pages 99–103.

is nonzero. If the first expression in the display is zero, then we find that (3.3.12) reduces to  $h + B_0^* B_0 h = 0$ , whence  $h = 0$  almost surely under  $\eta_0$ . Reinsert this into equation (3.3.12) to find that  $h$  is zero pointwise. This concludes the proof of the continuous invertibility of the operator  $I + K$  and hence of the continuous invertibility of  $\hat{\Psi}_0$ .

Thus, all conditions of Theorem 3.3.1 have been verified. We conclude that the sequence  $\sqrt{n}(\hat{\theta}_n - \theta_0, \hat{\eta}_n - \eta_0)$  is asymptotically normal in the space  $\mathbb{R} \times \ell^\infty(H)$ . In particular, the sequence of prime interest,  $\sqrt{n}(\hat{\theta}_n - \theta_0)$ , is asymptotically normal with mean zero and variance  $P_0(\dot{\ell}_0 - A_0(A_0^* A_0)^{-1} \dot{\ell}_0)^2$ .

## 3.4

# Rates of Convergence

This chapter gives some results on rates of convergence of  $M$ -estimators, including maximum likelihood estimators and least-squares estimators. We first state an abstract result, which is a generalization of the theorem on rates of convergence in Chapter 3.2, and next discuss some methods to establish the maximal inequalities needed for the application of this result. Our main interest is in  $M$ -estimators of infinite-dimensional parameters.

In the introduction the parameter set is denoted  $\Theta$  and may be thought of as a semimetric space  $\Theta$  with semimetric  $d$ . We consider *sieved  $M$ -estimators*, sieves  $\Theta_n$  being a sequence of subsets of the parameter space. Let  $\mathbb{P}_n$  denote the empirical measure of observations  $X_1, X_2, \dots, X_n$  taking their values in a measurable space  $(\mathcal{X}, \mathcal{A})$ . Given measurable maps  $m_{n,\theta}: \mathcal{X} \mapsto \mathbb{R}$ , “estimators”  $\hat{\theta}_n$  are maps defined on  $\mathcal{X}^n$  with values in the sieve  $\Theta_n$ , such that, for certain parameters  $\theta_n$ ,

$$\mathbb{P}_n m_{n,\hat{\theta}_n} \geq \mathbb{P}_n m_{n,\theta_n}.$$

The estimators may be thought of as maximizing the random criterion function  $\theta \mapsto \mathbb{P}_n m_{n,\theta}$  over  $\Theta_n$ . Then the preceding display is valid for every  $\theta_n \in \Theta_n$ . However, in the following it is not assumed that  $\hat{\theta}_n$  is defined as a maximizer of  $\theta \mapsto \mathbb{P}_n m_{n,\theta}$ . In fact, in the examples, the map  $\theta \mapsto \mathbb{P}_n m_{n,\theta}$  is often not observable. In these examples,  $\hat{\theta}_n$  satisfies the property as in the preceding display in virtue of the fact that it maximizes another criterion function, which is observable.

Generally, for the estimators  $\hat{\theta}_n$  to be consistent, the sieves  $\Theta_n$  must be constructed to grow dense in  $\Theta$  as  $n \rightarrow \infty$ . The simplest sequence of

sieves with this property is the whole parameter set  $\Theta_n = \Theta$ . In this case, a natural choice for  $\theta_n$  is the “true” parameter  $\theta_0$ . In any case the sieves must be large enough, because the rate of convergence of  $d(\hat{\theta}_n, \theta_0)$  to zero guaranteed by Theorem 3.4.1 ahead is never faster than  $d(\theta_n, \theta_0)$ . The latter quantity may be thought of as the distance of  $\theta_0$  to the sieve  $\Theta_n$ .

Maximizing a given criterion function over the whole parameter space may or may not be a good idea. On the one hand, the full maximum likelihood estimator is often a good (in particular, asymptotically efficient) estimator, even in the case of infinite-dimensional parameter spaces. On the other hand,  $\sup_\theta \mathbb{P}_n m_{n,\theta}$  may be infinite and no  $M$ -estimator is defined; or the criterion functions may have a unique point of maximum for every  $n$ , but the sequence of estimators may fail to converge to the true value, or do so at a suboptimal rate. While examples provide some insight as to when these possibilities occur, a general theory as to when a sequence of sieves is necessary or how to construct one is unavailable.

The observations  $X_1, \dots, X_n$  need not be i.i.d. For a flexible statement, we consider estimators  $\hat{\theta}_n$  such that

$$\mathbb{M}_n(\hat{\theta}_n) \geq \mathbb{M}_n(\theta_n),$$

for arbitrary stochastic processes  $\mathbb{M}_n$ . Then it is understood that each  $\hat{\theta}_n$  is an arbitrary map defined on the same probability space as  $\mathbb{M}_n$ . Corresponding to the criterion functions are centering functions  $\theta \mapsto M_n(\theta)$ . Typically these are the mean functions of the criterion functions, but the following theorem allows them to be arbitrary, even random. The value  $\theta_n$  may be thought of as maximizing the centering function  $M_n$  (but this is not an assumption), paralleling the maximization of  $\mathbb{M}_n$  by  $\hat{\theta}_n$ . Thus it is reasonable to expect that

$$M_n(\theta) - M_n(\theta_n) \leq -d_n^2(\theta, \theta_n).$$

This explains the first displayed condition of the following theorem. Here  $\theta \mapsto d_n(\theta, \theta_n)$  may be an arbitrary, nonnegative map on  $\Theta_n$ , though it is written in a form that suggests that it derives from a distance. In our discussions we shall think of it in this manner for a fixed distance  $d(\theta, \theta_n) = d_n(\theta, \theta_n)$ . Actually, the following theorem allows  $d_n$  as well as the centering functions  $M_n$  and the value  $\theta_n$  to be random (a map defined on the product of  $\Theta_n$  and the underlying probability space). In particular  $d_n^2(\theta, \theta_n)$  can be chosen equal to  $M_n(\theta_n) - M_n(\theta)$ , in which case the preceding display ceases to be a condition.

**3.4.1 Theorem (Rate of convergence).** *For each  $n$ , let  $\mathbb{M}_n$  and  $M_n$  be stochastic processes indexed by a set  $\Theta$ . Let  $\theta_n \in \Theta$  (possibly random) and  $0 \leq \delta_n < \eta$  be arbitrary,<sup>b</sup> and let  $\theta \mapsto d_n(\theta, \theta_n)$  be an arbitrary*

---

<sup>b</sup> In applications  $\delta_n$  is typically a multiple of  $d(\theta_n, \theta_0)$  and  $\eta = \infty$ .

map (possibly random) from  $\Theta$  to  $[0, \infty)$ . Suppose that, for every  $n$  and  $\delta_n < \delta \leq \eta$

$$\sup_{\delta/2 < d_n(\theta, \theta_n) \leq \delta, \theta \in \Theta_n} M_n(\theta) - M_n(\theta_n) \leq -\delta^2,$$

$$\mathbf{E}^* \sup_{\delta/2 < d_n(\theta, \theta_n) \leq \delta, \theta \in \Theta_n} \sqrt{n} [(\mathbb{M}_n - M_n)(\theta) - (\mathbb{M}_n - M_n)(\theta_n)]^+ \lesssim \phi_n(\delta),$$

for functions  $\phi_n$  such that  $\delta \mapsto \phi_n(\delta)/\delta^\alpha$  is decreasing on  $(\delta_n, \eta)$ , for some  $\alpha < 2$ . Let  $r_n \lesssim \delta_n^{-1}$  satisfy

$$r_n^2 \phi_n \left( \frac{1}{r_n} \right) \leq \sqrt{n}, \quad \text{for every } n.$$

If the sequence  $\hat{\theta}_n$  takes its values in  $\Theta_n$  and satisfies  $\mathbb{M}_n(\hat{\theta}_n) \geq \mathbb{M}_n(\theta_n) - O_P(r_n^{-2})$  and  $d_n(\hat{\theta}_n, \theta_n)$  converges to zero in outer probability, then  $r_n d_n(\hat{\theta}_n, \theta_n) = O_P^*(1)$ . If the displayed conditions are valid for  $\eta = \infty$ , then the condition that  $\hat{\theta}_n$  is consistent is unnecessary.

**Proof.** The proof is basically the same as that for Theorem 3.2.5, where  $\theta_n$  must be substituted for  $\theta_0$  throughout and  $\Theta_n$  for  $\Theta$ . ■

Under the conditions of the theorem, the distance of  $\hat{\theta}_n$  to the true value  $\theta_0$  satisfies

$$d(\hat{\theta}_n, \theta_0) = O_P^*(r_n^{-1}) + d(\theta_n, \theta_0).$$

The rate  $r_n$  is determined by the “modulus of continuity”  $\phi_n(\delta)$  of the centered processes  $\sqrt{n}(\mathbb{M}_n - M_n)$  over  $\Theta_n$ . Typically, small sieves  $\Theta_n$  lead to a small modulus, hence fast rates  $r_n$ . On the other hand, the distance of  $\theta_0$  to a small sieve will be large. Thus, the two terms on the right in the preceding display may be loosely understood as a “variance” (or rather, standard deviation) and “bias” term, which must be balanced to obtain a good rate of convergence. We note that in many problems an unsieved  $M$ -estimator actually performs very well, so the trade-off should not be understood too literally: it may work well to reduce the “bias” to zero.

The main challenge for the application of the preceding theorem is to derive the maximal inequalities for the modulus of the centered processes  $\sqrt{n}(\mathbb{M}_n - M_n)$ . For the case that  $\mathbb{M}_n(\theta) = \mathbb{P}_n m_{n,\theta}$  considered in the introduction with centering  $M_n(\theta) = Pm_{n,\theta}$ , this involves the empirical process indexed by the classes of functions

$$\mathcal{M}_{n,\delta} = \left\{ m_{n,\theta} - m_{n,\theta_n} : \theta \in \Theta_n, \frac{\delta}{2} < d_n(\theta, \theta_n) \leq \delta \right\}.$$

Write  $M_{n,\delta}$  for the envelope function of these classes. In several examples in Chapter 3.2 it works well to use an inequality of the type

$$\mathbf{E}_P^* \|\mathbb{G}_n\|_{\mathcal{M}_{n,\delta}} \lesssim J(1) (P^* M_{n,\delta}^2)^{1/2},$$

with  $J(1)$  a uniform bound on the uniform entropy or bracketing entropy integral of the classes  $\mathcal{M}_{n,\delta}$ . The rate of convergence is then driven by the sizes of the envelope functions as  $\delta \downarrow 0$ , and the size of the classes is important only to guarantee a finite entropy integral. In genuinely infinite-dimensional situations, this approach appears to be less useful. It is intuitively clear that the precise entropy must make a difference for the rate of convergence in general.

An alternative is to use the refined bracketing inequality given by Lemma 2.14.3, which bounds the modulus by a bracketing integral plus two remainder terms, which concern finite suprema. If a good (exponential) bound on all of the variables  $\mathbb{G}_n f_i$  and  $\mathbb{G}_n \sup_{f \in \mathcal{F}_i} |f - f_i|$  is available, then the remainder terms can be further bounded; for instance, by means of Lemma 2.2.2 or 2.2.10. We consider two examples of this approach. For a given norm  $\|\cdot\|$ , let

$$\tilde{J}_{[]}(\delta, \mathcal{F}, \|\cdot\|) = \int_0^\delta \sqrt{1 + \log N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)} d\varepsilon$$

be the bracketing integral of a class of functions  $\mathcal{F}$ .<sup>#</sup>

**3.4.2 Lemma.** *Let  $\mathcal{F}$  be class of measurable functions such that  $Pf^2 < \delta^2$  and  $\|f\|_\infty \leq M$  for every  $f$  in  $\mathcal{F}$ . Then*

$$E_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim \tilde{J}_{[]}(\delta, \mathcal{F}, L_2(P)) \left( 1 + \frac{\tilde{J}_{[]}(\delta, \mathcal{F}, L_2(P))}{\delta^2 \sqrt{n}} M \right).$$

The assumption that  $\mathcal{F}$  is uniformly bounded is strong, but it can be dropped if the  $L_2(P)$ -norm is replaced by the “norm”

$$\|f\|_{P,B} = \left( 2P(e^{|f|} - 1 - |f|) \right)^{1/2}.$$

This “Bernstein norm” neither is homogeneous nor satisfies the triangle inequality, but we shall nevertheless use it to measure the “size” of functions. The Bernstein “norm” dominates the  $L_2(P)$ -norm and allows the use of the refined form of Bernstein’s inequality given by Lemma 2.2.11, in a chaining argument, for the proof of Lemma 2.14.3 actually only requires the property that  $|f| \leq |g|$  implies  $\|f\| \leq \|g\|$ .

**3.4.3 Lemma.** *Let  $\mathcal{F}$  be a class of measurable functions such that  $\|f\|_{P,B} \leq \delta$  for every  $f \in \mathcal{F}$ . Then*

$$E_P^* \|\mathbb{G}_n\|_{\mathcal{F}} \lesssim \tilde{J}_{[]}(\delta, \mathcal{F}, \|\cdot\|_{P,B}) \left( 1 + \frac{\tilde{J}_{[]}(\delta, \mathcal{F}, \|\cdot\|_{P,B})}{\delta^2 \sqrt{n}} \right).$$

---

<sup>#</sup> In this chapter the notation  $\tilde{J}$  is used for entropy integrals that are not standardized relative to an envelope function. Thus  $\tilde{J}$  differs from  $J$  in Chapter 2.2.

**Proofs.** Apply Lemma 2.14.3 with  $\gamma = 0$  and  $p = 1$ . It suffices to bound the second and third terms on the right appropriately.

For the first lemma use Bernstein's inequality given by Lemma 2.2.9 and Lemma 2.2.10 (see (2.5.5)) to bound both the second and first terms by a multiple of

$$\left(1 + \log N_{[]}(\delta, \mathcal{F}, L_2(P))\right) \frac{M}{\sqrt{n}} + \sqrt{1 + \log N_{[]}(\delta, \mathcal{F}, L_2(P))} \delta.$$

Since the entropy is decreasing in  $\delta$ , the second term is smaller than the entropy integral. By the same reasoning the part in brackets of the first term is bounded by the square of the entropy integral divided by  $\delta^2$ .

The second lemma follows similarly, this time using the refined form of Bernstein's inequality, Lemma 2.2.11, in combination with the “norm”  $\|\cdot\|_{P,B}$ . ■

Since the Bernstein “norm” is not homogeneous, it may actually be beneficial to apply the second lemma to a multiple  $\sigma\mathcal{F}$  of the class  $\mathcal{F}$ , noting that the left side equals  $\sigma^{-1}\mathbb{E}_P^*\|\mathbb{G}_n\|_{\sigma\mathcal{F}}$  for every  $\sigma > 0$ . Even taking  $\sigma = 2$  may be helpful!

Another possibility is to use symmetrization by Rademacher variables followed by application of the sub-Gaussian maximal inequality given by Corollary 2.2.8, conditionally on the original observations. By Lemma 2.3.1 followed by Corollary 2.2.8,

$$\begin{aligned} \mathbb{E}^*\|\mathbb{G}_n\|_{\mathcal{M}_{n,\delta}} &\leq 2\mathbb{E}^*\left\|\frac{1}{\sqrt{n}}\sum_{i=1}^n \varepsilon_i m(X_i)\right\|_{\mathcal{M}_{n,\delta}} \\ &\lesssim \mathbb{E}^* \int_0^{\theta_n(\delta)} \sqrt{\log N(\varepsilon, \mathcal{M}_{n,\delta}, L_2(\mathbb{P}_n))} d\varepsilon. \end{aligned}$$

Here  $\theta_n(\delta)$  is the  $L_2(\mathbb{P}_n)$ -diameter of the set  $\mathcal{M}_{n,\delta}$ . This expression may be handled directly, but in most cases it allows further estimates only if the uniform entropy of the classes  $\mathcal{M}_{n,\delta}$  behaves well. Then the integral can be bounded by

$$\mathbb{E}^* \int_0^{\theta_n(\delta)/\|M_{n,\delta}\|_{\mathbb{P}_n,2}} \sup_Q \sqrt{\log N(\varepsilon \|M_{n,\delta}\|_{Q,2}, \mathcal{M}_{n,\delta}, L_2(Q))} d\varepsilon \sqrt{\mathbb{P}_n M_{n,\delta}^2}.$$

This is identical to the bound given by Theorem 2.14.1. For instance, if the uniform entropy is bounded by  $(1/\varepsilon)^V$  (uniformly in  $\delta$ ), we obtain the bound

$$\mathbb{E}^*\|\mathbb{G}_n\|_{\mathcal{M}_{n,\delta}} \lesssim \mathbb{E}^*\|\mathbb{P}_n\|_{\mathcal{M}_{n,\delta}^2}^{1-V/2} (\mathbb{P}_n M_{n,\delta}^2)^{V/4}.$$

A disadvantage of this approach is that further handling of the right-hand side still involves a maximal inequality, but this may be more straightforward than the original problem.

In certain applications the empirical process indexed by the class  $\mathcal{M}_{n,\delta}$  may also be sub-Gaussian or satisfy other tail bounds. Then the maximal inequalities of Chapter 2.2 may be applied directly, without the additional arguments used for the empirical process in Chapter 2.14.

In some examples the bracketing entropy integral diverges at zero. Then the preceding lemmas are useless, but the more general Lemma 2.14.3 can be used with a positive value of  $\gamma$ . The choice  $\gamma = c\delta^2 \wedge \delta/3$  for a small positive constant  $c$  is appropriate, because in Theorem 3.4.1 the random process  $\sqrt{n}(\mathbb{M}_n - M_n)$  could be allowed to have a quadratic drift as long as this is strictly smaller than the quadratic drift of the centering function. It suffices that the supremum of the empirical process indexed by  $\mathcal{M}_{n,\delta}$  can be bounded by  $Z_{n,\delta} + \frac{1}{2}\delta^2\sqrt{n}$  for a random variable  $Z_{n,\delta}$  such that

$$\mathbb{E}Z_{n,\delta}^+ \lesssim \phi_n(\delta),$$

(see Problem 3.4.2). Lemma 2.14.3 yields such a bound on the empirical processes shifted downward.

This argument shows that in the following the entropy integrals can be replaced by the integrals

$$\int_{c\delta^2 \wedge \delta/3}^{\delta} \sqrt{1 + \log N_{[]}(\varepsilon, \mathcal{F}, \|\cdot\|)} d\varepsilon.$$

Since this does not make a difference in most applications, this fact will not be stressed.

### 3.4.1 Maximum Likelihood

Let  $X_1, \dots, X_n$  be an i.i.d. sample from a density  $p$  that belongs to a set  $\mathcal{P}$  of densities with respect to a measure  $\mu$  on some measurable space.

The sieved maximum likelihood estimator  $\hat{p}_n$  based on  $X_1, \dots, X_n$  maximizes the loglikelihood  $p \mapsto \mathbb{P}_n \log p$ . In this section the parameter is the density  $p$  itself and is denoted by  $p$  rather than  $\theta$ . If  $\hat{p}_n$  maximizes the loglikelihood over a sieve  $\mathcal{P}_n$  and  $p_n \in \mathcal{P}_n$ , then by concavity of the logarithm,

$$\mathbb{P}_n \log \frac{\hat{p}_n + p_n}{2p_n} \geq 0 = \mathbb{P}_n \log \frac{p_n + p_n}{2p_n}.$$

Thus the set-up of the introduction applies with the criterion function

$$m_{n,p} = \log \frac{p + p_n}{2p_n}.$$

(Note that it is not claimed that the maximum likelihood estimator maximizes  $p \mapsto \mathbb{P}_n m_{n,p}$  over  $\mathcal{P}_n$ .) This choice is technically more convenient than the naive choice  $\log p$ , as it combines smoothly with the Hellinger distance

$$h(p, q) = \left( \int (p^{1/2} - q^{1/2})^2 d\mu \right)^{1/2}.$$

The key is the following pair of inequalities, which relate the Bernstein “norm” of the criterion functions  $m_{n,p}$  to the Hellinger distance of the densities  $p$ . For densities (nonnegative functions)  $p, q, p_n$ , and  $p_0$  such that  $p_0/p_n \leq M$  and  $p \leq q$ , we have

$$\begin{aligned}\|m_{n,p} - m_{n,p_n}\|_{P_0, B} &\lesssim h(p, p_n), \\ \|m_{n,p} - m_{n,q}\|_{P_0, B} &\lesssim h(p + p_n, q + p_n) \leq h(p, q),\end{aligned}$$

where the constant in  $\lesssim$  depends on the upper bound  $M$  only. These inequalities are proved ahead. Since the map  $p \mapsto m_{n,p}$  is monotone, the second inequality shows that a bracketing partition of a class of densities  $p$  for the Hellinger distance induces a bracketing partition of the class of criterion functions  $m_{n,p}$  for the Bernstein “norm” of essentially the same size. Thus, Lemma 3.4.3 becomes available for the classes of functions  $\mathcal{M}_{n,\delta}$  with the entropy bounded by the Hellinger entropy of the class of densities.

**3.4.4 Theorem.** *Let  $h$  denote the Hellinger distance on a class of densities  $\mathcal{P}$  and set  $m_{n,p} = \log \frac{1}{2}(p + p_n)/p_n$ . If  $p_n$  and  $p_0$  are probability densities with  $p_0/p_n \leq M$  pointwise, then*

$$P_0(m_{n,p} - m_{n,p_n}) \lesssim -h^2(p, p_n),$$

for every probability density  $p$  such that  $h(p, p_n) \geq 32M h(p_n, p_0)$ . Furthermore, for the class of functions  $\mathcal{M}_{n,\delta} = \{m_{n,p} - m_{n,p_n} : p \in \mathcal{P}_n, h(p, p_n) < \delta\}$ ,

$$E_{P_0}^* \|\mathbb{G}_n\|_{\mathcal{M}_{n,\delta}} \lesssim \tilde{J}_{[]}(\delta, \mathcal{P}_n, h) \left( 1 + \frac{\tilde{J}_{[]}(\delta, \mathcal{P}_n, h)}{\delta^2 \sqrt{n}} \right).$$

In both assertions the constants in  $\lesssim$  depend on  $M$  only. In  $\tilde{J}_{[]}(\delta, \mathcal{P}_n, h)$  the set  $\mathcal{P}_n$  may be replaced by  $\mathcal{P}_{n,\delta} = \{p \in \mathcal{P}_n : h(p, p_n) < \delta\}$  and the Hellinger distance may be replaced by the distance  $h_n(p, q) = h(p + p_n, q + p_n)$ .

**Proof.** Since  $\log x \leq 2(\sqrt{x} - 1)$  for every positive  $x$ ,

$$\begin{aligned}P_0 \log \frac{q}{p_n} &\leq 2P_0 \left( \frac{q^{1/2}}{p_n^{1/2}} - 1 \right) \\ &= 2P_0 \left( \frac{q^{1/2}}{p_n^{1/2}} - 1 \right) + 2 \int (q^{1/2} - p_n^{1/2})(p_0^{1/2} - p_n^{1/2}) \frac{p_0^{1/2} + p_n^{1/2}}{p_n^{1/2}} d\mu.\end{aligned}$$

The first term on the far right equals  $-h^2(q, p_n)$ ; the second can be bounded by the expression  $2h(q, p_n)h(p_0, p_n)(\sqrt{M} + 1)$  in view of the assumption on the quotient  $p_0/p_n$  and the Cauchy-Schwarz inequality. The sum is bounded by  $-\frac{1}{2}h^2(q, p_n)$  if  $h(p_0, p_n)2(\sqrt{M} + 1) \leq \frac{1}{2}h(q, p_n)$ .

The first statement of the theorem follows upon combining this with the inequalities

$$h(2p, p + q) \leq h(p, q) \leq 2h(2p, p + q).$$

These inequalities are valid for every pair of densities  $p$  and  $q$  and show that the Hellinger distance between  $p$  and  $q$  is equivalent to the Hellinger distance between  $p$  and  $\frac{1}{2}(p+q)$ . See Problem 3.4.4.

Next we establish the inequalities relating the Bernstein “norm” and the Hellinger distance given in the discussion preceding the statement of the theorem. Since  $e^{|x|} - 1 - |x| \leq 2(e^{x/2} - 1)^2$ , for every  $x \geq -2$  and  $m_{n,p} \geq -\log 2$ ,

$$\|m_{n,p} - m_{n,p_n}\|_{P_0,B}^2 \lesssim P_0(e^{m_{n,p}/2} - 1)^2 = P_0\left(\frac{(p + p_n)^{1/2}}{(2p_n)^{1/2}} - 1\right)^2.$$

Since  $p_0/p_n \leq M$ , the right side is bounded by  $\frac{1}{2}Mh^2(p + p_n, 2p_n)$ . Combination with the preceding display gives the first inequality. If  $p \leq q$ , then  $m_{n,q} - m_{n,p}$  is nonnegative. By the same inequality for  $e^x - 1 - x$  as before,

$$\|m_{n,p} - m_{n,q}\|_{P_0,B}^2 \lesssim P_0\left(e^{(m_{n,q}-m_{n,p})/2} - 1\right)^2 = P_0\left(\frac{(q + p_n)^{1/2}}{(p + p_n)^{1/2}} - 1\right)^2.$$

This is bounded by  $Mh^2(q + p_n, p + p_n)$  as before.

The maximal inequality is a consequence of Lemma 3.4.3. Each of the functions in  $\mathcal{M}_{n,\delta}$  has Bernstein “norm” bounded by a multiple of  $\delta$ , while a bracket  $[p^{1/2}, q^{1/2}]$  of densities of size  $\delta$  leads to a bracket  $[m_{n,p}, m_{n,q}]$  of Bernstein “norm” of size a multiple of  $\delta$ . ■

It follows that the conditions of Theorem 3.4.1 are satisfied with the Hellinger distance,  $\delta_n = h(p_n, p_0)$ , and

$$\phi_n(\delta) = \tilde{J}_{[]}(\delta, \mathcal{P}_n, h)\left(1 + \frac{\tilde{J}_{[]}(\delta, \mathcal{P}_n, h)}{\delta^2\sqrt{n}}\right),$$

where  $\tilde{J}_{[]}(\delta, \mathcal{P}_n, h)$  is the Hellinger bracketing integral of the sieve  $\mathcal{P}_n$ . (Usually this function has the property that  $\phi_n(\delta)/\delta^\alpha$  is increasing for some  $\alpha < 2$  as required by Theorem 3.4.1; otherwise  $\tilde{J}_{[]}(\delta, \mathcal{P}_n, h)$  must be replaced by a suitable upper bound.) The condition  $r_n^2 \phi_n(1/r_n) \lesssim \sqrt{n}$  is equivalent to

$$r_n^2 \tilde{J}_{[]} \left( \frac{1}{r_n}, \mathcal{P}_n, h \right) \lesssim \sqrt{n}.$$

For the unsieved maximum likelihood estimator the Hellinger integral is independent of  $n$  and any  $r_n$  solving the preceding display gives an upper bound on the rate. Under the condition that the true density  $p_0$  can be approximated by a sequence  $p_n \in \mathcal{P}_n$  such that  $p_0/p_n$  is uniformly bounded, the sieved maximum likelihood estimator that maximizes the likelihood over  $\mathcal{P}_n$  has at least the rate  $r_n$  satisfying both

$$r_n^2 \tilde{J}_{[]} \left( \frac{1}{r_n}, \mathcal{P}_n, h \right) \leq \sqrt{n} \quad \text{and} \quad r_n \lesssim h(p_n, p_0)^{-1}.$$

These are rates for the convergence of the maximum likelihood estimator  $\hat{p}_n$  to the true density  $p_0$  in the Hellinger distance:  $h(\hat{p}_n, p_0) = O_P^*(r_n^{-1})$ .

The last assertion of the theorem—that in the above the Hellinger distance may be replaced by  $h(p + p_n, q + p_n)$  when computing the bracketing entropy of  $\mathcal{P}_n$ —is of some interest, both because it tends to be hard to compute the entropy for the Hellinger distance and because this entropy may behave badly due to the infinite derivative at zero of the root transformation. We have

$$(3.4.5) \quad h_n^2(p, q) = h^2(p + p_n, q + p_n) \leq \int (p - q)^2 \frac{1}{p_n} d\mu.$$

Thus the distance  $h_n$  is bounded by the  $L_2$ -distance with respect to the measure with density  $p_n^{-1}$ . If the latter measure is finite, then the usual upper estimates for the  $L_2$ -entropy of the class of densities  $\mathcal{P}_n$  are relevant and give upper bounds on the entropy of the class of densities with respect to  $h_n$ . While the assumption that  $p_n^{-1}$  is integrable is not natural, at least it allows some conclusions in situations that appear otherwise hard to handle.

**3.4.6 Example (Monotone densities).** Suppose the observations take their values in a compact interval  $[0, T]$  in the real line and are sampled from a density that is known to be nonincreasing. According to Theorem 2.7.5, the set  $\mathcal{F}$  of all nonincreasing functions  $f: [0, T] \mapsto [0, 1]$  has bracketing entropy for the  $L_2(\lambda)$ -norm of the order  $1/\varepsilon$  for any finite measure  $\lambda$  on  $[0, T]$ , in particular Lebesgue measure. If a density  $p$  is nonincreasing, then so is its root  $p^{1/2}$ . Furthermore, the Hellinger distance on the densities is the  $L_2(\lambda)$ -distance on the root densities. Conclude that if  $\mathcal{P}$  is the set of all nonincreasing probability densities bounded by a constant  $C$ , then

$$\log N_{[]}(\varepsilon, \mathcal{P}, h) \leq N_{[]}(\varepsilon, \mathcal{F}, L_2(\lambda)) \lesssim \left(\frac{1}{\varepsilon}\right).$$

Thus  $\tilde{J}_{[]}(\delta, \mathcal{P}, h) \lesssim \delta^{1/2}$ , which yields a rate of convergence of at least  $r_n = n^{1/3}$  for the maximum likelihood estimator. The maximum likelihood estimator is the Grenander estimator. In Example 3.2.14 it is shown that the pointwise rate is exactly  $n^{-1/3}$  if the derivative of the true density is nonzero.

**3.4.7 Example (Convex densities).** Suppose the observations take their values in a compact convex subset of  $\mathbb{R}^d$  and are sampled from a density  $p$  that is known to be convex, uniformly bounded by a constant  $M$ , and uniformly Lipschitz  $|p(y) - p(x)| \leq M\|y - x\|$ . According to Theorem 2.7.10, the collection of all convex functions that are uniformly bounded and uniformly Lipschitz (without the restrictions that it is nonnegative and integrates to 1) has entropy of the order  $\varepsilon^{-d/2}$  for the uniform norm. In view of (3.4.5), the entropy of the class of densities for the distance  $h_1(p, q) = h(p + p_0, q + p_0)$  is bounded by the  $L_2$ -entropy with respect to the measure with Lebesgue

density  $p_0^{-1}$ . Thus, under the assumption that the true density satisfies  $\int p_0^{-1} d\lambda < \infty$ ,

$$\log N_{[]}(\varepsilon, \mathcal{P}, h_1) \lesssim \left(\frac{1}{\varepsilon}\right)^{d/2}.$$

For dimension  $d \leq 3$ , this leads to a converging entropy integral  $\tilde{J}_{[]}(\delta, \mathcal{P}, h_1) \lesssim \delta^{1-d/4}$  and yields a rate of convergence of at least  $r_n = n^{2/(d+4)}$ . For dimension  $d \geq 4$ , the entropy integral diverges, but according to the note following Theorem 3.4.1 (also see Problem 3.4.2), the relevant quantity is

$$\int_{c\delta^2 \wedge \delta/3}^{\delta} \sqrt{\log N_{[]}(\varepsilon, \mathcal{P}, h_1)} d\varepsilon \lesssim \begin{cases} \log\left(\frac{1}{\delta}\right), & \text{if } d = 4, \\ \left(\frac{1}{\delta}\right)^{d/2-2}, & \text{if } d > 4. \end{cases}$$

This leads to a rate of convergence of at least  $r_n = n^{1/4}/(\log n)^{1/2}$  if  $d = 4$  and a rate of convergence of at least  $r_n = n^{1/d}$  if  $d > 4$ .

### 3.4.2 Concave Parametrizations

Assume that the parameter  $\theta$  ranges over a subset of a linear space. Consider criterion functions  $\theta \mapsto \mathbb{P}_n m_\theta$  for measurable functions  $m_\theta$  such that the maps  $\theta \mapsto \exp m_\theta(x)$  are concave. We are mostly thinking of maximum likelihood estimators based on a sample  $X_1, \dots, X_n$  from a density  $p_\theta$  that depends linearly on the parameter. Consequently we denote  $\exp m_\theta(x)$  by  $p_\theta(x)$ . If  $\hat{\theta}_n$  maximizes  $\mathbb{P}_n \log p_\theta$  over a convex subset  $\Theta_n$ , then

$$\mathbb{P}_n \log \frac{p_{\hat{\theta}_n}}{(1-t)p_{\hat{\theta}_n} + tp_{\theta_n}} \geq 0,$$

for all  $0 \leq t \leq 1$  and every  $\theta_n \in \Theta_n$ . Differentiation at  $t = 0$  yields the inequality

$$(3.4.8) \quad \mathbb{P}_n \frac{p_{\theta_n}}{p_{\hat{\theta}_n}} \leq 1, \quad \text{for every } \theta_n \in \Theta_n.$$

If  $L: (0, \infty) \mapsto \mathbb{R}$  is increasing such that the function  $t \mapsto L(1/t)$  is convex, then Jensen's inequality gives

$$\mathbb{P}_n L\left(\frac{p_{\hat{\theta}_n}}{p_{\theta_n}}\right) \geq L\left(\frac{1}{\mathbb{P}_n(p_{\theta_n}/p_{\hat{\theta}_n})}\right) \geq L(1) = \mathbb{P}_n L\left(\frac{p_{\theta_n}}{p_{\theta_n}}\right).$$

Thus the set-up of the introduction applies with

$$m_{n,\theta} = L\left(\frac{p_\theta}{p_{\theta_n}}\right).$$

This conclusion depends essentially on the concavity of the map  $\theta \mapsto p_\theta$ , an assumption that is not made in the preceding section. We can take

advantage of the present structure, which is not uncommon in infinite-dimensional applications, by making a clever choice of the function  $L$ . (This may even depend on  $n$ .) Some possible choices are

$$\log t; \quad \frac{\log(1+t)}{2}; \quad 1 - t^{-\alpha}; \quad t^\alpha - 1; \quad \left(\frac{2t}{t+1}\right)^\alpha - 1; \quad \frac{t^\alpha - 1}{t^\alpha + 1},$$

each time with  $0 < \alpha \leq 1$ . The last two choices are of special interest because they are bounded functions. The first choice leads back to the original definition of  $\hat{\theta}_n$ , showing that (3.4.8) characterizes the  $M$ -estimator. The second choice is used in the preceding section.

### 3.4.3 Least Squares Regression

In this section we consider estimating a regression function  $\theta$  in the model  $Y = \theta(X) + e$  by the method of least-squares. The regression function is typically known to belong to an infinite-dimensional set (or sieve) only. The application of Theorem 3.4.1 to obtain rates of convergence is carried out separately for models with fixed and random design points, by different methods.

#### 3.4.3.1 Fixed Design

Given fixed “design” points  $x_1, \dots, x_n$  in a set  $\mathcal{X}$  and a map  $\theta_0: \mathcal{X} \mapsto \mathbb{R}$ , let

$$Y_i = \theta_0(x_i) + e_i,$$

for independent and identically distributed “error” variables  $e_1, \dots, e_n$ . The observations consist of the pairs  $(x_1, Y_1), \dots, (x_n, Y_n)$  and the unknown regression function  $\theta_0$  is known to belong to a set  $\Theta$ . The sieved least-squares estimator  $\hat{\theta}_n$  minimizes

$$\mathbb{P}_n(Y - \theta)^2 = \frac{1}{n} \sum_{i=1}^n (Y_i - \theta(x_i))^2$$

over a set  $\Theta_n$  of regressions functions. Inserting the expression for  $Y_i$  and calculating the square, we see that  $\hat{\theta}_n$  maximizes

$$2\mathbb{P}_n(\theta - \theta_0)e - \mathbb{P}_n(\theta - \theta_0)^2.$$

The latter criterion function is not observable but is of simpler character than the sum of squares. Note that the second term is assumed nonrandom, the randomness solely residing in the error terms. The set-up of the introduction is valid with

$$m_{n,\theta}(x, e) = 2(\theta - \theta_0)(x)e - \mathbb{P}_n(\theta - \theta_0)^2.$$

Under the assumption that the error variables have mean zero, the mean of this expression is given by  $M_n(\theta) = -\mathbb{P}_n(\theta - \theta_0)^2$  and can be used as a centering function. It satisfies

$$M_n(\theta) - M_n(\theta_n) \leq -\frac{1}{4}\mathbb{P}_n(\theta - \theta_n)^2,$$

for every  $\theta$  such that  $\mathbb{P}_n(\theta - \theta_n)^2 \geq 4\mathbb{P}_n(\theta_n - \theta_0)^2$  (Problem 3.4.5). Thus, Theorem 3.4.1 applies with  $d_n(\theta, \theta_n)$  equal to the  $L_2(\mathbb{P}_n)$ -distance on the set of regression functions. The necessary maximal inequality takes the form

$$\mathbf{E}^* \sup_{\mathbb{P}_n(\theta - \theta_n)^2 < \delta^2, \theta \in \Theta_n} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n (\theta - \theta_n)(x_i) e_i \right| \lesssim \phi_n(\delta).$$

Since the design points are nonrandom, this involves relatively simple multiplier processes, to which the abstract maximal inequalities of Chapter 2.2 may apply directly. In particular, if the error variables are Gaussian, then the stochastic process  $\{n^{-1/2} \sum_{i=1}^n \theta(x_i) e_i : \theta \in \Theta\}$  is sub-Gaussian for the  $L_2(\mathbb{P}_n)$ -semimetric on the set of regression functions. Corollary 2.2.8 shows that we may choose

$$\phi_n(\delta) = \int_0^\delta \sqrt{\log N(\varepsilon, \Theta_n, L_2(\mathbb{P}_n))} d\varepsilon.$$

This conclusion depends only the sub-Gaussianity of the stochastic process  $\{n^{-1/2} \sum_{i=1}^n \theta(x_i) e_i : \theta \in \Theta\}$  for the  $L_2(\mathbb{P}_n)$ -semimetric. By Proposition A.1.6, this is more generally true if the error variables are sub-Gaussian.

For errors with heavier than Gaussian tails, Theorem 2.2.4 may be applied with a different Orlicz norm. For instance, in case of errors with sub-exponential tails, we have by Proposition A.1.6, followed by Lemma 2.2.2:

$$\begin{aligned} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \theta(x_i) e_i \right\|_{\psi_1} &\lesssim \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \theta(x_i) e_i \right\|_1 + \left\| \max_{1 \leq i \leq n} n^{-1/2} \theta(x_i) e_i \right\|_{\psi_1} \\ &\leq \|\theta\|_{\mathbb{P}_n, 2} \|e\|_2 + \frac{\log n}{\sqrt{n}} \|\theta\|_{\mathbb{P}_n, \infty} \|e\|_{\psi_1} \\ &\lesssim \log n \|\theta\|_{\mathbb{P}_n, 2} \|e\|_{\psi_1}. \end{aligned}$$

Then Theorem 2.2.4 shows that we may choose

$$\phi_n(\delta) = \int_0^{\delta \log n} \log N(\varepsilon, \Theta_n, \log n L_2(\mathbb{P}_n)) d\varepsilon.$$

The  $\log n$  terms do not affect the upper bound on the rate of convergence much. However, taking the integral of the entropy, rather than the square-root entropy, makes much difference. For instance, the integral does not appear to converge for monotone regression functions. Thus, this approach does not necessarily lead to a sharp upper bound on the rate. The discrepancy might be due to the fact that the present application of Theorem 2.2.4 uses the special sum structure of the process  $\{n^{-1/2} \sum_{i=1}^n \theta(x_i) e_i : \theta \in \Theta\}$

only through their  $\psi_1$ -norms. The maximal inequalities of Chapter 2.2 are designed especially for sums of independent processes and may be expected to give sharper bounds. Another explanation of the discrepancy is that too much is lost in our attempt to bound the  $\psi_1$ -norm by the  $L_2(\mathbb{P}_n)$ -norm. Unfortunately, it appears hard to use better estimates in general, since we are interested in the modulus with respect to the  $L_2(\mathbb{P}_n)$ -norm.

Another approach is possible if the class of regression functions  $\Theta_n$  has bounded uniform entropy. Symmetrization by Rademacher variables (Lemma 2.3.1) followed by application of the sub-Gaussian maximal inequality (Corollary 2.2.8) conditionally on the observations yields the desired maximal inequality with

$$\phi_n(\delta) = \mathbb{E}^* \int_0^{d_n} \sup_Q \sqrt{\log N(\varepsilon \|M_n\|_{Q,2}, \Theta_n, L_2(Q))} d\varepsilon (\mathbb{P}_n M_n^2 e^2)^{1/2}.$$

Here the supremum is taken over all finitely discrete measures  $Q$ , the function  $M_n(x)$  is an envelope function for  $\Theta_n$ , and

$$d_n^2 = \sup_{\mathbb{P}_n(\theta - \theta_n)^2 < \delta^2, \theta \in \Theta_n} \frac{\mathbb{P}_n(\theta - \theta_n)^2 e^2}{\mathbb{P}_n M_n^2 e^2} \lesssim \frac{\delta^{2/p} (\mathbb{P}_n M_n^{(1-2/p)q} e^{2q})^{1/q}}{\mathbb{P}_n M_n^2 e^2}.$$

The last step follows by Hölder's inequality for every  $1/p + 1/q = 1$ . Assume that the uniform entropy (the square of the integrand) is bounded by  $\varepsilon^{-1/m}$  for  $m > 1/2$ . Then

$$\phi_n(\delta) \leq \delta^{1/p(1-1/2m)} \mathbb{E}^* \left( \mathbb{P}_n M_n^{(1-2/p)q} e^{2q} \right)^{(1-1/2m)/2q} \left( \mathbb{P}_n M_n^2 e^2 \right)^{1/4m}$$

If the expectation on the right-hand side is bounded as  $n \rightarrow \infty$ , then this leads to a rate of convergence of at least

$$r_n = n^{\frac{mp}{(4p-2)m+1}}.$$

For instance, if the class of regression functions is uniformly bounded, then this is a valid upper bound on the rate if  $\mathbb{E}|e|^{2q} < \infty$ . For every  $q$  the upper bound  $r_n$  is slower than the rate  $n^{m/(2m+1)}$  obtained under sub-Gaussian tails, although this rate is approached arbitrarily closely as  $q \rightarrow \infty$ . It seems not unreasonable to expect that slower rates may pertain under just polynomial tails of the error distribution, but from the present derivation it is unclear whether the present upper bound is sharp or is the outcome of the application of suboptimal inequalities.

**3.4.9 Example.** If the set of regression functions is known to belong to  $C_1^\alpha[K]$  for a compact convex subset  $K$  of  $\mathbb{R}^d$ , then the unsieved least-squares estimator attains a rate of convergence of at least

$$n^{\frac{\alpha p}{(4p-2)\alpha+d}}$$

if the errors have finite  $2q$ th moment. For errors with sub-Gaussian tails, the rate of convergence is at least  $n^{\alpha/(2\alpha+d)}$ , which is the limit as  $q \rightarrow \infty$  of the display.

A third approach to least-squares uses finite-dimensional sieves and can be analyzed by an elementary maximal inequality. For each  $n$  let  $\psi_{1,n}, \dots, \psi_{N_n,n}$  be functions from  $\mathcal{X}$  to  $\mathbb{R}$  that form an orthonormal system in  $L_2(\mathbb{P}_n^x)$ . The number of elements  $N_n \leq n$  of the system is to be determined later. Consider the sieves

$$\Theta_n = \text{lin}(\psi_{1,n}, \dots, \psi_{N_n,n}).$$

Expressing  $\theta - \theta_n$  in terms of the orthonormal system and applying the Cauchy-Schwarz inequality, we obtain (if  $\theta_n \in \Theta_n$ )

$$\begin{aligned} \mathbb{E}^* \sup \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \theta(x_i) e_i \right|^2 &\leq \mathbb{E}^* \left[ \sup \sum_{j=1}^{N_n} \langle \theta, \psi_{j,n} \rangle \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{j,n}(x_i) e_i \right) \right]^2, \\ &\leq \sup \sum_{j=1}^{N_n} \langle \theta, \psi_{j,n} \rangle^2 \mathbb{E}^* \left( \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi_{j,n}(x_i) e_i \right)^2, \\ &\leq \delta^2 N_n \mathbb{E} e^2, \end{aligned}$$

where each time the supremum is taken over all  $\theta \in \Theta_n - \theta_n$  such that  $\mathbb{P}_n \theta^2 < \delta^2$ . This approach gives useful results under just the assumption that the errors are mean zero with finite variance. The rate of convergence of the sieved least-squares estimator depends on the upper bound in combination with the approximation error  $\|\theta_0 - \Theta_n\|_{\mathbb{P}_n,2}$ .

**3.4.10 Example.** Consider the design points  $x_i = i/n$  when the set of regression functions  $\theta: [0, 1] \mapsto \mathbb{R}$  is known to be Lipschitz of order  $0 < \alpha \leq 1$ . The Lipschitz constants are not assumed uniformly bounded. A simple choice of the sieve is given by letting the base functions be multiples of the indicators of the cells  $((i-1)/N, i/N]$ . Discretization of  $\theta_0$  on the points  $i/N$  shows that

$$\|\theta_0 - \Theta_n\| \leq \|\theta_0\|_{\text{Lip}} \frac{1}{N^\alpha}.$$

In combination with the preceding maximal inequality, a rate of convergence for the sieved least-squares estimator in the  $L_2(\mathbb{P}_n)$ -semimetric is given by the solution of

$$r_n^2 \frac{1}{r_n} \sqrt{N_n} \leq \sqrt{n}; \quad \text{and} \quad \frac{1}{r_n} \geq \frac{1}{N_n^\alpha}.$$

The minimal choice  $N_n = r_n^{1/\alpha}$  leads to the rate  $r_n = n^{\alpha/(2\alpha+1)}$ . This rate is known to be optimal in this case.

### 3.4.3.2 Random Design

Another approach towards finding an upper bound on the rate of least-squares estimators uses the maximal inequality Lemma 3.4.3 based on the Bernstein “norm”. It is convenient to develop this approach for the situation wherein the design points are i.i.d. random variables. Thus  $X_1, \dots, X_n$  and  $e_1, \dots, e_n$  are independent i.i.d. samples and the observations consist of the pairs  $(X_1, Y_1), \dots, (X_n, Y_n)$ , where each  $Y_i$  is defined by

$$Y_i = \theta_0(X_i) + e_i.$$

As in the case of fixed design points considered in the preceding section, we consider the least-squares estimator  $\hat{\theta}_n$ , which minimizes the sum of squares  $n^{-1} \sum_{i=1}^n (Y_i - \theta(X_i))^2$  over a sieve  $\Theta_n$ . Equivalently, it maximizes

$$\mathbb{M}_n(\theta) = 2\mathbb{P}_n(\theta - \theta_0)e - \mathbb{P}_n(\theta - \theta_0)^2.$$

Under the assumption that the errors have mean zero, this leads to the centering function  $M(\theta) = -P(\theta - \theta_0)^2$ , which satisfies

$$M(\theta) - M(\theta_n) \leq -\frac{1}{4}P(\theta - \theta_n)^2,$$

for every  $\theta$  such that  $P(\theta - \theta_n)^2 \geq 4P(\theta_n - \theta_0)^2$ . Thus Theorem 3.4.1 applies with  $d_n$  equal to the  $L_2$ -distance on the regression functions. If the class of regression functions is uniformly bounded by  $M \geq 1/2$ , then

$$\begin{aligned} P\left(e^{t|\theta-\theta_n||e|} - 1 - t|\theta - \theta_n||e|\right) &= \sum_{m \geq 2} \frac{P|\theta - \theta_n|^m P|te|^m}{m!} \\ &\leq P(\theta - \theta_n)^2 \mathbb{E}e^{2Mt|e|}. \end{aligned}$$

If the error variables have subexponential tails, then it follows that the Bernstein “norm” of a multiple of the variables  $(\theta - \theta_n)(X_i)e_i$  is bounded by a multiple of the  $L_2$ -norm of  $\theta - \theta_n$ . A bracket  $[\theta_1, \theta_2]$  for regression functions leads to a bracket  $[\theta_1 e^+ - \theta_2 e^-, \theta_2 e^+ - \theta_1 e^-]$  for the variables  $\theta(X_i)e_i$ . In view of Lemma 3.4.3,

$$\begin{aligned} \mathbb{E}^* \sup_{P(\theta - \theta_n)^2 < \delta^2, \theta \in \Theta_n} &\left| \frac{1}{\sqrt{n}} \sum_{i=1}^n (\theta - \theta_n)(X_i) e_i \right| \\ &\lesssim \tilde{J}_{[]}(\delta, \Theta_n, L_2(P)) \left( 1 + \frac{\tilde{J}_{[]}(\delta, \Theta_n, L_2(P))}{\delta^2 \sqrt{n}} \right). \end{aligned}$$

This essentially gives the same rates of convergence as the approach based on Theorem 2.2.4 in the preceding section. Presently, the error variables are assumed to be subexponential rather than sub-Gaussian. On the other hand the set of regression functions is assumed uniformly bounded.

### 3.4.4 Least-Absolute-Deviation Regression

The results on least-squares regression can be easily generalized to other minimization schemes. Typically, the rates of convergence of the estimators are the same, but the conditions on the error variables necessary to obtain these rates vary strongly. This is illustrated in this section by the case of least-absolute-deviation estimators.

In the set-up of Section 3.4.3 with fixed design points, let  $\hat{\theta}_n$  minimize the sum

$$\frac{1}{n} \sum_{i=1}^n |Y_i - \theta(x_i)|$$

over the set of possible regression functions  $\Theta$ . If the error distribution has median zero and is smooth, then the map  $\mu \mapsto P|e + \mu|$  will be twice differentiable at its point of maximum  $\mu = 0$ . Then for  $\|\theta - \theta_0\|_\infty$  close enough to zero,

$$\frac{1}{n} \sum_{i=1}^n \left( P|e_i - (\theta - \theta_0)(x_i)| - P|e| \right) \lesssim -\frac{1}{n} \sum_{i=1}^n (\theta - \theta_0)^2(x_i).$$

In order to apply Theorem 3.4.1 with the empirical  $L_2$ -semimetric of the design points, it suffices to show that  $\|\hat{\theta}_n - \theta_0\|$  converges to zero (so that we can take  $\Theta_n$  to be inside a small ball in the uniform norm around  $\theta_0$ ) and to control

$$E^* \sup_{\mathbb{P}_n(\theta - \theta_0)^2 < \delta^2} \left| \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( |e_i - (\theta - \theta_0)(x_i)| - |e_i| - P|e - (\theta - \theta_0)(x_i)| + P|e| \right) \right|.$$

The centered variables within the large parentheses are bounded in absolute value by  $2|\theta - \theta_0|(x_i)$ . In view of Proposition A.1.6,

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \left( |e_i - (\theta - \theta_0)(x_i)| - |e_i| - P|e - \theta(x_i)| + P|e| \right) \right\|_{\psi_2} \lesssim (\mathbb{P}_n(\theta - \theta_0)^2)^{1/2}$$

Thus the  $\psi_2$ -norm of the increments is bounded by the  $L_2(\mathbb{P}_n)$ -norm of the regression functions. According to Corollary 2.2.8, Theorem 3.4.1 can be applied with

$$\phi_n(\delta) = \int_0^\delta \sqrt{\log N(\varepsilon, \Theta_n, L_2(\mathbb{P}_n))} d\varepsilon.$$

Note that this is true without any tail conditions on the error distribution. It can be concluded that the method of least-absolute-deviations estimation is very robust against outliers, also in the sense of stabilizing rates of convergence in nonparametric regression problems.

## Problems and Complements

1. If the conditions of Theorem 3.4.1 hold for  $\eta = \infty$ , then  $r_n d_n(\hat{\theta}_n, \theta_n)$  converges to zero in mean. If the second condition of the theorem is strengthened to

$$\left\| \sup_{\delta/2 < d_n(\theta, \theta_n) \leq \delta, \theta \in \Theta_n} \sqrt{n} [(\mathbb{M}_n - M_n)(\theta) - (\mathbb{M}_n - M_n)(\theta_n)]^+ \right\|_p \lesssim \phi_n(\delta),$$

then the sequence  $r_n d_n(\hat{\theta}_n, \theta_n)$  converges to zero in  $p$ th mean. Finally, to obtain the conclusion of the theorem (convergence in probability), it suffices that a probability version of this condition is valid.

[Hint: Bound  $E^* r_n^p d_n^p(\hat{\theta}_n, \theta_n)$  by the sum  $\sum 2^{jp} P^*(\hat{\theta}_n \in S_{j,n})$ .]

2. The conditions in Theorem 3.4.1 can be relaxed to

$$\sup_{\delta/2 < d_n(\theta, \theta_n) \leq \delta, \theta \in \Theta_n} M_n(\theta) - M_n(\theta_n) \leq -c\delta^2,$$

$$E^* \sup_{\delta/2 < d_n(\theta, \theta_n) \leq \delta, \theta \in \Theta_n} \sqrt{n} [(\mathbb{M}_n - M_n)(\theta) - (\mathbb{M}_n - M_n)(\theta_n) - d\delta^2]^+ \lesssim \phi_n(\delta),$$

for constants  $c > d \geq 0$ . If the other conditions of the theorem remain valid, so does the conclusion.

3. Under the conditions of Lemma 3.4.2 or 3.4.3, the expectation

$$E_P^* \sup_f (G_n f - \eta \sqrt{n})^+$$

is bounded by the right side of the lemma, but with the entropy integrals computed over the interval  $[\eta/8, \delta]$  rather than  $[0, \delta]$ .

4. For any nonnegative numbers  $p$  and  $q$ ,

$$|(2p)^{1/2} - (p+q)^{1/2}| \leq |p^{1/2} - q^{1/2}| \leq (1 + \frac{1}{2}\sqrt{2})|(2p)^{1/2} - (p+q)^{1/2}|.$$

Thus  $h(p, q)$  and  $h(2p, p+q)$  are equivalent up to constants.

[Hint: The lower inequality is trivial. The upper is valid with constant  $\sqrt{2}$  if  $q \geq p$  and with constant  $1 + \frac{1}{2}\sqrt{2}$  as stated if  $q \leq p$ .]

5. For any elements  $x$ ,  $y$ , and  $z$  in a normed space and  $c \geq 2$ , we have  $\|x - z\|^2 - \|y - z\|^2 \geq (1 - 2/c)^2 \|x - y\|^2$  whenever  $\|x - y\| \geq c\|z - y\|$ .

[Hint: By the triangle inequality,  $\|x - y\| \leq \|x - z\| + \|y - z\| + 2\|y - z\|$ . Bound  $\|z - y\|$  by  $1/c$  times the left side and thus obtain

$$(1 - 2/c)\|x - y\| \leq \|x - z\| + \|y - z\|.$$

Take squares and use the fact that the right-hand side is nonnegative, since  $c \geq 2$ , so that one difference in the square on the right can be replaced by a sum.]

6. Consider Example 3.2.15 with the additional assumption that the distribution  $G$  of the observation times is fixed and known. Consider the class  $\mathcal{P}$  of

densities with respect to  $\mu = G \times \nu$ , where  $\nu$  is the counting measure on  $\{0, 1\}$ , given by

$$\mathcal{P} = \{p_F(t, \delta) = F(t)^\delta (1 - F(t))^{1-\delta} : F \text{ is a d.f. on } \mathbb{R}\}.$$

Let  $\hat{F}$  denote the maximum likelihood estimator as described in Example 3.2.15, given a sample  $(T_1, \Delta_1), \dots, (T_n, \Delta_n)$  from  $p_{F_0}$ . Show that  $h(p_{\hat{F}_n}, p_{F_0}) = O_p(n^{-1/3})$ .

[Hint: See Van de Geer (1993a), pages 25 and 35.]

7. Does the result in Exercise 3.4.6 change if  $G$  is unknown?

## 3.5

# Random Sample Size, Poissonization and Kac Processes

It can be argued that in practice the number of available observations is random and perhaps dependent on the same random phenomenon. In the first section it is shown in fair generality that the empirical central limit theorem is valid also for random sample size.

The Kac process is a Poisson process obtained by taking the sample size in the empirical distribution to be a Poisson distributed variable independent of the data. We study the convergence of the Kac empirical process.

### 3.5.1 Random Sample Size

Suppose that at “time”  $n$  a random number  $N_n$  of observations from an infinite i.i.d. sequence  $X_1, X_2, \dots$  is available. Formally,  $N_n$  is any integer-valued, nonnegative map defined on the same probability space as the sequence of observations. Assume that the sequence  $N_n$  converges to infinity in the strong sense that

$$\frac{N_n}{c_n} \xrightarrow{\text{P}} \nu,$$

for a strictly positive random variable  $\nu$  and a deterministic sequence  $c_n \rightarrow \infty$ . Let  $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$  be the empirical process as before. Then  $\mathbb{G}_{N_n}$  converges in distribution whenever the sequence  $\mathbb{G}_n$  does, to the same limit.

**3.5.1 Theorem.** *Let  $\mathcal{F}$  be a Donsker class of measurable functions. Suppose that  $N_n$  is a sequence of positive, integer-valued random variables such*

that  $N_n/c_n \xrightarrow{P} \nu$  for a random variable  $\nu$  with  $P(\nu > 0) = 1$  and a deterministic sequence  $c_n \rightarrow \infty$ . Then the sequence  $\mathbb{G}_{N_n}$  converges in distribution in  $\ell^\infty(\mathcal{F})$  to a tight Brownian bridge as  $n \rightarrow \infty$ .

**Proof.** First, suppose that  $c_n = n$  and  $N_n/n \leq M$  for some number  $M$ . Let  $\mathbb{Z}_n(s, f) = n^{-1/2} \sum_{i=1}^{\lfloor ns \rfloor} (\delta_{X_i} - P)$  be the sequential empirical process. By a minor extension of Theorem 2.12.1, the sequence  $\mathbb{Z}_n$  converges in distribution in  $\ell^\infty([0, M] \times \mathcal{F})$  to a Kiefer-Müller process  $\mathbb{Z}$ .

The lemma below shows that the sequence  $(\mathbb{Z}_n, N_n/n)$  converges weakly in the space  $\ell^\infty([0, M] \times \mathcal{F}) \times \mathbb{R}$  to a pair  $(\mathbb{Z}, \nu)$  of independent random elements  $\mathbb{Z}$  and  $\nu$ . Now

$$\frac{1}{\sqrt{n}} \sum_{i=1}^{N_n} (\delta_{X_i} - P) f = \mathbb{Z}_n \left( \frac{N_n}{n}, f \right) = g \left( \mathbb{Z}_n, \frac{N_n}{n} \right) f,$$

where  $g: \ell^\infty([0, M] \times \mathcal{F}) \times \mathbb{R} \mapsto \ell^\infty(\mathcal{F})$  is defined by  $g(z, r) = z(r, \cdot)$ . The map  $g$  is continuous at every point  $(z, r)$  such that  $s \mapsto z(s, \cdot)$  is continuous with respect to  $\|\cdot\|_{\mathcal{F}}$ . The paths of the Kiefer-Müller process  $\mathbb{Z}$  possess the latter continuity property with probability 1. Thus, the continuous mapping theorem yields  $g(\mathbb{Z}_n, N_n/n) \rightsquigarrow g(\mathbb{Z}, \nu) = \mathbb{Z}(\nu, \cdot)$ . By the same argument,

$$\mathbb{G}_{N_n} = \frac{1}{\sqrt{N_n}} \sum_{i=1}^{N_n} (\delta_{X_i} - P) \rightsquigarrow \frac{1}{\sqrt{\nu}} \mathbb{Z}(\nu, \cdot).$$

Since  $\nu$  and  $\mathbb{Z}$  are independent and  $\nu^{-1/2} \mathbb{Z}(\nu, \cdot)$  is distributed as a Brownian bridge for every deterministic  $\nu$ , the variable on the right is distributed as a Brownian bridge. This concludes the proof under the special assumptions on the  $N_n$ .

If  $N_n/n$  is not bounded (but still  $c_n = n$ ), define  $N_{n,M} = N_n \wedge (Mn)$ . By the preceding argument,  $\mathbb{G}_{N_{n,M}} \rightsquigarrow \mathbb{G}$  for every  $M$ . The probability that  $N_{n,M} \neq N_n$  can be made arbitrarily small by choice of  $M$ . The theorem follows.

Finally, the case that  $c_n \neq n$  can be treated by relabeling the indexes. We may assume that each  $c_n$  is an integer. Furthermore, since it suffices to show that every subsequence of  $\{n\}$  has a further subsequence along which  $\mathbb{G}_{N_n}$  converges, it is not a loss of generality to assume that  $c_n$  is also strictly increasing. Define  $N'_k = N_n$  if  $c_n = k$  and define  $N'_k = k\nu$  if  $k \neq c_n$  for every  $n$ . Then  $N'_k/k \xrightarrow{P} \nu$ , hence  $\mathbb{G}_{N'_k} \rightarrow \mathbb{G}$ . The sequence  $\mathbb{G}_{N_n}$  is a subsequence. ■

**3.5.2 Lemma.** Let  $\mathcal{F}$  be a Donsker class and  $\nu_n$  a sequence of random variables such that  $\nu_n \xrightarrow{P} \nu$  for a random variable  $\nu$ . Then the sequence of sequential empirical processes  $\mathbb{Z}_n$  satisfies  $(\mathbb{Z}_n, \nu_n) \rightsquigarrow (\mathbb{Z}, \nu)$  in  $\ell^\infty([0, M] \times \mathcal{F}) \times \mathbb{R}$ , where  $\mathbb{Z}$  and  $\nu$  are independent.

**Proof.** Let  $k_n \rightarrow \infty$  slowly enough that  $k_n = o(\sqrt{n})$ . Set

$$\mathbb{Z}'_n(s, f) = \frac{1}{\sqrt{n}} \sum_{i=k_n+1}^{\lfloor ns \rfloor} (\delta_{X_i} - P).$$

Then the sequence  $\mathbb{Z}'_n - \mathbb{Z}_n$  converges to zero in probability in  $\ell^\infty([0, M] \times \mathcal{F})$ . By a version of Slutsky's lemma, the sequence  $(\mathbb{Z}'_n, \nu)$  has the same limit distribution as the sequence  $(\mathbb{Z}_n, \nu_n)$ . By Doob's martingale convergence theorem,  $P(\nu \in B | X_1, \dots, X_k) \rightarrow P(\nu \in B | X_1, X_2, \dots)$  in mean as  $k \rightarrow \infty$ . Conclude that

$$\lim_{n \rightarrow \infty} P^*(\mathbb{Z}'_n \in A, \nu \in B) = \lim_{n \rightarrow \infty} E\left(1_A(\mathbb{Z}'_n)^* P(\nu \in B | X_1, \dots, X_{k_n})\right).$$

Since  $\mathbb{Z}'_n$  is independent of  $X_1, \dots, X_{k_n}$ , the expectation in the right side can be factorized as  $P^*(\mathbb{Z}'_n \in A) P(\nu \in B)$ . This converges to  $P(\mathbb{Z} \in A) P(\nu \in B)$  for every continuity set  $A$ . ■

The assumption on the sequence  $N_n$  in the preceding theorem is remarkably weak. Under the stronger assumption that  $N_n/n \xrightarrow{P} 1$ , the sequences  $\mathbb{G}_{N_n}$  and  $\mathbb{G}_n$  are not only asymptotically equal in law, but also equivalent in probability.

**3.5.3 Theorem.** *Let  $\mathcal{F}$  be a Donsker class of measurable functions. Suppose that  $N_n$  is a sequence of positive, integer-valued random variables such that  $N_n/n \xrightarrow{P} 1$ . Then the sequence  $\mathbb{G}_{N_n} - \mathbb{G}_n$  converges in outer probability to zero in  $\ell^\infty(\mathcal{F})$  as  $n \rightarrow \infty$ .*

**Proof.** With the same notation as in the proof of Theorem 3.5.1, the sequence  $(\mathbb{Z}_n, N_n/n)$  converges in distribution to  $(\mathbb{Z}, 1)$  in the space  $\ell^\infty([0, 2] \times \mathcal{F}) \times \mathbb{R}$ . By the continuous mapping theorem,  $\mathbb{Z}_n(N_n/n, \cdot) - \mathbb{Z}_n(1, \cdot) \rightsquigarrow \mathbb{Z}(1, \cdot) - \mathbb{Z}(1, \cdot) = 0$  in  $\ell^\infty(\mathcal{F})$ . Convergence in distribution and in outer probability to a degenerate limit are equivalent. ■

### 3.5.2 Poissonization

Let the sample size  $N_n$  at “time”  $n$  be a Poisson variable with mean  $n$  and independent of the i.i.d. observations  $X_1, X_2, \dots$ . The random measure

$$\mathbb{N}_n = \sum_{i=1}^{N_n} \delta_{X_i}$$

is the *Kac empirical point process*. Straightforward calculations show that, for every measurable set  $C$ , the random variable  $\mathbb{N}_n(C)$  is Poisson-distributed with mean  $n P(C)$  (where  $P(C) = P(X_i \in C)$ ). Furthermore, for every finite collection  $C_1, \dots, C_k$  of pairwise-disjoint measurable sets,

the random variables  $\mathbb{N}_n(C_1), \dots, \mathbb{N}_n(C_k)$  are independent. Thus the Kac process  $\mathbb{N}_n$  is a (generalized) Poisson process on the sample space  $(\mathcal{X}, \mathcal{A})$  with intensity measure  $nP$ .

Given a class  $\mathcal{F}$  of measurable functions  $f: \mathcal{X} \mapsto \mathbb{R}$ , consider the process  $\{\mathbb{N}_n(f): f \in \mathcal{F}\}$ . Its mean and variance function equal  $E\mathbb{N}_n(f) = nPf = \text{var } \mathbb{N}_n(f)$ . A standardized version of this process is

$$\mathbb{Z}_n = \frac{1}{\sqrt{n}}(\mathbb{N}_n - nP) = \sqrt{\frac{N_n}{n}}\mathbb{G}_{N_n} + \sqrt{n}\left(\frac{N_n}{n} - 1\right)P,$$

where  $\mathbb{G}_n = \sqrt{n}(\mathbb{P}_n - P)$  is the empirical process. The collection  $\mathcal{F}$  is called  $P$ -Kac if the sequence  $\mathbb{Z}_n$  converges in distribution to a tight limit process in  $\ell^\infty(\mathcal{F})$ . (The sample paths of  $\mathbb{Z}_n$  are in  $\ell^\infty(\mathcal{F})$  if  $\mathcal{F}$  possesses a finite envelope function and  $\|P\|_{\mathcal{F}} < \infty$ .)

Since  $N_n/n \xrightarrow{P} 1$ , Theorem 3.5.3 yields that  $\mathbb{G}_{N_n} - \mathbb{G}_n \xrightarrow{P} 0$  if  $\mathcal{F}$  is Donsker. In that case the sequence  $\mathbb{Z}_n$  is equivalent to the sequence  $\mathbb{G}_n + \sqrt{n}(N_n/n - 1)P$  and converges in distribution to  $\mathbb{G} + ZP$  for an independent Brownian bridge  $\mathbb{G}$  and  $N(0, 1)$  variable  $Z$ . Thus the limit process  $\mathbb{Z} = \mathbb{G} + ZP$  in the Kac central limit theorem is a Brownian motion process and has covariance function

$$E\mathbb{Z}(f)\mathbb{Z}(g) = Pf g.$$

Each  $\mathbb{Z}_n$  has the same covariance function,  $E\mathbb{Z}_n(f)\mathbb{Z}_n(g) = Pf g$ .

We have shown that every Donsker class with  $\|P\|_{\mathcal{F}} < \infty$  is Kac. Actually, these two concepts are equivalent, which can be seen from the following concentration lemma.

Since  $\mathbb{N}_n$  is a Poisson process with intensity measure  $nP$ , it can be written as the sum of  $n$  i.i.d. Poisson processes of intensity measure  $P$ . Formally, let  $Y_1, Y_2, \dots$  be an i.i.d. sequence of  $\text{Poisson}(1)$  variables, and let  $X_{i,j}$  be an array of i.i.d. copies of  $X_1$ . Then the process

$$H_n = \sum_{i=1}^n \sum_{j=1}^{Y_i} (\delta_{X_{i,j}} - P)$$

is equal in distribution to  $H'_n = \sum_{i=1}^{N_n} (\delta_{X_i} - P)$  (in the sense that  $E^*h(H_n) = E^*h(H'_n)$  for every  $h$  if the  $Y_i$  and  $X_{i,j}$  are suitably defined as coordinate projections on a product space). It follows that the random-sample central limit theorem for  $\mathbb{G}_{N_n}$  is equivalent to the central limit theorem for a deterministic number of Poisson processes of the type  $\sum_{j=1}^{Y_i} (\delta_{X_{i,j}} - P)$ . Le Cam's lemma compares the concentration of these processes with the concentration of the empirical process.

**3.5.4 Lemma (Le Cam's Poissonization lemma).** Let  $N_n$  be Poisson-distributed with mean  $n$  and be independent of the i.i.d. zero-mean stochastic processes  $Z_1, \dots, Z_n$ . Then for any class of functions  $\mathcal{F}$

$$\left(1 - \frac{1}{e}\right) E^* \left\| \sum_{i=1}^n Z_i \right\|_{\mathcal{F}} \leq E^* \left\| \sum_{i=1}^{N_n} Z_i \right\|_{\mathcal{F}}.$$

**Proof.** A Poisson variable  $Y_i$  with mean 1 satisfies  $E(Y_i \wedge 1) = 1 - e^{-1}$ . Let the array  $Z_{i,j}$  consist of i.i.d. copies of  $Z_1$  independent of  $Y_1, \dots, Y_n$ . The left side of the lemma equals

$$\begin{aligned} E_Z^* \left\| E_Y \sum_{i=1}^n (Y_i \wedge 1) Z_i \right\|_{\mathcal{F}} &\leq E^* \left\| \sum_{i=1}^n (Y_i \wedge 1) Z_i \right\|_{\mathcal{F}} \\ &\leq E_Y E_Z^* \left\| \sum_{i=1}^n \sum_{j=1}^{Y_i} Z_{i,j} \right\|_{\mathcal{F}}. \end{aligned}$$

Here the factorization of  $E^*$  as  $E_Y E_Z^*$  is warranted by Lemma 1.2.7, and in the last step we replace 0 by 0 if  $Y_i = 0$ ,  $Z_i$  by  $Z_{i,1}$  if  $Y_i = 1$ ,  $Z_i$  by  $Z_{i,1} + Z_{i,2}$  if  $Y_i = 2$ , etc. The inequality is valid in view of the inequality  $E^* \|Z_1\| = E^* \|Z_1 + EZ_2\| \leq E^* \|Z_1 + Z_2\|$  for independent processes  $Z_1$  and  $Z_2$  with  $EZ_2 = 0$ .

By the discussion preceding the lemma, the double sum on the right side is equal in distribution to  $\sum_{i=1}^{N_n} Z_i$ . ■

**3.5.5 Theorem.** A class  $\mathcal{F}$  of measurable functions with  $\|P\|_{\mathcal{F}} < \infty$  is Kac if and only it is Donsker. In that case  $\|\mathbb{G}_{N_n} - \mathbb{G}_n\|_{\mathcal{F}}^* = O_P(n^{-1/4})$ .

**Proof.** Set  $\mathbb{G}'_n = n^{-1/2} \sum_{i=1}^{N_n} (\delta_{X_i} - P)$ . If  $N_n = n + k$ , then the difference  $\mathbb{G}_n - \mathbb{G}'_n$  is the sum of  $k$  terms  $n^{-1/2}(\delta_{X_i} - P)$ . Since  $P(|N_n - n| \geq M\sqrt{n}) \leq M^{-2}$  by Chebyshev's inequality, we have for every  $\varepsilon > 0$

$$\begin{aligned} P^*(\|\mathbb{G}_n - \mathbb{G}'_n\|_{\mathcal{F}} > \varepsilon) &\leq \frac{1}{M^2} + \frac{1}{\varepsilon} \sum_{|k| < M\sqrt{n}} P(N_n = n + k) E^* \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{|k|} (\delta_{X_i} - P) \right\|_{\mathcal{F}} \\ &\leq \frac{1}{M^2} + \frac{1}{\varepsilon\sqrt{n}} E^* \left\| \sum_{i=1}^{\lfloor M\sqrt{n} \rfloor} (\delta_{X_i} - P) \right\|_{\mathcal{F}}. \end{aligned}$$

Here the last step follows, since the outer expectations  $E^* \left\| \sum_{i=1}^{|k|} (\delta_{X_i} - P) \right\|_{\mathcal{F}}$  are non-decreasing in  $|k|$  by Jensen's inequality.

If  $\mathcal{F}$  is Donsker, then the sequence  $n^{-1/4} E^* \left\| \sum_{i=1}^{\lfloor M\sqrt{n} \rfloor} (\delta_{X_i} - P) \right\|_{\mathcal{F}}$  is bounded by Lemma 2.3.11. In that case the calculation in the preceding paragraph shows that the probability  $P^*(\|\mathbb{G}_n - \mathbb{G}'_n\|_{\mathcal{F}} > Kn^{-1/4})$  is

bounded by  $M^{-2} + K^{-1}O(1)$ . Its limsup as  $n \rightarrow \infty$  can be made arbitrarily small by choice of  $M$  and  $K$ . Thus  $\|\mathbb{G}_n - \mathbb{G}'_n\|_{\mathcal{F}}^* = O_P(n^{-1/4})$ . The theorem follows, because  $\|\mathbb{G}'_n - \mathbb{G}_{N_n}\|_{\mathcal{F}}^* = O_P(n^{-1/2})$ .

If  $\mathcal{F}$  is Kac, then  $\mathbb{G}'_n = n^{-1/2}(\mathbb{N}_n - \mathbb{N}_n(1)P)$  converges in distribution to a tight limit. In view of the preceding discussion, this is equivalent to the processes  $Z_i = \sum_{j=1}^{Y_i} (\delta_{X_{i,j}} - P)$  satisfying the central limit theorem. Thus the sequence of expectations  $n^{-1/2}\mathbb{E}^*\|\sum_{i=1}^n Z_i\|_{\mathcal{F}}$  is bounded by Lemma 2.3.11. By Le Cam's lemma these expectations dominate the outer expectations  $n^{-1/2}\mathbb{E}^*\|\sum_{i=1}^n (\delta_{X_i} - P)\|_{\mathcal{F}}$  up to a positive constant and the argument can be finished as before. ■

## 3.6

# The Bootstrap

In this chapter we first prove the asymptotic consistency of the empirical bootstrap estimator of the distribution of the empirical process. Next this result is generalized to more general, “exchangeable” bootstrap schemes. The results of this chapter are particularly interesting when combined with those of Section 3.9.3, which show that the consistency of the bootstrap is retained under application of the delta-method.

### 3.6.1 The Empirical Bootstrap

Let  $\mathbb{P}_n$  be the empirical measure of an i.i.d. sample  $X_1, \dots, X_n$  from a probability measure  $P$ . Given the sample values, let  $\hat{X}_1, \dots, \hat{X}_n$  be an i.i.d. sample from  $\mathbb{P}_n$ . The *bootstrap empirical distribution* is the empirical measure  $\hat{\mathbb{P}}_n = n^{-1} \sum_{i=1}^n \delta_{\hat{X}_i}$ , and the *bootstrap empirical process*  $\hat{\mathbb{G}}_n$  is

$$\hat{\mathbb{G}}_n = \sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (M_{ni} - 1) \delta_{X_i},$$

where  $M_{ni}$  is the number of times that  $X_i$  is “redrawn” from the original sample. Instead of  $n$  we can also “redraw”  $k$  bootstrap values  $\hat{X}_1, \dots, \hat{X}_k$ . The corresponding bootstrap empirical process is

$$\hat{\mathbb{G}}_{n,k} = \sqrt{k}(\hat{\mathbb{P}}_k - \mathbb{P}_n) = \frac{1}{\sqrt{k}} \sum_{i=1}^n \left( M_{ki} - \frac{k}{n} \right) \delta_{X_i}.$$

A more precise specification of taking “bootstrap samples”  $\hat{X}_1, \dots, \hat{X}_k$  should include a probability space on which these variables are defined.

(This would be particularly appropriate since the results in this chapter are formulated in terms of outer measure.) Alternatively, the definitions of the bootstrap processes can be made rigorous by simply defining them as the far right side of the preceding displays. This is more convenient for the purpose of this chapter. In the resulting definition of  $\hat{G}_n$  and  $\hat{G}_{n,k}$ , it is further specified that the vector  $(M_{k1}, \dots, M_{kn})$  is independent of  $X_1, \dots, X_n$  and multinomially distributed with parameters  $k$  and (probabilities)  $1/n, \dots, 1/n$ . For computing outer expectations independence is understood in terms of a product probability space. Let  $X_1, X_2, \dots$  be the coordinate projections on the first  $\infty$  coordinates of the product space  $(\mathcal{X}^\infty, \mathcal{A}^\infty, P^\infty) \times (\mathcal{Z}, \mathcal{C}, Q)$  and let the multinomial vectors  $M$  depend on the last factor only.<sup>†</sup> The Poisson variables introduced later are also assumed to depend on  $\mathcal{Z}$  only.

Let  $\mathcal{F}$  be a class of measurable functions with a finite envelope function. Then  $\hat{G}_n$  considered as a process indexed by  $\mathcal{F}$  can be viewed as a map in  $\ell^\infty(\mathcal{F})$ . The main result of this section is that the sequence  $\hat{G}_n$  converges in  $\ell^\infty(\mathcal{F})$  conditionally in distribution to a tight Brownian bridge process given almost all sequences  $X_1, X_2, \dots$  if and only if  $\mathcal{F}$  is Donsker and  $P^* \|f - Pf\|_{\mathcal{F}}^2 < \infty$ . In that case the difference between the “conditional random law” of  $\sqrt{n}(\hat{P}_n - P_n)$  and the “law” of  $\sqrt{n}(P_n - P)$  converges to zero almost surely. In statistical terms this means that the conditional law of  $\sqrt{n}(\hat{P}_n - P_n)$  is an asymptotically consistent estimator of the law of  $\sqrt{n}(P_n - P)$ . This result is particularly interesting in combination with the delta-method, by which the consistency result carries over to a large class of statistics which are smooth functionals of the empirical distribution (see Section 3.9.3).

The result on almost-sure weak convergence of the bootstrap empirical process is reminiscent of the multiplier central limit theorem discussed in Chapter 2.9. The difference is that in the bootstrap situation the multipliers  $M_{n1} - 1, \dots, M_{nn} - 1$  are dependent. Our line of proof is to remove the dependence by Poissonization and to show that the Poissonized process and the ordinary bootstrap process are asymptotically equivalent.

Instead of  $n$ , take a Poisson number  $N_n$  of replicates, where  $N_n$  has mean  $n$  and is independent of the original observations. Then  $M_{N_n,1}, \dots, M_{N_n,n}$  are i.i.d. Poisson variables with mean 1. The corresponding bootstrap empirical process can be written

$$\hat{G}_{n,N_n} = \frac{1}{\sqrt{N_n}} \sum_{i=1}^n (M_{N_n,i} - 1) (\delta_{X_i} - P) - \frac{N_n - n}{\sqrt{N_n}} (P_n - P).$$

Conditionally on  $X_1, X_2, \dots$ , the second term converges in distribution to zero almost surely if  $\mathcal{F}$  is Glivenko-Cantelli. Furthermore, by the results of Chapter 2.9, the first term converges in distribution to a Brownian bridge process almost surely if and only if  $\mathcal{F}$  is Donsker and  $P^* \|f - Pf\|_{\mathcal{F}}^2 < \infty$ . It follows under these conditions that the Poissonized bootstrap process

---

<sup>†</sup> Given this set-up, Problem 3.6.1 gives a way of defining bootstrap values  $\hat{X}_i$  on  $\mathcal{X}^\infty \times \mathcal{Z}$  such that the identities for  $\hat{G}_n$  and  $\hat{G}_{n,k}$  are valid.

$\hat{G}_{n,N_n}$  converges to a Brownian bridge. In the proof ahead, it is shown that the difference in distribution between  $\hat{G}_{n,N_n}$  and  $\hat{G}_n = \hat{G}_{n,n}$  is very small.

As in Chapter 2.9, (conditional) weak convergence is formulated in terms of the bounded Lipschitz metric.

**3.6.1 Theorem.** Let  $\mathcal{F}$  be a class of measurable functions with finite envelope function. Define  $\hat{Y}_n = n^{-1/2} \sum_{i=1}^n (M_{N_n,i} - 1) (\delta_{X_i} - P)$ . The following statements are equivalent:

- (i)  $\mathcal{F}$  is Donsker;
- (ii)  $\sup_{h \in \text{BL}_1} |\mathbb{E}_{M,N} h(\hat{Y}_n) - \mathbb{E} h(\mathbb{G})| \xrightarrow{\text{P}*} 0$  and  $\hat{Y}_n$  is asymptotically measurable;
- (iii)  $\sup_{h \in \text{BL}_1} |\mathbb{E}_M h(\hat{G}_n) - \mathbb{E} h(\mathbb{G})| \xrightarrow{\text{P}*} 0$  and  $\hat{G}_n$  is asymptotically measurable.

**3.6.2 Theorem.** Let  $\mathcal{F}$  be a class of measurable functions with finite envelope function. Define  $\hat{Y}_n = n^{-1/2} \sum_{i=1}^n (M_{N_n,i} - 1) (\delta_{X_i} - P)$ . The following statements are equivalent:

- (i)  $\mathcal{F}$  is Donsker and  $P^* \|f - Pf\|_{\mathcal{F}}^2 < \infty$ ;
- (ii)  $\sup_{h \in \text{BL}_1} |\mathbb{E}_{M,N} h(\hat{Y}_n) - \mathbb{E} h(\mathbb{G})| \xrightarrow{\text{as}*} 0$  and the sequence  $\mathbb{E}_{M,N} h(\hat{Y}_n)^* - \mathbb{E}_{M,N} h(\hat{Y}_n)_*$  converges almost surely to zero for every  $h \in \text{BL}_1$ ;
- (iii)  $\sup_{h \in \text{BL}_1} |\mathbb{E}_M h(\hat{G}_n) - \mathbb{E} h(\mathbb{G})| \xrightarrow{\text{as}*} 0$  and the sequence  $\mathbb{E}_M h(\hat{G}_n)^* - \mathbb{E}_M h(\hat{G}_n)_*$  converges almost surely to zero for every  $h \in \text{BL}_1$ .

Here the asterisks denote the measurable cover functions with respect to  $M$ ,  $N$ , and  $X_1, X_2, \dots$  jointly.

**Proofs.** The equivalence of (i) and (ii) in both theorems is an immediate consequence of the multiplier central limit theorems, Theorems 2.9.6 and 2.9.7. The proof that (i)+(ii) is equivalent to (iii) is based on a coupling of the bootstrap empirical process  $\hat{G}_n$  and its (partly) Poissonized version  $\hat{Y}_n$  by a special construction of multinomial variables. Let  $m_n^{(1)}, m_n^{(2)}, \dots$  be i.i.d. multinomial( $1, n^{-1}, \dots, n^{-1}$ ) variables independent of  $N_n$  and set

$$M_n = \sum_{i=1}^n m_n^{(i)}; \quad M_{N_n} = \sum_{i=1}^{N_n} m_n^{(i)}.$$

Define  $\hat{G}_n$  using  $M_n$  and  $\hat{Y}_n$  using  $M_{N_n}$ . Note that  $\mathbb{E}_M h(\hat{G}_n)$  and  $\mathbb{E}_M h(\hat{G}_n)^*$  do not depend on the probability space on which  $M_n$  is defined (up to null sets). The first depends on the distribution of  $M_n$  only, in the second  $h(\hat{G}_n)^*$  is equal to the measurable majorant with respect to  $X_1, \dots, X_n$  only by the remarks following Lemma 1.2.7.

The absolute difference  $|M_{N_n} - M_n|$  is the sum of  $|N_n - n|$  of the variables  $m_n^{(i)}$ . Given  $N_n = k$ , its  $i$ -th component  $|M_{N_n,i} - M_{n,i}|$  is binomially ( $|k - n|, n^{-1}$ )-distributed. For any  $\varepsilon > 0$ , there exists a sequence of integers

$\ell_n$  with  $\ell_n = O(\sqrt{n})$  such that  $P(|N_n - n| \geq \ell_n) \leq \varepsilon$  for every  $n$ . By direct calculation,

$$P\left(\max_{1 \leq i \leq n} |M_{N_n,i} - M_{n,i}| > 2\right) \leq \varepsilon + n P\left(\text{bin}(\ell_n, n^{-1}) > 2\right) \rightarrow \varepsilon.$$

It follows that for sufficiently large  $n$  all coordinates of the vector  $|M_{N_n} - M_n|$  are 0, 1, or 2 with probability at least  $1 - 2\varepsilon$ . (The 2 is not important for the following: any bound would do.)

We can write  $|M_{Ni} - M_{ni}| = \sum_{j=1}^{\infty} 1\{|M_{Ni} - M_{ni}| \geq j\}$ . Alternatively, with  $I_n^j$  the set of indexes  $i \in \{1, 2, \dots, n\}$  such that  $|M_{Ni} - M_{ni}| \geq j$ , we have  $M_{Ni} - M_{ni} = \text{sign}(N - n) \sum_{j=1}^{\infty} 1\{i \in I_n^j\}$ . Thus,

$$\begin{aligned}\hat{Y}_n - \hat{G}_n &= \frac{1}{\sqrt{n}} \sum_{i=1}^n (M_{Ni} - M_{ni})(\delta_{X_i} - P) \\ &= \text{sign}(N - n) \sum_{j=1}^{\infty} \frac{\#I_n^j}{\sqrt{n}} \left( \frac{1}{\#I_n^j} \sum_{i \in I_n^j} (\delta_{X_i} - P) \right).\end{aligned}$$

On the set where  $\max_{1 \leq i \leq n} |M_{Ni} - M_{ni}| \leq 2$ , only the first two terms of the sum over  $j$  can be nonzero. For any  $j$  we have  $j(\#I_n^j) \leq |N_n - n| = O_P(\sqrt{n})$ . Furthermore, the norm of the average between brackets on the far right side converges to zero outer almost surely for any  $j$  if  $\mathcal{F}$  is a Glivenko-Cantelli class (cf. Lemma 3.6.16 ahead). Conclude that in that case

$$P_{M,N}(\|\hat{Y}_n - \hat{G}_n\|_{\mathcal{F}}^* > \varepsilon) \rightarrow 0,$$

as  $n \rightarrow \infty$ , given almost all sequences  $X_1, X_2, \dots$ , for every  $\varepsilon > 0$ . Consequently, the difference  $\sup_h |\mathbf{E}_M h(\hat{G}_n)^* - \mathbf{E}_{M,N} h(\hat{Y}_n)^*|$  converges to zero outer almost surely. The same is true for this expression with the asterisks removed or moved to the bottom.

It follows that (i)+(ii) and (iii) are equivalent (in both theorems) provided  $\mathcal{F}$  is Glivenko-Cantelli. If (i)+(ii) holds, then  $\mathcal{F}$  is Donsker and certainly Glivenko-Cantelli. Thus, the proof of the theorem in the most interesting direction is complete. For the proof in the converse direction, it must be shown that (iii) implies that  $\mathcal{F}$  is Glivenko-Cantelli.

It is no loss of generality to assume that  $Pf = 0$  for every  $f$  in  $\mathcal{F}$ . Suppose that (iii) holds in probability, and assume for the moment that  $P^* F^p < \infty$  for some  $p > 1$ . Under (iii) the sequence  $n^{-1/2} \|\hat{G}_n\|_{\mathcal{F}}$  converges (unconditionally) to zero in outer probability. An application of Hölder's inequality yields

$$\mathbf{E}^*\left(\frac{1}{\sqrt{n}} \|\hat{G}_n\|_{\mathcal{F}}\right)^p \leq \mathbf{E}^* \frac{1}{n} \sum_{i=1}^n (M_{ni} + 1)^p F^p(X_i) = \mathbf{E}(M_{n1} + 1)^p P^* F^p.$$

This is uniformly bounded in  $n$ . Conclude that  $n^{-1/2} \|\hat{G}_n\|_{\mathcal{F}}$  converges to zero in outer mean. This implies the same for  $n^{-1/2} \|\hat{G}_n\|_{\mathcal{F}}$  in view of the

following inequalities. Since each  $M_{ni}$  is binomially  $(n, n^{-1})$  distributed,

$$\begin{aligned} \left(1 - \frac{1}{n}\right)^n E^* \sqrt{n} \|\mathbb{G}_n\|_{\mathcal{F}} &= E_X^* \left\| \sum_{i=1}^n E_M 1_{M_{ni}=0} (\delta_{X_i} - P) \right\|_{\mathcal{F}}^* \\ &\leq E^* \left\| \sum_{i=1}^n 1_{M_{ni}=0} (\delta_{X_i} - P) \right\|_{\mathcal{F}}^* \\ &= E_M E_X^* \left\| \sum_{i=1}^n 1_{M_{ni}=0} (1 - M_{ni}) (\delta_{X_i} - P) \right\|_{\mathcal{F}}^* \\ &\leq E^* \left\| \sum_{i=1}^n (1 - M_{ni}) (\delta_{X_i} - P) \right\|_{\mathcal{F}}^* = E^* \sqrt{n} \|\hat{\mathbb{G}}_n\|_{\mathcal{F}}, \end{aligned}$$

where the last inequality follows from the inequality  $E^* \|U\| \leq E^* \|U + V\|$  for independent, zero-mean processes  $U$  and  $V$  (applied conditionally given  $M$ ; also cf. Lemma 1.2.7). Conclude that  $\mathcal{F}$  is Glivenko-Cantelli in outer mean. Still under the assumption that  $P^* F < \infty$ , a martingale argument shows that  $\mathcal{F}$  is also Glivenko-Cantelli almost surely (cf. Lemma 2.4.5 and the proof of Theorem 2.4.3.)

Finally, it is shown that  $F$  satisfies  $P^* F^p < \infty$ , for every  $p < 2$ , under (iii) of the first theorem. The proof is based on the representation  $\hat{\mathbb{G}}_n = n^{-1/2} \sum_{i=1}^n (\delta_{\hat{X}_i} - \mathbb{P}_n)$  of the bootstrap empirical process, where for each  $n$  given the original observations  $\hat{X}_1, \dots, \hat{X}_n$  is an i.i.d. sample from  $\mathbb{P}_n$ . To give a precise meaning to the (conditional) outer probabilities, these bootstrap values can be defined as in Problem 3.6.1. As before,  $P_M$  denotes conditional probability given the original observations. Under assumption (iii), the sequence  $P_M(\|\hat{\mathbb{G}}_n\|^* > m_n)$  converges in probability to zero for every  $m_n \rightarrow \infty$ . Let  $\hat{Y}_1, \dots, \hat{Y}_n$  be an independent copy of  $\hat{X}_1, \dots, \hat{X}_n$  given the original observations based on multinomial variables  $M$ . Adapt the proof of the Lévy inequality given by Proposition A.1.2 to obtain

$$\begin{aligned} P_{M, \tilde{M}} \left( \max_{1 \leq i \leq n} \|f(\hat{X}_i) - f(\hat{Y}_i)\|_{\mathcal{F}}^* > m_n \sqrt{n} \right) \\ \leq 2P_{M, \tilde{M}} \left( \left\| \sum (f(\hat{X}_i) - f(\hat{Y}_i)) \right\|_{\mathcal{F}}^* > m_n \sqrt{n} \right) \xrightarrow{P} 0. \end{aligned}$$

In view of Problem 2.3.2, it follows that

$$nP_{M, \tilde{M}} \left( \|f(\hat{X}_i) - f(\hat{Y}_i)\|_{\mathcal{F}}^* > m_n \sqrt{n} \right) \xrightarrow{P} 0.$$

Since the processes  $f(\hat{X}_i)$  and  $f(\hat{Y}_i)$  are (conditionally) independent, the left side is an upper bound for

$$nP_M \left( \|f(\hat{X}_i)\|_{\mathcal{F}}^* > 2m_n \sqrt{n} \right) P_{\tilde{M}} \left( \|f(\hat{Y}_i)\|_{\mathcal{F}}^* \leq m_n \sqrt{n} \right).$$

Here  $\|f(\hat{X}_i)\|_{\mathcal{F}} = \sum F(X_j)1\{m_n^{(i)} = e_j\}$ , so that we can conclude that

$$n P_M \left( \|f(\hat{X}_i)\|_{\mathcal{F}}^* > 2m_n \sqrt{n} \right) = \sum_{j=1}^n 1\{F^*(X_j) > 2m_n \sqrt{n}\} \xrightarrow{P} 0.$$

Since the sums are binomially distributed, it follows that  $n P(F^* > m_2 n \sqrt{n}) = O(1)$ . ■

The preceding proof does not apply to the bootstrapped empirical process  $\hat{\mathbb{G}}_{n,k}$  based on  $k$  (possibly unequal to  $n$ ) replicates of the original sample  $X_1, \dots, X_n$ . However, the most important part of the theorem is true for arbitrary bootstrap sample sizes. If  $\mathcal{F}$  is Donsker, then the “sequence”  $\hat{\mathbb{G}}_{n,k}$  converges conditionally in distribution to a Brownian bridge process for every possible manner in which both  $n, k \rightarrow \infty$ .

Let  $\mathcal{F}_\delta = \{f - g : f, g \in \mathcal{F}, \rho_P(f - g) < \delta\}$ .

**3.6.3 Theorem.** Let  $\mathcal{F}$  be a Donsker class of measurable functions such that  $\mathcal{F}_\delta$  is measurable for every  $\delta > 0$ . Then

$$\sup_{h \in \text{BL}_1} |E_M h(\hat{\mathbb{G}}_{n,k_n}) - Eh(\mathbb{G})| \xrightarrow{P^*} 0,$$

as  $n \rightarrow \infty$ , for any sequence  $k_n \rightarrow \infty$ . Furthermore, the sequence  $E_M h(\hat{\mathbb{G}}_{n,k_n})^* - E_M h(\hat{\mathbb{G}}_{n,k_n})_*$  converges to zero in probability for every  $h \in \text{BL}_1$ . If  $P^* \|f - Pf\|_{\mathcal{F}}^2 < \infty$ , then the convergence is also outer almost surely.

**Proof.** As in the proof of the conditional multiplier theorems, Theorems 2.9.6 and 2.9.7, it suffices to establish almost-sure convergence of all the finite-dimensional marginals plus conditional convergence to zero of the sequence  $\|\hat{\mathbb{G}}_{n,k}\|_{\mathcal{F}_\delta}$ . Without loss of generality, assume that  $Pf = 0$  for every  $f \in \mathcal{F}$ .

The distribution of  $\hat{\mathbb{G}}_{n,k} f$  equals that of  $k^{-1/2}$  times the centered sum of the i.i.d. random variables  $f(\hat{X}_1), \dots, f(\hat{X}_k)$ . Weak convergence of the sequence  $\hat{\mathbb{G}}_{n,k} f$  can be established by the Lindeberg-Lévy theorem. By the law of large numbers for real variables,

$$\begin{aligned} E_{\hat{X}} f^2(\hat{X}_1) &= \frac{1}{n} \sum_{i=1}^n f^2(X_i) \xrightarrow{\text{as}} Pf^2, \\ E_{\hat{X}} f^2(\hat{X}_1) \{ |f(\hat{X}_1)| > \varepsilon \sqrt{k} \} &= \frac{1}{n} \sum_{i=1}^n f^2(X_i) \{ |f(X_i)| > \varepsilon \sqrt{k} \} \xrightarrow{\text{as}} 0. \end{aligned}$$

Therefore,  $\hat{\mathbb{G}}_{n,k} f$  converges in distribution to a  $N(0, Pf^2)$  distribution for almost every sequence  $X_1, X_2, \dots$ . This concludes the proof of finite-dimensional convergence.

Let  $\tilde{N}_1, \tilde{N}_2, \dots$  be i.i.d. symmetrized Poisson variables with parameter  $\frac{1}{2}k/n$ . By Lemma 3.6.6 (ahead),

$$(3.6.4) \quad E_{\hat{X}} \|\hat{G}_{n,k}\|_{\mathcal{F}_\delta} \leq 4E_{\tilde{N}} \left\| \frac{1}{\sqrt{k}} \sum_{i=1}^n \tilde{N}_i \delta_{X_i} \right\|_{\mathcal{F}_\delta}.$$

By the multiplier inequality Lemma 2.9.1, the outer expectation of the left side is for any  $1 \leq n_0 \leq n$  bounded up to a constant by

$$(n_0 - 1) E \max_{1 \leq i \leq n} \frac{\tilde{N}_i}{\sqrt{k}} P^* F + \sqrt{\frac{n}{k}} \|\tilde{N}_i\|_{2,1} \max_{n_0 \leq j \leq n} E^* \left\| \frac{1}{\sqrt{j}} \sum_{i=n_0}^j \varepsilon_i \delta_{X_i} \right\|_{\mathcal{F}_\delta}.$$

In view of Problem 3.6.4, the first term is bounded by  $(n_0 - 1)2\sqrt{2}(n \wedge k)^{-1/4}P^*F$ . It converges to zero as  $n, k \rightarrow \infty$  for every fixed  $n_0 > 1$ . By Problem 3.6.3,  $\sqrt{n/k}\|\tilde{N}_i\|_{2,1} \leq 2\sqrt{2}$  for all  $n, k$ . Hence the second term converges to 0 as  $n, k \rightarrow \infty$  followed by  $n_0 \rightarrow \infty$  and  $\delta \downarrow 0$  in view of Theorem 2.3.11. It follows that  $E^* \|\hat{G}_{n,k}\|_{\mathcal{F}_\delta} \rightarrow 0$  as  $n, k \rightarrow \infty$  followed by  $\delta \downarrow 0$ . This concludes the proof of the first statement of the theorem.

Assume that  $P^*F^2 < \infty$ . If  $\varepsilon_1, \varepsilon_2, \dots$  are i.i.d. Rademacher variables independent of  $\tilde{N}_1, \tilde{N}_2, \dots$  and  $X_1, X_2, \dots$ , then  $\tilde{N}_i$  in (3.6.4) can be replaced by  $\varepsilon_i |\tilde{N}_i|$ . Next, Lemma 3.6.7 applied with  $Z_i = \varepsilon_i \delta_{X_i}$  and  $X_i$  fixed yields

$$\begin{aligned} E_{\hat{X}} \|\hat{G}_{n,k}\|_{\mathcal{F}_\delta}^* &\leq (n_0 - 1) E \max_{1 \leq i \leq n} \frac{|\tilde{N}_i|}{\sqrt{k}} E_\varepsilon \frac{1}{n} \sum_{i=1}^n \|\varepsilon_i \delta_{X_i}\|_{\mathcal{F}_\delta}^* \\ &\quad + \sqrt{\frac{n}{k}} \|\tilde{N}_i\|_{2,1} \max_{n_0 \leq j \leq n} E_{\varepsilon, R} \left\| \frac{1}{\sqrt{j}} \sum_{i=n_0}^j \varepsilon_i \delta_{X_{R_i}} \right\|_{\mathcal{F}_\delta}^*. \end{aligned}$$

The first term on the right is bounded by  $(n_0 - 1)2\sqrt{2}(n \wedge k)^{-1/4}2\mathbb{P}_n F$ . Since  $\mathcal{F}$  is Glivenko-Cantelli, the first term converges to zero almost surely for every fixed  $n_0$ . To conclude the proof, it suffices to show that

$$(3.6.5) \quad \max_{n_0 \leq j \leq n} E_{\varepsilon, R} \left\| \frac{1}{\sqrt{j}} \sum_{i=n_0}^j \varepsilon_i \delta_{X_{R_i}} \right\|_{\mathcal{F}_\delta}^* \xrightarrow{\text{as}} 0,$$

as  $n, k \rightarrow \infty$  followed by  $n_0 \rightarrow \infty$  and  $\delta \downarrow 0$ . By Jensen's inequality, the expression is bounded by

$$\max_{n_0 \leq j \leq n} E_{\varepsilon, R} \left\| \frac{1}{\sqrt{j}} \sum_{i=1}^j \varepsilon_i \delta_{X_{R_i}} \right\|_{\mathcal{F}_\delta}^* = \max_{n_0 \leq j \leq n} E(U_j | \Sigma_n) \leq E\left(\max_{n_0 \leq j} U_j | \Sigma_n\right),$$

where  $U_j = E_\varepsilon \left\| j^{-1/2} \sum_{i=1}^j \varepsilon_i \delta_{X_i} \right\|_{\mathcal{F}_\delta}^*$  and  $\Sigma_n$  is the  $\sigma$ -field generated by all functions  $f: \mathcal{X}^\infty \mapsto \mathbb{R}$  that are symmetric in their first  $n$  coordinates. By Corollary 2.9.9, the variable  $\max_j U_j$  is integrable. The sequence  $\Sigma_n$

decreases to the “symmetric”  $\sigma$ -field  $\Sigma$ , which consists of sets of probability 0 or 1 only by the Hewitt-Savage zero-one law.<sup>†</sup> It follows that, as  $n \rightarrow \infty$ ,

$$\mathbb{E} \left( \max_{n_0 \leq j} U_j \mid \Sigma_n \right) \xrightarrow{\text{as}} \mathbb{E} \left( \max_{n_0 \leq j} U_j \mid \Sigma \right) = \mathbb{E} \max_{n_0 \leq j} U_j.$$

By Lemma 2.9.8,  $\limsup_{j \rightarrow \infty} U_j \leq K \mathbb{E} \|G\|_{\mathcal{F}_\delta}$  almost surely. Hence  $\max_{n_0 \leq j} U_j \rightarrow 0$  almost surely as  $n_0 \rightarrow \infty$  followed by  $\delta \downarrow 0$ . Its expectation converges to zero by dominated convergence. This proves (3.6.5). ■

**3.6.6 Lemma.** *For fixed elements  $x_1, \dots, x_n$  of a set  $\mathcal{X}$ , let  $\hat{X}_1, \dots, \hat{X}_k$  be an i.i.d. sample from  $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{x_i}$ . Then*

$$\mathbb{E}_{\hat{X}} \left\| \sum_{j=1}^k (\delta_{\hat{X}_j} - \mathbb{P}_n) \right\|_{\mathcal{F}} \leq 4 \mathbb{E}_{N, N'} \left\| \sum_{i=1}^n (N_i - N'_i) \delta_{x_i} \right\|_{\mathcal{F}},$$

for every class  $\mathcal{F}$  of functions  $f: \mathcal{X} \mapsto \mathbb{R}$  and i.i.d. Poisson variables  $N_1, N'_1, \dots, N_n, N'_n$  with mean  $\frac{1}{2}k/n$ .

**Proof.** Let  $\varepsilon_j$  and  $\varepsilon_{i,j}$  be i.i.d. Rademacher variables, independent of  $\hat{X}_1, \dots, \hat{X}_k$ . By the symmetrization lemma, Lemma 2.3.1, and Le Cam’s Poissonization’s lemma, Lemma 3.5.4,

$$\mathbb{E}_{\hat{X}} \left\| \sum_{j=1}^k (\delta_{\hat{X}_j} - \mathbb{P}_n) \right\|_{\mathcal{F}} \leq 2 \mathbb{E}_{\hat{X}} \left\| \sum_{j=1}^k \varepsilon_j \delta_{\hat{X}_j} \right\|_{\mathcal{F}} \leq 4 \mathbb{E}_{N, \hat{X}} \left\| \sum_{j=1}^N \varepsilon_j \delta_{\hat{X}_j} \right\|_{\mathcal{F}},$$

where  $N$  is Poisson( $k$ )-distributed, independent of the i.i.d. sequence  $\hat{X}_1, \hat{X}_2, \dots$ . Set

$$N_i = \#\{j \leq N : \hat{X}_j = x_i, \varepsilon_j = 1\}; \quad N'_i = \#\{j \leq N : \hat{X}_j = x_i, \varepsilon_j = -1\}.$$

Given  $N = m$ , the vector  $(N_1, N'_1, \dots, N_n, N'_n)$  is multinomially distributed with parameters  $(m, 1/2n, \dots, 1/2n)$ . Hence, unconditionally the coordinates of this vector are i.i.d. Poisson-distributed with mean  $k/2n$ . The lemma follows since  $\sum_{j=1}^N \varepsilon_j \delta_{\hat{X}_j} = \sum_{i=1}^n (N_i - N'_i) \delta_{x_i}$ . ■

**3.6.7 Lemma (Multiplier inequalities).** *For arbitrary stochastic processes  $Z_1, \dots, Z_n$ , every exchangeable random vector  $(\xi_1, \dots, \xi_n)$  that is independent of  $Z_1, \dots, Z_n$ , and any  $1 \leq n_0 \leq n$ ,*

$$\begin{aligned} \mathbb{E}_{\xi} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i Z_i \right\|_{\mathcal{F}}^* &\leq 2(n_0 - 1) \frac{1}{n} \sum_{i=1}^n \|Z_i\|_{\mathcal{F}}^* \mathbb{E}_{\xi} \max_{1 \leq i \leq n} \frac{|\xi_i|}{\sqrt{n}} \\ &\quad + 2 \|\xi_1\|_{2,1} \max_{n_0 \leq k \leq n} \mathbb{E}_R \left\| \frac{1}{\sqrt{k}} \sum_{i=n_0}^k Z_{R_i} \right\|_{\mathcal{F}}^*. \end{aligned}$$

---

<sup>†</sup> Cf. Dudley (1993).

Here  $(R_1, \dots, R_n)$  is uniformly distributed on the set of all permutations of  $\{1, 2, \dots, n\}$  and independent of  $Z_1, \dots, Z_n$  and the asterisks denote measurable covers with respect to all variables jointly. Furthermore,

$$\begin{aligned} \mathbb{E} \left\| \frac{1}{\sqrt{n}} \sum_{i=1}^n \xi_i Z_i \right\|_{\mathcal{F}}^* &\leq 2(n_0 - 1) \frac{1}{n} \sum_{i=1}^n \mathbb{E} \|Z_i\|_{\mathcal{F}}^* \mathbb{E}_{\xi} \max_{1 \leq i \leq n} \frac{|\xi_i|}{\sqrt{n}} \\ &\quad + 2 \|\xi_1\|_{2,1} \max_{n_0 \leq k \leq n} \mathbb{E} \left\| \frac{1}{\sqrt{k}} \sum_{i=n_0}^k Z_{R_i} \right\|_{\mathcal{F}}^*. \end{aligned}$$

For nonnegative variables  $\xi_i$ , the two constants 2 on the right can be deleted.

**Proof.** Since the  $\xi_i$  can be split into their positive and negative parts, it suffices to consider the case that they are nonnegative.

Choose the vector  $R = (R_1, \dots, R_n)$  independent of  $\xi_1, \dots, \xi_n$ . By exchangeability of  $\xi_1, \dots, \xi_n$ , the left side times  $\sqrt{n}$  equals

$$\mathbb{E}_{\xi} \left\| \sum_{i=1}^n \xi_{R_i} Z_i \right\|_{\mathcal{F}}^* = \mathbb{E}_{\xi, R} \left\| \sum_{i=1}^n \xi_i Z_{R_i} \right\|_{\mathcal{F}}^* = \mathbb{E}_{\xi, R} \left\| \sum_{i=1}^n \xi_{(i)} Z_{R_{S_i}} \right\|_{\mathcal{F}}^*,$$

where  $\xi_{(1)} \geq \dots \geq \xi_{(n)}$  are the reverse order statistics of  $\xi_1, \dots, \xi_n$  and  $S = (S_1, \dots, S_n)$  is a (random) permutation such that  $\xi_{(i)} = \xi_{S_i}$ . To determine  $S$  uniquely also in the presence of ties, we can require that  $S_i < S_{i+1}$  whenever  $\xi_{S_i} = \xi_{S_{i+1}}$ . Then the vector  $R \circ S$  (with  $i$ th component  $R_{S_i}$ ) is distributed as  $R$  and is independent of  $S$  and  $\xi_1, \dots, \xi_n$ . By the triangle inequality, the right-hand side of the last display is bounded by

$$\mathbb{E}_{\xi, R} \left\| \sum_{i=1}^{n_0-1} \xi_{(i)} Z_{R_{S_i}} \right\|_{\mathcal{F}}^* + \mathbb{E}_{\xi, R} \left\| \sum_{j=n_0}^n \xi_{(j)} Z_{R_{S_j}} \right\|_{\mathcal{F}}^*.$$

The first term is bounded by

$$\mathbb{E}_{\xi, R} \max_{1 \leq i \leq n} \xi_i \sum_{i=1}^{n_0-1} \|Z_{R_{S_i}}\|_{\mathcal{F}}^* \leq \mathbb{E}_{\xi} \max_{1 \leq i \leq n} \xi_i \frac{n_0 - 1}{n} \sum_{i=1}^n \|Z_i\|_{\mathcal{F}}^*.$$

The second term can be bounded by the same method as in the proof of Lemma 2.9.1.

The second inequality is proved in the same manner. ■

### 3.6.2 The Exchangeable Bootstrap

For each  $n$ , let  $(W_{n1}, \dots, W_{nn})$  be an exchangeable, nonnegative random vector. Consider the *weighted bootstrap empirical measure*

$$\hat{\mathbb{P}}_n = \frac{1}{n} \sum_{i=1}^n W_{ni} \delta_{X_i}.$$

The corresponding weighted bootstrap empirical process is defined as

$$\hat{\mathbb{G}}_n = \sqrt{n}(\hat{\mathbb{P}}_n - \bar{W}_n \mathbb{P}_n) = \frac{1}{\sqrt{n}} \sum_{i=1}^n (W_{ni} - \bar{W}_n) \delta_{X_i} = \frac{1}{\sqrt{n}} \sum_{i=1}^n W_{ni} (\delta_{X_i} - \mathbb{P}_n).$$

This corresponds to resampling  $W_{ni}$  times the variable  $X_i$ . However, it is not required that the weights  $W_{ni}$  be integer-valued. In the case that in total  $n$  variables are resampled, the average  $\bar{W}_n$  is equal to one and the weighted bootstrap process is simply  $\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n)$ . It is assumed that

$$(3.6.8) \quad \begin{aligned} & \sup_n \|W_{n1} - \bar{W}_n\|_{2,1} < \infty, \\ & n^{-1/2} \mathbf{E} \max_{1 \leq i \leq n} |W_{ni} - \bar{W}_n| \xrightarrow{\text{P}} 0, \\ & n^{-1} \sum_{i=1}^n (W_{ni} - \bar{W}_n)^2 \xrightarrow{\text{P}} c^2 > 0. \end{aligned}$$

Sufficient for the second condition is that the variables  $W_{ni} - \bar{W}_n$  are uniformly weak- $L_2$  (Problem 2.3.3). Both the first and second conditions are valid if  $\|W_{n1}\|_{2+\varepsilon}$  is uniformly bounded for some  $\varepsilon > 0$ .

The multinomial weights considered in the preceding section satisfy these conditions. The following theorem shows that the conditions (3.6.8) alone imply the conditional weak convergence of the bootstrap process. This gives a large number of different ways to obtain bootstrap estimators for the distribution of the empirical process.

**3.6.9 Example.** If  $Y_1, \dots, Y_n$  are i.i.d. nonnegative random variables with  $\|Y_1\|_{2,1} < \infty$ , then the weights  $W_{ni} = Y_i/\bar{Y}_n$  satisfy (3.6.8) with  $c = \sigma(Y_1)/\mathbf{E}Y_1$ . If the variables  $Y_i$  are exponentially distributed with mean 1, the resulting scheme is known as the *Bayesian bootstrap*.

**3.6.10 Example.** Multinomial vectors  $(W_{n1}, \dots, W_{nn})$  with parameters  $n$  and (probabilities)  $(1/n, \dots, 1/n)$  satisfy (3.6.8) with  $c = 1$ .

**3.6.11 Example.** The vectors  $(W_{n1}, \dots, W_{nn})$  equal to  $\sqrt{n/k}$  times multinomial vectors with parameters  $k$  and (probabilities)  $(1/n, \dots, 1/n)$  satisfy (3.6.8) with  $c = 1$  provided  $k \rightarrow \infty$ . Thus, the theorem in this section contains the consistency of the empirical bootstrap based on  $k$  replicates.

**3.6.12 Example.** If  $W_{n1}, \dots, W_{nn}$  are i.i.d. vectors with finite  $L_{2,1}$ -norm, then the conditions (3.6.8) are satisfied with  $c^2 = \text{var } W_{1,1}$ . In this case the total mass  $\bar{W}_n$  of  $\hat{\mathbb{P}}_n$  is a random variable. This choice of weights  $\{W_{ni}\}$  has been called the *wild bootstrap* by some authors.

**3.6.13 Theorem.** Let  $\mathcal{F}$  be a Donsker class of measurable functions such that  $\mathcal{F}_\delta$  is measurable for every  $\delta > 0$ . For each  $n$  let  $(W_{n1}, \dots, W_{nn})$  be an exchangeable, nonnegative random vector independent of  $X_1, X_2, \dots$  such that the conditions (3.6.8) are satisfied. Then as  $n \rightarrow \infty$ ,

$$\sup_{h \in \text{BL}_1} |\mathbb{E}_W h(\hat{\mathbb{G}}_n) - \mathbb{E} h(c \mathbb{G})| \xrightarrow{P^*} 0.$$

Furthermore, the sequence  $\mathbb{E}_W h(\hat{\mathbb{G}}_n)^* - \mathbb{E}_W h(\hat{\mathbb{G}}_n)_*$  converges to zero in outer probability. If  $P^* \|f - Pf\|_{\mathcal{F}}^2 < \infty$ , then the convergence is also outer almost surely.

**Proof.** Without loss of generality, assume that  $\bar{W}_n = 0$  and that  $Pf = 0$  for every  $n$  and  $f$ . As before, it suffices to prove the conditional almost-sure weak convergence of every marginal plus the conditional asymptotic equicontinuity in probability or almost surely.

The variables  $W_{ni}$  satisfy the conditions of Lemma 3.6.15 (ahead). Thus this lemma with  $a_{ni} = f(X_i) - \mathbb{P}_n f$  implies that conditionally on the sequence  $X_1, X_2, \dots$  the sequence  $\hat{\mathbb{G}}_n f$  is asymptotically normal  $N(0, c^2 Pf^2)$  given every sequence  $X_1, X_2, \dots$  such that, as  $n \rightarrow \infty$  followed by  $M \rightarrow \infty$ ,

$$\mathbb{P}_n(f - \mathbb{P}_n f)^2 \rightarrow Pf^2 \quad \text{and} \quad \mathbb{P}_n(f - \mathbb{P}_n f)^2 \{ |f - \mathbb{P}_n f| > M \} \rightarrow 0.$$

Both are valid almost surely, since  $Pf^2 < \infty$ . Combined with the Cramér-Wold device, this establishes finite-dimensional convergence.

Let  $\mathcal{F}_\delta$  be the set of differences  $f - g$  of elements  $f, g \in \mathcal{F}$  with  $P(f - g)^2 < \delta^2$ , as usual. By the multiplier inequalities given by Lemma 3.6.7 (applied with  $Z_i = \delta_{X_i} - \mathbb{P}_n$  fixed), the conditional expectation  $\mathbb{E}_W \|\hat{\mathbb{G}}_n\|_{\mathcal{F}_\delta}$  equals

$$\begin{aligned} \mathbb{E}_W \left\| \frac{1}{\sqrt{n}} \sum W_{ni} (\delta_{X_i} - \mathbb{P}_n) \right\|_{\mathcal{F}_\delta}^* &\lesssim \frac{n_0 - 1}{n} \sum_{i=1}^n \|\delta_{X_i} - \mathbb{P}_n\|_{\mathcal{F}_\delta}^* \mathbb{E} \max_{1 \leq i \leq n} \frac{|W_{ni}|}{\sqrt{n}} \\ &+ \|W_{n1}\|_{2,1} \max_{n_0 \leq k \leq n} \mathbb{E}_R \left\| \frac{1}{\sqrt{n}} \sum_{i=n_0}^k (\delta_{X_{R_i}} - \mathbb{P}_n) \right\|_{\mathcal{F}_\delta}^*. \end{aligned}$$

The first term on the right converges to zero outer almost surely for every fixed  $n_0$ . In the second term the vector  $(X_{R_1}, \dots, X_{R_n})$  is a random sample without replacement from  $X_1, \dots, X_n$ . By Hoeffding's inequality (Lemma A.1.9), this term increases if the vector is replaced by a sample with replacement  $\hat{X}_1, \dots, \hat{X}_n$ . Next, the sum can be extended to the range from 1 to  $k$  by Jensen's inequality. Conclude that the second term is bounded by

$$2 \sup_n \|W_{n1}\|_{2,1} \max_{n_0 \leq k \leq n} \mathbb{E}_{\hat{X}} \|\hat{\mathbb{G}}_{n,k}\|_{\mathcal{F}_\delta}.$$

Here  $\hat{\mathbb{G}}_{n,k} = k^{-1/2} \sum_{i=1}^k (\delta_{\hat{X}_i} - \mathbb{P}_n)$  is the multinomial bootstrap process encountered in the previous section. By (the proof of) Theorem 3.6.3, this expression converges to zero in outer probability as  $n_0, n \rightarrow \infty$  followed by  $\delta \downarrow 0$ . Under the additional condition on the envelope function, the convergence is outer almost surely. ■

**3.6.14 Example (Bootstrap without replacement).** The bootstrap without replacement is based on resampling  $k < n$  observations from  $X_1, \dots, X_n$  without replacement. This can be incorporated in the scheme of the theorem by letting  $(W_{n1}, \dots, W_{nn})$  be a row of  $k$  times the number  $n(n-k)^{-1/2}k^{-1/2}$  and  $n-k$  times the number 0, ordered at random, independent of the  $X$ 's. Then the conditions (3.6.8) on the weights are satisfied for  $c = 1$ , provided both  $k \rightarrow \infty$  and  $n-k \rightarrow \infty$ . (For this choice the sample standard deviation is even identically equal to  $c = 1$ .)

In this case the assertion of the theorem can be phrased in terms of the empirical measures

$$\tilde{\mathbb{P}}_{k,n} = \frac{1}{k} \sum_{i=1}^k \delta_{X_{R_{ni}}}; \quad \tilde{\mathbb{Q}}_{n-k,n} = \frac{1}{n-k} \sum_{i=k+1}^n \delta_{X_{R_{ni}}},$$

where  $(R_{n1}, \dots, R_{nn})$  is a random permutation of the numbers  $1, 2, \dots, n$ . If both  $k \rightarrow \infty$  and  $n-k \rightarrow \infty$ , then the sequence

$$\sqrt{\frac{nk}{n-k}} (\tilde{\mathbb{P}}_{k,n} - \mathbb{P}_n) = \sqrt{\frac{k(n-k)}{n}} (\tilde{\mathbb{P}}_{k,n} - \tilde{\mathbb{Q}}_{n-k,n})$$

converges conditionally in distribution to a tight Brownian bridge.

**3.6.15 Lemma.** For each  $n$ , let  $(a_{n1}, \dots, a_{nn})$  and  $(W_{n1}, \dots, W_{nn})$  be a vector of numbers and an exchangeable random vector such that

$$\begin{aligned} \bar{a}_{ni} &= 0; \quad \frac{1}{n} \sum_{i=1}^n a_{ni}^2 \rightarrow \sigma^2 > 0; \quad \lim_{M \rightarrow \infty} \limsup_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n a_{ni}^2 \{ |a_{ni}| > M \} = 0; \\ \bar{W}_{ni} &= 0; \quad \frac{1}{n} \sum_{i=1}^n W_{ni}^2 \xrightarrow{\text{P}} \tau^2 > 0; \quad \frac{1}{n} \max_{1 \leq i \leq n} W_{ni}^2 \xrightarrow{\text{P}} 0. \end{aligned}$$

Then the sequence  $n^{-1/2} \sum_{i=1}^n a_{ni} W_{ni}$  converges weakly to a  $N(0, \sigma^2 \tau^2)$ -distribution.

**Proof.** Without loss of generality, assume that both  $n^{-1} \sum a_{ni}^2$  and  $n^{-1} \sum_{i=1}^n W_{ni}^2$  are equal to 1 for every  $n$ . Then

$$\sum_i \sum_j \frac{a_{ni}^2 W_{nj}^2}{n n} \{ |a_{ni} W_{nj}| > \varepsilon \sqrt{n} \} \leq \frac{1}{n} \sum_i a_{ni}^2 \{ |a_{ni}| \max_j |W_{nj}| > \varepsilon \sqrt{n} \}.$$

Since  $\max_j |W_{nj}|/\sqrt{n}$ , converges to zero in probability and the array  $a_{ni}$  is “uniformly integrable”, this expression converges to zero in probability as  $n \rightarrow \infty$ , for every  $\varepsilon > 0$ .

Combination with the assumptions shows that every subsequence of  $\{n\}$  has a further subsequence along which, for every  $\varepsilon > 0$ ,

$$\max_{1 \leq j \leq n} \frac{|W_{nj}|}{\sqrt{n}} \xrightarrow{\text{as}} 0; \quad \sum_i \sum_j \frac{a_{ni}^2 W_{nj}^2}{n n} \{ |a_{ni} W_{nj}| > \varepsilon \sqrt{n} \} \xrightarrow{\text{as}} 0.$$

This means that for almost every realization of the  $W_{nj}$ , the conditions of Hájek’s rank central limit theorem (Proposition A.5.3) are satisfied along the subsequence. Hence conditionally on the  $W_{nj}$ , By the subsequence of rank statistics  $n^{-1/2} \sum_{i=1}^n a_{ni} W_{n,R_{ni}}$  is asymptotically standard normally distributed.

This argument implies that every subsequence of  $\{n\}$  has a further subsequence along which

$$\sup_{h \in BL_1} \left| E_R h \left( \frac{1}{\sqrt{n}} \sum a_{ni} W_{n,R_{ni}} \right) - Eh(\xi) \right| \rightarrow 0,$$

almost surely, for a standard normal variable  $\xi$ . Conclude that this bounded Lipschitz distance converges to zero in probability along the whole sequence  $\{n\}$ . Take the expectation with respect to the  $W_{ni}$  to see that the sequence  $n^{-1/2} \sum a_{ni} W_{n,R_{ni}}$  is unconditionally asymptotically normal as well. By the exchangeability of the  $W_{nj}$ , this sequence is equal in distribution to the sequence  $n^{-1/2} \sum a_{ni} W_{ni}$ . ■

Glivenko-Cantelli theorem with exchangeable multipliers. A special case of the lemma is used in the proof of Theorem 3.6.2.

**3.6.16 Lemma.** *Let  $\mathcal{F}$  be a Glivenko-Cantelli class of measurable functions. For each  $n$ , let  $(W_{n1}, \dots, W_{nn})$  be an exchangeable nonnegative random vector independent of  $X_1, X_2, \dots$  such that  $\sum_{i=1}^n W_{ni} = 1$  and  $\max_{1 \leq i \leq n} |W_{ni}|$  converges to zero in probability. Then, for every  $\varepsilon > 0$ , as  $n \rightarrow \infty$ ,*

$$P_W \left( \left\| \sum_{i=1}^n W_{ni} (\delta_{X_i} - P) \right\|_{\mathcal{F}}^{*r} > \varepsilon \right) \xrightarrow{\text{as*}} 0.$$

**Proof.** By a modification of the multiplier inequality given by Lemma 3.6.7 (the  $L_r$  form with  $r < 1$  with Glivenko-Cantelli normalization instead of the  $L_1$  form with Donsker normalization in the bound of the second term), with  $Z_i = \delta_{X_i} - P$ ,

$$\begin{aligned} E_W \left\| \sum_{i=1}^n W_{ni} Z_i \right\|_{\mathcal{F}}^{*r} &\leq (n_0 - 1) E \max_{1 \leq i \leq n} W_{ni}^r \frac{1}{n} \sum_{j=1}^n \|Z_j\|_{\mathcal{F}}^{*r} \\ &\quad + (n E W_{n1})^r \max_{n_0 \leq k \leq n} E_R \left\| \frac{1}{k} \sum_{j=n_0}^k Z_{R_j} \right\|_{\mathcal{F}}^{*r}. \end{aligned}$$

In the first term on the right, the average  $n^{-1} \sum_{i=1}^n \|Z_i\|_{\mathcal{F}}^{*r}$  is bounded by  $\mathbb{P}_n F^{*r} + P^* F^r$ , which converges almost surely to  $2P^* F^r$ . Since the variables  $W_{ni}$  take their values in  $[0, 1]$ , the sequence  $\max_{1 \leq i \leq n} W_{ni}^r$  converges to zero in mean by the dominated convergence theorem. Thus, the first term on the right converges almost surely to zero for every fixed  $n_0$ .

The factor  $nEW_{n1}$  in the second term equals 1 by exchangeability and the fact that the sum of the  $W_{ni}$  is 1. By the triangle inequality, the second term is bounded by

$$\begin{aligned} 2 \max_{n_0-1 \leq k \leq n} \mathbb{E}_R \left\| \frac{1}{k} \sum_{j=1}^k Z_{R_j} \right\|_{\mathcal{F}}^{*r} &= 2 \max_{n_0-1 \leq k \leq n} \mathbb{E}(U_k^r | \Sigma_n) \\ &\leq 2 \mathbb{E} \left( \max_{n_0-1 \leq k} U_k^r | \Sigma_n \right), \end{aligned}$$

where  $U_k = \|k^{-1} \sum_{j=1}^k Z_j\|_{\mathcal{F}}^*$ , and  $\Sigma_n$  is the  $\sigma$ -field generated by all functions  $f: \mathcal{X}^\infty \mapsto \mathbb{R}$  that are symmetric in their first  $n$  coordinates. Since  $\mathcal{F}$  is Glivenko-Cantelli, we have  $\mathbb{E} \max_k U_k^r < \infty$  by Problem 2.3.6 (also see Corollary A.1.8) and the sequence  $U_k$  converges almost surely to zero. The sequence  $\Sigma_n$  decreases to the symmetric  $\sigma$ -field  $\Sigma$ , which consists of sets of probability 0 or 1 by the Hewitt-Savage zero-one law. It follows that as  $n \rightarrow \infty$

$$\mathbb{E} \left( \max_{n_0-1 \leq j} U_j^r | \Sigma_n \right) \xrightarrow{\text{as}} \mathbb{E} \left( \max_{n_0-1 \leq j} U_j^r | \Sigma \right) = \mathbb{E} \max_{n_0-1 \leq j} U_j^r.$$

The right side converges to zero as  $n_0 \rightarrow \infty$ . Conclude that the left side converges to zero almost surely when  $n \rightarrow \infty$  followed by  $n_0 \rightarrow \infty$ .

Apply Markov's inequality to complete the proof. ■

## Problems and Complements

- (Formally defining bootstrap samples)** Let  $m_n^{(1)}, m_n^{(2)}, \dots$  be i.i.d.  $n$ -dimensional multinomial  $(1, 1/n, \dots, 1/n)$  variables defined on some probability space  $(\mathcal{Z}, \mathcal{C}, Q)$ . Let  $X_1, X_2, \dots$  be the coordinate projections of  $(\mathcal{X}^\infty, \mathcal{A}^\infty, P^\infty)$ . Then bootstrap values can be defined on  $\mathcal{X}^\infty \times \mathcal{Z}$  by  $\hat{X}_i = X_j$  if  $m_n^{(i)}$  equals the  $j$ th unit vector in  $\mathbb{R}^n$ . Then

$$P^\infty \times Q((\hat{X}_1, \dots, \hat{X}_n) = (X_{j_1}, \dots, X_{j_n}) | X_1, \dots, X_n) = \left( \frac{1}{n} \right)^n$$

for every vector  $j \in \{1, 2, \dots, n\}^n$ , and taking conditional expectation given  $X_1, \dots, X_n$  is the same as taking expectation with respect to  $m_n^{(1)}, m_n^{(2)}, \dots$  only.

2. For each  $n$ , let  $Z_{n1}, \dots, Z_{nn}$  be i.i.d. stochastic processes. Suppose that  $\left\| n^{-1/2} \sum_{i=1}^n (Z_{ni} - E Z_{ni}) \right\|_{\mathcal{F}}^* = O_P(1)$  and  $P^*(\|Z_{n1}\|_{\mathcal{F}} > \sqrt{n}) \rightarrow 0$  as  $n \rightarrow \infty$ . Then  $\limsup_{n \rightarrow \infty} n P^*(\|Z_{n1}\|_{\mathcal{F}} > t\sqrt{n})$  converges to zero as  $t \rightarrow \infty$ .

[Hint: Let  $\tilde{Z}_{n1}, \dots, \tilde{Z}_{nn}$  be an i.i.d. copy of  $Z_{n1}, \dots, Z_{nn}$ . Fix  $0 < \varepsilon < 1/4$ . For any sufficiently large  $t$ ,

$$\limsup \left( \left\| \sum_{i=1}^n (Z_{ni} - \tilde{Z}_{ni}) \right\|_{\mathcal{F}} > t\sqrt{n} \right) < \varepsilon.$$

By a Lévy inequality followed by Problem 2.3.2, there exists an  $N_t$  such that, for all  $n \geq N_t$ ,

$$n P^*(\|Z_{n1} - \tilde{Z}_{n1}\|_{\mathcal{F}}^* > t\sqrt{n}) < 4\varepsilon.$$

Independent random processes  $Z$  and  $\tilde{Z}$  satisfy for every  $t$  the inequality  $P^*(\|Z - \tilde{Z}\|_{\mathcal{F}} > t) \geq P^*(\|Z\|_{\mathcal{F}} > 2t)P^*(\|\tilde{Z}\|_{\mathcal{F}} \leq t).$ ]

3. A symmetrized Poisson variable  $\tilde{N}$  with parameter  $\lambda$  satisfies  $\|\tilde{N}/\sqrt{\lambda}\|_{2,1} \leq 4$ .

[Hint:  $E\tilde{N}^4 \leq 12\lambda^2 + 2\lambda.$ ]

4. The maximum of i.i.d. symmetrized Poisson variables  $\tilde{N}_1, \dots, \tilde{N}_n$  with parameter  $\lambda$  satisfies  $E \max_{1 \leq i \leq n} |\tilde{N}_i| \leq n^{1/4} (12\lambda^2 + 2\lambda)^{1/4}$ .

[Hint: This bound is based on Lemma 2.2.2 with the  $L_4$ -norm and can be greatly improved.]

5. If  $\tilde{N}_\lambda$  is a symmetrized Poisson variable with parameter  $\lambda$ , then the family  $\{\tilde{N}_\lambda/\sqrt{\lambda}: \lambda > 0\}$  is not uniformly square integrable (even though the family is uniformly bounded in the  $L_{2,1}$ -norm).

6. It appears that for the consistency of the bootstrap without replacement with  $n - k = o(n)$ , the fact that  $\mathcal{F}$  is Donsker is not used as much in the proof, as one would expect.

## 3.7

# The Two-Sample Problem

Let  $X_1, \dots, X_m$  and  $Y_1, \dots, Y_n$  be independent random samples from distributions  $P$  and  $Q$  on a measurable space  $(\mathcal{X}, \mathcal{A})$ . We wish to test the null hypothesis  $H_0: P = Q$  versus the alternative  $H_1: P \neq Q$ .

One possible test statistic is the *two-sample Kolmogorov-Smirnov statistic*. Given a class  $\mathcal{F}$  of measurable functions  $f: \mathcal{X} \mapsto \mathbb{R}$ , set

$$D_{m,n} = \sqrt{\frac{mn}{m+n}} \|\mathbb{P}_m - \mathbb{Q}_n\|_{\mathcal{F}},$$

where  $\mathbb{P}_m$  and  $\mathbb{Q}_n$  are the empirical measures of the  $X$ 's and  $Y$ 's, respectively. For real-valued observations and  $\mathcal{F}$  equal to the set of indicators of cells  $(-\infty, t]$ , this reduces to the classical Kolmogorov-Smirnov statistic, which is the supremum distance between the two empirical distribution functions. If the underlying distributions  $P$  and  $Q$  are a priori assumed to be atomless, then the classical Kolmogorov-Smirnov statistic is distribution-free under the null hypothesis and both its finite sample and asymptotic null distribution are well known.

In contrast, for more general sample spaces and indexing classes  $\mathcal{F}$ , even the asymptotic null distribution of the “sequence”  $D_{m,n}$  may depend on the common underlying measure  $P = Q$ . In this chapter we discuss two ways of setting critical values for the test and show that the test is asymptotically consistent against every alternative  $(P, Q)$  with  $\|P - Q\|_{\mathcal{F}} > 0$ , if  $\mathcal{F}$  is a Donsker class with respect to both  $P$  and  $Q$ .

If  $\mathcal{F}$  is a Donsker class under both  $P$  and  $Q$ , then the sequences  $\mathbb{G}_m = \sqrt{m}(\mathbb{P}_m - P)$  and  $\mathbb{G}'_n = \sqrt{n}(\mathbb{Q}_n - Q)$  converge jointly in distribution to

independent Brownian bridges  $\mathbb{G}_P$  and  $\mathbb{G}_Q$ . Set  $N = m + n$  and note that

$$D_{m,n} = \left\| \sqrt{\frac{n}{N}} \mathbb{G}_m - \sqrt{\frac{m}{N}} \mathbb{G}'_n + \sqrt{\frac{mn}{N}} (P - Q) \right\|_{\mathcal{F}}.$$

If  $\|P - Q\|_{\mathcal{F}} > 0$ , then  $D_{m,n} \rightarrow \infty$  in probability as  $m, n \rightarrow \infty$ . On the other hand, under the null hypothesis  $P = Q$  the sequence  $D_{m,n}$  converges in distribution to the random variable  $\|\sqrt{1-\lambda} \mathbb{G}_P - \sqrt{\lambda} \mathbb{G}_Q\|_{\mathcal{F}}$  if  $m, n \rightarrow \infty$  such that  $m/N \rightarrow \lambda$ . For  $P = Q$ , the limit variable  $\sqrt{1-\lambda} \mathbb{G}_P - \sqrt{\lambda} \mathbb{G}_Q$  possesses the same distribution as  $\mathbb{G}_P$ . It follows that the test that rejects the null hypothesis if  $D_{m,n} > c_{m,n}$  has asymptotic level  $\alpha$  if the critical values  $c_{m,n}$  can be chosen such that

$$c_{m,n} \rightarrow c_P = \inf \left\{ t : P(\|\mathbb{G}_P\|_{\mathcal{F}} > t) \leq \alpha \right\}.$$

Since  $c_P < \infty$ , the test is then consistent in the sense that  $P(D_{m,n} > c_{m,n}) \rightarrow 1$  against any alternative with  $\|P - Q\|_{\mathcal{F}} > 0$ .

In general, the upper  $\alpha$ -quantile,  $c_P$ , of the limiting distribution depends on the underlying measure,  $P$ . Then it is impossible to find numbers  $c_{m,n}$  with the given convergence property for every  $P$  in the null hypothesis. Instead we use data-dependent “critical points”  $\tilde{c}_{m,n} = c_{m,n}(X_1, \dots, X_m, Y_1, \dots, Y_n)$ , and properly speaking the test statistic is  $D_{m,n} - \tilde{c}_{m,n}$  rather than  $D_{m,n}$ . Similar reasoning as before shows that the sequence of tests that reject the null hypothesis if  $D_{m,n} > \tilde{c}_{m,n}$  is asymptotically consistent and of level  $\alpha$  if the sequence  $\tilde{c}_{m,n}$  behaves well in probability. More precisely, the asymptotic level is  $\alpha$  in the sense that  $\limsup P(D_{m,n} > \tilde{c}_{m,n}) \leq \alpha$  for  $P = Q$  if<sup>b</sup>

$$\tilde{c}_{m,n} \xrightarrow{P} c_P.$$

Furthermore, the sequence of tests is consistent against any alternative  $(P, Q)$  for which  $\tilde{c}_{m,n}$  is bounded in probability and  $\|P - Q\|_{\mathcal{F}} > 0$ .

“Critical values”  $\tilde{c}_{m,n}$  with these properties can be determined by a permutation or a bootstrap approach. These procedures amount to sampling without and with replacement, respectively, from the pooled data  $(Z_{N1}, \dots, Z_{NN}) = (X_1, \dots, X_m, Y_1, \dots, Y_n)$ . Let  $\lambda_N = m/N$ , and consider the *pooled empirical measure*

$$\mathbb{H}_N = \frac{1}{N} \sum_{i=1}^N \delta_{Z_{Ni}} = \lambda_N \mathbb{P}_m + (1 - \lambda_N) \mathbb{Q}_n.$$

Since  $\mathbb{P}_m - \mathbb{H}_N = (1 - \lambda_N)(\mathbb{P}_m - \mathbb{Q}_n)$ , the Kolmogorov-Smirnov test is equivalent to a test based on the discrepancy between  $\mathbb{P}_m$  and the pooled empirical measure.

<sup>b</sup> Since  $H_0$  is composite, it is perhaps more proper to call the sequence of tests asymptotically of level  $\alpha$  if  $\sup_{P=Q} P(D_{m,n} > \tilde{c}_{m,n})$  is asymptotically bounded by  $\alpha$ . We do not consider this stronger concept. In view of the results of Chapter 2.8 on uniform in  $P$  Donsker classes, the present test is asymptotically of level  $\alpha$  also in this stronger sense for many Donsker classes  $\mathcal{F}$ .

### 3.7.1 Permutation Empirical Processes

Sampling without replacement from the pooled empirical measure  $\mathbb{H}_N$  can be represented in terms of a random permutation. Let the random vector  $R = (R_1, \dots, R_N)$  be uniformly distributed on the set of all permutations of  $\{1, 2, \dots, N\}$  and be independent from  $X_1, \dots, X_m, Y_1, \dots, Y_n$ . (As usual, when outer expectations are involved, “independence” is understood in terms of a product probability space.) The *two-sample permutation empirical measures* are

$$\tilde{\mathbb{P}}_{m,N} = \frac{1}{m} \sum_{i=1}^m \delta_{Z_{NR_i}}; \quad \tilde{\mathbb{Q}}_{n,N} = \frac{1}{n} \sum_{i=m+1}^N \delta_{Z_{NR_i}}.$$

The *permutation empirical process* corresponding to the first sample is  $\sqrt{m}(\tilde{\mathbb{P}}_{m,N} - \mathbb{H}_N)$ . The following theorems establish the almost-sure and in-probability limiting behavior of this process.

**3.7.1 Theorem.** *Let  $\mathcal{F}$  be a class of measurable functions that is Donsker under both  $P$  and  $Q$  and satisfies both  $\|P\|_{\mathcal{F}} < \infty$  and  $\|Q\|_{\mathcal{F}} < \infty$ . If  $m, n \rightarrow \infty$  such that  $m/N \rightarrow \lambda \in (0, 1)$ , then  $\sqrt{m}(\tilde{\mathbb{P}}_{m,N} - \mathbb{H}_N) \rightsquigarrow \sqrt{1-\lambda} \mathbb{G}_H$  given  $X_1, X_2, \dots, Y_1, Y_2, \dots$  in probability. Here  $\mathbb{G}_H$  is a tight Brownian bridge process corresponding to the measure  $H = \lambda P + (1-\lambda)Q$ .*

**3.7.2 Theorem.** *If in addition  $\mathcal{F}$  possesses an envelope function  $F$  with both  $P^*F^2 < \infty$  and  $Q^*F^2 < \infty$ , then also  $\sqrt{m}(\tilde{\mathbb{P}}_{m,N} - \mathbb{H}_N) \rightsquigarrow \sqrt{1-\lambda} \mathbb{G}_H$  given almost every sequence  $X_1, X_2, \dots, Y_1, Y_2, \dots$ .*

**Proof.** Without loss of generality, assume that  $Pf = 0$  for every  $f$ . Otherwise, replace  $f$  by  $f^\circ = f - Pf$  for every  $f$ , in which case we still have  $\|Q\|_{\mathcal{F}^\circ} < \infty$ . Given fixed values of the original pooled sample, let  $\hat{Z}_{N1}, \dots, \hat{Z}_{NN}$  be an i.i.d. sample from  $\mathbb{H}_N$ .

For marginal convergence it suffices by the Cramér-Wold device to show that, for every  $f$  with  $Hf = 0$  and  $Hf^2 < \infty$ ,

$$\frac{1}{\sqrt{m}} \sum_{i=1}^m (f(Z_{NR_i}) - \mathbb{H}_N f) \rightsquigarrow N(0, (1-\lambda) Hf^2),$$

given almost every sequence  $X_1, X_2, \dots, Y_1, Y_2, \dots$ . This follows from the central limit theorem for two-sample rank statistics, Proposition A.5.3, applied with  $a_{N,i} = f(Z_{Ni})$  and  $b_{N,i}$  equal to 1 or 0 if  $i \leq m$  or  $i > m$ .

Set  $\mathcal{F}_\delta = \{f - g, f, g \in \mathcal{F}, H(f - g)^2 < \delta^2\}$ . The assumptions imply that  $\mathcal{F}$  is totally bounded in both  $L_2(P)$  and  $L_2(Q)$ . Then  $\mathcal{F}$  is also totally bounded in  $L_2(H)$ . By Hoeffding’s inequality, Proposition A.1.9,

$$E_R \left\| \frac{1}{\sqrt{m}} \sum_{i=1}^m (\delta_{Z_{NR_i}} - \mathbb{H}_N) \right\|_{\mathcal{F}_\delta} \leq E_{\hat{Z}} \left\| \frac{1}{\sqrt{m}} \sum_{i=1}^m (\delta_{\hat{Z}_{Ni}} - \mathbb{H}_N) \right\|_{\mathcal{F}_\delta}.$$

Let  $\tilde{N}_1, \dots, \tilde{N}_N$  be i.i.d. symmetrized Poisson( $m/2N$ ) variables. By the Poissonization inequality, Lemma 3.6.6, the right side is bounded by 4 times

$$\mathbb{E}_{\tilde{N}} \left\| \frac{1}{\sqrt{m}} \sum_{i=1}^N \tilde{N}_i \delta_{Z_{N,i}} \right\|_{\mathcal{F}_\delta} \leq \mathbb{E}_{\tilde{N}} \left\| \frac{1}{\sqrt{m}} \sum_{i=1}^m \tilde{N}_i \delta_{X_i} \right\|_{\mathcal{F}_\delta} + \mathbb{E}_{\tilde{N}} \left\| \frac{1}{\sqrt{m}} \sum_{i=1}^n \tilde{N}_i \delta_{Y_i} \right\|_{\mathcal{F}_\delta}.$$

By the closedness of the symmetrized Poisson family under convolution and Jensen's inequality, the parameter of the Poisson variables can be increased to  $1/2$ . According to the unconditional multiplier theorem, Theorem 2.9.2, the sequences of processes  $m^{-1/2} \sum_{i=1}^m \tilde{N}_i \delta_{X_i}$ , and  $n^{-1/2} \sum_{i=1}^n \tilde{N}_i \delta_{Y_i}$  converge in distribution to tight Gaussian processes  $Z_P$  and  $Z_Q$ , respectively.

By Lemma 2.3.11, the outer expectation with respect to the original observations  $X_1, X_2, \dots, Y_1, Y_2, \dots$  of the right side of the last display is asymptotically bounded by a multiple of the expression  $\mathbb{E}\|Z_P\|_{\mathcal{F}_\delta} + \sqrt{(1-\lambda)/\lambda} \mathbb{E}\|Z_Q\|_{\mathcal{F}_\delta}$  as  $m, n \rightarrow \infty$ . This converges to 0 as  $\delta \downarrow 0$  since both processes have uniformly continuous sample paths with respect to the  $L_2(H)$ -semimetric. This concludes the proof of the first theorem.

If  $\mathcal{F}$  possesses an envelope function that is square integrable under both  $P$  and  $Q$ , then the limsup as  $m, n \rightarrow \infty$  of the last display is bounded by a constant times  $\mathbb{E}\|Z_P\|_{\mathcal{F}_\delta} + \sqrt{(1-\lambda)/\lambda} \mathbb{E}\|Z_Q\|_{\mathcal{F}_\delta}$ , for almost every  $X_1, X_2, \dots, Y_1, Y_2, \dots$ , by Corollary 2.9.8 combined with Lemma 2.3.11. Again, this upper bound converges to 0 as  $\delta \rightarrow 0$ . ■

The preceding theorem implies that the sequence

$$\tilde{D}_{m,n} = \sqrt{\frac{mn}{m+n}} \|\tilde{\mathbb{P}}_{m,N} - \tilde{\mathbb{Q}}_{n,N}\|_{\mathcal{F}} = \frac{1}{\sqrt{1-\lambda_N}} \sqrt{m} \|\tilde{\mathbb{P}}_{m,N} - \mathbb{H}_N\|_{\mathcal{F}}$$

converges in distribution to  $\|\mathbb{G}_H\|_{\mathcal{F}}$  given the original observations. Consequently, the upper  $\alpha$ -quantiles

$$(3.7.3) \quad \tilde{c}_{m,n} = \inf \left\{ t : P_R(\tilde{D}_{m,n} > t) \leq \alpha \right\}$$

of the conditional distribution of  $\tilde{D}_{m,n}$  can be used as “critical values” for the Kolmogorov-Smirnov test. The distribution of the norm  $\mathbb{G}_H$  of a tight Brownian bridge is known to be absolutely continuous with a positive density.<sup>#</sup> Thus, under the conditions of the preceding theorems,

$$\tilde{c}_{m,n} \rightarrow c_H = \inf \left\{ t : P(\|\mathbb{G}_H\|_{\mathcal{F}} > t) \leq \alpha \right\}$$

in probability and almost surely, respectively.

If follows that the sequence of tests that reject the null hypothesis  $H_0: P = Q$  if  $D_{m,n} > \tilde{c}_{m,n}$ , is asymptotically of level  $\alpha$  for any  $P = Q$  for which the index set  $\mathcal{F}$  is Donsker. Furthermore, the sequence of tests is consistent against any alternative  $(P, Q)$  for which  $\|P - Q\|_{\mathcal{F}} > 0$  and for

---

<sup>#</sup> See Beran and Millar (1986), Proposition 2, page 442.

which  $\mathcal{F}$  is both  $P$ - and  $Q$ -Donsker. In most cases there exist simple index classes  $\mathcal{F}$  that satisfy these requirements for every pair  $P = Q$  and  $P \neq Q$  on the sample space, respectively.

**3.7.4 Example.** In Euclidean space any suitably measurable, uniformly bounded VC-class of functions is universally Donsker. The collection of all indicators of cells  $(-\infty, t]$  can be added to ensure that  $\|P - Q\|_{\mathcal{F}} > 0$  for every  $P \neq Q$ . The resulting class satisfies the two requirements, and the sequence of Kolmogorov-Smirnov tests is asymptotically consistent against every possible alternative.

While every reasonable VC-class will thus ensure consistency against every possible alternative, the particular choice will determine the power of the resulting test. See Chapter 3.10 for a study of empirical processes under contiguous alternatives.

**3.7.5 Example.** If the  $\sigma$ -field in the sample space is countably generated, then there exists a sequence of measurable sets  $A_i$  such that  $P \neq Q$  if and only if  $P(A_i) \neq Q(A_i)$  for at least one  $i$ . (Replace a given countable generator by the collection of all finite intersections. This is a countable generator that is intersection-stable and hence a measure determining class.) By Theorem 2.13.1, the sequence  $\{c_i 1_{A_i}\}$  is a universal Donsker class for every sequence  $c_i$  that decreases to zero sufficiently fast.

Thus, any sample space with a countably generated  $\sigma$ -field supports a class  $\mathcal{F}$  for which the Kolmogorov-Smirnov test is asymptotically consistent against every possible alternative. (The Donsker class in the preceding paragraph proves existence but is not necessarily recommended to obtain power against “reasonable” alternatives.)

The test that rejects the null hypothesis if  $D_{m,n} > \tilde{c}_{m,n}$  with  $\tilde{c}_{m,n}$  determined by (3.7.3) is exactly a classical *permutation test*. There are  $N!$  possible orderings of the pooled sample  $X_1, \dots, X_m, Y_1, \dots, Y_n$  and, counting ties by multiplicity, equally many corresponding values of the test statistic  $D_{m,n}$ . Given the values in the pooled sample, each possible value of  $D_{m,n}$  is equally probable under the null hypothesis (still counting ties by multiplicity). The definition (3.7.3) of  $\tilde{c}_{m,n}$  is such that the null hypothesis is rejected if the observed value  $D_{m,n}(x_1, \dots, x_m, y_1, \dots, y_n)$  is among the  $\alpha$ -fraction of largest possible values of  $D_{m,n}$ .

It follows that given the values of the pooled sample the probability of an error of the first kind is smaller than  $\alpha$ . This being true for every  $m, n$  and every pooled sample also yields the conclusion that the sequence of tests is asymptotically of level  $\alpha$ . The abstract approach given previously shows in addition that the sequence of tests is consistent and that the critical values  $\tilde{c}_{m,n}$  converge (in probability or almost surely) to a deterministic value. It can also be used to study the power of the sequence of tests under contiguous alternatives.

### 3.7.2 Two-Sample Bootstrap

Instead of sampling values without replacement from the pooled sample, we can sample with replacement. This leads to *two-sample bootstrap empirical measures*

$$\hat{\mathbb{P}}_{m,N} = \frac{1}{m} \sum_{i=1}^m \delta_{\hat{Z}_{N,i}}; \quad \hat{\mathbb{Q}}_{n,N} = \frac{1}{n} \sum_{i=1}^n \delta_{\hat{Z}_{N,m+i}},$$

where  $\hat{Z}_{N1}, \dots, \hat{Z}_{NN}$  is an i.i.d. sample from the pooled empirical measure  $\mathbb{H}_N$ . Unlike the permutation empirical measures, the bootstrap empirical measures corresponding to the two samples are independent.

**3.7.6 Theorem.** *Let  $\mathcal{F}$  be a class of measurable functions that is Donsker under both  $P$  and  $Q$  and satisfies both  $\|P\|_{\mathcal{F}} < \infty$  and  $\|Q\|_{\mathcal{F}} < \infty$ . If  $m, n \rightarrow \infty$  such that  $m/N \rightarrow \lambda \in (0, 1)$ , then  $\sqrt{m}(\hat{\mathbb{P}}_{m,N} - \mathbb{H}_N) \rightsquigarrow \mathbb{G}_H$  given  $X_1, X_2, \dots, Y_1, Y_2, \dots$  in probability. Here  $\mathbb{G}_H$  is a tight Brownian bridge process corresponding to the measure  $H = \lambda P + (1 - \lambda)Q$ .*

**3.7.7 Theorem.** *If in addition  $\mathcal{F}$  possesses an envelope function  $F$  with both  $P^*F^2 < \infty$  and  $Q^*F^2 < \infty$ , then also  $\sqrt{m}(\hat{\mathbb{P}}_{m,N} - \mathbb{H}_N) \rightsquigarrow \mathbb{G}_H$  given almost every sequence  $X_1, X_2, \dots, Y_1, Y_2, \dots$ .*

The theorems are proved by the same arguments as used in the proofs of the permutation theorems. In fact, the proofs are simpler, because Hoeffding's inequality is no longer needed and the finite-dimensional convergence follows from the Lindeberg theorem.

The theorems combined with the analogous results for the bootstrap process corresponding to the second sample imply that the sequence

$$\begin{aligned} \hat{D}_{m,n} &= \sqrt{\frac{mn}{N}} \|\hat{\mathbb{P}}_{m,N} - \hat{\mathbb{Q}}_{n,N}\|_{\mathcal{F}} \\ &= \|\sqrt{m(1 - \lambda_N)} (\hat{\mathbb{P}}_{m,N} - \mathbb{H}_N) - \sqrt{n\lambda_N} (\hat{\mathbb{Q}}_{n,N} - \mathbb{H}_N)\|_{\mathcal{F}} \end{aligned}$$

converges in distribution to  $\|\sqrt{1 - \lambda} \mathbb{G}_H - \sqrt{\lambda} \mathbb{G}'_H\|_{\mathcal{F}}$  for independent Brownian bridges  $\mathbb{G}_H$  and  $\mathbb{G}'_H$ . The process  $\sqrt{1 - \lambda} \mathbb{G}_H - \sqrt{\lambda} \mathbb{G}'_H$  is a version of an  $H$ -Brownian bridge itself. Consequently, the upper  $\alpha$ -quantiles

$$\hat{c}_{m,n} = \inf \left\{ t: \mathbb{P}_{\hat{\mathcal{Z}}}(\hat{D}_{m,n} > t) \leq \alpha \right\}$$

of the conditional distribution of  $\hat{D}_{m,n}$  can be used as “critical values” for the Kolmogorov-Smirnov test. Under the conditions of the preceding theorems,

$$\hat{c}_{m,n} \rightarrow c_H = \inf \left\{ t: \mathbb{P}(\|\mathbb{G}_H\|_{\mathcal{F}} > t) \leq \alpha \right\},$$

in probability and almost surely, respectively. The critical values set by the permutation method in the preceding section possess exactly the same behavior. Hence at this level of analysis the two methods are of equal accuracy.

## Problems and Complements

- If  $\mathcal{F}$  is  $P$ -Donsker and  $\mathbb{P}_m$  and  $\mathbb{Q}_n$  are the empirical measures of independent i.i.d. samples of sizes  $m$  and  $n$  from  $P$ , then  $\sqrt{mn/N}(\mathbb{P}_m - \mathbb{Q}_n)$  converges in distribution to a tight Brownian bridge if  $m, n \rightarrow \infty$  (also if  $m/n \rightarrow 0$  or does not converge).
- (The  $k$ -sample problem)** For each  $j = 1, \dots, k$ , let  $X_{j1}, \dots, X_{jn_j}$  be an i.i.d. sample from a probability measure  $P_j$ . Consider testing the null hypothesis  $H_0: P_1 = \dots = P_k$ . Let  $\mathbb{P}_{j,n_j}$  be the empirical measure of the  $j$ th sample, and let  $\mathbb{H}_n$  be the empirical measure of the pooled sample. If  $\mathcal{F}$  is a  $P$ -Donsker class and  $\|P\|_{\mathcal{F}} < \infty$  for every  $P$ , then under the null hypothesis

$$K = \sum_{j=1}^k n_j (\mathbb{P}_{j,n_j} - \mathbb{H}_N)^2 \sim \sum_{j=1}^{k-1} \mathbb{G}_j^2,$$

in  $\ell^\infty(\mathcal{F})$ , where  $\mathbb{G}_1, \dots, \mathbb{G}_{k-1}$  are i.i.d. tight  $P$ -Brownian bridge processes. Critical values for a test based on  $\|K\|_{\mathcal{F}}$  can be set both by a permutation and a bootstrap procedure.

- In the setting of the preceding problem, consider tests based on the statistics defined as  $\sum_{i \neq j} \sqrt{n_i n_j / N} \|\mathbb{P}_{i,n_i} - \mathbb{P}_{j,n_j}\|_{\mathcal{F}}$  and  $\sup_{i \neq j} \sqrt{n_i n_j / N} \|\mathbb{P}_{i,n_i} - \mathbb{P}_{j,n_j}\|_{\mathcal{F}}$ .
- (Deficient two-sample bootstrap)** Let  $\hat{\mathbb{P}}_m$  and  $\hat{\mathbb{Q}}_n$  be the bootstrap empirical measures based on independent samples of sizes  $m$  and  $n$  from  $\mathbb{P}_m$  and  $\mathbb{Q}_n$ , respectively. If  $\mathcal{F}$  is  $P$ -Donsker, then the sequence  $\hat{D}_{m,n} = \sqrt{mn/N} \|\hat{\mathbb{P}}_m - \hat{\mathbb{Q}}_n\|_{\mathcal{F}}$  converges under the null hypothesis  $P = Q$  in distribution to  $\|\mathbb{G}_P\|_{\mathcal{F}}$  conditionally on the original observations (in probability or almost surely). Thus the upper  $\alpha$ -quantile of the distribution of  $\hat{D}_{m,n}$  can be used as critical values for the test based on  $D_{m,n}$ . Is the resulting test consistent for every  $\|P - Q\|_{\mathcal{F}} > 0$ ?

[**Hint:** The variables  $\hat{D}_{m,n}$  defined this way mimic the variables  $\hat{D}_{m,n}$  under the alternative hypothesis as well as under the null hypothesis in the sense that  $D_{m,n}/\sqrt{mn/N} \xrightarrow{P} \|P - Q\|_{\mathcal{F}}$  and  $\hat{D}_{m,n}/\sqrt{mn/N} \xrightarrow{P} \|P - Q\|_{\mathcal{F}}$ . Hence  $\hat{c}_\alpha \xrightarrow{P} \infty$ .]

## 3.8

# Independence Empirical Processes

Let  $H$  be a probability measure on the measurable space  $(\mathcal{X} \times \mathcal{Y}, \mathcal{A} \times \mathcal{B})$  with marginal laws  $P$  and  $Q$  on  $(\mathcal{X}, \mathcal{A})$  and  $(\mathcal{Y}, \mathcal{B})$ , respectively. Given a sample  $(X_1, Y_1), \dots, (X_n, Y_n)$  of independently and identically distributed vectors from  $H$ , we want to test the null hypothesis of independence  $H_0: H = P \times Q$  versus the alternative hypothesis  $H_1: H \neq P \times Q$ . Let  $\mathbb{H}_n$  be the empirical measure of the observations, and let  $\mathbb{P}_n$  and  $\mathbb{Q}_n$  be its marginals. The latter are the empirical measures of the  $X_i$ 's and  $Y_i$ 's, respectively.

Given classes  $\mathcal{F}$  and  $\mathcal{G}$  of real-valued measurable functions defined on  $\mathcal{X}$  and  $\mathcal{Y}$ , respectively, let  $\mathcal{F} \times \mathcal{G}$  be the class of all functions  $f \times g$  from  $\mathcal{X} \times \mathcal{Y}$  to  $\mathbb{R}$  defined as

$$(f \times g)(x, y) = f(x)g(y),$$

when  $f$  and  $g$  range over  $\mathcal{F}$  and  $\mathcal{G}$ , respectively. Consider the test statistics

$$K_n = \sqrt{n} \|\mathbb{H}_n - \mathbb{P}_n \times \mathbb{Q}_n\|_{\mathcal{F} \times \mathcal{G}}.$$

The *independence empirical process*  $\mathbb{Z}_n$  is the process indexed by  $\mathcal{F} \times \mathcal{G}$  given by

$$\begin{aligned} \mathbb{Z}_n(f, g) &= \sqrt{n} [(\mathbb{H}_n - \mathbb{P}_n \times \mathbb{Q}_n)(f \times g) - (H - P \times Q)(f \times g)] \\ &= \sqrt{n}(\mathbb{H}_n - H)(f \times g) - \sqrt{n}(\mathbb{P}_n - P)f \mathbb{Q}_n g - Pf \sqrt{n}(\mathbb{Q}_n - Q)g \\ &= \sqrt{n}(\mathbb{H}_n - H)((f - Pf) \times (g - Qg)) - \sqrt{n}(\mathbb{P}_n - P)f (\mathbb{Q}_n - Q)g. \end{aligned}$$

Under the null hypothesis  $H = P \times Q$  the test statistic can be written as  $K_n = \|\mathbb{Z}_n\|_{\mathcal{F} \times \mathcal{G}}$ , while under the alternative hypothesis  $K_n = \|\mathbb{Z}_n + \sqrt{n}(H - P \times Q)\|_{\mathcal{F} \times \mathcal{G}}$ .

The second term in the third representation of  $\mathbb{Z}_n$  is asymptotically negligible. The following result follows from Slutsky's lemma.

**3.8.1 Theorem.** *Let  $\mathcal{F}$  and  $\mathcal{G}$  be classes of measurable functions on measurable spaces  $(\mathcal{X}, \mathcal{A})$  and  $(\mathcal{Y}, \mathcal{B})$ , respectively. If  $\mathcal{F} \times \mathcal{G}$ ,  $\mathcal{F}$ , and  $\mathcal{G}$  are  $H$ -Donsker and  $\|P\|_{\mathcal{F}} < \infty$  and  $\|Q\|_{\mathcal{G}} < \infty$ , then the sequence of independence processes  $\mathbb{Z}_n$  converges in distribution in  $\ell^\infty(\mathcal{F} \times \mathcal{G})$  to the Gaussian process  $\mathbb{Z}_H(f \times g) = \mathbb{G}_H((f - Pf) \times (g - Pg))$  for a tight  $H$ -Brownian bridge  $\mathbb{G}_H$ .*

Under the null hypothesis  $H = P \times Q$ , the covariance function of the limit process can be calculated as

$$\text{cov}(\mathbb{Z}_H(f_1, g_1), \mathbb{Z}_H(f_2, g_2)) = (Pf_1 f_2 - Pf_1 Pf_2)(Qg_1 g_2 - Qg_1 Qg_2).$$

Thus, the limiting process has the same mean and covariance function as the product  $\mathbb{G}_P(f)\mathbb{G}_Q(g)$  of two independent Brownian bridges (which is not Gaussian).

**3.8.2 Example (Completely tucked Brownian sheet).** When  $\mathcal{X} \times \mathcal{Y}$  is the unit square in the Euclidean plane with uniform measure and both  $\mathcal{F}$  and  $\mathcal{G}$  are the indicators of the cells  $\{[0, t]: 0 \leq t \leq 1\}$ , the limit process is the *completely tucked Brownian sheet*. The class  $\mathcal{F} \times \mathcal{G}$  is the set of all indicators of cells  $[0, t]$  in the unit square, and the limit process can be identified with a process  $\mathbb{Z}(s, t)$  indexed by the unit square. This is a zero-mean Gaussian process with covariance function

$$E\mathbb{Z}(s_1, t_1)\mathbb{Z}(s_2, t_2) = (s_1 \wedge s_2 - s_1 s_2)(t_1 \wedge t_2 - t_1 t_2).$$

A “completely tucked Brownian sheet” derives its name from the fact that  $\mathbb{Z}(s, t) = 0$  almost surely for all  $(s, t)$  in the boundary of the unit square. The distribution of the norm  $\|\mathbb{Z}\|_{[0,1]^2}$  appears to be unknown in closed form.

The same limiting distribution for the sequence  $K_n$  arises when  $\mathcal{X} \times \mathcal{Y} = \mathbb{R}^2$  and  $\mathcal{F}$  and  $\mathcal{G}$  are equal to the class of indicators of cells  $\{(-\infty, t]: t \in \mathbb{R}\}$ . Under the null hypothesis of independence, the weak limit of the sequence  $\mathbb{Z}_n$  can be represented as  $\mathbb{Z}(P(s), Q(t))$ , where  $\mathbb{Z}$  is a standard, completely tucked Brownian sheet. Under the assumption that the marginal distribution functions  $P(s)$  and  $Q(t)$  are continuous, the variable  $\|\mathbb{Z}(P(s), Q(t))\|_{\mathbb{R}^2}$  is distributed as the norm of a completely tucked Brownian sheet.

The conclusion of the preceding theorem immediately yields that under the null hypothesis the sequence of test statistics  $K_n = \sqrt{n}\|\mathbb{H}_n - \mathbb{P}_n \times \mathbb{Q}_n\|_{\mathcal{F} \times \mathcal{G}}$  converges in distribution to the norm  $\|\mathbb{Z}_H\|_{\mathcal{F} \times \mathcal{G}}$  of the limit independence process. Furthermore, since the test statistics can be written

$$K_n = \|\mathbb{Z}_n + \sqrt{n}(H - P \times Q)\|_{\mathcal{F} \times \mathcal{G}},$$

the sequence of test statistics converges in probability to  $+\infty$  under every  $H$  in the alternative hypothesis such that  $\|H - P \times Q\|_{\mathcal{F} \times \mathcal{G}} > 0$ .

To implement a test based on  $K_n$ , it remains to approximate the critical points of its distribution. Since the limit distribution of the sequence  $K_n$  depends on the unknown marginal distributions  $P$  and  $Q$  even under the null hypothesis, it is typically not practical to obtain critical points from the asymptotic distribution. However, critical points can be set by a bootstrap approach.

Under the null hypothesis, a natural estimate for the underlying distribution  $H$  of the observations is the product  $\mathbb{P}_n \times \mathbb{Q}_n$  of the empirical measures of the two samples. For the bootstrap test of independence, let  $(\hat{X}_1, \hat{Y}_1), \dots, (\hat{X}_n, \hat{Y}_n)$  be an i.i.d. sample from the measure  $\mathbb{P}_n \times \mathbb{Q}_n$ , and let  $\hat{\mathbb{H}}_n = n^{-1} \sum_{i=1}^n \delta_{(\hat{X}_i, \hat{Y}_i)}$  be the corresponding bootstrap empirical measure. The *bootstrap independence process* is defined by

$$\hat{Z}_n = \sqrt{n}(\hat{\mathbb{H}}_n - \hat{\mathbb{P}}_n \times \hat{\mathbb{Q}}_n).$$

Note that, given the original observations, the variable  $\hat{Z}_n(f \times g)$  is centered at mean zero since the bootstrap sample is taken from  $\mathbb{P}_n \times \mathbb{Q}_n$ , not  $\mathbb{H}_n$ . Thus, there is a symmetry in the definitions of  $\hat{Z}_n$  and  $Z_n$  only if the latter is considered under the null hypothesis  $H = P \times Q$ . We shall still be interested in the asymptotic behavior of  $\hat{Z}_n$  conditionally on the original observations, under both null and alternative hypotheses.

With  $\hat{K}_n = \|\hat{Z}_n\|_{\mathcal{F} \times \mathcal{G}}$ , define a critical point

$$\hat{c}_n = \inf\{t > 0 : \mathbb{P}_{\hat{X}, \hat{Y}}(\hat{K}_n > t) \leq \alpha\}.$$

Here  $\mathbb{P}_{\hat{X}, \hat{Y}}$  denotes the conditional law given the original observations  $(X_1, Y_1), \dots, (X_n, Y_n)$ . The bootstrap test of independence rejects the null hypothesis for values of  $K_n$  exceeding  $\hat{c}_n$ . This procedure is a *model-based* bootstrap in that the bootstrap resampling is done from the estimated model under the null hypothesis.

To derive the asymptotic properties of this test, the convergence of the process  $\hat{Z}_n$  must be established under both the null and alternative hypotheses. Although theorems as general as those of Chapter 3.7 are lacking for this process, a number of conclusions are possible using the methods developed in Sections 2.8 and 2.9. These methods apply to a wide range of problems involving model-based bootstrapping.

We apply the results of Section 2.8.3 with  $P_n$  of that section taken to be  $\mathbb{P}_n \times \mathbb{Q}_n$  for given values of the original observations.

**3.8.3 Theorem.** *Let  $\mathcal{F}$  and  $\mathcal{G}$  be separable classes of measurable functions on measurable spaces  $(\mathcal{X}, \mathcal{A})$  and  $(\mathcal{Y}, \mathcal{B})$ , respectively, such that  $\mathcal{F} \times \mathcal{G}$  satisfies the uniform entropy condition for envelope functions  $F$ ,  $G$ , and  $F \times G$  that are  $H$ -square integrable. Then*

$$\hat{\mathbb{G}}_n = \sqrt{n}(\hat{\mathbb{H}}_n - \mathbb{P}_n \times \mathbb{Q}_n) \rightsquigarrow \mathbb{G}_{P \times Q}, \quad \text{in } \ell^\infty(\mathcal{F} \times \mathcal{G}),$$

and

$$\hat{Z}_n = \sqrt{n}(\hat{\mathbb{H}}_n - \hat{\mathbb{P}}_n \times \hat{\mathbb{Q}}_n) \rightsquigarrow \mathbb{Z}_{P \times Q}, \quad \text{in } \ell^\infty(\mathcal{F} \times \mathcal{G}),$$

given  $H^\infty$ -almost every sequence  $(X_1, Y_1), (X_2, Y_2), \dots$ . Here  $\mathbb{G}_{P \times Q}$  is a  $P \times Q$ -Brownian bridge.

**Proof.** The bootstrap independence process can be rewritten as

$$\begin{aligned} \hat{Z}_n = & \sqrt{n}(\hat{\mathbb{H}}_n - \mathbb{P}_n \times \mathbb{Q}_n)(f \times g) - \sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n)f \mathbb{Q}_n g \\ & - \mathbb{P}_n f \sqrt{n}(\hat{\mathbb{Q}}_n - \mathbb{Q}_n)g - \sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n)f(\hat{\mathbb{Q}}_n - \mathbb{Q}_n)g. \end{aligned}$$

The assertions now follow from an empirical central limit theorem for a triangular array of observations. In the present case the  $n$ th row of the array is the sample  $(\hat{X}_1, \hat{Y}_1), \dots, (\hat{X}_n, \hat{Y}_n)$  from  $\mathbb{P}_n \times \mathbb{Q}_n$ . Since  $\mathcal{F} \times \mathcal{G}$  satisfies the uniform entropy condition, Theorem 2.8.9 shows that  $\hat{\mathbb{G}}_n$  converges in distribution to  $\mathbb{G}_{P \times Q}$  for every sequence of original observations for which

$$\begin{aligned} & (\mathbb{P}_n \times \mathbb{Q}_n)(F \times G)^2 = O(1), \\ & (\mathbb{P}_n \times \mathbb{Q}_n)(F \times G)^2 \{|F \times G| \geq \varepsilon \sqrt{n}\} \rightarrow 0, \quad \text{every } \varepsilon > 0, \\ & \sup_{h_1, h_2 \in \mathcal{F} \times \mathcal{G}} |\rho_{\mathbb{P}_n \times \mathbb{Q}_n}(h_1, h_2) - \rho_{P \times Q}(h_1, h_2)| \rightarrow 0. \end{aligned}$$

The quantities in this display are two-sample U-statistics (for dependent samples if  $H$  is not a product measure), but in view of the special structure of the kernels can be handled by the strong law of large numbers for i.i.d. variables. For instance, the second line of the display can be written as

$$\frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n F^2(X_i) G^2(Y_j) \{|F(X_i)G(Y_j)| \geq \varepsilon \sqrt{n}\}.$$

For every  $M^2 \leq \varepsilon \sqrt{n}$  this is bounded above by

$$\mathbb{P}_n F^2 \{F \geq M\} \mathbb{Q}_n G^2 + \mathbb{P}_n F^2 \mathbb{Q}_n G^2 \{G \geq M\}.$$

For  $n \rightarrow \infty$  this converges almost surely to a fixed value, which can be made arbitrarily small by choosing  $M$  large. This concludes the proof of the weak convergence of  $\hat{\mathbb{G}}_n$ .

By Slutsky's lemma and the continuous mapping theorem, the sequence  $\hat{Z}_n$  converges in distribution to  $\mathbb{G}_{P \times Q}(f \times g) - \mathbb{G}_{P \times Q}(f \times 1)Qg - Pf\mathbb{G}_{P \times Q}(1 \times g)$ , which is a mean zero Gaussian process with the same covariance function as  $\mathbb{Z}_{P \times Q}$  (Cf. the second representation of  $\mathbb{Z}_n$  in the introduction). ■

The preceding theorem implies that the sequence  $\hat{K}_n = \|\hat{Z}_n\|_{\mathcal{F} \times \mathcal{G}}$  converges almost surely conditionally in distribution to  $\|\mathbb{Z}_{P \times Q}\|_{\mathcal{F} \times \mathcal{G}}$  under

both null and alternative hypotheses. Thus, it has the same limiting distribution as the test statistic  $K_n$  under the null hypothesis  $H = P \times Q$ . Hence the upper  $\alpha$ -quantiles  $\hat{c}_n$  of its (conditional) distribution satisfy

$$\hat{c}_n \rightarrow \inf \{t > 0 : P(\|\mathbb{Z}_{P \times Q}\|_{\mathcal{F} \times \mathcal{G}} > t) \leq \alpha\}, \quad \text{a.s.}$$

It follows that the sequence of tests that reject the null hypothesis  $H_0: H = P \times Q$  if  $K_n > \hat{c}_n$  is asymptotically of level  $\alpha$  for any  $H = P \times Q$  for which the index set  $\mathcal{F} \times \mathcal{G}$  satisfies the conditions of the preceding theorem. Furthermore, the sequence of tests is consistent against any alternative  $H$  for which  $\|H - P \times Q\|_{\mathcal{F} \times \mathcal{G}} > 0$ .

An alternative to using the bootstrap is to set critical points by a permutation procedure. For a permutation test of independence let  $R = (R_1, \dots, R_n)$  be uniformly distributed on the set of permutations of  $\{1, \dots, n\}$  and independent of the original observations. Then  $\tilde{\mathbb{H}}_n = n^{-1} \sum_{i=1}^n \delta_{(X_i, Y_{R_i})}$  is the permutation empirical measure, and the process

$$\tilde{\mathbb{Z}}_n = \sqrt{n}(\tilde{\mathbb{H}}_n - \mathbb{P}_n \times \mathbb{Q}_n)$$

is called the *permutation independence process*. Now with  $\tilde{K}_n \equiv \|\tilde{\mathbb{Z}}_n\|_{\mathcal{F} \times \mathcal{G}}$  define as critical point

$$\tilde{c}_n = \inf \{t > 0 : P_R(\tilde{K}_n > t) < \alpha\}.$$

The permutation test of independence rejects the null hypothesis for values of  $\tilde{K}_n$  larger than  $\tilde{c}_n$ . A proof of the consistency of this procedure appears to be unavailable at this point, although it is likely that the permutation independence process has asymptotic behavior similar to the bootstrap process considered previously.

## Problems and Complements

- Suppose that  $\mathcal{F} \times \mathcal{G}$  is  $H$ -Donsker where  $H$  is a distribution on  $\mathcal{X} \times \mathcal{Y}$ . Then the finite-dimensional distributions of the permutation independence process  $\tilde{\mathbb{Z}}_n$  converge in distribution to those of  $\mathbb{Z}_{P \times Q}$ , given  $H^\infty$ -almost every sequence  $(X_1, Y_1), (X_2, Y_2), \dots$

[Hint: Use the rank central limit theorem, Proposition A.5.3.]

## 3.9

# The Delta-Method

After giving the general principle of the delta-method, we consider the special case of Gaussian limits and the “conditional” delta-method, which applies to the bootstrap. The chapter closes with a large number of examples.

### 3.9.1 Main Result

Suppose  $X_n$  is a sequence of real-valued maps with  $\sqrt{n}(X_n - \theta) \rightsquigarrow X$  for some constant  $\theta$ . If  $\phi: \mathbb{R} \mapsto \mathbb{R}$  is differentiable at  $\theta$ , then by Slutsky’s theorem

$$\sqrt{n}(\phi(X_n) - \phi(\theta)) = \frac{\phi(X_n) - \phi(\theta)}{X_n - \theta} \sqrt{n}(X_n - \theta) \rightsquigarrow \phi'(\theta) X.$$

This is the simplest form of the *delta-method*.

A more general form of the delta-method is valid for maps  $\phi: \mathbb{D} \mapsto \mathbb{E}$  between normed spaces  $\mathbb{D}$  and  $\mathbb{E}$  or, even more generally: metrizable, topological vector spaces.<sup>†</sup> The appropriate form of differentiability of  $\phi$  is Hadamard differentiability. A map  $\phi: \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$  is called *Hadamard-differentiable* at  $\theta \in \mathbb{D}_\phi$  if there is a continuous linear map  $\phi'_\theta: \mathbb{D} \mapsto \mathbb{E}$  such that

$$\frac{\phi(\theta + t_n h_n) - \phi(\theta)}{t_n} \rightarrow \phi'_\theta(h), \quad n \rightarrow \infty,$$

---

<sup>†</sup> A *topological vector space* is a vector space for which addition and scalar multiplication are continuous operations: if  $x_n \rightarrow x$ ,  $y_n \rightarrow y$ , and  $c_n \rightarrow c$ , then  $x_n + y_n \rightarrow x + y$  and  $c_n x_n \rightarrow cx$ . Every normed space is a topological vector space; another example is  $\mathbb{R}^\infty$  with product topology.

for all converging sequences  $t_n \rightarrow 0$  and  $h_n \rightarrow h$  such that  $\theta + t_n h_n \in \mathbb{D}_\phi$  for every  $n$ . An equivalent definition is that (for normed spaces  $\mathbb{E}$ )

$$(3.9.1) \quad \sup_{h \in K, \theta + th \in \mathbb{D}_\phi} \left\| \frac{\phi(\theta + th) - \phi(\theta)}{t} - \phi'_\theta(h) \right\| \rightarrow 0, \quad t \rightarrow 0,$$

for every compact  $K \subset \mathbb{D}$ . The first definition may be refined to *Hadamard-differentiable tangentially* to a set  $\mathbb{D}_0 \subset \mathbb{D}$  by requiring that every  $h_n \rightarrow h$  in the definition has  $h \in \mathbb{D}_0$ ; the derivative need then be defined on  $\mathbb{D}_0$  only. The second definition more or less shows why Hadamard differentiability is the appropriate type of differentiability for our purposes: the compacts  $K$  match up with the asymptotic tightness of weakly convergent sequences.

The domain  $\mathbb{D}_\phi$  is allowed to be an arbitrary subset of  $\mathbb{D}$ . In particular, it is not necessary that it is all of  $\mathbb{D}$  and it need not be an open subset. This is often important for applications, for instance when considering functions of distribution functions viewed as element of a Skorohod space.

**3.9.2 Example (Hadamard and Fréchet differentiability).** For maps  $\phi: \mathbb{D}_\phi \subset \mathbb{R}^k \mapsto \mathbb{R}^m$ , Hadamard differentiability is equivalent to the usual differentiability of calculus, and the derivative  $\phi'_\theta: \mathbb{R}^k \mapsto \mathbb{R}^m$  is given by the matrix of partial derivatives  $(\partial \phi_i / \partial \theta_j)$  evaluated at  $\theta$ . In general, a map  $\phi: \mathbb{D}_\phi \mapsto \mathbb{E}$  is called *Fréchet-differentiable* if there exists a continuous, linear map  $\phi'_\theta: \mathbb{D} \mapsto \mathbb{E}$  such that (3.9.1) holds uniformly in  $h$  in bounded subsets of  $\mathbb{D}_\phi$ . For normed spaces, this is equivalent to

$$\|\phi(\theta + h) - \phi(\theta) - \phi'_\theta(h)\| = o(\|h\|), \quad \|h\| \rightarrow 0.$$

Since a compact set is bounded, Fréchet differentiability is a stronger requirement than Hadamard differentiability. For normed spaces, they are equivalent if and only if the unit ball is compact, that is,  $\mathbb{D}$  is finite-dimensional.

The chain rule asserts that the composition of two differentiable maps is differentiable. It makes possible a calculus of differentiable maps, in which complicated maps can be decomposed into more basic maps. We shall use it frequently when dealing with examples.

**3.9.3 Lemma (Chain rule).** If  $\phi: \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}_\psi$  is Hadamard-differentiable at  $\theta \in \mathbb{D}_\phi$  tangentially to  $\mathbb{D}_0$  and  $\psi: \mathbb{E}_\psi \mapsto \mathbb{F}$  is Hadamard-differentiable at  $\phi(\theta)$  tangentially to  $\phi'_\theta(\mathbb{D}_0)$ , then  $\psi \circ \phi: \mathbb{D}_\phi \mapsto \mathbb{F}$  is Hadamard-differentiable at  $\theta$  tangentially to  $\mathbb{D}_0$  with derivative  $\psi'_{\phi(\theta)} \circ \phi'_\theta$ .

**Proof.** Rewrite the difference  $\psi \circ \phi(\theta + th_t) - \psi \circ \phi(\theta)$  as  $\psi(\phi(\theta) + tk_t) - \psi(\phi(\theta))$ , with  $k_t$  given by  $(\phi(\theta + th_t) - \phi(\theta))/t$ . First apply Hadamard differentiability of  $\phi$  to see that  $k_t$  converges to  $\phi'_\theta(h)$ , and next Hadamard differentiability of  $\psi$ . ■

For measurable maps  $X_n$ , the delta-method is an almost immediate consequence of the almost-sure representation theorem. If  $r_n \rightarrow \infty$ ,  $r_n(X_n - \theta) \rightsquigarrow X$  in  $\mathbb{D}$ , and every  $X_n$  is Borel measurable, then there are Borel measurable  $\tilde{X}_n$  that are equal in law to  $X_n$ , take their values in the range of  $X_n$ , and satisfy  $r_n(\tilde{X}_n - \theta) \xrightarrow{\text{as}} \tilde{X}$ . (Apply the representation theorem to  $r_n(X_n - \theta)$  and transform.) If  $\phi$  is Hadamard-differentiable at  $\theta$ , then

$$r_n(\phi(\tilde{X}_n) - \phi(\theta)) = \frac{\phi(\theta + r_n^{-1}r_n(\tilde{X}_n - \theta)) - \phi(\theta)}{r_n^{-1}} \xrightarrow{\text{as}} \phi'_\theta(\tilde{X}).$$

If  $\phi$  is measurable the left side is equal in distribution to  $r_n(\phi(X_n) - \phi(\theta))$ . Moreover, the almost sure convergence implies convergence in law. Thus, we obtain that the sequence  $r_n(\phi(X_n) - \phi(\theta))$  converges in distribution to  $\phi'_\theta(\tilde{X})$ .

Unfortunately, this argument breaks down if the measurability assumptions fail. It can be fixed by using the (refined) continuous mapping theorem for outer almost-sure convergence, or by a direct argument. The latter approach appears preferable. In any case, the delta-method “works” without the need of explicit measurability assumptions.

**3.9.4 Theorem (Delta-method).** *Let  $\mathbb{D}$  and  $\mathbb{E}$  be metrizable topological vector spaces. Let  $\phi: \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$  be Hadamard-differentiable at  $\theta$  tangentially to  $\mathbb{D}_0$ . Let  $X_n: \Omega_n \mapsto \mathbb{D}_\phi$  be maps with  $r_n(X_n - \theta) \rightsquigarrow X$  for some sequence of constants  $r_n \rightarrow \infty$ , where  $X$  is separable and takes its values in  $\mathbb{D}_0$ . Then  $r_n(\phi(X_n) - \phi(\theta)) \rightsquigarrow \phi'_\theta(X)$ . If  $\phi'_\theta$  is defined and continuous on the whole of  $\mathbb{D}$ , then the sequence  $r_n(\phi(X_n) - \phi(\theta)) - \phi'_\theta(r_n(X_n - \theta))$  converges to zero in outer probability.*

**Proof.** The map  $g_n(h) = r_n(\phi(\theta + r_n^{-1}h) - \phi(\theta))$  is defined on the domain  $\mathbb{D}_n = \{h: \theta + r_n^{-1}h \in \mathbb{D}_\phi\}$  and satisfies  $g_n(h_n) \rightarrow \phi'_\theta(h)$  for every  $h_n \rightarrow h \in \mathbb{D}_0$ . By the extended continuous mapping theorem, Theorem 1.11.1, one has  $g_n(r_n(X_n - \theta)) \rightsquigarrow \phi'_\theta(X)$ . This is the first assertion of the theorem.

The second assertion follows upon applying the first to the map  $\psi: \mathbb{D}_\phi \mapsto \mathbb{E} \times \mathbb{E}$  defined by  $\psi(d) = (\phi(d), \phi'_\theta(d))$ . Since this map is Hadamard-differentiable at  $(\theta, \theta)$  with derivative  $(\phi'_\theta, \phi''_\theta)$ , it follows that

$$(r_n(\phi(X_n) - \phi(\theta)), r_n(\phi'_\theta(X_n) - \phi'_\theta(\theta))) \rightsquigarrow (\phi'_\theta(X), \phi''_\theta(X)).$$

By the continuous mapping theorem, the difference of the coordinates of these vectors converges weakly to  $\phi''_\theta(X) - \phi'_\theta(X) = 0$ . ■

A remarkable fact about the previous theorem is that it suffices that  $\phi$  is differentiable at just the single point  $\theta$ . This is certainly convenient for applications to abstract-valued random elements, such as the empirical process, when continuous differentiability can easily fail.

A stronger form of differentiability may be needed for a “uniform” delta-method. Suppose  $\theta_n \rightarrow \theta$  and  $r_n(X_n - \theta_n) \rightsquigarrow X$ . Under what condition does it follow that  $r_n(\phi(X_n) - \phi(\theta_n)) \rightsquigarrow \phi'_\theta(X)$ ? Consider two cases.

First, if  $\theta_n$  converges sufficiently fast to  $\theta$ —specifically if the sequence  $r_n(\theta_n - \theta)$  is relatively compact—then the conditions of the previous theorem apply. Simply write the object of interest as the sum  $r_n(\phi(X_n) - \phi(\theta)) + r_n(\phi(\theta) - \phi(\theta_n))$  and apply the previous theorem to the first term and the definition of Hadamard differentiability to the second. If  $r_n(\theta_n - \theta) \rightarrow h$  (along a subsequence), this yields  $r_n(\phi(X_n) - \phi(\theta_n)) \rightsquigarrow \phi'_\theta(X + h) + \phi'_\theta(-h) = \phi'_\theta(X)$ .

Second, for general  $\theta_n$ , it suffices that  $\phi: \mathbb{D}_\phi \mapsto \mathbb{E}$  be uniformly differentiable.

**3.9.5 Theorem (Delta-method).** *Let  $\mathbb{D}$  and  $\mathbb{E}$  be metrizable topological vector spaces and  $r_n$  constants with  $r_n \rightarrow \infty$ . Let  $\phi: \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$  satisfy*

$$r_n(\phi(\theta_n + r_n^{-1}h_n) - \phi(\theta_n)) \rightarrow \phi'_\theta(h),$$

for every converging sequence  $h_n$  with  $\theta_n + r_n^{-1}h_n \in \mathbb{D}_\phi$  for all  $n$  and  $h_n \rightarrow h \in \mathbb{D}_0$  and some arbitrary map  $\phi'_\theta$  on  $\mathbb{D}_0$ . If  $X_n: \Omega_n \mapsto \mathbb{D}_\phi$  are maps with  $r_n(X_n - \theta_n) \rightsquigarrow X$ , where  $X$  is separable and takes its values in  $\mathbb{D}_0$ , then  $r_n(\phi(X_n) - \phi(\theta_n)) \rightsquigarrow \phi'_\theta(X)$ . If  $\phi'_\theta$  is defined, linear, and continuous on the whole of  $\mathbb{D}$ , then the sequence  $r_n(\phi(X_n) - \phi(\theta_n)) - \phi'_\theta(r_n(X_n - \theta))$  converges to zero in outer probability.

This theorem is proved in exactly the same manner as the previous theorem. A sufficient condition for uniform differentiability is continuous differentiability in a neighborhood of  $\theta$ . For convex domains  $\mathbb{D}_\phi$ , a weak form of this is that  $\phi: \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$  be differentiable at every  $\vartheta \in \mathbb{D}_\phi$  that is sufficiently close to  $\theta$ , where the derivatives  $\phi'_\vartheta$  satisfy

$$(3.9.6) \quad \begin{aligned} \lim_{\substack{\eta \rightarrow \vartheta \\ \eta \in \mathbb{D}_\phi}} \phi'_\eta(h) &= \phi'_\vartheta(h), && \text{for every } h \in \text{lin } \mathbb{D}_\phi, \\ \lim_{\substack{\vartheta \rightarrow \theta \\ \vartheta \in \mathbb{D}_\phi}} \phi'_\vartheta(h) &= \phi'_\theta(h), && \text{uniformly in } h \in K, \end{aligned}$$

for every totally bounded subset  $K$  of  $\text{lin } \mathbb{D}_\phi$ .

**3.9.7 Lemma.** *Let  $\mathbb{D}$  and  $\mathbb{E}$  be normed spaces and  $\mathbb{D}_\phi \subset \mathbb{D}$  convex. Let  $\phi: \mathbb{D}_\phi \mapsto \mathbb{E}$  be Hadamard-differentiable at every  $\vartheta \in \mathbb{D}_\phi$  in a neighborhood around  $\theta$ , tangentially to  $\mathbb{D}_0$ , where the derivatives satisfy (3.9.6). Then  $\phi$  is uniformly differentiable along every sequence  $\theta_n \rightarrow \theta$  in  $\mathbb{D}_\phi$ , tangentially to  $\mathbb{D}_0$ .*

**Proof.** Let  $t_n \downarrow 0$  be given. Fix sufficiently large  $n$  and  $h$  such that  $\theta_n + t_n h \in \mathbb{D}_\phi$ , and define a map  $f_n(t) = \phi(\theta_n + th)$  on the interval  $0 \leq t \leq t_n$ .

Since  $\mathbb{D}_\phi$  is convex, this is well defined, and by Hadamard differentiability of  $\phi$ ,

$$\frac{f_n(t+g) - f_n(t)}{g} \rightarrow \phi'_{\theta_n + th}(h), \quad \text{as } g \rightarrow 0,$$

for every  $t$ . Thus  $f_n$  is differentiable on the interval  $[0, t_n]$ . By assumption its derivative is continuous in  $t$ . Hence for every element  $e^*$  from the dual space of  $\mathbb{E}$ , the fundamental theorem of calculus allows one to write  $e^*(f_n(t_n) - f_n(0))$  as the integral of its derivative. In other words

$$e^*(\phi(\theta_n + t_n h) - \phi(\theta_n)) = t_n \int_0^1 e^* \phi'_{\theta_n + t t_n h}(h) dt.$$

For every  $h_n \rightarrow h$  such that  $\theta_n + t_n h_n \in \mathbb{D}_\phi$ , we have by assumption that

$$\int_0^1 |e^* \phi'_{\theta_n + t t_n h_n}(h_n) - e^* \phi'_\theta(h_n)| dt \rightarrow 0,$$

uniformly in  $\|e^*\| \leq 1$ . The combination of the last two equations shows that the sequence  $t_n^{-1}(\phi(\theta_n + t_n h_n) - \phi(\theta_n))$  has the same limit as the sequence  $\phi'_\theta(h_n)$ . ■

### 3.9.2 Gaussian Limits

The delta-method yields the conclusion that the transformed sequence of variables  $r_n(\phi(X_n) - \phi(\theta))$  converges to a limit variable  $\phi'_\theta(X)$  for a certain continuous, linear map  $\phi'_\theta: \mathbb{D} \mapsto \mathbb{E}$ . In the case that  $\mathbb{D} = \mathbb{R}^k$  and  $\mathbb{E} = \mathbb{R}^m$ , the map  $\phi'_\theta$  is an  $(m \times k)$ -matrix. If the limit  $X$  of the original sequence is normally  $N_k(\mu, \Sigma)$ -distributed, then  $\phi'_\theta(X)$  is  $N_m(\phi'_\theta \mu, \phi'_\theta \Sigma (\phi'_\theta)^t)$ -distributed. In particular, the asymptotic normality is retained. In this section it is noted that the same conclusion is true in the infinite-dimensional situation.

By definition, a Borel measurable random element  $X$  in a Banach space  $\mathbb{D}$  is normally distributed (or Gaussian) if the real-valued random variable  $d^* X$  is normally distributed for every element  $d^*$  of the dual space (the collection of continuous, linear real maps on  $\mathbb{D}$ ). From this definition it is immediate that  $\phi'_\theta(X)$  is normally distributed in  $\mathbb{E}$  for every continuous, linear map  $\phi'_\theta: \mathbb{D} \mapsto \mathbb{E}$ , whenever  $X$  is normally distributed in  $\mathbb{D}$ . It suffices to note that  $e^* \circ \phi'_\theta$  is an element of the dual space of  $\mathbb{D}$  for every element  $e^*$  of the dual space of  $\mathbb{E}$ .

Many applications concern stochastic processes with bounded sample paths. By definition, a stochastic process  $\{X_t: t \in T\}$  is Gaussian if every one of its finite-dimensional marginals  $(X_{t_1}, \dots, X_{t_k})$  is multivariately normally distributed. If the stochastic process is also defined as a Borel measurable map into  $\ell^\infty(T)$ , then the definition of Gaussianity of a Banach-valued random element given in the preceding paragraph may be applied as well. Thus,  $X$  is Gaussian if  $d^* X$  is normally distributed for every element  $d^*$  of

the dual space of  $\ell^\infty(T)$ . It is not immediately clear that these two definitions are equivalent. The second definition appears to be more stringent, as the first definition requires that  $d^*X$  is normally distributed for every linear combination  $d^* = \sum \alpha_i \pi_{t_i}$  of coordinate projections, but not necessarily for every element of the dual space of  $\ell^\infty(T)$ . The following lemma shows that the two definitions are equivalent if  $X$  is tight.

**3.9.8 Lemma.** *Let  $X$  be a tight, Borel measurable map into  $\ell^\infty(T)$  such that the vector  $(X_{t_1}, \dots, X_{t_k})$  is multivariately normally distributed for every finite set  $t_1, \dots, t_k$  in  $T$ . Then  $\phi(X)$  is normally distributed for every continuous, linear map  $\phi: \ell^\infty(T) \mapsto \mathbb{E}$  into a Banach space  $\mathbb{E}$ .*

**Proof.** It suffices to consider the case that  $\mathbb{E}$  is the real line. By Lemma 1.5.9 there exists a semimetric  $\rho$  on  $T$  that makes  $T$  totally bounded and such that almost all sample paths  $t \mapsto X_t$  are uniformly continuous. Every bounded, uniformly continuous function  $z: T \mapsto \mathbb{R}$  has a unique extension  $\bar{z}$  to a continuous function on the completion  $\bar{T}$  of  $T$ . This completion is compact for (the extension of)  $\rho$ . By a minor modification of the Riesz representation theorem,<sup>†</sup> there exists a signed Borel measure  $\mu$  on the completion such that

$$\phi(z) = \int_{\bar{T}} \bar{z}(t) d\mu(t).$$

By discretization we can construct a sequence of maps of the form  $\phi_m = \sum \alpha_{mi} \pi_{t_{mi}}$ , with every  $t_{mi} \in T$ , that converges pointwise to  $\phi$  on  $UC(T, \rho)$ . In particular,  $\phi_m(X) \rightarrow \phi(X)$  almost surely. Since  $\phi_m(X)$  is normally distributed by assumption, so is  $\phi(X)$ . ■

It is clear from the preceding proof that in applications of the delta-method to maps  $\phi: \ell^\infty(T) \mapsto \mathbb{R}$ , the resulting limit variable  $\phi'_\theta(X)$  can be written as  $\int_{\bar{T}} X(t) d\mu(t)$ , whenever  $X$  is tight. The preceding lemma shows that  $\phi'_\theta(X)$  is normally distributed. With the help of Fubini's theorem, its mean and variance can be computed as

$$\int_{\bar{T}} \mathbb{E}\bar{X}(t) d\mu(t); \quad \int_{\bar{T}} \int_{\bar{T}} \mathbb{E}\bar{X}(s)\bar{X}(t) d\mu(s)d\mu(t),$$

respectively. Often mean and variance can be computed more easily from the asymptotic linear expansion guaranteed by the second assertion of Theorem 3.9.4.

### 3.9.3 The Delta-Method for the Bootstrap

The various bootstrap processes discussed in Chapter 3.6 were shown to converge conditionally in distribution given the “original” observations.

---

<sup>†</sup> E.g., Rudin (1966).

In this case, application of the delta-method should lead to a statement concerning the conditional weak convergence of the transformed processes. Proper care of measurability issues requires a separate discussion.

Consider sequences of random elements  $\mathbb{P}_n = \mathbb{P}_n(X_n)$  and  $\hat{\mathbb{P}}_n = \hat{\mathbb{P}}_n(X_n, M_n)$  in a normed space  $\mathbb{D}$  such that the sequence  $\sqrt{n}(\mathbb{P}_n - P)$  converges unconditionally and the sequence  $\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n)$  converges conditionally given  $X_n$  in distribution to a tight random element  $\mathbb{G}$ . A precise formulation of the second is that

$$(3.9.9) \quad \begin{aligned} & \sup_{h \in \text{BL}_1(\mathbb{D})} |\mathbb{E}_M h(\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n)) - \mathbb{E} h(\mathbb{G})| \rightarrow 0, \\ & \mathbb{E}_M h(\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n))^* - \mathbb{E}_M h(\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n))_* \rightarrow 0, \end{aligned}$$

in outer probability or outer almost surely, with  $h$  ranging over the bounded Lipschitz functions. Given a map  $\phi$ , we wish to show that

$$(3.9.10) \quad \begin{aligned} & \sup_{h \in \text{BL}_1(\mathbb{E})} \left| \mathbb{E}_M h\left(\sqrt{n}(\phi(\hat{\mathbb{P}}_n) - \phi(\mathbb{P}_n))\right) - \mathbb{E} h(\phi'_P(\mathbb{G})) \right| \rightarrow 0, \\ & \mathbb{E}_M h\left(\sqrt{n}(\phi(\hat{\mathbb{P}}_n) - \phi(\mathbb{P}_n))\right)^* - \mathbb{E}_M h\left(\sqrt{n}(\phi(\hat{\mathbb{P}}_n) - \phi(\mathbb{P}_n))\right)_* \rightarrow 0, \end{aligned}$$

in outer probability or outer almost surely.

For statistical purposes, consistency in probability appears to be sufficient. This is retained under just Hadamard differentiability at the single point  $P$ .

**3.9.11 Theorem (Delta-method for bootstrap in probability).** *Let  $\mathbb{D}$  and  $\mathbb{E}$  be normed spaces. Let  $\phi: \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$  be Hadamard-differentiable at  $P$  tangentially to a subspace  $\mathbb{D}_0$ . Let  $\mathbb{P}_n$  and  $\hat{\mathbb{P}}_n$  be maps as indicated previously with values in  $\mathbb{D}_\phi$  such that  $\sqrt{n}(\mathbb{P}_n - P) \rightsquigarrow \mathbb{G}$  and (3.9.9) holds in outer probability, where  $\mathbb{G}$  is separable and takes its values in  $\mathbb{D}_0$ . Then (3.9.10) holds in outer probability.*

Thus, the weak consistency of the bootstrap estimators considered in Chapter 3.6 carries over to any Hadamard-differentiable functional: the sequence of “conditional random laws” (given  $X_1, X_2, \dots$ ) of  $\sqrt{n}(\phi(\hat{\mathbb{P}}_n) - \phi(\mathbb{P}_n))$  is asymptotically consistent in probability for estimating the laws of the random elements  $\sqrt{n}(\phi(\mathbb{P}_n) - \phi(P))$ .

For almost-sure consistency, we need a stronger differentiability property of  $\phi$ . The next two results impose uniform Hadamard differentiability and Fréchet differentiability with a rate, respectively. First, suppose that

$$\sqrt{n}(\phi(\mathbb{P}_n + n^{-1/2}h_n) - \phi(\mathbb{P}_n)) \rightarrow \phi'_P(h)$$

for almost every sequence  $\mathbb{P}_n$  and every converging sequence  $h_n \rightarrow h \in \mathbb{D}_0$ . Then Theorem 3.9.5 yields

$$\sqrt{n}(\phi(\hat{\mathbb{P}}_n) - \phi(\mathbb{P}_n)) \rightsquigarrow \phi'_P(\mathbb{G}),$$

given almost every sequence  $\mathbb{P}_n$  for which  $\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n) \rightsquigarrow \mathbb{G}$ . Thus, without further effort we have obtained an almost-sure version of the preceding theorem. However, the result implies at best that the first expression in (3.9.10) converges almost surely to zero, not necessarily outer almost surely. (In particular, it does not imply the in-outer-probability statement of the preceding theorem.) In a given application the measurability of the transformed bootstrap process could be argued directly. Alternatively, outer-almost-sure convergence in (3.9.10) can be obtained directly under a stronger type of differentiability. Assume that

$$(3.9.12) \quad \frac{\phi(P_n + t_n h_n) - \phi(P_n)}{t_n} \rightarrow \phi'_P(h),$$

for all sequences  $t_n \downarrow 0$ ,  $P_n \rightarrow P$ ,  $h_n \rightarrow h \in \mathbb{D}_0$  such that  $P_n$  and  $P_n + t_n h_n \in \mathbb{D}_\phi$  for every  $n$  and a continuous, linear map  $\phi'_P: \mathbb{D}_0 \mapsto \mathbb{E}$ .

**3.9.13 Theorem (Delta-method for bootstrap almost surely).** Let  $\mathbb{D}$  and  $\mathbb{E}$  be normed spaces. Let  $\phi: \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$  satisfy (3.9.12) for a given measurable subspace  $\mathbb{D}_0$ . Let  $\mathbb{P}_n$  and  $\hat{\mathbb{P}}_n$  be maps as indicated previously with values in  $\mathbb{D}_\phi$  such that  $\sqrt{n}(\mathbb{P}_n - P) \rightsquigarrow \mathbb{G}$ ,  $\|\mathbb{P}_n - P\| \rightarrow 0$  outer almost surely and (3.9.9) holds outer almost surely, where  $\mathbb{G}$  is tight and takes its values in  $\mathbb{D}_0$ . Then (3.9.10) holds outer almost surely.

The uniform Hadamard differentiability needed for the preceding theorem is generally satisfied by functionals on finite-dimensional spaces, but can easily fail in infinite dimensions. An alternative is to use Fréchet differentiability. It suffices to consider Fréchet differentiability at a fixed point, but we need control over the order of the remainder term. Assume that

$$\phi(P + h) - \phi(P) = \phi'_P(h) + O(q(\|h\|)), \quad h \rightarrow 0,$$

for a continuous, linear map  $\phi'_P: \mathbb{D} \mapsto \mathbb{E}$  and a given monotone function  $q$  with (at least)  $q(t) = o(t)$ . The extra control over the remainder term can be exploited if there is also additional information on the almost-sure behavior of  $\|\mathbb{P}_n - P\|$ . Assume that

$$(3.9.14) \quad \limsup_{n \rightarrow \infty} \frac{\sqrt{n} \|\mathbb{P}_n - P\|^*}{b_n} \leq 1 \quad \text{a.s.,}$$

for some sequence  $b_n \rightarrow \infty$ . For  $\mathbb{P}_n$  equal to the empirical distribution function and  $b_n = \sqrt{2 \log \log n} \|P(f - Pf)^2\|_{\mathcal{F}}^{1/2}$ , this follows from the law of the iterated logarithm, which asserts equality in the preceding display. This is well known to be valid for a Donsker class  $\mathcal{F}$  with square-integrable envelope function.<sup>b</sup> The following theorem requires that  $b_n \rightarrow \infty$  sufficiently slowly that  $\sqrt{n} q(c b_n / \sqrt{n}) \rightarrow 0$  for some  $c > 1$ . For  $b_n \sim \sqrt{\log \log n}$ , this condition is satisfied already for  $q(t) = t / (\log \log(1/t))^r$  and  $r > 1/2$ .

<sup>b</sup> Cf. Dudley and Philipp (1983), Theorem 1.3, together with Kuelbs (1976); see also Ledoux and Talagrand (1991).

**3.9.15 Theorem (Delta-method for bootstrap almost surely).** Let  $\mathbb{D}$  and  $\mathbb{E}$  be normed spaces. Suppose  $\phi: \mathbb{D}_\phi \subset \mathbb{D} \mapsto \mathbb{E}$  is Fréchet-differentiable at  $P$  with rate function  $q$ . Let  $\mathbb{P}_n$  and  $\hat{\mathbb{P}}_n$  be maps as indicated previously with values in  $\mathbb{D}_\phi$  such that  $\sqrt{n}(\mathbb{P}_n - P) \rightsquigarrow \mathbb{G}$ , and (3.9.9) holds outer almost surely, where  $\mathbb{G}$  takes its values in a separable subspace  $\mathbb{D}_0$ . Furthermore let (3.9.14) hold for a sequence  $b_n$  such that  $b_n = o(\sqrt{n})$  and  $\sqrt{n}q(c b_n / \sqrt{n}) \rightarrow 0$  for some  $c > 1$ . Then (3.9.10) holds outer almost surely.

**Proofs.** Without loss of generality assume, that the derivative  $\phi'_P: \mathbb{D} \mapsto \mathbb{E}$  is defined and continuous on the whole space. (Otherwise, replace  $\mathbb{E}$  by its second dual  $\mathbb{E}^{**}$  and the derivative by an extension  $\phi'_P: \mathbb{D} \mapsto \mathbb{E}^{**}$ .) For every  $h \in \text{BL}_1(\mathbb{E})$ , the function  $h \circ \phi'_P$  is contained in  $\text{BL}_{\|\phi'_P\|}(\mathbb{D})$ . Thus (3.9.9) implies

$$\sup_{h \in \text{BL}_1(\mathbb{E})} \left| \mathbb{E}_M h\left(\phi'_P(\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n))\right) - \mathbb{E} h(\phi'_P(\mathbb{G})) \right| \rightarrow 0,$$

in outer probability or outer almost surely, corresponding to which of the two assumptions is made in (3.9.9). Next

$$(3.9.16) \quad \begin{aligned} & \sup_{h \in \text{BL}_1(\mathbb{E})} \left| \mathbb{E}_M h\left(\sqrt{n}(\phi(\hat{\mathbb{P}}_n) - \phi(\mathbb{P}_n))\right) - \mathbb{E}_M h\left(\phi'_P(\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n))\right) \right| \\ & \leq \varepsilon + 2\mathbb{P}_M \left( \left\| \sqrt{n}(\phi(\hat{\mathbb{P}}_n) - \phi(\mathbb{P}_n)) - \phi'_P(\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n)) \right\|^* > \varepsilon \right). \end{aligned}$$

The three theorems are proved once it has been shown that the conditional probability on the right converges to zero in outer probability (Theorem 3.9.11) or outer almost surely (Theorem 3.9.13 and 3.9.15).

Both sequences  $\sqrt{n}(\mathbb{P}_n - P)$  and  $\sqrt{n}(\hat{\mathbb{P}}_n - P)$  converge (unconditionally) in distribution to separable random elements that concentrate on the space  $\mathbb{D}_0$ . By Theorem 3.9.4,

$$\begin{aligned} \sqrt{n}(\phi(\hat{\mathbb{P}}_n) - \phi(P)) &= \phi'_P(\sqrt{n}(\hat{\mathbb{P}}_n - P)) + o_P^*(1), \\ \sqrt{n}(\phi(\mathbb{P}_n) - \phi(P)) &= \phi'_P(\sqrt{n}(\mathbb{P}_n - P)) + o_P^*(1). \end{aligned}$$

Subtract these equations to conclude that the sequence  $\sqrt{n}(\phi(\hat{\mathbb{P}}_n) - \phi(\mathbb{P}_n)) - \phi'_P(\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n))$  converges (unconditionally) to zero in outer probability. Thus, the conditional probability on the right in (3.9.16) converges to zero in outer mean. This concludes the proof of the first theorem.

For the proof of the second theorem, fix  $\varepsilon > 0$  and choose a compact set  $K \subset \mathbb{D}_0$  such that  $\mathbb{P}(\mathbb{G} \notin K) < \varepsilon$ . By the uniform Hadamard differentiability of  $\phi$ , there exist  $\delta, \eta > 0$  such that for every  $P' \in \mathbb{D}_\phi$ ,  $\|P' - P\| < \eta$ ,  $t < \eta$ ,  $P' + th \in \mathbb{D}_\phi$ , and  $h \in K^\delta$ :

$$\left\| \frac{\phi(P' + th) - \phi(P')}{t} - \phi'_P(h) \right\| < \varepsilon.$$

Consequently, for  $n \geq 1/\eta^2$ , the right side of (3.9.16) is bounded by

$$\varepsilon + 2\mathbf{E}_M 1_{\mathbb{D}-K^\delta} (\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n))^* + \{\|\mathbb{P}_n - P\| \geq \eta\}^*.$$

The last term converges to zero almost surely by assumption. The function  $h(z) = \delta^{-1}(d(z, K) \wedge \delta)$  is bounded and Lipschitz and satisfies  $1_{\mathbb{D}-K^\delta} \leq h \leq 1_{\mathbb{D}-K}$ . Hence the conditional expectation in the middle term is bounded by

$$\mathbf{E}_M h(\sqrt{n}(\hat{\mathbb{P}}_n - \mathbb{P}_n))^* \xrightarrow{\text{as}} \mathbf{E} h(\mathbb{G}) \leq \mathbf{P}(\mathbb{G} \notin K) < \varepsilon.$$

Thus, the conditional probability in the right side of (3.9.16) converges to zero almost surely. This concludes the proof of the second theorem.

If  $\text{rem}(h) = \phi(P+h) - \phi(P) - \phi'_P(h)$ , then the norm in the right side of (3.9.16) can be written  $\sqrt{n}\|\text{rem}(\hat{\mathbb{P}}_n - P) - \text{rem}(\mathbb{P}_n - P)\|$  and can be bounded above by the triangle inequality. By assumption, there exist  $K, \eta > 0$  such that  $\text{rem}(h) \leq Kq(\|h\|)$  for every  $\|h\| \leq \eta$ . Thus

$$\begin{aligned} \mathbf{P}_M \left( \sqrt{n}\|\text{rem}(\mathbb{P}_n - P)\| > \varepsilon \right)^* \\ \leq \{\|\mathbb{P}_n - P\|^* > \eta\} + \left\{ \sqrt{n}q(\|\mathbb{P}_n - P\|) > \frac{\varepsilon}{K} \right\}^*. \end{aligned}$$

This converges to zero almost surely by assumption. Next, for  $\eta_n/\sqrt{n} + \eta/2 < \eta$ , the triangle inequality and monotonicity of  $q$  yield

$$\begin{aligned} \mathbf{P}_M \left( \sqrt{n}\|\text{rem}(\hat{\mathbb{P}}_n - P)\|^* > \varepsilon \right) \leq \mathbf{P}_M \left( \sqrt{n}\|\hat{\mathbb{P}}_n - \mathbb{P}_n\|^* > \eta_n \right) \\ + \left\{ \|\mathbb{P}_n - P\|^* > \frac{\eta}{2} \right\} + \left\{ \sqrt{n}q(\eta_n/\sqrt{n} + \|\mathbb{P}_n - P\|) > \frac{\varepsilon}{K} \right\}^*. \end{aligned}$$

The first two terms on the right converge to zero almost surely for every  $\eta_n \rightarrow \infty$  and  $\eta > 0$ . The third term converges to zero for  $\eta_n = o(b_n)$ . ■

### 3.9.4 Examples of the Delta-Method

Throughout this section,  $D[a, b]$  is the Banach space of all cadlag functions  $z: [a, b] \mapsto \mathbb{R}$  on an interval  $[a, b] \subset \bar{\mathbb{R}}$  equipped with the uniform norm. The notation  $\int |dA|$  is used for the total variation of a function  $A$ , and  $\text{BV}_M[a, b]$  is the set of all cadlag functions of total variation bounded by  $M$ . Product spaces, such as  $D[a, b] \times D[a, b]$ , are always equipped with a product norm. Given an arbitrary set  $\mathcal{X}$  and Banach space  $\mathcal{Y}$ , the Banach space  $\ell^\infty(\mathcal{X}, \mathcal{Y})$  is the set of all maps  $z: \mathcal{X} \mapsto \mathcal{Y}$  that are uniformly norm-bounded equipped with the norm  $\|z\| = \sup_x \|z(x)\|_{\mathcal{Y}}$ .

#### 3.9.4.1 The Wilcoxon Statistic

Given a cadlag function  $A$  and a function of bounded variation  $B$  on an interval  $[a, b] \subset \bar{\mathbb{R}}$ , define

$$\phi(A, B) = \int_{(a,b]} A dB \quad \text{and} \quad \psi(A, B)(t) = \int_{(a,t]} A dB.$$

These maps are Hadamard-differentiable if the domain is restricted to pairs  $(A, B)$  such that  $B$  is of total variation bounded by some fixed constant.

**3.9.17 Lemma.** For each fixed  $M$ , the maps  $\phi: D[a, b] \times \text{BV}_M[a, b] \mapsto \mathbb{R}$  and  $\psi: D[a, b] \times \text{BV}_M[a, b] \mapsto D[a, b]$  are Hadamard-differentiable at each  $(A, B) \in D_\phi$  such that  $\int |dA| < \infty$ . The derivatives are given by

$$\phi'_{A,B}(\alpha, \beta) = \int A d\beta + \int \alpha dB; \quad \psi'_{A,B}(\alpha, \beta)(t) = \int_{(a,t]} A d\beta + \int_{(a,t]} \alpha dB,$$

where  $\int A d\beta$  is defined via integration by parts<sup>#</sup> if  $\beta$  is not of bounded variation.

**Proof.** For  $\alpha_t \rightarrow \alpha$  and  $\beta_t \rightarrow \beta$ , define  $A_t = A + t\alpha_t$  and  $B_t = B + t\beta_t$ . By convention, we consider only perturbations such that  $(A_t, B_t)$  is contained in the given domain. In particular, the variation of  $B_t$  is bounded by  $M$ . Write

$$\frac{\int A_t dB_t - \int A dB}{t} - \phi'_{A,B}(\alpha_t, \beta_t) = \int \alpha d(B_t - B) + \int (\alpha_t - \alpha) d(B_t - B).$$

Since  $A$  is of bounded variation, the derivative map as stated is continuous (with respect to the uniform norm). It suffices to show that the expression in the display converges to zero. The second term on the right is bounded in absolute value by  $\|\alpha_t - \alpha\|_\infty 2M$  and hence converges to zero. Since  $\alpha$  is cadlag on  $[a, b]$ , there exists a partition  $a = t_0 < t_1 < \dots < t_m = b$  such that  $\alpha$  varies less than  $\varepsilon$  on each interval  $[t_{i-1}, t_i]$ . Let  $\tilde{\alpha}$  be the discretization that is constant and equal to  $\alpha(t_{i-1})$  on  $[t_{i-1}, t_i]$ . Then

$$\begin{aligned} \left| \int \alpha d(B_t - B) \right| &\leq \|\alpha - \tilde{\alpha}\|_\infty 2M + \sum_{i=1}^m |\alpha(t_{i-1})| |(B_t - B)[t_{i-1}, t_i]| \\ &\quad + |\alpha(b)| |(B_t - B)\{b\}|. \end{aligned}$$

The first term on the right can be made arbitrarily small by the choice of  $\varepsilon$ . The second term is bounded by  $(2m+1)\|B_t - B\|_\infty \|\alpha\|_\infty$  and hence converges to zero for every fixed partition.

The proof for  $\psi$  is basically the same as that for  $\phi$ . ■

The preceding lemma has many corollaries. Here we discuss two simple statistical consequences.

**3.9.18 Example (Wilcoxon statistic).** Let  $X_1, \dots, X_m$  and  $Y_1, \dots, Y_n$  be independent samples from distribution functions  $F$  and  $G$  on the real line, respectively. If  $\mathbb{F}_m$  and  $\mathbb{G}_n$  are the empirical distribution functions of the two samples, then

$$\phi(\mathbb{F}_m, \mathbb{G}_n) = \int \mathbb{F}_m d\mathbb{G}_n$$

---

<sup>#</sup> Thus  $\int A d\beta = (A\beta)(b) - (A\beta)(a) - \int (\beta-) dA$ .

is the Mann-Whitney form of the Wilcoxon statistic. It is an estimator of  $\phi(F, G) = \int F dG = P(X \leq Y)$ . As a consequence of the differentiability of  $\phi$ , if  $m/(m+n) \rightarrow \lambda$ , then

$$\begin{aligned} \sqrt{\frac{mn}{m+n}} \left( \int \mathbb{F}_m d\mathbb{G}_n - \int F dG \right) &\rightsquigarrow \phi'_{F,G}(\sqrt{1-\lambda}\mathbb{G}_F, \sqrt{\lambda}\mathbb{G}_G) \\ &= \sqrt{\lambda} \int F d\mathbb{G}_G + \sqrt{1-\lambda} \int \mathbb{G}_F dG, \end{aligned}$$

where  $\mathbb{G}_F$  and  $\mathbb{G}_G$  are independent tight,  $F$ - and  $G$ -Brownian bridge processes. The limit variable is zero-mean normally distributed with variance  $\lambda \text{ var } F(Y) + (1-\lambda) \text{ var } G(X)$ . An elementary way to see this is to use the stronger assertion of Theorem 3.9.4, which gives the weak representation

$$\begin{aligned} \sqrt{\frac{mn}{m+n}} \left( \int \mathbb{F}_m d\mathbb{G}_n - \int F dG \right) \\ = \sqrt{\lambda} \int F d\mathbb{G}_{n,G} + \sqrt{1-\lambda} \int \mathbb{G}_{m,F} dG + o_P(1). \end{aligned}$$

The variable on the right is asymptotically normal by the central limit theorem and Slutsky's lemma. (Also see Problems 3.9.2 and 3.9.3.)

**3.9.19 Example (Nelson-Aalen).** The Nelson-Aalen estimator of a cumulative hazard function based on censored data is a Hadamard-differentiable map of the empirical distribution function of the observations. Let  $X_1, \dots, X_n$  be i.i.d. "failure times" distributed according to the distribution function  $F$  and let  $C_1, \dots, C_n$  be i.i.d. "censoring times" distributed according to the distribution function  $G$ . Failure times and censoring times are assumed independent. Observed are the pairs  $(Z_1, \Delta_1), \dots, (Z_n, \Delta_n)$ , where  $Z_i = X_i \wedge C_i$ , and  $\Delta_i = 1\{X_i \leq C_i\}$  indicates whether a failure time is censored or not. By definition, the cumulative hazard function of the failure times is

$$\Lambda(t) = \int_{[0,t]} \frac{1}{\bar{F}} dF = \int_{[0,t]} \frac{1}{\bar{H}} dH^{uc},$$

where  $\bar{F}(t) = P(X \geq t)$  and  $\bar{H}(t) = P(Z \geq t)$  are (left-continuous) survival distributions, and  $H^{uc}(t) = P(Z \leq t, \Delta = 1)$  is the subdistribution function of the uncensored observations. The *Nelson-Aalen estimator* is given by

$$\Lambda_n(t) = \int_{[0,t]} \frac{1}{\bar{\mathbb{H}}_n} d\mathbb{H}_n^{uc},$$

where

$$\mathbb{H}_n^{uc}(t) = \frac{1}{n} \sum_{i=1}^n \Delta_i 1\{Z_i \leq t\} \quad \text{and} \quad \bar{\mathbb{H}}_n(t) = \frac{1}{n} \sum_{i=1}^n 1\{Z_i \geq t\}$$

are the empirical subdistribution functions of the uncensored failure times and survival function of the observation times, respectively. The Nelson-Aalen estimator depends on the pair  $(\mathbb{H}_n^{uc}, \bar{\mathbb{H}}_n)$  through the two maps

$$(A, B) \mapsto \left( A, \frac{1}{B} \right) \mapsto \int_{[0,t]} \frac{1}{B} dA.$$

The composition map is Hadamard-differentiable on a domain of the type  $\{(A, B) : \int |dA| \leq M, B \geq \varepsilon\}$  for given  $M$  and  $\varepsilon > 0$ , at every point  $(A, B)$  such that  $1/B$  is of bounded variation. If  $t$  is restricted to an interval  $[0, \tau]$  such that  $H(\tau) < 1$ , then the pair  $(\mathbb{H}_n^{uc}, \bar{\mathbb{H}}_n)$  is contained in this domain with probability tending to 1 for  $M \geq 1$  and sufficiently small  $\varepsilon$ . The derivative map is given by  $(\alpha, \beta) \mapsto \int (1/B) d\alpha - \int (\beta/B^2) dA$ .

The pair  $(\mathbb{H}_n^{uc}, \bar{\mathbb{H}}_n)$  can be identified with the empirical distribution of the observations indexed by the functions  $\delta 1\{z \leq t\}$  and  $1\{z > t\}$ , when  $t$  ranges over  $\mathbb{R}$ . Any of the main empirical central limit theorems yields

$$\sqrt{n}(\mathbb{H}_n^{uc} - H^{uc}, \bar{\mathbb{H}}_n - \bar{H}) \rightsquigarrow (\mathbb{G}^{uc}, \bar{\mathbb{G}}), \quad \text{in } (D[0, \tau])^2,$$

where  $(\mathbb{G}^{uc}, \bar{\mathbb{G}})$  is a tight, zero-mean Gaussian process with covariance structure

$$E\mathbb{G}^{uc}(s)\mathbb{G}^{uc}(t) = H^{uc}(s \wedge t) - H^{uc}(s)H^{uc}(t),$$

$$E\bar{\mathbb{G}}(s)\bar{\mathbb{G}}(t) = \bar{H}(s \vee t) - \bar{H}(s)\bar{H}(t),$$

$$E\mathbb{G}^{uc}(s)\bar{\mathbb{G}}(t) = (H^{uc}(s) - H^{uc}(t-))1\{t \leq s\} - H^{uc}(s)\bar{H}(t).$$

By the delta-method we conclude that

$$\sqrt{n}(\Lambda_n - \Lambda) \rightsquigarrow \int_{[0,t]} \frac{1}{\bar{H}} d\mathbb{G}^{uc} - \int_{[0,t]} \frac{\bar{\mathbb{G}}}{\bar{H}^2} dH^{uc},$$

where the first term on the right is to be understood via partial integration. The covariance function of the Gaussian limit may be evaluated by martingale calculus.<sup>†</sup>

The process  $\mathbb{M}^{uc}(t) = \mathbb{G}^{uc}(t) - \int_{[0,t]} \bar{\mathbb{G}} d\Lambda$  is a zero-mean Gaussian martingale with covariance function

$$E\mathbb{M}^{uc}(s)\mathbb{M}^{uc}(t) = \int_{[0,s \wedge t]} \bar{H}(1 - \Delta\Lambda) d\Lambda.$$

The limit variable can be expressed as the stochastic integral  $\int_{[0,t]} 1/\bar{H} d\mathbb{M}^{uc}$  and therefore is a martingale. It is distributed as  $\mathbb{Z}(C)$  for a standard Brownian motion  $\mathbb{Z}$  and the function  $C$  given by

$$C(t) = \int_{[0,t]} \frac{(1 - \Delta\Lambda)}{\bar{H}} d\Lambda.$$

The final conclusion is that the sequence  $\sqrt{n}(\Lambda_n - \Lambda)$  converges in distribution to  $\mathbb{Z}(C)$  in  $D[0, \tau]$  for every  $\tau$  such that  $H(\tau) < 1$ .

---

<sup>†</sup> See, for instance, Revuz and Yor (1994), Chapter 4, Section 2, page 138.

### 3.9.4.2 The Inverse Map

For a nondecreasing function  $A \in D[a, b]$  and fixed  $p \in \mathbb{R}$ , let  $\phi(A) \in [a, b]$  be an arbitrary point in  $[a, b]$  such that

$$A(\phi(A)-) \leq p \leq A(\phi(A)).$$

The natural domain  $\mathbb{D}_\phi$  of the resulting map  $\phi$  is the set of all nondecreasing  $A$  such that there exists a solution to the pair of inequalities. If there exists more than one solution, then the precise choice of  $\phi(A)$  is irrelevant. For instance,  $\phi(A)$  may be taken equal to  $\inf\{t: A(t) \geq p\}$ .

**3.9.20 Lemma.** *Let  $A \in \mathbb{D}_\phi$  be differentiable at a point  $\xi_p \in (a, b)$  such that  $A(\xi_p) = p$ , with strictly positive derivative. Then  $\phi: \mathbb{D}_\phi \subset D[a, b] \mapsto \mathbb{R}$  is Hadamard-differentiable at  $A$  tangentially to the set of functions  $\alpha \in D[a, b]$  that are continuous at  $\xi_p$ . The derivative is given by  $\phi'_A(\alpha) = -\alpha(\xi_p)/A'(\xi_p)$ .*

**Proof.** Let  $\alpha_t \rightarrow \alpha$  uniformly on  $[a, b]$  for a function  $\alpha$  that is continuous at  $\xi_p$ . Write  $\xi_{pt}$  for  $\phi(A + t\alpha_t)$ . By the definition of  $\phi$ , we have

$$(A + t\alpha_t)(\xi_{pt} - \varepsilon_t) \leq p \leq (A + t\alpha_t)(\xi_{pt}),$$

for every  $\varepsilon_t > 0$ . Choose  $\varepsilon_t$  positive and such that  $\varepsilon_t = o(t)$ . Since  $\alpha_t$  converges uniformly to a bounded function, the sequence  $\alpha_t$  is uniformly bounded. Conclude that  $A(\xi_{pt} - \varepsilon_t) + O(t) \leq p \leq A(\xi_{pt}) + O(t)$ . By assumption, the function  $A$  is monotone and bounded away from  $p$  outside any interval  $(\xi_p - \varepsilon, \xi_p + \varepsilon)$  around  $\xi_p$ . To satisfy the preceding inequalities, the numbers  $\xi_{pt}$  must be to the right of  $\xi_p - \varepsilon$  eventually, and the numbers  $\xi_{pt} - \varepsilon_t$  must be to the left of  $\xi_p + \varepsilon$  eventually. In other words,  $\xi_{pt} \rightarrow \xi_p$ .

By the uniform convergence of  $\alpha_t$  and the continuity of the limit,  $\alpha_t(\xi_{pt} - \varepsilon_t) \rightarrow \alpha(\xi_p)$  for every  $\varepsilon_t \rightarrow 0$ . Using this and Taylor's formula on the preceding display yields

$$\begin{aligned} p + (\xi_{pt} - \xi_p)A'(\xi_p) - o(\xi_{pt} - \xi_p) + O(\varepsilon_t) + t\alpha(\xi_p) - o(t) \\ \leq p \leq p + (\xi_{pt} - \xi_p)A'(\xi_p) + o(\xi_{pt} - \xi_p) + O(\varepsilon_t) + t\alpha(\xi_p) + o(t). \end{aligned}$$

Conclude first that  $\xi_{pt} - \xi_p = O(t)$ . Next, use this to replace the  $o(\xi_{pt} - \xi_p)$  in the display by  $o(t)$ -terms and conclude that  $(\xi_{pt} - \xi_p)/t \rightarrow -(\alpha/A')(\xi_p)$ . ■

**3.9.21 Example (Empirical quantiles).** Fix  $0 < p < 1$  and consider the map that assigns to each cumulative distribution function its  $p$ th quantile  $F^{-1}(p) = \inf\{x: F(x) \geq p\}$ . This map is Hadamard-differentiable at every cumulative distribution  $F$  that is differentiable at  $F^{-1}(p)$  with strictly positive derivative  $f(F^{-1}(p))$ , tangentially to the set of functions that are continuous at  $F^{-1}(p)$ .

If  $\mathbb{F}_n$  is the empirical distribution function of an i.i.d. sample of size  $n$  from  $F$ , then the sequence  $\sqrt{n}(\mathbb{F}_n - F)$  converges in  $D(\bar{\mathbb{R}})$  to the process

$\mathbb{G} \circ F$ , where  $\mathbb{G}$  is a standard Brownian bridge. Since  $F$  is continuous at  $F^{-1}(p)$ , almost all the sample paths of this process are continuous at the point  $F^{-1}(p)$ . Thus, the delta-method yields

$$\sqrt{n}(\mathbb{F}_n^{-1}(p) - F^{-1}(p)) \rightsquigarrow \phi'_F(\mathbb{G} \circ F) = -\frac{\mathbb{G}(p)}{f(F^{-1}(p))}.$$

This variable on the right is zero-mean normally distributed with its variance given by  $p(1-p)/f^2(F^{-1}(p))$ .

**3.9.22 Example.** While the delta-method yields an elegant proof of the asymptotic normality of the empirical quantile function, this result can easily be obtained directly. The real attraction of the delta-method is that it separates the analysis and the probability and can be applied without much work in a variety of situations. In the preceding derivation, the empirical distribution can be replaced by any estimator  $\tilde{F}_n$  such that the sequence  $\sqrt{n}(\tilde{F}_n - F)$  converges weakly in  $D[a, b]$  for an interval  $[a, b]$  that contains  $F^{-1}(p)$  as an interior point. The estimator  $\tilde{F}_n$  need not be based on an i.i.d. sample, and  $F$  may be an arbitrary centering function.

Two concrete applications are the inverse of the Nelson-Aalen estimator and the quantiles of the Kaplan-Meier estimator (see Examples 3.9.19 and 3.9.31).

Next consider the inverse function  $\phi(F) = F^{-1}$  as a map from the set of distribution functions into the space  $\ell^\infty(p, q)$  for given  $0 < p < q < 1$ . For definiteness, let  $F^{-1}$  denote the left-continuous inverse  $F^{-1}(p) = \inf\{t: F(t) \geq p\}$ . Given an interval  $[a, b] \subset \mathbb{R}$  let  $\mathbb{D}_1$  be the set of all restrictions of distribution functions on  $\mathbb{R}$  to  $[a, b]$  and let  $\mathbb{D}_2$  be the subset of  $\mathbb{D}_1$  of distribution functions of measures that concentrate on  $(a, b]$ .

### 3.9.23 Lemma.

- (i) Let  $0 < p < q < 1$ , and let  $F$  be continuously differentiable on the interval  $[a, b] = [F^{-1}(p) - \varepsilon, F^{-1}(q) + \varepsilon]$  for some  $\varepsilon > 0$ , with strictly positive derivative  $f$ . Then the inverse map  $G \mapsto G^{-1}$  as a map  $\mathbb{D}_1 \subset D[a, b] \mapsto \ell^\infty[p, q]$  is Hadamard-differentiable at  $F$  tangentially to  $C[a, b]$ .
- (ii) Let  $F$  have compact support  $[a, b]$  and be continuously differentiable on its support with strictly positive derivative  $f$ . Then the inverse map  $G \mapsto G^{-1}$  as a map  $\mathbb{D}_2 \subset D[a, b] \mapsto \ell^\infty(0, 1)$  is Hadamard-differentiable at  $F$  tangentially to  $C[a, b]$ .

In both cases the derivative is the map  $\alpha \mapsto -(\alpha/f) \circ F^{-1}$ .

**Proof.** The proof follows the same pattern as the proof for the preceding lemma, but with care to keep the needed uniformity in  $p$ . We give the proof of (ii) only.

Let  $\alpha_t \rightarrow \alpha$  uniformly in  $D[a, b]$ , where  $\alpha$  is continuous and  $F + t\alpha_t$  is contained in  $\mathbb{D}_2$  for all  $t$ . Abbreviate  $F^{-1}(p)$  and  $(F + t\alpha_t)^{-1}(p)$  to  $\xi_p$  and  $\xi_{pt}$ , respectively. Since  $F$  and  $F + t\alpha_t$  are concentrated on  $(a, b]$  (by assumption), we have  $a < \xi_{pt}, \xi_p \leq b$  for all  $0 < p < 1$ . Thus the numbers  $\varepsilon_{pt} = t^2 \wedge (\xi_{pt} - a)$  are positive, whence by definition

$$(F + t\alpha_t)(\xi_{pt} - \varepsilon_{pt}) \leq p \leq (F + t\alpha_t)(\xi_{pt}).$$

By the smoothness of  $F$  we have  $F(\xi_p) = p$  and  $F(\xi_{pt} - \varepsilon_{pt}) = F(\xi_{pt}) + O(\varepsilon_{pt})$ , uniformly in  $0 < p < 1$ . It follows that

$$-t\alpha(\xi_{pt}) + o(t) \leq F(\xi_{pt}) - F(\xi_p) \leq -t\alpha(\xi_{pt} - \varepsilon_{pt}) + o(t).$$

The  $o(t)$ -terms are uniform in  $0 < p < 1$ . The far left side and the far right side are  $O(t)$ , while the middle is bounded above and below by a constant times  $|\xi_{pt} - \xi_p|$ . Conclude that  $|\xi_{pt} - \xi_p| = O(t)$ , uniformly in  $p$ . Next, the lemma follows by the uniform differentiability of  $F$ . ■

**3.9.24 Example (Empirical quantile process).** Suppose that  $F$  is a distribution function with continuous and positive derivative  $f$  on the interval  $[F^{-1}(p) - \varepsilon, F^{-1}(q) + \varepsilon]$  for some  $\varepsilon > 0$ . The empirical distribution function  $\mathbb{F}_n$  of an i.i.d. sample of size  $n$  from  $F$  satisfies  $\sqrt{n}(\mathbb{F}_n - F) \rightsquigarrow \mathbb{G} \circ F$  in  $D(\bar{\mathbb{R}})$  for a standard Brownian bridge  $\mathbb{G}$ . Almost all sample paths of the limit process are continuous on the interval  $[F^{-1}(p) - \varepsilon, F^{-1}(q) + \varepsilon]$ . Since the inverse map is Hadamard-differentiable at  $F$  tangentially to the subspace of functions that are continuous on this interval, the delta-method yields

$$\sqrt{n}(\mathbb{F}_n^{-1} - F^{-1}) \rightsquigarrow -\frac{\mathbb{G} \circ F(F^{-1})}{f(F^{-1})}, \quad \text{in } \ell^\infty[p, q].$$

The limit process is Gaussian with zero-mean and covariance function

$$\frac{s \wedge t - st}{f(F^{-1}(s)) f(F^{-1}(t))}, \quad s, t \in \mathbb{R}.$$

The second part of the preceding lemma may be used to obtain uniform convergence of the whole quantile process for distributions with compact support and strictly positive, continuous density, such as the uniform distribution.

### 3.9.4.3 Composition

Let  $g: \mathbb{R} \mapsto \mathbb{R}$  be a fixed map. Given an arbitrary set  $\mathcal{X}$ , consider the map  $\phi: \ell^\infty(\mathcal{X}) \mapsto \ell^\infty(\mathcal{X})$  given by  $\phi(A)(x) = g(A(x))$ . The natural domain of this map is the set of elements of  $\ell^\infty(\mathcal{X})$  that take their values in the domain of  $g$ .

**3.9.25 Lemma.** Let  $g: (a, b) \subset \mathbb{R} \mapsto \mathbb{R}$  be differentiable with uniformly continuous and bounded derivative, and let  $\mathbb{D}_\phi = \{A \in \ell^\infty(\mathcal{X}): a < A < b\}$ . Then the map  $A \mapsto g \circ A$  is Hadamard-differentiable as a map  $\mathbb{D}_\phi \subset \ell^\infty(\mathcal{X}) \mapsto \ell^\infty(\mathcal{X})$  at every  $A \in \mathbb{D}_\phi$ . The derivative is given by  $\phi'_A(\alpha) = g'(A(x))\alpha(x)$ .

**3.9.26 Example.** The map  $A \mapsto 1/A$  is differentiable on the domain of functions that are bounded away from zero.

The proof of the preceding lemma is easy. Consider the more complicated composition map into which the function  $g$ , which is fixed in the preceding result, is a second variable. Given maps  $A: \mathcal{X} \mapsto \mathcal{Y}$  and  $B: \mathcal{Y} \mapsto \mathcal{Z}$  define the composition map  $\phi(A, B): \mathcal{X} \mapsto \mathcal{Z}$ , by

$$\phi(A, B)(x) = B \circ A(x) = B(A(x)).$$

We shall assume that  $\mathcal{Y}$  and  $\mathcal{Z}$  are subsets of normed spaces. If  $B$  is a uniformly norm-bounded map from  $\mathcal{Y} \mapsto \mathcal{Z}$ , then  $\phi(A, B)$  is a uniformly norm-bounded map from  $\mathcal{X} \mapsto \mathcal{Z}$ . Consider  $\phi$  as a map with domain  $\ell^\infty(\mathcal{X}, \mathcal{Y}) \times \ell^\infty(\mathcal{Y}, \mathcal{Z})$  equipped with the norm  $\|(A, B)\|_\infty = \sup_x \|A(x)\|_{\mathcal{Y}} \vee \sup_y \|B(y)\|_{\mathcal{Z}}$ .

**3.9.27 Lemma.** Suppose that  $B: \mathcal{Y} \mapsto \mathcal{Z}$  is Fréchet-differentiable uniformly in  $y$  in the range of  $A$  with derivatives  $B'_y$  such that  $y \mapsto B'_y$  is uniformly norm-bounded.<sup>†</sup> Then the composition map  $\phi: \ell^\infty(\mathcal{X}, \mathcal{Y}) \times \ell^\infty(\mathcal{Y}, \mathcal{Z}) \rightarrow \ell^\infty(\mathcal{X}, \mathcal{Z})$  is Hadamard-differentiable at  $(A, B)$  tangentially to the set  $\ell^\infty(\mathcal{X}, \mathcal{Y}) \times \text{UC}(\mathcal{Y}, \mathcal{Z})$ . The derivative is given by

$$\phi'_{A,B}(\alpha, \beta)(x) = \beta \circ A(x) + B'_{A(x)}(\alpha(x)), \quad x \in \mathcal{X}.$$

**Proof.** Let  $\alpha_t \rightarrow \alpha$  and  $\beta_t \rightarrow \beta$  in  $\ell^\infty(\mathcal{X}, \mathcal{Y})$  and  $\ell^\infty(\mathcal{Y}, \mathcal{Z})$ , respectively. Write

$$\begin{aligned} & \frac{(B + t\beta_t) \circ (A + t\alpha_t) - B \circ A}{t} - \beta \circ A - B'_A(\alpha) \\ &= (\beta_t - \beta) \circ (A + t\alpha_t) + \beta(A + t\alpha_t) - \beta(A) \\ &+ \frac{B(A + t\alpha_t) - B(A)}{t} - B'_A(\alpha_t) + B'_A(\alpha_t - \alpha). \end{aligned}$$

The first term converges to zero in  $\ell^\infty(\mathcal{X}, \mathcal{Z})$ , since  $\beta_t \rightarrow \beta$  uniformly in  $y$ . The second term converges to zero for every  $\beta \in \text{UC}(\mathcal{Y}, \mathcal{Z})$ . The third term converges to zero by uniform Fréchet differentiability. The fourth term converges to zero since the map  $y \mapsto B'_y$  is norm-bounded. ■

<sup>†</sup> Differentiability at each  $y$  in a convex set plus the uniform norm continuity of the derivatives  $y \mapsto B'_y$  implies uniform Fréchet differentiability. See Problem 3.9.1.

### 3.9.4.4 Copula Function

For a bivariate distribution function  $H$ , denote the marginals by  $F(x) = H(x, \infty)$  and  $G(y) = H(\infty, y)$ . Consider the map  $\phi$  from bivariate distribution functions  $H$  on  $\mathbb{R}^2$  to bivariate distribution functions on  $[0, 1]^2$  defined by

$$\phi(H)(u, v) = H(F^{-1}(u), G^{-1}(v)), \quad (u, v) \in [0, 1]^2.$$

Here  $F^{-1}$  and  $G^{-1}$  are the left-continuous quantile functions corresponding to the marginal distribution functions  $F$  and  $G$ , respectively. The function  $C = \phi(H)$  is called the *copula function* corresponding to  $H$ .

**3.9.28 Lemma.** Fix  $0 < p < q < 1$ , and suppose that  $H$  is a distribution function on  $\mathbb{R}^2$  with marginal distribution functions  $F$  and  $G$  that are continuously differentiable on the intervals  $[F^{-1}(p) - \varepsilon, F^{-1}(q) + \varepsilon]$  and  $[G^{-1}(p) - \varepsilon, G^{-1}(q) + \varepsilon]$  with positive derivatives  $f$  and  $g$ , respectively, for some  $\varepsilon > 0$ . Furthermore, assume that  $\partial H/\partial x$  and  $\partial H/\partial y$  exist and are continuous on the product of these intervals. Then the map  $\phi: D(\bar{\mathbb{R}}^2) \mapsto \ell^\infty([p, q]^2)$  is Hadamard-differentiable at  $H$  tangentially to  $C(\bar{\mathbb{R}}^2)$ . The derivative is given by

$$\begin{aligned} \phi'_H(h)(u, v) &= h(F^{-1}(u), G^{-1}(v)) \\ &\quad - \frac{\partial H}{\partial x}(F^{-1}(u), G^{-1}(v)) \frac{h(F^{-1}(u), \infty)}{f(F^{-1}(u))} \\ &\quad - \frac{\partial H}{\partial y}(F^{-1}(u), G^{-1}(v)) \frac{h(\infty, G^{-1}(v))}{g(G^{-1}(v))}. \end{aligned}$$

**Proof.** Mapping a distribution function into its copula distribution can be decomposed as

$$H \mapsto (H, F, G) \mapsto (H, F^{-1}, G^{-1}) \mapsto H \circ (F^{-1}, G^{-1}).$$

The first map is linear and continuous, hence Hadamard-differentiable. The second map is Hadamard-differentiable by Section 3.9.4.2, and the third map is Hadamard-differentiable by the result of the preceding section, applied with  $\mathcal{X} = [p, q]^2$ ,  $\mathcal{Y} = \mathbb{R}^2$  and  $\mathcal{Z} = \mathbb{R}$ . The conclusion follows from the chain rule. ■

**3.9.29 Example.** Suppose that  $(X_1, Y_1), \dots, (X_n, Y_n)$  are i.i.d. vectors with distribution function  $H$ . The empirical estimator for the copula function  $C(u, v) = H(F^{-1}(u), G^{-1}(v))$  is

$$\mathbb{C}_n(u, v) = \mathbb{H}_n(\mathbb{F}_n^{-1}(u), \mathbb{G}_n^{-1}(v)),$$

where  $\mathbb{H}_n$ ,  $\mathbb{F}_n$ , and  $\mathbb{G}_n$  are the joint and marginal empirical distribution functions of the observations. If  $H$  satisfies the hypotheses of the preceding lemma, then the sequence  $\sqrt{n}(\mathbb{C}_n - C)$  converges in distribution in

$D([a, b]^2)$  to a  $\phi'_H(\mathbb{G}_H)$  for a tight Brownian bridge  $\mathbb{G}_H$ . The limit variable is distributed as  $\mathbb{G}_H(\phi)$  for the functions  $\phi$  given by

$$\begin{aligned}\dot{\phi}_{u,v}(x, y) &= 1\{x \leq F^{-1}(u), y \leq G^{-1}(v)\} \\ &\quad - \frac{\partial H}{\partial x}(F^{-1}(u), G^{-1}(v)) \frac{1\{x \leq F^{-1}(u)\}}{f(F^{-1}(u))} \\ &\quad - \frac{\partial H}{\partial y}(F^{-1}(u), G^{-1}(v)) \frac{1\{y \leq G^{-1}(v)\}}{g(G^{-1}(v))}.\end{aligned}$$

If  $H(u, v) = uv$  is the uniform distribution function on the unit square, the limit process is a “completely tucked” Brownian sheet with covariance function

$$\text{cov}(\mathbb{Z}(s, t), \mathbb{Z}(u, v)) = (s \wedge u - su)(t \wedge v - tv).$$

In general, the covariance function has a more complicated structure.

### 3.9.4.5 The Product Integral

For a function  $A \in D(0, b]$ , let  $\Delta A(t) = A(t) - A(t-)$  and  $A^c(t) \equiv A(t) - \sum_{0 < s \leq t} \Delta A(s)$  be the jump part and the continuous part, respectively. The *product integral* is defined as

$$\phi(A)(t) \equiv \prod_{0 < s \leq t} (1 + dA(s)) = \prod_{0 < s \leq t} (1 + \Delta A(s)) \exp(A^c(t)).$$

The middle expression is simply notation, which is motivated by the fact that

$$\phi(A)(t) = \lim_{\max_i |t_i - t_{i-1}| \rightarrow 0} \prod_i \left(1 + (A(t_i) - A(t_{i-1}))\right)$$

if the limit is taken over partitions  $0 = t_0 < t_1 < \dots < t_n = t$  with mesh-width decreasing to zero. Alternatively, the product integral can be defined in terms of a Peano series or as the unique solution of a Volterra integral equation (see Problem 3.9.5). As additional notation, we shall use

$$\phi(A)(s, t] = \prod_{s < u \leq t} (1 + dA(u)) = \frac{\phi(A)(t)}{\phi(A)(s)}, \quad s < t.$$

Here both the first and second expressions are defined by the right-hand side, although their definition in terms of the jump and continuous parts of  $A$  restricted to the interval  $(s, t]$  is clear.

Some basic properties of product integrals, such as the “forward” and “backward” equation and the *Duhamel equation*, are given in the Problems and Complements section (Problems 3.9.6 and 3.9.7). The Duhamel equation is the key to proving the Hadamard differentiability of the product integral. It asserts that

$$(\phi(B) - \phi(A))(t) = \int_{(0, t]} \phi(A)(0, u) \phi(B)(u, t] (B - A)(du).$$

Integration by parts shows that this difference is bounded by a constant times  $\|A - B\|_\infty$  (also see Problem 3.9.8). Thus, product integration is uniformly continuous.

**3.9.30 Lemma.** *For fixed, finite, positive constants  $b$  and  $M$ , the product integral map  $\phi: \text{BV}_M[0, b] \subset D[0, b] \mapsto D[0, b]$  is Hadamard-differentiable with derivative*

$$\phi'_A(\alpha)(t) = \int_{(0,t]} \phi(A)(0, u) \phi(A)(u, t] d\alpha(u).$$

Here the integral with respect to  $\alpha$  is defined by integration by parts (as given in Problem 3.9.8) if  $\alpha$  is of unbounded variation.

**Proof.** Set  $A_n = A + t_n \alpha_n$  for a sequence  $\alpha_n \rightarrow \alpha$ . In view of the Duhamel equation, it suffices to show that

$$\int_{(0,t]} \phi(A)(0, u) \phi(A_n)(u, t] d\alpha_n(u) \rightarrow \int_{(0,t]} \phi(A)(0, u) \phi(A)(u, t] d\alpha(u),$$

uniformly in  $0 \leq t \leq b$ . In each side the error if  $\alpha_n$  or  $\alpha$  is replaced by a function  $\tilde{\alpha}$  is bounded by a constant times  $\|\alpha_n - \tilde{\alpha}\|_\infty$  or  $\|\alpha - \tilde{\alpha}\|_\infty$ , respectively. This follows by integration by parts. Choose a function  $\tilde{\alpha}$  of bounded variation that is close to  $\alpha$  in the uniform norm. Then it is also close to  $\alpha_n$  for sufficiently large  $n$ , and the error is small on both sides. Conclude that it suffices to show that

$$\int_{(0,t]} \phi(A)(0, u) \phi(A_n)(u, t] d\tilde{\alpha}(u) \rightarrow \int_{(0,t]} \phi(A)(0, u) \phi(A)(u, t] d\tilde{\alpha}(u),$$

for every function of bounded variation  $\tilde{\alpha}$ . This follows since the product integrals  $\phi(A_n)$  converge uniformly to their limit  $\phi(A)$ , by a second application of the Duhamel equation. ■

The product integral is important in statistics because it transforms cumulative hazard functions into the corresponding survival functions. This can be seen most conveniently from the fact that the product integral is the unique solution of the Volterra equation defined by the negative of the cumulative hazard function. The cumulative hazard function corresponding to the distribution function  $F$  is defined as

$$\Lambda(t) = \int_{[0,t]} \frac{1}{1 - F(s-)} dF(s).$$

This immediately yields

$$1 - \int_{[0,t]} (1 - F(s-)) d\Lambda(s) = 1 - F(t).$$

Thus the survival function  $S = 1 - F$  solves the Volterra equation  $B = 1 + \int B(s-) d(-\Lambda)(s)$  induced by  $-\Lambda$ . According to Problem 3.9.5,

$$1 - F(t) = \prod_{s \leq t} (1 - d\Lambda(s)).$$

The Hadamard differentiability of the product integral now implies that the asymptotic normality of estimators for the cumulative hazard function carries over to the asymptotic normality of the corresponding estimators of the survival function.

Let the estimator  $\widehat{\Lambda}_n$  have the property that the sequence  $\sqrt{n}(\widehat{\Lambda}_n - \Lambda)$  converges in distribution in  $D[0, \tau]$  to a limit  $\mathbb{Z}$ . Let  $1 - \mathbb{F}_n = \phi(-\widehat{\Lambda}_n)$  be the corresponding estimator of  $1 - F = \phi(\Lambda)$ . If the estimators  $\widehat{\Lambda}_n$  are of uniformly bounded total variation (with probability tending to 1), then the delta-method, gives in  $D[0, \tau]$ ,

$$\begin{aligned} \sqrt{n}(\mathbb{F}_n - F) &\rightsquigarrow -\phi'_{-\Lambda}(-\mathbb{Z}) = \int_{(0,t]} S(u-) \frac{S(t)}{S(u)} d\mathbb{Z}(u) \\ &= S(t) \int_{(0,t]} \frac{1}{1 - \Delta\Lambda(u)} d\mathbb{Z}(u). \end{aligned}$$

**3.9.31 Example (Kaplan-Meier).** The Nelson-Aalen estimator has limit distribution  $\mathbb{Z} = \mathbb{B} \circ C$  for a standard Brownian motion  $\mathbb{B}$  and the function  $C$  as given in Example 3.9.19. In that case, the zero-mean Gaussian process  $\phi'_{-\Lambda}(\mathbb{Z})$  in the preceding display has covariance function

$$S(s)S(t) \int_{[0,s \wedge t]} \frac{1}{(1 - \Delta\Lambda) \bar{H}} d\Lambda, \quad 0 \leq s, t \leq \tau.$$

The corresponding estimator for the survival function is the *Kaplan-Meier estimator*.

### 3.9.4.6 Multivariate Trimming

Throughout this section fix a number  $0 < \alpha < 1/2$  and let  $\mathcal{H}$  denote the collection of all closed half-spaces in  $\mathbb{R}^d$ . For a given probability distribution  $P$  on  $\mathbb{R}^d$ , define a compact, convex set by

$$K_P = \cap \{H \in \mathcal{H}: P(H) \geq 1 - \alpha\}.$$

It will be assumed that  $K_P$  contains a neighborhood of the origin. (If  $K_P$  has a nonempty interior, this can always be achieved by translation of  $P$ .) We shall be interested in the distance of the origin to the boundary of  $K_P$ . The distance in the direction  $u \in S^{d-1}$  equals

$$R_P(u) = \inf \{r \geq 0: ru \notin K\}.$$

Define a functional  $\phi$  from the probability measures  $\mathcal{M}$  on  $\mathbb{R}^d$  to  $\ell^\infty(S^{d-1})$  by setting  $\phi(P)(u)$  equal to  $R_P(u)$ . Under suitable conditions, this map is Hadamard-differentiable.

The set  $K_P$  at the true  $P$  is assumed to be regular in the sense that it has a unique supporting hyperplane at each of its boundary points  $R_P(u)u$ . Let  $V_P(u)$  be the outward unit-normal vectors of these supporting hyperplanes.

**3.9.32 Lemma.** *Let  $P$  be a probability distribution on  $\mathbb{R}^d$  such that  $K_P$  is regular and the maps  $u \mapsto R_P(u)$  and  $u \mapsto V_P(u)$  are continuous. Furthermore, assume that  $u'X$  has a uniformly bounded density  $p(\cdot; u)$  such that  $p(z; V_P(u))$  is positive and continuous in  $u \in S^{d-1}$  and  $z$  near  $R_P(u)u'V_P(u)$ . Then  $\phi: \mathcal{M} \subset \ell^\infty(\mathcal{H}) \mapsto \ell^\infty(S^{d-1})$  is Hadamard-differentiable at  $P$  tangentially to  $\text{UC}(\mathcal{H}, \rho_P)$  with derivative given by*

$$\phi'_P(h)(u) = -\frac{h(H_P(u))}{u'V_P(u)p(u'V_P(u)R_P(u); V_P(u))},$$

for  $h \in \text{UC}(\mathcal{H}, \rho_P)$  and  $u \in S^{d-1}$ . Here  $H_P(u)$  is the half-space containing  $K_P$  whose boundary is equal to the supporting hyperplane at  $R_P(u)u$ .

**Proof.** Write  $H(u, r)$  for the half-space  $\{x: u'x \leq r\}$ . Let  $P_t = P + th_t$  be a sequence of probability measures viewed as elements of  $\ell^\infty(\mathcal{H})$  such that  $h_t \rightarrow h \in \text{UC}(\mathcal{H}, \rho_P)$ . Let  $K_t$  be the corresponding compact, convex sets, and define for every  $u \in S^{d-1}$

$$\begin{aligned} R_t(u) &= \inf\{r \geq 0: ru \notin K_t\}, \\ Q_t(u) &= \inf\{r \geq 0: P_t H(u, r) \geq 1 - \alpha\}. \end{aligned}$$

Furthermore, let  $V_t(u)$  be the outward unit-normal of some supporting hyperplane to  $K_t$  at the point  $R_t(u)u$ . Thus

$$K_t \subset \{x: V_t(u)'x \leq R_t(u)V_t(u)'u\} = H(V_t(u), R_t(u)V_t(u)'u).$$

For  $t = 0$ , the supporting hyperplanes are unique by assumption. If there are more candidates, then any one will do in the following argument. For simplicity of notation, we drop the index  $t$  if  $t = 0$ .

The half-space  $H(v, Q_t(v))$  is the minimal half-space in the direction  $v$  that contains  $K_t$ . Combining the fact that it contains  $K_t$  with the fact that  $R_t(u)u \in K_t$  immediately gives

$$R_t(u)v'u \leq Q_t(v), \quad \text{for every } u, v \in S^{d-1}.$$

For the normal  $v = V_t(u)$  of a supporting hyperplane at  $R_t(u)u$ , this inequality becomes the equality

$$R_t(u)V_t(u)'u = Q_t(V_t(u)), \quad \text{for every } u \in S^{d-1}.$$

Otherwise, the half-space  $H(v, Q_t(v))$  would not be minimal, in view of the fact that  $K_t \subset H(V_t(u), R_t(u)V_t(u)'u)$  by the definition of a supporting hyperplane.

The combination of the two preceding displays shows that

$$\frac{Q_t(V_t(u)) - Q(V_t(u))}{u'V_t(u)} \leq R_t(u) - R(u) \leq \frac{Q_t(V(u)) - Q(V(u))}{u'V(u)}.$$

The left and right sides divided by  $t$  will be shown to converge to the same limit.

First we show consistency of  $Q_t$  and  $V_t$ . The numbers  $Q_t(u)$  are the  $(1-\alpha)$ -quantiles of the random variables  $u'X_t$  if  $X_t$  is distributed according to  $P_t$ . The assumption  $P_t \rightarrow P$  in  $\ell^\infty(\mathcal{H})$  shows that the distribution functions of these variables converge uniformly, uniformly in  $u$ . A standard argument shows that  $Q_t \rightarrow Q$  in  $\ell^\infty(S^{d-1})$  (Problem 3.9.10).

Since the density  $p(z; V(u))$  of  $u'X$  is bounded away from zero for  $z$  close to  $Q(u)$ , uniformly in  $u$ , it follows that  $\sup_u PH(u, \varepsilon) < 1 - \alpha$  for  $\varepsilon < \inf_u Q(u)$ . Here  $\inf_u Q(u)$  is positive, because  $K$  contains an interior point. Conclude that for sufficiently small  $t$  we have  $\sup_u P_t H(u, \varepsilon) < 1 - \alpha$ , whence  $K_t$  contains the ball of radius  $\varepsilon$ . According to Problem 3.9.11,  $u'V_t(u)$  is bounded away from zero uniformly in  $u$ , for sufficiently small  $t$ . Conclude that the denominators in the preceding display are bounded away from zero and next that  $R_t \rightarrow R$  uniformly in  $u$ .

For the proof that  $V_t \rightarrow V$  in  $\ell^\infty(S^{d-1}, \mathbb{R}^d)$ , consider arbitrary sequences  $u_n$  and  $t_n \rightarrow 0$ . By compactness of  $S^{d-1}$ , there exists a subsequence along which  $u_n \rightarrow u$  and  $V_{t_n}(u_n) \rightarrow v$  for some  $u$  and  $v$  in  $S^{d-1}$ . By the definition of  $R$  and  $R_t$ , we have for every  $n$

$$\left( \inf_u \frac{R_{t_n}(u)}{R(u)} \right) K \subset K_{t_n} \subset H(V_{t_n}(u_n), R_{t_n}(u_n)V_{t_n}(u_n)'u_n).$$

The left side approaches  $K$  and the right side  $H(v, R(u)v'u)$ . Conclude that  $K \subset H(v, R(u)v'u)$ . By regularity, it follows that  $v = V(u)$ , and next by continuity of  $V$ , we have  $V_{t_n}(u_n) - V(u_n) \rightarrow 0$  along the subsequence.

The absolute value of the difference  $P_t H(v, Q_t(v)) - (1-\alpha)$  is bounded by the jump size of the distribution function of  $v'X_t$  at the point  $Q_t(v)$ . By assumption,  $v'X$  has no atoms and, since  $v'X$  has a bounded density and  $h$  is uniformly  $\rho_P$ -continuous,  $r \mapsto h(H(v, r))$  is continuous. Since  $P_t = P + th_t$  we can conclude that the difference is  $o(t)$ . Thus, uniformly in  $v \in S^{d-1}$ ,

$$\begin{aligned} o(t) &= P_t H(v, Q_t(v)) - PH(v, Q(v)) \\ &= PH(v, Q_t(v)) - PH(v, Q(v)) + th_t(H(v, Q_t(v))) \\ &= \left( p(Q(v); v) + o(1) \right) (Q_t(v) - Q(v)) + th(H(v, Q(v))) + o(t). \end{aligned}$$

This may be rewritten as

$$\frac{Q_t(v) - Q(v)}{t u' v} = \frac{-h(H(v, Q(v))) + o(1)}{u' v (p(Q(v); v) + o(1))}.$$

Apply this with  $v = V_t(u)$  and  $v = V(u)$  to complete the proof. ■

**3.9.33 Example.** Let  $X_1, \dots, X_n$  be an i.i.d. sample from the distribution  $P$  on  $\mathbb{R}^d$  with empirical distribution  $\mathbb{P}_n$ . The collection of half-spaces is a separable Vapnik-Červonenkis class and hence a Donsker class. Thus, the empirical process  $\sqrt{n}(\mathbb{P}_n - P)$  converges in distribution in  $\ell^\infty(\mathcal{H})$  to a tight Brownian  $\mathbb{G}_P$ . Define a random compact, convex set  $\mathbb{K}_n$  by

$$\mathbb{K}_n = \cap \{H \in \mathcal{H}: \mathbb{P}_n(H) \geq 1 - \alpha\}.$$

If  $P$  satisfies the hypotheses of the preceding theorem, then

$$\sqrt{n}(R_{\mathbb{P}_n} - R_P) \rightsquigarrow - \frac{\mathbb{G}_P(H_P(u))}{u' V_P(u) p(u' V_P(u) R_P(u); V_P(u))}, \quad \text{in } \ell^\infty(S^{d-1}).$$

The limit process is zero-mean Gaussian with covariance function

$$\frac{PH_P(u) \cap H_P(v) - (1 - \alpha)^2}{g(u)g(v)}, \quad u, v \in S^{d-1},$$

where  $g(u) = u' V_P(u) f(R(u) u' V(u); V(u))$ .

A *multivariate trimmed mean* of a probability measure  $P$  could be defined as

$$T(P) = \frac{1}{\lambda(K_P)} \int_{K_P} x d\lambda(x).$$

The empirical trimmed mean is obtained by replacing  $P$  by the empirical distribution  $\mathbb{P}_n$ . We can prove the asymptotic normality of this estimator, and more generally the asymptotic normality of the trimmed mean of estimators that converge in  $\ell^\infty(\mathcal{H})$ , by application of the preceding lemma. For this we write the trimmed mean in the form

$$T(P) = \frac{\int_{S^{d-1}} \int_0^{R_P(u)} r u J(u, r) dr du}{\int_{S^{d-1}} \int_0^{R_P(u)} J(u, r) dr du},$$

where  $J(u, r)$  is the Jacobian of the transformation  $x \mapsto (u, r)$  and  $S^{d-1}$  can be identified with an interval in  $\mathbb{R}^{d-1}$ . For instance, for  $d = 2$ , we can use polar coordinates and obtain

$$T(P) = \frac{\int_0^{2\pi} R_P(u)^3 / 3 du}{\int_0^{2\pi} R_P(u)^2 / 2 du}.$$

It is easy to see that the map  $R_P \mapsto T(P)$  is Hadamard-differentiable. Combining this with Lemma 3.9.32 and the chain rule, we see that the map  $P \mapsto T(P)$  is Hadamard-differentiable as well.

### 3.9.4.7 Z-functionals

Estimators defined by equations,  $Z$ -estimators, are considered in Chapter 3.2, and their asymptotic distribution may be established by Theorem 3.3.1. Solving an estimating equation is the same as assigning a zero to a random function. In this section it is shown that the functional that assigns a zero is Hadamard-differentiable. Thus, the delta-method may be used for deriving the asymptotic distribution of  $Z$ -estimators. This route is not necessarily recommended, however; while the packaging of the argument by Hadamard differentiability is elegant, it seems to require conditions that are stronger than those of the more direct Theorem 3.3.1.

Given an arbitrary subset  $\Theta$  of a Banach space and another Banach space  $\mathbb{L}$ , let  $\ell^\infty(\Theta, \mathbb{L})$  be the Banach space of all uniformly norm-bounded functions  $z: \Theta \mapsto \mathbb{L}$ . Let  $Z(\Theta, \mathbb{L})$  be the subset consisting of all maps with at least one zero. Let  $\phi: Z(\Theta, \mathbb{L}) \mapsto \Theta$  be a map that assigns one of its zeros  $\phi(z)$  to each element  $z \in Z(\Theta, \mathbb{L})$ . In case of multiple zeros the precise choice of a zero is irrelevant.

**3.9.34 Lemma.** *Assume  $\Psi: \Theta \mapsto \mathbb{L}$  is uniformly norm-bounded, is one-to-one, possesses a  $\theta_0$  and has an inverse (with domain  $\Psi(\Theta)$ ) that is continuous at 0. Let  $\Psi$  be Fréchet-differentiable at  $\theta_0$  with derivative  $\dot{\Psi}_{\theta_0}$ , which is one-to-one and continuously invertible on  $\text{lin } \Theta$ . Then the map  $\phi: Z(\Theta, \mathbb{L}) \subset \ell^\infty(\Theta, \mathbb{L}) \mapsto \Theta$  is Hadamard-differentiable at  $\Psi$  tangentially to the set of  $z \in \ell^\infty(\Theta, \mathbb{L})$  that are continuous at  $\theta_0$ . The derivative is given by  $\phi'_{\Psi}(z) = -\dot{\Psi}_{\theta_0}^{-1}(z(\theta_0))$ .*

**Proof.** Let  $z_t \rightarrow z$  uniformly on  $\Theta$  for a map  $z: \Theta \mapsto \mathbb{L}$  that is continuous at  $\theta_0$ . By definition the element  $\theta_t = \phi(\Psi + tz_t)$  satisfies  $\Psi(\theta_t) + tz_t(\theta_t) = 0$ . Conclude that  $\Psi(\theta_t) = O(t)$ . Since  $\Psi$  is one-to-one and has an inverse that is continuous at zero, it follows that  $\theta_t = \Psi^{-1}(\Psi(\theta_t)) \rightarrow \Psi^{-1}(0) = \theta_0$ . By the Fréchet differentiability of  $\Psi$ ,

$$\liminf_{t \downarrow 0} \frac{\|\Psi(\theta_t) - \Psi(\theta_0)\|}{\|\theta_t - \theta_0\|} \geq \liminf_{t \downarrow 0} \frac{\|\dot{\Psi}_{\theta_0}(\theta_t - \theta_0)\|}{\|\theta_t - \theta_0\|} \geq \inf_{\|h\|=1} \|\dot{\Psi}_{\theta_0}(h)\|,$$

where  $h$  ranges over  $\text{lin } \Theta$ . Since the inverse of  $\dot{\Psi}_{\theta_0}$  is continuous, the right side is positive. Conclude that there exists a positive constant  $c$  such that  $\|\theta_t - \theta_0\| < c\|\Psi(\theta_t) - \Psi(\theta_0)\| = c\|tz_t(\theta_t)\|$  for all sufficiently small  $t > 0$ . Conclude that  $\|\theta_t - \theta_0\| = O(t)$ . By Fréchet differentiability,

$$\Psi(\theta_t) - \Psi(\theta_0) = \dot{\Psi}_{\theta_0}(\theta_t - \theta_0) + o(\|\theta_t - \theta_0\|).$$

The remainder term is  $o(t)$ . Conclude that

$$\frac{\theta_t - \theta_0}{t} = \dot{\Psi}_{\theta_0}^{-1}\left(\frac{\Psi(\theta_t) - \Psi(\theta_0)}{t} + o(1)\right) \rightarrow \dot{\Psi}_{\theta_0}^{-1}(-z(\theta_0)),$$

since  $t^{-1}(\Psi(\theta_t) - \Psi(\theta_0)) = -z_t(\theta_t) \rightarrow z(\theta_0)$ . ■

**3.9.35 Example ( $Z$ -estimators).** Let  $\hat{\theta}_n$  be “estimators” solving the estimating equation  $\Psi_n(\theta) = 0$  for given random criterion functions  $\Psi_n: \Theta \mapsto \mathbb{L}$ . Assume that the sequence  $\sqrt{n}(\Psi_n - \Psi)$  converges in distribution to a process  $\mathbb{Z}$  in the space  $\ell^\infty(\Theta, \mathbb{L})$ . If  $\Psi$  satisfies the conditions of the preceding lemma and almost all sample paths  $\theta \mapsto \mathbb{Z}(\theta)$  are continuous at  $\theta_0$ , then the delta-method yields

$$\sqrt{n}(\hat{\theta}_n - \theta_0) \rightsquigarrow \phi'_\Psi(\mathbb{Z}) = -\dot{\Psi}_{\theta_0}^{-1}(\mathbb{Z}(\theta_0)).$$

This is the same conclusion as given by Theorem 3.3.1.

It should be noted that the uniformity in convergence with respect to  $\Theta$  together with the requirement that the limit process is continuous at  $\theta_0$  implies condition (3.3.2). Thus, Theorem 3.3.1 gives the same result under weaker conditions, except that consistency of  $\hat{\theta}_n$  is part of the assumptions.

For finite-dimensional  $\theta$ , the maps  $\Psi_n$  will usually take their values in Euclidean space. The weak convergence (in  $\ell^\infty(\Theta, \mathbb{L})$ ) needed for the preceding argument is then the weak convergence of each coordinate function in  $\ell^\infty(\Theta)$ . For  $\mathbb{L} = \ell^\infty(\mathcal{F})$  for some set  $\mathcal{F}$ , the space  $\ell^\infty(\Theta, \mathbb{L})$  can be isometrically identified with  $\ell^\infty(\Theta \times \mathcal{F})$  through  $z(\theta)f \leftrightarrow z(\theta, f)$ , and the weak convergence can be interpreted accordingly.

## Problems and Complements

- (Uniform Fréchet differentiability)** Let  $\mathbb{D}$  and  $\mathbb{E}$  be normed spaces and  $\mathbb{D}_\phi \subset \mathbb{D}$  convex. Let  $\phi: \mathbb{D}_\phi \mapsto \mathbb{E}$  be Fréchet-differentiable at every  $\theta \in \mathbb{D}_\phi$ , where the derivatives  $\theta \mapsto \phi'_\theta$  are uniformly norm-continuous. Then  $\phi$  is uniformly Fréchet-differentiable on  $\mathbb{D}_\phi$ .
- [Hint: The first part of the argument is the same as the argument in the proof of Lemma 3.9.7.]

- Arbitrary distribution functions  $F$  and  $G$  on the real line satisfy the identity

$$\iint [G(s \wedge t) - G(s)G(t)] dF(s)dF(t) = \int F^2 dG - (\int F dG)^2.$$

- If  $\mathbb{G}_G$  is a  $G$ -Brownian bridge process indexed by the half-lines  $(-\infty, t]$ , then

$$-\int \mathbb{G}_G dF \sim \mathbb{G}_G(F),$$

where the right side denotes a  $G$ -Brownian bridge process indexed by the single function  $F$ .

[Hint: Use the previous exercise.]

4. (**Quantile-quantile**) The *quantile-quantile transformation*  $\phi(F, G) = F \circ G^{-1}$  for distribution functions  $F$  and  $G$  on  $\mathbb{R}$  is Hadamard-differentiable tangentially to  $D[a, b] \times C[a, b]$  as a map  $D(\bar{\mathbb{R}})^2 \mapsto D[a, b]$  is Hadamard-differentiable tangentially to  $D[a, b] \times C[a, b]$ . This may be used to derive the limit distribution of  $\mathbb{F}_n \circ \mathbb{G}_n^{-1}$  for the empirical distribution functions  $\mathbb{F}_n$  and  $\mathbb{G}_n$  based on independent samples from  $F$  and  $G$ . [Dudley (1992) establishes Fréchet differentiability of this map  $\phi$  with other norms on the domain and range spaces.]

5. (**Volterra equation and Peano series**) The product integral  $B = \phi(A)$  is the unique solution of the *Volterra equation*

$$B(t) = 1 + \int_{(0,t]} B(s-) dA(s).$$

This may be used to write the product integral as the *Peano series*

$$B(t) = 1 + \sum_{n=1}^{\infty} \int \dots \int_{0 < t_1 < \dots < t_n \leq t} dA(t_1) \dots dA(t_n).$$

6. (**Forward and backward equation**) The product integral  $B = \phi(A)$  satisfies the forward equation

$$B(s, t] - 1 = \int_{(s,t]} B(s, u) dA(u)$$

and the backward equation

$$B(s, t] - 1 = \int_{(s,t]} dA(u) B(u, t].$$

7. (**Duhamel equation**) Product integrals  $B_1 = \phi(A_1)$  and  $B_2 = \phi(A_2)$  satisfy the *Duhamel equation*

$$B_2(s, t] - B_1(s, t] = \int_{(s,t]} B_1(s, u) B_2(u, t] (A_2 - A_1)(du).$$

[**Hint:** Define a Borel measure  $A_{1,2}$  on  $\mathbb{R}_+^{m+n}$  by

$$A_{1,2}(C_1 \times \dots \times C_m \times D_1 \times \dots \times D_n) = A_1(C_1) \dots A_1(C_m) A_2(D_1) \dots A_2(D_n).$$

Then integrate the indicator of the set  $\{s < u_1 < \dots < u_{m+n} \leq t\}$  with respect to  $A_{1,2}$ , apply Fubini's theorem, and sum both sides of the resulting equality over  $m \geq 1$  and  $n \geq 1$ .]

8. (**Integration by parts for the Duhamel equation**) Suppose that  $A, B \in D(\mathbb{R}^+)$  and that  $h \in D(\mathbb{R}^+)$  is of bounded variation. (When  $h$  is of unbounded

variation, the left side is *defined by* the right side!) Then

$$\begin{aligned}
 & \int_{(0,t]} \overline{\prod}((1 + dB)(0, u) dh(u)) \overline{\prod}((1 + dA)(u, t]) \\
 &= h(t) + \int_{(0,t]} \overline{\prod}((1 + dB)(0, s) dB(s)) [h(t) - h(s)] \\
 &\quad + \int_{(0,t]} h(r-) dA(r) \overline{\prod}((1 + A)(r, t]) \\
 &\quad + \int \int_{0 < s < r \leq t} \overline{\prod}((1 + dB)(0, s) dB(s)) [h(r-) - h(s)] \\
 &\quad \times \overline{\prod}((1 + A)(r, t]) dA(r).
 \end{aligned}$$

[Hint: Use the forward and backward equations to replace product integrals on the left side, and then use Fubini's theorem.]

- 9. (Existence of zeros)** Lemma 3.9.34 takes the domain of the zero-functional equal to the set of maps that have at least one zero. This avoids some technical problems. In many situations it can be shown that the domain is open by the following result. Let  $\Psi: \Theta \mapsto \mathbb{R}^p$  be a homeomorphism of a neighborhood of  $\theta_0 \in \mathbb{R}^p$  onto a neighborhood of  $0 \in \mathbb{R}^p$ . Then every continuous  $z: \Theta \mapsto \mathbb{R}^p$  for which  $\|z - \Psi\|_\infty$  is sufficiently small has at least one zero.

[Hint: Without loss of generality the neighborhood of 0 can be chosen to be the ball  $B(0, r)$ . Let  $G = \{z \in C(\Theta): \|z - \Psi\|_\infty < r\}$ . The function  $x \mapsto z \circ \Psi^{-1}(x) - x$  maps the ball  $B(0, r)$  into itself, since

$$\|z \circ \Psi^{-1}(x) - x\| \leq \|z - \Psi\|_\infty.$$

If  $z$  is continuous, then so is the map  $x \mapsto z \circ \Psi^{-1}(x) - x$ . By Brouwer's fixed-point theorem, it has a fixed point  $x_z$ . This satisfies  $z(\Psi^{-1}(x_z)) = 0$ .]

- 10.** Let  $x \mapsto F_n(x; u)$  be distribution functions that converge uniformly in  $x$  and  $u$  to distribution functions  $x \mapsto F(x; u)$ . If there exist points  $\xi_{pu}$  such that for every  $\delta > 0$

$$\sup_u F(\xi_{pu} - \delta; u) < p < \inf_u F(\xi_{pu} + \delta; u),$$

then  $F_n^{-1}(p; u) \rightarrow F^{-1}(p; u)$  uniformly in  $u$ . The condition is satisfied if every  $F(\cdot; u)$  has a density that is uniformly bounded away from zero in a neighborhood of its  $p$ th quantile  $\xi_{pu}$ .

- 11.** For a compact, convex set  $K \subset \mathbb{R}^d$ , let  $V(u)$  be the outward normal of a supporting hyperplane at the point  $R(u)u$ , where  $R(u) = \inf\{r: ru \notin K\}$ . If  $K$  contains the closed ball of radius  $\varepsilon$  around the origin, then  $u'V(u) \geq \varepsilon / \text{diam } K$  for every  $u \in S^{d-1}$ .

[Hint: The convex hull of the ball of radius  $\varepsilon$  and the point  $R(u)u$  is contained in  $K$ . Hence, by the definition of a supporting hyperplanes  $V(u)'(\frac{1}{2}R(u)u + \frac{1}{2}\varepsilon V(u)) \leq V(u)'R(u)u$ , since  $\varepsilon V(u)$  is contained in the ball of radius  $\varepsilon$ .]

12. In Lemma 3.9.32 the distance  $\rho_P$  may be replaced by  $d(H(u, r), H(v, s)) = \|u - v\| + |r - s|$ .

## 3.10

# Contiguity

Let  $P$  and  $Q$  be probability measures on a measurable space  $(\Omega, \mathcal{A})$ . If  $Q$  is absolutely continuous with respect to  $P$ , then the  $Q$ -law of a measurable map  $X: \Omega \mapsto \mathbb{D}$  can be calculated from the  $P$ -law of the pair  $(X, dQ/dP)$  through the formula

$$E_Q f(X) = E_P f(X) \frac{dQ}{dP}.$$

With  $M$  equal to the induced  $P$ -law of the pair  $(X, dQ/dP)$ , this identity can also be expressed as

$$Q(X \in B) = \int_{B \times \mathbb{R}} v \, dM(x, v).$$

The validity of these formulas depends essentially on the absolute continuity of  $Q$  with respect to  $P$ . Indeed, it is clear that no  $P$ -law contains information about the part of  $Q$  that is singular with respect to  $P$ .

Consider an asymptotic version of the problem. Let  $(\Omega_\alpha, \mathcal{A}_\alpha)$  be measurable spaces, each with a pair of probability measures  $P_\alpha$  and  $Q_\alpha$  on it. Under what conditions can a  $Q_\alpha$ -limit law of maps  $X_\alpha: \Omega_\alpha \mapsto \mathbb{D}$  be obtained from suitable  $P_\alpha$ -limit laws? In view of the above, it is necessary that  $Q_\alpha$  is asymptotically absolutely continuous with respect to  $P_\alpha$  in a suitable sense.

**3.10.1 Definition.** Let  $P_\alpha$  and  $Q_\alpha$  be nets of probability measures on measurable spaces  $(\Omega_\alpha, \mathcal{A}_\alpha)$ . Then  $Q_\alpha$  is *contiguous* with respect to  $P_\alpha$  if  $P_\alpha(A_\alpha) \rightarrow 0$  along a subnet implies  $Q_\alpha(A_\alpha) \rightarrow 0$  along this subnet for every choice of measurable sets  $A_\alpha$  (and every subnet). This is denoted

$Q_\alpha \triangleleft P_\alpha$ . The nets  $P_\alpha$  and  $Q_\alpha$  are *two-sided contiguous* if both  $P_\alpha \triangleleft Q_\alpha$  and  $Q_\alpha \triangleleft P_\alpha$ . This is denoted  $P_\alpha \bowtie Q_\alpha$ .

Contiguity can be characterized by the asymptotic behavior of the *likelihood ratios* of  $P_\alpha$  and  $Q_\alpha$ . Given two measures  $P$  and  $Q$ , write  $dQ/dP$  for the Radon-Nikodym density of the part of  $Q$  that is  $P$ -absolutely continuous. This function is only  $P$ -a.s. uniquely defined, but since only its  $P$ -law will be studied here, it is not necessary to remove the ambiguity. One way of forming  $dQ/dP$  is to take densities  $p$  and  $q$  of  $P$  and  $Q$  with respect to some measure for which there exists such densities (for instance, a  $\sigma$ -finite measure that dominates both, such as  $P + Q$ ) and set  $dQ/dP = q/p$ .

Loglikelihood ratios are always nonnegative and integrable. In particular, for nets of probability measures  $P_\alpha$  and  $Q_\alpha$ ,

$$\mathbb{E}_{Q_\alpha} \frac{dP_\alpha}{dQ_\alpha} \leq 1 \quad \text{and} \quad \mathbb{E}_{P_\alpha} \frac{dQ_\alpha}{dP_\alpha} \leq 1.$$

Thus, the nets of likelihood ratios  $dQ_\alpha/dP_\alpha$  and  $dP_\alpha/dQ_\alpha$  are uniformly tight as maps into  $[0, \infty)$  under  $Q_\alpha$  and  $P_\alpha$ , respectively. The properties of their limit points determine contiguity.

**3.10.2 Theorem.** Let  $P_\alpha$  and  $Q_\alpha$  be nets of probability measures on measurable spaces  $(\Omega_\alpha, \mathcal{A}_\alpha)$ . Then the following statements are equivalent:

- (i)  $Q_\alpha \triangleleft P_\alpha$ ;
- (ii) if  $dP_\alpha/dQ_\alpha \xrightarrow{Q_\alpha} L$  along a subnet, then  $L(0, \infty) = 1$ ;
- (iii) if  $dQ_\alpha/dP_\alpha \xrightarrow{P_\alpha} V$  along a subnet, then  $EV = 1$ ;
- (iv) for any  $T_\alpha: \Omega_\alpha \rightarrow \mathbb{R}$ , if  $T_\alpha \xrightarrow{P_\alpha} 0$  along a subnet, then  $T_\alpha \xrightarrow{Q_\alpha} 0$  along this subnet.

**Proof.** The equivalence of (i) and (iv) is an easy exercise.

(i)  $\Rightarrow$  (ii). Suppose, without loss of generality, that  $dP_\alpha/dQ_\alpha \xrightarrow{Q_\alpha} L$ . Then by the portmanteau theorem,  $\liminf Q_\alpha(dP_\alpha/dQ_\alpha < \varepsilon) \geq L[0, \varepsilon]$  for every  $\varepsilon > 0$ . For  $i \in \mathbb{N}$ , take  $\alpha_i$  such that, for all  $\alpha \geq \alpha_i$ ,

$$(3.10.3) \quad Q_\alpha \left( \frac{dP_\alpha}{dQ_\alpha} < \frac{1}{i} \right) \geq L[0, \frac{1}{2i}] - \frac{1}{i}.$$

Next, define for any  $\alpha$  a number  $\varepsilon_\alpha = \inf\{1/i: \alpha \geq \alpha_i\}$ . (Let the infimum over the empty set be 2.) Then  $\varepsilon_\alpha \leq 1/j$  for  $\alpha \geq \alpha_j$ ; hence the net  $\varepsilon_\alpha$  satisfies  $\varepsilon_\alpha \rightarrow 0$ . Moreover, for  $\alpha$  sufficiently large (already for  $\alpha \geq \alpha_1$ ), either  $\varepsilon_\alpha = 1/i$  for some  $i$  or  $\varepsilon_\alpha = 0$ . In the first case,  $\alpha \geq \alpha_i$ , so that by (3.10.3),

$$(3.10.4) \quad Q_\alpha \left( \frac{dP_\alpha}{dQ_\alpha} \leq \varepsilon_\alpha \right) \geq L[0, \frac{1}{2}\varepsilon_\alpha] - \varepsilon_\alpha.$$

In the second case, (3.10.3) holds for infinitely many  $i$ ; again conclude that (3.10.4) holds. Now

$$P_\alpha \left( \frac{dP_\alpha}{dQ_\alpha} \leq \varepsilon_\alpha \wedge dQ_\alpha > 0 \right) = \int_{dP_\alpha/dQ_\alpha \leq \varepsilon_\alpha} \frac{dP_\alpha}{dQ_\alpha} dQ_\alpha \leq \int \varepsilon_\alpha dQ_\alpha \rightarrow 0.$$

If  $Q_\alpha$  is contiguous with respect to  $P_\alpha$ , then the  $Q_\alpha$ -probability of the set on the left goes to zero too. Combination with (3.10.4) shows that  $L\{0\} = 0$ .

(iii)  $\Rightarrow$  (i). Suppose  $P_\alpha(A_\alpha) \rightarrow 0$  along some subnet. By Prohorov's theorem, every further subnet has a further subnet along which  $(dQ_\alpha/dP_\alpha, 1_{\Omega_\alpha - A_\alpha}) \rightsquigarrow (V, 1)$  under  $P_\alpha$  for some weak limit  $V$ . Along the subnet, by the portmanteau theorem,

$$\liminf Q_\alpha(\Omega_\alpha - A_\alpha) \geq \liminf \int_{\Omega_\alpha - A_\alpha} \frac{dQ_\alpha}{dP_\alpha} dP_\alpha \geq E1 \cdot V,$$

since the continuous function  $(v, t) \mapsto vt$  is bounded from below on  $[0, \infty) \times \{0, 1\}$ . If (iii) holds, then  $EV = 1$ , so  $Q_\alpha(A_\alpha) \rightarrow 0$  along the subnet.

(ii)  $\Leftrightarrow$  (iii). In view of Prohorov's theorem, given any subnet there is a further subnet such that along it

$$\frac{dP_\alpha}{dQ_\alpha} \xrightarrow{Q_\alpha} L \quad ; \quad \frac{dQ_\alpha}{dP_\alpha} \xrightarrow{P_\alpha} V \quad ; \quad \frac{dP_\alpha}{dP_\alpha + dQ_\alpha} \xrightarrow{P_\alpha+Q_\alpha} W,$$

for some weak limits  $L$ ,  $V$ , and  $W$ , where  $W$  is a random element of total mass 2 with  $EW = 1$ . Assume for simplicity that this statement is true for the whole net. It must be shown that  $L\{0\} = 0$  and  $EV = 1$  are equivalent. By the last of the three convergences and the continuous mapping theorem,

$$\left( \frac{dP_\alpha}{dQ_\alpha}, \frac{dQ_\alpha}{dP_\alpha}, \frac{dP_\alpha}{dP_\alpha + dQ_\alpha} \right) \xrightarrow{P_\alpha+Q_\alpha} \left( \frac{W}{1-W}, \frac{1-W}{W}, W \right),$$

in  $[0, \infty] \times [0, \infty] \times [0, 1]$  (with  $1/0 = \infty$ ; with this convention the functions  $w \mapsto w/(1-w)$  and  $w \mapsto (1-w)/w$  are continuous from  $[0, 1]$  into  $[0, \infty]$ ). Let  $f: [0, \infty] \mapsto \mathbb{R}$  be bounded and continuous. Then so is the function  $(x, v) \mapsto f(x)v$ , defined on  $[0, \infty] \times [0, 1]$ . Consequently,

$$E_{Q_\alpha} f \left( \frac{dP_\alpha}{dQ_\alpha} \right) = E_{P_\alpha+Q_\alpha} f \left( \frac{dP_\alpha}{dQ_\alpha} \right) \frac{dQ_\alpha}{dP_\alpha + dQ_\alpha} \rightarrow Ef \left( \frac{W}{1-W} \right) (1-W).$$

Conclude that  $L(A) = E1_A(W/(1-W))(1-W)$ ; so  $L\{0\} = P(W=0)$ . By a similar argument,  $Ef(V) = Ef((1-W)/W)W$ ; so  $EV = E((1-W)/W)W = E(1-W)1_{W>0} = 1 - P(W=0)$ . Hence the three statements  $L\{0\} = 0$ ,  $EV = 1$ , and  $P(W=0) = 0$  are equivalent. ■

Since subnets of sequences are not necessarily subsequences, the preceding theorem is unnecessarily complicated if the directed set is equal to the natural numbers. For sequences the theorem can be simplified to the following theorem.

**3.10.5 Theorem.** Let  $P_n$  and  $Q_n$  be sequences of probability measures on measurable spaces  $(\Omega_n, \mathcal{A}_n)$ . Then the following statements are equivalent:

- (i)  $Q_n \triangleleft P_n$ ;
- (ii) if  $dP_n/dQ_n \xrightarrow{Q_n} L$  along a subsequence, then  $L(0, \infty) = 1$ ;
- (iii) if  $dQ_n/dP_n \xrightarrow{P_n} V$  along a subsequence, then  $EV = 1$ ;
- (iv) for any  $T_n: \Omega_n \rightarrow \mathbb{R}$ , if  $T_n \xrightarrow{P_n} 0$ , then  $T_n \xrightarrow{Q_n} 0$ ;
- (v) for any  $A_n$  in  $\mathcal{A}_n$ : if  $P_n(A_n) \rightarrow 0$ , then  $Q_n(A_n) \rightarrow 0$ .

**Proof.** The equivalence of (ii) through (v) can be shown by the same arguments as before. Furthermore, in every metric space a limit point of a sequence (possibly along a subnet) is always also the limit of a subsequence. ■

**3.10.6 Example (Lognormality).** Let  $P_\alpha$  and  $Q_\alpha$  be probability measures on arbitrary measurable spaces. Suppose

$$\frac{dP_\alpha}{dQ_\alpha} \xrightarrow{Q_\alpha} e^{N(\mu, \sigma^2)}.$$

Then  $Q_\alpha \triangleleft P_\alpha$  and  $Q_\alpha \bowtie P_\alpha$  if and only if  $\mu = -\frac{1}{2}\sigma^2$ . For the last it suffices to note that  $E \exp\{N(\mu, \sigma^2)\} = 1$  if and only if  $\mu = -\frac{1}{2}\sigma^2$ .

The following theorem solves the problem posed in the introduction. It is a version of *Le Cam's third lemma*.

**3.10.7 Theorem (Le Cam's third lemma).** Let  $P_\alpha$  and  $Q_\alpha$  be nets of probability measures on measurable spaces  $(\Omega_\alpha, \mathcal{A}_\alpha)$  and  $X_\alpha: \Omega_\alpha \mapsto \mathbb{D}$  maps with values in a metric space. Let  $Q_\alpha \triangleleft P_\alpha$  and

$$(X_\alpha, \frac{dQ_\alpha}{dP_\alpha}) \xrightarrow{P_\alpha} (X, V).$$

Then  $L(B) = E1_B(X)V$  defines a probability measure and  $X_\alpha \xrightarrow{Q_\alpha} L$ . If  $X$  is tight or separable, then so is  $L$ .

**Proof.** Clearly  $V \geq 0$  and by contiguity  $EV = 1$ ; so  $L$  is a probability measure. It is immediate from the definition of  $L$  that  $\int f dL = Ef(X)V$  for every indicator function or nonnegative  $f$ . Conclude, in steps, that the same is true for every simple function  $f$  and any integrable measurable function.

If  $f: \mathbb{D} \mapsto \mathbb{R}$  is lower semicontinuous and nonnegative, then so is the function  $(t, v) \mapsto f(t)v$  on  $\mathbb{D} \times [0, \infty)$ . Thus by the portmanteau theorem,

$$\liminf E_{Q_\alpha,*} f(X_\alpha) \geq \liminf \int f(X_\alpha)_* \frac{dQ_\alpha}{dP_\alpha} dP_\alpha \geq Ef(X)V.$$

Finally, apply the portmanteau theorem a second time, in the converse direction.

The last statement of the theorem follows immediately from the representation of  $L$ : if  $X$  concentrates on  $S$ , so does  $L$ . ■

**3.10.8 Example (Le Cam's third lemma).** The name *Le Cam's third lemma* is often reserved for the following special case of the previous theorem. Suppose

$$\left( X_\alpha, \log \frac{dQ_\alpha}{dP_\alpha} \right) \xrightarrow{P_\alpha} N_{k+1} \left( \begin{pmatrix} \mu \\ -\frac{1}{2}\sigma^2 \end{pmatrix}, \begin{pmatrix} \Sigma & \tau \\ \tau' & \sigma^2 \end{pmatrix} \right).$$

Then  $X_\alpha \xrightarrow{Q_\alpha} N_k(\mu + \tau, \Sigma)$ . To see this, let  $(X, W)$  have the given  $(k+1)$ -dimensional normal distribution. Then  $\int e^{it'x} dL(x) = Ee^{it'X} e^W$  equals the characteristic function of this normal distribution at the point  $(t, -i)$ , that is,

$$\int e^{it'x} dL(x) = e^{it'\mu - \frac{1}{2}\sigma^2 - \frac{1}{2}(t', -i)} \begin{pmatrix} \Sigma & \tau \\ \tau' & \sigma^2 \end{pmatrix} \begin{pmatrix} t \\ -i \end{pmatrix} = e^{it'(\mu + \tau) - \frac{1}{2}t'\Sigma t}.$$

Two-sided contiguity can of course be characterized by “doubling” any of the characterizations of Theorem 3.10.2, but it is customary to write a characterization in terms of the logarithms of the likelihood ratios. If it is understood that  $\log 0 = -\infty$ , then

$$\log \frac{dP_\alpha}{dQ_\alpha} : \mathcal{X}_\alpha \mapsto \bar{\mathbb{R}}$$

is a well-defined map, although it is only  $Q_\alpha$ -a.s. unique. Similarly,

$$\log \frac{dQ_\alpha}{dP_\alpha} : \mathcal{X}_\alpha \mapsto \bar{\mathbb{R}}$$

is  $P_\alpha$ -a.s. uniquely defined. We are interested only in the “laws” of these maps under  $Q_\alpha$  and  $P_\alpha$ , respectively, so for now the ambiguity of definition need not be removed. The extended logarithm  $x \mapsto \log x$  is a continuous bijection from  $[0, \infty)$  onto  $[-\infty, \infty)$ , with a continuous inverse. Consequently, a likelihood ratio converges weakly in  $[0, \infty)$  if and only if a loglikelihood ratio converges weakly in  $[-\infty, \infty)$  and the limits are related through taking logarithms and exponentials. This leads to the following reformulation of Theorem 3.10.2.

**3.10.9 Theorem.** Let  $P_\alpha$  and  $Q_\alpha$  be nets of probability measures on measurable spaces  $(\Omega_\alpha, \mathcal{A}_\alpha)$ . Then the following statements are equivalent:

- (i)  $Q_\alpha \not\Phi P_\alpha$ ;
- (ii) if  $\log dP_\alpha/dQ_\alpha \xrightarrow{Q_\alpha} L$  and  $\log dQ_\alpha/dP_\alpha \xrightarrow{P_\alpha} M$  in  $\bar{\mathbb{R}}$  along a subnet, then  $L(\mathbb{R}) = M(\mathbb{R}) = 1$ ;
- (iii) if  $\log dP_\alpha/dQ_\alpha \xrightarrow{Q_\alpha} L$  in  $\bar{\mathbb{R}}$  along a subnet, then  $L(\mathbb{R}) = 1$  and  $\int e^x dL(x) = 1$ .

**Proof.** Use characterization (ii) of Theorem 3.10.2 to get equivalence of the three statements

$Q_\alpha \triangleleft P_\alpha$ ,  
if  $dP_\alpha/dQ_\alpha \stackrel{Q_\alpha}{\rightsquigarrow} V$  along a subnet, then  $P(V \in (0, \infty)) = 1$ ,  
if  $\log dP_\alpha/dQ_\alpha \stackrel{Q_\alpha}{\rightsquigarrow} W$ , then  $P(W \in \mathbb{R}) = 1$ .

Next use characterization (iii) to obtain the equivalence of

$P_\alpha \triangleleft Q_\alpha$ ,  
if  $dP_\alpha/dQ_\alpha \stackrel{Q_\alpha}{\rightsquigarrow} V$  along a subnet, then  $EV = 1$ ,  
if  $\log dP_\alpha/dQ_\alpha \stackrel{Q_\alpha}{\rightsquigarrow} W$ , then  $Ee^W = 1$ .

Together this gives the equivalence of (i) and (iii) of the present theorem.

To obtain the equivalence of (i) and (ii), apply characterization (ii) of Theorem 3.10.2 twice, to both  $P_\alpha \triangleleft Q_\alpha$  and  $Q_\alpha \triangleleft P_\alpha$ . ■

### 3.10.1 The Empirical Process

For each  $n$ , let  $X_{n1}, \dots, X_{nn}$  be i.i.d. random elements in a measurable space  $(\mathcal{X}, \mathcal{A})$ . Under the “null hypothesis”, the common law is a fixed measure  $P$ , whereas under the “alternative hypothesis”, the common law is  $P_n$ . Set  $\mathbb{P}_n = n^{-1} \sum_{i=1}^n \delta_{X_{ni}}$ . In many studies of the asymptotic efficiency of estimators and tests, it is of interest to study the behavior of statistics based on  $X_{n1}, \dots, X_{nn}$  under the assumption that  $P_n$  converges to  $P$  in the sense that

$$(3.10.10) \quad \int \left[ \sqrt{n} (dP_n^{1/2} - dP^{1/2}) - \frac{1}{2} h dP^{1/2} \right]^2 \rightarrow 0,$$

for some measurable function  $h: \mathcal{X} \mapsto \mathbb{R}$ . In this section we derive the asymptotic distribution of the empirical process  $\sqrt{n}(\mathbb{P}_n - P)$  under both the null and alternative hypotheses.

Sequences  $P_n$  as above are often referred to as “contiguous alternatives” to  $P$ . A more precise statement is that the sequence of distributions  $P_n^n$  and  $P^n$  (of  $(X_{n1}, \dots, X_{nn})$  on  $\mathcal{X}^n$ ) are contiguous. In fact, the sequence of loglikelihood ratios allows a linear expansion, and contiguity follows from the central limit theorem and Example 3.10.6.

**3.10.11 Lemma.** *Let the sequence of probability measures  $P_n$  satisfy (3.10.10). Then necessarily  $Ph = 0$  and  $Ph^2 < \infty$  and*

$$\sum_{i=1}^n \log \frac{dP_n}{dP}(X_{ni}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n h(X_{ni}) - \frac{1}{2} Ph^2 + R_n,$$

where the sequence  $R_n$  converges to zero in probability under both  $P^n$  and  $P_n^n$ .

**Proof.** See, e.g., Hájek and Šidák (1967), Chapter VI, Section 1, pages 201–209; Bickel, Klaassen, Ritov, and Wellner (1993), Appendix A.9, page 498–513; or Van der Vaart (1988), Proposition A.8, page 182. ■

Let  $\xrightarrow{P}$  denote weak convergence under the assumption that  $(X_{n1}, \dots, X_{nn})$  are distributed according to  $P^n$ . The linear expansion of the loglikelihood ratio  $\Lambda_n(P_n, P) = \log dP_n^n/dP^n$  given by the preceding lemma combined with Slutsky's lemma and the multivariate central limit theorem gives, for any  $f$  with  $Pf^2 < \infty$ ,

$$\left( \sqrt{n}(\mathbb{P}_n - P)f, \Lambda_n(P_n, P) \right) \xrightarrow{P} N \left( \begin{pmatrix} 0 \\ -\frac{1}{2}Ph^2 \end{pmatrix}, \begin{pmatrix} P(f - Pf)^2 & Pfh \\ Pfh & Ph^2 \end{pmatrix} \right).$$

According to Le Cam's third lemma, this implies that

$$\sqrt{n}(\mathbb{P}_n - P)f \xrightarrow{P} N(Pfh, P(f - Pf)^2).$$

It follows that the naturally centered empirical process  $\sqrt{n}(\mathbb{P}_n - P_n)f$  converges marginally under  $P_n$  if and only if the deterministic sequence  $\sqrt{n}(\mathbb{P}_n - P)f$  converges to a limit. This is not necessarily true under (3.10.10), but for instance if  $\|P_n f^2\|_{\mathcal{F}} = O(1)$ , then the sequence  $\sqrt{n}(\mathbb{P}_n - P)f$  converges to  $Pfh$ . In that case  $\sqrt{n}(\mathbb{P}_n - P_n)f$  converges under  $P_n$  in distribution to a  $N(0, P(f - Pf)^2)$ -distribution, the same limit distribution as  $\sqrt{n}(\mathbb{P}_n - P)f$  under  $P$ , as expected. The following theorem records the uniform version of this result.

**3.10.12 Theorem.** *Let  $\mathcal{F}$  be a  $P$ -Donsker class of measurable functions with  $\|P\|_{\mathcal{F}} < \infty$ . If the sequence of probability measures  $P_n$  satisfies (3.10.10), then the sequence  $\sqrt{n}(\mathbb{P}_n - P)$  converges under  $P_n$  in distribution in  $\ell^\infty(\mathcal{F})$  to the process  $f \mapsto \mathbb{G}(f) + Pfh$ , where  $\mathbb{G}$  is a tight Brownian bridge. Moreover, if  $\|P_n f^2\|_{\mathcal{F}} = O(1)$ , then  $\|\sqrt{n}(\mathbb{P}_n - P)f - Pfh\|_{\mathcal{F}} \rightarrow 0$  and the sequence  $\sqrt{n}(\mathbb{P}_n - P_n)$  converges under  $P_n$  in distribution to  $\mathbb{G}$ .*

**Proof.** According to the discussion preceding the theorem, the sequence  $\sqrt{n}(\mathbb{P}_n - P)$  converges under  $P_n$  marginally to the process  $f \mapsto \mathbb{G}(f) + Pfh$ . Since the two marginal sequences of  $(\sqrt{n}(\mathbb{P}_n - P), \Lambda_n(P_n, P))$  are asymptotically tight under  $P$  in  $\ell^\infty(\mathcal{F})$  and  $\mathbb{R}$ , respectively, this sequence is jointly asymptotically tight in  $\ell^\infty(\mathcal{F}) \times \mathbb{R}$  under  $P$ . By Le Cam's third lemma and the converse part of Prohorov's theorem, the sequence  $\sqrt{n}(\mathbb{P}_n - P)$  is asymptotically tight in  $\ell^\infty(\mathcal{F})$  under  $P_n$ . Conclude that  $\sqrt{n}(\mathbb{P}_n - P)$  converges under  $P_n$  in distribution to the process  $f \mapsto \mathbb{G}(f) + Pfh$ .

The proof is complete if it is shown that the sequence  $\sqrt{n}(\mathbb{P}_n - P)f - Pfh$  converges to zero uniformly in  $f$ . For any  $f$ ,

$$\begin{aligned} \sqrt{n}(\mathbb{P}_n - P)f - Pfh &= \frac{1}{2} \int f h dP^{1/2} (dP_n^{1/2} - dP^{1/2}) \\ &\quad + \int f \left[ \sqrt{n}(dP_n^{1/2} - dP^{1/2}) - \frac{1}{2} h dP^{1/2} \right] [dP_n^{1/2} + dP^{1/2}]. \end{aligned}$$

By the Cauchy-Schwarz inequality, the second term on the right converges to zero uniformly in  $f$ , because  $\sup_f \int f^2 [dP_n^{1/2} + dP^{1/2}]^2 = O(1)$  by assumption. The first term is bounded by

$$\frac{1}{2} M \int h dP^{1/2} |dP_n^{1/2} - dP^{1/2}| + \frac{1}{2} \left( \int_{F > M} h^2 dP \right)^{1/2} \left( \int f^2 (dP_n + dP) \right)^{1/2}.$$

This converges to zero for  $M = M_n \rightarrow \infty$  sufficiently slowly. ■

**3.10.13 Example (Power of the Kolmogorov-Smirnov test).** A test of the null hypothesis  $H_0: P = P_0$ , that  $X_1, \dots, X_n$  are i.i.d. according to a fixed probability measure  $P_0$ , can be based on the (generalized) Kolmogorov-Smirnov statistic  $\|\mathbb{P}_n - P_0\|_{\mathcal{F}}$  for a given  $P_0$ -Donsker class  $\mathcal{F}$ . If  $c_\alpha$  is the upper  $\alpha$ -quantile of the distribution of the norm  $\|\mathbb{G}\|_{\mathcal{F}}$  of a tight Brownian bridge  $\mathbb{G}$ , then the test that rejects the null hypothesis if  $\sqrt{n} \|\mathbb{P}_n - P_0\|_{\mathcal{F}} > c_\alpha$  has asymptotic level  $\alpha$ . The sequence of tests is consistent against every alternative  $P$  such that  $\mathcal{F}$  is  $P$ -Donsker and  $\|P - P_0\|_{\mathcal{F}} > 0$ . The power against a sequence of alternatives  $P_n$  satisfying (3.10.10) equals

$$P_{P_n}^* (\sqrt{n} \|\mathbb{P}_n - P_0\|_{\mathcal{F}} > c_\alpha) \rightarrow P(\|\mathbb{G}(f) + P_0 f h\|_{\mathcal{F}} > c_\alpha).$$

Unfortunately, the last expression can rarely be evaluated explicitly.

### 3.10.2 Change-Point Alternatives

For each  $n$ , let  $X_{n1}, \dots, X_{nn}$  be independent random elements in a measurable space  $(\mathcal{X}, \mathcal{A})$ . Under the “null hypothesis”, the random elements are i.i.d. with common law  $P$ , whereas under the alternative hypothesis,  $X_{ni}$  has law  $P_{ni}$ . Suppose that the marginal laws  $P_{ni}$  converge to the measure  $P$  in the following sense:

$$(3.10.14) \quad n^{-1} \sum_{i=1}^n \int \left[ \sqrt{n} (dP_{ni}^{1/2} - dP^{1/2}) - \frac{1}{2} h\left(\cdot, \frac{i}{n}\right) dP^{1/2} \right]^2 \rightarrow 0,$$

for a function  $h \in L_2(\mathcal{X} \times [0, 1], P \times \lambda)$  such that  $\int h(x, t) dP(x) = 0$  for every  $t \in [0, 1]$ . Assume that the functions  $t \mapsto h(x, t)$  are continuous in the sense that the sequence of discretizations  $h_n(x, t) = \sum_{i=1}^n h(x, i/n) 1_{((i-1)/n, i/n]}(t)$  converges in second mean to  $h$ :

$$(3.10.15) \quad (P \times \lambda)(h_n - h)^2 \rightarrow 0.$$

Under these assumptions, the following extension of the local asymptotic normality lemma, Lemma 3.10.11, holds.

**3.10.16 Lemma.** Let the sequence of probability measures  $P_n = P_{n1} \times \dots \times P_{nn}$  satisfy (3.10.14) and (3.10.15). Then

$$\sum_{i=1}^n \log \frac{dP_{ni}}{dP}(X_{ni}) = \frac{1}{\sqrt{n}} \sum_{i=1}^n h\left(X_{ni}, \frac{i}{n}\right) - \frac{1}{2}(P \times \lambda)h^2 + R_n,$$

where the sequence  $R_n$  converges to zero in probability under both  $P^n$  and  $P_n$  and the sequence  $n^{-1/2} \sum_{i=1}^n h(X_{ni}, i/n)$  converges under  $P^n$  in distribution to a normal distribution with mean zero and variance  $(P \times \lambda)h^2$ .

**Proof.** Condition (3.10.15) implies the convergence of variances as well as the Lindeberg condition for the tangent vectors  $h(\cdot, i/n)$ :

$$(P \times \lambda)(h_n^2) = \frac{1}{n} \sum_{i=1}^n Ph^2\left(\cdot, \frac{i}{n}\right) \rightarrow (P \times \lambda)h^2$$

$$\frac{1}{n} \sum_{i=1}^n Ph^2\left(\cdot, \frac{i}{n}\right) \left\{ \left| h\left(\cdot, \frac{i}{n}\right) \right| \geq \varepsilon \sqrt{n} \right\} \rightarrow 0, \quad \text{for every } \varepsilon > 0.$$

Next the lemma follows from, for instance, Proposition A.8, page 182, Van der Vaart (1988); or Hájek and Šidák (1967), Chapter VI, Section 1, pages 201–209; or Bickel, Klaassen, Ritov, and Wellner (1993), Appendix A.9, page 498–513. ■

Let  $\mathbb{Z}_n$  be the sequential empirical process  $n^{-1/2} \sum_{i=1}^{\lfloor ns \rfloor} (f(X_{ni}) - Pf)$  under the null hypothesis. The linear expansion of the loglikelihood ratio  $\Lambda_n(P_n, P^n) = \log dP_n/dP^n$  given by the preceding lemma combined with Slutsky's lemma and the Lindeberg-Feller central limit theorem gives, for any  $s$  and any  $f$  with  $Pf^2 < \infty$ ,

$$(\mathbb{Z}_n(s, f), \Lambda_n(P_n, P^n)) \xrightarrow{P_n} N\left(\begin{pmatrix} 0 \\ -\frac{1}{2}\sigma^2 \end{pmatrix}, \begin{pmatrix} sP(f - Pf)^2 & c(s, f) \\ c(s, f) & (P \times \lambda)h^2 \end{pmatrix}\right).$$

Here  $c(s, f)$  is the asymptotic covariance of  $\mathbb{Z}_n$  and the log likelihood ratios and is given by

$$c(s, f) = (P \times \lambda)fh[0, s] = \int_0^s \int_{\mathcal{X}} f(x)h(x, t) dP(x) dt.$$

In view of Le Cam's third lemma, this implies that under the alternative hypothesis, for every  $(s, f)$  in  $[0, 1] \times \mathcal{F}$ ,

$$\mathbb{Z}_n(s, f) \xrightarrow{P_n} N(c(s, f), sP(f - Pf)^2).$$

The following theorem yields the corresponding conclusion for the whole process  $\mathbb{Z}_n$  in  $\ell^\infty([0, 1] \times \mathcal{F})$ .

**3.10.17 Theorem.** Let  $\mathcal{F}$  be a  $P$ -Donsker class of measurable functions with  $\|P\|_{\mathcal{F}} < \infty$ . If the triangular array of probability measures  $P_{ni}$  satisfies (3.10.14) and (3.10.15), then the sequence  $\mathbb{Z}_n$  converges weakly in  $\ell^\infty([0, 1] \times \mathcal{F})$  to the process

$$(s, f) \mapsto \mathbb{Z}(s, f) + (P \times \lambda)fh[0, s],$$

where  $\mathbb{Z}$  is a tight  $P$ -Kiefer-Müller process. If, in addition, the sequence of average measures  $\bar{P}_n = n^{-1} \sum_1^n P_{ni}$  satisfies  $\|\bar{P}_n f^2\|_{\mathcal{F}} = O(1)$ , then the sequence of processes  $n^{-1/2} \sum_{i=1}^{\lfloor ns \rfloor} (\delta_{X_{ni}} - P_{ni})f$  converges in distribution under  $P_n = P_{n1} \times \dots \times P_{nn}$  in  $\ell^\infty([0, 1] \times \mathcal{F})$  to  $\mathbb{Z}$ .

**Proof.** The first assertion follows by combining Prohorov's theorem and Le Cam's third lemma as in the proof of Theorem 3.10.12.

For the second assertion it suffices to show that

$$\left\| \frac{1}{\sqrt{n}} \sum_{i=1}^{\lfloor ns \rfloor} (P_{ni} - P)f - (P \times \lambda)fh[0, s] \right\|_{[0,1] \times \mathcal{F}} \rightarrow 0.$$

By extension of the argument in the proof of Theorem 3.10.12 the sequence  $n^{-1} \sum_{i=1}^{\lfloor ns \rfloor} [\sqrt{n}(P_{ni} - P)f - Pfh(\cdot, i/n)]$  converges to zero uniformly in  $(s, f)$ . Furthermore, since  $|\int_0^s h_n(x, t) dt - n^{-1} \sum_{i=1}^{\lfloor ns \rfloor} h(x, i/n)|$  is bounded above by  $n^{-1}|h(x, (\lfloor ns \rfloor + 1)/n)|$ , it follows that for every  $(s, f)$

$$\begin{aligned} & \left| \frac{1}{n} \sum_{i=1}^{\lfloor ns \rfloor} Pfh\left(\cdot, \frac{i}{n}\right) - (P \times \lambda)fh1_{[0,s]} \right| \\ & \leq \frac{1}{n} P \sup_i \left| h\left(\cdot, \frac{i}{n}\right) f \right| + (P \times \lambda) |fh_n - fh|. \end{aligned}$$

In view of condition (3.10.15), the sequence  $n^{-1} \sum_{i=1}^n Ph^2(\cdot, i/n) = (P \times \lambda)h_n^2$  is bounded. Two applications of the Cauchy-Schwarz inequality show that the expressions in the last display converge to zero uniformly in  $(s, f)$ . ■

**3.10.18 Example (Change-point alternatives).** Let  $\{P_n\}$  be a given sequence that satisfies (3.10.10), and let  $t \in (0, 1)$  be fixed. Define  $P_{ni}$  to be  $P$  for  $1 \leq i \leq \lfloor nt \rfloor$  and define it to be  $P_n$  for  $\lfloor nt \rfloor < i \leq n$ . Thus the distribution governing the observations changes from  $P$  to  $P_n$  between  $i = \lfloor nt \rfloor$  and  $i = \lfloor nt \rfloor + 1$ .

These "change-point alternatives" satisfy the conditions of the preceding theorem, with  $h(x, s) = h(x)1\{s > t\}$ .

## Problems and Complements

- 1. (Contiguity and limit experiments)** Let  $P_\alpha$  and  $Q_\alpha$  be nets of probability measures on measurable spaces  $(\mathcal{X}_\alpha, \mathcal{B}_\alpha)$ . There exist probability measures  $P$  and  $Q$  on a measurable space  $(\mathcal{X}, \mathcal{B})$  and a subnet such that  $dP_\alpha/dQ_\alpha \rightsquigarrow dP/dQ$  under  $Q_\alpha$  and  $dQ_\alpha/dP_\alpha \rightsquigarrow dQ/dP$  under  $P_\alpha$  along the subnet. (Here the laws of the limits are computed under  $Q$  and  $P$ , respectively.) Contiguity  $Q_\alpha \triangleleft P_\alpha$  holds if and only if for every such subnet and pairs  $P$  and  $Q$  one has  $Q \ll P$ . For sequences this is already true with subsequence replacing subnet.

[**Hint:** Let  $W$  be a weak limit point of the maps  $dP_\alpha/(dP_\alpha + dQ_\alpha) : \mathcal{X}_\alpha \mapsto [0, 1]$  under  $P_\alpha + Q_\alpha$ . ( $W$  has total mass 2.) By the continuous mapping theorem applied in  $\bar{\mathbb{R}}$ , one has  $dP_\alpha/dQ_\alpha \rightsquigarrow W/(1 - W)$  and  $dQ_\alpha/dP_\alpha \rightsquigarrow (1 - W)/W$  (take  $1/0 = \infty$ ). On  $\mathcal{X} = [0, 1]$  with Borel sets, define  $P(A) = \text{E}1_A(W)W$  and  $Q(A) = \text{E}1_A(W)(1 - W)$ . Then  $(P + Q)(A) = \text{E}1_A(W)$ , so that  $dP/dQ(w) = w/(1 - w)$  and  $dQ/dP(w) = (1 - w)/w$ . The second assertion follows from the first, the characterization of contiguity in terms of likelihood ratios and the equivalence of the three statements:  $Q \ll P$ ;  $Q(dP/dQ = 0) = 0$ ; and  $\text{E}_P dQ/dP = 1$ .]

- 2.** Let  $P_\alpha$  and  $Q_\alpha$  be nets of probability measures on measurable spaces  $(\Omega_\alpha, \mathcal{A}_\alpha)$  such that  $P_\alpha(A_\alpha) \rightarrow 0$  implies  $Q_\alpha(A_\alpha) \rightarrow 0$  for every choice of measurable sets  $A_\alpha$ . Then not necessarily  $Q_\alpha \triangleleft P_\alpha$ .

[**Hint:** Subnets are tricky. See Le Cam (1986), page 87.]

- 3. (Conditions that imply the Lindeberg-Feller condition)**

Let  $A_n(s) = n^{-1} \sum_{i=1}^{\lfloor ns \rfloor} h^2(X_{ni}, i/n)$  and  $A(s) = \int_0^s Ph^2(\cdot, t)dt$ .

- (i) If  $A_n(1) \xrightarrow{P} A(1)$  and  $h_n \rightarrow h$  in  $P \times \lambda$ -measure, then the Lindeberg condition holds.
- (ii) If  $A_n(s) \xrightarrow{P} A(s)$  for each  $0 \leq s \leq 1$  and  $h_n \rightarrow h$  in  $L_2(P \times \lambda)$ , then the Lindeberg condition holds.

[**Hint:** Greenwood and Shirayev (1985), remark 1, page 116.]

- 4.** Use Theorem 3.10.12 to investigate the local asymptotic power of the bootstrap and permutation independence tests defined in Chapter 3.8.

- 5.** Suppose that  $X_1, \dots, X_n$  are i.i.d. with distribution function  $F$  on  $[0, 1]$ . Let  $F_0$  be the Uniform  $[0, 1]$  distribution function. Consider testing the null hypothesis  $F = F_0$  versus the alternative that  $F(x) \geq F_0(x)$  for all  $x$  with strict inequality for some  $x$ , based on the statistics

$$A_n^+ = \int_0^1 \sqrt{n} [\mathbb{F}_n(t) - F_0(t)]^+ dt \quad \text{and} \quad B_n = \int_0^1 \sqrt{n} [\mathbb{F}_n(t) - F_0(t)] dt.$$

Investigate the local asymptotic power of these two tests.

## 3.11

# Convolution and Minimax Theorems

Let  $H$  be a linear subspace of a Hilbert space with inner product  $\langle \cdot, \cdot \rangle$  and norm  $\| \cdot \|$ . For each  $n \in \mathbb{N}$  and  $h \in H$ , let  $P_{n,h}$  be a probability measure on a measurable space  $(\mathcal{X}_n, \mathcal{A}_n)$ . Consider the problem of estimating a “parameter”  $\kappa_n(h)$  given an “observation”  $X_n$  with law  $P_{n,h}$ . The convolution theorem and the minimax theorem give a lower bound on how well  $\kappa_n(h)$  can be estimated asymptotically as  $n \rightarrow \infty$ . Suppose the sequence of statistical experiments  $(\mathcal{X}_n, \mathcal{A}_n, P_{n,h}: h \in H)$  is “asymptotically normal” and the sequence of parameters is “regular”. Then the limit distribution of every “regular” estimator sequence is the convolution of a certain Gaussian distribution and a noise factor. Furthermore, the maximum risk of any estimator sequence is bounded below by the “risk” of this Gaussian distribution. These concepts are defined as follows.

For ease of notation, let  $\{\Delta_h: h \in H\}$  be the “iso-Gaussian process” with zero mean and covariance function  $E\Delta_{h_1}\Delta_{h_2} = \langle h_1, h_2 \rangle$ . The sequence of experiments  $(\mathcal{X}_n, \mathcal{A}_n, P_{n,h}: h \in H)$  is called *asymptotically (shift) normal* if

$$\log \frac{dP_{n,h}}{dP_{n,0}} = \Delta_{n,h} - \frac{1}{2}\|h\|^2,$$

for stochastic processes  $\{\Delta_{n,h}: h \in H\}$  such that

$$\Delta_{n,h} \xrightarrow{^0} \Delta_h \quad \text{marginally.}$$

Here  $\xrightarrow{^h}$  denotes weak convergence under  $P_{n,h}$ . This terminology arises from the theory of limiting experiments due to Le Cam, which, however, we shall not use in this chapter.

The sequence of parameters  $\kappa_n(h)$  is assumed to belong to a Banach space  $\mathbf{B}$ . It should be regular in the sense that

$$r_n(\kappa_n(h) - \kappa_n(0)) \rightarrow \dot{\kappa}(h), \quad \text{for every } h \in H,$$

for a continuous, linear map  $\dot{\kappa}: H \mapsto \mathbf{B}$  and certain linear maps  $r_n: \mathbf{B} \mapsto \mathbf{B}$  (“norming operators”). Any maps  $T_n: \mathcal{X}_n \mapsto \mathbf{B}$  are considered estimators of the parameter. A sequence of estimators  $T_n$  is *regular* with respect to the norming operators  $r_n$  if

$$r_n(T_n - \kappa_n(h)) \xrightarrow{h} L, \quad \text{for every } h \in H,$$

for a fixed, tight, Borel probability measure  $L$  on  $\mathbf{B}$ .

This set-up covers many examples. In the simplest and most common examples, the observation at time  $n$  is a sample of  $n$  independent observations from a fixed measure  $P$ .

**3.11.1 Example (I.i.d. observations).** Let  $X_1, \dots, X_n$  be an i.i.d. sample from a probability measure  $P$  on the measurable space  $(\mathcal{X}, \mathcal{A})$ . The common law  $P$  is known to belong to a class  $\mathcal{P}$  of probability measures. It is required to estimate the parameter  $\kappa(P)$ .

Sequences of asymptotically normal experiments arise as “localized” or “rescaled” experiments. Fix some  $P \in \mathcal{P}$  and set  $P_{n,0} = P^n$ . Consider one-dimensional submodels  $t \mapsto P_t$  (maps from  $[-1, 1] \subset \mathbb{R}$  to  $\mathcal{P}$ ) such that

$$\int \left[ \frac{dP_t^{1/2} - dP^{1/2}}{t} - \frac{1}{2}h dP^{1/2} \right]^2 \rightarrow 0, \quad t \rightarrow 0,$$

for some measurable function  $h: \mathcal{X} \mapsto \mathbb{R}$ . Single out those paths such that also

$$\frac{\kappa(P_t) - \kappa(P)}{t} \rightarrow \dot{\kappa}(h),$$

for some continuous linear map  $\dot{\kappa}: L_2(P) \mapsto \mathbf{B}$ . A function  $h \in L_2(P)$  for which there exists a path  $t \mapsto P_t$  with the first property is called a *score function*, and the collection of all score functions is called the *tangent set* at  $P$  (relative to  $\mathcal{P}$ ). Let  $H$  be a linear subspace of the tangent set such that for every  $h \in H$  there exists a path  $t \mapsto P_t$  with both properties. For each  $h$ , set  $P_{n,h} = P_{1/\sqrt{n}}$  for a corresponding path  $t \mapsto P_t$ . Then Lemma 3.10.11 implies that the sequence of experiments  $(\mathcal{X}^n, \mathcal{A}^n, P_{n,h}: h \in H)$  is asymptotically normal. (Since the experiments are “local” the sequence is often called *locally asymptotically normal* (LAN).) Furthermore, the sequence of parameters  $\kappa(P_{n,h})$  is regular with respect to the norming operators  $\sqrt{n}$  (the maps  $b \rightarrow \sqrt{n}b$ ).

A sequence of estimators  $T_n = T_n(X_1, \dots, X_n)$  is regular if

$$\sqrt{n}(T_n - \kappa(P_{n,h})) \xrightarrow{h} L, \quad \text{for every } h \in H,$$

for a some fixed, tight Borel measure  $L$ . The weak convergence in this display refers to a sequence of laws  $P_{1/\sqrt{n}}$  that converges to  $P$ . Thus, in this situation regularity comes down to a certain uniformity in reaching the limit measure. In particular, regularity is weaker than weak convergence uniformly in  $P'$  ranging through a “neighborhood” of  $P$  combined with continuity of the limiting measures as  $P'$  approaches  $P$ .

The continuous, linear map  $\dot{\kappa}: H \mapsto \mathbf{B}$  has an adjoint map  $\dot{\kappa}^*: \mathbf{B}^* \mapsto \bar{H}$ , which maps the dual space of  $\mathbf{B}$  into the completion of  $H$ . This is determined by  $\langle \dot{\kappa}^* b^*, h \rangle = b^* \dot{\kappa}(h)$ .

**3.11.2 Theorem (Convolution).** *Let the sequence of statistical experiments  $(\mathcal{X}_n, \mathcal{A}_n, P_{n,h}: h \in H)$  be asymptotically normal and the sequences of parameters  $\kappa_n(h)$  and estimators  $T_n$  be regular. Then the limit distribution  $L$  of the sequence  $r_n(T_n - \kappa_n(0))$  equals the distribution of a sum  $G + W$  of independent, tight, Borel measurable random elements in  $\mathbf{B}$  such that*

$$b^* G \sim N(0, \|\dot{\kappa}^* b^*\|^2), \quad \text{for every } b^* \in \mathbf{B}^*.$$

Here  $\dot{\kappa}^*: \mathbf{B}^* \mapsto \bar{H}$  is the adjoint of  $\dot{\kappa}$ . The law of  $G$  concentrates on the closure of  $\dot{\kappa}(H)$ .

**Proof.** (a) Assume first that the dimension of  $H$  is finite. Let  $h_1, \dots, h_k$  be an orthonormal base and set

$$\begin{aligned} Z_{n,a} &= r_n(T_n - \kappa_n(\sum a_i h_i)), \\ \Lambda_n(a) &= \log \frac{dP_{n,h}}{dP_{n,0}} = \Delta_{n,h} - \frac{1}{2} \|h\|^2, \quad \text{for } h = \sum a_i h_i. \end{aligned}$$

By assumption, the sequence  $Z_{n,0}$  and each sequence  $\Delta_{n,h}$  converge in distribution in  $\mathbf{B}$  and  $\bar{\mathbb{R}}$ , respectively. By Prohorov’s theorem, there exists a subsequence of  $\{n\}$  such that

$$(Z_{n',0}, \Delta_{n',h_1}, \dots, \Delta_{n',h_k}) \xrightarrow{0} (Z, \Delta_{h_1}, \dots, \Delta_{h_k}),$$

in  $\mathbf{B} \times \mathbb{R}^k$ . By assumption,  $Z$  has (marginal) distribution equal to  $L$ .

The random variable  $\Delta_{\sum a_i h_i} - \sum a_i \Delta_{h_i}$  has second moment zero. Hence this variable is zero almost surely. Conclude that the sequence  $\Delta_{n,\sum a_i h_i} - \sum a_i \Delta_{n,h_i}$  converges to zero in probability under  $P_{n,0}$ . Combination with the preceding displays and the regularity of the sequence of parameters yield, for every  $a \in \mathbb{R}^k$ ,

$$(Z_{n',a}, \Lambda_{n'}(a)) \xrightarrow{0} \left( Z - \sum a_i \dot{\kappa}(h_i), \sum a_i \Delta_{h_i} - \frac{1}{2} \|a\|^2 \right).$$

Apply Le Cam’s third lemma, Theorem 3.10.7, to see that  $Z_{n',a} \xrightarrow{a} Z_a$ , where  $\xrightarrow{a}$  denotes weak convergence under  $P_{n,h}$  for  $h = \sum a_i h_i$ , and  $Z_a$  is distributed according to

$$(3.11.3) \quad P(Z_a \in B) = E 1_B \left( Z - \sum a_i \dot{\kappa}(h_i) \right) e^{\sum a_i \Delta_{h_i} - \frac{1}{2} \|a\|^2}.$$

Since the sequence  $T_n$  is regular, the left side equals  $L(B)$  for each  $a$ . Average the left side and the right side over  $a$  with respect to a  $N_k(0, \lambda^{-1}I)$  weight function. Straightforward calculations yield

$$L(B) = \int \mathbb{E} 1_B \left( Z - \frac{\sum \Delta_{h_i} \dot{\kappa}(h_i)}{1 + \lambda} - \frac{\sum a_i \dot{\kappa}(h_i)}{(1 + \lambda)^{1/2}} \right) c_\lambda(\Delta) dN_k(0, I)(a),$$

where  $c_\lambda(\Delta) = (1 + \lambda^{-1})^{k/2} \exp\left(\frac{1}{2}(1 + \lambda)^{-1} \sum \Delta_{h_i}^2\right)$ . Conclude that  $L$  can be written as the law of the sum  $G_\lambda + W_\lambda$  of independent random elements  $G_\lambda$  and  $W_\lambda$ , where  $G_\lambda = \sum A_i \dot{\kappa}(h_i)/(1 + \lambda)^{1/2}$  for a  $N_k(0, I)$ -distributed vector  $(A_1, \dots, A_k)$  and  $W_\lambda$  is distributed according to

$$\mathbb{P}(W_\lambda \in B) = \mathbb{E} 1_B \left( Z - \frac{\sum \Delta_{h_i} \dot{\kappa}(h_i)}{1 + \lambda} \right) c_\lambda(\Delta).$$

As  $\lambda \downarrow 0$ , we have  $G_\lambda \rightsquigarrow G = \sum A_i \dot{\kappa}(h_i)$ . The variable  $b^* G = \sum A_i b^* \dot{\kappa}(h_i)$  is normally distributed with zero mean and variance

$$\mathbb{E} b^* G_\lambda^2 = \sum (b^* \dot{\kappa}(h_i))^2 = \|b^* \dot{\kappa}\|^2.$$

By the converse part of Prohorov's theorem, the variables  $\{G_\lambda : 0 < \lambda < 1\}$  are uniformly tight. Combined with the tightness of  $L$  and the convolution statement, it follows that the variables  $\{W_\lambda : 0 < \lambda < 1\}$  are uniformly tight (Problem 3.11.4). If  $W_{\lambda_m} \rightsquigarrow W$  for a sequence  $\lambda_m \downarrow 0$ , then  $(G_{\lambda_m}, W_{\lambda_m}) \rightsquigarrow (G, W)$ , where  $G$  and  $W$  are independent and  $G + W$  is distributed according to  $L$ . This concludes the proof of the theorem for finite-dimensional  $H$ .

(b) Let  $H$  be arbitrary. For any finite orthonormal set  $h_1, \dots, h_k$ , the previous argument yields tight independent processes  $G_k$  and  $W_k$  such that  $G_k + W_k$  is distributed according to  $L$  and  $G_k$  is zero-mean Gaussian with

$$\mathbb{E} b^* G_k^2 = \sum_{i=1}^k \langle \dot{\kappa}^* b^*, h_i \rangle^2.$$

The set of all variables  $G_k$  and  $W_k$  so obtained is uniformly tight. Indeed, by tightness of  $L$ , there exists for any given  $\varepsilon > 0$  a compact set  $K$  such that  $L(K) = \int \mathbb{P}(G_k \in K - x) dP^{W_k}(x) > 1 - \varepsilon$ . Thus there exists  $x_0$  with  $\mathbb{P}(G_k \in K - x_0) > 1 - \varepsilon$ . By symmetry,  $\mathbb{P}(G_k \in x_0 - K) > 1 - \varepsilon$ , whence  $\mathbb{P}(G_k \in \frac{1}{2}(K - K)) > 1 - 2\varepsilon$ . Next, the uniform tightness of  $L$  and the collection  $G_k$  imply the uniform tightness of the collection  $W_k$ .

Direct the finite-dimensional subspaces of  $H$  by inclusion, and construct variables  $(G_k, W_k)$  for every subspace. Every weak limit point  $(G, W)$  of the net of laws  $(G_k, W_k)$  satisfies the requirements of the theorem. ■

In the convolution theorem the distribution of the Gaussian variable  $G$  is interpreted as the optimal limit law. The distribution of  $G$  is completely

determined by the model, through the properties of  $\kappa$  and  $H$ . The variable  $W$  is interpreted as a noise factor that is zero for optimal estimator sequences.

The convolution of a zero-mean Gaussian variable with an arbitrary second variable leads to a loss of concentration. This is most easily understood in terms of variances: convolution increases variance. The following lemma makes the statement precise in terms of general subconvex risk functions. A nonnegative map  $\ell: \mathbf{B} \mapsto \mathbb{R}$  is *subconvex* if the set  $\{y: \ell(y) \leq c\}$  is closed, convex, and symmetric for every  $c$ .

**3.11.4 Lemma.** *Let  $\ell: \mathbf{B} \mapsto \mathbb{R}$  be subconvex. Let  $G$  be a tight, zero-mean, Borel measurable Gaussian variable and  $W$  be an arbitrary, tight, Borel measurable variable independent of  $G$ . Then*

$$\mathbb{E}\ell(G + W) \geq \mathbb{E}\ell(G).$$

In particular,  $\mathbb{P}(\|G + W\| \leq c) \leq \mathbb{P}(\|G\| \leq c)$  for every  $c$ .

**Proof.** For a finite-dimensional Banach space, this assertion is a special case of Anderson's lemma. The general case can be proved by the approximation of  $\ell$  from below by cylinder functions. See step (b) of the proof of the minimax theorem, Theorem 3.11.5. ■

A subconvex function is certainly lower semicontinuous. Therefore, the portmanteau theorem, the lemma, and the convolution theorem can be combined to yield, for any regular estimator sequence  $T_n$  and subconvex loss function  $\ell$ ,

$$\liminf_{n \rightarrow \infty} \mathbb{E}_{0*}\ell(r_n(T_n - \kappa_n(0))) \geq \mathbb{E}\ell(G).$$

This inequality may fail for nonregular estimator sequences. However, according to the minimax theorem, the maximum risk  $\sup_h \mathbb{E}_{h*}\ell(r_n(T_n - \kappa_n(h)))$  can never asymptotically fall below  $\mathbb{E}\ell(G)$  under just some measurability conditions on the estimator sequence  $T_n$ . The following theorem gives a slightly stronger result.

A little (asymptotic) measurability is the only requirement on  $T_n$ , but measurability can be restrictive, so we need to be careful about it. Let  $\mathbf{B}'$  be a given subspace of  $\mathbf{B}^*$  that separates points of  $\mathbf{B}$ , and let  $\tau(\mathbf{B}')$  be the weak topology induced on  $\mathbf{B}$  by the maps  $b': \mathbf{B} \mapsto \mathbb{R}$  when  $b'$  ranges over  $\mathbf{B}'$ . A map  $\ell: \mathbf{B} \mapsto \mathbb{R}$  is called  $\tau(\mathbf{B}')$ -*subconvex* if the sets  $\{y: \ell(y) \leq c\}$  are  $\tau(\mathbf{B}')$ -closed, convex, and symmetric for every  $c$ . An estimator sequence  $T_n$  is *asymptotically  $\mathbf{B}'$ -measurable* if

$$\mathbb{E}_0^* f(r_n(T_n - \kappa_n(0))) - \mathbb{E}_{0*} f(r_n(T_n - \kappa_n(0))) \rightarrow 0,$$

for every function  $f$  of the form  $f(y) = g(b'_1 y, \dots, b'_k y)$  with  $b'_1, \dots, b'_k$  in  $\mathbf{B}'$  and  $g: \mathbb{R}^k \mapsto \mathbb{R}$  continuous and bounded. Clearly, every (asymptotically)

measurable sequence is asymptotically  $\tau(\mathbf{B}')$ -measurable. A sequence of stochastic processes in a function space is  $\tau(\mathbf{B}')$ -measurable for  $\mathbf{B}'$  equal to the set of coordinate projections.

**3.11.5 Theorem (Minimax theorem).** *Let the sequence of experiments  $(\mathcal{X}_n, \mathcal{A}_n, P_{n,h}: h \in H)$  be asymptotically normal and the sequence of parameters  $\kappa_n(h)$  be regular. Suppose a tight, Borel measurable Gaussian element  $G$ , as in the statement of the convolution theorem, exists. Then for every asymptotically  $\mathbf{B}'$ -measurable estimator sequence  $T_n$  and  $\tau(\mathbf{B}')$ -subconvex function  $\ell$ ,*

$$\sup_{I \subset H} \liminf_{n \rightarrow \infty} \sup_{h \in I} \mathbb{E}_{h*} \ell \left( r_n(T_n - \kappa_n(h)) \right) \geq \mathbb{E} \ell(G).$$

Here the first supremum is taken over all finite subsets  $I$  of  $H$ .

**Proof.** (a). Assume first that the loss function can be written in the special form  $\ell(y) = \sum_{i=1}^r 1_{K_i^c}(b'_{i,1}y, \dots, b'_{i,p_i}y)$  for compact, convex, symmetric subsets  $K_i \subset \mathbb{R}^{p_i}$  and arbitrary elements  $b'_{i,j}$  of  $\mathbf{B}'$ . Fix an arbitrary orthonormal set  $h_1, \dots, h_k$  in  $H$ , and set

$$Z_{n,a}^i = (b'_{i,1}, \dots, b'_{i,p_i}) \circ r_n(T_n - \kappa_n(\sum a_i h_i)), \quad 1 \leq i \leq r.$$

Considered as maps into the one-point compactification of  $\mathbb{R}^{p_i}$ , the sequences  $Z_{n,a}^i$  are certainly asymptotically tight. The sequences are asymptotically measurable by assumption.

Direct the finite subsets of  $H$  by inclusion. There exists a subnet  $\{n_I: I \subset H, \text{finite}\}$  such that the left side of the statement of the theorem equals

$$\text{minimax risk} = \limsup_{I} \sup_{h \in I} \mathbb{E}_{h*} \ell \left( r_n(T_n - \kappa_n(h)) \right).$$

By the same arguments as in the proof of the convolution theorem there is a further subnet  $\{n'\} \subset \{n_I\}$  such that  $Z_{n',a}^i \xrightarrow{a} Z_a^i$  in the one-point compactifications, for every  $a \in \mathbb{R}^k$  and every  $i$ . Here the limiting processes satisfy, for each  $i$ ,

$$\int \mathcal{L}(Z_a^i) dN_k(0, \lambda^{-1}I) \sim G_\lambda^i + W_\lambda^i,$$

for independent elements  $G_\lambda^i$  and  $W_\lambda^i$  such that

$$G_\lambda^i = (b'_{i,1}, \dots, b'_{i,p_i}) \circ G_\lambda = (b'_{i,1}, \dots, b'_{i,p_i}) \circ \frac{\sum A_i \dot{\kappa}(h_i)}{(1+\lambda)^{1/2}},$$

for a  $N_k(0, I)$ -distributed vector  $(A_1, \dots, A_k)$ . By the portmanteau theorem,

$$\text{minimax risk} \geq \liminf_{n'} \sum_{i=1}^r P_{a*}(Z_{n',a}^i \notin K_i) \geq \sum_{i=1}^r P(Z_a^i \notin K_i).$$

Since this is true for every  $a$ , the left side is also bounded below by the average of the right side. Combination with Lemma 3.11.4 (only the finite-dimensional case is needed here) yields

$$\text{minimax risk} \geq \sum_{i=1}^r P(G_\lambda^i + W_\lambda^i \notin K_i) \geq \sum_{i=1}^r P(G_\lambda^i \notin K_i) = E\ell(G_\lambda).$$

Finish the proof for this special form of loss function by letting  $\lambda \downarrow 0$  followed by taking the limit along finite-dimensional subspaces of  $H$ .

(b). An arbitrary subconvex  $\ell$  can be approximated from below by a sequence of functions  $\ell_r$  of the type as in (a). More precisely, the sequence  $\ell_r$  can be constructed such that  $0 \leq \ell_r \leq \ell$  everywhere and  $\ell_r \uparrow \ell$   $G$ -almost surely. It follows that the minimax risk decreases when  $\ell$  is replaced by  $\ell_r$ . Combination with (a) shows that the minimax risk is bounded below by  $E\ell_r(G)$  for every  $r$ . The theorem follows by letting  $r \rightarrow \infty$ .

For the construction of  $\ell_r$ , note first that

$$0 \leq 2^{-r} \sum_{i=1}^{2^{2r}} 1\{y: \ell(y) > i2^{-r}\} \uparrow \ell(y), \quad \text{for every } y.$$

Each of the sets  $\{y: \ell(y) > i/r\}$  is convex,  $\tau(\mathbf{B}')$ -closed, and symmetric. Thus, it suffices to approximate functions  $\ell$  of the type  $1_{C^c}$  for a convex,  $\tau(\mathbf{B}')$ -closed, and symmetric set  $C$ .

By the Hahn-Banach theorem, any such set  $C$  can be written

$$C = \bigcap_{b' \in \mathbf{B}'} \{y: |b'y| \leq c_{b'}\}.$$

Thus the complement of  $C$  intersected with the support  $S$  of the limit variable  $G$  is the union of the sets  $\{y \in S: |b'y| > c_{b'}\}$ . These sets are relatively open in  $S$  and  $S$  is separable. Since a separable set is Lindelöf, the possibly uncountable union can be replaced by a countable subunion. Thus there exists a sequence  $b'_i$  in  $\mathbf{B}'$  and numbers  $c_i$  such that  $C^c \cap S = \cup_{i=1}^{\infty} \{y \in S: |b'_i y| > c_i\}$ . This implies that

$$1_{C^c \cap S} = \sup_r 1_{K_r^c}(b'_1 y, \dots, b'_r y),$$

for the subsets of  $\mathbb{R}^r$  defined by  $K_r = \cap_{i=1}^r \{x \in \mathbb{R}^r: |x_i| \leq c_i\}$ . ■

**3.11.6 Example (Separable Banach space).** A convex set in a Banach space is closed with respect to the norm if and only if it is weakly closed. This means that “ $\tau(\mathbf{B}^*)$ -subconvex” is identical to “subconvex”.

We conclude that the minimax theorem is valid for Borel measurable estimator sequences  $T_n$  and every subconvex loss function. The restriction that each  $T_n$  be Borel measurable is reasonable in separable Banach spaces, but less so in nonseparable spaces.

**3.11.7 Example (Skorohod space).** Consider the Skorohod space  $D[a, b]$  for a given interval  $[a, b] \subset \bar{\mathbb{R}}$ , equipped with the uniform norm. The dual space consists of maps of the form

$$d^*(z) = \int z(u) d\mu(u) + \sum_{i=1}^{\infty} \alpha_i (z(u_i) - z(u_i-)),$$

for a finite signed measure  $\mu$  on  $[a, b]$ , an arbitrary sequence  $u_i$  in  $(a, b]$ , and a sequence  $\alpha_i$  with  $\sum |\alpha_i| < \infty$ .<sup>b</sup> Each such  $d^*$  is the pointwise limit of a sequence of linear combinations of coordinate projections. Thus, the  $\sigma$ -field generated by the dual space equals the  $\sigma$ -field generated by the coordinate projections.

It follows that an estimator sequence is  $\tau(D[a, b]^*)$ -measurable if and only if it is a stochastic process. Since “ $\tau(D[a, b]^*)$ -subconvex” is identical to “subconvex with respect to the norm” (Problem 3.11.3), the minimax theorem is valid for any sequence of stochastic processes  $T_n$  and subconvex loss function  $\ell$ .

Examples of subconvex loss functions include

$$\begin{aligned} z &\mapsto \ell_0(\|z\|_\infty), \\ z &\mapsto \int |z|^p(t) d\mu(t), \end{aligned}$$

for a nondecreasing, left-continuous function  $\ell_0: \mathbb{R} \mapsto \mathbb{R}$ , a finite Borel measure  $\mu$ , and  $p \geq 1$ .

**3.11.8 Example (Bounded functions).** On the space  $\ell^\infty(\mathcal{F})$ , functions of the type

$$z \mapsto \ell_0\left(\left\|\frac{z}{q}\right\|_{\mathcal{F}}\right),$$

for a nondecreasing, left-continuous function  $\ell_0: \mathbb{R} \mapsto \mathbb{R}$  and an arbitrary map  $q: \mathcal{F} \mapsto \mathbb{R}$  are subconvex with respect to the linear space spanned by the coordinate projections  $z \mapsto z(f)$ . Indeed, for any  $c$  there exists  $d$  such that

$$\left\{ z: \ell_0\left(\left\|\frac{z}{q}\right\|_{\mathcal{F}}\right) \leq c \right\} = \left\{ z: \left\|\frac{z}{q}\right\|_{\mathcal{F}} \leq d \right\} = \bigcap_{f \in \mathcal{F}} \left\{ z: |z(f)| \leq d q(f) \right\}.$$

Thus, the minimax theorem is valid for any estimator sequence  $T_n$  that is coordinatewise measurable and any loss function of this type.

For general loss functions that are subconvex with respect to the norm, the preceding minimax theorem applies only under strong measurability conditions on the estimator sequences. It is of interest that these measurability conditions are satisfied by sequences  $T_n$  such that  $T_n(f)$  is measurable for every  $f$  and such that the sequence  $r_n(T_n - \kappa_n(0))$  is asymptotically

---

<sup>b</sup> Van der Vaart (1988), pages 81–85.

tight under  $P_{n,0}$ . Indeed, according to Lemma 1.5.2, such sequences are asymptotically  $\tau(\ell^\infty(\mathcal{F})^*)$ -measurable. It follows that, given any subconvex loss function, the minimax theorem may be used to designate optimal estimator sequences among the asymptotically tight sequences.

### 3.11.1 Efficiency of the Empirical Distribution

Let  $\mathcal{F}$  be a  $P$ -Donsker class of functions on a measurable space  $(\mathcal{X}, \mathcal{A})$ . In this section it is shown that the empirical distribution  $\mathbb{P}_n$  of a sample of size  $n$  from a measure  $P$  is an asymptotically efficient estimator for  $P$ , when both are (and can be) viewed as elements  $f \mapsto \mathbb{P}_n f$  and  $f \mapsto Pf$  in  $\ell^\infty(\mathcal{F})$  and when  $P$  is considered completely unknown.

The efficiency here is understood in the sense of best (locally) regular at  $P$  and locally asymptotically minimax at  $P$  within the context of Example 3.11.1. For local efficiency at all  $P$  in a collection  $\mathcal{P}$ , we would assume that  $\mathcal{F}$  is  $P$ -Donsker with  $\|P\|_{\mathcal{F}} < \infty$  for every  $P \in \mathcal{P}$ . At the end of this section we also briefly discuss global minimaxity over a class of underlying measures  $\mathcal{P}$ .

We use the set-up of Example 3.11.1 with the “model”  $\mathcal{P}$  equal to the class of all probability measures on  $(\mathcal{X}, \mathcal{A})$ . Then the tangent set can be shown to be equal to all square-integrable measurable functions  $h$  such that  $Ph = 0$ . We do not really need a result as strong as this (and would actually not be able to use it either, because the parameter might be irregular with respect to this large set). Instead we note that for every bounded, measurable function  $h: \mathcal{X} \mapsto \mathbb{R}$ , the measures with densities  $dP_t(x) = (1 + th(x)) dP(x)$  are well defined for sufficiently small  $|t|$  and define probability measures if  $Ph = 0$ . These paths can be seen to be differentiable in quadratic mean by elementary arguments. We define the set  $H$  to be equal to all bounded functions with mean zero and let  $P_{n,h}$  be the measure  $P_{1/\sqrt{n}}$ .

The parameter to be estimated is the map  $f \mapsto Pf$ , which is assumed to belong to  $\ell^\infty(\mathcal{F})$ . Since  $P_tf = Pf + tPfh$  for every  $f$ , the derivative of this parameter is the map  $\dot{\kappa}: h \mapsto (f \mapsto Pfh)$ . Since  $\mathcal{F}$  is a Donsker class, it is bounded in  $L_2(P)$ , which implies that the derivative  $\dot{\kappa}$  is continuous and the parameter regular with respect to our choice of  $H$ .

Since  $\mathcal{F}$  is assumed a Donsker class, the sequence  $\sqrt{n}(\mathbb{P}_n - P)$  converges in distribution under  $P$  to a Brownian bridge process  $\mathbb{G}_P$ . By Theorem 3.10.12, the sequence  $\sqrt{n}(\mathbb{P}_n - P_{n,h})$  has the same limit distribution if the observations are sampled from  $P_{n,h}$ . Thus, the sequence of empirical distributions is regular (at  $P$ ). Furthermore, for every bounded, continuous function  $\ell: \ell^\infty(\mathcal{F}) \rightarrow [0, \infty)$ ,

$$(3.11.9) \quad \sup_{I \subset H} \limsup_{n \rightarrow \infty} \sup_{h \in I} \mathbb{E}_{h*} \ell(\sqrt{n}(\mathbb{P}_n - P_{n,h})) = \mathbb{E} \ell(\mathbb{G}_P).$$

Thus, we may conclude that the empirical distribution is asymptotically efficient in terms of both the convolution and the minimax theorems, (for

certain loss functions) if the tight Gaussian element  $G$  in these theorems equals the Brownian bridge. For our choice of  $H$ , this is indeed the case and easy and to prove.

It suffices to calculate the covariance function of  $G$  in Theorem 3.11.2. The projection  $\pi_f: \ell^\infty(\mathcal{F}) \mapsto \mathbb{R}$  on the  $f$ -coordinate, given by  $z \mapsto z(f)$ , is an element of the dual space of  $\ell^\infty(\mathcal{F})$ . By definition of  $\kappa^*$ , we have

$$P(\kappa^* \pi_f)h = \pi_f(\kappa h) = Pfh = P(f - Pf)h, \quad \text{for every } h \in H.$$

Since the completion of  $H$  consists of all zero-mean functions and  $\kappa^*$  is required to map  $\ell^\infty(\mathcal{F})$  into  $\tilde{H}$ , this means that  $\kappa^* \pi_f$  is equal to  $f - Pf$ . Conclude that  $G(f)$  in Theorem 3.11.2 has variance  $P(f - Pf)^2$ . A similar argument applied to linear combinations of projections shows that the covariance function of  $G$  is equal to the covariance of a Brownian bridge  $\mathbb{G}_P$ .

The condition that  $\ell$  is bounded and continuous can be relaxed in many situations. In view of Theorem 1.11.3, equation (3.11.9) is valid for every function  $\ell$  that is continuous almost surely under the distribution of  $\mathbb{G}_P$  and such that the sequence  $\ell(\sqrt{n}(\mathbb{P}_n - P_{n,h}))$  is asymptotically equi-integrable under  $P_{n,h}$ .

The result (3.11.9) asserts that the empirical distribution is asymptotically minimax in a local sense. If  $\mathcal{F}$  is uniformly Donsker in a class  $\mathcal{P}$  of underlying measures and  $\ell$  is bounded and Lipschitz-continuous, then

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{E}_P \ell(\sqrt{n}(\mathbb{P}_n - P)) = \sup_{P \in \mathcal{P}} \mathbb{E} \ell(\mathbb{G}_P).$$

From consideration of the lower bounds for the local minimax risks, the right-hand side is the minimum value attainable for any (asymptotically measurable) estimator sequence if the local submodels  $P_{n,h}$  corresponding to the measure(s)  $P$  for which the right-hand side is (nearly) maximal also belong to  $\mathcal{P}$ . This is particularly the case if  $\mathcal{P}$  is the collection of all probability measures. In this situation the empirical measure is “globally asymptotically minimax over  $\mathcal{P}$ ” with respect to the loss  $\ell$ .

In the classical situation of cells  $(-\infty, t]$  in Euclidean space, the global asymptotic minimax character has been shown to be valid also for loss functions that are not continuous or bounded. (See the Notes to this chapter.) In more general situations this question appears not to have been investigated.

## Problems and Complements

- Both the convolution and the minimax theorem remain valid if the set  $H$  is not a linear space but does contain a convex cone that has the same closed linear span as  $H$ .

[Hint: Van der Vaart (1988).]

2. Suppose that  $\mathcal{P} = \{P_\theta : \theta \in \Theta\}$  is a parametric model indexed by an open subset of Euclidean space such that

$$\int \left[ dP_\theta^{1/2} - dP_{\theta_0}^{1/2} - \frac{1}{2}(\theta - \theta_0)' \dot{\ell}_{\theta_0} dP_{\theta_0}^{1/2} \right]^2 d\mu = o(\|\theta - \theta_0\|^2),$$

for some measurable, vector-valued function  $\dot{\ell}_{\theta_0}$ . Derive the optimal variance for estimating a parameter  $q(\theta)$  given a differentiable map  $q: \Theta \mapsto \mathbb{R}$ . Is this parameter always regular?

3. A convex set in a Banach space is closed with respect to the norm if and only if it is weakly closed. Thus, “ $\tau(\mathbf{B}^*)$ -subconvex” is identical to “subconvex”.  
**[Hint:** Apply the Hahn-Banach theorem, or see Rudin (1973).]
4. Suppose  $L_n$  and  $M_n$  are uniformly tight sequences of Borel probability measures on a metric space. If  $L_n = M_n * N_n$  for every  $n$ , then the sequence  $N_n$  is uniformly tight.

# 3

## Notes

**3.2.** Versions of the continuous mapping theorem for the argmax functional have already been used by Prakasa Rao (1969), Le Cam (1970a), and Ibragimov and Has'minskii (1981). An explicit statement of a slightly more special theorem than the theorem presented here occurs in Kim and Pollard (1990). It is clear from the proof that uniform weak convergence of the criterion functions can be relaxed to the weak convergence of pairs of maxima of the form  $(\sup_{h \in A} M_n(h), \sup_{h \in B} M_n(h))$ . This is slightly weaker, but in general still much stronger than needed.

For general notes on results on rates of convergence, see the notes for the next chapter. Theorem 3.2.10 is essentially due to Kim and Pollard (1990). The maximum likelihood estimator for a monotone decreasing density (or concave distribution function) is known as the Grenander estimator. Its limit behavior was derived by Prakasa Rao (1969). In his study of the concave majorant of Brownian motion Groeneboom (1983) introduced an “inverse process” similar to  $\hat{s}_n$  and the corresponding basic switching identity which is the basis of our proof. Groeneboom (1985) goes on to initiate the study of global functionals of  $\hat{f}_n$ , in particular the  $L_1$  distance from  $\hat{f}_n$  to  $f$  using the structure of the inverse process  $\hat{s}_n$ . The limit distribution of the maximum likelihood estimator based on interval censored observations was first obtained by Groeneboom (1987). The present example is an adaptation of the presentation in Part 2 of Groeneboom and Wellner (1992). Groeneboom (1988) gave a systematic study of the “inverse process” associated with Brownian motion minus a parabola, including a determination of its infinitesimal generator. As a corollary of this deep and detailed anal-

ysis, Groeneboom characterized the distribution of the point of maximum of Brownian motion minus a parabola.

Rigorous proofs of the asymptotic normality of the maximum likelihood estimator were first given in the 1930s and 1940s. For a direct proof under the classical smoothness conditions, see, for instance, Cramér (1946). General  $M$ -estimators were studied systematically by Huber (1967), who also introduced bracketing conditions to ensure their asymptotic consistency and normality. Pollard (1984, 1985) shows the gains of applying results from empirical processes, as in this monograph, to this problem. Theorems 3.2.16 and 3.3.1 are generalizations of results in this work. Hoffmann-Jørgensen (1994) also aims at applying techniques from probability in Banach spaces to maximum likelihood estimators (in mostly parametric models).

Example 3.2.12 follows Huang (1993). Example 3.2.24 is taken from lecture notes by the first author. The Lipschitz condition on  $\theta \mapsto \log p_\theta$  can of course be relaxed, but it seems to cover all the standard examples. For one-dimensional parameters, a better approach is possible based on the maximal inequality given by Theorem 2.2.4, applied with  $\psi(\varepsilon) = \varepsilon^2$  and the Hellinger distance. Then the pointwise-Lipschitz condition can be replaced by the condition  $\int [p_{\theta_1}^{1/2} - p_{\theta_2}^{1/2}]^2 \leq |\theta_1 - \theta_2|^2$ . See Le Cam (1970a) and Ibragimov and Has'minskii (1981) for results in this direction.

**3.3.** Estimators satisfying estimating equations were studied systematically by Huber (1967). Pollard (1984, 1985) gives the benefits of applying results from empirical processes, as in this monograph, to this problem. Theorems 3.3.1 is a generalization of results in his work to infinite dimensions. Also, see Bickel, Klaassen, Ritov, and Wellner (1993) for results of this type. Theorem 3.3.1 is used by Van der Vaart (1994a, 1994c, 1995) and Murphy (1995) to prove the asymptotic efficiency of the maximum likelihood estimator in some semiparametric models. Similar results can be found in Bickel, Klaassen, Ritov, and Wellner (1993). Example 3.3.10 is based on Van der Vaart (1994a, 1994c). For other applications of empirical processes in obtaining the limit distribution of estimators in semiparametric models, see Van der Laan (1993), Huang (1996), and Van der Vaart (1996).

**3.4.** A major part of this chapter is based on the exposition by Van de Geer (1993b), though the introduction of  $L_1$ -inequalities and the “Bernstein norm” are new and simplify the treatment.

Van de Geer (1990) studies least-squares and least-absolute-deviation estimators for regression functions. In the case of least-squares estimators, she uses chaining for sub-Gaussian processes, under the assumption that the error distribution is sub-Gaussian. The results on least-squares with random design were obtained by Birgé and Massart (1993). Van de Geer (1993b) considers maximum likelihood estimators for densities under the

assumption that the model is convex or that the loglikelihood ratios are uniformly bounded. Her arguments are based on chaining by bracketing.

Birgé and Massart (1993) suggest using the functions  $\log(p+p_0)/2p_0$  as criterion functions rather than the loglikelihood ratios. They also introduce the use of the refined version of Bernstein's inequality in the bracketing-chaining argument, both for regression and maximum likelihood estimators. Using this refined inequality, Wong and Shen (1995), removed unnecessary boundedness conditions and obtain a version of Theorem 3.4.4. Other references in this area are Wong and Severini (1991) and Shen and Wong (1994).

The maximum likelihood estimator for a monotone density was introduced by Grenander (1956). The rate at a point for the maximum likelihood estimator of a convex regression function is  $n^{-2/5}$ , as shown by Mammen (1991) and Wang (1993). Jongbloed (1995) obtains the same pointwise rate for the maximum likelihood estimator of a convex density.

Related papers that are not treated in this chapter are Birgé and Massart (1994), who discuss the use of many different types of sieves, and Barron, Birgé, and Massart (1995), who consider model selection via penalized minimum-contrast estimators.

**3.5.** The basic method of Poissonization was used by Kolmogorov (1933) and Donsker (1952) in their classic papers; see, e.g., Shorack and Wellner (1986), Chapter 8. Poissonization techniques in the study of Banach-space-valued random variables began in Le Cam (1970b) and were developed further in Araujo and Giné (1980). Kac (or Poissonized) empirical processes have also played an important role in obtaining rates of convergence for limit theorems and invariance principles; see, e.g., Massart (1989), in particular Lemma 2, and Dudley (1984), Section 8.3, where Poissonization is used to study rates of convergence for classes of sets too large to satisfy the central limit theorem. Durst and Dudley (1981) prove a partial result in the direction of Theorem 3.1 for classes of sets. In the classical one-dimensional case  $\mathcal{X} = \mathbb{R}$ , there is a substantial literature concerned with finite sample and asymptotic distributions of statistics connected with the Kac empirical process; see S. Csörgő (1981) and the references therein.

Pyke (1968) proves Theorem 3.5.1 for the one-dimensional distribution function under the assumption that  $\nu$  equals 1 almost surely. Billingsley (1968) and S. Csörgő (1974), following Blum, Hanson and Rosenblatt (1963), show that the result continues to hold under the present conditions, which allow a general positive limit random variable  $\nu$ . These results were extended by Wichura (1968) and Fernandez (1970) to allow, for instance, the  $q$ -metrics mentioned by Pyke (1968) in his closing remark. Theorem 3.5.1 contains their results as well as many others.

Part of Theorem 3.5.5 is contained in Theorem 3.4.9 of Araujo and Giné (1980); see also their Exercise 2 on page 122, and note that their  $X_{nj}$  are assumed symmetric. Evarist Giné has shown us a proof of Theo-

rem 3.5.5 via symmetrization and desymmetrization. The present statement of Theorem 3.5.5 is from Klaassen and Wellner (1992).

**3.6.** The bootstrap central limit theorems, Theorems 3.6.1 and 3.6.2, follow essentially from the bootstrap results of Giné and Zinn (1990), who prove the equivalence of (i) and (iii) under measurability assumptions on the class  $\mathcal{F}$ , the equivalence of (i) and (ii) being immediate from the multiplier central limit theorems. A new detail in the present statement is the absence of measurability assumptions.

The proofs of Giné and Zinn use symmetrization by Rademacher variables. If we write  $s(i)$ ,  $s(ii)$ , and  $s(iii)$  for the corresponding symmetrized parts of Theorem 3.6.1 and  $s(ii)u$  for the unconditional symmetrized version of (ii), then the proof of Giné and Zinn (1990) consists of the steps  $s(i) \Rightarrow s(ii)u \Rightarrow s(iii) \Rightarrow s(i)$ . Here the first implication follows from Lemma 2.9 of Giné and Zinn (1984); the second follows from the almost-sure multiplier central limit theorem of Ledoux and Talagrand (1988) and Proposition 2.2 of Giné and Zinn (1990); the last implication comes from Proposition 2.2 again. Giné and Zinn (1990) next show that the  $s$ 's can be removed from the string of implications. (They do not have  $s(ii)$  or  $s(ii)u$  in the statement of their main result, but only in their Lemma 2.1 and Proposition 2.2.)

The finite-dimensional convergence of the exchangeable bootstrap empirical process is proved by Mason and Newton (1990), who also establish Theorem 3.6.13 for the special case that  $\mathcal{F}$  is the class of indicators of cells in the real line. The general version of Theorem 3.6.13 (under somewhat stronger assumptions, including that  $\sum_{i=1}^n W_{ni} = n$ ) is from Praestgaard and Wellner (1993), as is the almost-sure part of Theorem 3.6.3. The improbability part of Theorem 3.6.3 was noted by Arcones and Giné (1992).

**3.8.** For observations in the unit square, an alternative to the Kolmogorov-Smirnov statistic is

$$B_n = \int \int n(\mathbb{H}_n(s, t) - \mathbb{P}_n(s)\mathbb{Q}_n(t))^2 d\mathbb{H}_n(s, t).$$

Blum, Kiefer and Rosenblatt (1961) show that under independence and continuity of the marginal distributions  $P$  and  $Q$

$$B_n \rightsquigarrow \iint \mathbb{Z}_{P \times Q}^2(s, t) dP(s)dQ(t) \sim \int_0^1 \int_0^1 \mathbb{Z}_{[0,1]^2}^2(u, v) du dv.$$

They also compute the characteristic function of the limit variable and use this to compute and tabulate its distribution. For higher-dimensional generalizations and extensions, see Dugue (1975) and Deheuvels (1981). The sequence  $B_n$  is asymptotically equivalent to a statistic proposed by Hoeffding (1948).

For material on distributions related to the completely tucked Brownian sheet, see, e.g., Adler (1990), Section V.5.

**3.9.** The use of Hadamard instead of Fréchet differentiability to derive limiting distributions of transformed statistics was developed systematically by Reeds (1975), although there are earlier applications in work by Le Cam.

A number of the examples can be found in Gill (1989). Comprehensive reviews of the product integral can be found in Gill and Johansen (1990) and Gill (1994). We borrowed the symbol for the product integral from the latter author. They work in the more general framework of  $p \times p$  matrices of elements of  $D(\mathbb{R}^+)$ , in which case the definition given here is appropriate only in the “commutative case.” The limit behavior of estimators connected to multivariate trimming is studied by Nolan (1992). We have reformulated her result in terms of Hadamard differentiability, under slightly simpler conditions. The result on the Hadamard differentiability of  $M$ -functionals is taken from Heesterman and Gill (1992) and Van der Vaart (1995), although this is also one of the main examples in Reeds (1975). For still more applications of the delta-method, see, e.g., Grübel and Pitts (1992, 1993) and Pitts (1994).

The weak convergence of the sample quantile process was proved by Bickel (1967). Substantial refinements, which involve further interplay between analysis and probability, are due to Shorack (1972) and Csörgő and Révész (1978); see Shorack and Wellner (1986), Chapter 18. See also Shorack and Wellner (1986), Chapter 15, for connections with the well-known Bahadur-Kiefer theorems. For further applications of the approach taken here, see Doss and Gill (1992).

The use of Hadamard differentiability does not allow conclusions about the speed of convergence. On the other hand, Fréchet differentiability, with a rate on the remainder term plus information on the speed of convergence of the original sequence, gives a rate on the linear approximation for the transformed statistics. In a series of papers, Dudley (1990, 1992, 1994) explores Fréchet differentiability of some of the standard functionals with respect to nonstandard norms, such as  $p$ -variation on distribution functions.

**3.10.** Contiguity was introduced and studied by Le Cam. See Le Cam (1960, 1969). Also see Le Cam (1986), pages 85–90, for connections with limiting experiments and different characterizations. Differentiability in quadratic mean was introduced by Le Cam as well. Le Cam’s third lemma derives its name from being listed as the third in a list of basic lemmas. For the other lemmas, see the original papers by Le Cam or see Hájek and Šidák (1967).

**3.11.** The convolution and minimax theorems for locally asymptotically normal parametric models were established by Hájek (1970, 1972). Le Cam (1972) shows the connection with approximation of statistical experiments and states similar theorems for non-Gaussian limit experiments. The proofs for the Gaussian case in the present chapter have the benefit of being short

and easy, but they lack the insight provided by Le Cam's method, which is to show the convergence of a sequence of experiments to a limit in a general sense first and next obtain concrete results (such as a convolution theorem) from an analysis of the limit experiment. The convolution theorem for infinite-dimensional parameters was proved in increasing generality by Levit (1978), Millar (1983, 1985), Van der Vaart (1988), and Van der Vaart and Wellner (1989), among others. The assumption of asymptotic tightness and measurability can be relaxed considerably, as well as the assumption that the tangent space is linear, see Van der Vaart (1988). For many examples of tangent spaces in semi-parametric models, see Bickel, Klaassen, Ritov, and Wellner (1993).

Anderson (1963) proves a concentration lemma for symmetric distributions on Euclidean space. Lemma 3.11.4, taken from Millar (1983), extends his result to infinite-dimensional Gaussian variables.

Kiefer and Wolfowitz (1959) shows the global asymptotic minimax character of the empirical distribution function in Euclidean space. Here "global" means that they took the supremum over all probability measures, rather than over shrinking neighborhoods as in the present chapter. This is better in terms of attainment of the bound, but less interesting for the lower bound. Millar (1979) shows that the empirical distribution is also asymptotically globally minimax in certain models not consisting of all probability distributions on the sample space. His results can be further refined to local minimaxity, for which the key property is that the tangent space is sufficiently large. See Problem 3.11.1.

PART A

# Appendix

# A.1

## Inequalities

This section presents several inequalities for sums of independent stochastic processes. For a stochastic process  $\{X_t: t \in T\}$  indexed by some arbitrary index set, the notation  $\|X\|$  is an abbreviation for the supremum  $\sup_t |X_t|$ .

The stochastic processes need not be measurable maps into a Banach space. Independence of the stochastic processes  $X_1, X_2, \dots$  is understood in the sense that each of the processes is defined on a product probability space  $\prod_{i=1}^{\infty} (\Omega_i, \mathcal{U}_i, P_i)$  with  $X_i$  depending on the  $i$ th coordinate of  $(\omega_1, \omega_2, \dots)$  only. The process  $X_i$  is called symmetric if  $X_i$  and  $-X_i$  have the same distribution. In the case of nonmeasurability, the symmetry of independent processes  $X_1, X_2, \dots$  may be understood in the sense that outer probabilities remain the same if one or more  $X_i$  are replaced by  $-X_i$ .<sup>†</sup>

Throughout the chapter  $S_n$  equals the partial sum  $X_1 + \dots + X_n$  of given stochastic processes  $X_1, X_2, \dots$ .

**A.1.1 Proposition (Ottaviani's inequality).** *Let  $X_1, \dots, X_n$  be independent stochastic processes indexed by an arbitrary set. Then for  $\lambda, \mu > 0$ ,*

$$P^*\left(\max_{k \leq n} \|S_k\| > \lambda + \mu\right) \leq \frac{P^*(\|S_n\| > \lambda)}{1 - \max_{k \leq n} P^*(\|S_n - S_k\| > \mu)}.$$

**Proof.** Let  $A_k$  be the event that  $\|S_k\|^*$  is the first  $\|S_j\|^*$  that is strictly greater than  $\lambda + \mu$ . The event on the left is the disjoint union of  $A_1, \dots, A_n$ .

---

<sup>†</sup> This is the case, for instance, if each  $(\Omega_i, \mathcal{U}_i, P_i)$  is a product  $(\mathcal{X}_i, \mathcal{A}_i, P_i)^2$  of two identical probability spaces and  $X_i(x_1, x_2) = Z_i(x_1) - Z_i(x_2)$  for some stochastic process  $Z_i$ .

Since  $\|S_n - S_k\|^*$  is independent of  $\|S_1\|^*, \dots, \|S_k\|^*$ ,

$$\begin{aligned} P(A_k) \min_{j \leq n} P(\|S_n - S_j\|^* \leq \mu) &\leq P(A_k, \|S_n - S_k\|^* \leq \mu) \\ &\leq P(A_k, \|S_n\|^* > \lambda), \end{aligned}$$

since  $\|S_k\|^* > \lambda + \mu$  on  $A_k$ . Sum up over  $k$  to obtain the result. ■

**A.1.2 Proposition (Lévy's inequalities).** *Let  $X_1, \dots, X_n$  be independent, symmetric stochastic processes indexed by an arbitrary set. Then for every  $\lambda > 0$  we have the inequalities*

$$P^*\left(\max_{k \leq n} \|S_k\| > \lambda\right) \leq 2P^*(\|S_n\| > \lambda),$$

$$P^*\left(\max_{k \leq n} \|X_k\| > \lambda\right) \leq 2P^*(\|S_n\| > \lambda).$$

**Proof.** Let  $A_k$  be the event that  $\|S_k\|^*$  is the first  $\|S_j\|^*$  that is strictly greater than  $\lambda$ . The event on the left in the first inequality is the disjoint union of  $A_1, \dots, A_n$ . Write  $T_n$  for the sum of the sequence  $X_1, \dots, X_k, -X_{k+1}, \dots, -X_n$ . By the triangle inequality,  $2\|S_k\|^* \leq \|S_n\|^* + \|T_n\|^*$ . It follows that

$$P(A_k) \leq P(A_k, \|S_n\|^* > \lambda) + P(A_k, \|T_n\|^* > \lambda) = 2P(A_k, \|S_n\|^* > \lambda),$$

since  $X_1, \dots, X_n$  are symmetric. Sum up over  $k$  to obtain the first inequality.

For the second inequality, let  $A_k$  be the event that  $\|X_k\|^*$  is the first  $\|X_j\|^*$  that is strictly greater than  $\lambda$ . Write  $T_n$  for the sum of the variables  $-X_1, \dots, -X_{k-1}, X_k, -X_{k+1}, \dots, -X_n$ . By the triangle inequality,  $2\|X_k\|^* \leq \|S_n\|^* + \|T_n\|^*$ . Proceed as before. ■

An interesting consequence of Lévy's inequalities is the following theorem concerning the convergence of random series of independent processes.

**A.1.3 Proposition (Lévy-Ito-Nisio).** *Let  $X_1, X_2, \dots$  be independent stochastic processes with sample paths in  $\ell^\infty(T)$ . Then the following statements are equivalent:*

- (i)  $\{S_n\}$  converges outer almost surely;
- (ii)  $\{S_n\}$  converges in outer probability;
- (iii)  $\{S_n\}$  converges weakly.

**Proof.** The implications (i)  $\Rightarrow$  (ii)  $\Rightarrow$  (iii) are true for general random sequences. We must prove the implications in the converse direction.

(iii)  $\Rightarrow$  (ii). Since any Cauchy sequence is convergent, it suffices to show that  $S_{n_{k+1}} - S_{n_k}$  converges in outer probability to zero as  $k \rightarrow \infty$  for any sequence  $n_1 < n_2 < \dots$ . The sequence  $(S_{n_{k+1}}, S_{n_k})$  is asymptotically tight and asymptotically measurable. By Prohorov's theorem, every subsequence

has a further subsequence along which  $Y_k = S_{n_{k+1}} - S_{n_k}$  converges in distribution to a tight limit  $Y$ . Then  $Y_n(t) \rightsquigarrow Y(t)$  for every  $t$ . For every real number  $s$  at which the characteristic function of the weak limit of  $S_n(t)$  is nonzero, in particular for every  $s$  sufficiently close to zero,

$$\mathbb{E}e^{isY_k(t)} = \frac{\mathbb{E}e^{isS_{n_{k+1}}(t)}}{\mathbb{E}e^{isS_{n_k}(t)}} \rightarrow 1.$$

Thus the characteristic function of  $Y(t)$  is 1 in a neighborhood of zero. Conclude that  $Y = 0$  almost surely.

(ii)  $\Rightarrow$  (i). Write  $S$  for the limit in probability of  $S_n$ . First assume that the processes  $X_j$  are symmetric. There exists a subsequence  $n_1 < n_2 < \dots$  such that  $P^*(\|S_{n_k} - S\| > 2^{-k}) < 2^{-k}$  for every  $k$ . By the Borel-Cantelli lemma,  $S_{n_k} \xrightarrow{\text{as*}} S$  as  $k \rightarrow \infty$ . By a Lévy inequality,

$$P^*\left(\max_{n_k < n \leq n_{k+1}} \|S_n - S_{n_k}\| > 2^{-k+1}\right) \leq 2P^*\left(\|S_{n_{k+1}} - S_{n_k}\| > 2^{-k+1}\right).$$

The right side is smaller than a multiple of  $2^{-k}$ . Hence  $\max_{n_k < n \leq n_{k+1}} \|S_n - S_{n_k}\|^*$  converges almost surely to zero by the Borel-Cantelli lemma. This concludes the proof that  $S_n$  converges outer almost surely for symmetric  $X_i$ .

Given general processes, construct an independent copy  $Y_1, Y_2, \dots$  defined on a copy of the original probability space  $(\Omega, \mathcal{U}, P)$ , and let  $T_n$  be the corresponding partial sums. Then elements of the sequence  $S_n - T_n$  are the partial sums of the symmetric variables  $X_i - Y_i$  and converges in outer probability. (It is formally defined on  $(\Omega, \mathcal{U}, P) \times (\Omega, \mathcal{U}, P)$ .) By the preceding paragraph,  $S_n - T_n$  converges outer almost surely. By Fubini's theorem there exists  $\omega$  such that  $S_n - T_n(\omega)$  converges outer almost surely. Then it converges also in distribution. Since  $S_n$  converges in distribution as well, it follows that the sequence  $T_n(\omega)$  is convergent. ■

**A.1.4 Proposition (Hoffmann-Jørgensen inequalities).** *Let  $X_1, \dots, X_n$  be independent stochastic processes indexed by an arbitrary set. Then for any  $\lambda, \eta > 0$ ,*

$$P^*\left(\max_{k \leq n} \|S_k\| > 3\lambda + \eta\right) \leq P^*\left(\max_{k \leq n} \|S_k\| > \lambda\right)^2 + P^*\left(\max_{k \leq n} \|X_k\| > \eta\right).$$

If  $X_1, \dots, X_n$  are independent and symmetric, then also

$$P^*(\|S_n\| > 2\lambda + \eta) \leq 4P^*(\|S_n\| > \lambda)^2 + P^*\left(\max_{k \leq n} \|X_k\| > \eta\right).$$

**Proof.** Let  $A_k$  be the event that  $\|S_k\|^*$  is the first  $\|S_j\|^*$  that is strictly greater than  $\lambda$ . The (disjoint) union of  $A_1, \dots, A_n$  is the event that  $\max_{k \leq n} \|S_k\|^*$  is greater than  $\lambda$ . By the triangle inequality,  $\|S_j\|^* \leq$

$\|S_{k-1}\|^* + \|X_k\|^* + \|S_j - S_k\|^*$  for every  $j \geq k$ . On  $A_k$  the first term on the right is bounded by  $\lambda$ . Conclude that on  $A_k$

$$\max_{j \geq k} \|S_j\|^* \leq \lambda + \max_{k \leq n} \|X_k\|^* + \max_{j > k} \|S_j - S_k\|^*.$$

On  $A_k$  this remains true if the maximum on the left is taken over all  $\|S_j\|^*$ . Since the processes are independent, we obtain for every  $k$

$$\begin{aligned} & P\left(A_k, \max_{k \leq n} \|S_k\|^* > 3\lambda + \eta\right) \\ & \leq P\left(A_k, \max_{k \leq n} \|X_k\|^* > \eta\right) + P(A_k)P\left(\max_{j > k} \|S_j - S_k\|^* > 2\lambda\right). \end{aligned}$$

In the probability on the far right the variable  $\max_{j > k} \|S_j - S_k\|^*$  can be bounded by  $2 \max_{k \leq n} \|S_k\|^*$ . Next sum over  $k$  to obtain the first inequality of the proposition.

To prove the second inequality, first establish by the same method that

$$P(A_k, \|S_n\|^* > 2\lambda + \eta) \leq P(A_k, \max_{k \leq n} \|X_k\|^* > \eta) + P(A_k)P(\|S_n - S_k\|^* > \lambda).$$

In the probability on the far right, bound the variable  $\|S_n - S_k\|^*$  by  $\max_{k \leq n} \|S_n - S_k\|^*$ . Next sum over  $k$  to obtain

$$\begin{aligned} & P(\|S_n\|^* > 2\lambda + \eta) \\ & \leq P\left(\max_{k \leq n} \|X_k\|^* > \eta\right) + P\left(\max_{k \leq n} \|S_k\|^* > \lambda\right)P\left(\max_{k \leq n} \|S_n - S_k\|^* > \lambda\right). \end{aligned}$$

The processes  $S_k$  and  $S_n - S_k$  are the partial sums of the symmetric processes  $X_1, \dots, X_n$  and  $X_n, \dots, X_2$ , respectively. Apply Lévy's inequality to both probabilities on the far right to conclude the proof. ■

**A.1.5 Proposition (Hoffmann-Jørgensen inequalities for moments).** Let  $0 < p < \infty$  and suppose that  $X_1, \dots, X_n$  are independent stochastic processes indexed by an arbitrary index set  $T$ . Then there exist constants  $C_p$  and  $0 < u_p < 1$  such that

$$E^* \max_{k \leq n} \|S_k\|^p \leq C_p \left( E^* \max_{k \leq n} \|X_k\|^p + F^{-1}(u_p)^p \right),$$

where  $F^{-1}$  is the quantile function of the random variable  $\max_{k \leq n} \|S_k\|^*$ . Furthermore, if  $X_1, \dots, X_n$  are symmetric, then there exist constants  $K_p$  and  $0 < v_p < 1$  such that

$$E^* \|S_n\|^p \leq K_p \left( E^* \max_{k \leq n} \|X_k\|^p + G^{-1}(v_p)^p \right),$$

where  $G^{-1}$  is the quantile function of the random variable  $\|S_n\|^*$ . For  $p \geq 1$ , the last inequality is also valid for mean-zero processes (with different constants).

**Proof.** Take  $\lambda = \eta = t$  in the first inequality of the preceding proposition to find that, for any  $\tau > 0$ ,

$$\begin{aligned} E^* \max_{k \leq n} \|S_k\|^p &= 4^p \int P\left(\max_{k \leq n} \|S_k\|^* > 4t\right) d(t^p) \\ &\leq (4\tau)^p + 4^p \int_{\tau}^{\infty} P\left(\max_{k \leq n} \|S_k\|^* > t\right)^2 d(t^p) \\ &\quad + 4^p \int_{\tau}^{\infty} P\left(\max_{k \leq n} \|X_k\|^* > t\right) d(t^p) \\ &\leq (4\tau)^p + 4^p P\left(\max_{k \leq n} \|S_k\|^* > \tau\right) E^* \max_{k \leq n} \|S_k\|^p \\ &\quad + 4^p E^* \max_{k \leq n} \|X_k\|^p. \end{aligned}$$

Choosing  $\tau$  such that  $4^p P(\max_{k \leq n} \|S_k\|^* > \tau)$  is bounded by  $\frac{1}{2}$ , and rearranging terms, yield the claimed inequality. The second inequality can be proved in a similar manner, this time using the second inequality of the preceding proposition.

The inequality for mean-zero processes follows from the inequality for symmetric processes by symmetrization and desymmetrization: by Jensen's inequality,  $E^* \|S_n\|^p$  is bounded by  $E^* \|S_n - T_n\|^p$  if  $T_n$  is the sum of  $n$  independent copies of  $X_1, \dots, X_n$ . ■

Hoffmann-Jørgensen's inequality strengthens probability statements about sums to statements concerning expectations or  $p$ -moments, provided the corresponding moment of the maximum of the individual terms can be controlled. Thus it can be used to "invert" the consequences of Markov-type inequalities under an additional condition. A typical application is to a sequence of sums  $\sum_{i=1}^n X_{ni}$ . Boundedness in probability  $\|\sum_{i=1}^n X_{ni}\|^* = O_P(1)$  implies that  $G_n^{-1}(u) = O(1)$  for the sequence of quantile functions of the variables  $\|\sum_{i=1}^n X_{ni}\|^*$ . Conclude that the sequence of expectations  $E^* \|\sum_{i=1}^n X_{ni}\|^p$  is  $O(1)$  if the same is true for the sequence  $E^* \max_{1 \leq i \leq n} \|X_{ni}\|^p$ .

Hoffmann-Jørgensen's inequality may also be used to bound higher moments of sums in terms of a first moment plus a higher moment of the maximum of the individual terms. The first part of the following proposition is in this spirit, although its constant is much smaller than the constant obtainable from Hoffmann-Jørgensen's inequality.

**A.1.6 Proposition.** Let  $X_1, \dots, X_n$  be independent, mean-zero stochastic processes indexed by an arbitrary index set  $T$ . Then

$$\begin{aligned} \|\|S_n\|^*\|_{P,p} &\leq K \frac{p}{\log p} \left[ \|\|S_n\|^*\|_{P,1} + \|\max_{1 \leq i \leq n} \|X_i\|^*\|_{P,p} \right] \quad (p > 1), \\ \|\|S_n\|^*\|_{\psi_p} &\leq K_p \left[ \|\|S_n\|^*\|_{P,1} + \|\max_{1 \leq i \leq n} \|X_i\|^*\|_{\psi_p} \right] \quad (0 < p \leq 1), \\ \|\|S_n\|^*\|_{\psi_p} &\leq K_p \left[ \|\|S_n\|^*\|_{P,1} + \left( \sum_{i=1}^n \|\|X_i\|^*\|_{\psi_p}^q \right)^{1/q} \right] \quad (1 < p \leq 2). \end{aligned}$$

Here  $1/p + 1/q = 1$ , and  $K$  and  $K_p$  are a universal constant and a constant depending on  $p$  only, respectively.

**Proof.** The first part with the inferior constant  $24(2^{1/p})(3^p)$  follows from the preceding Hoffmann-Jørgensen inequality by noting that  $(1 - v)G^{-1}(v) \leq \int_v^1 G^{-1}(s) ds \leq E^* \|S_n\|$  for every  $v$ , and then substitution of  $v_p = 1 - 3^{-p}/8$  and  $K_p = 2(3^p)$ . The proofs of the first part with the improved constant  $p/\log p$  and of the second and third parts are long and rely on the isoperimetric methods developed by Talagrand (1989). See Ledoux and Talagrand (1991), pages 172–175. ■

Another consequence of Hoffmann-Jørgensen's inequality is the following equivalence of moments of the supremum of sums of independent stochastic processes and the same moment of the maximal summand.

**A.1.7 Proposition.** Let  $0 < p < \infty$  and suppose that  $X_1, X_2, \dots$  are independent stochastic processes indexed by an arbitrary set  $T$ . If  $\sup_{n \geq 1} \|S_n\|^* < \infty$  almost surely, then

$$E^* \sup_{n \geq 1} \|S_n\|^p < \infty$$

if and only if

$$E^* \sup_{n \geq 1} \|X_n\|^p < \infty.$$

**Proof.** Since  $\|X_n\| \leq \|S_n\| + \|S_{n-1}\|$ , finiteness of the first expectation clearly implies finiteness of the second expectation. For any fixed  $n$ , it follows from Proposition A.1.5 that

$$E^* \max_{k \leq n} \|S_k\|^p \leq C_p \left( E^* \max_{k \leq n} \|X_k\|^p + F^{-1}(u_p)^p \right),$$

where  $F^{-1}$  is the quantile function of the random variable  $\max_{k \leq n} \|S_k\|^*$ , and  $0 < u_p < 1$ . Since  $\sup_{n \geq 1} \|S_n\|$  is assumed finite almost surely, there is an  $M$  so that  $F^{-1}(u_p) \leq M$  for all  $n$ . Hence letting  $n \rightarrow \infty$  in the last display shows that finiteness of the second expectation implies finiteness of the first expectation. ■

An interesting corollary of this proposition for normalized sums of independent stochastic processes is as follows.

**A.1.8 Corollary.** Let  $0 < p < \infty$  and let  $\{a_n\}$  be a sequence of positive numbers that increases to infinity. Let  $X_1, X_2, \dots$  be independent stochastic processes indexed by an arbitrary set  $T$ . If  $\sup_{n \geq 1} (\|S_n\|/a_n)^* < \infty$  almost surely, then

$$E^* \sup_{n \geq 1} \frac{\|S_n\|^p}{a_n^p} < \infty$$

if and only if

$$\mathbb{E}^* \sup_{n \geq 1} \frac{\|X_n\|^p}{a_n^p} < \infty.$$

**Proof.** See Ledoux and Talagrand (1991), page 159. ■

Verification of the second statement in the preceding corollary can be carried out by use of Problem 2.3.5.

**A.1.9 Proposition (Hoeffding's inequality).** Let  $\{c_1, \dots, c_N\}$  be elements of an arbitrary vector space  $\mathbb{V}$ , and let  $U_1, \dots, U_n$  and  $V_1, \dots, V_n$  denote samples of size  $n \leq N$  drawn without and with replacement, respectively, from  $\{c_1, \dots, c_N\}$ . Then, for every convex function  $\phi: \mathbb{V} \rightarrow \mathbb{R}$ ,

$$\mathbb{E}\phi\left(\sum_{j=1}^n U_j\right) \leq \mathbb{E}\phi\left(\sum_{j=1}^n V_j\right).$$

**Proof.** See Hoeffding (1963) and Marshall and Olkin (1979), Corollary A.2.e, page 339. ■

**A.1.10 Proposition (Contraction principle).** Let  $X_1, \dots, X_n$  be arbitrary stochastic processes and  $\gamma_1, \dots, \gamma_n$  arbitrary, real-valued (measurable) random variables with  $0 \leq \gamma_i \leq 1$ . Let  $\xi_1, \dots, \xi_n$  be independent, real random variables with zero means independent of  $(X_1, \dots, X_n, \gamma_1, \dots, \gamma_n)$ . Then

$$\mathbb{E}^* \left\| \sum_{i=1}^n \xi_i \gamma_i X_i \right\| \leq \mathbb{E}^* \left\| \sum_{i=1}^n \xi_i X_i \right\|.$$

**Proof.** Since the  $\gamma_i$  can be taken out one at a time, it suffices to show that

$$\mathbb{E}^* \|\xi \gamma X + Y\| \leq \mathbb{E}^* \|\xi X + Y\|$$

whenever  $\xi$  has zero mean and is independent of  $(\gamma, X, Y)$ . By the triangle inequality, the left side is bounded by

$$\mathbb{E}^* \|(\xi X + Y)\gamma\| + \mathbb{E}^* \|Y(1 - \gamma)\| = \mathbb{E}^* \|(\xi X + Y)\| \gamma + \mathbb{E}^* \|Y\| (1 - \gamma).$$

The second term on the right can be written

$$\mathbb{E}_{Y,\gamma}^* \|Y + (\mathbb{E}_\xi \xi) X\| (1 - \gamma).$$

By Jensen's inequality and Fubini's theorem, it is bounded by  $\mathbb{E}^* \|Y + \xi X\| (1 - \gamma)$ . The result follows since  $\gamma$  is measurable. ■

## Problems and Complements

- In the first inequality of Proposition A.1.5, the constants  $C_p = 2(4^p)$  and  $u_p = 1 - 4^{-p}/2$  will do. The second inequality is true for symmetric processes with  $K_p = 2(3^p)$  and  $v_p = 1 - 3^{-p}/8$ ; for mean-zero processes with  $K_p = 4(6^p)$  and  $v_p = 1 - 3^{-p}/16$ .

## A.2

# Gaussian Processes

Since Gaussian processes arise as the limits in distribution of empirical processes, their properties are of importance for many of the developments in Parts 2 and 3. Inequalities for Gaussian processes have also been used in a number of proofs involving multiplier processes with Gaussian multipliers. In this chapter we discuss the most important results.

### A.2.1 Inequalities and Gaussian Comparison

Throughout this section  $X$  and  $Y$  denote separable, Gaussian stochastic processes indexed by a semimetric space  $T$ , and  $\|X\|$  is the supremum  $\sup_{t \in T} |X_t|$ . The process  $X$  is *mean-zero* if  $\mathbb{E}X_t = 0$  for all  $t \in T$ . Let  $M(X)$  denote a median of  $\|X\|$ , defined by the two requirements

$$\mathbb{P}(\|X\| \leq M(X)) \geq \frac{1}{2} \quad ; \quad \mathbb{P}(\|X\| \geq M(X)) \geq \frac{1}{2}.$$

(In the proof of the following proposition, it is shown that  $M(X)$  is unique.) Furthermore, define

$$\sigma^2(X) = \sup_{t \in T} \text{var } X_t.$$

The following proposition shows that the distribution of the supremum of a zero-mean Gaussian process has sub-Gaussian tails, whenever it is finite.

**A.2.1 Proposition (Borell's inequality).** Let  $X$  be a mean-zero, separable Gaussian process  $X$  with finite median. Then for every  $\lambda > 0$ ,

$$\begin{aligned} P(|\|X\| - M(X)| \geq \lambda) &\leq \exp\left(-\frac{\lambda^2}{2\sigma^2(X)}\right), \\ P(|\|X\| - E\|X\|| \geq \lambda) &\leq 2 \exp\left(-\frac{\lambda^2}{2\sigma^2(X)}\right), \\ P(\|X\| \geq \lambda) &\leq 2 \exp\left(-\frac{\lambda^2}{8E\|X\|^2}\right). \end{aligned}$$

**Proof.** The proof is based on finite-dimensional approximation and the concentration inequalities for the finite-dimensional, standard normal distribution given in the lemma ahead.

First, consider the case that the index set  $T$  consists of finitely many points  $t_1, \dots, t_k$ . Then the process  $X$  can be represented as  $AZ$ , for  $Z$  a standard normal vector and  $A$  the symmetric square root of the covariance matrix of  $X$ . By the Cauchy-Schwarz inequality,

$$\max_{1 \leq i \leq k} |Az|_i \leq \max_{1 \leq i \leq k} \|A_i\| \|z\| = \max_{1 \leq i \leq k} (A_{ii}^2)^{1/2} \|z\|.$$

This implies that the function  $f(z) = \max_i |Az|_i$  is Lipschitz of norm bounded by  $\sup_t \sigma(X_t) = \sigma(X)$ . Apply Lemma A.2.2 to both  $f$  and  $-f$ , and combine the results to obtain the theorem for the case where the index set is finite.

Since  $X$  is separable by assumption, the supremum  $\|X\|$  is the almost-sure, monotone limit of a sequence of finite suprema. The supremum  $M$  of the sequence of medians can be seen to be a median (the smallest) of  $\|X\|$ . Approximate  $\|X\| - M(X)$  by a sequence of similar objects for finite suprema to obtain the first inequality of the theorem with  $M = M(X)$ . The proof of this inequality is complete if it is shown that  $M$  is the only median of  $\|X\|$ .

Since the median of  $|X_t|$  is bounded above by the median of  $\|X\|$ , it follows that  $P(|X_t| \leq M(X)) \geq \frac{1}{2}$  for every  $t$ . Taking into account the normal distribution of  $X_t$ , we obtain that  $\sigma(X_t)$  is bounded above by  $M(X)/\Phi^{-1}(3/4)$ . Therefore,  $\sigma(X)$  is finite and the exponential inequality obtained previously is nontrivial. The argument actually shows that both the left-tail probability  $P(\|X\| \leq M - \lambda)$  and the right-tail probability  $P(\|X\| \geq M + \lambda)$  are bounded by 1/2 times the exponential upper bound. This means that these probabilities are strictly less than 1/2 for  $\lambda > 0$ , whence  $M$  is a unique median of  $\|X\|$ .

To obtain the second inequality of the theorem, we note first that  $E\|X\|$  is finite in view of the exponential tail bound for  $\|X\| - M(X)$ . Next we use the following lemma and take limits along finite subsets as before.

The third inequality is trivially satisfied if  $0 \leq \lambda < 2E\|X\|$ , because in that case the exponential is larger than  $\exp(-\frac{1}{2}) \geq 0.6$ . For  $\lambda > 2E\|X\|$ ,

the probability is bounded by  $P(\|X\| > E\|X\| + \lambda/2)$ , which can be further bounded by the second inequality. ■

**A.2.2 Lemma.** *Let  $Z$  be a  $d$ -dimensional random vector with the standard normal distribution. Then for every Lipschitz function  $f: \mathbb{R}^d \mapsto \mathbb{R}$  with  $\|f\|_{\text{Lip}} \leq 1$ ,*

$$\begin{aligned} P(f(Z) - \text{med } f(Z) > \lambda) &\leq \frac{1}{2} \exp(-\frac{1}{2}\lambda^2), \\ P(f(Z) - Ef(Z) > \lambda) &\leq \exp(-\frac{1}{2}\lambda^2). \end{aligned}$$

**Proof.** For the first inequality consider, the set  $A = \{z: f(z) \leq \text{med } f(Z)\}$ . Since  $f$  has Lipschitz norm bounded by 1, the set  $A^\lambda$  of points at distance at most  $\lambda$  from  $A$  is contained in the set  $\{z: f(z) \leq \text{med } f(Z) + \lambda\}$ . It follows that

$$P(f(Z) \leq \text{med } f(Z) + \lambda) \geq P(Z \in A^\lambda).$$

By definition of the median, the set  $A$  has probability at least 1/2 under the standard normal distribution. According to the isoperimetric inequality for the normal distribution,<sup>‡</sup> a half-space  $H$  with boundary at the origin is an extreme set in the sense that for any set  $A$  with probability at least 1/2 we have  $P(Z \in A^\lambda) \geq P(Z \in H^\lambda)$  for every  $\lambda > 0$ . The proof of the first inequality of the lemma is complete upon noting that  $P(Z \in H^\lambda) = \Phi(\lambda)$  and  $1 - \Phi(\lambda) \leq \frac{1}{2} \exp(-\frac{1}{2}\lambda^2)$ .

For the proof of the second inequality, assume without loss of generality that  $Ef(Z) = 0$ . An arbitrary Lipschitz function  $f$  can be approximated by arbitrarily smooth functions with compact supports and of no larger Lipschitz norms. Therefore, it is no loss of generality to assume that  $f$  is sufficiently regular. In particular assume without loss of generality that  $f$  is differentiable with gradient  $\nabla f$  uniformly norm bounded by 1.

Let  $Z_t$  be normally distributed with mean zero and covariance matrix  $(1 - e^{-2t})I$ . For  $t \geq 0$ , define functions  $P_t f$  by

$$P_t f(x) = Ef(e^{-t}x + Z_t).$$

The operators  $P_t$  form a semigroup ( $P_0 = I$  and  $P_s P_t = P_{s+t}$ ) on a domain that includes all sufficiently regular functions: the Ornstein-Uhlenbeck or Hermite semigroup defined via Mehler's formula. The *generator* of the group is the derivative  $A = d/dt P_t|_{t=0}$  and has the property that  $d/dt P_t = AP_t$ . We shall need the integration-by-parts formula

$$-Ef(Z)Ag(Z) = E\langle \nabla f(Z), \nabla g(Z) \rangle.$$

This is valid for sufficiently regular functions  $f$  and  $g$ .

For fixed  $r > 0$ , define  $G(t) = E\exp(rP_t f(Z))$ . Then  $G(0) = E\exp(rf(Z))$  is the moment-generating function of  $f(Z)$  and  $G(\infty) = 1$ .

<sup>‡</sup> See Ledoux (1995) for an introduction.

Differentiating under the expectation and using the integration-by-parts formula, we obtain

$$\begin{aligned} -G'(t) &= -r \mathbb{E} e^{rP_t f(Z)} AP_t f(Z) \\ &= r \mathbb{E} \langle \nabla e^{rP_t f(Z)}, \nabla P_t f(Z) \rangle = r^2 \mathbb{E} e^{rP_t f(Z)} \|\nabla P_t f(Z)\|^2. \end{aligned}$$

Differentiating under the expectation in the definition of  $P_t f$ , we see that the gradient  $\nabla P_t f(x)$  is equal to  $\mathbb{E} \nabla f(e^{-t}x + Z_t)e^{-t}$  and its norm is bounded above by  $\sup_x \|\nabla f(x)\| e^{-t}$ . Substitution in the preceding display yields the inequality  $-G'(t) \leq r^2 e^{-2t} G(t)$ . Equivalently,

$$(\log G)'(t) \geq (\frac{1}{2}r^2 e^{-2t})'.$$

Since the functions  $\log G$  and  $\frac{1}{2}r^2 e^{-2t}$  are both zero at  $\infty$  and hence equal, it follows that  $\log G$  is smaller than  $\frac{1}{2}r^2 e^0$  at zero. This concludes the proof that  $\mathbb{E} \exp r f(Z) = G(0) \leq \exp(\frac{1}{2}r^2)$ .

By Markov's inequality, it follows that  $\mathbb{P}(f(Z) \geq \lambda) \leq \exp(\frac{1}{2}r^2 - r\lambda)$  for every  $r > 0$ . Optimize to conclude the proof of the lemma. ■

The process  $X$  having bounded sample paths is the same as  $\|X\|$  being a finite random variable. In that case the median  $M(X)$  is certainly finite and  $\sigma(X)$  is finite by the argument in the proof of the preceding proposition. Next, the inequalities in the preceding proposition show that  $\|X\|$  has moments of all orders. In fact, we have the following proposition.

**A.2.3 Proposition.** *Let  $X$  be a mean-zero, separable Gaussian process such that  $\|X\|$  is finite almost surely. Then*

$$\mathbb{E} \exp(\alpha \|X\|^2) < \infty \quad \text{if and only if} \quad \alpha 2\sigma^2(X) < 1.$$

**Proof.** The sufficiency of the condition  $\alpha 2\sigma^2(X) < 1$  follows by integrating Borell's inequality. The necessity may be proved by considering the individual  $X_t$ , whose normal distributions can be handled explicitly. ■

Thus, for Gaussian processes, finiteness of lower moments implies finiteness of higher moments. Higher moments can even be bounded by lower ones up to universal constants. The following proposition reverses the usual Liapunov inequality for moments.

**A.2.4 Proposition.** *There exist constants  $K_{p,q}$  depending on  $0 < p \leq q < \infty$  only such that*

$$(\mathbb{E} \|X\|^q)^{1/q} \leq K_{p,q} (\mathbb{E} \|X\|^p)^{1/p},$$

for any separable Gaussian process  $X$  for which  $\|X\|$  is finite almost surely.

For a centered Gaussian process  $\{X_t: t \in T\}$ , let  $\rho$  denote its standard deviation semimetric on  $T$  defined by

$$\rho(s, t) = \sigma(X_s - X_t), \quad s, t \in T.$$

Let  $N(\varepsilon, T, \rho)$  be the covering number of  $T$  with respect to  $\rho$ . By Corollary 2.2.8, the expectation  $E\|X\|$  can be bounded above by an entropy integral with respect to this metric. Sudakov's inequality gives a bound in the opposite direction.

**A.2.5 Proposition (Sudakov's inequality).** *For every mean-zero, separable Gaussian process  $X$ ,*

$$\sup_{\varepsilon > 0} \varepsilon \sqrt{\log N(\varepsilon, T, \rho)} \leq 3E\|X\|.$$

Moreover, if  $X$  has almost all sample paths bounded and uniformly continuous on  $(T, \rho)$ , then  $\varepsilon \sqrt{\log N(\varepsilon, T, \rho)} \rightarrow 0$  as  $\varepsilon \rightarrow 0$ .

**Proofs.** See Ledoux and Talagrand (1991), pages 79–81. For the second part, also see Lemma 2.10.15. ■

The distribution of a mean-zero Gaussian process is completely determined by its covariance function. According to Anderson's lemma (Lemma 3.11.4), a smaller covariance function indicates that the process is more concentrated near zero. The following theorem is in the same spirit and shows that smaller variances of differences imply a stochastically smaller maximum value.

**A.2.6 Proposition (Slepian, Fernique, Marcus, and Shepp).** *Let  $X$  and  $Y$  be separable, mean-zero Gaussian processes indexed by a common index set  $T$  such that*

$$E(X_s - X_t)^2 \leq E(Y_s - Y_t)^2, \quad \text{for all } s, t \in T.$$

Then

$$P\left(\sup_{t \in T} X_t \geq \lambda\right) \leq P\left(\sup_{t \in T} Y_t \geq \lambda\right), \quad \text{for all } \lambda > 0.$$

Consequently,  $E \sup_{t \in T} X_t \leq E \sup_{t \in T} Y_t$  and  $E\|X\| \leq 2E\|Y\|$ . If  $T$  is a compact metric space and  $Y$  has a version with continuous sample paths, then so does  $X$ .

**Proof.** See Ledoux and Talagrand (1991). ■

### A.2.2 Exponential Bounds

To obtain exponential bounds for  $\|X\| = \sup_{t \in T} |X_t|$  without the centering by  $M(X)$  or  $E\|X\|$  as in Proposition A.2.1, requires hypotheses giving control of  $E\|X\|$  (or equivalently of  $M(X)$ ). The following theorem, due to Talagrand, is an example of results of this type.

For a given Gaussian process  $X$  indexed by an arbitrary set  $T$ , we denote by  $\rho$  its “natural semimetric”  $\rho(s, t) = \sigma(X_s - X_t)$ . Recall that  $\sigma(X)$  is the supremum of the standard deviations  $\sigma(X_t)$ . Let  $\bar{\Phi}(z) = \int_z^\infty \phi(x) dx \leq z^{-1}\phi(z)$  be the tail probability of a standard normal variable.

**A.2.7 Proposition.** *Let  $X$  be a separable, mean-zero Gaussian process such that for some  $K > \sigma(X)$ , some  $V > 0$ , and some  $0 < \varepsilon_0 \leq \sigma(X)$ ,*

$$N(\varepsilon, T, \rho) \leq \left( \frac{K}{\varepsilon} \right)^V, \quad 0 < \varepsilon < \varepsilon_0.$$

Then there exists a universal constant  $D$  such that, for all  $\lambda \geq \sigma^2(X)(1 + \sqrt{V})/\varepsilon_0$ ,

$$P\left(\sup_{t \in T} X_t \geq \lambda\right) \leq \left( \frac{DK\lambda}{\sqrt{V}\sigma^2(X)} \right)^V \bar{\Phi}\left(\frac{\lambda}{\sigma(X)}\right).$$

This is closely related to Theorems 2.14.9 and 2.14.13 for empirical processes. Another result of Talagrand for Gaussian processes which is parallel to Theorem 2.14.14 in the case of empirical processes, is as follows.

**A.2.8 Proposition.** *Let  $X$  be a separable, zero-mean Gaussian process, and for  $\delta > 0$ , let*

$$T_\delta = \{t \in T : EX_t^2 \geq \sigma^2(X) - \delta^2\}.$$

Let  $V \geq W \geq 1$ , and suppose that, for all  $0 < \delta^2 \leq \sigma^2(X)$  and all  $0 < \epsilon \leq \delta(1 + \sqrt{V})/\sqrt{W}$ ,

$$(A.2.9) \quad N(\epsilon, T_\delta, \rho) \leq K\delta^W \epsilon^{-V}.$$

Then there exists a universal constant  $D$  such that, for  $\lambda \geq 2\sigma(X)\sqrt{W}$ ,

$$P\left(\sup_{t \in T} X_t \geq \lambda\right) \leq K \frac{W^{W/2}}{V^{V/2}} D^{V+W} \left( \frac{\lambda}{\sigma^2(X)} \right)^{V-W} \bar{\Phi}\left(\frac{\lambda}{\sigma(X)}\right).$$

For more results in this vein, see Talagrand (1994), Samorodnitsky (1991), Adler (1990) and Adler and Brown (1986). For particular Gaussian random fields, both the renewal theoretic methods of Siegmund (1988, 1992), and the “Poisson clump heuristic” methods of Aldous (1989), yield results on the asymptotic behavior of  $P(\sup_{t \in T} X_t \geq \lambda)$ , as  $\lambda \rightarrow \infty$ , that often seem to provide accurate approximations for  $\lambda$  on the order of  $1.5\sigma(X)$ . Here are some particular examples of the above theorems with connections to the results of Hogan and Siegmund (1986), Siegmund (1988, 1992), and Aldous (1989).

**A.2.10 Example (Brownian sheet on  $[0, 1]^2$ ).** The standard Brownian sheet  $B$  is a zero-mean Gaussian process indexed by  $T = [0, 1]^2$ , with covariance function

$$\text{cov}(B(s), B(t)) = (s_1 \wedge t_1)(s_2 \wedge t_2).$$

Then  $\sigma^2(B) = 1$ , and (A.2.9) holds with  $V = 4$  and  $W = 4$ . Hence, for some constant  $M$ , and  $\lambda > 0$ ,

$$P\left(\sup_{t \in T} B_t \geq \lambda\right) \leq M\lambda^{-1} \exp\left(-\frac{\lambda^2}{2}\right).$$

In fact, it was shown by Goodman (1976) that, for all  $\lambda > 0$ ,

$$4 \int_{-\lambda}^{\infty} z \bar{\Phi}(z) dz \leq P\left(\sup_{t \in T} B_t \geq \lambda\right) \leq 4\bar{\Phi}(\lambda).$$

These inequalities imply that, for  $\lambda \rightarrow \infty$ ,

$$P\left(\sup_{t \in T} B_t \geq \lambda\right) \sim 4\bar{\Phi}(\lambda) = 4P(B(1, 1) \geq \lambda) \sim \sqrt{\frac{8}{\pi}}\lambda^{-1} \exp\left(-\frac{\lambda^2}{2}\right).$$

**A.2.11 Example (Brownian bridge on  $[0, 1]^2$ ).** The Brownian sheet  $B^0$  pinned down to 0 at  $t = (1, 1)$  is a zero-mean Gaussian process indexed by  $[0, 1]^2$ , with covariance function

$$\text{cov}(B^0(s), B^0(t)) = (s_1 \wedge t_1)(s_2 \wedge t_2) - s_1 s_2 t_1 t_2.$$

Alternatively,  $B^0$  can be defined from the Brownian sheet of Example A.2.10, by setting

$$B^0(t_1, t_2) = B(t_1, t_2) - t_1 t_2 B(1, 1).$$

This is just the  $P$ -Brownian bridge process  $\mathbb{G}_P$  indexed by the collection of sets  $\{[0, t] : t \in [0, 1]\}$ , with  $P$  equal to the uniform distribution (Lebesgue measure) on  $[0, 1]^2$ . Then  $\sigma^2(B^0) = 1/4$ , and this supremum is achieved for every  $t \in [0, 1]^2$  with  $t_1 t_2 = 1/2$ . It can be shown that (A.2.9) holds with  $V = 4$  and  $W = 1$ . Therefore, for some constant  $M$ ,

$$P\left(\sup_{t \in T} B_t^0 \geq \lambda\right) \leq M\lambda^2 \exp(-2\lambda^2).$$

It has been shown by Hogan and Siegmund (1986) and also by Aldous (1989), page 202, that as  $\lambda \rightarrow \infty$ ,

$$P\left(\sup_{t \in T} B_t^0 \geq \lambda\right) \sim (4 \log 2)\lambda^2 \exp(-2\lambda^2).$$

In this case Goodman (1976) established the lower bound

$$P\left(\sup_{t \in T} B_t^0 \geq \lambda\right) \geq (1 + 2\lambda^2) \exp(-2\lambda^2), \quad \lambda > 0.$$

The asymptotic formula of Hogan and Siegmund becomes greater than Goodman's lower bound for  $\lambda \geq (2(\log 4 - 1))^{-1/2} = 1.138\dots$ . For this and some numerical comparisons, see Adler (1990), Chapter V.

**A.2.12 Example (Tucked brownian sheet on  $[0, 1]^2$ ).** The tucked Brownian sheet is the zero-mean Gaussian process indexed by  $[0, 1]^2$ , with covariance function

$$\text{cov}(B^{00}(s), B^{00}(t)) = (s_1 \wedge t_1 - s_1 t_1)(s_2 \wedge t_2 - s_2 t_2).$$

This can be obtained from the Brownian sheet given in Example A.2.10, by pinning it down to 0 on the entire boundary of  $[0, 1]^2$ , i.e.

$$B^{00}(t_1, t_2) = B(t_1, t_2) - t_1 B(1, t_2) - t_2 B(t_1, 1) + t_1 t_2 B(1, 1).$$

This process arises in testing independence and in connection with the estimation of copula functions; see Chapters 3.8 and 3.9. In this case  $\sigma^2(B^{00}) = 1/16$ , and this supremum is uniquely achieved for  $t = (1/2, 1/2)$ . It can be shown that (A.2.9) holds with  $V = 4$  and  $W = 2$ . Therefore, for some constant  $M$ ,

$$P\left(\sup_{t \in T} B_t^{00} \geq \lambda\right) \leq M \lambda \exp(-8\lambda^2).$$

It follows from the methods of Siegmund (1988, 1992), or alternatively from Aldous (1989), that as  $\lambda \rightarrow \infty$ ,

$$P\left(\sup_{t \in T} B_t^{00} \geq \lambda\right) \sim (4\sqrt{2\pi})\lambda \exp(-8\lambda^2).$$

**A.2.13 Example (Kiefer-Müller process on  $[0, 1]^2$ ).** The Kiefer-Müller process  $Z$  on  $[0, 1]^2$  can be obtained from a Brownian sheet  $B$  by way of

$$Z(t_1, t_2) = B(t_1, t_2) - t_2 B(t_1, 1).$$

Then  $Z$  has covariance function

$$\text{cov}(Z(s_1, t_1), Z(s_2, t_2)) = (s_1 \wedge s_2)(t_1 \wedge t_2 - t_1 t_2).$$

This process is a special case of the  $P$ -Kiefer process in Chapter 2.12 with  $P$  the uniform distribution on  $[0, 1]$  and  $\mathcal{F} = \{1_{[0,t]} : t \in [0, 1]\}$ . In this case  $\sigma^2(Z) = 1/4$ , and this is uniquely achieved for  $(s, t) = (1, 1/2)$ . It can be shown that (A.2.9) holds with  $V = 4$  and  $W = 3$ . Therefore, for some constant  $M$ ,

$$P\left(\sup_{t \in T} Z_t \geq \lambda\right) \leq M \exp(-2\lambda^2).$$

It follows from the methods of Siegmund (1988, 1992), or alternatively from Aldous (1989), that as  $\lambda \rightarrow \infty$ ,

$$P\left(\sup_{t \in T} Z_t \geq \lambda\right) \sim 2 \exp(-2\lambda^2).$$

**A.2.14 Example (Brownian bridge indexed by convex subsets of  $[0, 1]^2$ ).** Consider the  $P$ -Brownian bridge process with  $P$  uniform on  $[0, 1]^2$  indexed by the collection  $\mathcal{C}$  of all convex subsets of  $[0, 1]^2$ . It follows from Corollary 2.7.9 (with  $d = r = 2$ ) that

$$(A.2.15) \quad \log N_{[]}(\varepsilon, \mathcal{C}, L_2(P)) \leq K\left(\frac{1}{\varepsilon}\right),$$

and hence  $\mathcal{C}$  is  $P$ -preGaussian. Although neither Theorem A.2.7 nor A.2.8 apply in this case, the methods of Samorodnitsky (1991) apply, and show that for some constants  $C$  and  $D$ ,

$$(A.2.16) \quad P\left(\sup_{C \in \mathcal{C}} \mathbb{G}_P(C) \geq \lambda\right) \leq C \exp(D\lambda^{2/3}) \exp(-2\lambda^2).$$

More generally, the methods of Samorodnitsky (1991) show that if the power 1 in (A.2.15) is replaced by  $V$ , then the power  $2/3$  in (A.2.16) must be replaced by  $2V/(V + 2)$ .

### A.2.3 Majorizing Measures

A separable, zero-mean Gaussian process  $\{X_t: t \in T\}$  indexed by an arbitrary index set  $T$  is sub-Gaussian with respect to its standard deviation semimetric  $\rho(s, t) = \sigma(X_s - X_t)$ . Thus Corollary 2.2.8 yields for every  $\delta > 0$

$$E \sup_{\rho(s,t) < \delta} |X_s - X_t| \lesssim \int_0^\delta \sqrt{\log N(\varepsilon, T, \rho)} d\varepsilon,$$

where  $N(\varepsilon, T, \rho)$  is the covering number of  $T$  with respect to  $\rho$ . If the entropy integral on the right is finite, then it follows immediately that the process is uniformly  $\rho$ -continuous in mean. This conclusion may be strengthened to the uniform continuity of almost all sample paths with the help of the Borel-Cantelli lemma (Problem 2.2.17).

While a finite entropy integral is sufficient for the sample-path continuity of a Gaussian process, it is not necessary. *Majorizing measures* may be considered a refinement of entropy numbers and yield a necessary and sufficient condition for sample-path continuity.

**A.2.17 Proposition.** Let  $\{X_t: t \in T\}$  be a separable zero-mean Gaussian process with standard deviation semimetric  $\rho$ . Then almost all sample paths  $t \rightarrow X_t$  are uniformly  $\rho$ -continuous and bounded if and only if

$$(A.2.18) \quad \limsup_{\delta \downarrow 0} \int_0^\delta \sqrt{\log \frac{1}{\mu(B(t, \varepsilon))}} d\varepsilon = 0,$$

for some Borel probability measure  $\mu$  on  $(T, \rho)$ . Here  $B(t, \varepsilon)$  is the  $\rho$ -ball of radius  $\varepsilon$  around  $t$ .

A measure  $\mu$  for which the integral in the theorem is finite is called a *majorizing measure*. A further result on majorizing measures is that finiteness of the integral characterizes the boundedness of the sample paths. Among the bounded processes the uniformly continuous processes are characterized by majorizing measures with the additional property that the majorizing measure integral is continuous at zero, as in (A.2.18). A chaining argument may establish that for any probability measure  $\mu$  and every  $\delta, \eta > 0$ ,<sup>b</sup>

$$\begin{aligned} \mathbb{E} \sup_t |X_t - X_{t_0}| &\lesssim \sup_t \int_0^\infty \sqrt{\log \frac{1}{\mu(B(t, \varepsilon))}} d\varepsilon, \\ \mathbb{E} \sup_{\rho(s, t) < \delta} |X_s - X_t| &\lesssim \sup_t \int_0^\eta \sqrt{\log \frac{1}{\mu(B(t, \varepsilon))}} d\varepsilon + \delta \sqrt{N(\eta, T, \rho)}. \end{aligned}$$

This explains the name “majorizing measure” and readily yields one direction of the preceding proposition. In the other direction, the proposition is harder to prove. This converse part is used in the proof of Theorem 2.11.11 in combination with the following lemma.

**A.2.19 Lemma.** *Let  $T$  be a semimetric space and  $\mu$  be a Borel probability measure on  $T$  such that (A.2.18) holds. Then there exists a probability measure  $m$  on  $T$  and a sequence of nested, measurable partitions  $T = \cup_i T_{qi}$  such that  $\text{diam } T_{qi} \leq 2^{-q}$  for every  $i$  and*

$$\lim_{q_0 \rightarrow \infty} \sup_{t \in T} \sum_{q > q_0} 2^{-q} \sqrt{\log \frac{1}{m(T_q t)}} = 0,$$

where  $T_q t$  is the partitioning set  $T_{qi}$  at level  $q$  to which  $t$  belongs.

**Proof.** The set  $T$  is totally bounded: the existence of infinitely many disjoint balls  $B(t_i, \delta)$  of fixed radius  $\delta > 0$  would require that  $\mu(B(t_i, \delta)) \rightarrow 0$  as  $i \rightarrow \infty$ . However,

$$\sup_t \delta \sqrt{\log \frac{1}{\mu(B(t, \delta))}} \leq \sup_t \int_0^\delta \sqrt{\log \frac{1}{\mu(B(t, \varepsilon))}} d\varepsilon.$$

The right-hand side is finite by assumption.

Fix  $\delta > 0$ . Find a point  $t_1$  that (nearly) maximizes  $\mu(B(t, \delta))$  over  $t \in T$ . Precisely, let  $\mu(B(t_1, \delta)) \geq r \sup_t \mu(B(t, \delta))$  for some fixed  $0 < r \leq 1$ . Set  $m\{t_1\} = \mu(B(t_1, \delta))$ , and let  $T_1 = B(t_1, 2\delta)$  be the ball of radius  $2\delta$  around  $t_1$ . For every  $t \in T_1$  (even for every  $t \in T$ ),

$$r \mu(B(t, \delta)) \leq \mu(B(t_1, \delta)) = m\{t_1\} \leq m(T_1).$$

---

<sup>b</sup> Ledoux and Talagrand (1991), pages 320–321, and (11.15) on page 317.

Next, find a  $t_2$  that nearly maximizes  $\mu(B(t, \delta))$  over  $t \in T - T_1$ . Set  $m\{t_2\} = \mu(B(t_2, \delta))$ , and let  $T_2 = B(t_2, 2\delta) - T_1$ . For every  $t \in T_2$

$$r \mu(B(t, \delta)) \leq \mu(B(t_2, \delta)) = m\{t_2\} \leq m(T_2).$$

Continue this process, constructing disjoint sets  $T_1, T_2, \dots, T_I$  until the set of  $t$  with  $\rho(t, t_i) \geq 2\delta$  for every  $i$  is empty. Then  $T = \cup_i T_i$  is a finite partition of  $T$  in sets of diameter at most  $2\delta$  and  $m(T_i) \geq r \mu(B(t, \delta))$  for every  $t \in T_i$ .

Apply this construction for  $\delta = 2^{-q-1}$  and every natural number  $q$ . This yields a sequence of partitions  $T = \cup_i \bar{T}_{qi}$  and measures  $\bar{m}_q$  such that

$$\sum_{q>q_0} 2^{-q} \sqrt{\log \frac{1}{\bar{m}_q(\bar{T}_q t)}} \leq \sum_{q>q_0} 2^{-q} \sqrt{\log \frac{1}{r \mu(B(t, 2^{-q-1}))}}.$$

The right side is bounded by a multiple of

$$\int_0^{2^{-q_0}} \sqrt{\log 1/\mu(B(t, \varepsilon))} d\varepsilon$$

if  $\mu(B(t, \varepsilon))$  is bounded away from 1, which may be assumed.

The present sequence of partitions need not be nested. Replace the  $q$ th partition by the partition in the sets  $\cap_{p=1}^q \bar{T}_{p,i_p}$ , where  $i_1, \dots, i_q$  range over all possible values. For each set in the  $q$ th partition, define

$$m_q\left(\cap_{p=1}^q \bar{T}_{p,i_p}\right) = \prod_{p=1}^q \bar{m}_p(\bar{T}_{p,i_p}).$$

(The exact location of the mass is irrelevant.) Next, set  $m = \sum_q 2^{-q} m_q$ . Then  $m$  is a subprobability measure and

$$\sum_{q>q_0} 2^{-q} \sqrt{\log \frac{1}{m(T_q t)}} \leq \sum_{q>q_0} 2^{-q} \sqrt{\log 2^q + \sum_{p=1}^q \log \frac{1}{\bar{m}_p(\bar{T}_p t)}}.$$

This converges to zero uniformly in  $t$  as  $q_0 \rightarrow \infty$ . ■

#### A.2.4 Further Results

Many  $M$ -estimators are asymptotically distributed as the point of maximum of a Gaussian process. This point of maximum is not always easy to evaluate. The following proposition shows that a point of absolute maximum is unique, if it exists.

**A.2.20 Proposition.** *Let  $\{X_t: t \in T\}$  be a Gaussian process with continuous sample paths, indexed by a  $\sigma$ -compact metric space  $T$ . If  $\text{var}(X_s - X_t) \neq 0$  for every  $s \neq t$ , then almost every sample path achieves its supremum at at most one point.*

**Proof.** See Kim and Pollard (1990), Lemma 2.6. ■

## Problems and Complements

1. For every separable Gaussian process,  $\sigma^2(X) \leq M^2(X)/\Phi^{-1}(3/4)^2$ .

[Hint: See the proof of Proposition A.2.1.]

2. For every separable Gaussian process  $E\|X\|^p \leq K_p M(X)^p$  for a constant depending on  $p$  only.

[Hint: Integrate Borell's inequality to bound  $E\|X - M(X)\|^p$ . Next, use the preceding problem.]

3. For every separable Gaussian process,  $|E\|X\| - M(X)| \leq \sqrt{\pi/2} \sigma(X)$ .

4. Suppose that  $\mathcal{F}$  and  $\mathcal{G}$  are pre-Gaussian classes of measurable functions. Use majorizing measures to show that  $\mathcal{F} \cup \mathcal{G}$  is pre-Gaussian.

5. Every tight, Borel measurable centered, Gaussian variable  $X$  in  $\ell^\infty(T)$  satisfies  $P(\|X\| < \varepsilon) > 0$  for every  $\varepsilon > 0$ .

[Hint: By Anderson's lemma,  $P(\|X - x\| \leq \varepsilon) \leq P(\|X\| \leq \varepsilon)$  for every  $x$ . The support of  $X$  can be covered by countably many balls.]

6. Verify the claim made in Example A.2.10 that (A.2.9) holds with  $V = 4$ ,  $W = 4$ . Do the same for Example A.2.11, Example A.2.12, and Example A.2.13.

[Hint: First determine the size and shapes of the  $\varepsilon$ -balls with respect to the semimetric  $\rho$ , and note that  $V = 4$  is consistent with Example 2.6.1 and Theorem 2.6.4. Then determine the size and shape of the set(s)  $T_\delta$ .]

7. Use Lévy's inequality twice to show that for the Brownian sheet  $B$ ,

$$\begin{aligned} P\left(\sup_{0 \leq t_i \leq 1} B(t_1, t_2) \geq \lambda\right) &\leq 2P\left(\sup_{0 \leq t_2 \leq 1} B(1, t_2) \geq \lambda\right) \\ &\leq 4P\left(B(1, 1) \geq \lambda\right) = 4\bar{\Phi}(\lambda). \end{aligned}$$

(In fact, from known results about the suprema of Brownian motion (the reflection principle), the second inequality holds with equality.) Similarly, use Lévy's inequality to show that for the Kiefer-Müller process  $Z$

$$P\left(\sup_{0 \leq t_i \leq 1} Z(t_1, t_2) \geq \lambda\right) \leq 2P\left(\sup_{0 \leq t_2 \leq 1} Z(1, t_2) \geq \lambda\right) = 2\exp(-2\lambda^2),$$

since the process  $\{Z(1, t_2): 0 \leq t_2 \leq 1\}$  is a standard Brownian bridge process on  $[0, 1]$ , for which the distribution of the one-sided supremum is known from Doob (1949); see Shorack and Wellner (1986), page 34.

## A.3 Rademacher Processes

Let  $\varepsilon_1, \dots, \varepsilon_n$  be Rademacher variables and  $x_1, \dots, x_n$  arbitrary real functions on some “index” set  $T$ . The *Rademacher process*  $\{\sum_{i=1}^n \varepsilon_i x_i(t) : t \in T\}$  shares several properties with Gaussian processes. In this chapter we restrict ourselves to two propositions that have been used in Part 2. See, for instance, Ledoux and Talagrand (1991) for a further discussion.

Let  $\|x\|$  be the supremum norm  $\sup_{t \in T} |x(t)|$  for a given function  $x: T \rightarrow \mathbb{R}$ . Define

$$\sigma^2 = \sup_{t \in T} \text{var} \sum_{i=1}^n \varepsilon_i x_i(t) = \|\sum x_i^2\|.$$

The tail probabilities of the norm of a Rademacher process are comparable to those of Gaussian processes. The following inequalities are similar to the Borell inequalities, although it appears likely that the constants can be improved.

**A.3.1 Proposition.** *Let  $M$  be a median of the norm of the Rademacher process  $\sum_{i=1}^n \varepsilon_i x_i$ . Then there exists a universal constant  $C$  such that, for every  $\lambda > 0$ ,*

$$\begin{aligned} P\left(\left|\|\sum \varepsilon_i x_i\| - M\right| > \lambda\right) &\leq 4 \exp\left(-\frac{\lambda^2}{8\sigma^2}\right), \\ P\left(\left|\|\sum \varepsilon_i x_i\| - E\|\sum \varepsilon_i x_i\|\right| > \lambda\right) &\leq C \exp\left(-\frac{\lambda^2}{9\sigma^2}\right). \end{aligned}$$

**Proof.** For the first inequality, see Ledoux and Talagrand (1991), Theorem 4.7, on page 100.

To obtain the second inequality set  $\mu = \mathbb{E}\|\sum \varepsilon_i x_i\|$ . Integrate over the first inequality to obtain  $|\mu - M| \leq 4\sqrt{2\pi}\sigma$ . For  $c\lambda > |\mu - M|$ , the event in the second inequality is contained in the event  $\{\|\sum \varepsilon_i x_i\| - M > (1 - c)\lambda\}$ . Choose  $c > 0$  sufficiently small to bound the probability of this event by  $4\exp(-\lambda^2/9\sigma^2)$ . For  $c\lambda \leq |\mu - M|$ , the right side in the second inequality is never smaller than  $C\exp(-(\mu - M)^2/9c^2\sigma^2)$ , which is at least 1 for sufficiently large  $C$ . ■

**A.3.2 Proposition (Contraction principle).** *For each  $i$ , let  $x_i$  take its values in the interval  $[-1, 1]$ , and let  $\phi_i: [-1, 1] \rightarrow \mathbb{R}$  be Lipschitz of norm 1 with  $\phi_i(0) = 0$ . Then*

$$\mathbb{E}\left\|\sum \varepsilon_i \phi_i(x_i)\right\| \leq 2\mathbb{E}\left\|\sum \varepsilon_i x_i\right\|.$$

**Proof.** See Ledoux and Talagrand (1991), Theorem 4.12, on page 112. ■

The last proposition with the choices  $\phi_i(x) = \frac{1}{2}x^2$  is applied in Chapter 2.14.2 to obtain  $\mathbb{E}\left\|\sum \varepsilon_i x_i^2\right\| \leq 4\mathbb{E}\left\|\sum \varepsilon_i x_i\right\|$ .

# A.4

## Isoperimetric Inequalities for Product Measures

Let  $(\mathcal{X}^n, \mathcal{A}^n)$  be the product of  $n$  copies of a measurable space  $(\mathcal{X}, \mathcal{A})$ . Given points  $x = (x_1, \dots, x_n)$  and  $y = (y_1, \dots, y_n)$ , say that  $y$  controls a subset of coordinates  $\{x_i : i \in I\}$  of  $x$  if the corresponding coordinates of  $y$  are the same:  $y_i = x_i$  for every  $i \in I$ . Given vectors  $y^1, \dots, y^q$ , let  $f(y^1, \dots, y^q, x)$  be the number of coordinates of  $x$  that are not controlled by any  $y^j$ . Given subsets  $A_1, \dots, A_q$  of  $\mathcal{X}^n$ , define  $f(A_1, \dots, A_q, x)$  to be the minimum of the numbers  $f(y^1, \dots, y^q, x)$  when the  $y^j$  range over  $A_j$ . Thus  $n - f(A_1, \dots, A_q, x)$  is the maximal number of coordinates of  $x$  that can be controlled by the clever choice of  $y^j$  from  $A_j$ . Also, if  $f(A_1, \dots, A_q, x) = k$ , then there exist points  $y^j \in A_j$  such that the number of indices  $i \in \{1, \dots, n\}$  such that  $x_i \notin \{y_i^1, \dots, y_i^q\}$  is  $k$  (and a choice of  $y^j$  giving more control is impossible).

A basic inequality giving an upper bound on the size of the control functions  $f(A_1, \dots, A_q, x)$  is as follows.

**A.4.1 Proposition.** *If the coordinates of  $X = (X_1, \dots, X_n)$  are i.i.d. random elements in  $\mathcal{X}$ , then for any measurable sets  $A_1, \dots, A_q$  in  $\mathcal{A}^n$  and any integer  $q \geq 1$*

$$\mathbb{E}^* q^{f(A_1, \dots, A_q, X)} \leq \left( \prod_{l=1}^q \mathbb{P}(X \in A_l) \right)^{-1}.$$

Consequently, for every set  $A \in \mathcal{A}^n$  with  $\mathbb{P}(X \in A) \geq 1/2$

$$\mathbb{P}^*(f(A, \dots, A, X) \geq k) \leq \frac{2^q}{q^k} \leq \left(\frac{2}{q}\right)^k, \quad k \geq q.$$

Here  $q$  copies of  $A$  appear in the left side.

**Proof.** It will first be shown that, for any random variables  $U_1, \dots, U_q$  taking values in  $[0, 1]$ ,

$$\mathbb{E}\left(q \wedge \min_{1 \leq l \leq q} U_l^{-1}\right) \prod_{l=1}^q \mathbb{E}U_l \leq 1.$$

Here  $1/0$  is understood to be infinite. Indeed, the random variable  $U$  defined by  $U^{-1} = q \wedge \min_{1 \leq l \leq q} U_l^{-1}$  satisfies  $1/q \leq U \leq 1$ . Since  $U_l \leq U$  for every  $l$ , the left side is bounded above by  $\mathbb{E}U^{-1}(\mathbb{E}U)^q$ . Since  $\log x \leq x - 1$ , the logarithm of this expression is bounded by  $\mathbb{E}U^{-1} - 1 + q\mathbb{E}U - q$ . This is bounded above by zero, because on the interval  $[1/q, 1]$  the function  $z \rightarrow z^{-1} + qz$  attains a maximum value of  $1 + q$  (at the boundary points).

The proof of the theorem proceeds by induction on  $n$ . For  $n = 1$  the control number  $f(A_1, \dots, A_q, X)$  equals 0 if  $X$  is contained in some  $A_l$  and 1 otherwise. Thus

$$q^{f(A_1, \dots, A_q, X)} = q \wedge \min_{1 \leq l \leq q} 1_{A_l}(X)^{-1}.$$

The inequality in the first paragraph yields the result.

Suppose the theorem is true for  $n$ , and consider a vector  $X$  and subsets  $A_l$  in  $\mathcal{X}^{n+1}$ . Write  $Y$  and  $Z$  for the first  $n$  and the last coordinate of  $X = (Y, Z)$ , respectively. Furthermore, for a subset  $A$  of  $\mathcal{A}^{n+1}$ , define the sections and projection relative to the first  $n$  coordinates by

$$\begin{aligned} A(Z) &= \{Y \in \mathcal{X}^n : (Y, Z) \in A\}, \\ B &= \{Y \in \mathcal{X}^n : (Y, Z) \in A, \text{ for some } Z\}. \end{aligned}$$

If  $y^1, \dots, y^q$  control all except  $k$  coordinates of  $Y$ , then  $(y^1, z^1), \dots, (y^q, z^q)$  control all except at most  $k+1$  coordinates of  $(Y, Z)$ ; they control all except  $k$  coordinates of  $(Y, Z)$  if one  $z^j$  is chosen equal to  $Z$ . Therefore, it follows that, for every  $l \leq q$ ,

$$\begin{aligned} f(A_1, \dots, A_q, X) &\leq 1 + f(B_1, \dots, B_q, Y) \\ f(A_1, \dots, A_q, X) &\leq f(B_1, \dots, B_{l-1}, A_l(Z), B_{l+1}, \dots, B_q, Y). \end{aligned}$$

Write  $\mathbb{E}_Y$  for the expectation taken with respect to the first  $n$  coordinates only and let  $P$  be the distribution of  $Y$ . The first inequality of the display yields

$$\mathbb{E}_Y q^{f(A_1, \dots, A_q, X)} \leq q \mathbb{E}_Y q^{f(B_1, \dots, B_q, Y)} \leq q \left( \prod_{l=1}^q P(B_l) \right)^{-1},$$

by the induction hypothesis. Similarly, the second inequality together with the induction hypothesis implies that, for every  $1 \leq l \leq q$ ,

$$\mathbb{E}_Y q^{f(A_1, \dots, A_q, X)} \leq \left( \prod_{j \neq l} P(B_j) P(A_l(Z)) \right)^{-1}.$$

Combine the last two displays to see that their common left side is bounded above by  $q \wedge \min_{1 \leq l \leq q} Z_l^{-1}$  times  $(\prod_{l=1}^q P(B_l))^{-1}$  for the random variables  $Z_l = P(A_l(Z))/P(B_l)$ . An application of the inequality in the first paragraph completes the proof.

Actually the preceding proof ignores issues of measurability and is false in general, because maps of the type  $X \mapsto f(A_1, \dots, A_q, X)$  need not be measurable, as required for the application of Fubini's theorem.

With some effort it can be seen that this problem does not arise in the case that the sample space is Polish with Borel  $\sigma$ -field and compact sets  $A_1, \dots, A_q$ . This follows from the identity

$$\{x: n - f(A_1, \dots, A_q, x) \leq k\} = \bigcap_{|\cup I_i| > k} \bigcup_{i=1}^q \{x: \pi_{I_i} x \notin \pi_{I_i} A_i\},$$

where the intersection is taken over all collections of  $q$  subsets  $I_1, \dots, I_q$  of  $\{1, 2, \dots, n\}$  and  $\pi_I$  is the projection  $\pi_I: \mathcal{X}^n \mapsto \mathcal{X}^I$  on the  $I$ th coordinates. (Note that  $\pi_I x \in \pi_I A$  if and only if there exists  $y \in A$  that controls  $\{x_i: i \in I\}$ .) Projections of compact sets are compact, hence measurable. Thus, in this special situation the proof is correct.

Next, the case of general Borel sets in a Polish (product) space can be obtained by approximation from within by compact sets. By replacing every set  $A_i$  by a compact  $K_i \subset A_i$ , the control numbers increase (there is less control), whence the left side of the theorem increases. Next, apply the theorem and note that the resulting upper bound decreases to the right side of the theorem if the probabilities  $P(A_i - K_i)$  are made to decrease to zero. The latter is possible by the regularity of Borel measures on Polish spaces.

The case of Polish spaces is sufficient for most applications but can be extended to arbitrary sample spaces. We sketch the extension.

Since any event in  $\mathcal{A}^n$  is contained in the completion of the product of sub- $\sigma$ -fields of  $\mathcal{A}$  that are countably generated, it is no loss of generality to assume that  $\mathcal{A}$  is countably generated, say with generator the sets  $A_1, A_2, \dots$ . Then the map  $\phi: \mathcal{X} \mapsto \{0, 1\}^\infty$  given by

$$\phi(x) = (1_{A_1}(x), 1_{A_2}(x), \dots)$$

is Borel measurable into the Polish space  $\{0, 1\}^\infty \equiv R$  and  $\mathcal{A} = \phi^{-1}(\mathcal{B})$  for the Borel sets  $\mathcal{B}$ . Let  $\phi_n: \mathcal{X}^n \mapsto R^n$  be the map  $(x_1, \dots, x_n) \mapsto (\phi(x_1), \dots, \phi(x_n))$ , and write  $P$  for the underlying measure on  $\mathcal{X}$ .

By construction there exists for every  $A \in \mathcal{A}^n$  a Borel set  $B \in \mathcal{B}^n$  with  $A = \phi_n^{-1}(B)$ . Suppose that there exists a Borel set  $B' \subset B$  of the same measure under  $(P \circ \phi^{-1})^n$  with the property that for every  $z' \in B'$ , there exists  $z \in B \cap \phi(\mathcal{X})^n$  such that  $z_i = z'_i$  for all  $i$  such that  $z'_i \in \phi(\mathcal{X})$ . Thus, the coordinates of  $z'$  that are not in  $\phi(\mathcal{X})$  can be changed, meanwhile leaving the other coordinates the same, so as to obtain a point in  $B$  with

all coordinates in  $\phi(\mathcal{X})$ . A vector in  $B$  with all coordinates in  $\phi(\mathcal{X})$  is the image of some element in  $A$  and it can be seen that

$$\left\{ x : f(B'_1, \dots, B'_q, \phi_n(x)) \leq k \right\} \subset \left\{ x : f(A_1, \dots, A_q, x) \leq k \right\}.$$

Consequently,

$$(P^n)^*(f(A_1, \dots, A_q, x) > k) \leq ((P \circ \phi^{-1})^n)^*(z : f(B'_1, \dots, B'_q, z) > k).$$

Since  $(P \circ \phi^{-1})^n(B') = P^n(A)$ , the problem has been reduced to the case of a Polish sample space.

The existence of the sets  $B'$  can be argued as follows. For every partition  $\{1, \dots, n\} = I \cup J$ , define

$$B_{I,J} = \left\{ x \in R^n : (P \circ \phi^{-1})^J(B(\pi_I x)) = 0 \right\}.$$

Here  $B(\pi_I x)$  is the section of  $B$  at  $\pi_I x$ , which is contained in  $R^J$ . By Fubini's theorem, each set  $B_{I,J}$  is a  $(P \circ \phi^{-1})^n$  null set. Thus,  $B' = B - \cup_{I,J} B_{I,J}$  has the same measure as  $B$ . If  $z' \in B'$ , then

$$(P \circ \phi^{-1})^J(B(\pi_I z')) > 0; \quad ((P \circ \phi^{-1})^J)^*(\phi(\mathcal{X})^J) = 1.$$

This means that the set  $B(\pi_I z') \cap \phi(\mathcal{X})^J$  cannot be empty. If the coordinates  $\{z'_i : i \in J\}$  are not contained in  $\phi(\mathcal{X})$ , then they can be changed in the desired manner. ■

A typical application of the control numbers  $f(A_1, \dots, A_q, X)$  is as follows. For natural numbers  $n$ , let  $S_n : \mathcal{X}^n \rightarrow \mathbb{R}$  be permutation-symmetric functions that satisfy

$$(A.4.2) \quad \begin{aligned} S_n(x) &\leq S_{n+m}(x, y), && \text{for every } x \in \mathcal{X}^n, y \in \mathcal{X}^m, \\ S_{n+m}(x, y) &\leq S_n(x) + S_m(y), && \text{for every } x \in \mathcal{X}^n, y \in \mathcal{X}^m. \end{aligned}$$

An example is the functions  $S_n = E_\varepsilon \|\sum_{i=1}^n \varepsilon_i f(x_i)\|_{\mathcal{F}}$  for Rademacher random variables  $\varepsilon_i$ . Given sets  $A_1, \dots, A_q$  in  $\mathcal{X}^n$ , the elements of a random sample  $X_1, \dots, X_n$  in  $\mathcal{X}$  can be split into  $q+1$  sets, the first one consisting of the  $k = f(A_1, \dots, A_q, X)$  variables not controlled by the sets  $A_l$  and the other  $q$  sets consisting of variables that also occur in some  $Y^l \in A_l$  for  $l = 1, \dots, q$ . By the “triangle inequality” in (A.4.2), the number  $S_n(X_1, \dots, X_n)$  can be bounded by the sum of  $q+1$  terms. By the first inequality in (A.4.2), the last  $q$  terms can be bounded by  $S_n(Y^l)$  for  $l = 1, \dots, q$ . It follows that on the set  $f(A_1, \dots, A_q, X) = k$ ,

$$S_n(X) \leq \sup_{x \in \mathcal{X}^k} S_k(x) + \sum_{l=1}^q \sup_{y \in A_l} S_n(y).$$

The first term on the right should be small if  $k$  is small, while the size of the second term can be influenced by the choice of the control sets  $A_l$ .<sup>#</sup>

<sup>#</sup> The full force of the definition of  $f(A_1, \dots, A_q, X)$  is not used for this application. The control numbers could be reduced to the minimal  $k$  such that there exist  $y^l \in A_l$  with at most  $k$  of the  $X_i$  not occurring in the set  $\{y_i^l : l = 1, \dots, q, i = 1, \dots, n\}$ .

**A.4.3 Lemma.** Let  $S_n: \mathcal{X}^n \rightarrow [0, n]$  be permutation-symmetric functions satisfying (A.4.2). Then, for every  $t > 0$  and  $n$ ,

$$P^*(S_n \geq t) \leq \exp\left(-\frac{1}{2}t \log \frac{t}{12(E^*S_n \vee 1)}\right).$$

This is true for any  $S_n = S_n(X)$  and any vector  $X = (X_1, \dots, X_n)$  with i.i.d. coordinates.

**Proof.** Set  $\mu = E^*S_n$ . By Markov's inequality, the set  $A = \{S_n \leq 2\mu\}_*$  has probability at least  $\frac{1}{2}$ . Since  $S_n \leq n$  by assumption, the argument preceding the lemma shows that  $S_n(X) \leq f(A, \dots, A, X) + q2\mu$  for any integer  $q \geq 1$ . Thus, the probability in the lemma is bounded by

$$P^*\left(f(A, \dots, A, X) \geq t - 2q\mu\right) \leq \frac{2^q}{q^{t-2q\mu}} \leq e^{q \log q(2\mu + 1) - t \log q},$$

for  $q \geq 2$  by Theorem A.4.1. For  $t \geq 12(\mu \vee 1)$  and  $q = \lfloor t/6(\mu \vee 1) \rfloor$ , the exponent is bounded above by  $-\frac{1}{2}t \log \lfloor t/6(\mu \vee 1) \rfloor$ . The lemma follows since  $\lfloor x \rfloor \geq \frac{1}{2}x$  for  $x \geq 1$ . For  $t \leq 12(\mu \vee 1)$  the lemma is trivially true, because the right side is larger than 1. ■

## A.5

# Some Limit Theorems

In this chapter we present the strong law of large numbers and the central limit theorem uniformly in the underlying measure, as well as the rank central limit theorem.

Let  $\bar{X}_n$  be the average of the first  $n$  variables from a sequence  $X_1, X_2, \dots$  of independent and identically distributed random vectors in  $\mathbb{R}^d$ . Let  $\mathcal{P}$  be a class of underlying probability measures. For instance, let  $X_i$  be the  $i$ th coordinate projection of  $(\mathbb{R}^d)^\infty$ , and let  $\mathcal{P}$  consist of Borel probability measures on  $\mathbb{R}^d$ .

**A.5.1 Proposition.** *Let  $X_1, X_2, \dots$  be i.i.d. random vectors with distribution  $P \in \mathcal{P}$  such that*

$$\lim_{M \rightarrow \infty} \sup_{P \in \mathcal{P}} \mathbb{E}_P |X_1| \{ |X_1| \geq M \} = 0.$$

*Then the strong law of large numbers holds uniformly in  $P \in \mathcal{P}$  in the sense that, for every  $\varepsilon > 0$ ,*

$$\lim_{n \rightarrow \infty} \sup_{P \in \mathcal{P}} P_P \left( \sup_{m \geq n} |\bar{X}_m - \mathbb{E}_P X_1| \geq \varepsilon \right) = 0.$$

**Proof.** See Chung (1951). ■

The *Prohorov distance*  $\pi$  between probability laws on  $\mathbb{R}^d$  dominates twice the bounded Lipschitz distance and generates the weak topology. The following theorem gives an explicit upper bound on the Prohorov distance between the distribution of  $\sqrt{n}\bar{X}_n$  and the limiting normal distribution.

This can easily be converted into a central limit theorem uniformly in  $P$  ranging over  $\mathcal{P}$ . Write  $\pi(X, Y)$  for the Prohorov distance between the distributions of random vectors  $X$  and  $Y$ .

**A.5.2 Proposition.** *Let  $X_1, \dots, X_n$  be i.i.d. random vectors with mean zero and finite covariance matrix  $\Sigma$ . Then for every  $\varepsilon > 0$ ,*

$$\begin{aligned}\pi(\sqrt{n}\bar{X}_n, N(0, \Sigma)) &\leq 2\varepsilon^{-2}g(\varepsilon\sqrt{n}) \vee \varepsilon + 4^{1/3}g(\varepsilon\sqrt{n})^{1/3} \\ &\quad + C(d\varepsilon g(0))^{1/4} \left(1 + \left|\log \frac{\varepsilon g(0)}{d}\right|^{1/2}\right).\end{aligned}$$

Here  $C$  is an absolute constant and  $g(\varepsilon) = E\|X\|^2\{\|X\| \geq \varepsilon\}$ .

**Proof.** For a given  $\varepsilon \in (0, 1)$  and  $1 \leq i \leq n$ , let  $Y_i$  be the truncated variable  $X_i 1\{|X_i| \leq \varepsilon\sqrt{n}\}$  and  $Z_i$  the centered variable  $Y_i - EY_i$ . By Strassen's theorem, it follows that

$$\pi(\sqrt{n}\bar{X}_n, \sqrt{n}\bar{Y}_n) \vee \pi(\sqrt{n}\bar{Y}_n, \sqrt{n}\bar{Z}_n) \leq (\varepsilon^{-2} \vee \varepsilon^{-1})E\|X\|^2\{\|X\| \geq \varepsilon\sqrt{n}\} \vee \varepsilon.$$

Furthermore,  $E\|Z_i\|^3 \leq 8E\|Y_i\|^3 \leq 8\varepsilon\sqrt{n}E\|X_i\|^2$ . Thus, by Theorem 1 of Yurinskii (1977), as corrected in Theorem B on page 395 of Dehling (1983),

$$\pi(\sqrt{n}\bar{Z}_n, N(0, \Sigma_Z)) \lesssim (d\varepsilon g(0))^{1/4} \left(1 + \left|\log \frac{8\varepsilon g(0)}{d}\right|^{1/2}\right),$$

where  $\Sigma_Z$  is the covariance matrix of  $Z_1$ . Finally,

$$\pi(N(0, \Sigma_Z), N(0, \Sigma))^3 \leq 4E\|X\|^2\{\|X\| \geq \varepsilon\sqrt{n}\},$$

by Lemma 2.1, page 402, of Dehling (1983). ■

For each  $n$ , let  $a_{n1}, \dots, a_{nn}$  and  $b_{n1}, \dots, b_{nn}$  be real numbers, and let  $(R_{n1}, \dots, R_{nn})$  be a random vector that is uniformly distributed on the  $n!$  permutations of  $\{1, \dots, n\}$ . Consider the rank statistic

$$S_n = \sum_{i=1}^n b_{ni} a_{n, R_{ni}}.$$

The mean and variance of  $S_n$  are equal to  $ES_n = n\bar{a}_n\bar{b}_n$  and  $\text{var } S_n = A_n^2 B_n^2 / (n-1)$ , where  $A_n^2$  and  $B_n^2$  are the sums of squared deviations from the mean of the numbers  $a_{n1}, \dots, a_{nn}$  and  $b_{n1}, \dots, b_{nn}$ , respectively. Thus  $A_n^2 = \sum_{i=1}^n (a_{ni} - \bar{a}_n)^2$ .

**A.5.3 Proposition (Rank central limit theorem).** Suppose that

$$\max_{1 \leq i \leq n} \frac{|a_{ni} - \bar{a}_n|}{A_n} \rightarrow 0; \quad \max_{1 \leq i \leq n} \frac{|b_{ni} - \bar{b}_n|}{B_n} \rightarrow 0.$$

Then the sequence  $(S_n - \mathbb{E}S_n)/\sigma(S_n)$  converges in distribution to a standard normal distribution if and only if

$$\sum_{(i,j): \sqrt{n}|a_{ni} - \bar{a}_n| | b_{nj} - \bar{b}_n| > \varepsilon A_n B_n} \frac{|a_{ni} - \bar{a}_n|^2 |b_{nj} - \bar{b}_n|^2}{A_n^2 B_n^2} \rightarrow 0, \quad \text{for every } \varepsilon > 0.$$

**Proof.** See Hájek (1961). ■

# A.6

## More Inequalities

This chapter collects well-known and less-known inequalities for binomial, multinomial, and Rademacher random variables. Some of the results are related to the inequalities obtained by other means in Chapters 2.2 and 2.14.

### A.6.1 Binomial Random Variables

Throughout this subsection  $\bar{Y}_n$  is the average of independent Bernoulli variables  $Y_1, \dots, Y_n$  with success probability  $p$ .

One of the simplest inequalities for the tails of sums of bounded random variables—in particular, for the binomial distribution—is that of Hoeffding (1963).

**A.6.1 Proposition (Hoeffding's inequality).** *Let  $X_1, \dots, X_n$  be independent random variables taking values in  $[0, 1]$ . Let  $\mu = E\bar{X}_n$ . Then for  $0 < t < 1 - \mu$ ,*

$$\begin{aligned} P(\bar{X}_n - \mu \geq t) &\leq \left( \left( \frac{\mu}{\mu + t} \right)^{\mu+t} \left( \frac{1-\mu}{1-\mu-t} \right)^{1-\mu-t} \right)^n \\ &\leq \exp -nt^2 g(\mu) \\ &\leq \exp -2nt^2. \end{aligned}$$

Here  $g(\mu)$  is equal to  $(1 - 2\mu)^{-1} \log(1/\mu - 1)$ , for  $0 < \mu < 1/2$ , and  $(2\mu(1 - \mu))^{-1}$ , for  $1/2 \leq \mu < 1$ .

**Proof.** See Hoeffding (1963). ■

Applied to Bernoulli variables, Hoeffding's inequality yields

$$P(\sqrt{n}|\bar{Y}_n - p| \geq \lambda) \leq 2 \exp -2\lambda^2, \quad \lambda > 0.$$

This inequality is valid for any  $n$  and  $p$ . For  $p$  close to zero or one, it is possible to improve on the factor 2 in the exponent considerably.

**A.6.2 Proposition (Bennett's inequality).** *For all  $n$  and  $0 < p < 1$ ,*

$$P(\sqrt{n}|\bar{Y}_n - p| \geq \lambda) \leq 2 \exp -\frac{\lambda^2}{2p} \psi\left(\frac{\lambda}{\sqrt{np}}\right), \quad \lambda > 0.$$

Here  $\psi(x) = 2h(1+x)/x^2$ , for  $h(x) = x(\log x - 1) + 1$ .

**Proof.** See Bennett (1962), Hoeffding (1963), or Shorack and Wellner (1986), page 440. ■

**A.6.3 Corollary (Kiefer's inequality).** *For all  $n$  and  $0 < p < e^{-1}$ ,*

$$P(\sqrt{n}|\bar{Y}_n - p| \geq \lambda) \leq 2 \exp -\lambda^2(\log(1/p) - 1).$$

**Proof.** Since the probability on the left side is zero if  $\lambda > \sqrt{n}$ , we may assume that  $\lambda \leq \sqrt{n}$ . Then  $\psi(\lambda/\sqrt{np}) \geq \psi(1/p)$ , because  $\psi$  is decreasing. Now apply Bennett's inequality and finish the proof by noting that  $\psi(1/p)/(2p) \geq \log(1/p) - 1$  for  $0 < p < e^{-1}$ . ■

For  $p \downarrow 0$ , the function  $\log(1/p) - 1$  increases to infinity. Thus Kiefer's inequality gives bounds of the type  $2 \exp -C\lambda^2$  for large constants  $C$  for sufficiently small  $p$ . For example,  $2 \exp -11\lambda^2$  for  $p \leq e^{-12}$ .

For  $p$  bounded away from zero and one, Hoeffding's inequality can be improved as well—at least for large values of  $\lambda$ .

**A.6.4 Proposition (Talagrand's inequality).** *There exist constants  $K_1$  and  $K_2$  depending only on  $p_0$ , such that for  $p_0 \leq p \leq 1 - p_0$  and all  $n$ ,*

$$\begin{aligned} P(\sqrt{n}(\bar{Y}_n - p) = \lambda) &\leq \frac{K_1}{\sqrt{n}} \exp -\left(2\lambda^2 + \frac{\lambda^4}{4n}\right), \quad \lambda > 0, \\ P(\sqrt{n}(\bar{Y}_n - p) \geq t) &\leq \frac{K_2}{\lambda} \exp -\left(2\lambda^2 + \frac{\lambda^4}{4n}\right) \exp 5\lambda(\lambda - t), \quad 0 < t < \lambda, \\ P(\sqrt{n}(\bar{Y}_n - p) \geq \lambda) &\leq \frac{K_2}{\lambda} \exp -\left(2\lambda^2 + \frac{\lambda^4}{4n}\right), \quad \lambda > 0. \end{aligned}$$

**Proof.** Stirling's formula with bounds asserts that

$$\sqrt{2\pi n} \left(\frac{n}{e}\right)^n e^{1/(12n+1)} \leq n! \leq \sqrt{2\pi n} \left(\frac{n}{e}\right)^n e^{1/(12n)}.$$

Hence it follows, that for  $1 \leq k \leq n - 1$ ,

$$(A.6.5) \quad \binom{n}{k} p^k (1-p)^{n-k} \leq \frac{1}{\sqrt{2\pi}} \left( \frac{n}{k(n-k)} \right)^{1/2} e^{-n\Psi(u,p)}.$$

Here set  $u = k/n - p$  and define

$$\Psi(u,p) = (u+p) \log\left(\frac{u+p}{p}\right) + (1-(u+p)) \log\left(\frac{1-(u+p)}{1-p}\right).$$

The function  $\Psi$  satisfies  $\Psi(0,p) = 0$ ,  $\partial/\partial u \Psi(0,p) = 0$ , and

$$\frac{\partial^2}{\partial u^2} \Psi(u,p) = \frac{4}{1 - 4(u - (\frac{1}{2} - p))^2} \geq 4 \left( 1 + 4(u - (\frac{1}{2} - p))^2 \right).$$

Integrate this inequality to obtain

$$\frac{\partial}{\partial u} \Psi(u,p) \geq 4u + \frac{16}{3} \left( (u - (\frac{1}{2} - p))^3 + (\frac{1}{2} - p)^3 \right) \geq 4u + \frac{4}{3}u^3.$$

In the last step, use the fact that the function  $\beta \mapsto (u - \beta)^3 + \beta^3$  takes its minimal value at  $\beta = u/2$ . Integrate again to obtain the inequality  $\Psi(u,p) \geq 2u^2 + u^4/3$ .

For the proof of the first part of the proposition, put  $k = \sqrt{n}\lambda + np$ . Then  $k \geq np_0$ . Furthermore, for  $k \leq n(1 - p_0/2)$ , we have  $n - k \geq np_0/2$ , and the right-hand side of (A.6.5) is bounded above by

$$\frac{1}{\sqrt{2\pi}} \frac{1}{\sqrt{np_0^2/2}} e^{-n(2u^2+u^4/3)}.$$

For  $n(1 - p_0/2) \leq k < n$ , we have  $u = k/n - p \geq p_0/2$ , and the right-hand side of (A.6.5) is bounded by

$$\frac{1}{\sqrt{2\pi(1-p_0/2)}} e^{-n(2u^2+u^4/3)} \leq \frac{K}{\sqrt{n}} e^{-n(2u^2+u^4/4)} \sup_{n \geq 1} \sqrt{n} e^{-n(p_0/2)^4/12}.$$

Finally, for  $k = n$ , the inequality of the proposition can be reduced to

$$-\log(1/p) + 2(1-p)^2 + \frac{1}{4}(1-p)^4 \leq \frac{1}{n} \log\left(\frac{K_1}{\sqrt{n}}\right).$$

The function  $h(p)$  on the left side of this last inequality is negative for  $0 < p < 1$ , with  $h(0) = -\infty$  and  $h(1) = 0$ , and has a local maximum of  $-0.160144\dots$  at  $0.344214\dots$  and a local minimum of  $-0.184428\dots$  at  $0.598428\dots$ . It follows that the supremum of the left side over any interval  $p_0 \leq p \leq 1 - p_0$  is strictly less than zero. On the other hand, the infimum of the right side over  $n$  is bounded below by  $-1/(2eK_1^2)$ . Thus, the inequality is valid for sufficiently large  $K_1$ .

For the proof of the second part of the proposition, consider the function  $h(x) = 2x^2 + x^4/4$ . The derivative  $h'(x) = 4x + x^3$  is bounded below by  $4x$  and is bounded above by  $5x$  for  $x \leq 1$ . Since  $h$  is convex,

$h(x) \geq h(u) + (x - u)h'(u)$  for all  $x$ . Apply the first inequality of the proposition to find that, with  $k_0 = \lceil np + \sqrt{nt} \rceil$ ,

$$\begin{aligned} P(\sqrt{n}(\bar{Y}_n - p) \geq t) &\leq \sum_{k \geq k_0} \frac{K_1}{\sqrt{n}} e^{-n(h(u) + (k/n - p - u)h'(u))} \\ &= \frac{K_1}{\sqrt{n}} e^{-nh(u)} \sum_{k \geq k_0} e^{(nu - (k - np))h'(u)} \\ &\leq \frac{K_1}{\sqrt{n}} e^{-nh(u)} \frac{e^{(nu - (k_0 - np))h'(u)}}{1 - e^{-h'(u)}} \\ &\leq \frac{K_1}{\sqrt{n}} e^{-nh(u)} \frac{1}{4u} e^{5nu(u-t)}. \end{aligned}$$

This concludes the proof of the proposition. ■

### A.6.2 Multinomial Random Vectors

In this section  $(N_1, \dots, N_k)$  is a multinomially distributed vector with parameters  $n$  and  $(p_1, \dots, p_k)$ . It is helpful to relate this to empirical processes. Let  $\{C_i\}_{i=1}^k$  be a partition of a set  $\mathcal{X}$  and  $P$  the probability measure on the  $\sigma$ -field  $\mathcal{C}$  generated by the partition such that  $P(C_i) = p_i$ . If  $\mathbb{P}_n$  is the empirical measure corresponding to a sample of size  $n$  from  $P$ , then the vector  $(n\mathbb{P}_n(C_1), \dots, n\mathbb{P}_n(C_k))$  has the same distribution as  $(N_1, \dots, N_k)$ .

The first two propositions concern the “ $L_1$ ” or “total variation distance”  $\sum_{i=1}^k |N_i - np_i|$ . In the representation using the empirical process, this is equivalent to the Kolmogorov-Smirnov distance

$$\|\mathbb{P}_n - P\|_{\mathcal{C}} = \sup_{C \in \mathcal{C}} (\mathbb{P}_n - P)(C) = \frac{1}{2} \sum_{i=1}^k |\mathbb{P}_n(C_i) - P(C_i)|.$$

**A.6.6 Proposition (Bretagnolle–Huber–Carol inequality).** *If the random vector  $(N_1, \dots, N_k)$  is multinomially distributed with parameters  $n$  and  $(p_1, \dots, p_k)$ , then*

$$P\left(\sum_{i=1}^k |N_i - np_i| \geq 2\sqrt{n}\lambda\right) \leq 2^k \exp -2\lambda^2, \quad \lambda > 0.$$

**Proof.** In terms of the Kolmogorov-Smirnov statistic, the left side is equal to  $P(\|\mathbb{G}_n\|_{\mathcal{C}} \geq \lambda)$ . Each of the probabilities  $P((\mathbb{P}_n - P)(C) > \lambda)$  can be bounded by  $\exp(-2\lambda^2)$  by a one-sided version of Hoeffding’s inequality. There are  $2^k$  sets in  $\mathcal{C}$ . Alternatively, first reduce  $\mathcal{C}$  by dropping all sets with probability more than  $1/2$ . This does not change  $\|\mathbb{P}_n - P\|_{\mathcal{C}}$ . Now the two-sided version of Hoeffding’s inequality can be applied to the  $2^{k-1}$  remaining sets. ■

The following inequality results from combining Kiefer’s inequality, Corollary A.6.3, and Talagrand’s inequality, Proposition A.6.4.

**A.6.7 Proposition.** *If the random vector  $(N_1, \dots, N_k)$  is multinomially distributed with parameters  $n$  and  $(p_1, \dots, p_k)$ , then there is a universal constant  $K$  such that*

$$P\left(\sum_{i=1}^k |N_i - np_i| \geq 2\sqrt{n}\lambda\right) \leq \frac{K}{\lambda} 2^k \exp -2\lambda^2, \quad \lambda > 0.$$

**Proof.** The proof is slightly easier in the language of empirical processes as introduced above. Let  $\mathcal{C}_0$  be the class of sets  $C \in \mathcal{C}$  such that  $P(C) < e^{-12}$ , and let  $\mathcal{C}_1$  be the class of sets with  $e^{-12} \leq P(C) \leq 1/2$ . Then  $\|\mathbb{G}_n\|_C = \|\mathbb{G}_n\|_{\mathcal{C}_0} \vee \|\mathbb{G}_n\|_{\mathcal{C}_1}$ . By the remark following Kiefer's inequality, Corollary A.6.3, applied to  $\|\mathbb{G}_n\|_{\mathcal{C}_0}$  and Talagrand's inequality, Proposition A.6.4, applied to  $\|\mathbb{G}_n\|_{\mathcal{C}_1}$ , there exists a universal constant  $K_2$  such that

$$P(\|\mathbb{G}_n\|_C \geq \lambda) \leq 2^k 2 \exp -11\lambda^2 + 2^k \frac{K_2}{\lambda} \exp -2\lambda^2.$$

The proposition follows easily. ■

The last inequality for multinomial random vectors concerns a slight modification of the classical chi-square statistic (or weighted  $L_2$ -distance).

**A.6.8 Proposition (Mason and van Zwet inequality).** *Let the random vector  $(N_1, \dots, N_k)$  be multinomially distributed with parameters  $n$  and  $(p_1, \dots, p_k)$  such that  $p_i > 0$  for  $i < k$ . Then for every  $C > 0$  and  $\delta > 0$  there exist constants  $a, b, c > 0$ , such that for all  $n \geq 1$  and  $\lambda, p_1, \dots, p_{k-1}$  satisfying  $\lambda \leq Cn \min\{p_i : 1 \leq i \leq k-1\}$  and  $\sum_{i=1}^{k-1} p_i \leq 1 - \delta$ , we have*

$$P\left(\sum_{i=1}^{k-1} \frac{(N_i - np_i)^2}{np_i} > \lambda\right) \leq a \exp(bk - c\lambda).$$

We do not give a full derivation of this inequality here, but note that the chi-square statistic is an example of a supremum over an elliptical class of functions. Therefore, the preceding inequality can be deduced from Talagrand's general empirical process inequality, Theorem 2.14.24.

### A.6.3 Rademacher Sums

The following inequality improves Hoeffding's inequality for Rademacher sums, Lemma 2.2.7, in much the same way that Talagrand's inequality improves on Hoeffding's inequality. Let  $\varepsilon_1, \dots, \varepsilon_n$  denote independent Rademacher variables.

**A.6.9 Proposition (Pinelis' inequality).** *For any numbers  $a_1, \dots, a_n$  with  $\sum_{i=1}^n a_i^2 = 1$ ,*

$$P\left(\left|\sum_{i=1}^n a_i \varepsilon_i\right| \geq \lambda\right) \leq \frac{4e^3}{9} (1 - \Phi(\lambda)) \leq \frac{4e^3}{9\lambda} \phi(\lambda), \quad \lambda > 0.$$

**A.6.10 Proposition (Khinchine's inequality).** For any real numbers  $a_1, \dots, a_n$  and any  $p \geq 1$ , there exist constants  $A_p$  and  $B_p$  such that

$$A_p \|a\|_2 \leq \left\| \sum_{i=1}^n a_i \varepsilon_i \right\|_p \leq B_p \|a\|_2.$$

## Problems and Complements

1. The function  $g(p) = \psi(1/p)/(2p)$  is bounded below by  $\log(1/p) - 1$  for  $0 < p < e^{-1}$ . It is not bounded below by  $\log(1/p) - c$  for any  $c < 1$ .

[Hint:  $\psi(1/p)/(2p) - (\log(1/p) - 1) \rightarrow 0$  as  $p \rightarrow 0$ .]

2. (Bennett's inequality plus Jensen equals Bernstein's inequality) Show that the function  $\psi$  in Proposition A.6.2 is decreasing,  $\psi(0) = 1$ ,  $x\psi(x)$  is increasing, and  $\psi(x) \geq 1/(1+x/3)$ .

[Hint: Let  $f(x) = h(1+x) = (x+1)\log(x+1) - x$ , and write

$$f(x) = \int_0^1 xf'(xt)dt = \int_0^1 x^2 \int_0^1 1\{0 \leq s \leq t \leq 1\} f''(xs) ds dt,$$

where  $f'(x) = \log(1+x)$  and  $f''(x) = 1/(1+x)$ . Hence

$$\psi(x) = 2 \int_0^1 \int_0^1 1\{0 \leq s \leq t \leq 1\} f''(xs) ds dt.$$

To prove the last inequality, interpret the right side of this identity as an expectation.]

3. The probability in Kiefer's inequality is bounded below by  $\exp[-\lambda^2(1-p)^{-2}\log(1/p)]$  when  $\lambda = \sqrt{n}(1-p)$ .
4. The constant  $K$  resulting from the proof of Proposition A.6.7 may be minimized by choosing an optimal cut-off point instead of  $e^{-12}$ . This involves understanding the dependence of the constant  $K_2$  in Proposition A.6.4 on  $p_0$ .
5. Independent binomial  $(1, p_i)$  variables  $X_1, \dots, X_n$  satisfy

$$P\left(\sum_{i=1}^n X_i > k\right) \leq \exp\left(-n\bar{p}_n h\left(\frac{k}{n\bar{p}_n}\right)\right) < \left(\frac{e n \bar{p}_n}{k}\right)^k,$$

for the function  $h(x) = x(\log x - 1) + 1$  and  $\bar{p}_n$  the average of the success probabilities.

# A

## Notes

**A.1.** Ottaviani (1939) proves his inequality for the case of real-valued random variables. Hoffmann-Jørgensen (1974) proves his inequalities for Banach-valued random variables. For both cases Dudley (1984) gives careful proofs for random elements in nonseparable Banach spaces.

Proposition A.1.6 is due to Talagrand (1989).

**A.2.** The first inequality of Proposition A.2.1 is a consequence of finer isoperimetric results obtained independently by Borell (1975) and Sudakov and Tsirel'son (1978). The second inequality was found by Ibragimov, Sudakov, and Tsirel'son (1976) and rediscovered by Maurey and Pisier (see Pisier (1986)). The present proof is based on the exposition by Ledoux (1995). See Ledoux and Talagrand (1991), page 59, for the third inequality of Proposition A.2.1 as well as other interesting results.

The moment inequality Proposition A.2.4 is given by Pisier (1986).

Adler (1990) gives a good summary of many of the Gaussian comparison results, including Sudakov's minorization inequality, Dudley's majorization inequality, and the more recent work on majorizing measures. The inequalities of Proposition A.2.6 are due to Fernique after earlier work by Slepian (1962), Marcus and Shepp (1971) and Landau and Shepp (1970). The last assertion of Proposition A.2.6 was noted by Marcus and Shepp and is stated explicitly in Jain and Marcus (1978).

The section on exponential bounds for Gaussian processes is based mostly on Talagrand (1994). The interested reader should also consult Samorodnitsky (1991), Adler and Samorodnitsky (1987), Adler (1990), and

Goodman (1976). The renewal theoretic large deviation methods of Hogan and Siegmund (1986), Siegmund (1988, 1992) give precise results on the asymptotic tail behavior for the suprema of particular Gaussian fields, especially those satisfying what Aldous (1989) calls the “uncorrelated orthogonal increments” property. See Aldous (1989), Chapter J, pages 190–219 for a variety of fascinating asymptotic formulas. Computation of the asymptotic behavior in Example A.2.12 was carried out by David Siegmund (personal communication).

The section on majorizing measures records work by Preston (1972), Fernique (1974), and Talagrand (1987c). The sufficiency of the existence of a (continuous) majorizing measure for the boundedness (and continuity) of a process was established by Fernique (1974) following preliminary results of Preston (1972). The necessity was proved for stationary Gaussian processes by Fernique (1974) and for general Gaussian processes by Talagrand (1987c). A relatively simple proof is given by Talagrand (1992). Lemma A.2.19 is adapted from Talagrand (1987c) and Ledoux and Talagrand (1991), who formulate similar results in terms of ultrametrics. For complete expositions of majorizing measures, see Adler (1990), pages 80ff., and Ledoux and Talagrand (1991), Chapters 11 and 12.

**A.4.** The basic inequality in this chapter was first proved by Talagrand (1989) and Ledoux and Talagrand (1991). The present elegant proof by induction is taken from Talagrand (1995), who discusses many extensions and applications. The name “isoperimetric inequality” appears not entirely appropriate. It resulted from the analogy with exponential inequalities for Gaussian variables, which can be derived from isoperimetric inequalities for the standard normal distribution. This asserts that, among all sets  $A$  with  $P(A)$  equal to a fixed number, the probability  $P(A^\varepsilon)$  increases the least as  $\varepsilon$  increases from zero for  $A$  of the form  $\{x: |x| > c\}$ . See Ledoux and Talagrand (1991) and Ledoux (1995) for further references.

**A.5.** In connection to Proposition A.5.2, consult the papers by Yurinskii (1977) and Dehling (1983).

**A.6.** Bennett’s inequality was proved by Bennett (1962); the function  $\psi$  was introduced by Shorack (1980). For Kiefer’s inequality, see Kiefer (1961); the present proof from Bennett’s inequality is new. For Talagrand’s inequality, see Talagrand (1994), and for the  $L_1$ -inequality for multinomial variables, see Bretagnolle and Huber (1978). Mason and Van Zwet (1987) prove inequality A.6.8 and use it to obtain an interesting refinement of the Komlos-Major-Tusnady inequality. Proposition A.6.9 is due to Pinelis (1994); it improves on conjectures of Eaton (1974). Problem A.6.2 was communicated to us by David Pollard.

# References

- Adler, R. J. (1990). An introduction to continuity, extrema, and related topics for general Gaussian processes. *IMS Lecture Notes—Monograph Series* **12**. Institute of Mathematical Statistics, Hayward, CA.
- Adler, R. J. and Brown, L. D. (1986). Tail behaviour for suprema of empirical processes. *Annals of Probability* **14**, 1–30.
- Adler, R. J. and Samorodnitsky, G. (1987). Tail behavior for the suprema of Gaussian processes with applications to empirical processes. *Annals of Probability* **15**, 1339–1351.
- Aldous, D. (1989). *Probability Approximations via the Poisson Clumping Heuristic*. Springer-Verlag, New York.
- Alexander, K. S. (1984). Probability inequalities for empirical processes and a law of the iterated logarithm. *Annals of Probability* **12**, 1041–1067. (Correction: *Annals of Probability* **15**, 428–430.)
- Alexander, K. S. (1985). The non-existence of a universal multiplier moment for the central limit theorem. *Lecture Notes in Mathematics* **1153**, 15–16, (eds., A. Beck, R. Dudley, M. Hahn, J. Kuelbs, and M. Marcus). Springer-Verlag, New York.
- Alexander, K. S. (1987a). The central limit theorem for weighted empirical processes indexed by sets. *Journal of Multivariate Analysis* **22**, 313–339.

- Alexander, K. S. (1987b). Central limit theorems for stochastic processes under random entropy conditions. *Probability Theory and Related Fields* **75**, 351–378.
- Alexander, K. S. (1987c). The central limit theorem for empirical processes on Vapnik-Červonenkis classes. *Annals of Probability* **15**, 178–203.
- Alexander, K. S. (1987d). Rates of growth and sample moduli for weighted empirical processes indexed by sets. *Probability Theory and Related Fields* **75**, 379–423.
- Alexander, K. S. and Pyke, R. (1986). A uniform central limit theorem for set-indexed partial-sum processes with finite variance. *Annals of Probability* **14**, 582–597.
- Andersen, N. T. (1985). The central limit theorem for non-separable valued functions. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **70**, 445–455.
- Andersen, N. T. and Dobrić, V. (1987). The central limit theorem for stochastic processes. *Annals of Probability* **15**, 164–177.
- Andersen, N. T. and Dobrić, V. (1988). The central limit theorem for stochastic processes II. *Journal of Theoretical Probability* **1**, 287–303.
- Andersen, N. T., Giné, E., Ossiander, M., and Zinn, J. (1988). The central limit theorem and the law of iterated logarithm for empirical processes under local conditions. *Probability Theory and Related Fields* **77**, 271–305.
- Anderson, T. W. (1963). Asymptotic theory for principal component analysis. *Annals of Mathematical Statistics* **34**, 122–148.
- Araujo, A. and Giné, E. (1980). *The Central Limit Theorem for Real and Banach Valued Random Variables*. John Wiley, New York.
- Arcones, M. and Giné, E. (1992). Bootstrap of  $M$  estimators and related statistical functionals. *Exploring the Limits of Bootstrap*, 14–47, (eds., R. LePage and L. Billard). Wiley, New York.
- Arcones, M. and Giné, E. (1993). Limit theorems for U-processes. *Annals of Probability* **21**, 1494–1452.
- Assouad, P. (1981). Sur les classes de Vapnik-Červonenkis. *Comptes Rendus des Séances de l'Académie des Sciences Series A* **292**, 921–924.
- Assouad, P. (1983). Densité et dimension. *Annales de l'Institut Fourier (Grenoble)* **33**(3), 233–282.
- Ball, K. and Pajor, A. (1990). The entropy of convex bodies with “few” extreme points. Geometry of Banach Spaces, Proceedings of the conference held in Strobl, Austria, 1989 (eds., P.F.X. Müller and W. Schachermayer). *London Mathematical Society Lecture Note Series* **158**, 25–32.

- Barron, A., Birgé, L., and Massart, P. (1995). Risk bounds for model selection via penalization. *Preprint*.
- Bass, R. F. and Pyke, R. (1985). The space  $D(A)$  and weak convergence for set-indexed processes. *Annals of Probability* **13**, 860–884.
- Bauer, H. (1981). *Probability Theory and Elements of Measure Theory*. Holt, Rinehart, and Winston, New York.
- Bennett, G. (1962). Probability inequalities for the sum of independent random variables. *Journal of the American Statistical Association* **57**, 33–45.
- Beran, R. and Millar, P. W. (1986). Confidence sets for a multivariate distribution. *Annals of Statistics* **14**, 431–443.
- Bickel, P. J. (1967). Some contributions to the theory of order statistics. Proc. Fifth Berkeley Symp. Math. Statist. Prob. **1**, 575–591. University of California Press, Berkeley.
- Bickel, P. J. (1969). A distribution free version of the Smirnov two sample test in the  $p$ -variate case. *Annals of Statistics* **50**, 1–23.
- Bickel, P. J., Klaassen, C. A. J., Ritov, Y., and Wellner, J. A. (1993). *Efficient and Adaptive Estimation for Semiparametric Models*. Johns Hopkins University Press, Baltimore.
- Billingsley, P. (1968). *Convergence of Probability Measures*. John Wiley, New York.
- Billingsley, P. (1971). *Weak Convergence of Measures: Applications in Probability*. Regional Conference Series in Mathematics **5**. Society for Industrial and Applied Mathematics, Philadelphia.
- Birgé, L. and Massart, P. (1993). Rates of convergence for minimum contrast estimators. *Probability Theory and Related Fields* **97**, 113–150.
- Birgé, L. and Massart, P. (1994). Minimum contrast estimators on sieves. Report **34**. Mathématiques, Université de Paris-Sud, Orsay.
- Birman, M. S. and Solomjak, M. Z. (1967). Piecewise-polynomial approximation of functions of the classes  $W_p$ . *Mathematics of the USSR Sbornik* **73**, 295–317.
- Blum, J. R. (1955). On the convergence of empiric distribution functions. *Annals of Mathematical Statistics* **26**, 527–529.
- Blumberg, H. (1935). The measurable boundaries of an arbitrary function. *Acta Mathematica* **65**, 263–282.
- Blum, J. R., Hanson, D. L., and Rosenblatt, J. I. (1963). On the central limit theorem for the sum of a random number of independent random variables. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **1**, 389–393.

- Blum, J. R., Kiefer, J., and Rosenblatt, M. (1961). Distribution free tests of fit based on the sample distribution function. *Annals of Mathematical Statistics* **32**, 485–498.
- Bolthausen, E. (1978). Weak convergence of an empirical process indexed by the closed convex subsets of  $I^2$ . *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **43**, 173–181.
- Borell, C. (1975). The Brunn-Minkowski inequality in Gauss space. *Inventiones Mathematicae* **30**, 205–216.
- Bretagnolle, J. and Huber, C. (1978). Lois empiriques et distance de Prokhorov. *Lecture Notes in Mathematics* **649**, 332–341.
- Bronštein, E. M. (1976). Epsilon-entropy of convex sets and functions. *Siberian Mathematics Journal* **17**, 393–398.
- Cantelli, F. P. (1933). Sulla determinazione empirica delle leggi di probabilità. *Giornale dell'Istituto Italiano degli Attuari* **4**, 421–424.
- Carl, B. and Stephani, I. (1990). *Entropy, Compactness, and the Approximation of Operators*. Cambridge University Press, Cambridge, England.
- Chibisov, D. M. (1965). An investigation of the asymptotic power of the tests of fit. *Theory of Probability and Its Applications* **10**, 421–437.
- Chow, Y. S. and Teicher, H. (1978). *Probability Theory*. Springer-Verlag, New York.
- Chung, K. L. (1951). The strong law of large numbers. Proc. Second Berkeley Symp. Math. Statist. Prob., 342–352, (eds., L.M. LeCam, J. Neyman, and E. Scott). University of California Press, Berkeley.
- Cohn, D. L. (1980). *Measure Theory*. Birkhäuser, Boston.
- Cramér, H. (1946). *Mathematical Methods of Statistics*. Princeton University Press, Princeton, NJ.
- Csörgő, S. (1974). On weak convergence of the empirical process with random sample size. *Acta Scientiarum Mathematicarum* **36**, 17–25. (Correction: *Acta Scientiarum Mathematicarum* **36**, 375–376.)
- Csörgő, S. (1981). Strong approximation of empirical Kac processes. *Annals of the Institute of Statistical Mathematics* **33**, 417–423.
- Csörgő, M. and Révész, P. (1978). Strong approximations of the quantile process. *Annals of Statistics* **6**, 882–894.
- Dehardt, J. (1971). Generalizations of the Glivenko-Cantelli theorem. *Annals of Mathematical Statistics* **42**, 2050–2055.
- Deheuvels, P. (1981). An asymptotic decomposition for multivariate distribution-free tests of independence. *Journal of Multivariate Analysis* **11**, 102–113.
- Dehling, H. (1983). Limit theorems for sums of weakly dependent Banach space valued random variables. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **63**, 393–432.

- Devroye, L. (1982). Bounds for the uniform deviations of empirical measures. *Journal of Multivariate Analysis* **13**, 72–79.
- 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.
- Doss, H. and Gill, R. D. (1992). A method for obtaining weak convergence results for quantile processes, with applications to censored survival data. *Journal of the American Statistical Association* **87**, 869–877.
- Dudley, R. M. (1966). Weak convergence of measures on nonseparable metric spaces and empirical measures on Euclidean spaces. *Illinois Journal of Mathematics* **10**, 109–126.
- Dudley, R. M. (1967a). Measures on nonseparable metric spaces. *Illinois Journal of Mathematics* **11**, 449–453.
- Dudley, R. M. (1967b). The sizes of compact subsets of Hilbert spaces 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 boundaries. *Journal of Approximation Theory* **10**, 227–236. (Correction: *Journal of Approximation Theory* **26** (1979), 192–193.)
- Dudley, R. M. (1976). *Probabilities and Metrics: Convergence of Laws on Metric Spaces*. Mathematics Institute Lecture Note Series **45**. Aarhus University, Aarhus, Denmark.
- Dudley, R. M. (1978a). Central limit theorems for empirical measures. *Annals of Probability* **6**, 899–929. (Correction: *Annals of Probability* **7**, 909–911.)
- Dudley, R. M. (1978b). Review of Sudakov (1976). *Mathematical Reviews* **55**, 606–607.
- Dudley, R. M. (1979). Balls in  $R^k$  do not cut all subsets of  $k + 2$  points. *Advances in Mathematics* **31**, 306–308.
- Dudley, R. M. (1984). A course on empirical processes (École d'Été de Probabilités de Saint-Flour XII-1982). *Lecture Notes in Mathematics* **1097**, 2–141, (ed., P.L. Hennequin). Springer-Verlag, New York.

- Dudley, R. M. (1985). An extended Wichura theorem, definition of Donsker class, and weighted empirical distributions. *Lecture Notes in Mathematics* **1153**, 141–178. Springer-Verlag, New York.
- Dudley, R. M. (1987). Universal Donsker classes and metric entropy. *Annals of Probability* **15**, 1306–1326.
- Dudley, R. M. (1989). *Real Analysis and Probability*. Wadsworth, Pacific Grove.
- Dudley, R. M. (1990). Nonlinear functionals of empirical measures and the bootstrap. Probability in Banach Spaces VII, Progress in Probability **21**, 63–82, (eds., E. Eberlein, J. Kuelbs and M.B. Marcus). Birkhäuser, Boston.
- Dudley, R. M. (1992). Fréchet differentiability,  $p$ -variation and uniform Donsker classes. *Annals of Probability* **20**, 1968–1982.
- Dudley, R. M. (1993). *Real Analysis and Probability*. Second Printing (corrected). Chapman and Hall, New York.
- Dudley, R. M. (1994). The order of the remainder in derivatives of composition and inverse operators for  $p$ -variation norms. *Annals of Statistics* **22**, 1–20.
- Dudley, R. M., Giné, E., and Zinn, J. (1991). Uniform and universal Glivenko-Cantelli classes. *Journal of Theoretical Probability* **4**, 485–510.
- Dudley, R. M. and Koltchinskii, V. (1994). Envelope moment conditions and Donsker classes. *Probability Theory and Mathematical Statistics* **51**, 39–49.
- Dudley, R. M. and Philipp, W. (1983). Invariance principles for sums of Banach space valued random elements and empirical processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **62**, 509–552.
- Dugue, D. (1975). Sur des tests d'indépendance indépendants de la loi. *Comptes Rendus des Séances de l'Académie des Sciences Series A* **281**, 1103–1104.
- Durst, M. and Dudley, R. M. (1981). Empirical processes, Vapnik-Chervonenkis classes and Poisson processes. *Probability and Mathematical Statistics* **1**, 109–115.
- Dvoretzky, A., Kiefer, J., and Wolfowitz, J. (1956). Asymptotic minimax character of the sample distribution function and of the classical multinomial estimator. *Annals of Mathematical Statistics* **27**, 642–669.
- Eames, W. and May, L. E. (1967). Measurable cover functions. *Canadian Mathematics Bulletin* **10**, 519–523.
- Eaton, M. L. (1974). A probability inequality for linear combinations of bounded random variables. *Annals of Statistics* **2**, 609–614.

- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *Annals of Statistics* **7**, 1–26.
- Erdős, P. and Kac, M. (1946). On certain limit theorems in the theory of probability. *Bulletin of the American Mathematical Society* **52**, 292–302.
- Fernandez, P. J. (1970). A weak convergence theorem for random sums of independent random variables. *Annals of Mathematical Statistics* **41**, 710–712.
- Fernique, X. (1974). Régularité des trajectoires des fonctions aléatoires gaussiennes. *Lecture Notes in Mathematics* **480**, 1–96. Springer-Verlag, Berlin.
- Fernique, X. (1978). Caractérisation des processus à trajectoires majorées ou continues. *Lecture Notes in Mathematics* **649**, 691–706. Springer-Verlag, Berlin.
- Fernique, X. (1985). Sur la convergence étroite des mesures gaussiennes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **68**, 331–336.
- Fortet, R. and Mourier, E. (1953). Convergence de la répartition empirique vers la répartition théorique. *Annales Scientifiques de l’École Normale Supérieure* **70**, 266–285.
- Frankl, P. (1983). On the trace of finite sets. *Journal of Combinatorial Theory A* **34**, 41–45.
- Gaenssler, P. (1983). *Empirical Processes*. Institute of Mathematical Statistics, Hayward, CA.
- Gill, R. D. (1989). Non- and semi-parametric maximum likelihood estimators and the von-Mises method (part I). *Scandinavian Journal of Statistics* **16**, 97–128.
- Gill, R. D. (1994). Lectures on survival analysis (École d’Été de Probabilités de Saint-Flour XXII-1992). *Lecture Notes in Mathematics* **1581**, 115–241, (ed., P. Bernard). Springer-Verlag, New York.
- Gill, R. D. and Johansen, S. (1990). A survey of product-integration with a view towards application in survival analysis. *Annals of Statistics* **18**, 1501–1555.
- Giné, E. and Zinn, J. (1984). Some limit theorems for empirical processes. *Annals of Probability* **12**, 929–989.
- Giné, E. and Zinn, J. (1986a). Lectures on the central limit theorem for empirical processes. *Lecture Notes in Mathematics* **1221**, 50–11. Springer-Verlag, Berlin.
- Giné, E. and Zinn, J. (1986b). Empirical processes indexed by Lipschitz functions. *Annals of Probability* **14**, 1329–1338.
- Giné, E. and Zinn, J. (1990). Bootstrapping general empirical measures. *Annals of Probability* **18**, 851–869.

- Giné, E. and Zinn, J. (1991). Gaussian characterization of uniform Donsker classes of functions. *Annals of Probability* **19**, 758–782.
- Glivenko, V. (1933). Sulla determinazione empirica della leggi di probabilità. *Giornale dell'Istituto Italiano degli Attuari* **4**, 92–99.
- Gnedenko, B. V. and Kolmogorov, A. N. (1954). *Limit Distributions for Sums of Independent Random Variables*. Addison Wesley, Reading, MA (revised version, 1968).
- Goodman, V. (1976). Distribution estimates for functionals of the two-parameter Wiener process. *Annals of Probability* **4**, 977–982.
- Greenwood, P. E. and Shirayev, A. N. (1985). *Contiguity and the Statistical Invariance Principle*. Gordon and Breach, New York.
- Grenander, U. (1956). On the theory of mortality measurement, part II. *Skandinavisk Aktuarietidskrift* **39**, 125–153.
- Groeneboom, P. (1983). The concave majorant of Brownian motion. *Annals of Probability* **11**, 1016–1027.
- Groeneboom, P. (1985). Estimating a monotone density. Proceedings of the Berkeley Conference in Honor of Jerzy Neyman and Jack Kiefer, Vol. II, 539–555, (eds., Lucien M. LeCam and Richard A. Olshen). Wadsworth, Monterey, CA.
- Groeneboom, P. (1987). Asymptotics for interval censored observations. Report **87-18**. Department of Mathematics, University of Amsterdam.
- Groeneboom, P. (1988). Brownian Motion with a parabolic drift and Airy functions. *Probability Theory and Related Fields* **81**, 79–109.
- Groeneboom, P. and Wellner, J. A. (1992). *Information Bounds and Nonparametric Maximum Likelihood Estimation*. Birkhäuser, Basel.
- Grübel, R. and Pitts, S. M. (1992). A functional approach to the stationary waiting time and idle period distributions of the GI/G/1 queue. *Annals of Probability* **20**, 1754–1778.
- Grübel, R. and Pitts, S. M. (1993). Nonparametric estimation in renewal theory I: The empirical renewal function. *Annals of Statistics* **21**, 1431–1451.
- Hájek, J. (1961). Some extensions of the Wald-Wolfowitz-Noether theorem. *Annals of Mathematical Statistics* **32**, 506–523.
- Hájek, J. (1970). A characterization of limiting distributions of regular estimates. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **14**, 323–330.
- Hájek, J. (1972). Local asymptotic minimax and admissibility in estimation. Proc. Sixth Berkeley Symp. Math. Statist. Prob. **1**, 175–194, (eds., L.M. LeCam, J. Neyman, and E. Scott).
- Hájek, J. and Šidák, Z. (1967). *Theory of Rank Tests*. Academic Press, New York.

- Hammersley, J. M. (1952). An extension of the Slutsky-Fréchet theorem. *Acta Mathematica* **87**, 243–257.
- Haussler, D. (1995). Sphere packing numbers for subsets of the Boolean  $n$ -cube with bounded Vapnik-Chervonenkis dimension. *Journal of Combinatorial Theory A* **69**, 217–232.
- Heesterman, C. C. and Gill, R. D. (1992). A central limit theorem for  $M$ -estimators by the von Mises method. *Statistica Neerlandica* **46**, 165–177.
- Hjort, N. L. and Fenstad, G. (1992). On the last time and the number of times an estimator is more than epsilon from its target value. *Annals of Statistics* **20**, 469–489.
- Hoeffding, W. (1948). A non-parametric test of independence. *Annals of Mathematical Statistics* **19**, 546–557.
- Hoeffding, W. (1963). Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association* **58**, 13–30.
- Hoffmann-Jørgensen, J. (1970). *The Theory of Analytic Spaces*. Various Publications Series **10**. Matematisk Institut, Aarhus Universitet, Aarhus, Denmark.
- Hoffmann-Jørgensen, J. (1973). Sums of independent Banach space valued random variables. *Aarhus Univ. Preprint Series 1972/73* **15**.
- Hoffmann-Jørgensen, J. (1974). Sums of independent Banach space valued random variables. *Studia Mathematica* **52**, 159–186.
- Hoffmann-Jørgensen, J. (1976). Probability in Banach spaces. *Lecture Notes in Mathematics* **598**, 164–229. Springer-Verlag, New York.
- Hoffmann-Jørgensen, J. (1984). *Stochastic Processes on Polish Spaces*. unpublished.
- Hoffmann-Jørgensen, J. (1991). *Stochastic Processes on Polish Spaces*. Various Publication Series **39**. Aarhus Universitet, Aarhus, Denmark.
- Hoffmann-Jørgensen, J. (1994). *Probability with a View Towards Statistics*. Chapman and Hall, New York.
- Hogan, M. L. and Siegmund, D. (1986). Large deviations for the maxima of some random fields. *Advances in Applied Mathematics* **7**, 2–22.
- Huang, J. (1993). Central limit theorems for  $M$ -estimates. Report **251**. Department of Statistics, University of Washington, Seattle.
- Huang, J. (1996). Efficient estimation for the Cox model with interval censoring. *Annals of Statistics*, to appear.
- Huber, P. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. Proc. Fifth Berkeley Symp. Math. Statist. Probab. **1**, 221–233, (eds., L.M. LeCam and J. Neyman). University of California Press, Berkeley.

- Ibragimov, I. A. and Has'minskii, R. Z. (1981). *Statistical Estimation: Asymptotic Theory*. Springer-Verlag, New York.
- Ibragimov, I. A., Sudakov, V. N., and Tsirel'son, B. S. (1976). Norms of Gaussian sample functions. Proceedings of the Third Japan-USSR Symposium on Probability Theory. *Lecture Notes in Mathematics* **550**, 20–41. Springer-Verlag, New York.
- Jain, N. and Marcus, M. (1975). Central limit theorems for  $C(S)$ -valued random variables. *Journal Functional Analysis* **19**, 216–231.
- Jain, N. and Marcus, M. (1978). Continuity of subgaussian processes. *Probability on Banach Spaces, Advances in Probability* **4**, 81–196. Dekker, New York.
- Jameson, J. O. (1974). *Topology and Normed Spaces*. Chapman and Hall, London.
- Jongbloed, G. (1995). *Three Statistical Inverse Problems*. Department of Mathematics, Delft University, Delft, Netherlands.
- Jupp, P. E. and Spurr, B. D. (1985). Sobolev tests for independence of directions. *Annals of Statistics* **13**, 1140–1155.
- Kac, M. (1949). On deviations between theoretical and empirical distributions. *Proceedings of the National Academy of Sciences USA* **35**, 252–257.
- Kahane, J.-P. (1968). *Some Random Series of Functions*. Heath, Lexington, (2nd edition: Cambridge University Press, Cambridge, England, 1985).
- Kelley, J. L. (1955). *General Topology*. Van Nostrand, Princeton, NJ.
- Kiefer, J. (1961). On large deviations of the empiric d.f. of vector chance variables and a law of iterated logarithm. *Pacific Journal of Mathematics* **11**, 649–660.
- Kiefer, J. (1969). On the deviations in the Skorokhod-Strassen approximation scheme. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **13**, 321–332.
- Kiefer, J. and Wolfowitz, J. (1958). On the deviations of the empiric distribution function of vector chance variables. *Transactions of the American Mathematical Society* **87**, 173–186.
- Kiefer, J. and Wolfowitz, J. (1959). Asymptotic minimax character of the sample distribution function for vector chance variables. *Annals of Mathematical Statistics* **30**, 463–489.
- Kim, J. and Pollard, D. (1990). Cube root asymptotics. *Annals of Statistics* **18**, 191–219.
- Klaassen, C. A. J. and Wellner, J. A. (1992). Kac empirical processes and the bootstrap. Proceedings of the Eighth International Conference on Probability in Banach Spaces, 411–429, (eds., M. Hahn and J. Kuelbs). Birkhäuser, New York.

- Kolčinskii, V. I. (1981). On the central limit theorem for empirical measures. *Theory of Probability and Mathematical Statistics* **24**, 71–82.
- Kolmogorov, A. N. (1933). Sulla determinazione empirica di una legge di distribuzione. *Giornale dell'Istituto Italiano degli Attuari* **4**, 83–91.
- Kolmogorov, A. N. and Tikhomirov, V. M. (1961). Epsilon-entropy and epsilon-capacity of sets in function spaces. *American Mathematical Society Translations, series 2* **17**, 277–364.
- Koul, H. (1970). Some convergence theorems for ranks and weighted empirical cumulatives. *Annals of Mathematical Statistics* **41**, 1768–1773.
- Kuelbs, J. (1976). A strong convergence theorem for Banach space valued random variables. *Annals of Probability* **4**, 744–771.
- Landau, H. and Shepp, L. A. (1970). On the supremum of a Gaussian process. *Sankhyà A* **32**, 369–378.
- Le Cam, L. (1957). Convergence in distribution of stochastic processes. *University of California Publications in Statistics* **2**, 207–236.
- Le Cam, L. (1960). Locally asymptotically normal families of distributions. *University of California Publications in Statistics* **3**, 37–98.
- Le Cam, L. (1969). *Théorie Asymptotique de la Décision Statistique*. Les Presses de l'Université de Montréal.
- Le Cam, L. (1970a). On the assumptions used to prove asymptotic normality of maximum likelihood estimates. *Annals of Mathematical Statistics* **41**, 802–828.
- Le Cam, L. (1970b). Remarques sur le théorème limite central dans les espaces localement convexes. *Les Probabilités sur les Structures Algébriques*, 233–249. Centre National des Recherches Scientifiques, Paris.
- Le Cam, L. (1972). Limits of experiments. Proc. Sixth Berkeley Symp. Math. Statist. Probab. **1**, 245–261, (eds., L.M. LeCam, J. Neyman, and E. Scott). University of California Press, Berkeley.
- Le Cam, L. (1986). *Asymptotic Methods in Statistical Decision Theory*. Springer-Verlag, New York.
- Le Cam, L. (1989). On measurability and convergence in distribution. Report **211**. Department of Statistics, University of California, Berkeley.
- Ledoux, M. (1995). Isoperimetry and Gaussian analysis (École d'Été de Probabilités de Saint-Flour XXIV-1994). *Lecture Notes in Mathematics to appear* (ed., P. Bernard). Springer-Verlag, New York.
- Ledoux, M. and Talagrand, M. (1986). Conditions d'intégrabilité pour les multiplicateurs dans le TLC Banachique. *Annals of Probability* **14**, 916–921.

- Ledoux, M. and Talagrand, M. (1988). Un critère sur les petite boules dans le théorème limite central. *Probability Theory and Related Fields* **77**, 29–47.
- Ledoux, M. and Talagrand, M. (1991). *Probability in Banach Spaces: Isoperimetry and Processes*. Springer-Verlag, Berlin.
- Lehmann, E. L. (1983). *Theory of Point Estimation*. Wiley, New York.
- Levit, B. Ya. (1978). Infinite-dimensional informational lower bounds. *Theory of Probability and Its Applications* **23**, 388–394.
- 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.
- Lorentz, G. G. (1966). *Approximation of Functions*. Holt, Rhinehart, Winston, New York.
- Mammen, E. (1991). Nonparametric regression under qualitative smoothness assumptions. *Annals of Statistics* **19**, 741–759.
- Marcus, M. B. and Shepp, L. A. (1971). Sample behaviour of Gaussian processes. Proc. Sixth Berkeley Symp. Math. Statist. Probab. **2**, 423–442.
- Marcus, M. and Zinn, J. (1984). The bounded law of the iterated logarithm for the weighted empirical distribution in the non-iid case. *Annals of Probability* **12**, 334–360.
- Marshall, A. W. and Olkin, I. (1979). *Inequalities: Theory of Majorization and Its Applications*. Academic Press, New York.
- Mason, D. M. and Newton, M. (1990). A rank statistics approach to the consistency of a general bootstrap. *Annals of Statistics* **20**, 1611–1624.
- Mason, D. M. and Van Zwet, W. R. (1987). A refinement of the KMT inequality for the uniform empirical process. *Annals of Probability* **15**, 871–894.
- Massart, P. (1986). Rates of convergence in the central limit theorem for empirical processes. *Annales Institut Henri Poincaré* **22**, 381–423.
- Massart, P. (1989). Strong approximation for multivariate empirical and related processes, via KMT constructions. *Annals of Probability* **17**, 266–291.
- Massart, P. (1990). The tight constant in the Dvoretzky-Kiefer-Wolfowitz inequality. *Annals of Probability* **18**, 1269–1283.
- May, L. E. (1973). Separation of functions. *Canadian Mathematics Bulletin* **16**, 245–250.
- Millar, P. W. (1979). Asymptotic minimax theorems for the sample distribution function. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **48**, 233–252.

- Millar, P. W. (1983). The minimax principle in asymptotic statistical theory. École d'Été de Probabilités de St. Flour XI. *Lecture Notes in Mathematics* **976**, 76–267. Springer-Verlag, New York.
- Millar, P. W. (1985). Non-parametric applications of an infinite dimensional convolution theorem. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **68**, 545–556.
- Mitoma, I. (1983). Tightness of probabilities on  $C([0, 1]; S')$  and  $D([0, 1]; S')$ . *Annals of Probability* **11**, 989–999.
- Müller, D. W. (1968). Verteilungs-Invarianzprinzipien für das starke Gesetz der grossen Zahl. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **10**, 173–192.
- Munkres, J. (1975). *Topology; a First Course*. Prentice-Hall, Englewood Cliffs, NJ.
- Murphy, S. A. (1994). Consistency in a proportional hazards model incorporating a random effect. *Annals of Statistics* **22**, 712–731.
- Murphy, S. A. (1995). Asymptotic theory for the frailty model. *Annals of Statistics* **23**, 182–198.
- Nolan, D. (1992). Asymptotics for multivariate trimming. *Stochastic Processes and Their Applications* **42**, 157–169.
- Ossiander, M. (1987). A central limit theorem under metric entropy with  $L_2$  bracketing. *Annals of Probability* **15**, 897–919.
- Ottaviani, G. (1939). Sulla teoria astratta del calcolo delle probabilità proposta dal Cantelli. *Giornale dell'Istituto Italiano degli Attuari* **10**, 10–40.
- Parthasarathy, K. R. (1967). *Probability Measures on Metric Spaces*. Academic Press, New York.
- Pinelis, I. (1994). Extremal probabilistic problems and Hotelling's  $T^2$  test under a symmetry condition. *Annals of Statistics* **22**, 357–368.
- Pisier, G. (1981). Remarques sur un résultat non publié de B. Maurey. Séminaire d'analyse Fonctionnelle, 1980–1981, Exposé No. 5. École Polytechnique, Palaiseau.
- Pisier, G. (1983). Some applications of the metric entropy condition to harmonic analysis. Banach spaces, Harmonic Analysis and Probability. *Lecture Notes in Mathematics* **995**, 123–154. Springer-Verlag, New York.
- Pisier, G. (1984). Remarques sur les classes de Vapnik-Červonenkis. *Annales Institut Henri Poincaré* **20**, 287–298.
- Pisier, G. (1986). Probabilistic methods in the geometry of Banach space. *Lecture Notes in Mathematics* **1206**, 167–241. Springer-Verlag, Berlin.
- Pisier, G. (1989). *The Volume of Convex Bodies and Banach Space Geometry*. Cambridge University Press, Cambridge, England.

- Pitts, S. M. (1994). Nonparametric estimation of the stationary waiting time distribution function for the GI/G/1 queue. *Annals of Statistics* **22**, 1428–1446.
- Pollard, D. (1981). Limit theorems for empirical processes. *Zeitschrift für Wahrscheinlichkeitstheorie und Verwandte Gebiete* **57**, 181–195.
- Pollard, D. (1982). A central limit theorem for empirical processes. *Journal of the Australian Mathematical Society A* **33**, 235–248.
- Pollard, D. (1984). *Convergence of Stochastic Processes*. Springer-Verlag, New York.
- Pollard, D. (1985). New ways to prove central limit theorems. *Econometric Theory* **1**, 295–314.
- Pollard, D. (1989a). A maximal inequality for sums of independent processes under a bracketing condition. *Preprint*.
- Pollard, D. (1989b). Asymptotics via empirical processes. *Statistical Science* **4**, 341–366.
- Pollard, D. (1990). *Empirical Processes: Theory and Applications. NSF-CBMS Regional Conference Series in Probability and Statistics* **2**. Institute of Mathematical Statistics and American Statistical Association.
- Pollard, D. (1995). Uniform ratio limit theorems for empirical processes. *Scandinavian Journal of Statistics* **22**, 271–278.
- Praestgaard, J. (1995). Permutation and bootstrap Kolmogorov-Smirnov tests for the equality of two distributions. *Scandinavian Journal of Statistics* **22**, 305–322.
- Praestgaard, J. and Wellner, J. A. (1993). Exchangeably weighted bootstraps of the general empirical process. *Annals of Probability* **21**, 2053–2086.
- Prakasa Rao, B. L. S. (1969). Estimation of a unimodal density. *Sankya Series A* **31**, 23–36.
- Preston, C. (1972). Continuity properties of some Gaussian processes. *Annals of Mathematical Statistics* **43**, 285–292.
- Prohorov, Yu. V. (1956). Convergence of random processes and limit theorems in probability. *Theory of Probability and Its Applications* **1**, 157–214.
- Pyke, R. (1968). The weak convergence of the empirical process with random sample size. *Proceedings of the Cambridge Philosophical Society* **64**, 155–160.
- Pyke, R. (1969). Applications of almost surely convergent constructions of weakly convergent processes. *Lecture Notes in Mathematics* **89**, 187–200. Springer-Verlag, New York.

- Pyke, R. (1970). Asymptotic results for rank statistics. *Nonparametric Techniques in Statistical Inference*, 21–37, (ed., M.L. Puri). Cambridge University Press, Cambridge, England.
- Pyke, R. (1983). A uniform central limit theorem for partial-sum processes indexed by sets. *Probability, Statistics and Analysis*, 219–240, (eds., J.F.C. Kingman and G.E.H. Reuter). Cambridge University Press, Cambridge, England.
- Pyke, R. (1984). Asymptotic results for empirical and partial sum processes: a review. *Canadian Journal of Statistics* **12**, 241–264.
- Pyke, R. (1992). Probability in mathematics and statistics: A century's predictor of future directions. *Jahresbericht Deutschen Mathematiker-Vereinigung, Jubiläumstagung 1990*, 239–264, (ed., W.-D. Geyer). Teubner, Stuttgart.
- Pyke, R. and Shorack, G. R. (1968). Weak convergence of a two-sample empirical process and a new approach to Chernoff-Savage theorems. *Annals of Mathematical Statistics* **39**, 755–771.
- Quiroz, A. J., Nakamura, M., and Perez, F. J. (1995). Estimation of a multivariate Box-Cox transformation to elliptical symmetry via the empirical characteristic function. *Preprint*.
- Ranga Rao, R. (1962). Relations between weak and uniform convergence of measures with applications. *Annals of Mathematical Statistics* **33**, 659–680.
- Reeds, J. A. (1976). *On the Definition of von Mises Functionals*. Ph.D. dissertation, Department of Statistics, Harvard University, Cambridge, MA.
- Révész, P. (1976). Three theorems of multivariate empirical process. *Lecture Notes in Mathematics* **566**, 106–126. Springer-Verlag, New York.
- Revuz, D. and Yor, M. (1994). *Continuous Martingales and Brownian Motion, Second Edition*. Springer-Verlag, New York.
- Robertson, T., Wright, F. T., and Dykstra, R. L. (1988). *Order Restricted Statistical Inference*. Wiley, New York.
- Roussas, G. G. (1972). *Contiguity of Probability Measures*. Cambridge University Press, London.
- Rudin, W. (1966). *Real and Complex Analysis*. McGraw-Hill, New York.
- Rudin, W. (1973). *Functional Analysis*. McGraw-Hill, New York.
- Runnenburg, J. Th. and Vervaat, W. (1969). Asymptotical independence of the lengths of subintervals of a randomly partitioned interval; a sample from S. Ikeda's work. *Statistica Neerlandica* **23**, 67–77.
- Samorodnitsky, G. (1991). Probability tails of Gaussian extrema. *Stochastic Processes and Their Applications* **38**, 55–84.

- Sauer, N. (1972). On the density of families of sets. *Journal of Combinatorial Theory A* **13**, 145–147.
- Sheehy, A. and Wellner, J. A. (1992). Uniform Donsker classes of functions. *Annals of Probability* **20**, 1983–2030.
- Shen, X. and Wong, W. H. (1994). Convergence rate of sieve estimates. *Annals of Statistics* **22**, 580–615.
- Shorack, G. R. (1972). Convergence of quantile and spacings processes with applications. *Annals of Mathematical Statistics* **43**, 1400–1411.
- Shorack, G. R. (1973). Convergence of reduced empirical and quantile processes with application to functions of order statistics in the non-iid case. *Annals of Statistics* **1**, 146–152.
- Shorack, G. R. (1979). The weighted empirical process of row independent random variables with arbitrary distribution functions. *Statistica Neerlandica* **33**, 169–189.
- Shorack, G. R. (1980). Some law of the iterated logarithm type results for the empirical process. *Australian Journal of Statistics* **22**, 50–59.
- Shorack, G. R. and Beirlant, J. (1986). The appropriate reduction of the weighted empirical process. *Statistica Neerlandica* **40**, 123–128.
- Shorack, G. R. and Wellner, J. A. (1986). *Empirical Processes with Applications to Statistics*. Wiley, New York.
- Siegmund, D. (1988). Approximate tail probabilities for the maxima of some random fields. *Annals of Probability* **16**, 487–501.
- Siegmund, D. (1992). Tail approximations for maxima of random fields. *Proceedings of the 1989 Singapore Probability Conference*, 147–158. Walter de Gruyter, Berlin.
- Skorokhod, A. V. (1956). Limit theorems for stochastic processes. *Theory of Probability and Its Applications* **1**, 261–290.
- Slepian, D. (1962). The one-sided barrier problem for Gaussian noise. *Bell System Technical Journal* **42**, 463–501.
- Smith, D. and Dudley, R. M. (1992). Exponential bounds in Vapnik-Červonenkis classes of index 1. *Probability in Banach Spaces VIII, Progress in Probability*, 451–465, (eds., R.M. Dudley, M.G. Hahn, and J. Kuelbs).
- Smolyanov, O. G. and Fomin, S. V. (1976). Measure on linear topological spaces. *Russian Mathematical Surveys* **31**, 1–53.
- Stengle, G. and Yukich, J. E. (1989). Some new Vapnik-Chervonenkis classes. *Annals of Statistics* **17**, 1441–1446.
- Stone, C. (1963). Weak convergence of stochastic processes defined on semi-infinite time intervals. *Proceedings of the American Mathematical Society* **14**, 694–696.

- Strassen, V. and Dudley, R. M. (1969). The central limit theorem and epsilon-entropy. *Lecture Notes in Mathematics* **89**, 224–231. Springer-Verlag, New York.
- Strobl, F. (1992). On the reversed sub-martingale property of empirical discrepancies in arbitrary sample spaces. Report **53**. Universität München.
- Sudakov, V. N. (1969). Gauss and Cauchy measures and  $\epsilon$ -entropy. *Doklady Akademii Nauk SSSR* **185**, 51–53.
- Sudakov, V. N. (1971). Gaussian random processes and measures of solid angles in Hilbert spaces. *Soviet Mathematics-Doklady* **12**, 412–415.
- Sudakov, V. N. (1976). *Geometric Problems of the Theory of Infinite-Dimensional Probability Distributions*. Trudy Matematicheskogo Instituta im. V.A. Steklova **141**.
- Sudakov, V. N. and Tsirel'son, B. S. (1978). Extremal properties of half-spaces for spherically invariant measures. *Journal of Soviet Mathematics* **9**, 9–18.
- Sun, T. G. and Pyke, R. (1982). Weak convergence of empirical processes. Report **19**. Department of Statistics, University of Washington, Seattle.
- Talagrand, M. (1982). Closed convex hull of set of measurable functions, Riemann-measurable functions and measurability of translations. *Annales de l'Institut Fourier (Grenoble)* **32**, 39–69.
- Talagrand, M. (1987a). Measurability problems for empirical processes. *Annals of Probability* **15**, 204–212.
- Talagrand, M. (1987b). Donsker classes and random geometry. *Annals of Probability* **15**, 1327–1338.
- Talagrand, M. (1987c). Regularity of Gaussian processes. *Acta Mathematica* **159**, 99–149.
- Talagrand, M. (1988). Donsker classes of sets. *Probability Theory and Related Fields* **78**, 169–191.
- Talagrand, M. (1989). Isoperimetry and integrability of the sum of independent Banach-space valued random variables. *Annals of Probability* **17**, 1546–1570.
- Talagrand, M. (1992). A simple proof of the majorizing measure theorem. *Geometric and Functional Analysis* **2**, 118–125.
- Talagrand, M. (1994a). Sharper bounds for Gaussian and empirical processes. *Annals of Probability* **22**, 28–76.
- Talagrand, M. (1995). Concentration of measure and isoperimetric inequalities in product spaces. *Publications Mathématiques IHES* **81**, 73–205.

- Topsoe, F. (1967a). On the connection between  $P$ -continuity and  $P$ -uniformity in weak convergence. *Theory of Probability and its Applications* **12**, 281–290.
- Topsoe, F. (1967b). Preservation of weak convergence under mappings. *Annals of Mathematical Statistics* **38**, 1661–1665.
- Tsirel'son, V. S. (1975). The density of the distribution of the maximum of a Gaussian process. *Theory of Probability and Its Applications* **20**, 847–856.
- Van de Geer, S. (1990). Estimating a regression function. *Annals of Statistics* **18**, 907–924.
- Van de Geer, S. (1991). The entropy bound for monotone functions. Report **91-10**. University of Leiden.
- Van de Geer, S. (1993a). Hellinger-consistency of certain nonparametric maximum likelihood estimators. *Annals of Statistics* **21**, 14–44.
- Van de Geer, S. (1993b). The method of sieves and minimum contrast estimators. Report **93-06**. University of Leiden.
- Van der Laan, M. (1993). *Efficient and Inefficient Estimation in Semiparametric Models*. Ph.D. dissertation, University of Utrecht.
- Van der Vaart, A. W. (1988). *Statistical Estimation in Large Parameter Spaces*. CWI Tracts **44**. Center for Mathematics and Computer Science, Amsterdam.
- Van der Vaart, A. W. (1991). Efficiency and Hadamard differentiability. *Scandinavian Journal of Statistics* **18**, 63–75.
- Van der Vaart, A. W. (1993). New Donsker classes. *Annals of Probability*, to appear.
- Van der Vaart, A. W. (1994a). Maximum likelihood estimation with partially censored data. *Annals of Statistics* **22**, 1896–1916.
- Van der Vaart, A. W. (1994b). Bracketing smooth functions. *Stochastic Processes and Their Applications* **52**, 93–105.
- Van der Vaart, A. W. (1994c). On a model of Hasminskii and Ibragimov. Proceedings Kolmogorov Semester at the Euler International Mathematical Institute, St. Petersburg (ed., A. Yu. Zaitsev). North Holland, Amsterdam.
- Van der Vaart, A. W. (1995). Efficiency of infinite dimensional  $M$ -estimators. *Statistica Neerlandica* **49**, 9–30.
- Van der Vaart, A. W. (1996). Efficient estimation in semiparametric models. *Annals of Statistics*, to appear.
- Van der Vaart, A. W. and Wellner, J. A. (1989). Prohorov and continuous mapping theorems in the Hoffmann-Jorgensen weak convergence theory with applications to convolution and asymptotic minimax theorems. *Preprint*. Department of Statistics, University of Washington.

- Van Zuijlen, M. (1978). Properties of the empirical distribution function for independent not identically distributed random variables. *Annals of Probability* **6**, 250–266.
- Vapnik, V. N. (1982). *Estimation of Dependences*. Springer-Verlag, New York.
- Vapnik, V. N. and Červonenkis, A. Ya. (1968). On the uniform convergence of frequencies of occurrence events to their probabilities. *Soviet Mathematics-Doklady* **9**, 915–918.
- Vapnik, V. N. and Červonenkis, 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 Červonenkis, A. Ya. (1974). *The Theory of Pattern Recognition*. Nauka, Moscow.
- Vapnik, V. N. and Červonenkis, A. Ya. (1981). Necessary and sufficient conditions for uniform convergence of means to mathematical expectations. *Theory of Probability and Its Applications* **26**, 147–163.
- Varadarajan, V. S. (1961). Measures on topological spaces. *Mathematics of the USSR Sbornik* **55**, 35–100. (Translation: (1965). *American Mathematical Society Translations, series 2* **48**, 161–228, American Mathematical Society, Providence.)
- Wang, Y. (1993). The limiting distribution in concave regression. *Preprint*. Department of Statistics, University of Missouri, Columbia.
- Wellner, J. A. (1977). Distributions related to linear bounds for the empirical distribution function. *Annals of Statistics* **5**, 1003–1016.
- Wellner, J. A. (1989). Discussion of Gill's paper “Non- and semi-parametric maximum likelihood estimators and the von Mises method (part I)”. *Scandinavian Journal of Statistics* **16**, 124–127.
- Wellner, J. A. (1992). Empirical processes in action: A review. *International Statistical Review* **60**, 247–269.
- Wenocur, R. S. and Dudley, R. M. (1981). Some special Vapnik-Cervonenkis classes. *Discrete Mathematics* **33**, 313–318.
- 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 Operations Research* **5**, 67–85.
- Wichura, M. J. (1968). *On the Weak Convergence of Non-Borel Probabilities on a Metric Space*. Ph.D. dissertation, Columbia University, New York.
- Wichura, M. J. (1970). On the construction of almost uniformly convergent random variables with given weakly convergent image laws. *Annals of Mathematical Statistics* **41**, 284–291.

- Wong, W. H. and Severini, T. A. (1991). On maximum likelihood estimation in infinite dimensional parameter spaces. *Annals of Statistics* **19**, 603–632.
- Wong, W. H. and Shen, X. (1995). Probability inequalities for likelihood ratios and convergence rates of sieve MLE's. *Annals of Statistics* **23**, to appear.
- Yurinskii, V. V. (1974). Exponential bounds for large deviations. *Theory of Probability and Its Applications* **19**, 154–155.
- Yurinskii, V. V. (1977). On the error of the Gaussian approximation for convolutions. *Theory of Probability and Its Applications* **22**, 236–247.

# Author Index

- Adler, R. J. . . . . 248, 275, 426, 442, 443, 465, 466  
Aldous, D. . . . . 442–444, 466  
Alexander, K. S. . . . . 142, 272–274, 275  
Andersen, N. T. . . . . 76, 270, 273  
Anderson, T. W. . . . . 77, 428  
Araujo, A. . . . . 76, 425  
Arcones, M. . . . . 270, 426  
Assouad, P. . . . . 153, 271
- Ball, K. . . . . 270, 271  
Barron, A. . . . . 425  
Bass, R. F. . . . . 274  
Bauer, H. . . . . 22, 26  
Beirlant, J. . . . . 273  
Bennett, G. . . . . 103, 269, 460  
Beran, R. . . . . 363  
Bickel, P. J. . . . . 406, 409, 424, 427, 428  
Billingsley, P. . . . . 2, 3, 5, 75–77, 274, 425  
Birgé, L. . . . . 275, 424, 425  
Birman, M. S. . . . . 271  
Blum, J. R. . . . . 270, 425, 426  
Blumberg, H. . . . . 75  
Bolthausen, E. . . . . 271  
Borell, C. . . . . 465

Bretagnolle, J.	466
Bronštein, E. M.	163, 271
Brown, L. D.	248, 275, 442
Cantelli, F. P.	270
Carl, B.	271
Červonenkis, A. Ya.	270, 272, 274
Chibisov, D. M.	3, 76
Chow, Y. S.	75, 245
Chung, K. L.	456
Cohn, D. L.	47, 54, 76
Cramér, H.	424
Csörgő, M.	427
Csörgő, S.	425
Dehardt, J.	270
Deheuvels, P.	426
Dehling, H.	457, 466
Devroye, L.	249
Dobrić, V.	76
Donsker, M. D.	75, 225, 274, 425
Doob, J. L.	75, 274, 448
Doss, H.	427
Dudley, R. M.	3, 10, 22, 75–78, 146, 150–152, 163, 234, 248, 269–275, 352, 379, 398, 425, 427, 465
Dugue, D.	426
Durst, M.	425
Dvoretzky, A.	248, 272, 274
Dykstra, R. L.	298
Eames, W.	75
Eaton, M. L.	466
Efron, B.	280
Erdős, P.	274
Fenstad, G.	231
Fernandez, P. J.	425
Fernique, X.	273, 466
Fomin, S. V.	76
Fortet, R.	77
Frankl, P.	270
Gaenssler, P.	75, 76
Gill, R. D.	427
Giné, E.	76, 236, 270, 271, 273, 274, 281, 425, 426

- |  |   |
|--|---|
| Glivenko, V.                           | 270   |
| Gnedenko, B. V.                        | 77  |
| Goodman, V.                            | 443, 466  |
| Greenwood, P. E.                       | 411   |
| Grenander, U.                          | 425   |
| Groeneboom, P.                         | 423   |
| Grübel, R.                             | 427   |
| Hájek, J.                              | 406, 409, 427, 458  |
| Hammersley, J. M.                      | 77  |
| Hanson, D. L.                          | 425   |
| Has'minskii, R. Z.                     | 423, 424  |
| Haussler, D.                           | 270   |
| Heesterman, C. C.                      | 427   |
| Hjort, N. L.                           | 231   |
| Hoeffding, W.                          | 426, 436, 459, 460  |
| Hoffmann-Jørgensen, J.                 | 75, 76, 270, 424, 465   |
| Hogan, M. L.                           | 442, 443, 466   |
| Huang, J.                              | 424   |
| Huber, C.                              | 466   |
| Huber, P.                              | 424   |
| Ibragimov, I. A.                       | 423, 424, 465   |
| Jain, N.                               | 273, 465  |
| Jameson, J. O.                         | 23, 25, 72  |
| Johansen, S.                           | 427   |
| Jongbloed, G.                          | 425   |
| Jupp, P. E.                            | 76  |
| Kac, M.                                | 274   |
| Kahane, J-P.                           | 270   |
| Kiefer, J.                             | 248, 270, 272, 274, 426, 448, 466                                       |
| Kim, J.                                | 76, 274, 423, 447   |
| Klaassen, C. A. J.                     | 406, 409, 424, 426, 428   |
| Kolčinskii, V. I. (= Koltchinskii, V.) | 146, 270, 272   |
| Kolmogorov, A. N.                      | 77, 165, 271, 425   |
| Koul, H.                               | 273   |
| Kuelbs, J.                             | 379   |
| Landau, H.                             | 465   |
| Le Cam, L.                             | 75–77, 411, 423–425, 427  |
| Ledoux, M.                             | 93, 269, 271–274, 379, 426, 435, 436, 439, 441, 446, 449, 450, 465, 466 |
| Lehmann, E. L.                         | 313   |

Levit, B. Ya.	428
Lindvall, T.	76
Lorentz, G. G.	271
Mammen, E.	425
Marcus, M. B.	273, 465
Marshall, A. W.	436
Mason, D. M.	426, 466
Massart, P.	272, 274, 275, 424, 425
May, L. E.	75
Millar, P. W.	363, 428
Mitoma, I.	76
Mourier, E.	77
Müller, D. W.	231, 274
Munkres, J.	65
Murphy, S. A.	424
Nakamura, M.	271
Newton, M.	426
Nolan, D.	427
Olkin, I.	436
Ossiander, M.	270, 273
Ottaviani, G.	465
Pajor, A.	270, 271
Parthasarathy, K. R.	75, 76
Perez, F. J.	271
Philipp, W.	75, 77, 274, 379
Pinelis, I.	466
Pisier, G.	269, 271, 465
Pitts, S. M.	427
Pollard, D.	48, 75, 76, 102, 270–274, 423, 424, 447
Praestgaard, J.	426
Prakasa Rao, B. L. S.	423
Preston, C.	466
Prohorov, Yu. V.	75, 76, 77
Pyke, R.	3, 77, 274, 425
Quiroz, A. J.	271
Ranga Rao, R.	77
Reeds, J. A.	427
Révész, P.	427
Revuz, D.	384

- Ritov, Y. . . . . 406, 409, 424, 428  
 Robertson, T. . . . . 298  
 Rosenblatt, J. I. . . . . 425  
 Rosenblatt, M. . . . . 426  
 Rudin, W. . . . . 319, 377, 422  
 Runnenburg, J. Th. . . . . 283
- Samorodnitsky, G. . . . . 442, 445, 465  
 Sauer, N. . . . . 270  
 Severini, T. A. . . . . 425  
 Sheehy, A. . . . . 272, 274  
 Shen, X. . . . . 425  
 Shepp, L. A. . . . . 465  
 Shirayev, A. N. . . . . 411  
 Shorack, G. R. . . . . 3, 102, 215, 249, 250, 273, 425, 427, 448, 460, 466  
 Šidák, Z. . . . . 406, 409, 427  
 Siegmund, D. . . . . 442, 443, 444, 466  
 Skorokhod, A. V. . . . . 3, 75, 77  
 Slepian, D. . . . . 465  
 Smith, D. . . . . 248, 275  
 Smolyanov, O. G. . . . . 76  
 Solomjak, M. Z. . . . . 271  
 Spurr, B. D. . . . . 76  
 Stengle, G. . . . . 271  
 Stephan, I. . . . . 271  
 Stone, C. . . . . 76  
 Strassen, V. . . . . 269  
 Strobl, F. . . . . 270  
 Sudakov, V. N. . . . . 269, 465
- Talagrand, M. . 93, 118, 236, 248, 269–275, 379, 426, 435, 436, 441, 442,  
                     446, 449, 450, 465, 466  
 Teicher, H. . . . . 75, 245  
 Tikhomirov, V. M. . . . . 165, 271  
 Topsøe, F. . . . . 77  
 Tsirel'son, V. S. . . . . 465
- Van de Geer, S. . . . . 271, 338, 424  
 Van der Laan, M. . . . . 424  
 Van der Vaart, A. W. . 76, 78, 271, 273, 406, 409, 419, 421, 424, 427, 428  
 Van Zuijlen, M. . . . . 273  
 Van Zwet, W. R. . . . . 466  
 Vapnik, V. N. . . . . 270, 272, 274  
 Varadarajan, V. S. . . . . 76  
 Vervaat, W. . . . . 283

- Wang, Y. . . . . 425  
Wellner, J. A. 76–78, 102, 215, 249, 250, 272–274, 283, 423–427, 448, 460  
Whitt, W. . . . . 76, 77  
Wichura, M. J. . . . . 76, 77, 425  
Wolfowitz, J. . . . . 248, 272, 274, 428  
Wong, W. H. . . . . 425  
Wright, F. T. . . . . 298  
  
Yor, M. . . . . 384  
Yukich, J. E. . . . . 271  
Yurinskii, V. V. . . . . 273, 457, 466  
  
Zinn, J. . . . . 236, 270–274, 281, 426

# Subject Index

abstract integral . . . . .	22
algebra . . . . .	25
alternative	
definition of $\ \cdot\ _{\psi_p}$ for small $p$ . . . . .	266
hypothesis . . . . .	360, 408
proof of Corollary 2.9.9 . . . . .	179, 187
proofs of the Donsker theorems . . . . .	243
analytic . . . . .	47
Anderson-Darling statistic . . . . .	51, 235
argmax continuous mapping . . . . .	286
Arzelà-Ascoli . . . . .	41
asymptotically	
equicontinuous uniformly in . . . . .	169
finite-dimensional . . . . .	49
independent . . . . .	31
measurable . . . . .	20
$\mathbf{B}'$ -measurable . . . . .	416
(shift) normal . . . . .	412
strongly, measurable . . . . .	55
tight . . . . .	21
uniformly integrable . . . . .	69
uniformly $\rho$ -equicontinuous in probability . . . . .	37
asymptotic and uniform tightness of a sequence . . . . .	27
asymptotic equicontinuity . . . . .	67

a.s. representations . . . . .	58, 59
Baire $\sigma$ -field . . . . .	26
ball $\sigma$ -field . . . . .	45
Bayesian bootstrap . . . . .	354
Bennett's inequality . . . . .	460
Bernstein norm . . . . .	324
Bernstein's inequality . . . . .	102, 103
between graphs . . . . .	152
block bracketing . . . . .	126
bootstrap	
Bayesian . . . . .	354
delta-method for . . . . .	378, 379, 380
empirical distribution . . . . .	345
empirical process . . . . .	345
exchangeable . . . . .	353
independence process . . . . .	369
measure . . . . .	345
model-based . . . . .	369
nonparametric . . . . .	345
samples . . . . .	345, 358
two-sample empirical measures . . . . .	365
wild . . . . .	354
without replacement . . . . .	356
Borel measure . . . . .	16, 73
Borel $\sigma$ -field . . . . .	16
Borell's inequality . . . . .	438
bounded functions . . . . .	419
bounded Lipschitz metric . . . . .	73
bounded VC-major classes . . . . .	145
bracket . . . . .	83
$\varepsilon$ -bracket . . . . .	83
bracketing	
block . . . . .	126
by Gaussian hypotheses . . . . .	212
central limit theorem . . . . .	211
entropy with . . . . .	83
number . . . . .	83
with Hellinger metric . . . . .	327
Bretagnolle and Huber-Carol inequality . . . . .	462
Brownian	
bridge . . . . .	82, 89, 443
motion . . . . .	93
sheet . . . . .	230, 443
tucked sheet . . . . .	231, 368, 444

cadlag functions . . . . .	46
canonically formed . . . . .	83
canonical representation . . . . .	27
cells in $\mathbb{R}^k$ . . . . .	129, 133, 135
central limit theorem . . . . .	50
bracketing . . . . .	130, 211, 221
bracketing by Gaussian hypotheses . . . . .	212
conditonal multiplier . . . . .	182, 183
Jain-Marcus . . . . .	213
Lindeberg-Feller . . . . .	411
multiplier . . . . .	176, 179
rank . . . . .	458
rate of convergence in multivariate . . . . .	457
chain rule . . . . .	373
chaining . . . . .	97, 131, 241
chaining method . . . . .	90
change-point alternatives . . . . .	410
classical smoothness . . . . .	313
closed convex sets . . . . .	202
comparing $\psi_p$ -norms . . . . .	105
completely regular . . . . .	48
completely regular points . . . . .	48
completely tucked Brownian sheet . . . . .	368
completeness of the bounded Lipschitz metric . . . . .	71
completion . . . . .	14, 17
conditional multiplier central limit theorem . . . . .	176, 181, 183
conditions that imply the Lindeberg-Feller condition . . . . .	411
consistency . . . . .	287
consistency and Lipschitz criterion functions . . . . .	308
consistency and concave criterion functions . . . . .	308
consistent lifting . . . . .	118
contiguity and limit experiments . . . . .	411
contiguous . . . . .	401
continuity set . . . . .	19, 60
continuous	
argmax, mapping . . . . .	286
extended, mapping . . . . .	67
functions . . . . .	34
mapping . . . . .	20, 54
contraction principle . . . . .	436, 450
convergence	
almost surely . . . . .	52
almost uniformly . . . . .	52
for nets . . . . .	4, 17, 21, 58, 411, 417

in outer probability . . . . .	52, 57
in probability . . . . .	56
outer almost surely . . . . .	52
weakly . . . . .	17
convex	
densities . . . . .	329
functions . . . . .	445
sets . . . . .	202
convolution . . . . .	414
copula function . . . . .	389
covering number . . . . .	83, 90, 98
Cramér-von Mises . . . . .	234
current status . . . . .	298
deficient two-sample bootstrap . . . . .	366
delta-method . . . . .	372, 374, 375
for bootstrap almost surely . . . . .	379, 380
for bootstrap in probability . . . . .	378
differentiable	
Fréchet . . . . .	373
functions . . . . .	154, 202
Hadamard . . . . .	373
Hadamard tangentially . . . . .	373
uniform Fréchet . . . . .	397
directed set . . . . .	4
dominated convergence . . . . .	13
Donsker	
class . . . . .	81
functional . . . . .	226
uniformly in . . . . .	168
Duhamel equation . . . . .	390, 398
efficiency . . . . .	269, 399
Egorov's theorem . . . . .	53
elliptical classes . . . . .	233, 234
empirical	
bootstrap measure . . . . .	345
bootstrap process . . . . .	345
distribution function . . . . .	82
measure . . . . .	80
process . . . . .	80
process indexed by sets . . . . .	82
quantiles . . . . .	385
quantile process . . . . .	387
entropy	

numbers . . . . .	83, 98
uniform . . . . .	81
with bracketing . . . . .	83
without bracketing . . . . .	83
envelope function . . . . .	84
essential infimum . . . . .	12
estimator	
asymptotically efficient . . . . .	420
asymptotically optimal . . . . .	416
Kaplan-Meier . . . . .	392
$M$ - . . . . .	284
maximum likelihood . . . . .	286, 288, 296, 315, 316, 322, 326
Nelson-Aalen . . . . .	365
regular . . . . .	413
sieved $M$ - . . . . .	306
$Z$ - . . . . .	284, 309, 397
extended continuous mapping theorem . . . . .	67
$E^*T$ is not always $ET^*$ . . . . .	12
formally defining bootstrap samples . . . . .	345, 358
forward and backward equations . . . . .	379
Fréchet-differentiable . . . . .	373
Fredholm operator . . . . .	319
Fubini's theorem . . . . .	11
functional Donsker class . . . . .	226
Gaussian . . . . .	40, 212
Gaussian-dominated . . . . .	212
Gaussianization . . . . .	194
generator . . . . .	439
Glivenko-Cantelli class . . . . .	81
Glivenko-Cantelli, uniformly in . . . . .	167
Hadamard and Fréchet differentiability . . . . .	373
Hadamard-differentiable . . . . .	372
Hadamard-differentiable tangentially . . . . .	373
half-spaces . . . . .	152
Hamel base . . . . .	54
Hamming metric . . . . .	137
Hausdorff distance . . . . .	162
Hellinger metric . . . . .	280, 327
hereditary . . . . .	136
Hoeffding's inequality . . . . .	100, 266, 436, 459
Hoffmann-Jørgensen inequalities . . . . .	432
Hoffmann-Jørgensen inequalities for moments . . . . .	433

independence empirical process . . . . .	367
i.i.d. observations . . . . .	208, 413
inequality	
Bennett's . . . . .	460
Bernstein's . . . . .	102, 103
Borell's . . . . .	438
Bretagnolle and Huber-Carol . . . . .	462
Hoeffding's exponential . . . . .	100, 266, 439
Hoeffding's finite sampling . . . . .	436
Hoffmann-Jørgensen's . . . . .	432
Hoffmann-Jørgensen's for moments . . . . .	433
isoperimetric . . . . .	184, 451
Khinchine's . . . . .	464
Kiefer's . . . . .	460
Le Cam's Poissonization . . . . .	343
Lévy's . . . . .	431
Mason and van Zwet's . . . . .	463
maximal . . . . .	90
multiplier . . . . .	177, 352
Ottaviani's . . . . .	430
Pinelis's . . . . .	463
Sudakov's . . . . .	441
symmetrization . . . . .	108
Talagrand's . . . . .	460
inner integral . . . . .	6
inner probability . . . . .	6
inner regular . . . . .	26
integration by parts for the Duhamel equation . . . . .	398
intrinsic semimetric . . . . .	91
isoperimetric inequalities . . . . .	184, 451
Jain-Marcus theorem . . . . .	213
Kac empirical point process . . . . .	341
Kaplan-Meier . . . . .	392
Kaplan-Meier estimator . . . . .	392
Khinchine's inequality . . . . .	464
Kiefer process . . . . .	226, 444
Kiefer's inequality . . . . .	460
Kolmogorov statistic . . . . .	360
$k$ -sample problem . . . . .	366
law of large numbers . . . . .	456
uniform in $P \in \mathcal{P}$ . . . . .	456

strong . . . . .	456
least-absolute-deviation regression . . . . .	305
Le Cam's Poissonization lemma . . . . .	343
Le Cam's third lemma . . . . .	404, 405
Lévy-Ito-Nisio . . . . .	431
Lévy's inequalities . . . . .	431
lifting . . . . .	118
likelihood ratios . . . . .	402
Lipschitz functions . . . . .	74, 163, 312
Lipschitz in parameter . . . . .	294, 305
Lipschitz transformations . . . . .	197, 198
locally asymptotically normal . . . . .	413
lognormality . . . . .	404
$L_{2,1}$ -condition . . . . .	88
majorizing measure . . . . .	445, 446
marginals . . . . .	35
Mason and van Zwet's inequality . . . . .	463
maximal inequalities . . . . .	90
maximal measurable minorant . . . . .	7
maximum likelihood . . . . .	286, 288, 296, 315, 316, 322, 326
measurable	
asymptotically . . . . .	20
asymptotically $\mathbf{B}'$ - . . . . .	416
asymptotically strongly . . . . .	55
class . . . . .	110
cover function . . . . .	6
majorant . . . . .	6
minorant . . . . .	7
processes with Suslin index set . . . . .	47
stochastic process . . . . .	47
measure	
bootstrap empirical . . . . .	345
Borel . . . . .	16, 73
discrete . . . . .	23
empirical . . . . .	80
-like . . . . .	209
majorizing . . . . .	445
outer measure . . . . .	14
pooled empirical . . . . .	361
Radon . . . . .	26
regular . . . . .	26
trace . . . . .	14
two-sample bootstrap empirical . . . . .	365
two-sample permutation empirical . . . . .	362

median and median deviation . . . . .	314
metric	
Bernstein . . . . .	324
bounded Lipschitz . . . . .	73
Hamming . . . . .	137
Hausdorff . . . . .	162
Hellinger . . . . .	280, 327
intrinsic semi- . . . . .	91
$L_r$ . . . . .	84
$L_{2,1}$ . . . . .	177
$L_{2,\infty}$ . . . . .	130
Prohorov . . . . .	456
semi- . . . . .	38, 39
space . . . . .	16
Skorokhod . . . . .	3
uniform . . . . .	34
$M$ -estimators . . . . .	270
minimal measurable majorant . . . . .	7
minimax theorem . . . . .	417
monotone	
convergence . . . . .	13
convergence for Orlicz norms . . . . .	105
densities . . . . .	296, 329
functions . . . . .	202
processes . . . . .	215
multiplier	
central limit theorem . . . . .	176, 179, 182, 183
inequalities . . . . .	177, 352
multipliers	
exchangeable . . . . .	353
Gaussian . . . . .	194
Rademacher . . . . .	107, 111, 112, 123, 127
Poisson . . . . .	346, 352, 363
multivariate trimmed mean . . . . .	395
necessity of integrability of the envelope . . . . .	120
Nelson-Aalen estimator . . . . .	383
net convergence . . . . .	4, 17, 21, 58, 411, 417
nontrivial . . . . .	58
norm	
Bernstein . . . . .	324
Orlicz . . . . .	95, 244
uniform . . . . .	34
$L_{2,1}$ . . . . .	177
$L_{2,\infty}$ . . . . .	130

open and closed subgraphs . . . . .	146
Orlicz norm . . . . .	95
Ottaviani's inequality . . . . .	430
outer	
integral . . . . .	6
measure . . . . .	14
probability . . . . .	6
regular . . . . .	26
packing number . . . . .	98
parametric maximum likelihood . . . . .	288
partial-sum processes . . . . .	88
partitioning . . . . .	262
<i>P</i> -Brownian bridge . . . . .	82
<i>P</i> -completion . . . . .	14
<i>P</i> -Donsker . . . . .	81
<i>P</i> -Kac . . . . .	342
Peano series . . . . .	398
perfect . . . . .	10
permutation	
empirical process . . . . .	362
independence process . . . . .	371
test . . . . .	364
picks out . . . . .	85
Pinelis's inequality . . . . .	463
<i>P</i> -measurable class . . . . .	110
pointwise	
measurable classes . . . . .	110
separable class . . . . .	110, 116
separable processes . . . . .	46
separable version . . . . .	116
Polish . . . . .	17
polynomial classes . . . . .	86
pooled empirical measure . . . . .	361
portmanteau theorem . . . . .	18
power of the Kolmogorov-Smirnov test . . . . .	408
<i>P</i> -pre-Gaussian . . . . .	89
pre-Gaussian . . . . .	89
pre-Gaussian uniformly in . . . . .	169
pre-tight . . . . .	17
process	
bootstrap empirical . . . . .	345
empirical . . . . .	80
empirical, indexed by sets . . . . .	80

empirical quantile . . . . .	387
Gaussian . . . . .	40, 212
independence empirical . . . . .	367
Kac empirical point . . . . .	341
Kiefer-Müller . . . . .	226, 340, 444
measurable stochastic . . . . .	47
measurable, with Suslin index set . . . . .	47
partial-sum . . . . .	88
permutation empirical . . . . .	362
permutation independence . . . . .	371
pointwise-separable . . . . .	46
Poisson . . . . .	342
Rademacher . . . . .	101, 449
sequential empirical . . . . .	88, 225
stochastic . . . . .	34
sub-Gaussian . . . . .	91, 101
two-sample bootstrap empirical . . . . .	365
two-sample permutation empirical . . . . .	362
weighted empirical . . . . .	210
product	
integral . . . . .	390
space . . . . .	29
topology . . . . .	29
Prohorov distance . . . . .	456
Prohorov's theorem . . . . .	21
projection $\sigma$ -field . . . . .	34
properties of the ball $\sigma$ -field . . . . .	46
quasiperfect . . . . .	10
quantile-quantile transformation . . . . .	398
Rademacher . . . . .	100
Rademacher process . . . . .	101, 449
Radon measure . . . . .	26
rank central limit theorem . . . . .	458
rate of convergence . . . . .	289, 322
regular . . . . .	413
regularity of Borel measures . . . . .	26
relatively compact . . . . .	23, 73
$\rho_P$ -totally bounded uniformly . . . . .	169
sample path . . . . .	34
score function . . . . .	413
semicontinuous	

lower . . . . .	18, 416
upper . . . . .	18, 286
separable . . . . .	17, 98, 115
Banach space . . . . .	418
modification . . . . .	119
version . . . . .	116
$\varepsilon$ -separated . . . . .	98
separates points . . . . .	25
sequences . . . . .	232
sequential empirical process . . . . .	88, 225
shatter . . . . .	85, 135
sieved $M$ -estimators . . . . .	321
$\sigma$ -field	
Baire . . . . .	26
ball . . . . .	44
Borel . . . . .	16
Skorohod space . . . . .	3, 46, 419
Slepian, Fernique, Marcus, and Shepp lemma . . . . .	441
Slutsky's lemma . . . . .	32
Slutsky's theorem . . . . .	32
smooth functions . . . . .	202
space	
metric . . . . .	16
Banach . . . . .	91, 376, 413
Hilbert . . . . .	49
normed . . . . .	372
Skorohod . . . . .	3, 46, 419
topological vector . . . . .	372
stability of the Glivenko-Cantelli property . . . . .	120
standard Brownian sheet . . . . .	231
statistic	
Anderson-Darling . . . . .	51, 235
Cramér-von Mises . . . . .	234
Kolmogorov-Smirnov . . . . .	234, 408
two-sample Kolmogorov . . . . .	360
Watson . . . . .	235
Wilcoxon . . . . .	382
stochastic process . . . . .	34
strongly asymptotically measurable . . . . .	55
subconvex . . . . .	416
sub-Gaussian . . . . .	91, 101
subgraph . . . . .	141
Sudakov's inequality . . . . .	441
Suslin . . . . .	47
symmetric convex hull . . . . .	87, 142

symmetrization . . . . .	108
symmetrization for probabilities . . . . .	112
Talagrand's inequality . . . . .	460
taking the supremum over all $Q$ . . . . .	133
tangent set . . . . .	413
$\tau(\mathbf{B}')$ -subconvex . . . . .	416
tight . . . . .	16
time reversal . . . . .	231
topological vector space . . . . .	372
total boundedness in $L_2(P)$ . . . . .	90
totally bounded . . . . .	17
trace measure . . . . .	15
truncation . . . . .	208
tucked Brownian sheet . . . . .	444
two-sample	
bootstrap empirical measures . . . . .	365
Kolmogorov-Smirnov statistic . . . . .	360
permutation empirical measures . . . . .	362
two-sided Brownian motion . . . . .	295
two-sided contiguous . . . . .	402
uniform distance . . . . .	34
uniform entropy condition . . . . .	127
uniform entropy numbers . . . . .	84
uniform Fréchet differentiability . . . . .	397
uniform small variance exponential bound . . . . .	257
uniformly tight . . . . .	21, 27, 73
uniform tightness and weak compactness of a set of Borel measures .	71
universally Donsker . . . . .	82
universally measurable . . . . .	47
using $\ \cdot\ _{P,2}$ instead of $\rho_P$ . . . . .	93
Vapnik-Červonenkis or VC	
-class . . . . .	85, 86, 135, 141
-class of sets . . . . .	85
-hull class . . . . .	87, 142
-hull class for sets . . . . .	142
-index . . . . .	85, 135
-major class . . . . .	145
-subgraph class . . . . .	86, 141
vector lattice . . . . .	25
Volterra equation . . . . .	398
weak convergence . . . . .	17

weak convergence of discrete measures . . . . .	24
weighted bootstrap empirical measure . . . . .	353
weighted empirical distribution function . . . . .	214
weighted empirical processes . . . . .	210
Wilcoxon statistic . . . . .	382
wild bootstrap . . . . .	354
<i>Z</i> -estimators . . . . .	284, 309, 397
Zermelo-Frankel . . . . .	24

# List of Symbols

$\mathbf{B}^*$	dual space of the Banach space $\mathbf{B}$	414
$\text{BL}(\mathbb{D})$	the set of bounded, Lipschitz functions on $\mathbb{D}$	73
$C_b(\mathbb{D})$	bounded, continuous, real functions on $\mathbb{D}$	16
$C(T)$	continuous functions from $T$ to $\mathbb{R}$	34
$C_M^\alpha(\mathcal{X})$	continuous functions from $\mathcal{X} \subset \mathbb{R}^d$ to $\mathbb{R}$ with $\ f\ _\alpha \leq M$	154
$\text{conv}(\mathcal{F})$	convex hull of $\mathcal{F}$	142
$\text{cov}(X)$	covariance matrix of a random vector $X$	181
$\delta B$	boundary of the set $B$	19
$(\mathbb{D}, d), (\mathbb{E}, e)$	metric spaces	16
$D[a, b]$	cadlag functions on $[a, b]$	46
$\delta_x$	point mass at $x$ or Dirac measure	80
$E^*$	outer integral	6
$\mathcal{F}, \mathcal{G}, \mathcal{H}$	collections of functions	86, 87
$\mathbb{F}_n$	empirical distribution function	51
$\mathbb{G}$	$P$ -Brownian bridge process	80, 81
$\mathbb{G}_n$	empirical process	78
$\mathbb{H}$	real Hilbert space	49
$\ell^\infty(T)$	set of all uniformly bounded real functions on $T$	34
$\ell^\infty(T_1, T_2, \dots)$	set of all locally bounded real functions on $T = \cup_{i=1}^\infty T_i$	43
$\mathcal{L}_r$	$r$ -integrable functions	83, 84
$L_r$	equivalence classes of $r$ -integrable functions	83, 84
$\mathcal{L}_\infty$	essentially bounded functions	118
$L_\infty$	equivalence classes of essentially bounded functions	118

$\mathcal{L}(\cdot)$	law or distribution of a random variable . . . . .	417
$\lambda$	Lebesgue measure . . . . .	3, 141
lin	linear span . . . . .	50
${}^2 \log m$	logarithm base 2 . . . . .	185
$N(\mu, \sigma^2)$	normal (Gaussian) distribution on $\mathbb{R}$ . . . . .	81, 194
$N_k(\mu, \Sigma)$	normal (Gaussian) distribution on $\mathbb{R}^k$ . . . . .	81
$P, Q, H$	probability measures on underlying sample space $\mathcal{X}$ .	6, 361
$\mathbb{P}_n, \mathbb{Q}_n$	empirical measures . . . . .	80, 361
$P^*$	outer probability . . . . .	6
$\mathcal{P}$	collection of probability measures . . . . .	9
$\mathbb{Q}$	set of rational numbers . . . . .	17
$\mathbb{R}$	real numbers . . . . .	7
$\bar{\mathbb{R}}$	extended real numbers, $[-\infty, \infty]$ . . . . .	6
$\mathbb{R}^k$	$k$ -dimensional Euclidean space . . . . .	20
$\rho, \rho_0$	semimetrics on a set $T$ or $\mathcal{F}$ . . . . .	37, 39
$\rho$	a lifting . . . . .	118
$\rho_P$	variance semimetric . . . . .	89
$sconv(\mathcal{F})$	symmetric convex hull of $\mathcal{F}$ . . . . .	142, 191
$T$	index set for a stochastic process . . . . .	34
$(\bar{T}, \rho)$	$\rho$ -completion of $T$ . . . . .	40
$T^*$	minimal measurable majorant of the map/function $T$ , or measurable cover function of $T$ . . . . .	6
$UC(T, \rho)$	uniformly continuous functions from $(T, \rho)$ to $\mathbb{R}$ . . . . .	41
$\text{var}(X)$	variance of a real random variable $X$ . . . . .	102
$\sigma(X)$	standard deviation of a real random variable $X$ . . . . .	194
$\Omega, \Omega_n$	sample spaces . . . . .	6, 17

## Symbols not connected to Greek or Roman letters

$\ll$	absolutely continuous with respect to . . . . .	9
$\kappa^*$	adjoint of the linear map $\kappa$ . . . . .	444
$\approx$	approximately equals . . . . .	300
$\#$	cardinality of a finite set . . . . .	80, 82, 99
$\times$	Cartesian product . . . . .	29
$\triangleleft$	contiguous with respect to . . . . .	402
$\xrightarrow{\text{as}}$	convergence almost surely . . . . .	52
$\xrightarrow{\text{au}}$	convergence almost uniformly . . . . .	52
$\xrightarrow{P^*}$	convergence in outer probability . . . . .	52
$\xrightarrow{P}$	convergence in probability . . . . .	52, 57
$\rightarrow$	convergence of real numbers . . . . .	2
$\xrightarrow{\text{as*}}$	convergence outer almost surely . . . . .	52
■	end-of-proof symbol . . . . .	7
$\emptyset$	empty set . . . . .	14
$\sim$	equal in distribution to . . . . .	426
$\equiv$	equals by definition . . . . .	87
$\mapsto$	function specifier . . . . .	6

$\underline{x}$	greatest integer less than or equal to $x$	154
$1_A$	indicator function of the set $A$	7
$\lesssim$	left side bounded by a constant times the right side	128
$\sim$	left side bounded above and below by constants times the right side	130
$\ f\ _{Q,r}$	$L_r(Q)$ norm of $f$	84
$\ \xi\ _{2,1}$	$L_{2,1}$ norm of the random variable $\xi$	177
$\vee$	maximum	25, 147
$\wedge$	minimum	7, 25, 147
$\ X\ _\psi$	Orlicz norm of the random variable $X$	95
$\square$	pairwise intersections of two classes of sets	147
$\sqcup$	pairwise unions of two classes of sets	147
$\ f\ _\infty$	supremum norm of $f$	73
$\triangle$	symmetric difference	150
$\bowtie$	two-sided contiguous	402
$\ \cdot\ _{\mathcal{F}}$	uniform norm for maps from $\mathcal{F}$ to $\mathbb{R}$	34, 81
$\rightsquigarrow$	weak convergence	18
$\overset{h}{\rightsquigarrow}$	weak convergence under $P_{n,h}$	412
$\overset{a}{\rightsquigarrow}$	weak convergence under $P_{n,h}$ for $h = \sum a_i h_i$	414
$\overset{Q_\alpha}{\rightsquigarrow}$	weak convergence under $Q_\alpha$	403

# **Springer Series in Statistics**

---

(continued from p. ii)

*Read/Cressie*: Goodness-of-Fit Statistics for Discrete Multivariate Data.

*Reinsel*: Elements of Multivariate Time Series Analysis.

*Reiss*: A Course on Point Processes.

*Reiss*: Approximate Distributions of Order Statistics: With Applications to Non-parametric Statistics.

*Rieder*: Robust Asymptotic Statistics.

*Rosenbaum*: Observational Studies.

*Ross*: Nonlinear Estimation.

*Sachs*: Applied Statistics: A Handbook of Techniques, 2nd edition.

*Särndal/Swensson/Wretman*: Model Assisted Survey Sampling.

*Schervish*: Theory of Statistics.

*Seneta*: Non-Negative Matrices and Markov Chains, 2nd edition.

*Shao/Tu*: The Jackknife and Bootstrap.

*Siegmund*: Sequential Analysis: Tests and Confidence Intervals.

*Simonoff*: Smoothing Methods in Statistics.

*Tanner*: Tools for Statistical Inference: Methods for the Exploration of Posterior Distributions and Likelihood Functions, 2nd edition.

*Tong*: The Multivariate Normal Distribution.

*van der Vaart/Wellner*: Weak Convergence and Empirical Processes: With Applications to Statistics.

*Vapnik*: Estimation of Dependences Based on Empirical Data.

*Weerahandi*: Exact Statistical Methods for Data Analysis.

*West/Harrison*: Bayesian Forecasting and Dynamic Models.

*Wolter*: Introduction to Variance Estimation.

*Yaglom*: Correlation Theory of Stationary and Related Random Functions I: Basic Results.

*Yaglom*: Correlation Theory of Stationary and Related Random Functions II: Supplementary Notes and References.